



(12) **DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION**

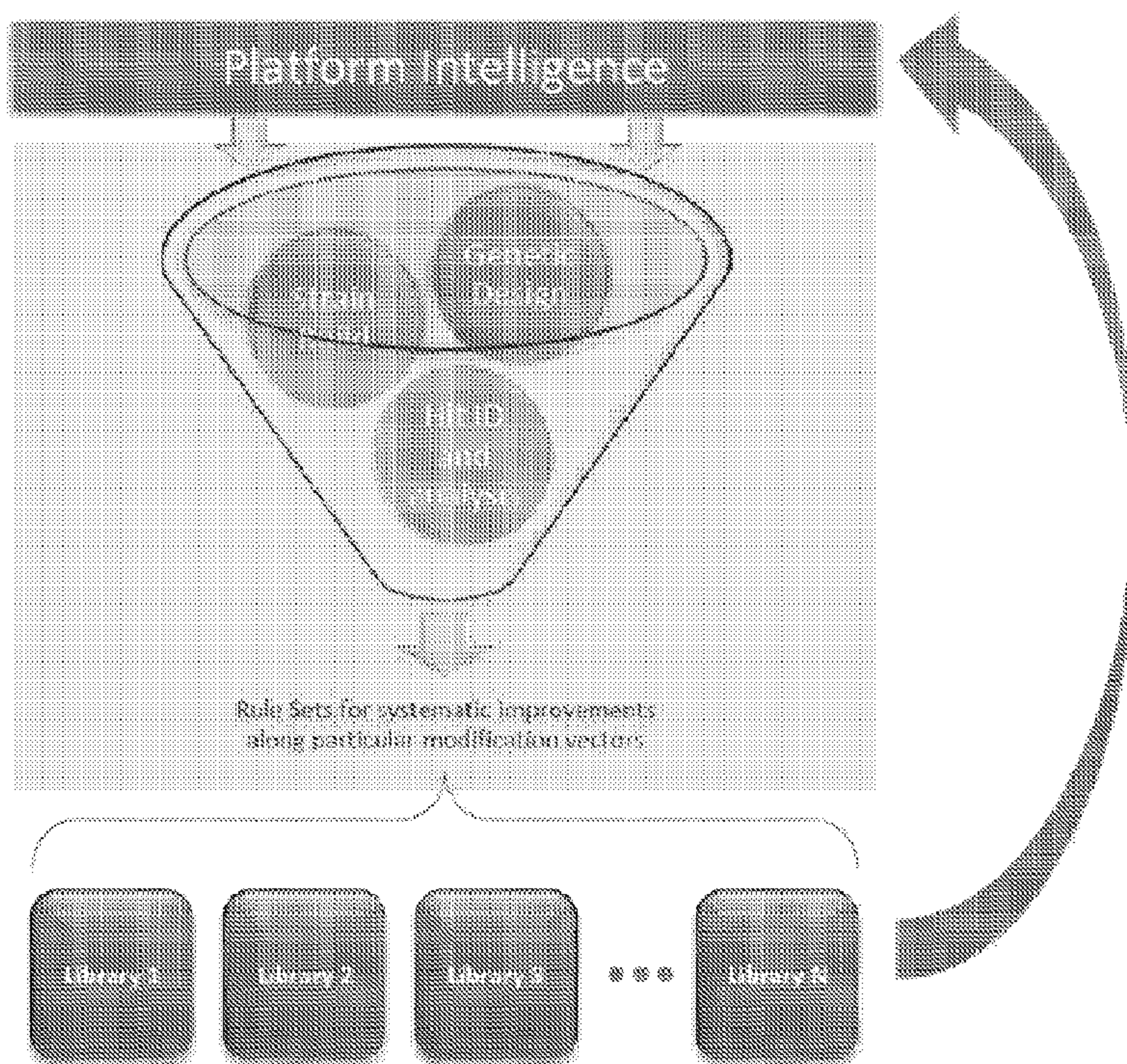
(13) **A1**

(86) Date de dépôt PCT/PCT Filing Date: 2018/06/06
 (87) Date publication PCT/PCT Publication Date: 2018/12/13
 (85) Entrée phase nationale/National Entry: 2019/11/21
 (86) N° demande PCT/PCT Application No.: US 2018/036333
 (87) N° publication PCT/PCT Publication No.: 2018/226880
 (30) Priorité/Priority: 2017/06/06 (US62/515,870)

(51) Cl.Int./Int.Cl. *C12N 15/10* (2006.01),
C12N 15/73 (2006.01)
 (71) Demandeur/Applicant:
ZYMERGEN INC., US
 (72) Inventeurs/Inventors:
DAVIS, MATTHEW, US;
WISNEWSKI, CHRISTY, US;
WESTFALL, PATRICK, US;
SERBER, ZACH, US;
DEAN, ERIK JEDEDIAH, US;
MANCHESTER, SHAWN, US;
GORA, KATHERINE, US;
...
 (74) Agent: ROBIC

(54) Titre : PLATE-FORME D'INGENIERIE GENOMIQUE HTP PERMETTANT D'AMELIORER ESCHERICHIA COLI
 (54) Title: A HTP GENOMIC ENGINEERING PLATFORM FOR IMPROVING ESCHERICHIA COLI

FIGURE 5



(57) **Abrégé/Abstract:**

The present disclosure provides a HTP genomic engineering platform for improving Escherichia coli. that is computationally driven and integrates molecular biology, automation, and advanced machine learning protocols. This integrative platform utilizes a suite of

(72) **Inventeurs(suite)/Inventors(continued)**: SHELLMAN, ERIN, US; KIMBALL, AARON, US; SZYJKA, SHAWN, US; FREWEN, BARBARA, US; TREYNOR, THOMAS, US; FLASHMAN, MICHAEL, US; HAUSHALTER, ROBERT, US; MORGAN, STACY-ANNE, US; BLAISSE, MICHAEL, US; RAMAKRISHNAN, PRABHA, US; ROTHSCHILD-MANCINELLI, KYLE, US; KIM, YOUNGNYUN, US

(57) **Abrégé(suite)/Abstract(continued)**:

HTP molecular tool sets to create HTP genetic design libraries, which are derived from, inter alia, scientific insight and iterative pattern recognition.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau(10) International Publication Number
WO 2018/226880 A1(43) International Publication Date
13 December 2018 (13.12.2018)

(51) International Patent Classification:

C12N 15/10 (2006.01) C12N 15/73 (2006.01)

Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(21) International Application Number:

PCT/US2018/036333

(72) Inventors: SERBER, Zach; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(22) International Filing Date:

06 June 2018 (06.06.2018)

DEAN, Erik Jedediah; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(25) Filing Language:

English

MANCHESTER, Shawn; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(26) Publication Language:

English

GORA, Katherine; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(30) Priority Data:

62/515,870 06 June 2017 (06.06.2017) US

SHELLMAN, Erin; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(71) Applicant: ZYMERGEN INC. [US/US]; 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

KIMBALL, Aaron; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(72) Inventors; and

(71) Applicants: DAVIS, Matthew [US/US]; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). WISNEWSKI, Christy [US/US]; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). WESTFALL, Patrick [US/US]; c/o

SZYJKA, Shawn; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

FREWEN, Barbara; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

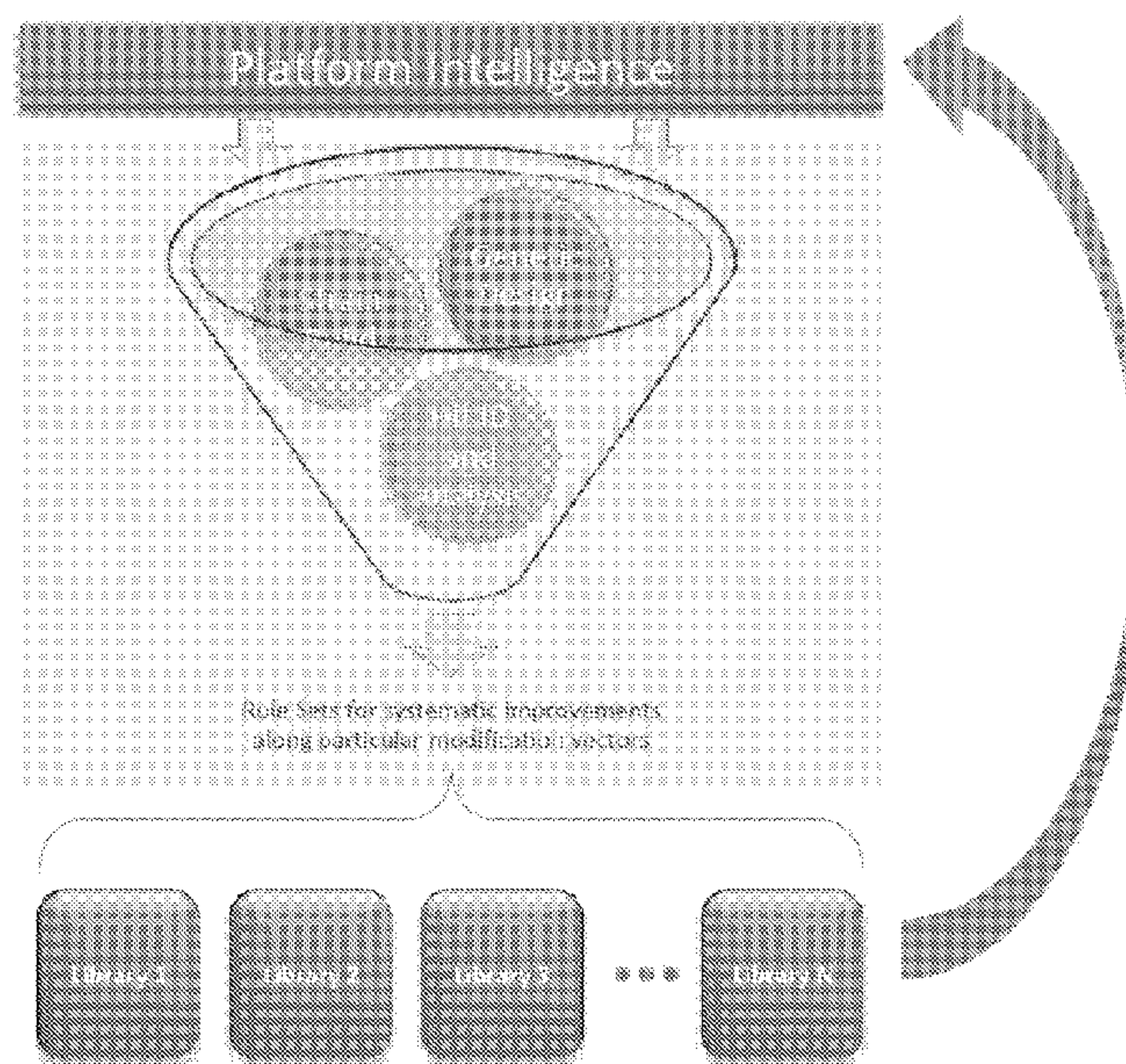
TREYNOR, Thomas; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

FLASHMAN, Michael; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

HAUSHALTER, Robert; c/o Zymergen

(54) Title: A HTP GENOMIC ENGINEERING PLATFORM FOR IMPROVING ESCHERICHIA COLI

FIGURE 5

(57) Abstract: The present disclosure provides a HTP genomic engineering platform for improving *Escherichia coli*. that is computationally driven and integrates molecular biology, automation, and advanced machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets to create HTP genetic design libraries, which are derived from, *inter alia*, scientific insight and iterative pattern recognition.

[Continued on next page]



WO 2018/226880 A1

WO 2018/226880 A1

Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). **MORGAN, Stacy-Anne**; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). **BLAISSE, Michael**; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). **RAMAKRISHNAN, Prabha**; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). **ROTHSCHILD-MANCINELLI, Kyle**; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US). **KIM, Youngnyun**; c/o Zymergen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).

(74) **Agent: HOLLY, David Christopher** et al.; Cooley LLP, 1299 Pennsylvania Avenue, N.W., Suite 700, Washington, District of Columbia 20004 (US).

(81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
- *with sequence listing part of description (Rule 5.2(a))*

**IN THE UNITED STATES PATENT & TRADEMARK
RECEIVING OFFICE**

INTERNATIONAL PCT PATENT APPLICATION

**A HTP GENOMIC ENGINEERING PLATFORM FOR IMPROVING ESCHERICHIA
COLI**

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority to U.S. Provisional Application Serial No. 62/515,870, filed June 6, 2017, which is herein incorporated by reference in its entirety for all purposes.

FIELD

[0002] The present disclosure is directed to high-throughput (HTP) microbial genomic engineering for *Escherichia coli*. The disclosed HTP genomic engineering platform is computationally driven and integrates molecular biology, automation, and advanced machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets to create HTP genetic design libraries, which are derived from, *inter alia*, scientific insight and iterative pattern recognition.

STATEMENT REGARDING SEQUENCE LISTING

[0003] The Sequence Listing associated with this application is provided in text format in lieu of a paper copy, and is hereby incorporated by reference into the specification. The name of the text file containing the Sequence Listing is ZYMR_012_01WO_SeqList_ST25.txt. The text file is \approx 127 KB, was created on June 6, 2018, and is being submitted electronically *via* EFS-Web.

BACKGROUND

[0004] Humans have been harnessing the power of microbial cellular biosynthetic pathways for millennia to produce products of interest, the oldest examples of which include alcohol, vinegar, cheese, and yogurt. These products are still in large demand today and have also been accompanied by an ever increasing repertoire of products producible by microbes. The advent of genetic

engineering technology has enabled scientists to design and program novel biosynthetic pathways into a variety of organisms to produce a broad range of industrial, medical, and consumer products. Indeed, microbial cellular cultures are now used to produce products ranging from small molecules, antibiotics, vaccines, insecticides, enzymes, fuels, and industrial chemicals.

[0005] Given the large number of products produced by modern industrial microbes, it comes as no surprise that engineers are under tremendous pressure to improve the speed and efficiency by which a given microorganism is able to produce a target product. A variety of approaches have been used to improve the economy of biologically-based industrial processes by “improving” the microorganism involved. For example, many pharmaceutical and chemical industries rely on microbial strain improvement programs in which the parent strains of a microbial culture are continuously mutated through exposure to chemicals or UV radiation and are subsequently screened for performance increases, such as in productivity, yield and titer. This mutagenesis process is extensively repeated until a strain demonstrates a suitable increase in product performance. The subsequent “improved” strain is then utilized in commercial production. The identification of improved industrial microbial strains through mutagenesis is time consuming and inefficient. The process, by its very nature, is haphazard and relies on stumbling upon a mutation that has a desirable outcome on product output. Not only are traditional microbial strain improvement programs inefficient, but the process can also lead to industrial strains with a high degree of detrimental mutagenic load. The accumulation of mutations in industrial strains subjected to these types of programs can become significant and may lead to an eventual stagnation in the rate of performance improvement.

[0006] Perhaps there is no better example of the stagnation resultant from traditional strain improvement programs than with *E. coli*, which is one of the most engineered microbial host systems in existence. The microbe has been subjected to the aforementioned traditional methods of microbial strain improvement for decades. Despite the vast amount of effort that has been devoted to engineering *E. coli*, the microbe still possess an enormous amount of untapped potential. This is because *E. coli* presents unique challenges for researchers attempting to improve the microbe for production purposes. These challenges have hampered the field of genomic engineering in *E. coli* and prevented researchers from harnessing the full potential of this microbial system.

[0007] In particular, the industry has not yet developed a high-throughput system for genomic engineering in *E. coli*. It is clear that traditional methods of strain improvement have reached a plateau with respect to this organismal system, but yet researchers do not have the genomic engineering tools that are needed to traverse this plateau.

[0008] Thus, there is a great need in the art for new methods of engineering *E. coli* for production purposes, which do not suffer from the aforementioned drawbacks inherent with traditional strain improvement programs. Specifically, a high-throughput system for discovering and consolidating beneficial mutations in *E. coli* would revolutionize the field and allow researchers to tap the full potential of this organism.

SUMMARY OF THE DISCLOSURE

[0009] The present disclosure provides a high-throughput (HTP) genomic engineering platform for *E. coli* that does not suffer from the myriad of problems associated with traditional microbial strain improvement programs.

[0010] Further, the HTP platform taught herein is able to rehabilitate *E. coli* strains that have accumulated non-beneficial mutations through decades of random mutagenesis-based strain improvement programs.

[0011] The disclosure also provides for unique genomic engineering toolsets and procedures, which undergird the HTP platform's functionality in an *E. coli* system.

[0012] The disclosed HTP genomic engineering platform is computationally driven and integrates molecular biology, automation, and advanced machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets to create HTP genetic design libraries, which are derived from, *inter alia*, scientific insight and iterative pattern recognition.

[0013] The taught HTP genetic design libraries function as drivers of the genomic engineering process, by providing libraries of particular genomic alterations for testing in *E. coli*. The microbes engineered utilizing a particular library, or combination of libraries, are efficiently screened in a HTP manner for a resultant outcome, *e.g.* production of a product of interest. This process of utilizing the HTP genetic design libraries to define particular genomic alterations for testing in a microbe and then subsequently screening host microbial genomes harboring the alterations is implemented in an efficient and iterative manner. In some aspects, the iterative cycle or "rounds"

of genomic engineering campaigns can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more iterations/cycles/rounds.

[0014] Thus, in some aspects, the present disclosure teaches methods of conducting at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000 or more “rounds” of HTP genetic engineering (e.g., rounds of SNP swap, PRO swap, STOP swap, or combinations thereof) in an *E. coli* host system.

[0015] In some embodiments, the present disclosure teaches a linear approach, in which each subsequent HTP genetic engineering round is based on genetic variation identified in the previous round of genetic engineering. In other embodiments the present disclosure teaches a non-linear approach, in which each subsequent HTP genetic engineering round is based on genetic variation identified in any previous round of genetic engineering, including previously conducted analysis, and separate HTP genetic engineering branches.

[0016] The data from these iterative cycles enables large scale data analytics and pattern recognition, which is utilized by the integrative platform to inform subsequent rounds of HTP genetic design library implementation. Consequently, the HTP genetic design libraries utilized in the taught platform are highly dynamic tools that benefit from large scale data pattern recognition algorithms and become more informative through each iterative round of microbial engineering. Such a system has never been developed for *E. coli* and is desperately needed in the art.

[0017] In some embodiments, the genetic design libraries of the present disclosure comprise at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000 or more individual genetic changes (e.g., at least X number of promoter:gene combinations in the PRO swap library).

[0018] In some embodiments, the present disclosure teaches a high-throughput (HTP) method of genomic engineering to evolve an *E. coli* strain to acquire a desired phenotype, comprising: a) perturbing the genomes of an initial plurality of *E. coli* strains having the same strain background, to thereby create an initial HTP genetic design *E. coli* strain library comprising individual strains with unique genetic variations; b) screening and selecting individual strains of the initial HTP genetic design *E. coli* strain library for the desired phenotype; c) providing a subsequent plurality of *E. coli* microbes that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent HTP genetic design *E. coli* strain library; d) screening and selecting individual *E. coli* strains of the subsequent HTP genetic design *E. coli* strain library for the desired phenotype; e) repeating steps c)-d) one or more times, in a linear or non-linear fashion, until an *E. coli* strain has acquired the desired phenotype, wherein each subsequent iteration creates a new HTP genetic design *E. coli* strain library comprising individual *E. coli* strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual *E. coli* strains of a preceding HTP genetic design *E. coli* strain library.

[0019] In some embodiments, the present disclosure teaches that the initial HTP genetic design *E. coli* strain library is at least one selected from the group consisting of a promoter swap microbial strain library, SNP swap microbial strain library, start/stop codon microbial strain library, optimized sequence microbial strain library, a terminator swap microbial strain library, a protein solubility tag microbial strain library, a protein degradation tag microbial strain library or any combination thereof.

[0020] In some embodiments, the present disclosure teaches methods of making a subsequent plurality of *E. coli* strains that each comprise a unique combination of genetic variations, wherein each of the combined genetic variations is derived from the initial HTP genetic design *E. coli* strain library or the HTP genetic design *E. coli* strain library of the preceding step.

[0021] In some embodiments, the combination of genetic variations in the subsequent plurality of *E. coli* strains will comprise a subset of all the possible combinations of the genetic variations in the initial HTP genetic design *E. coli* strain library or the HTP genetic design *E. coli* strain library of the preceding step.

[0022] In some embodiments, the present disclosure teaches that the subsequent HTP genetic design *E. coli* strain library is a full combinatorial strain library derived from the genetic variations in the initial HTP genetic design *E. coli* strain library or the HTP genetic design *E. coli* strain library of the preceding step.

[0023] For example, if the prior HTP genetic design *E. coli* strain library only had genetic variations A, B, C, and D, then a partial combinatorial of said variations could include a subsequent HTP genetic design *E. coli* strain library comprising three strains with each comprising either the AB, AC, or AD unique combinations of genetic variations (order in which the mutations are represented is unimportant). A full combinatorial *E. coli* strain library derived from the genetic variations of the HTP genetic design library of the preceding step would include six microbes, each comprising either AB, AC, AD, BC, BD, or CD unique combinations of genetic variations.

[0024] In some embodiments, the methods of the present disclosure teach perturbing the genome of *E. coli* utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, or any combination thereof.

[0025] In some embodiments of the presently disclosed methods, the initial plurality of *E. coli* comprise unique genetic variations derived from an industrial production *E. coli* strain.

[0026] In some embodiments of the presently disclosed methods, the initial plurality of *E. coli* comprise industrial production *E. coli* strains denoted S₁Gen₁ and any number of subsequent microbial generations derived therefrom denoted S_nGen_n.

[0027] In some embodiments, the present disclosure teaches a method for generating a SNP swap *E. coli* strain library, comprising the steps of: a) providing a reference *E. coli* strain and a second *E. coli* strain, wherein the second *E. coli* strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, which are not present in the reference strain; b) perturbing the genome of either the reference strain, or the second strain, to thereby create an initial SNP swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations corresponds to a single genetic variation selected from the plurality of identified genetic variations between the reference strain and the second strain.

[0028] In some embodiments of a SNP swap library, the genome of the reference *E. coli* strain is perturbed to add one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the second *E. coli* strain.

[0029] In some embodiments of a SNP swap library, the genome of the second *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the reference *E. coli* strain.

[0030] In some embodiments, the genetic variations of the SNP swap library will comprise a subset of all the genetic variations identified between the reference *E. coli* strain and the second *E. coli* strain.

[0031] In some embodiments, the genetic variations of the SNP swap library will comprise all of the identified genetic variations identified between the reference *E. coli* strain and the second *E. coli* strain.

[0032] In some embodiments, the present disclosure teaches a method for rehabilitating and improving the phenotypic performance of an industrial *E. coli* strain, comprising the steps of: a) providing a parental lineage *E. coli* strain and an industrial *E. coli* strain derived therefrom, wherein the industrial strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, not present in the parental lineage strain; b) perturbing the genome of either the parental lineage strain, or the industrial strain, to thereby create an initial SNP swap *E. coli* strain library comprising a plurality of individual strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations corresponds to a single genetic variation selected from the plurality of identified genetic variations between the parental lineage strain and the industrial strain; c) screening and selecting individual strains of the initial SNP swap *E. coli* strain library for phenotype performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer said *E. coli* strains with phenotype performance improvements; d) providing a subsequent plurality of *E. coli* strains that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual strains screened in the preceding step, to thereby create a subsequent SNP swap *E. coli* strain library; e) screening and selecting individual strains of the subsequent SNP swap *E. coli* strain library for phenotype performance improvements over the reference strain, thereby identifying unique combinations of genetic variation that confer said *E. coli* strains with

additional phenotype performance improvements; and f) repeating steps d)-e) one or more times, in a linear or non-linear fashion, until a strain exhibits a desired level of improved phenotype performance compared to the phenotype performance of the industrial *E. coli* strain, wherein each subsequent iteration creates a new SNP swap *E. coli* strain library comprising individual microbial strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual microbial strains of a preceding SNP swap *E. coli* strain library.

[0033] In some embodiments, the present disclosure teaches methods for rehabilitating and improving the phenotypic performance of an industrial *E. coli* strain, wherein the genome of the parental lineage *E. coli* strain is perturbed to add one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the industrial *E. coli* strain.

[0034] In some embodiments, the present disclosure teaches methods for rehabilitating and improving the phenotypic performance of an industrial *E. coli* strain, wherein the genome of the industrial *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the parental lineage *E. coli* strain.

[0035] In some embodiments, the present disclosure teaches a method for generating a promoter swap *E. coli* strain library, said method comprising the steps of: a) providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain; b) engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations comprises one of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

[0036] In some embodiments, the present disclosure teaches a promoter swap method of genomic engineering to evolve an *E. coli* strain to acquire a desired phenotype, said method comprising the steps of: a) providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain; b) engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual

E. coli strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations comprises one of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain; c) screening and selecting individual strains of the initial promoter swap *E. coli* strain library for the desired phenotype; d) providing a subsequent plurality of *E. coli* strains that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual strains screened in the preceding step, to thereby create a subsequent promoter swap *E. coli* strain library; e) screening and selecting individual strains of the subsequent promoter swap *E. coli* strain library for the desired phenotype; f) repeating steps d)-e) one or more times, in a linear or non-linear fashion, until a microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new promoter swap *E. coli* strain library comprising individual strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual strains of a preceding promoter swap *E. coli* strain library.

[0037] In some embodiments, the present disclosure teaches a method for generating a terminator swap *E. coli* strain library, said method comprising the steps of: a) providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain; b) engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations comprises one of the target genes endogenous to the base *E. coli* strain operably linked to one or more of the terminators from the terminator ladder.

[0038] In some embodiments, the present disclosure teaches a terminator swap method of genomic engineering to evolve an *E. coli* strain to acquire a desired phenotype, said method comprising the steps of: a) providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain; b) engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations comprises one of the target

genes endogenous to the base *E. coli* strain operably linked to one or more of the terminators from the terminator ladder; c) screening and selecting individual microbial strains of the initial terminator swap *E. coli* strain library for the desired phenotype; d) providing a subsequent plurality of *E. coli* strains that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual strains screened in the preceding step, to thereby create a subsequent terminator swap *E. coli* strain library; e) screening and selecting individual strains of the subsequent terminator swap *E. coli* strain library for the desired phenotype; f) repeating steps d)-e) one or more times, in a linear or non-linear fashion, until a microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new terminator swap *E. coli* strain library comprising individual strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual strains of a preceding terminator swap *E. coli* strain library.

[0039] In some embodiments, the present disclosure teaches iteratively improving the design of candidate *E. coli* strains by (a) accessing a predictive model populated with a training set comprising (1) inputs representing genetic changes to one or more background *E. coli* strains and (2) corresponding performance measures; (b) applying test inputs to the predictive model that represent genetic changes, the test inputs corresponding to candidate *E. coli* strains incorporating those genetic changes; (c) predicting phenotypic performance of the candidate *E. coli* strains based at least in part upon the predictive model; (d) selecting a first subset of the candidate *E. coli* strains based at least in part upon their predicted performance; (e) obtaining measured phenotypic performance of the first subset of the candidate *E. coli* strains; (f) obtaining a selection of a second subset of the candidate *E. coli* strains based at least in part upon their measured phenotypic performance; (g) adding to the training set of the predictive model (1) inputs corresponding to the selected second subset of candidate *E. coli* strains, along with (2) corresponding measured performance of the selected second subset of candidate *E. coli* strains; and (h) repeating (b)-(g) until measured phenotypic performance of at least one candidate *E. coli* strain satisfies a performance metric. In some cases, during a first application of test inputs to the predictive model, the genetic changes represented by the test inputs comprise genetic changes to the one or more background *E. coli* strains; and during subsequent applications of test inputs, the genetic changes represented by the test inputs comprise genetic changes to candidate *E. coli* strains within a previously selected second subset of candidate *E. coli* strains.

[0040] In some embodiments, selection of the first subset may be based on epistatic effects. This may be achieved by: during a first selection of the first subset: determining degrees of dissimilarity between performance measures of the one or more background *E. coli* strains in response to application of a plurality of respective inputs representing genetic changes to the one or more background *E. coli* strains; and selecting for inclusion in the first subset at least two candidate *E. coli* strains based at least in part upon the degrees of dissimilarity in the performance measures of the one or more background *E. coli* strains in response to application of genetic changes incorporated into the at least two candidate *E. coli* strains.

[0041] In some embodiments, the present invention teaches applying epistatic effects in the iterative improvement of candidate *E. coli* strains, the method comprising: obtaining data representing measured performance in response to corresponding genetic changes made to at least one *E. coli* background strain; obtaining a selection of at least two genetic changes based at least in part upon a degree of dissimilarity between the corresponding responsive performance measures of the at least two genetic changes, wherein the degree of dissimilarity relates to the degree to which the at least two genetic changes affect their corresponding responsive performance measures through different biological pathways; and designing genetic changes to an *E. coli* background strain that include the selected genetic changes. In some cases, the *E. coli* background strain for which the at least two selected genetic changes are designed is the same as the at least one *E. coli* background strain for which data representing measured responsive performance was obtained.

[0042] In some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing only a single type of genetic library. For example, in some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing only SNP swap libraries. In other embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing only PRO swap libraries. In some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing only STOP swap libraries. In some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing only Start/Stop Codon swap libraries.

[0043] In other embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing two or more types of genetic libraries. For example, in some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods combining SNP swap and PRO swap libraries. In some embodiments, the present disclosure teaches HTP *E. coli* strain

improvement methods combining SNP swap and STOP swap libraries. In some embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods combining PRO swap and STOP swap libraries.

[0044] In other embodiments, the present disclosure teaches HTP *E. coli* strain improvement methods utilizing multiple types of genetic libraries (see, for example, Figure 5). In some embodiments, the genetic libraries are combined to produce combination mutations (e.g., promoter/terminator combination ladders applied to one or more genes). In yet other embodiments, the HTP *E. coli* strain improvement methods of the present disclosure can be combined with one or more traditional strain improvement methods.

[0045] In some embodiments, the HTP *E. coli* strain improvement methods of the present disclosure result in an improved *E. coli* host cell. That is, the present disclosure teaches methods of improving one or more *E. coli* host cell properties. In some embodiments the improved *E. coli* host cell property is selected from the group consisting of: volumetric productivity, specific productivity, yield or titre, of a product of interest produced by the *E. coli* host cell. In some embodiments, the improved *E. coli* host cell property is volumetric productivity. In some embodiments, the improved *E. coli* host cell property is specific productivity. In some embodiments, the improved *E. coli* host cell property is yield.

[0046] In some embodiments, the HTP *E. coli* strain improvement methods of the present disclosure result in an an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the HTP strain improvements methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween). In some embodiments, the HTP *E. coli* strain improvement methods of the present disclosure are selected from the group consisting of SNP swap, PRO swap, STOP swap, SOLUBILITY TAG swap, DEGRADATION TAG swap, and combinations thereof.

[0047] Thus, in some embodiments, the SNP swap methods of the present disclosure result in an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the SNP swap methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween).

[0048] Thus, in some embodiments, the PRO swap methods of the present disclosure result in an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the PRO swap methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween).

[0049] Thus, in some embodiments, the TERMINATOR swap methods of the present disclosure result in an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the TERMINATOR

swap methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween).

[0050] Thus, in some embodiments, the SOLUBILITY TAG swap methods of the present disclosure result in an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the SOLUBILITY TAG swap methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween).

[0051] Thus, in some embodiments, the DEGRADATION TAG swap methods of the present disclosure result in an *E. coli* host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one *E. coli* host cell property over a control *E. coli* host cell that is not subjected to the DEGRADATION TAG swap methods (e.g, an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therebetween).

[0052] In some embodiments, the present disclosure teaches a method for generating a protein solubility tag swap *E. coli* strain library, comprising the steps of: a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different solubility profiles in the base *E. coli* strain; and b. engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains,

wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

[0053] In some embodiments, the present disclosure teaches a protein solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of: providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different expression profiles in the base *E. coli* strain; engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain; screening and selecting individual *E. coli* strains of the initial solubility tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements; providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent solubility tag swap *E. coli* strain library; screening and selecting individual *E. coli* strains of the subsequent solubility tag swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new solubility tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

[0054] In some embodiments, the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of the initial solubility tag swap *E. coli* strain library.

[0055] In some embodiments, the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial solubility tag swap *E. coli* strain library.

[0056] In some embodiments, the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.

[0057] In some embodiments, the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.

[0058] In some embodiments, steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

[0059] In some embodiments, steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

[0060] In some embodiments, the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.

[0061] In some embodiments, the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.

[0062] In some embodiments, the present disclosure teaches a method for generating a protein degradation tag swap *E. coli* strain library, comprising the steps of: a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different solubility profiles in the base *E. coli* strain; and b. engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation

tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

[0063] In some embodiments, the present disclosure teaches a protein degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of: providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different expression profiles in the base *E. coli* strain; engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain; screening and selecting individual *E. coli* strains of the initial degradation tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements; providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent degradation tag swap *E. coli* strain library; screening and selecting individual *E. coli* strains of the subsequent degradation tag swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new degradation tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

[0064] In some embodiments, the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of the initial degradation tag swap *E. coli* strain library.

[0065] In some embodiments, the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial degradation tag swap *E. coli* strain library.

[0066] In some embodiments, the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.

[0067] In some embodiments, the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.

[0068] In some embodiments, steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent degradation tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

[0069] In some embodiments, steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent degradation tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

[0070] In some embodiments, the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.

[0071] In some embodiments, the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.

[0072] In some embodiments, the present disclosure teaches a chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of lambda phage p_R promoter.

[0073] In some embodiments, nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site

(RBS) portion of the lambda phage p_R promoter are selected from the nucleic acid sequences found in Table 1.5.

[0074] In some embodiments, the present disclosure teaches a chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene.

[0075] In some embodiments, nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene are selected from the nucleic acid sequences found in Table 1.5.

[0076] In some embodiments, the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 132-152, 159-160, 162, 165, 174-175, 188, 190, 199-201 or 207.

[0077] In some embodiments, the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 153-158, 161, 163-164, 166-173, 176-187, 189, 191-198 or 202-206.

[0078] In some embodiments, the microbial host cell is *E. coli*.

[0079] In some embodiments, the heterologous gene encodes a protein product of interest found in Table 2.

[0080] In some embodiments, the heterologous gene is a gene that is part of a lysine biosynthetic pathway.

[0081] In some embodiments, the heterologous gene is selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene, the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene.

[0082] In some embodiments, the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.

[0083] In some embodiments, the heterologous gene is selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA*

gene, the ispB gene, the crtE gene, the crtB gene, the crtI gene, the crtY gene, the ymgA gene, the dxr gene, the elbA gene, the gdhA gene, the appY gene, the elbB gene, or the ymgB gene.

[0084] In some embodiments, the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.

[0085] In some embodiments, the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natrecor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).

[0086] In some embodiments, the present disclosure teaches a heterologous gene operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.

[0087] In some embodiments, the heterologous gene encodes a protein product of interest found in Table 2.

[0088] In some embodiments, the heterologous gene is a gene that is part of a lysine biosynthetic pathway.

[0089] In some embodiments, the heterologous gene is selected from the asd gene, the ask gene, the hom gene, the dapA gene, the dapB gene, the dapD gene, the ddh gene, the argD gene, the dapE gene, the dapF gene, the lysA gene, the lysE gene, the zwf gene, the pgi gene, the ktk gene, the fbp gene, the ppc gene, the pck gene, the ddx gene, the pyc gene or the icd gene.

[0090] In some embodiments, the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.

[0091] In some embodiments, the heterologous gene is selected from the dxs gene, the ispC gene, the ispE gene, the ispD gene, the ispF gene, the ispG gene, the ispH gene, the idi gene, the ispA gene, the ispB gene, the crtE gene, the crtB gene, the crtI gene, the crtY gene, the ymgA gene, the dxr gene, the elbA gene, the gdhA gene, the appY gene, the elbB gene, or the ymgB gene.

[0092] In some embodiments, the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.

[0093] In some embodiments, the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).

BRIEF DESCRIPTION OF THE FIGURES

[0094] **FIGURE 1** depicts a DNA recombination method of the present disclosure for increasing variation in diversity pools. DNA sections, such as genome regions from related species, can be cut *via* physical or enzymatic/chemical means. The cut DNA regions are melted and allowed to reanneal, such that overlapping genetic regions prime polymerase extension reactions. Subsequent melting/extension reactions are carried out until products are reassembled into chimeric DNA, comprising elements from one or more starting sequences.

[0095] **FIGURE 2** outlines methods of the present disclosure for generating new host *E. coli* strains with selected sequence modifications (*e.g.*, 100 SNPs to swap). Briefly, the method comprises (1) desired DNA inserts are designed and generated by combining one or more synthesized oligos in an assembly reaction, (2) DNA inserts are cloned into transformation plasmids, (3) completed plasmids are transferred into desired production strains, where they are integrated into the host strain genome, and (4) selection markers and other unwanted DNA elements are looped out of the host strain. Each DNA assembly step may involve additional quality control (QC) steps, such as cloning plasmids into *E. coli* bacteria for amplification and sequencing.

[0096] **FIGURE 3** depicts assembly of transformation plasmids of the present disclosure, and their integration into a host *E. coli* genome. The insert DNA is generated by combining one or more synthesized oligos in an assembly reaction. DNA inserts containing the desired sequence are flanked by regions of DNA homologous to the targeted region of the genome. These homologous regions facilitate genomic integration, and, once integrated, form direct repeat regions designed for looping out vector backbone DNA in subsequent steps. Assembled plasmids contain the insert DNA, and optionally, one or more selection markers.

[0097] **FIGURE 4** depicts a procedure for looping-out selected regions of DNA from host *E. coli* strains. Direct repeat regions of the inserted DNA and host genome can “loop out” in a recombination event. Cells counter selected for the selection marker contain deletions of the loop DNA flanked by the direct repeat regions.

[0098] **FIGURE 5** depicts an embodiment of the *E. coli* strain improvement process of the present disclosure. Host strain sequences containing genetic modifications (Genetic Design) are tested for strain performance improvements in various strain backgrounds (Strain Build). Strains exhibiting beneficial mutations are analyzed (Hit ID and Analysis) and the data is stored in libraries for further analysis (e.g., SNP swap libraries, PRO swap libraries, and combinations thereof, among others). Selection rules of the present disclosure generate new proposed *E. coli* host strain sequences based on the predicted effect of combining elements from one or more libraries for additional iterative analysis.

[0099] **FIGURE 6A-B** depicts the DNA assembly, transformation, and *E. coli* strain screening steps of one of the embodiments of the present disclosure. **FIGURE 6A** depicts the steps for building DNA fragments, cloning said DNA fragments into vectors, transforming said vectors into host *E. coli* strains, and looping out selection sequences through counter selection. **FIGURE 6B** depicts the steps for high-throughput culturing, screening, and evaluation of selected *E. coli* host strains. This figure also depicts the optional steps of culturing, screening, and evaluating selected *E. coli* strains in culture tanks.

[0100] **FIGURE 7** depicts one embodiment of the automated system of the present disclosure. The present disclosure teaches use of automated robotic systems with various modules capable of cloning, transforming, culturing, screening and/or sequencing host *E. coli*.

[0101] **FIGURE 8** depicts an overview of an embodiment of the *E. coli* strain improvement program of the present disclosure.

[0102] **FIGURE 9** is a representation of the genome of *Corynebacterium glutamicum*, comprising around 3.2 million base pairs.

[0103] **FIGURE 10** depicts the results of a transformation experiment of the present disclosure. DNA inserts ranging from 0.5kb to 5.0kb were targeted for insertion into various regions (shown as relative positions 1-24) of the genome of *Corynebacterium glutamicum*. Light color indicates successful integration, while darker color indicates insertion failure.

[0104] **FIGURE 11** depicts the results of a second round HTP engineering PRO swap program. Top promoter::gene combinations identified during the first PRO swap round were analyzed according to the methods of the present disclosure to identify combinations of said mutations that would be likely to exhibit additive or combinatorial beneficial effects on host performance. Second round PRO swap mutants thus comprised pair combinations of various promoter::gene mutations. The resulting second round mutants were screened for differences in host cell yield of a selected biomolecule. A combination pair of mutations that had been predicted to exhibit beneficial effects is emphasized with a circle.

[0105] **FIGURE 12** depicts the results of an experiment testing successful plasmid assembly for plasmids transformed into *E. coli*. Picking four colonies is sufficient to achieve 13% failure rate for plasmids containing 1 and 2kb insertion sequences. Larger insertions may require additional colony screening to achieve consistent results.

[0106] **FIGURE 13** depicts results of an experiment testing successful transformation of *Corynebacterium glutamicum* with insertion vectors. DNA insert sizes of 2 and 5 kb exhibited high transformation rates with low assembly failure rates.

[0107] **FIGURE 14** depicts results of loop out selections in *Corynebacterium glutamicum*. Sucrose resistance of transformed bacteria indicates loop out of sacB selection marker. DNA insert size does not appear to impact loop out efficiency.

[0108] **FIGURE 15** is a similarity matrix computed using the correlation measure. The matrix is a representation of the functional similarity between SNP variants. The consolidation of SNPs with low functional similarity is expected to have a higher likelihood of improving strain performance, as opposed to the consolidation of SNPs with higher functional similarity.

[0109] **FIGURE 16A-B** depicts the results of an epistasis mapping experiment. Combination of SNPs and PRO swaps with low functional similarities yields improved strain performance. **FIGURE 16A** depicts a dendrogram clustered by functional similarity of all the SNPs/PRO swaps. **FIGURE 16B** depicts host strain performance of consolidated SNPs as measured by product yield. Greater cluster distance correlates with improved consolidation performance of the host strain.

[0110] **FIGURE 17A-B** depicts SNP differences among strain variants in the diversity pool. **FIGURE 17A** depicts the relationship among the strains of this experiment. Strain A is the wild-type host strain. Strain B is an intermediate engineered strain. Strain C is the industrial production strain. **FIGURE 17B** is a graph identifying the number of unique and shared SNPs in each strain.

[0111] **FIGURE 18** depicts a first-round SNP swapping experiment according to the methods of the present disclosure. (1) all the SNPs from C will be individually and/or combinatorially cloned into the base A strain (“wave up” A to C). (2) all the SNPs from C will be individually and/or combinatorially removed from the commercial strain C (“wave down” C to A). (3) all the SNPs from B will be individually and/or combinatorially cloned into the base A strain (wave up A to B). (4) all the SNPs from B will be individually and/or combinatorially removed from the commercial strain B (wave down B to A). (5) all the SNPs unique to C will be individually and/or combinatorially cloned into the commercial B strain (wave up B to C). (6) all the SNPs unique to C will be individually and/or combinatorially removed from the commercial strain C (wave down C to B).

[0112] **FIGURE 19** illustrates example gene targets to be utilized in a promoter swap process. The 4 underlined are diverting genes that can be targeted for downregulation, while the remaining 19 on pathway genes can be targeted for overexpression.

[0113] **FIGURE 20** illustrates an exemplary promoter library that is being utilized to conduct a promoter swap process for the identified gene targets. Promoters utilized in the PRO swap (*i.e.* promoter swap) process are P₁-P₈, the sequences and identity of which can be found in Table 1.

[0114] **FIGURE 21** illustrates the different available approaches to promoter swapping depending on whether the targeted gene comprises its own promoter, or is part of an operon.

[0115] **FIGURE 22** depicts exemplary HTP promoter swapping data showing modifications that significantly affect performance on lysine yield. The X-axis represents different strains within the promoter swap genetic design microbial strain library, and the Y-axis includes relative lysine yield values for each strain. Each letter on the graph represents a PRO swap target gene. Each data point represents a replicate. The data demonstrates that a molecular tool adapted for HTP applications, as described herein (*i.e.* PRO swap), is able to efficiently create and optimize microbial strain performance for the production of a compound or molecule of interest. In this case, the compound of interest was lysine; however, the taught PRO swap molecular tool can be utilized to optimize and/or increase the production of any compound of interest. One of skill in the art would understand how to choose target genes, encoding the production of a desired compound, and then utilize the taught PRO swap procedure. One of skill in the art would readily appreciate that the demonstrated data exemplifying lysine yield increases taught herein, along with the detailed

disclosure presented in the application, enables the PRO swap molecular tool to be a widely applicable advancement in HTP genomic engineering.

[0116] **FIGURE 23** illustrates the distribution of relative strain performances for the input data under consideration. A relative performance of zero indicates that the engineered strain performed equally well to the in-plate base strain. The processes described herein are designed to identify the strains that are likely to perform significantly above zero.

[0117] **FIGURE 24** illustrates the linear regression coefficient values, which depict the average change (increase or decrease) in relative strain performance associated with each genetic change incorporated into the depicted strains.

[0118] **FIGURE 25** illustrates the composition of changes for the top 100 predicted strain designs. The x-axis lists the pool of potential genetic changes (dss mutations are SNP swaps, and Pcg mutations are PRO swaps), and the y-axis shows the rank order. Black cells indicate the presence of a particular change in the candidate design, while white cells indicate the absence of that change. In this particular example, all of the top 100 designs contain the changes pcg3121_pgi, pcg1860_pyc, dss_339, and pcg0007_39_lysa. Additionally, the top candidate design contains the changes dss_034, dss_009.

[0119] **FIGURE 26** depicts the DNA assembly and transformation steps of one of the embodiments of the present disclosure. The flow chart depicts the steps for building DNA fragments, cloning said DNA fragments into vectors, transforming said vectors into host *E. coli* strains, and looping out selection sequences through counter selection.

[0120] **FIGURE 27** depicts the steps for high-throughput culturing, screening, and evaluation of selected host *E. coli* strains. This figure also depicts the optional steps of culturing, screening, and evaluating selected *E. coli* strains in culture tanks.

[0121] **FIGURE 28** depicts expression profiles of illustrative promoters exhibiting a range of regulatory expression, according to the promoter ladders of the present disclosure. Promoter A expression peaks at the lag phase of bacterial cultures, while promoter B and C peak at the exponential and stationary phase, respectively.

[0122] **FIGURE 29** depicts expression profiles of illustrative promoters exhibiting a range of regulatory expression, according to the promoter ladders of the present disclosure. Promoter A expression peaks immediately upon addition of a selected substrate, but quickly returns to undetectable levels as the concentration of the substrate is reduced. Promoter B expression peaks

immediately upon addition of the selected substrate and lowers slowly back to undetectable levels together with the corresponding reduction in substrate. Promoter C expression peaks upon addition of the selected substrate, and remains highly expressed throughout the culture, even after the substrate has dissipated.

[0123] **FIGURE 30** depicts expression profiles of illustrative promoters exhibiting a range of constitutive expression levels, according to the promoter ladders of the present disclosure. Promoter A exhibits the lowest expression, followed by increasing expression levels promoter B and C, respectively.

[0124] **FIGURE 31** diagrams an embodiment of LIMS system of the present disclosure for *E. coli* strain improvement.

[0125] **FIGURE 32** diagrams a cloud computing implementation of embodiments of the LIMS system of the present disclosure.

[0126] **FIGURE 33** depicts an embodiment of the iterative predictive strain design workflow of the present disclosure.

[0127] **FIGURE 34** diagrams an embodiment of a computer system, according to embodiments of the present disclosure.

[0128] **FIGURE 35** depicts the workflow associated with the DNA assembly according to one embodiment of the present disclosure. This process is divided up into 4 stages: parts generation, plasmid assembly, plasmid QC, and plasmid preparation for transformation. During parts generation, oligos designed by Laboratory Information Management System (LIMS) are ordered from an oligo sequencing vendor and used to amplify the target sequences from the host organism via PCR. These PCR parts are cleaned to remove contaminants and assessed for success by fragment analysis, *in silico* quality control comparison of observed to theoretical fragment sizes, and DNA quantification. The parts are transformed into yeast along with an assembly vector and assembled into plasmids via homologous recombination. Assembled plasmids are isolated from yeast and transformed into *E. coli* for subsequent assembly quality control and amplification. During plasmid assembly quality control, several replicates of each plasmid are isolated, amplified using Rolling Circle Amplification (RCA), and assessed for correct assembly by enzymatic digest and fragment analysis. Correctly assembled plasmids identified during the QC process are hit picked to generate permanent stocks and the plasmid DNA extracted and quantified prior to transformation into the target host organism.

[0129] **FIGURE 36** depicts the results of an experiment characterizing the effects of Terminators T1-T8 in two media over two time points. Conditions A and C represent the two time points for the BHI media, while the B and D points represent the two time points for the HTP test media.

[0130] **FIGURE 37** depicts the results of an experiment comparing the effectiveness of traditional strain improvement approaches such as UV mutagenesis against the HTP engineering methodologies of the present disclosure. The vast majority of UV mutations produced no noticeable increase in host cell performance. In contrast, PRO swap methodologies of the present disclosure produced a high proportion of mutants exhibiting 1.2 to 2 fold increases in host cell performance.

[0131] **FIGURE 38** depicts the results of a first round HTP engineering SNP swap program. 186 individual SNP mutations were identified and individually cloned onto a base strain. The resulting mutants were screened for differences in host cell yield of a selected biomolecule.

[0132] **FIGURE 39** depicts the results of a second round HTP engineering SNP swap program. 176 individual SNP mutations from a first round SNP swap program were individually cloned into a second round host cell strain containing a beneficial SNP identified during a first round SNP program. The resulting mutants thus represent the effect of two mutation combination pairs. Screening results for differences in host cell yield (Y-axis) and productivity (X-axis) for the selected biomolecule are shown.

[0133] **FIGURE 40** depicts the results of a tank fermentation validation experiment. The top mutation pairs from the second round of HTP SNP swap were cultured in fermentation tanks. Results for host cell yield and productivity for the selected biomolecule (*i.e.* lysine) are shown. As can be seen, in one round of genomic engineering the inventors utilized the PRO swap procedure to determine that a particular PRO swap mutant (*zwf*) exhibited increased yield of a selected biomolecule compared to base strain (*i.e.* compare base strain to base strain + *zwf*). Then, the inventors performed another round of genomic engineering, wherein a SNP swap procedure was used to determine beneficial SNP mutations that could affect yield of the biomolecule, when combined with said PRO swap mutant. The combination of the PRO swap procedure and SNP swap procedure created mutants with even higher yields than the previous PRO swap only mutants (*i.e.* compare base strain + *zwf* + SNP121 to the previously discussed base strain + *zwf*). This figure illustrates the dramatic improvements in yield that can be achieved by combining the PRO swap and SNP swap procedures of the disclosure. In aspects, combining a PRO swap genomic

engineering campaign with a SNP swap genomic engineering campaign can lead to increased yield and/or productivity of a biomolecule/product of interest by a factor of 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 40%, 45%, 50%, or more, relative to a base strain.

[0134] FIGURE 41 depicts the results of a first round HTP engineering PRO swap program. Selected genes believed to be associated with host performance were combined with a promoter ladder to create a first round PRO swap library, according to the methods of the present disclosure. The resulting mutants were screened for differences in host cell yield of a selected biomolecule (*i.e.* lysine).

[0135] FIGURE 42 is a flowchart illustrating the consideration of epistatic effects in the selection of mutations for the design of a microbial strain, according to embodiments of the disclosure.

[0136] FIGURE 43 depicts a Bicistronic Design (BCD) regulatory sequence, according to the present disclosure. In some embodiments, the present disclosure teaches that BCDs can be used in place of traditional promoters in order to improve expression consistency between different promoter::target genes combinations in PRO-swaps. In some embodiments, BCDs comprise a promoter, a first ribosome binding site (SD1), a first cistronic sequence (Cis1), a second ribosome binding site (SD2), operably linked to a target gene of interest (Cis2). In some embodiments, the present disclosure teaches that Cis1 can be any peptide-coding sequence. Additional information about BCD design and use is provided in later sections of the specification.

[0137] FIGURE 44 is an illustration of pathway enzyme co-localization via recombinant DNA binding domains. An engineered cell encodes pathway enzymes Enz1-3 with DNA binding domains. When expressed, these enzymes bind to a scaffold DNA or other target location, which comprises DNA motifs that are recognized by the recombinant DNA binding domains fused to the pathway enzymes. When the fused DNA binding domains attach to their cognate DNA motifs on the scaffold plasmid, the enzymes are constrained close to one another in space, which can improve productivity of the pathway.

[0138] FIGURE 45 is a schematic diagram for incorporation of nucleotide sequences encoding DNA binding domains into pathway enzymes. GOI encodes a pathway enzyme. *E. coli* cells are transformed with a plasmid encoding a mutant version of GOI that includes a nucleotide sequence that encodes a DNA binding domain (denoted with a star). The plasmid also encodes an antibiotic resistance marker (Ab) that allows for selection of loop in cells, and a counterselection marker (Counter) that allows for subsequent counterselection of loop out cells. In the “Loop-in” step, the

entire plasmid, including the mutant GOI, is incorporated into the genome via homologous recombination (HR). During the “Loop-out” step, some of the cells will revert to the native GOI sequence via HR, while others will undergo a HR event that leaves the mutant GOI in the genome.

[0139] **FIGURE 46** is a dot plot for the predicted performance vs measured performance of training data for a yield model of the present disclosure. The underlying model is a Kernel Ridge Regression model (with 4th order polynomial kernel). The model is trained on 1864 unique genetic constructs and associated phenotypic performance. The fitted model has an r^2 value of 0.52.

[0140] **FIGURE 47** depicts the genetic makeup of candidate designs generated by the prediction algorithms of the present disclosure. These candidate designs were submitted for HTP build and analysis. Here the candidate design is defined as the combination of parent strain id and introduced mutation(s).

[0141] **FIGURE 48** is a dot plot of the predicted performance vs. measured performance of candidate designs generated by the prediction algorithms of the present disclosure, and built according the HTP build methods of the present disclosure. This figure demonstrates that the model may predict candidate strain performance within an acceptable degree of accuracy.

[0142] **FIGURE 49** is a box and whiskers plot depicting the yield percent change of candidate strains with respect to parent strains. On the y-axis, a value of 0.01 corresponds to 1%. This figure demonstrates that strains designed by a computer model (light gray) achieve measureable improvement over their corresponding parent strains. Additionally, the figure demonstrates that these model base strain improvements are comparable in magnitude to improvements achieved by human expert designed strains.

[0143] **FIGURE 50** illustrates the yield performance distribution for strains designed by the computer model (dark grey) and by a human expert (light grey). Computer-designed strains exhibited tighter distributions with higher median gains.

[0144] **FIGURE 51** is a box and whiskers plot depicting the absolute yield of candidate strains generated by the computer (light grey) or by a human expert (dark grey). Results are aggregated by parent strain.

[0145] **FIGURE 52** is a representation of the genome of *Escherichia coli*, comprising around 4.6 million base pairs.

[0146] **FIGURE 53** illustrates the effect of insulator and terminator parts in vector backbones on the efficiency of transformation and plasmid integration.

[0147] **FIGURE 54** illustrates the combinatorial design of synthetic promoter-5'UTR sequences from Table 1.4.

[0148] **FIGURE 55** depicts the plasmid map illustrating the components of the vector 1 backbone.

[0149] **FIGURE 56** depicts the plasmid map illustrating the components of the vector 2 backbone.

[0150] **FIGURE 57** depicts the plasmid map illustrating the components of the vector 3 backbone.

[0151] **FIGURE 58** depicts the plasmid map illustrating the components of the vector 4 backbone.

[0152] **FIGURE 59** depicts the E. coli lycopene biosynthetic pathway.

[0153] **FIGURE 60** depicts terminator edits at lycopene pathway targets *idi* and *yngA*. The terminator TyjBE demonstrates decreased strain performance relative to the control, thus highlighting the utility of these library types for identifying critical pathway targets.

[0154] **FIGURE 61** depicts terminator edits at multiple lycopene pathway targets.

[0155] **FIGURE 62** depicts promoter (for comparison), degradation tag, and terminator swaps at lycopene pathway target *dxs*. The *ssrA_LAA* degradation tag demonstrates improved strain performance relative to the control. This is unexpected as this strain is a combination of a PROSWP with a degradation tag at a single pathway target. The initial PROSWP is expected to increase protein abundance, and the degradation tag is expected to decrease protein abundance, thus demonstrating the utility of combinations of library types for tuning optimal strain performance.

[0156] **FIGURE 63** depicts solubility tag, promoter, and terminator swaps at lycopene pathway target *gdhA*. The solubility tag FH8 demonstrates improved strain performance relative to the control, but the GB1 solubility tag does not, thus demonstrating the necessity for evaluating libraries of each modification type.

DETAILED DESCRIPTION

Definitions

[0157] While the following terms are believed to be well understood by one of ordinary skill in the art, the following definitions are set forth to facilitate explanation of the presently disclosed subject matter.

[0158] The term “a” or “an” refers to one or more of that entity, *i.e.* can refer to a plural referents. As such, the terms “a” or “an”, “one or more” and “at least one” are used interchangeably herein. In addition, reference to “an element” by the indefinite article “a” or “an” does not exclude the

possibility that more than one of the elements is present, unless the context clearly requires that there is one and only one of the elements.

[0159] As used herein the terms “cellular organism” “microorganism” or “microbe” should be taken broadly. These terms are used interchangeably and include, but are not limited to, the two prokaryotic domains, Bacteria and Archaea, as well as certain eukaryotic fungi and protists. In some embodiments, the disclosure refers to the “microorganisms” or “cellular organisms” or “microbes” of lists/tables and figures present in the disclosure. This characterization can refer to not only the identified taxonomic genera of the tables and figures, but also the identified taxonomic species, as well as the various novel and newly identified or designed strains of any organism in said tables or figures. The same characterization holds true for the recitation of these terms in other parts of the Specification, such as in the Examples.

[0160] The term “prokaryotes” is art recognized and refers to cells which contain no nucleus or other cell organelles. The prokaryotes are generally classified in one of two domains, the Bacteria and the Archaea. The definitive difference between organisms of the Archaea and Bacteria domains is based on fundamental differences in the nucleotide base sequence in the 16S ribosomal RNA.

[0161] The term “Archaea” refers to a categorization of organisms of the division Mendosicutes, typically found in unusual environments and distinguished from the rest of the prokaryotes by several criteria, including the number of ribosomal proteins and the lack of muramic acid in cell walls. On the basis of ssrRNA analysis, the Archaea consist of two phylogenetically-distinct groups: Crenarchaeota and Euryarchaeota. On the basis of their physiology, the Archaea can be organized into three types: methanogens (prokaryotes that produce methane); extreme halophiles (prokaryotes that live at very high concentrations of salt (NaCl); and extreme (hyper) thermophilus (prokaryotes that live at very high temperatures). Besides the unifying archaeal features that distinguish them from Bacteria (*i.e.*, no murein in cell wall, ester-linked membrane lipids, etc.), these prokaryotes exhibit unique structural or biochemical attributes which adapt them to their particular habitats. The Crenarchaeota consists mainly of hyperthermophilic sulfur-dependent prokaryotes and the Euryarchaeota contains the methanogens and extreme halophiles.

[0162] “Bacteria” or “eubacteria” refers to a domain of prokaryotic organisms. Bacteria include at least 11 distinct groups as follows: (1) Gram-positive (gram+) bacteria, of which there are two major subdivisions: (1) high G+C group (*Actinomycetes, Mycobacteria, Micrococcus*, others) (2)

low G+C group (*Bacillus*, *Clostridia*, *Lactobacillus*, *Staphylococci*, *Streptococci*, *Mycoplasmas*); (2) Proteobacteria, *e.g.*, Purple photosynthetic+non-photosynthetic Gram-negative bacteria (includes most “common” Gram-negative bacteria); (3) Cyanobacteria, *e.g.*, oxygenic phototrophs; (4) Spirochetes and related species; (5) Planctomyces; (6) *Bacteroides*, *Flavobacteria*; (7) *Chlamydia*; (8) Green sulfur bacteria; (9) Green non-sulfur bacteria (also anaerobic phototrophs); (10) Radioresistant micrococci and relatives; (11) *Thermotoga* and *Thermosipho thermophiles*.

[0163] A “eukaryote” is any organism whose cells contain a nucleus and other organelles enclosed within membranes. Eukaryotes belong to the taxon Eukarya or Eukaryota. The defining feature that sets eukaryotic cells apart from prokaryotic cells (the aforementioned Bacteria and Archaea) is that they have membrane-bound organelles, especially the nucleus, which contains the genetic material, and is enclosed by the nuclear envelope.

[0164] The terms “genetically modified host cell,” “recombinant host cell,” and “recombinant strain” are used interchangeably herein and refer to host cells that have been genetically modified by the cloning and transformation methods of the present disclosure. Thus, the terms include a host cell (*e.g.*, bacteria, yeast cell, fungal cell, CHO, human cell, etc.) that has been genetically altered, modified, or engineered, such that it exhibits an altered, modified, or different genotype and/or phenotype (*e.g.*, when the genetic modification affects coding nucleic acid sequences of the microorganism), as compared to the naturally-occurring organism from which it was derived. It is understood that in some embodiments, the terms refer not only to the particular recombinant host cell in question, but also to the progeny or potential progeny of such a host cell

[0165] The term “wild-type microorganism” or “wild-type host cell” describes a cell that occurs in nature, *i.e.* a cell that has not been genetically modified.

[0166] The term “genetically engineered” may refer to any manipulation of a host cell’s genome (*e.g.* by insertion, deletion, mutation, or replacement of nucleic acids).

[0167] The term “control” or “control host cell” refers to an appropriate comparator host cell for determining the effect of a genetic modification or experimental treatment. In some embodiments, the control host cell is a wild type cell. In other embodiments, a control host cell is genetically identical to the genetically modified host cell, save for the genetic modification(s) differentiating the treatment host cell. In some embodiments, the present disclosure teaches the use of parent strains as control host cells (*e.g.*, the S₁ strain that was used as the basis for the strain improvement

program). In other embodiments, a host cell may be a genetically identical cell that lacks a specific promoter or SNP being tested in the treatment host cell.

[0168] As used herein, the term “allele(s)” means any of one or more alternative forms of a gene, all of which alleles relate to at least one trait or characteristic. In a diploid cell, the two alleles of a given gene occupy corresponding loci on a pair of homologous chromosomes.

[0169] As used herein, the term “locus” (loci plural) means a specific place or places or a site on a chromosome where for example a gene or genetic marker is found.

[0170] As used herein, the term “genetically linked” refers to two or more traits that are co-inherited at a high rate during breeding such that they are difficult to separate through crossing.

[0171] A “recombination” or “recombination event” as used herein refers to a chromosomal crossing over or independent assortment.

[0172] As used herein, the term “phenotype” refers to the observable characteristics of an individual cell, cell culture, organism, or group of organisms which results from the interaction between that individual’s genetic makeup (*i.e.*, genotype) and the environment.

[0173] As used herein, the term “chimeric” or “recombinant” when describing a nucleic acid sequence or a protein sequence refers to a nucleic acid, or a protein sequence, that links at least two heterologous polynucleotides, or two heterologous polypeptides, into a single macromolecule, or that re-arranges one or more elements of at least one natural nucleic acid or protein sequence. For example, the term “recombinant” can refer to an artificial combination of two otherwise separated segments of sequence, *e.g.*, by chemical synthesis or by the manipulation of isolated segments of nucleic acids by genetic engineering techniques.

[0174] As used herein, a “synthetic nucleotide sequence” or “synthetic polynucleotide sequence” is a nucleotide sequence that is not known to occur in nature or that is not naturally occurring. Generally, such a synthetic nucleotide sequence will comprise at least one nucleotide difference when compared to any other naturally occurring nucleotide sequence.

[0175] As used herein, the term “nucleic acid” refers to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides, or analogs thereof. This term refers to the primary structure of the molecule, and thus includes double- and single-stranded DNA, as well as double- and single-stranded RNA. It also includes modified nucleic acids such as methylated and/or capped nucleic acids, nucleic acids containing modified bases, backbone modifications, and the like. The terms “nucleic acid” and “nucleotide sequence” are used interchangeably.

[0176] As used herein, the term “DNA scaffold” or “nucleic acid scaffold” refers to a nucleic acid scaffold that is either artificially produced or a naturally occurring sequence that is repurposed as a scaffold. In one embodiment of the present invention, the nucleic acid scaffold is a synthetic deoxyribonucleic acid scaffold. The deoxyribonucleotides of the synthetic scaffold may comprise purine and pyrimidine bases or other natural, chemically or biochemically modified, non-natural, or derivatized deoxyribonucleotide bases. As described in more detail herein, the nucleic acid scaffold of the present invention is utilized to spatially and temporally assemble and immobilize two or more proteins involved in a biological pathway, i.e. biosynthetic enzymes, to create a functional complex. The assembly and immobilization of each biological pathway protein on the scaffold occurs via the binding interaction between one of the protein-binding sequences, i.e., protein docking sites, of the scaffold and a corresponding DNA-binding portion of a chimeric biosynthetic enzyme. Accordingly, the nucleic acid scaffold comprises one or more subunits, each subunit comprising two or more protein-binding sequences to accommodate the binding of two or more different chimeric biological pathway proteins.

[0177] As used herein, a “DNA binding sequence” or “DNA binding site” refers to a specific nucleic acid sequence that is recognized and bound by a DNA-binding domain portion of a chimeric biosynthetic gene (e.g., chimeric biosynthetic enzyme) encoded by modified genes of the present disclosure. Many DNA-binding domains and their cognate binding partner DNA recognition sites (i.e., DNA binding sites) are well known in the art. For example, numerous zinc finger binding domains and their corresponding DNA binding target sites are known in the art and suitable for use in the present invention. Other DNA binding domains include, without limitation, leucine zipper binding domains and their corresponding DNA binding sites, winged helix DNA binding domains and their corresponding DNA binding sites, winged helix-turn-helix DNA binding domains and their corresponding DNA binding sites, HMG-box DNA binding domains and their corresponding DNA binding sequences, helix-loop-helix DNA binding domains and their corresponding DNA binding sequences, and helix-turn-helix DNA binding domains and their corresponding DNA binding sequences. Other known DNA binding domains with known DNA binding sequences include the immunoglobulin DNA domain, B3 DNA binding domain, and TAL effector DNA binding domains. Nucleic acid scaffold subunits of the present invention may comprise any two or more of the aforementioned DNA binding sites.

[0178] As used herein, the term “gene” refers to any segment of DNA associated with a biological function. Thus, genes include, but are not limited to, coding sequences and/or the regulatory sequences required for their expression. Genes can also include non-expressed DNA segments that, for example, form recognition sequences for other proteins. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters.

[0179] As used herein, the term “homologous” or “homologue” or “ortholog” or “orthologue” is known in the art and refers to related sequences that share a common ancestor or family member and are determined based on the degree of sequence identity. The terms “homology,” “homologous,” “substantially similar” and “corresponding substantially” are used interchangeably herein. They refer to nucleic acid fragments wherein changes in one or more nucleotide bases do not affect the ability of the nucleic acid fragment to mediate gene expression or produce a certain phenotype. These terms also refer to modifications of the nucleic acid fragments of the instant disclosure such as deletion or insertion of one or more nucleotides that do not substantially alter the functional properties of the resulting nucleic acid fragment relative to the initial, unmodified fragment. It is therefore understood, as those skilled in the art will appreciate, that the disclosure encompasses more than the specific exemplary sequences. These terms describe the relationship between a gene found in one species, subspecies, variety, cultivar or strain and the corresponding or equivalent gene in another species, subspecies, variety, cultivar or strain. For purposes of this disclosure homologous sequences are compared. “Homologous sequences” or “homologues” or “orthologs” are thought, believed, or known to be functionally related. A functional relationship may be indicated in any one of a number of ways, including, but not limited to: (a) degree of sequence identity and/or (b) the same or similar biological function. Preferably, both (a) and (b) are indicated. Homology can be determined using software programs readily available in the art, such as those discussed in Current Protocols in Molecular Biology (F.M. Ausubel *et al.*, eds., 1987) Supplement 30, section 7.718, Table 7.71. Some alignment programs are MacVector (Oxford Molecular Ltd, Oxford, U.K.), ALIGN Plus (Scientific and Educational Software, Pennsylvania) and AlignX (Vector NTI, Invitrogen, Carlsbad, CA). Another alignment program is Sequencher (Gene Codes, Ann Arbor, Michigan), using default parameters.

[0180] As used herein, the term “endogenous” or “endogenous gene,” refers to the naturally occurring gene, in the location in which it is naturally found within the host cell genome. In the

context of the present disclosure, operably linking a heterologous promoter to an endogenous gene means genetically inserting a heterologous promoter sequence in front of an existing gene, in the location where that gene is naturally present. An endogenous gene as described herein can include alleles of naturally occurring genes that have been mutated according to any of the methods of the present disclosure.

[0181] As used herein, the term “exogenous” is used interchangeably with the term “heterologous,” and refers to a substance coming from some source other than its native source. For example, the terms “exogenous protein,” or “exogenous gene” refer to a protein or gene from a non-native source or location, and that have been artificially supplied to a biological system.

[0182] As used herein, the term “nucleotide change” refers to, *e.g.*, nucleotide substitution, deletion, and/or insertion, as is well understood in the art. For example, mutations contain alterations that produce silent substitutions, additions, or deletions, but do not alter the properties or activities of the encoded protein or how the proteins are made.

[0183] As used herein, the term “protein modification” refers to, *e.g.*, amino acid substitution, amino acid modification, deletion, and/or insertion, as is well understood in the art.

[0184] As used herein, the term “at least a portion” or “fragment” of a nucleic acid or polypeptide means a portion having the minimal size characteristics of such sequences, or any larger fragment of the full length molecule, up to and including the full length molecule. A fragment of a polynucleotide of the disclosure may encode a biologically active portion of a genetic regulatory element. A biologically active portion of a genetic regulatory element can be prepared by isolating a portion of one of the polynucleotides of the disclosure that comprises the genetic regulatory element and assessing activity as described herein. Similarly, a portion of a polypeptide may be 4 amino acids, 5 amino acids, 6 amino acids, 7 amino acids, and so on, going up to the full length polypeptide. The length of the portion to be used will depend on the particular application. A portion of a nucleic acid useful as a hybridization probe may be as short as 12 nucleotides; in some embodiments, it is 20 nucleotides. A portion of a polypeptide useful as an epitope may be as short as 4 amino acids. A portion of a polypeptide that performs the function of the full-length polypeptide would generally be longer than 4 amino acids.

[0185] Variant polynucleotides also encompass sequences derived from a mutagenic and recombinogenic procedure such as DNA shuffling. Strategies for such DNA shuffling are known in the art. *See*, for example, Stemmer (1994) PNAS 91:10747-10751; Stemmer (1994) Nature

370:389-391; Crameri *et al.*(1997) Nature Biotech. 15:436-438; Moore *et al.*(1997) J. Mol. Biol. 272:336-347; Zhang *et al.*(1997) PNAS 94:4504-4509; Crameri *et al.*(1998) Nature 391:288-291; and U.S. Patent Nos. 5,605,793 and 5,837,458.

[0186] For PCR amplifications of the polynucleotides disclosed herein, oligonucleotide primers can be designed for use in PCR reactions to amplify corresponding DNA sequences from cDNA or genomic DNA extracted from any organism of interest. Methods for designing PCR primers and PCR cloning are generally known in the art and are disclosed in Sambrook *et al.*(2001) Molecular Cloning: A Laboratory Manual (3rd ed., Cold Spring Harbor Laboratory Press, Plainview, New York). See also Innis *et al.*, eds. (1990) PCR Protocols: A Guide to Methods and Applications (Academic Press, New York); Innis and Gelfand, eds. (1995) PCR Strategies (Academic Press, New York); and Innis and Gelfand, eds. (1999) PCR Methods Manual (Academic Press, New York). Known methods of PCR include, but are not limited to, methods using paired primers, nested primers, single specific primers, degenerate primers, gene-specific primers, vector-specific primers, partially-mismatched primers, and the like.

[0187] The term “primer” as used herein refers to an oligonucleotide which is capable of annealing to the amplification target allowing a DNA polymerase to attach, thereby serving as a point of initiation of DNA synthesis when placed under conditions in which synthesis of primer extension product is induced, *i.e.*, in the presence of nucleotides and an agent for polymerization such as DNA polymerase and at a suitable temperature and pH. The (amplification) primer is preferably single stranded for maximum efficiency in amplification. Preferably, the primer is an oligodeoxyribonucleotide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the agent for polymerization. The exact lengths of the primers will depend on many factors, including temperature and composition (A/T vs. G/C content) of primer. A pair of bi-directional primers consists of one forward and one reverse primer as commonly used in the art of DNA amplification such as in PCR amplification.

[0188] As used herein, “promoter” refers to a DNA sequence capable of controlling the expression of a coding sequence or functional RNA. In some embodiments, the promoter sequence consists of proximal and more distal upstream elements, the latter elements often referred to as enhancers. Accordingly, an “enhancer” is a DNA sequence that can stimulate promoter activity, and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be

composed of different elements derived from different promoters found in nature, or even comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of some variation may have identical promoter activity.

[0189] As used herein, the phrases “recombinant construct”, “expression construct”, “chimeric construct”, “construct”, and “recombinant DNA construct” are used interchangeably herein. A recombinant construct comprises an artificial combination of nucleic acid fragments, *e.g.*, regulatory and coding sequences that are not found together in nature. For example, a chimeric construct may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. Such construct may be used by itself or may be used in conjunction with a vector. If a vector is used then the choice of vector is dependent upon the method that will be used to transform host cells as is well known to those skilled in the art. For example, a plasmid vector can be used. The skilled artisan is well aware of the genetic elements that must be present on the vector in order to successfully transform, select and propagate host cells comprising any of the isolated nucleic acid fragments of the disclosure. The skilled artisan will also recognize that different independent transformation events will result in different levels and patterns of expression (Jones *et al.*, (1985) EMBO J. 4:2411-2418; De Almeida *et al.*, (1989) Mol. Gen. Genetics 218:78-86), and thus that multiple events must be screened in order to obtain lines displaying the desired expression level and pattern. Such screening may be accomplished by Southern analysis of DNA, Northern analysis of mRNA expression, immunoblotting analysis of protein expression, or phenotypic analysis, among others. Vectors can be plasmids, viruses, bacteriophages, pro-viruses, phagemids, transposons, artificial chromosomes, and the like, that replicate autonomously or can integrate into a chromosome of a host cell. A vector can also be a naked RNA polynucleotide, a naked DNA polynucleotide, a polynucleotide composed of both DNA and RNA within the same strand, a poly-lysine-conjugated DNA or RNA, a peptide-conjugated DNA or RNA, a liposome-conjugated DNA, or the like, that is not autonomously replicating. As used herein, the term “expression” refers to the production of a functional end-product *e.g.*, an mRNA or a protein (precursor or mature).

[0190] “Operably linked” means in this context the sequential arrangement of the promoter polynucleotide according to the disclosure with a further oligo- or polynucleotide, resulting in transcription of said further polynucleotide.

[0191] The term "product of interest" or "biomolecule" as used herein refers to any product produced by microbes from feedstock. In some cases, the product of interest may be a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, etc. For example, the product of interest or biomolecule may be any primary or secondary extracellular metabolite. The primary metabolite may be, inter alia, ethanol, citric acid, lactic acid, glutamic acid, glutamate, lysine, threonine, tryptophan and other amino acids, vitamins, polysaccharides, etc. The secondary metabolite may be, inter alia, an antibiotic compound like penicillin, or an immunosuppressant like cyclosporin A, a plant hormone like gibberellin, a statin drug like lovastatin, a fungicide like griseofulvin, etc. The product of interest or biomolecule may also be any intracellular component produced by a microbe, such as: a microbial enzyme, including: catalase, amylase, protease, pectinase, glucose isomerase, cellulase, hemicellulase, lipase, lactase, streptokinase, and many others. The intracellular component may also include recombinant proteins, such as: insulin, hepatitis B vaccine, interferon, granulocyte colony-stimulating factor, streptokinase and others.

[0192] The term “carbon source” generally refers to a substance suitable to be used as a source of carbon for cell growth. Carbon sources include, but are not limited to, biomass hydrolysates, starch, sucrose, cellulose, hemicellulose, xylose, and lignin, as well as monomeric components of these substrates. Carbon sources can comprise various organic compounds in various forms, including, but not limited to polymers, carbohydrates, acids, alcohols, aldehydes, ketones, amino acids, peptides, etc. These include, for example, various monosaccharides such as glucose, dextrose (D-glucose), maltose, oligosaccharides, polysaccharides, saturated or unsaturated fatty acids, succinate, lactate, acetate, ethanol, etc., or mixtures thereof. Photosynthetic organisms can additionally produce a carbon source as a product of photosynthesis. In some embodiments, carbon sources may be selected from biomass hydrolysates and glucose.

[0193] The term “feedstock” is defined as a raw material or mixture of raw materials supplied to a microorganism or fermentation process from which other products can be made. For example, a carbon source, such as biomass or the carbon compounds derived from biomass are a feedstock for a microorganism that produces a product of interest (*e.g.* small molecule, peptide, synthetic

compound, fuel, alcohol, etc.) in a fermentation process. However, a feedstock may contain nutrients other than a carbon source.

[0194] The term “volumetric productivity” or “production rate” is defined as the amount of product formed per volume of medium per unit of time. Volumetric productivity can be reported in gram per liter per hour (g/L/h).

[0195] The term “specific productivity” is defined as the rate of formation of the product. Specific productivity is herein further defined as the specific productivity in gram product per gram of cell dry weight (CDW) per hour (g/g CDW/h). Using the relation of CDW to OD₆₀₀ for the given microorganism specific productivity can also be expressed as gram product per liter culture medium per optical density of the culture broth at 600 nm (OD) per hour (g/L/h/OD).

[0196] The term “yield” is defined as the amount of product obtained per unit weight of raw material and may be expressed as g product per g substrate (g/g). Yield may be expressed as a percentage of the theoretical yield. “Theoretical yield” is defined as the maximum amount of product that can be generated per a given amount of substrate as dictated by the stoichiometry of the metabolic pathway used to make the product.

[0197] The term “titre” or “titer” is defined as the strength of a solution or the concentration of a substance in solution. For example, the titre of a product of interest (*e.g.* small molecule, peptide, synthetic compound, fuel, alcohol, etc.) in a fermentation broth is described as g of product of interest in solution per liter of fermentation broth (g/L).

[0198] The term “total titer” is defined as the sum of all product of interest produced in a process, including but not limited to the product of interest in solution, the product of interest in gas phase if applicable, and any product of interest removed from the process and recovered relative to the initial volume in the process or the operating volume in the process

[0199] As used herein, the term “HTP genetic design library” or “library” refers to collections of genetic perturbations according to the present disclosure. In some embodiments, the libraries of the present invention may manifest as i) a collection of sequence information in a database or other computer file, ii) a collection of genetic constructs encoding for the aforementioned series of genetic elements, or iii) host cell strains comprising said genetic elements. In some embodiments, the libraries of the present disclosure may refer to collections of individual elements (*e.g.*, collections of promoters for PRO swap libraries, collections of terminators for STOP swap libraries, collections of protein solubility tags for SOLUBILITY TAG swap libraries, or

collections of protein degradation tags for DEGRADATION TAG swap libraries). In other embodiments, the libraries of the present disclosure may also refer to combinations of genetic elements, such as combinations of promoter:genes, gene:terminator, or even promoter:gene:terminators. In some embodiments, the libraries of the present disclosure may also refer to combinations of promoters, terminators, protein solubility tags and/or protein degradation tags. In some embodiments, the libraries of the present disclosure further comprise meta data associated with the effects of applying each member of the library in host organisms. For example, a library as used herein can include a collection of promoter::gene sequence combinations, together with the resulting effect of those combinations on one or more phenotypes in a particular species, thus improving the future predictive value of using said combination in future promoter swaps.

[0200] As used herein, the term “SNP” refers to Small Nuclear Polymorphism(s). In some embodiments, SNPs of the present disclosure should be construed broadly, and include single nucleotide polymorphisms, sequence insertions, deletions, inversions, and other sequence replacements. As used herein, the term “non-synonymous” or non-synonymous SNPs” refers to mutations that lead to coding changes in host cell proteins

[0201] A “high-throughput (HTP)” method of genomic engineering may involve the utilization of at least one piece of automated equipment (*e.g.* a liquid handler or plate handler machine) to carry out at least one step of said method.

Traditional Methods of Strain Improvement

[0202] Traditional approaches to strain improvement can be broadly categorized into two types of approaches: directed strain engineering, and random mutagenesis.

[0203] Directed engineering methods of strain improvement involve the planned perturbation of a handful of genetic elements of a specific organism. These approaches are typically focused on modulating specific biosynthetic or developmental programs, and rely on prior knowledge of the genetic and metabolic factors affecting said pathways. In its simplest embodiments, directed engineering involves the transfer of a characterized trait (*e.g.*, gene, promoter, or other genetic element capable of producing a measurable phenotype) from one organism to another organism of the same, or different species.

[0204] Random approaches to strain engineering involve the random mutagenesis of parent strains, coupled with extensive screening designed to identify performance improvements. Approaches to generating these random mutations include exposure to ultraviolet radiation, or

mutagenic chemicals such as Ethyl methanesulfonate. Though random and largely unpredictable, this traditional approach to strain improvement had several advantages compared to more directed genetic manipulations. *First*, many industrial organisms were (and remain) poorly characterized in terms of their genetic and metabolic repertoires, rendering alternative directed improvement approaches difficult, if not impossible.

[0205] *Second*, even in relatively well characterized systems, genotypic changes that result in industrial performance improvements are difficult to predict, and sometimes only manifest themselves as epistatic phenotypes requiring cumulative mutations in many genes of known and unknown function.

[0206] *Additionally*, for many years, the genetic tools required for making directed genomic mutations in a given industrial organism were unavailable, or very slow and/or difficult to use.

[0207] The extended application of the traditional strain improvement programs, however, yield progressively reduced gains in a given strain lineage, and ultimately lead to exhausted possibilities for further strain efficiencies. Beneficial random mutations are relatively rare events, and require large screening pools and high mutation rates. This inevitably results in the inadvertent accumulation of many neutral and/or detrimental (or partly detrimental) mutations in “improved” strains, which ultimately create a drag on future efficiency gains.

[0208] Another limitation of traditional cumulative improvement approaches is that little to no information is known about any particular mutation’s effect on any strain metric. This fundamentally limits a researcher’s ability to combine and consolidate beneficial mutations, or to remove neutral or detrimental mutagenic “baggage.”

[0209] Other approaches and technologies exist to randomly recombine mutations between strains within a mutagenic lineage. For example, some formats and examples for iterative sequence recombination, sometimes referred to as DNA shuffling, evolution, or molecular breeding, have been described in U.S. patent application Ser. No. 08/198,431, filed Feb. 17, 1994, Serial No. PCT/US95/02126, filed, Feb. 17, 1995, Ser. No. 08/425,684, filed Apr. 18, 1995, Ser. No. 08/537,874, filed Oct. 30, 1995, Ser. No. 08/564,955, filed Nov. 30, 1995, Ser. No. 08/621,859, filed Mar. 25, 1996, Ser. No. 08/621,430, filed Mar. 25, 1996, Serial No. PCT/US96/05480, filed Apr. 18, 1996, Ser. No. 08/650,400, filed May 20, 1996, Ser. No. 08/675,502, filed Jul. 3, 1996, Ser. No. 08/721, 824, filed Sep. 27, 1996, and Ser. No. 08/722,660 filed Sep. 27, 1996; Stemmer, *Science* 270:1510 (1995); Stemmer *et al.*, *Gene* 164:49-53 (1995);

Stemmer, *Bio/Technology* 13:549-553 (1995); Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* 91:10747-10751 (1994); Stemmer, *Nature* 370:389-391 (1994); Cramer *et al.*, *Nature Medicine* 2(1):1-3 (1996); Cramer *et al.*, *Nature Biotechnology* 14:315-319 (1996), each of which is incorporated herein by reference in its entirety for all purposes.

[0210] These include techniques such as protoplast fusion and whole genome shuffling that facilitate genomic recombination across mutated strains. For some industrial microorganisms such as yeast and filamentous fungi, natural mating cycles can also be exploited for pairwise genomic recombination. In this way, detrimental mutations can be removed by ‘back-crossing’ mutants with parental strains and beneficial mutations consolidated. Moreover, beneficial mutations from two different strain lineages can potentially be combined, which creates additional improvement possibilities over what might be available from mutating a single strain lineage on its own. However, these approaches are subject to many limitations that are circumvented using the methods of the present disclosure.

[0211] For example, traditional recombinant approaches as described above are slow and rely on a relatively small number of random recombination crossover events to swap mutations, and are therefore limited in the number of combinations that can be attempted in any given cycle, or time period. In addition, although the natural recombination events in the prior art are essentially random, they are also subject to genome positional bias.

[0212] Most importantly, the traditional approaches also provide little information about the influence of individual mutations and due to the random distribution of recombined mutations many specific combinations cannot be generated and evaluated.

[0213] To overcome many of the aforementioned problems associated with traditional strain improvement programs, the present disclosure sets forth a unique HTP genomic engineering platform that is computationally driven and integrates molecular biology, automation, data analytics, and machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets that are used to construct HTP genetic design libraries. These genetic design libraries will be elaborated upon below. Figure 8 depicts an overview of an embodiment of the *E. coli* strain improvement program of the present disclosure.

[0214] The presently disclosed HTP platform and its unique microbial genetic design libraries fundamentally shift the paradigm of microbial strain development and evolution. For example, traditional mutagenesis-based methods of developing an industrial microbial strain will eventually

lead to microbes burdened with a heavy mutagenic load that has been accumulated over years of random mutagenesis.

[0215] The ability to solve this issue (*i.e.* remove the genetic baggage accumulated by these microbes) has eluded microbial researchers for decades. However, utilizing the HTP platform disclosed herein, these industrial strains can be “rehabilitated,” and the genetic mutations that are deleterious can be identified and removed. Congruently, the genetic mutations that are identified as beneficial can be kept, and in some cases improved upon. The resulting microbial strains demonstrate superior phenotypic traits (*e.g.*, improved production of a compound of interest), as compared to their parental strains.

[0216] Furthermore, the HTP platform taught herein is able to identify, characterize, and quantify the effect that individual mutations have on microbial strain performance. This information, *i.e.* what effect does a given genetic change x have on host cell phenotype y (*e.g.*, production of a compound or product of interest), is able to be generated and then stored in the microbial HTP genetic design libraries discussed below. That is, sequence information for each genetic permutation, and its effect on the host cell phenotype are stored in one or more databases, and are available for subsequent analysis (*e.g.*, epistasis mapping, as discussed below). The present disclosure also teaches methods of physically saving/storing valuable genetic permutations in the form of genetic insertion constructs, or in the form of one or more host cell organisms containing said genetic permutation (*e.g.*, see libraries discussed below.)

[0217] When one couples these HTP genetic design libraries into an iterative process that is integrated with a sophisticated data analytics and machine learning process a dramatically different methodology for improving host cells emerges. The taught platform is therefore fundamentally different from the previously discussed traditional methods of developing host cell strains. The taught HTP platform does not suffer from many of the drawbacks associated with the previous methods. These and other advantages will become apparent with reference to the HTP molecular tool sets and the derived genetic design libraries discussed below.

Genetic Design & Microbial Engineering: A Systematic Combinatorial Approach to Strain Improvement Utilizing a Suite of HTP Molecular Tools and HTP Genetic Design Libraries

[0218] As aforementioned, the present disclosure provides a novel HTP platform and genetic design strategy for engineering microbial organisms through iterative systematic introduction and removal of genetic changes across strains. The platform is supported by a suite of molecular tools,

which enable the creation of HTP genetic design libraries and allow for the efficient implementation of genetic alterations into a given host strain.

[0219] The HTP genetic design libraries of the disclosure serve as sources of possible genetic alterations that may be introduced into a particular microbial strain background. In this way, the HTP genetic design libraries are repositories of genetic diversity, or collections of genetic perturbations, which can be applied to the initial or further engineering of a given microbial strain. Techniques for programming genetic designs for implementation to host strains are described in pending US Patent Application, Serial No. 15/140,296, and pending International Patent Application Serial No. PCT/US17/29725, entitled “Microbial Strain Design System and Methods for Improved Large Scale Production of Engineered Nucleotide Sequences,” each of which is incorporated by reference in its entirety herein.

[0220] The HTP molecular tool sets utilized in this platform may include, *inter alia*: (1) Promoter swaps (PRO Swap), (2) SNP swaps, (3) Start/Stop codon exchanges, (4) STOP swaps, (5) Sequence optimization, (6) SOLUBILITY TAG swaps and (7) DEGRADATION TAG swaps. The HTP methods of the present disclosure also teach methods for directing the consolidation/combinatorial use of HTP tool sets, including (8) Epistasis mapping protocols. As aforementioned, this suite of molecular tools, either in isolation or combination, enables the creation of HTP genetic design host cell libraries.

[0221] As will be demonstrated, utilization of the aforementioned HTP genetic design libraries in the context of the taught HTP microbial engineering platform enables the identification and consolidation of beneficial “causative” mutations or gene sections and also the identification and removal of passive or detrimental mutations or gene sections. This new approach allows rapid improvements in strain performance that could not be achieved by traditional random mutagenesis or directed genetic engineering. The removal of genetic burden or consolidation of beneficial changes into a strain with no genetic burden also provides a new, robust starting point for additional random mutagenesis that may enable further improvements.

[0222] In some embodiments, the present disclosure teaches that as orthogonal beneficial changes are identified across various, discrete branches of a mutagenic strain lineage, they can also be rapidly consolidated into better performing strains. These mutations can also be consolidated into strains that are not part of mutagenic lineages, such as strains with improvements gained by directed genetic engineering.

[0223] In some embodiments, the present disclosure differs from known strain improvement approaches in that it analyzes the genome-wide combinatorial effect of mutations across multiple disparate genomic regions, including expressed and non-expressed genetic elements, and uses gathered information (e.g., experimental results) to predict mutation combinations expected to produce strain enhancements.

[0224] In some embodiments, the present disclosure teaches: i) industrial microorganisms, and other host cells amenable to improvement *via* the disclosed inventions, ii) generating diversity pools for downstream analysis, iii) methods and hardware for high-throughput screening and sequencing of large variant pools, iv) methods and hardware for machine learning computational analysis and prediction of synergistic effects of genome-wide mutations, and v) methods for high-throughput strain engineering.

[0225] The following molecular tools and libraries are discussed in terms of illustrative microbial examples. Persons having skill in the art will recognize that the HTP molecular tools of the present disclosure are compatible with any host cell, including eukaryotic cellular, and higher life forms. Furthermore, many of the illustrated embodiments are conducted in *Corynebacterium*; however, the same principles and process can be deployed in *Escherichia coli*.

[0226] Each of the identified HTP molecular tool sets—which enable the creation of the various HTP genetic design libraries utilized in the microbial engineering platform—will now be discussed.

1. Promoter Swaps: A Molecular Tool for the Derivation of Promoter Swap Microbial Strain Libraries

[0227] In some embodiments, the present disclosure teaches methods of selecting promoters with optimal expression properties to produce beneficial effects on overall-host strain phenotype (e.g., yield or productivity).

[0228] For example, in some embodiments, the present disclosure teaches methods of identifying one or more promoters and/or generating variants of one or more promoters within a host cell, which exhibit a range of expression strengths (e.g. promoter ladders discussed *infra*), or superior regulatory properties (e.g., tighter regulatory control for selected genes). A particular combination of these identified and/or generated promoters can be grouped together as a promoter ladder, which is explained in more detail below.

[0229] The promoter ladder in question is then associated with a given gene of interest. Thus, if one has promoters P₁-P₈ (representing eight promoters that have been identified and/or generated to exhibit a range of expression strengths) and associates the promoter ladder with a single gene of interest in a microbe (*i.e.* genetically engineer a microbe with a given promoter operably linked to a given target gene), then the effect of each combination of the eight promoters can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered microbes have an otherwise identical genetic background except the particular promoter(s) associated with the target gene.

[0230] The resultant microbes that are engineered *via* this process form HTP genetic design libraries.

[0231] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed *via* this process, with each member strain being representative of a given promoter operably linked to a particular target gene, in an otherwise identical genetic background, said library being termed a “promoter swap microbial strain library.” In the specific context of *E. coli*, the library can be termed a “promoter swap *E. coli* strain library,” but the terms can be used synonymously, as *E. coli* is a specific example of a microbe.

[0232] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given promoter *x* operably linked to a given gene *y*—said collection being termed a “promoter swap library.”

[0233] Further, one can utilize the same promoter ladder comprising promoters P₁-P₈ to engineer microbes, wherein each of the 8 promoters is operably linked to 10 different gene targets. The result of this procedure would be 80 microbes that are otherwise assumed genetically identical, except for the particular promoters operably linked to a target gene of interest. These 80 microbes could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any data storage construct, including a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could be, for example, a given promoter's (*e.g.* P₁-P₈) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more of promoters P₁-P₈ to a given gene target.

[0234] The aforementioned examples of eight promoters and 10 target genes is merely illustrative, as the concept can be applied with any given number of promoters that have been grouped together based upon exhibition of a range of expression strengths and any given number of target genes. Persons having skill in the art will also recognize the ability to operably link two or more promoters in front of any gene target. Thus, in some embodiments, the present disclosure teaches promoter swap libraries in which 1, 2, 3 or more promoters from a promoter ladder are operably linked to one or more genes.

[0235] In summary, utilizing various promoters to drive expression of various genes in an organism is a powerful tool to optimize a trait of interest. The molecular tool of promoter swapping, developed by the inventors, uses a ladder of promoter sequences that have been demonstrated to vary expression of at least one locus under at least one condition. This ladder is then systematically applied to a group of genes in the organism using high-throughput genome engineering. This group of genes is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity. In some embodiments, the selection of genes can include all the genes in a given host. In other embodiments, the selection of genes can be a subset of all genes in a given host, chosen randomly.

[0236] The resultant HTP genetic design microbial strain library of organisms containing a promoter sequence linked to a gene is then assessed for performance in a high-throughput screening model, and promoter-gene linkages which lead to increased performance are determined and the information stored in a database. The collection of genetic perturbations (*i.e.* given promoter x operably linked to a given gene y) form a “promoter swap library,” which can be utilized as a source of potential genetic alterations to be utilized in microbial engineering processing. Over time, as a greater set of genetic perturbations is implemented against a greater diversity of host cell backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest.

[0237] Transcription levels of genes in an organism are a key point of control for affecting organism behavior. Transcription is tightly coupled to translation (protein expression), and which proteins are expressed in what quantities determines organism behavior. Cells express thousands

of different types of proteins, and these proteins interact in numerous complex ways to create function. By varying the expression levels of a set of proteins systematically, function can be altered in ways that, because of complexity, are difficult to predict. Some alterations may increase performance, and so, coupled to a mechanism for assessing performance, this technique allows for the generation of organisms with improved function.

[0238] In the context of a small molecule synthesis pathway, enzymes interact through their small molecule substrates and products in a linear or branched chain, starting with a substrate and ending with a small molecule of interest. Because these interactions are sequentially linked, this system exhibits distributed control, and increasing the expression of one enzyme can only increase pathway flux until another enzyme becomes rate limiting.

[0239] Metabolic Control Analysis (MCA) is a method for determining, from experimental data and first principles, which enzyme or enzymes are rate limiting. MCA is limited however, because it requires extensive experimentation after each expression level change to determine the new rate limiting enzyme. Promoter swapping is advantageous in this context, because through the application of a promoter ladder to each enzyme in a pathway, the limiting enzyme is found, and the same thing can be done in subsequent rounds to find new enzymes that become rate limiting. Further, because the read-out on function is better production of the small molecule of interest, the experiment to determine which enzyme is limiting is the same as the engineering to increase production, thus shortening development time. In some embodiments the present disclosure teaches the application of PRO swap to genes encoding individual subunits of multi-unit enzymes. In yet other embodiments, the present disclosure teaches methods of applying PRO swap techniques to genes responsible for regulating individual enzymes, or whole biosynthetic pathways.

[0240] In some embodiments, the promoter swap tool of the present disclosure is used to identify optimum expression of a selected gene target. In some embodiments, the goal of the promoter swap may be to increase expression of a target gene to reduce bottlenecks in a metabolic or genetic pathway. In other embodiments, the goal of the promoter swap may be to reduce the expression of the target gene to avoid unnecessary energy expenditures in the host cell, when expression of said target gene is not required.

[0241] In the context of other cellular systems like transcription, transport, or signaling, various rational methods can be used to try and find out, *a priori*, which proteins are targets for expression

change and what that change should be. These rational methods reduce the number of perturbations that must be tested to find one that improves performance, but they do so at significant cost. Gene deletion studies identify proteins whose presence is critical for a particular function, and important genes can then be over-expressed. Due to the complexity of protein interactions, this is often ineffective at increasing performance. Different types of models have been developed that attempt to describe, from first principles, transcription or signaling behavior as a function of protein levels in the cell. These models often suggest targets where expression changes might lead to different or improved function. The assumptions that underlie these models are simplistic and the parameters difficult to measure, so the predictions they make are often incorrect, especially for non-model organisms. With both gene deletion and modeling, the experiments required to determine how to affect a certain gene are different than the subsequent work to make the change that improves performance. Promoter swapping sidesteps these challenges, because the constructed strain that highlights the importance of a particular perturbation is also, already, the improved strain.

[0242] Thus, in particular embodiments, promoter swapping is a multi-step process comprising:

[0243] 1. Selecting a set of “x” promoters to act as a “ladder.” Ideally these promoters have been shown to lead to highly variable expression across multiple genomic loci, but the only requirement is that they perturb gene expression in some way.

[0244] 2. Selecting a set of “n” genes to target. This set can be every open reading frame (ORF) in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps or previous SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection. In other embodiments, the “n” targeted genes can comprise non-protein coding genes, including non-coding RNAs.

[0245] 3. High-throughput strain engineering to rapidly-and in some embodiments, in parallel-carry out the following genetic modifications: When a native promoter exists in front of target gene n and its sequence is known, replace the native promoter with each of the x promoters in the ladder. When the native promoter does not exist, or its sequence is unknown, insert each of the x promoters in the ladder in front of gene n (*see e.g.*, Figure 21). In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the

library is an instance of x promoter operably linked to n target, in an otherwise identical genetic context. As previously described combinations of promoters can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0246] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0247] This foundational process can be extended to provide further improvements in strain performance by, *inter alia*: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then sequential regeneration of the library of perturbations into an improved member or members of the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches (see Figure 20).

[0248] The molecular tool, or technique, discussed above is characterized as promoter swapping, but is not limited to promoters and can include other sequence changes that systematically vary the expression level of a set of targets. Other methods for varying the expression level of a set of genes could include: a) a ladder of ribosome binding sites (or Kozak sequences in eukaryotes); b) replacing the start codon of each target with each of the other start codons (*i.e.* start/stop codon exchanges discussed *infra*); c) attachment of various mRNA stabilizing or destabilizing sequences to the 5' or 3' end, or at any other location, of a transcript, d) attachment of various protein stabilizing or destabilizing sequences at any location in the protein (*i.e.*, degradation or solubilization tag exchanges discussed *infra*).

[0249] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

2. *SNP Swaps: A Molecular Tool for the Derivation of SNP Swap Microbial*

Strain Libraries

[0250] In certain embodiments, SNP swapping is not a random mutagenic approach to improving a microbial strain, but rather involves the systematic introduction or removal of individual Small Nuclear Polymorphism nucleotide mutations (*i.e.* SNPs) (hence the name “SNP swapping”) across strains.

[0251] The resultant microbes that are engineered *via* this process form HTP genetic design libraries.

[0252] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed *via* this process, with each member strain being representative of the presence or absence of a given SNP, in an otherwise identical genetic background, said library being termed a “SNP swap microbial strain library.” In the specific context of *E. coli*, the library can be termed a “SNP swap *E. coli* strain library,” but the terms can be used synonymously, as *E. coli* is a specific example of a microbe.

[0253] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given SNP being present or a given SNP being absent—said collection being termed a “SNP swap library.”

[0254] In some embodiments, SNP swapping involves the reconstruction of host organisms with optimal combinations of target SNP “building blocks” with identified beneficial performance effects. Thus, in some embodiments, SNP swapping involves consolidating multiple beneficial mutations into a single strain background, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations.

[0255] In other embodiments, SNP swapping also involves removing multiple mutations identified as detrimental from a strain, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations. In some embodiments, the SNP swapping methods of the present disclosure include both the addition of beneficial SNPs, and removing detrimental and/or neutral mutations.

[0256] SNP swapping is a powerful tool to identify and exploit both beneficial and detrimental mutations in a lineage of strains subjected to mutagenesis and selection for an improved trait of

interest. SNP swapping utilizes high-throughput genome engineering techniques to systematically determine the influence of individual mutations in a mutagenic lineage. Genome sequences are determined for strains across one or more generations of a mutagenic lineage with known performance improvements. High-throughput genome engineering is then used systematically to recapitulate mutations from improved strains in earlier lineage strains, and/or revert mutations in later strains to earlier strain sequences. The performance of these strains is then evaluated and the contribution of each individual mutation on the improved phenotype of interest can be determined. As aforementioned, the microbial strains that result from this process are analyzed/characterized and form the basis for the SNP swap genetic design libraries that can inform microbial strain improvement across host strains.

[0257] Removal of detrimental mutations can provide immediate performance improvements, and consolidation of beneficial mutations in a strain background not subject to mutagenic burden can rapidly and greatly improve strain performance. The various microbial strains produced *via* the SNP swapping process form the HTP genetic design SNP swapping libraries, which are microbial strains comprising the various added/deleted/or consolidated SNPs, but with otherwise identical genetic backgrounds.

[0258] As discussed previously, random mutagenesis and subsequent screening for performance improvements is a commonly used technique for industrial strain improvement, and many strains currently used for large scale manufacturing have been developed using this process iteratively over a period of many years, sometimes decades. Random approaches to generating genomic mutations such as exposure to UV radiation or chemical mutagens such as ethyl methanesulfonate were a preferred method for industrial strain improvements because: 1) industrial organisms may be poorly characterized genetically or metabolically, rendering target selection for directed improvement approaches difficult or impossible; 2) even in relatively well characterized systems, changes that result in industrial performance improvements are difficult to predict and may require perturbation of genes that have no known function, and 3) genetic tools for making directed genomic mutations in a given industrial organism may not be available or very slow and/or difficult to use.

[0259] However, despite the aforementioned benefits of this process, there are also a number of known disadvantages. Beneficial mutations are relatively rare events, and in order to find these mutations with a fixed screening capacity, mutations rates must be sufficiently high. This often

results in unwanted neutral and partly detrimental mutations being incorporated into strains along with beneficial changes. Over time this ‘mutagenic burden’ builds up, resulting in strains with deficiencies in overall robustness and key traits such as growth rates. Eventually ‘mutagenic burden’ renders further improvements in performance through random mutagenesis increasingly difficult or impossible to obtain. Without suitable tools, it is impossible to consolidate beneficial mutations found in discrete and parallel branches of strain lineages.

[0260] SNP swapping is an approach to overcome these limitations by systematically recapitulating or reverting some or all mutations observed when comparing strains within a mutagenic lineage. In this way, both beneficial (‘causative’) mutations can be identified and consolidated, and/or detrimental mutations can be identified and removed. This allows rapid improvements in strain performance that could not be achieved by further random mutagenesis or targeted genetic engineering.

[0261] Removal of genetic burden or consolidation of beneficial changes into a strain with no genetic burden also provides a new, robust starting point for additional random mutagenesis that may enable further improvements.

[0262] In addition, as orthogonal beneficial changes are identified across various, discrete branches of a mutagenic strain lineage, they can be rapidly consolidated into better performing strains. These mutations can also be consolidated into strains that are not part of mutagenic lineages, such as strains with improvements gained by directed genetic engineering.

[0263] Other approaches and technologies exist to randomly recombine mutations between strains within a mutagenic lineage. These include techniques such as protoplast fusion and whole genome shuffling that facilitate genomic recombination across mutated strains. For some industrial microorganisms such as yeast and filamentous fungi, natural mating cycles can also be exploited for pairwise genomic recombination. In this way, detrimental mutations can be removed by ‘back-crossing’ mutants with parental strains and beneficial mutations consolidated. However, these approaches are subject to many limitations that are circumvented using the SNP swapping methods of the present disclosure.

[0264] For example, as these approaches rely on a relatively small number of random recombination crossover events to swap mutations, it may take many cycles of recombination and screening to optimize strain performance. In addition, although natural recombination events are essentially random, they are also subject to genome positional bias and some mutations may be

difficult to address. These approaches also provide little information about the influence of individual mutations without additional genome sequencing and analysis. SNP swapping overcomes these fundamental limitations as it is not a random approach, but rather the systematic introduction or removal of individual mutations across strains.

[0265] In some embodiments, the present disclosure teaches methods for identifying the SNP sequence diversity present among the organisms of a diversity pool. A diversity pool can be a given number n of microbes utilized for analysis, with said microbes' genomes representing the "diversity pool."

[0266] In particular aspects, a diversity pool may be an original parent strain (S_1) with a "baseline" or "reference" genetic sequence at a particular time point ($S_1\text{Gen}_1$) and then any number of subsequent offspring strains (S_{2-n}) that were derived/developed from said S_1 strain and that have a different genome ($S_{2-n}\text{Gen}_{2-n}$), in relation to the baseline genome of S_1 .

[0267] For example, in some embodiments, the present disclosure teaches sequencing the microbial genomes in a diversity pool to identify the SNPs present in each strain. In one embodiment, the strains of the diversity pool are historical microbial production strains. Thus, a diversity pool of the present disclosure can include for example, an industrial reference strain, and one or more mutated industrial strains produced *via* traditional strain improvement programs.

[0268] In some embodiments, the SNPs within a diversity pool are determined with reference to a "reference strain." In some embodiments, the reference strain is a wild-type strain. In other embodiments, the reference strain is an original industrial strain prior to being subjected to any mutagenesis. The reference strain can be defined by the practitioner and does not have to be an original wild-type strain or original industrial strain. The base strain is merely representative of what will be considered the "base," "reference" or original genetic background, by which subsequent strains that were derived, or were developed from said reference strain, are to be compared.

[0269] Once all SNPs in the diversity pool are identified, the present disclosure teaches methods of SNP swapping and screening methods to delineate (*i.e.* quantify and characterize) the effects (*e.g.* creation of a phenotype of interest) of SNPs individually and/or in groups.

[0270] In some embodiments, the SNP swapping methods of the present disclosure comprise the step of introducing one or more SNPs identified in a mutated strain (*e.g.*, a strain from amongst $S_{2-n}\text{Gen}_{2-n}$) to a reference strain ($S_1\text{Gen}_1$) or wild-type strain ("wave up").

[0271] In other embodiments, the SNP swapping methods of the present disclosure comprise the step of removing one or more SNPs identified in a mutated strain (*e.g.*, a strain from amongst $S_{2-n}Gen_{2-n}$) (“wave down”).

[0272] In some embodiments, each generated strain comprising one or more SNP changes (either introducing or removing) is cultured and analyzed under one or more criteria of the present disclosure (*e.g.*, production of a chemical or product of interest). Data from each of the analyzed host strains is associated, or correlated, with the particular SNP, or group of SNPs present in the host strain, and is recorded for future use. Thus, the present disclosure enables the creation of large and highly annotated HTP genetic design microbial strain libraries that are able to identify the effect of a given SNP on any number of microbial genetic or phenotypic traits of interest. The information stored in these HTP genetic design libraries informs the machine learning algorithms of the HTP genomic engineering platform and directs future iterations of the process, which ultimately leads to evolved microbial organisms that possess highly desirable properties/traits.

3. *Start/Stop Codon Exchanges: A Molecular Tool for the Derivation of Start/Stop Codon Microbial Strain Libraries*

[0273] In some embodiments, the present disclosure teaches methods of swapping start and stop codon variants. For example, typical stop codons for *S. cerevisiae* and mammals are TAA (UAA) and TGA (UGA), respectively. The typical stop codon for monocotyledonous plants is TGA (UGA), whereas insects and *E. coli* commonly use TAA (UAA) as the stop codon (Dalphin *et al.* (1996) Nucl. Acids Res. 24: 216-218). In other embodiments, the present disclosure teaches use of the TAG (UAG) stop codons.

[0274] The present disclosure similarly teaches swapping start codons. In some embodiments, the present disclosure teaches use of the ATG (AUG) start codon utilized by most organisms (especially eukaryotes). In some embodiments, the present disclosure teaches that prokaryotes use ATG (AUG) the most, followed by GTG (GUG) and TTG (UUG).

[0275] In other embodiments, the present invention teaches replacing ATG start codons with TTG. In some embodiments, the present invention teaches replacing ATG start codons with GTG. In some embodiments, the present invention teaches replacing GTG start codons with ATG. In some embodiments, the present invention teaches replacing GTG start codons with TTG. In some embodiments, the present invention teaches replacing TTG start codons with ATG. In some embodiments, the present invention teaches replacing TTG start codons with GTG.

[0276] In other embodiments, the present invention teaches replacing TAA stop codons with TAG. In some embodiments, the present invention teaches replacing TAA stop codons with TGA. In some embodiments, the present invention teaches replacing TGA stop codons with TAA. In some embodiments, the present invention teaches replacing TGA stop codons with TAG. In some embodiments, the present invention teaches replacing TAG stop codons with TAA. In some embodiments, the present invention teaches replacing TAG stop codons with TGA.

4. *Stop swap: A Molecular Tool for the Derivation of STOP Swap Microbial Strain Libraries*

[0277] In some embodiments, the present disclosure teaches methods of improving host cell productivity through the optimization of cellular gene transcription. Gene transcription is the result of several distinct biological phenomena, including transcriptional initiation (RNAP recruitment and transcriptional complex formation), elongation (strand synthesis/extension), and transcriptional termination (RNAP detachment and termination). Although much attention has been devoted to the control of gene expression through the transcriptional modulation of genes (*e.g.*, by changing promoters, or inducing regulatory transcription factors), comparatively few efforts have been made towards the modulation of transcription via the modulation of gene terminator sequences.

[0278] The most obvious way that transcription impacts on gene expression levels is through the rate of Pol II initiation, which can be modulated by combinations of promoter or enhancer strength and trans-activating factors (Kadonaga, JT. 2004 “Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors” *Cell*. 2004 Jan 23; 116(2):247-57). In eukaryotes, elongation rate may also determine gene expression patterns by influencing alternative splicing (Cramer P. et al., 1997 “Functional association between promoter structure and transcript alternative splicing.” *Proc Natl Acad Sci U S A*. 1997 Oct 14; 94(21):11456-60). Failed termination on a gene can impair the expression of downstream genes by reducing the accessibility of the promoter to Pol II (Greger IH. et al., 2000 “Balancing transcriptional interference and initiation on the GAL7 promoter of *Saccharomyces cerevisiae*.” *Proc Natl Acad Sci U S A*. 2000 Jul 18; 97(15):8415-20). This process, known as transcriptional interference, is particularly relevant in lower eukaryotes, as they often have closely spaced genes.

[0279] Termination sequences can also affect the expression of the genes to which the sequences belong. For example, studies show that inefficient transcriptional termination in eukaryotes results

in an accumulation of unspliced pre-mRNA (see West, S., and Proudfoot, N.J., 2009 “Transcriptional Termination Enhances Protein Expression in Human Cells” *Mol Cell*. 2009 Feb 13; 33(3-9); 354-364). Other studies have also shown that 3' end processing, can be delayed by inefficient termination (West, S et al., 2008 “Molecular dissection of mammalian RNA polymerase II transcriptional termination.” *Mol Cell*. 2008 Mar 14; 29(5):600-10.). Transcriptional termination can also affect mRNA stability by releasing transcripts from sites of synthesis. Further, strong termination sequences may increase mRNA stability, thus increasing protein abundance and overall pathway activity.

Termination of transcription mechanism in eukaryotes

[0280] Transcriptional termination in eukaryotes operates through terminator signals that are recognized by protein factors associated with the RNA polymerase II. In some embodiments, the cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) transfer from the carboxyl terminal domain of RNA polymerase II to the poly-A signal. In some embodiments, the CPSF and CstF factors also recruit other proteins to the termination site, which then cleave the transcript and free the mRNA from the transcription complex. Termination also triggers polyadenylation of mRNA transcripts. Illustrative examples of validated eukaryotic termination factors, and their conserved structures are discussed in later portions of this document.

Termination of transcription in prokaryotes

[0281] In prokaryotes, two principal mechanisms, termed Rho-independent and Rho-dependent termination, mediate transcriptional termination. Rho-independent termination signals do not require an extrinsic transcription-termination factor, as formation of a stem-loop structure in the RNA transcribed from these sequences along with a series of Uridine (U) residues promotes release of the RNA chain from the transcription complex. Rho-dependent termination, on the other hand, requires a transcription-termination factor called Rho and cis-acting elements on the mRNA. The initial binding site for Rho, the Rho utilization (rut) site, is an extended (~70 nucleotides, sometimes 80-100 nucleotides) single-stranded region characterized by a high cytidine/low guanosine content and relatively little secondary structure in the RNA being synthesized, upstream of the actual terminator sequence. When a polymerase pause site is encountered, termination occurs, and the transcript is released by Rho's helicase activity.

Terminator Swapping (STOP swap)

[0282] In some embodiments, the present disclosure teaches methods of selecting termination sequences (“terminators”) with optimal expression properties to produce beneficial effects on overall-host strain productivity.

[0283] For example, in some embodiments, the present disclosure teaches methods of identifying one or more terminators and/or generating variants of one or more terminators within a host cell, which exhibit a range of expression strengths (*e.g.* terminator ladders discussed *infra*). A particular combination of these identified and/or generated terminators can be grouped together as a terminator ladder, which is explained in more detail below.

[0284] The terminator ladder in question is then associated with a given gene of interest. Thus, if one has terminators T₁-T₈ (representing eight terminators that have been identified and/or generated to exhibit a range of expression strengths when combined with one or more promoters) and associates the terminator ladder with a single gene of interest in a host cell (*i.e.* genetically engineer a host cell with a given terminator operably linked to the 3' end of to a given target gene), then the effect of each combination of the terminators can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered host cells have an otherwise identical genetic background except the particular terminator(s) associated with the target gene. The resultant host cells that are engineered *via* this process form HTP genetic design libraries.

[0285] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed *via* this process, with each member strain being representative of a given terminator operably linked to a particular target gene, in an otherwise identical genetic background, said library being termed a “terminator swap microbial strain library” or “STOP swap microbial strain library.” In the specific context of *E. coli*, the library can be termed a “terminator swap *E. coli* strain library,” or “STOP swap *E. coli* strain library,” but the terms can be used synonymously, as *E. coli* is a specific example of a microbe.

[0286] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given terminator *x* operably linked to a given gene *y*—said collection being termed a “terminator swap library” or “STOP swap library.”

[0287] Further, one can utilize the same terminator ladder comprising terminators T₁-T₈ to engineer microbes, wherein each of the eight terminators is operably linked to 10 different gene

targets. The result of this procedure would be 80 host cell strains that are otherwise assumed genetically identical, except for the particular terminators operably linked to a target gene of interest. These 80 host cell strains could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any database, including without limitation, a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could include, for example, a given terminators' (e.g., T₁-T₈) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more promoters (e.g., T₁-T₈) to a given gene target.

[0288] The aforementioned examples of eight terminators and 10 target genes is merely illustrative, as the concept can be applied with any given number of terminators that have been grouped together based upon exhibition of a range of expression strengths and any given number of target genes. For example, another set of terminators that can be used in the methods provided herein (e.g., STOP Swapping) is the set of terminators found in Table 1.2 with nucleic acid SEQ ID Nos 225, 226, 227, 228, 229, or 230.

[0289] In summary, utilizing various terminators to modulate expression of various genes in an organism is a powerful tool to optimize a trait of interest. The molecular tool of terminator swapping, developed by the inventors, uses a ladder of terminator sequences that have been demonstrated to vary expression of at least one locus under at least one condition. This ladder is then systematically applied to a group of genes in the organism using high-throughput genome engineering. This group of genes is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity.

[0290] The resultant HTP genetic design microbial library of organisms containing a terminator sequence linked to a gene is then assessed for performance in a high-throughput screening model, and terminator-gene linkages which lead to increased performance are determined and the information stored in a database. The collection of genetic perturbations (*i.e.* given terminator *x* linked to a given gene *y*) form a "terminator swap library," which can be utilized as a source of potential genetic alterations to be utilized in microbial engineering processing. Over time, as a

greater set of genetic perturbations is implemented against a greater diversity of microbial backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest. That is in some embodiments, the present disclosures teaches introduction of one or more genetic changes into a host cell based on previous experimental results embedded within the meta data associated with any of the genetic design libraries of the invention.

[0291] Thus, in particular embodiments, terminator swapping is a multi-step process comprising:

[0292] 1. Selecting a set of “x” terminators to act as a “ladder.” Ideally these terminators have been shown to lead to highly variable expression across multiple genomic loci, but the only requirement is that they perturb gene expression in some way.

[0293] 2. Selecting a set of “n” genes to target. This set can be every ORF in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps, STOP swaps, SOLUBILITY TAG swaps, DEGRADATION TAG swaps or SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection. In other embodiments, the “n” targeted genes can comprise non-protein coding genes, including non-coding RNAs.

[0294] 3. High-throughput strain engineering to rapidly and in parallel carry out the following genetic modifications: When a native terminator exists at the 3’ end of target gene n and its sequence is known, replace the native terminator with each of the x terminators in the ladder. When the native terminator does not exist, or its sequence is unknown, insert each of the x terminators in the ladder after the gene stop codon.

[0295] In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the library is an instance of x terminator linked to n target, in an otherwise identical genetic context. As previously described, combinations of terminators can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0296] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0297] This foundational process can be extended to provide further improvements in strain performance by, *inter alia*: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then sequential regeneration of the library of perturbations into an improved member or members of the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches.

[0298] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

5. *Sequence Optimization: A Molecular Tool for the Derivation of Optimized Sequence Microbial Strain Libraries*

[0299] In one embodiment, the methods of the disclosure comprise codon optimizing one or more genes expressed by the host organism. Methods for optimizing codons to improve expression in various hosts are known in the art and are described in the literature (*see* U.S. Pat. App. Pub. No. 2007/0292918, incorporated herein by reference in its entirety). Optimized coding sequences containing codons preferred by a particular prokaryotic or eukaryotic host (*see* also, Murray *et al.* (1989) Nucl. Acids Res. 17:477-508) can be prepared, for example, to increase the rate of translation or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, as compared with transcripts produced from a non-optimized sequence.

[0300] Protein expression is governed by a host of factors including those that affect transcription, mRNA processing, and stability and initiation of translation. Optimization can thus address any of a number of sequence features of any particular gene. As a specific example, a rare codon induced translational pause can result in reduced protein expression. A rare codon induced translational pause includes the presence of codons in the polynucleotide of interest that are rarely used in the

host organism may have a negative effect on protein translation due to their scarcity in the available tRNA pool.

[0301] Alternate translational initiation also can result in reduced heterologous protein expression. Alternate translational initiation can include a synthetic polynucleotide sequence inadvertently containing motifs capable of functioning as a ribosome binding site (RBS). These sites can result in initiating translation of a truncated protein from a gene-internal site. One method of reducing the possibility of producing a truncated protein, which can be difficult to remove during purification, includes eliminating putative internal RBS sequences from an optimized polynucleotide sequence.

[0302] Repeat-induced polymerase slippage can result in reduced heterologous protein expression. Repeat-induced polymerase slippage involves nucleotide sequence repeats that have been shown to cause slippage or stuttering of DNA polymerase which can result in frameshift mutations. Such repeats can also cause slippage of RNA polymerase. In an organism with a high G+C content bias, there can be a higher degree of repeats composed of G or C nucleotide repeats. Therefore, one method of reducing the possibility of inducing RNA polymerase slippage, includes altering extended repeats of G or C nucleotides.

[0303] Interfering secondary structures also can result in reduced heterologous protein expression. Secondary structures can sequester the RBS sequence or initiation codon and have been correlated to a reduction in protein expression. Stemloop structures can also be involved in transcriptional pausing and attenuation. An optimized polynucleotide sequence can contain minimal secondary structures in the RBS and gene coding regions of the nucleotide sequence to allow for improved transcription and translation.

[0304] For example, the optimization process can begin by identifying the desired amino acid sequence to be expressed by the host. From the amino acid sequence a candidate polynucleotide or DNA sequence can be designed. During the design of the synthetic DNA sequence, the frequency of codon usage can be compared to the codon usage of the host expression organism and rare host codons can be removed from the synthetic sequence. Additionally, the synthetic candidate DNA sequence can be modified in order to remove undesirable enzyme restriction sites and add or remove any desired signal sequences, linkers or untranslated regions. The synthetic DNA sequence can be analyzed for the presence of secondary structure that may interfere with the translation process, such as G/C repeats and stem-loop structures.

6. SOLUBILITY TAG swap: A Molecular Tool for the Derivation of SOLUBILITY

TAG Swap Microbial Strain Libraries

[0305] In some embodiments, the present disclosure teaches methods of improving host cell productivity through the optimization of post-translational mechanisms. Traditional strain improvement can often be accomplished through overexpression of pathway genes that produce some molecule of interest. Typically, known pathway genes can be duplicated, or strong promoters can be inserted to drive expression of these genes, and thus increase mRNA transcript levels with the goal of increasing pathway protein abundance to achieve improved rate, titer, or yield from a given pathway. This approach can be applied systematically at the whole-genome scale to identify all genes that can improve strain performance. Another frequently applied approach can be deletion of potentially competing pathway genes with the goal of completely eliminating protein products that may divert carbon from the desired pathway. However, these overexpression and/or deletion strain improvement approaches known in the art can have several limitations.

[0306] Beginning with pathway duplication or strong promoter insertion, the expected effect of increased mRNA transcript levels may not necessarily result in increased protein abundance. Various protein products may have various rate-limiting steps in their production, and this rate-limiting step may not be mRNA transcript levels. In scenarios where mRNA transcription is not the rate limiting step, it is possible that post-translational mechanisms may be impacting overall protein abundance. For example, the presence of protein solubility tags may be used to increase the abundance of correctly folded, active protein that can contribute to production of a target molecule, whereas simply increasing mRNA transcript levels may lead only to an increase in misfolded, inactive protein. Effects exerted protein solubility tags can also be made to be tunable depending on the sequence of the solubility tag that is used, enabling precise optimization of the target phenotype.

Protein Solubility Tag Swapping (SOLUBILITY TAG swap)

[0307] In some embodiments, the present disclosure teaches methods of selecting protein solubility tag sequences (“solubility tags”) with optimal protein solubility properties to produce beneficial effects on overall-host strain productivity.

[0308] For example, in some embodiments, the present disclosure teaches methods of identifying one or more protein solubility tags and/or generating variants of one or more protein solubility tags within a host cell, which exhibit a range of solubility strengths (*e.g.* protein solubility tags

discussed *infra*). A particular combination of these identified and/or generated protein solubility tags can be grouped together as a protein solubility tag ladder, which is explained in more detail below.

[0309] The protein solubility tag ladder in question is then associated with a given gene of interest. Thus, if one has protein solubility tags PST₁-PST₄ (*see* Table 17) representing a subset of protein solubility tags that have been identified from Costa et al., *Front Microbiol.* 2014; 5: 63 to enhance protein solubility and also be smaller than 100 amino acids and associates the protein solubility tag ladder with a single gene of interest in a host cell (*i.e.* genetically engineer a host cell with a given protein solubility tag operably linked to a given target gene to generate a target protein tagged at either the N-terminus or the C-terminus). The effect of each combination of the protein solubility tag can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered host cells have an otherwise identical genetic background except the particular solubility tag(s) associated with the target gene. The resultant host cells that are engineered *via* this process form HTP genetic design libraries.

[0310] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed *via* this process, with each member strain being representative of a given protein solubility tag operably linked to a particular target protein, in an otherwise identical genetic background, said library being termed a “solubility tag swap microbial strain library” or “SOLUBILITY TAG swap microbial strain library.” In the specific context of *E. coli*, the library can be termed a “SOLUBILITY TAG swap *E. coli* strain library,” or “SOLUBILITY TAG swap *E. coli* strain library,” but the terms can be used synonymously, as *E. coli* is a specific example of a microbe.

[0311] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given protein solubility tag *x* operably linked to a given gene *y*—said collection being termed a “protein solubility tag swap library” or “SOLUBILITY TAG swap library.”

[0312] Further, one can utilize the same protein solubility tag ladder comprising protein solubility tag PST₁-PST₄ to engineer microbes, wherein each of the four protein solubility tags is operably linked to 10 different gene targets. The result of this procedure would be 40 host cell strains that are otherwise assumed genetically identical, except for the particular protein solubility tags operably linked to a target gene of interest. These 40 host cell strains could be appropriately

screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any database, including without limitation, a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could include, for example, a given protein solubility tag (*e.g.*, PST₁-PST₄) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more solubility tags (*e.g.*, PST₁-PST₄) to a given gene target.

[0313] The aforementioned examples of four protein solubility tags and 10 target genes is merely illustrative, as the concept can be applied with any given number of protein solubility tags that have been grouped together based upon exhibition of a range of solubility strengths and any given number of target genes.

[0314] In summary, utilizing various protein solubility tags to modulate solubility of various proteins in an organism is a powerful tool to optimize a trait of interest. The molecular tool of protein solubility tag swapping, developed by the inventors, uses a ladder of protein solubility tag sequences that have been demonstrated to vary solubility (*e.g.*, enhance) of at least one protein under at least one condition. This ladder is then systematically applied to a group of genes in the organism using high-throughput genome engineering. This group of genes is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity.

[0315] The resultant HTP genetic design microbial library of organisms containing a protein solubility tag sequence linked to a gene is then assessed for performance in a high-throughput screening model, and protein solubility tag -gene linkages which lead to increased performance are determined and the information stored in a database. The collection of genetic perturbations (*i.e.* given protein solubility tag *x* linked to a given gene *y*) form a “protein solubility tag swap library,” which can be utilized as a source of potential genetic alterations to be utilized in microbial engineering processing. Over time, as a greater set of genetic perturbations is implemented against a greater diversity of microbial backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest. That is in some embodiments, the present disclosures

teaches introduction of one or more genetic changes into a host cell based on previous experimental results embedded within the meta data associated with any of the genetic design libraries of the invention.

[0316] Thus, in particular embodiments, protein solubility tag swapping is a multi-step process comprising:

[0317] 1. Selecting a set of “x” protein solubility tags to act as a “ladder.” Ideally these protein solubility tags have been shown to lead to enhanced protein solubility across multiple genomic loci, but the only requirement is that they perturb solubility in some way.

[0318] 2. Selecting a set of “n” genes to target. This set can be every ORF in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps, STOP swaps, DEGRADATION TAG swaps or SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection.

[0319] 3. High-throughput strain engineering to rapidly and in parallel carry out the following genetic modifications: When a native protein solubility tag exists within a target gene n and its sequence is known, replace the native protein solubility tag with each of the x protein solubility tags in the ladder. When the native protein solubility tag does not exist, or its sequence is unknown, insert each of the x protein solubility tags in the ladder.

[0320] In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the library is an instance of x protein solubility tag linked to n target, in an otherwise identical genetic context. As previously described, combinations of protein solubility tags can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0321] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0322] This foundational process can be extended to provide further improvements in strain performance by, *inter alia*: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then

sequential regeneration of the library of perturbations into an improved member or members of the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches.

[0323] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

7. DEGRADATION TAG swap: A Molecular Tool for the Derivation of DEGRADATION TAG Swap Microbial Strain Libraries

[0324] Further to the above embodiments regarding methods for improving host cell productivity through the optimization of post-translational mechanisms, a strategy of gene deletion may also have drawbacks that can be addressed by the protein degradation tags (as well as terminators and protein solubility tags) of the present invention. Wholesale deletion of a gene and its corresponding protein product can in some cases impose a drastic modification to the cell. A more precise and tunable response may be achieved through libraries of protein degradation tags that target a protein for degradation at variable rates. This approach can also have the benefit of allowing modulation of protein products that may be essential for cell survival and would not be viable if completely deleted. As these degradation tags also function at a post-translational level, they may be able to address scenarios where altered mRNA transcript levels do not result in altered protein levels as described above.

Protein Degradation Tag Swapping (DEGRADATION TAG swap)

[0325] In some embodiments, the present disclosure teaches methods of selecting protein degradation tag sequences (“degradation tags”) with optimal protein degradation or protein level modulation properties to produce beneficial effects on overall-host strain productivity.

[0326] For example, in some embodiments, the present disclosure teaches methods of identifying one or more protein degradation tags and/or generating variants of one or more protein degradation tags within a host cell, which exhibit a range of degradation strengths or modulate the levels of target proteins (*e.g.* protein degradation tags discussed *infra*). A particular combination of these

identified and/or generated protein degradation tags can be grouped together as a protein degradation tag ladder, which is explained in more detail below.

[0327] The protein degradation tag ladder in question is then associated with a given gene of interest. Thus, if one has protein degradation tags PDT₁-PDT₈ (*see* Table 18) representing a subset of protein degradation tags that have been identified from various sources as detailed in Table 18) and associates the protein degradation tag ladder with a single gene of interest in a host cell (*i.e.* genetically engineer a host cell with a given protein degradation tag operably linked to a given target gene), then the effect of each combination of the protein degradation tag can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered host cells have an otherwise identical genetic background except the particular degradation tag(s) associated with the target gene. The resultant host cells that are engineered *via* this process form HTP genetic design libraries.

[0328] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed *via* this process, with each member strain being representative of a given protein degradation tag operably linked to a particular target protein, in an otherwise identical genetic background, said library being termed a “degradation tag swap microbial strain library” or “DEGRADATION TAG swap microbial strain library.” In the specific context of *E. coli*, the library can be termed a “DEGRADATION TAG swap *E. coli* strain library,” or “DEGRADATION TAG swap *E. coli* strain library,” but the terms can be used synonymously, as *E. coli* is a specific example of a microbe.

[0329] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given protein degradation tag *x* operably linked to a given gene *y*—said collection being termed a “protein degradation tag swap library” or “DEGRADATION TAG swap library.”

[0330] Further, one can utilize the same protein degradation tag ladder comprising protein degradation tag PDT₁-PDT₈ to engineer microbes, wherein each of the eight protein degradation tags is operably linked to 10 different gene targets. The result of this procedure would be 80 host cell strains that are otherwise assumed genetically identical, except for the particular protein degradation tags operably linked to a target gene of interest. These 80 host cell strains could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information

and data that can be stored in any database, including without limitation, a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could include, for example, a given protein degradation tag (e.g., PDT₁-PDT₈) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more degradation tags (e.g., PDT₁-PDT₈) to a given gene target.

[0331] The aforementioned examples of eight protein degradation tags and 10 target genes is merely illustrative, as the concept can be applied with any given number of protein degradation tags that have been grouped together based upon exhibition of a range of degradation strengths and any given number of target genes.

[0332] In summary, utilizing various protein degradation tags to modulate degradation of various proteins in an organism is a powerful tool to optimize a trait of interest. The molecular tool of protein degradation tag swapping, developed by the inventors, uses a ladder of protein degradation tag sequences that have been demonstrated to vary degradation (e.g., enhance) of at least one protein under at least one condition. This ladder is then systematically applied to a group of genes in the organism using high-throughput genome engineering. This group of genes is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity.

[0333] The resultant HTP genetic design microbial library of organisms containing a protein degradation tag sequence linked to a gene is then assessed for performance in a high-throughput screening model, and protein degradation tag -gene linkages which lead to increased performance are determined and the information stored in a database. The collection of genetic perturbations (*i.e.* given protein degradation tag *x* linked to a given gene *y*) form a “protein degradation tag swap library,” which can be utilized as a source of potential genetic alterations to be utilized in microbial engineering processing. Over time, as a greater set of genetic perturbations is implemented against a greater diversity of microbial backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest. That is in some embodiments, the present disclosures teaches introduction of one or more genetic changes into a host cell based on previous experimental

results embedded within the meta data associated with any of the genetic design libraries of the invention.

[0334] Thus, in particular embodiments, protein degradation tag swapping is a multi-step process comprising:

[0335] 1. Selecting a set of “x” protein degradation tags to act as a “ladder.” Ideally these protein degradation tags have been shown to lead to enhanced protein degradation across multiple genomic loci, but the only requirement is that they perturb degradation in some way.

[0336] 2. Selecting a set of “n” genes to target. This set can be every ORF in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps, STOP swaps, SOLUBILITY TAG swaps or SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection.

[0337] 3. High-throughput strain engineering to rapidly and in parallel carry out the following genetic modifications: When a native protein degradation tag exists within a target gene n and its sequence is known, replace the native protein degradation tag with each of the x protein degradation tags in the ladder. When the native protein degradation tag does not exist, or its sequence is unknown, insert each of the x protein degradation tags in the ladder.

[0338] In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the library is an instance of x protein degradation tag linked to n target, in an otherwise identical genetic context. As previously described, combinations of protein degradation tags can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0339] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0340] This foundational process can be extended to provide further improvements in strain performance by, *inter alia*: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then sequential regeneration of the library of perturbations into an improved member or members of

the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches.

[0341] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

8. *Epistasis Mapping – A Predictive Analytical Tool Enabling Beneficial Genetic Consolidations*

[0342] In some embodiments, the present disclosure teaches epistasis mapping methods for predicting and combining beneficial genetic alterations into a host cell. The genetic alterations may be created by any of the aforementioned HTP molecular tool sets (*e.g.*, promoter swaps, SNP swaps, start/stop codon exchanges, sequence optimization, protein solubility tag swaps, protein degradation tag swaps and STOP swaps) and the effect of those genetic alterations would be known from the characterization of the derived HTP genetic design microbial strain libraries. Thus, as used herein, the term epistasis mapping includes methods of identifying combinations of genetic alterations (*e.g.*, beneficial SNPs or beneficial promoter/target gene associations) that are likely to yield increases in host performance.

[0343] In embodiments, the epistasis mapping methods of the present disclosure are based on the idea that the combination of beneficial mutations from two different functional groups is more likely to improve host performance, as compared to a combination of mutations from the same functional group. *See, e.g.*, Costanzo, The Genetic Landscape of a Cell, *Science*, Vol. 327, Issue 5964, Jan. 22, 2010, pp. 425-431 (incorporated by reference herein in its entirety).

[0344] Mutations from the same functional group are more likely to operate by the same mechanism, and are thus more likely to exhibit negative or neutral epistasis on overall host performance. In contrast, mutations from different functional groups are more likely to operate by independent mechanisms, which can lead to improved host performance and in some instances synergistic effects. For example, referring to Figure 19, *lysA* and *zwf* are genes that operate in different pathways to achieve the production of lysine. Based upon the dissimilarity in the

individual performance of those genes, genetic changes using those genes should result in additive consolidation effects. This was borne out in the actual measurement of the consolidated effects of the combination of *lysA* and *zwf*, as shown in Figure 16B and Example 6.

[0345] Thus, in some embodiments, the present disclosure teaches methods of analyzing SNP mutations to identify SNPs predicted to belong to different functional groups. In some embodiments, SNP functional group similarity is determined by computing the cosine similarity of mutation interaction profiles (similar to a correlation coefficient, *see* Figure 16A). The present disclosure also illustrates comparing SNPs *via* a mutation similarity matrix (*see* Figure 15 for example analysis conducted in *Corynebacterium*) or dendrogram (*see* Figure 16A for example analysis conducted in *Corynebacterium*).

[0346] Thus, the epistasis mapping procedure provides a method for grouping and/or ranking a diversity of genetic mutations applied in one or more genetic backgrounds for the purposes of efficient and effective consolidations of said mutations into one or more genetic backgrounds.

[0347] In aspects, consolidation is performed with the objective of creating novel strains which are optimized for the production of target biomolecules. Through the taught epistasis mapping procedure, it is possible to identify functional groupings of mutations, and such functional groupings enable a consolidation strategy that minimizes undesirable epistatic effects.

[0348] As previously explained, the optimization of microbes for use in industrial fermentation is an important and difficult problem, with broad implications for the economy, society, and the natural world. Traditionally, microbial engineering has been performed through a slow and uncertain process of random mutagenesis. Such approaches leverage the natural evolutionary capacity of cells to adapt to artificially imposed selection pressure. Such approaches are also limited by the rarity of beneficial mutations, the ruggedness of the underlying fitness landscape, and more generally underutilize the state of the art in cellular and molecular biology.

[0349] Modern approaches leverage new understanding of cellular function at the mechanistic level and new molecular biology tools to perform targeted genetic manipulations to specific phenotypic ends. In practice, such rational approaches are confounded by the underlying complexity of biology. Causal mechanisms are poorly understood, particularly when attempting to combine two or more changes that each has an observed beneficial effect. Sometimes such consolidations of genetic changes yield positive outcomes (measured by increases in desired phenotypic activity), although the net positive outcome may be lower than expected and in some

cases higher than expected. In other instances, such combinations produce either net neutral effect or a net negative effect. This phenomenon is referred to as epistasis, and is one of the fundamental challenges to microbial engineering (and genetic engineering generally).

[0350] As aforementioned, the present HTP genomic engineering platform solves many of the problems associated with traditional microbial engineering approaches. The present HTP platform uses automation technologies to perform hundreds or thousands of genetic mutations at once. In particular aspects, unlike the rational approaches described above, the disclosed HTP platform enables the parallel construction of thousands of mutants to more effectively explore large subsets of the relevant genomic space, as disclosed in U.S. Application No. 15/140,296, entitled Microbial Strain Design System And Methods For Improved Large-Scale Production Of Engineered Nucleotide Sequences, incorporated by reference herein in its entirety. By trying “everything,” the present HTP platform sidesteps the difficulties induced by our limited biological understanding.

[0351] However, at the same time, the present HTP platform faces the problem of being fundamentally limited by the combinatorial explosive size of genomic space, and the effectiveness of computational techniques to interpret the generated data sets given the complexity of genetic interactions. Techniques are needed to explore subsets of vast combinatorial spaces in ways that maximize non-random selection of combinations that yield desired outcomes.

[0352] Somewhat similar HTP approaches have proved effective in the case of enzyme optimization. In this niche problem, a genomic sequence of interest (on the order of 1000 bases), encodes a protein chain with some complicated physical configuration. The precise configuration is determined by the collective electromagnetic interactions between its constituent atomic components. This combination of short genomic sequence and physically constrained folding problem lends itself specifically to greedy optimization strategies. That is, it is possible to individually mutate the sequence at every residue and shuffle the resulting mutants to effectively sample local sequence space at a resolution compatible with the Sequence Activity Response modeling.

[0353] However, for full genomic optimizations for biomolecules, such residue-centric approaches are insufficient for some important reasons. *First*, because of the exponential increase in relevant sequence space associated with genomic optimizations for biomolecules. *Second*, because of the added complexity of regulation, expression, and metabolic interactions in

biomolecule synthesis. The present inventors have solved these problems *via* the taught epistasis mapping procedure.

[0354] The taught method for modeling epistatic interactions, between a collection of mutations for the purposes of more efficient and effective consolidation of said mutations into one or more genetic backgrounds, is groundbreaking and highly needed in the art.

[0355] When describing the epistasis mapping procedure, the terms “more efficient” and “more effective” refers to the avoidance of undesirable epistatic interactions among consolidation strains with respect to particular phenotypic objectives.

[0356] As the process has been generally elaborated upon above, a more specific workflow example will now be described.

[0357] *First*, one begins with a library of M mutations and one or more genetic backgrounds (*e.g.*, parent bacterial strains). Neither the choice of library nor the choice of genetic backgrounds is specific to the method described here. But in a particular implementation, a library of mutations may include exclusively, or in combination: SNP swap libraries, Promoter swap libraries, or any other mutation library described herein.

[0358] In one implementation, only a single genetic background is provided. In this case, a collection of distinct genetic backgrounds (microbial mutants) will first be generated from this single background. This may be achieved by applying the primary library of mutations (or some subset thereof) to the given background for example, application of a HTP genetic design library of particular SNPs or a HTP genetic design library of particular promoters to the given genetic background, to create a population (perhaps 100's or 1,000's) of microbial mutants with an identical genetic background except for the particular genetic alteration from the given HTP genetic design library incorporated therein. As detailed below, this embodiment can lead to a combinatorial library or pairwise library.

[0359] In another implementation, a collection of distinct known genetic backgrounds may simply be given. As detailed below, this embodiment can lead to a subset of a combinatorial library.

[0360] In a particular implementation, the number of genetic backgrounds and genetic diversity between these backgrounds (measured in number of mutations or sequence edit distance or the like) is determined to maximize the effectiveness of this method.

[0361] A genetic background may be a natural, native or wild-type strain or a mutated, engineered strain. N distinct background strains may be represented by a vector **b**. In one example, the

background \mathbf{b} may represent engineered backgrounds formed by applying N primary mutations $\mathbf{m}_0 = (m_1, m_2, \dots, m_N)$ to a wild-type background strain b_0 to form the N mutated background strains $\mathbf{b} = \mathbf{m}_0 b_0 = (m_1 b_0, m_2 b_0, \dots, m_N b_0)$, where $m_i b_0$ represents the application of mutation m_i to background strain b_0 .

[0362] In either case (*i.e.* a single provided genetic background or a collection of genetic backgrounds), the result is a collection of N genetically distinct backgrounds. Relevant phenotypes are measured for each background.

[0363] *Second*, each mutation in a collection of M mutations \mathbf{m}_1 is applied to each background within the collection of N background strains \mathbf{b} to form a collection of $M \times N$ mutants. In the implementation where the N backgrounds were themselves obtained by applying the primary set of mutations \mathbf{m}_0 (as described above), the resulting set of mutants will sometimes be referred to as a combinatorial library or a pairwise library. In another implementation, in which a collection of known backgrounds has been provided explicitly, the resulting set of mutants may be referred to as a subset of a combinatorial library. Similar to generation of engineered background vectors, in embodiments, the input interface 202 (see, Figure 31) receives the mutation vector \mathbf{m}_1 and the background vector \mathbf{b} , and a specified operation such as cross product.

[0364] Continuing with the engineered background example above, forming the $M \times N$ combinatorial library may be represented by the matrix formed by $\mathbf{m}_1 \times \mathbf{m}_0 b_0$, the cross product of \mathbf{m}_1 applied to the N backgrounds of $\mathbf{b} = \mathbf{m}_0 b_0$, where each mutation in \mathbf{m}_1 is applied to each background strain within \mathbf{b} . Each i th row of the resulting $M \times N$ matrix represents the application of the i th mutation within \mathbf{m}_1 to all the strains within background collection \mathbf{b} . In one embodiment, $\mathbf{m}_1 = \mathbf{m}_0$ and the matrix represents the pairwise application of the same mutations to starting strain b_0 . In that case, the matrix is symmetric about its diagonal ($M=N$), and the diagonal may be ignored in any analysis since it represents the application of the same mutation twice.

[0365] In embodiments, forming the $M \times N$ matrix may be achieved by inputting into the input interface 202 (see, Figure 31) the compound expression $\mathbf{m}_1 \times \mathbf{m}_0 b_0$. The component vectors of the expression may be input directly with their elements explicitly specified, via one or more DNA specifications, or as calls to the library 206 to enable retrieval of the vectors during interpretation by interpreter 204. As described in U.S. Patent Application, Serial No. 15/140,296, entitled “Microbial Strain Design System and Methods for Improved Large Scale Production of Engineered Nucleotide Sequences,” via the interpreter 204, execution engine 207, order placement

engine 208, and factory 210, the LIMS system 200 generates the microbial strains specified by the input expression.

[0366] *Third*, with reference to Figure 42, the analysis equipment 214 (see, Figure 31) measures phenotypic responses for each mutant within the MxN combinatorial library matrix (4202). As such, the collection of responses can be construed as an M x N Response Matrix \mathbf{R} . Each element of \mathbf{R} may be represented as $r_{ij} = y(m_i, m_j)$, where y represents the response (performance) of background strain b_j within engineered collection \mathbf{b} as mutated by mutation m_i . For simplicity, and practicality, we assume pairwise mutations where $\mathbf{m}_1 = \mathbf{m}_0$. Where, as here, the set of mutations represents a pairwise mutation library, the resulting matrix may also be referred to as a gene interaction matrix or, more particularly, as a mutation interaction matrix.

[0367] Those skilled in the art will recognize that, in some embodiments, operations related to epistatic effects and predictive strain design may be performed entirely through automated means of the LIMS system 200, e.g., by the analysis equipment 214 (see, Figure 31), or by human implementation, or through a combination of automated and manual means. When an operation is not fully automated, the elements of the LIMS system 200, e.g., analysis equipment 214, may, for example, receive the results of the human performance of the operations rather than generate results through its own operational capabilities. As described elsewhere herein, components of the LIMS system 200, such as the analysis equipment 214, may be implemented wholly or partially by one or more computer systems. In some embodiments, in particular where operations related to predictive strain design are performed by a combination of automated and manual means, the analysis equipment 214 may include not only computer hardware, software or firmware (or a combination thereof), but also equipment operated by a human operator such as that listed in Table 5 below, e.g., the equipment listed under the category of “Evaluate performance.”

[0368] *Fourth*, the analysis equipment 212 (see, Figure 31) normalizes the response matrix. Normalization consists of a manual and/or, in this embodiment, automated processes of adjusting measured response values for the purpose of removing bias and/or isolating the relevant portions of the effect specific to this method. With respect to Figure 42, the first step 4202 may include obtaining normalized measured data. In general, in the claims directed to predictive strain design and epistasis mapping, the terms “performance measure” or “measured performance” or the like may be used to describe a metric that reflects measured data, whether raw or processed in some manner, e.g., normalized data. In a particular implementation, normalization may be performed by

subtracting a previously measured background response from the measured response value. In that implementation, the resulting response elements may be formed as $r_{ij} = y(m_i, m_j) - y(m_j)$, where $y(m_j)$ is the response of the engineered background strain b_j within engineered collection \mathbf{b} caused by application of primary mutation m_j to parent strain \mathbf{b}_0 . Note that each row of the normalized response matrix is treated as a response profile for its corresponding mutation. That is, the i th row describes the relative effect of the corresponding mutation m_i applied to all the background strains b_j for $j=1$ to N .

[0369] With respect to the example of pairwise mutations, the combined performance/response of strains resulting from two mutations may be greater than, less than, or equal to the performance/response of the strain to each of the mutations individually. This effect is known as “epistasis,” and may, in some embodiments, be represented as $e_{ij} = y(m_i, m_j) - (y(m_i) + y(m_j))$. Variations of this mathematical representation are possible, and may depend upon, for example, how the individual changes biologically interact. As noted above, mutations from the same functional group are more likely to operate by the same mechanism, and are thus more likely to exhibit negative or neutral epistasis on overall host performance. In contrast, mutations from different functional groups are more likely to operate by independent mechanisms, which can lead to improved host performance by reducing redundant mutative effects, for example. Thus, mutations that yield dissimilar responses are more likely to combine in an additive manner than mutations that yield similar responses. This leads to the computation of similarity in the next step.

[0370] *Fifth*, the analysis equipment 214 measures the similarity among the responses—in the pairwise mutation example, the similarity between the effects of the i th mutation and j th (e.g., primary) mutation within the response matrix (4204). Recall that the i th row of \mathbf{R} represents the performance effects of the i th mutation m_i on the N background strains, each of which may be itself the result of engineered mutations as described above. Thus, the similarity between the effects of the i th and j th mutations may be represented by the similarity s_{ij} between the i th and j th rows, ρ_i and ρ_j , respectively, to form a similarity matrix \mathbf{S} , an example of which is illustrated in Figure 15. Similarity may be measured using many known techniques, such as cross-correlation or absolute cosine similarity, e.g., $s_{ij} = \text{abs}(\cos(\rho_i, \rho_j))$.

[0371] As an alternative or supplement to a metric like cosine similarity, response profiles may be clustered to determine degree of similarity. Clustering may be performed by use of a distance-based clustering algorithms (e.g. k-mean, hierarchical agglomerative, etc.) in conjunction with

suitable distance measure (e.g. Euclidean, Hamming, etc). Alternatively, clustering may be performed using similarity based clustering algorithms (e.g. spectral, min-cut, etc.) with a suitable similarity measure (e.g. cosine, correlation, etc). Of course, distance measures may be mapped to similarity measures and vice-versa via any number of standard functional operations (e.g., the exponential function). In one implementation, hierarchical agglomerative clustering may be used in conjunction absolute cosine similarity. (See Figure 16A for example analysis conducted in *Corynebacterium*).

[0372] As an example of clustering, let C be a clustering of mutations m_i into k distinct clusters. Let C be the cluster membership matrix, where c_{ij} is the degree to which mutation i belongs to cluster j , a value between 0 and 1. The cluster-based similarity between mutations i and j is then given by $C_i \times C_j$ (the dot product of the i th and j th rows of C). In general, the cluster-based similarity matrix is given by CC^T (that is, C times C -transpose). In the case of hard-clustering (a mutation belongs to exactly one cluster), the similarity between two mutations is 1 if they belong to the same cluster and 0 if not.

[0373] As is described in Costanzo, The Genetic Landscape of a Cell, *Science*, Vol. 327, Issue 5964, Jan. 22, 2010, pp. 425-431 (incorporated by reference herein in its entirety), such a clustering of mutation response profiles relates to an approximate mapping of a cell's underlying functional organization. That is, mutations that cluster together tend to be related by an underlying biological process or metabolic pathway. Such mutations are referred to herein as a "functional group." The key observation of this method is that if two mutations operate by the same biological process or pathway, then observed effects (and notably observed benefits) may be redundant. Conversely, if two mutations operate by distant mechanism, then it is less likely that beneficial effects will be redundant.

[0374] *Sixth*, based on the epistatic effect, the analysis equipment 214 selects pairs of mutations that lead to dissimilar responses, e.g., their cosine similarity metric falls below a similarity threshold, or their responses fall within sufficiently separated clusters, (e.g., in Figure 15 and Figure 16A for example analyses conducted in *Corynebacterium*) as shown in Figure 42 (4206). Based on their dissimilarity, the selected pairs of mutations should consolidate into background strains better than similar pairs.

[0375] Based upon the selected pairs of mutations that lead to sufficiently dissimilar responses, the LIMS system (e.g., all of or some combination of interpreter 204, execution engine 207, order

placer 208, and factory 210) may be used to design microbial strains having those selected mutations (4208). In embodiments, as described below and elsewhere herein, epistatic effects may be built into, or used in conjunction with the predictive model to weight or filter strain selection.

[0376] It is assumed that it is possible to estimate the performance (a.k.a. score) of a hypothetical strain obtained by consolidating a collection of mutations from the library into a particular background *via* some preferred predictive model. A representative predictive model utilized in the taught methods is provided in the below section entitled “Predictive Strain Design” that is found in the larger section of: “Computational Analysis and Prediction of Effects of Genome-Wide Genetic Design Criteria.”

[0377] When employing a predictive strain design technique such as linear regression, the analysis equipment 214 may restrict the model to mutations having low similarity measures by, e.g., filtering the regression results to keep only sufficiently dissimilar mutations. Alternatively, the predictive model may be weighted with the similarity matrix. For example, some embodiments may employ a weighted least squares regression using the similarity matrix to characterize the interdependencies of the proposed mutations. As an example, weighting may be performed by applying the "kernel" trick to the regression model. (To the extent that the "kernel trick" is general to many machine learning modeling approaches, this re-weighting strategy is not restricted to linear regression.)

[0378] Such methods are known to one skilled in the art. In embodiments, the kernel is a matrix having elements $1 - w * s_{ij}$ where 1 is an element of the identity matrix, and w is a real value between 0 and 1 . When $w = 0$, this reduces to a standard regression model. In practice, the value of w will be tied to the accuracy (r^2 value or root mean square error (RMSE)) of the predictive model when evaluated against the pairwise combinatorial constructs and their associate effects $y(m_i, m_j)$. In one simple implementation, w is defined as $w = 1 - r^2$. In this case, when the model is fully predictive, $w = 1 - r^2 = 0$ and consolidation is based solely on the predictive model and epistatic mapping procedure plays no role. On the other hand, when the predictive model is not predictive at all, $w = 1 - r^2 = 1$ and consolidation is based solely on the epistatic mapping procedure. During each iteration, the accuracy can be assessed to determine whether model performance is improving.

[0379] It should be clear that the epistatic mapping procedure described herein does not depend on which model is used by the analysis equipment 214. Given such a predictive model, it is

possible to score and rank all hypothetical strains accessible to the mutation library *via* combinatorial consolidation.

[0380] In some embodiments, to account for epistatic effects, the dissimilar mutation response profiles may be used by the analysis equipment 214 to augment the score and rank associated with each hypothetical strain from the predictive model. This procedure may be thought of broadly as a re-weighting of scores, so as to favor candidate strains with dissimilar response profiles (e.g., strains drawn from a diversity of clusters). In one simple implementation, a strain may have its score reduced by the number of constituent mutations that do not satisfy the dissimilarity threshold or that are drawn from the same cluster (with suitable weighting). In a particular implementation, a hypothetical strain's performance estimate may be reduced by the sum of terms in the similarity matrix associated with all pairs of constituent mutations associated with the hypothetical strain (again with suitable weighting). Hypothetical strains may be re-ranked using these augmented scores. In practice, such re-weighting calculations may be performed in conjunction with the initial scoring estimation.

[0381] The result is a collection of hypothetical strains with score and rank augmented to more effectively avoid confounding epistatic interactions. Hypothetical strains may be constructed at this time, or they may be passed to another computational method for subsequent analysis or use.

[0382] Those skilled in the art will recognize that epistasis mapping and iterative predictive strain design as described herein are not limited to employing only pairwise mutations, but may be expanded to the simultaneous application of many more mutations to a background strain. In another embodiment, additional mutations may be applied sequentially to strains that have already been mutated using mutations selected according to the predictive methods described herein. In another embodiment, epistatic effects are imputed by applying the same genetic mutation to a number of strain backgrounds that differ slightly from each other, and noting any significant differences in positive response profiles among the modified strain backgrounds.

Organisms Amenable to Genetic Design

[0383] The disclosed HTP genomic engineering platform is exemplified with industrial microbial cell cultures (e.g., *Corynebacterium*), but is applicable to any host cell organism where desired traits can be identified in a population of genetic mutants.

[0384] Further, as set forth in the introduction, the current disclosure provides for a HTP genomic engineering platform to improve host cell characteristics in *E. coli* systems and solves many problems that have previously prevented the development of such a system in *E. coli*.

[0385] Thus, as used herein, the term “microorganism” should be taken broadly. It includes, but is not limited to, the two prokaryotic domains, Bacteria and Archaea, as well as certain eukaryotic fungi and protists. However, in certain aspects, “higher” eukaryotic organisms such as insects, plants, and animals can be utilized in the methods taught herein.

[0386] Suitable host cells include, but are not limited to: bacterial cells, algal cells, plant cells, fungal cells, insect cells, and mammalian cells. In one illustrative embodiment, suitable host cells include *E. coli* (e.g., SHuffle™ competent *E. coli* available from New England BioLabs in Ipswich, Mass.). The *E. coli* genome is 4,646,332 bp in size (see Figure 52).

[0387] Suitable host strains of the *E. coli* species comprise: Enterotoxigenic *E. coli* (ETEC), Enteropathogenic *E. coli* (EPEC), Enteroinvasive *E. coli* (EIEC), Enterohemorrhagic *E. coli* (EHEC), Uropathogenic *E. coli* (UPEC), Verotoxin-producing *E. coli*, *E. coli* O157:H7, *E. coli* O104:H4, *Escherichia coli* O121, *Escherichia coli* O104:H21, *Escherichia coli* K1, and *Escherichia coli* NC101. In some embodiments, the present disclosure teaches genomic engineering of *E. coli* K12, *E. coli* B, and *E. coli* C.

[0388] In some embodiments, the present disclosure teaches genomic engineering of *E. coli* strains NCTC 12757, NCTC 12779, NCTC 12790, NCTC 12796, NCTC 12811, ATCC 11229, ATCC 25922, ATCC 8739, DSM 30083, BC 5849, BC 8265, BC 8267, BC 8268, BC 8270, BC 8271, BC 8272, BC 8273, BC 8276, BC 8277, BC 8278, BC 8279, BC 8312, BC 8317, BC 8319, BC 8320, BC 8321, BC 8322, BC 8326, BC 8327, BC 8331, BC 8335, BC 8338, BC 8341, BC 8344, BC 8345, BC 8346, BC 8347, BC 8348, BC 8863, and BC 8864.

[0389] In some embodiments, the present disclosure teaches verocytotoxigenic *E. coli* (VTEC), such as strains BC 4734 (O26:H11), BC 4735 (O157:H-), BC 4736, BC 4737 (n.d.), BC 4738 (O157:H7), BC 4945 (O26:H-), BC 4946 (O157:H7), BC 4947 (O111:H-), BC 4948 (O157:H), BC 4949 (O5), BC 5579 (O157:H7), BC 5580 (O157:H7), BC 5582 (O3:H), BC 5643 (O2:H5), BC 5644 (O128), BC 5645 (O55:H-), BC 5646 (O69:H-), BC 5647 (O101:H9), BC 5648 (O103:H2), BC 5850 (O22:H8), BC 5851 (O55:H-), BC 5852 (O48:H21), BC 5853 (O26:H11), BC 5854 (O157:H7), BC 5855 (O157:H-), BC 5856 (O26:H-), BC 5857 (O103:H2), BC 5858 (O26:H11), BC 7832, BC 7833 (O raw form:H-), BC 7834 (ONT:H-), BC 7835 (O103:H2), BC

7836 (O57:H-), BC 7837 (ONT:H-), BC 7838, BC 7839 (O128:H2), BC 7840 (O157:H-), BC 7841 (O23:H-), BC 7842 (O157:H-), BC 7843, BC 7844 (O157:H-), BC 7845 (O103:H2), BC 7846 (O26:H11), BC 7847 (O145:H-), BC 7848 (O157:H-), BC 7849 (O156:H47), BC 7850, BC 7851 (O157:H-), BC 7852 (O157:H-), BC 7853 (O5:H-), BC 7854 (O157:H7), BC 7855 (O157:H7), BC 7856 (O26:H-), BC 7857, BC 7858, BC 7859 (ONT:H-), BC 7860 (O129:H-), BC 7861, BC 7862 (O103:H2), BC 7863, BC 7864 (O raw form:H-), BC 7865, BC 7866 (O26:H-), BC 7867 (O raw form:H-), BC 7868, BC 7869 (ONT:H-), BC 7870 (O113:H-), BC 7871 (ONT:H-), BC 7872 (ONT:H-), BC 7873, BC 7874 (O raw form:H-), BC 7875 (O157:H-), BC 7876 (O111:H-), BC 7877 (O146:H21), BC 7878 (O145:H-), BC 7879 (O22:H8), BC 7880 (O raw form:H-), BC 7881 (O145:H-), BC 8275 (O157:H7), BC 8318 (O55:K-:H-), BC 8325 (O157:H7), and BC 8332 (ONT), BC 8333.

[0390] In some embodiments, the present disclosure teaches enteroinvasive *E. coli* (EIEC), such as strains BC 8246 (O152:K-:H-), BC 8247 (O124:K(72):H3), BC 8248 (O124), BC 8249 (O112), BC 8250 (O136:K(78):H-), BC 8251 (O124:H-), BC 8252 (O144:K-:H-), BC 8253 (O143:K:H-), BC 8254 (O143), BC 8255 (O112), BC 8256 (O28a.e), BC 8257 (O124:H-), BC 8258 (O143), BC 8259 (O167:K-:H5), BC 8260 (O128a.c.:H35), BC 8261 (O164), BC 8262 (O164:K-:H-), BC 8263 (O164), and BC 8264 (O124).

[0391] In some embodiments, the present disclosure teaches enterotoxigenic *E. coli* (ETEC), such as strains BC 5581 (O78:H11), BC 5583 (O2:K1), BC 8221 (O118), BC 8222 (O148:H-), BC 8223 (O111), BC 8224 (O110:H-), BC 8225 (O148), BC 8226 (O118), BC 8227 (O25:H42), BC 8229 (O6), BC 8231 (O153:H45), BC 8232 (O9), BC 8233 (O148), BC 8234 (O128), BC 8235 (O118), BC 8237 (O111), BC 8238 (O110:H17), BC 8240 (O148), BC 8241 (O6H16), BC 8243 (O153), BC 8244 (O15:H-), BC 8245 (O20), BC 8269 (O125a.c:H-), BC 8313 (O6:H6), BC 8315 (O153:H-), BC 8329, BC 8334 (O118:H12), and BC 8339.

[0392] In some embodiments, the present disclosure teaches enteropathogenic *E. coli* (EPEC), such as strains BC 7567 (O86), BC 7568 (O128), BC 7571 (O114), BC 7572 (O119), BC 7573 (O125), BC 7574 (O124), BC 7576 (O127a), BC 7577 (O126), BC 7578 (O142), BC 7579 (O26), BC 7580 (OK26), BC 7581 (O142), BC 7582 (O55), BC 7583 (O158), BC 7584 (O-), BC 7585 (O-), BC 7586 (O-), BC 8330, BC 8550 (O26), BC 8551 (O55), BC 8552 (O158), BC 8553 (O26), BC 8554 (O158), BC 8555 (O86), BC 8556 (O128), BC 8557 (OK26), BC 8558 (O55), BC 8560 (O158), BC 8561 (O158), BC 8562 (O114), BC 8563 (O86), BC 8564 (O128), BC 8565 (O158),

BC 8566 (O158), BC 8567 (O158), BC 8568 (O111), BC 8569 (O128), BC 8570 (O114), BC 8571 (O128), BC 8572 (O128), BC 8573 (O158), BC 8574 (O158), BC 8575 (O158), BC 8576 (O158), BC 8577 (O158), BC 8578 (O158), BC 8581 (O158), BC 8583 (O128), BC 8584 (O158), BC 8585 (O128), BC 8586 (O158), BC 8588 (O26), BC 8589 (O86), BC 8590 (O127), BC 8591 (O128), BC 8592 (O114), BC 8593 (O114), BC 8594 (O114), BC 8595 (O125), BC 8596 (O158), BC 8597 (O26), BC 8598 (O26), BC 8599 (O158), BC 8605 (O158), BC 8606 (O158), BC 8607 (O158), BC 8608 (O128), BC 8609 (O55), BC 8610 (O114), BC 8615 (O158), BC 8616 (O128), BC 8617 (O26), BC 8618 (O86), BC 8619, BC 8620, BC 8621, BC 8622, BC 8623, BC 8624 (O158), and BC 8625 (O158).

[0393] In some embodiments, the present disclosure also teaches methods for the engineering of *Shigella* organisms, including *Shigella flexneri*, *Shigella dysenteriae*, *Shigella boydii*, and *Shigella sonnei*.

Generating Genetic Diversity Pools for Utilization in the Genetic Design & HTP Microbial Engineering Platform

[0394] In some embodiments, the methods of the present disclosure are characterized as genetic design. As used herein, the term genetic design refers to the reconstruction or alteration of a host organism's genome through the identification and selection of the most optimum variants of a particular gene, portion of a gene, promoter, stop codon, 5'UTR, 3'UTR, or other DNA sequence to design and create new superior host cells.

[0395] In some embodiments, a first step in the genetic design methods of the present disclosure is to obtain an initial genetic diversity pool population with a plurality of sequence variations from which a new host genome may be reconstructed.

[0396] In some embodiments, a subsequent step in the genetic design methods taught herein is to use one or more of the aforementioned HTP molecular tool sets (*e.g.* SNP swapping, promoter swapping, terminator swapping, protein solubility tag swapping or protein degradation tag swapping) to construct HTP genetic design libraries, which then function as drivers of the genomic engineering process, by providing libraries of particular genomic alterations for testing in a host cell.

Harnessing Diversity Pools From Existing Wild-type Strains

[0397] In some embodiments, the present disclosure teaches methods for identifying the sequence diversity present among microbes of a given wild-type population. Therefore, a diversity pool can

be a given number n of wild-type microbes utilized for analysis, with said microbes' genomes representing the "diversity pool."

[0398] In some embodiments, the diversity pools can be the result of existing diversity present in the natural genetic variation among said wild-type microbes. This variation may result from strain variants of a given host cell or may be the result of the microbes being different species entirely. Genetic variations can include any differences in the genetic sequence of the strains, whether naturally occurring or not. In some embodiments, genetic variations can include SNPs swaps, PRO swaps, Start/Stop Codon swaps, SOLUBILITY TAG swaps, DEGRADATION TAG swaps or STOP swaps, among others.

Harnessing Diversity Pools From Existing Industrial Strain Variants

[0399] In other embodiments of the present disclosure, diversity pools are strain variants created during traditional strain improvement processes (*e.g.*, one or more host organism strains generated *via* random mutation and selected for improved yields over the years). Thus, in some embodiments, the diversity pool or host organisms can comprise a collection of historical production strains.

[0400] In particular aspects, a diversity pool may be an original parent microbial strain (S_1) with a "baseline" genetic sequence at a particular time point (S_1Gen_1) and then any number of subsequent offspring strains (S_2, S_3, S_4, S_5 , etc., generalizable to S_{2-n}) that were derived/developed from said S_1 strain and that have a different genome ($S_{2-n}Gen_{2-n}$), in relation to the baseline genome of S_1 .

[0401] For example, in some embodiments, the present disclosure teaches sequencing the microbial genomes in a diversity pool to identify the SNP's present in each strain. In one embodiment, the strains of the diversity pool are historical microbial production strains. Thus, a diversity pool of the present disclosure can include for example, an industrial base strain, and one or more mutated industrial strains produced *via* traditional strain improvement programs.

[0402] Once all SNPs in the diversity pool are identified, the present disclosure teaches methods of SNP swapping and screening methods to delineate (*i.e.* quantify and characterize) the effects (*e.g.* creation of a phenotype of interest) of SNPs individually and in groups. Thus, as aforementioned, an initial step in the taught platform can be to obtain an initial genetic diversity pool population with a plurality of sequence variations, *e.g.* SNPs. Then, a subsequent step in the taught platform can be to use one or more of the aforementioned HTP molecular tool sets (*e.g.* SNP swapping) to construct HTP genetic design libraries, which then function as drivers of the

genomic engineering process, by providing libraries of particular genomic alterations for testing in a microbe.

[0403] In some embodiments, the SNP swapping methods of the present disclosure comprise the step of introducing one or more SNPs identified in a mutated strain (*e.g.*, a strain from amongst $S_{2-n}Gen_{2-n}$) to a base strain (S_1Gen_1) or wild-type strain.

[0404] In other embodiments, the SNP swapping methods of the present disclosure comprise the step of removing one or more SNPs identified in a mutated strain (*e.g.*, a strain from amongst $S_{2-n}Gen_{2-n}$).

Creating Diversity Pools via Mutagenesis

[0405] In some embodiments, the mutations of interest in a given diversity pool population of cells can be artificially generated by any means for mutating strains, including mutagenic chemicals, or radiation. The term “mutagenizing” is used herein to refer to a method for inducing one or more genetic modifications in cellular nucleic acid material.

[0406] The term “genetic modification” refers to any alteration of DNA. Representative gene modifications include nucleotide insertions, deletions, substitutions, and combinations thereof, and can be as small as a single base or as large as tens of thousands of bases. Thus, the term “genetic modification” encompasses inversions of a nucleotide sequence and other chromosomal rearrangements, whereby the position or orientation of DNA comprising a region of a chromosome is altered. A chromosomal rearrangement can comprise an intrachromosomal rearrangement or an interchromosomal rearrangement.

[0407] In one embodiment, the mutagenizing methods employed in the presently claimed subject matter are substantially random such that a genetic modification can occur at any available nucleotide position within the nucleic acid material to be mutagenized. Stated another way, in one embodiment, the mutagenizing does not show a preference or increased frequency of occurrence at particular nucleotide sequences.

[0408] The methods of the disclosure can employ any mutagenic agent including, but not limited to: ultraviolet light, X-ray radiation, gamma radiation, N-ethyl-N-nitrosourea (ENU), methylnitrosourea (MNU), procarbazine (PRC), triethylene melamine (TEM), acrylamide monomer (AA), chlorambucil (CHL), melphalan (MLP), cyclophosphamide (CPP), diethyl sulfate (DES), ethyl methane sulfonate (EMS), methyl methane sulfonate (MMS), 6-mercaptopurine (6-MP), mitomycin-C (MMC), N-methyl-N'-nitro-N-nitrosoguanidine (MNNG), 3H_2O , and urethane

(UR) (See *e.g.*, Rinchik, 1991; Marker *et al.*, 1997; and Russell, 1990). Additional mutagenic agents are well known to persons having skill in the art, including those described in http://www.iephb.nw.ru/~spirov/hazard/mutagen_lst.html.

[0409] The term “mutagenizing” also encompasses a method for altering (*e.g.*, by targeted mutation) or modulating a cell function, to thereby enhance a rate, quality, or extent of mutagenesis. For example, a cell can be altered or modulated to thereby be dysfunctional or deficient in DNA repair, mutagen metabolism, mutagen sensitivity, genomic stability, or combinations thereof. Thus, disruption of gene functions that normally maintain genomic stability can be used to enhance mutagenesis. Representative targets of disruption include, but are not limited to DNA ligase I (Bentley *et al.*, 2002) and casein kinase I (U.S. Pat. No. 6,060,296).

[0410] In some embodiments, site-specific mutagenesis (*e.g.*, primer-directed mutagenesis using a commercially available kit such as the Transformer Site Directed mutagenesis kit (Clontech)) is used to make a plurality of changes throughout a nucleic acid sequence in order to generate nucleic acid encoding a cleavage enzyme of the present disclosure.

[0411] The frequency of genetic modification upon exposure to one or more mutagenic agents can be modulated by varying dose and/or repetition of treatment, and can be tailored for a particular application.

[0412] Thus, in some embodiments, “mutagenesis” as used herein comprises all techniques known in the art for inducing mutations, including error-prone PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, and iterative sequence recombination by any of the techniques described herein.

Single Locus Mutations to Generate Diversity

[0413] In some embodiments, the present disclosure teaches mutating cell populations by introducing, deleting, or replacing selected portions of genomic DNA. Thus, in some embodiments, the present disclosure teaches methods for targeting mutations to a specific locus. In other embodiments, the present disclosure teaches the use of gene editing technologies such as ZFNs, TALENS, Lambda Red or CRISPR, to selectively edit target DNA regions.

[0414] In other embodiments, the present disclosure teaches mutating selected DNA regions outside of the host organism, and then inserting the mutated sequence back into the host organism. For example, in some embodiments, the present disclosure teaches mutating native or synthetic promoters to produce a range of promoter variants with various expression properties (*see*

promoter ladder *infra*). In other embodiments, the present disclosure is compatible with single gene optimization techniques, such as ProSAR (Fox *et al.* 2007. “Improving catalytic function by ProSAR-driven enzyme evolution.” Nature Biotechnology Vol 25 (3) 338-343, incorporated by reference herein).

[0415] In some embodiments, the selected regions of DNA are produced *in vitro* via gene shuffling of natural variants, or shuffling with synthetic oligos, plasmid-plasmid recombination, virus plasmid recombination, virus-virus recombination. In other embodiments, the genomic regions are produced via error-prone PCR.

[0416] In some embodiments, generating mutations in selected genetic regions is accomplished by “reassembly PCR.” Briefly, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of a nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length. Each of the segments is amplified with a set of such primers. The PCR products are then “reassembled” according to assembly protocols. In brief, in an assembly protocol, the PCR products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTP's) and appropriate buffer salts in the absence of additional primers (“self-priming”). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes.

[0417] In some embodiments of the disclosure, mutated DNA regions, such as those discussed above, are enriched for mutant sequences so that the multiple mutant spectrum, *i.e.* possible combinations of mutations, is more efficiently sampled. In some embodiments, mutated sequences are identified via a mutS protein affinity matrix (Wagner *et al.*, Nucleic Acids Res. 23(19):3944-3948 (1995); Su *et al.*, Proc. Natl. Acad. Sci. (U.S.A.), 83:5057-5061(1986)) with a preferred step of amplifying the affinity-purified material *in vitro* prior to an assembly reaction. This amplified material is then put into an assembly or reassembly PCR reaction as described in later portions of this application.

Promoter Ladders

[0418] Promoters regulate the rate at which genes are transcribed and can influence transcription in a variety of ways. Constitutive promoters, for example, direct the transcription of their

associated genes at a constant rate regardless of the internal or external cellular conditions, while regulatable promoters increase or decrease the rate at which a gene is transcribed depending on the internal and/or the external cellular conditions, *e.g.* growth rate, temperature, responses to specific environmental chemicals, and the like. Promoters can be isolated from their normal cellular contexts and engineered to regulate the expression of virtually any gene, enabling the effective modification of cellular growth, product yield and/or other phenotypes of interest.

[0419] In some embodiments, the present disclosure teaches methods for producing promoter ladder libraries for use in downstream genetic design methods. For example, in some embodiments, the present disclosure teaches methods of identifying one or more promoters and/or generating variants of one or more promoters within a host cell, which exhibit a range of expression strengths, or superior regulatory properties. A particular combination of these identified and/or generated promoters can be grouped together as a promoter ladder, which is explained in more detail below.

[0420] In some embodiments, the present disclosure teaches the use of promoter ladders. In some embodiments, the promoter ladders of the present disclosure comprise promoters exhibiting a continuous range of expression profiles. For example, in some embodiments, promoter ladders are created by: identifying natural, native, or wild-type promoters that exhibit a range of expression strengths in response to a stimuli, or through constitutive expression (*see e.g.*, Figure 20 and Figures 28-30). These identified promoters can be grouped together as a promoter ladder.

[0421] In other embodiments, the present disclosure teaches the creation of promoter ladders exhibiting a range of expression profiles across different conditions. For example, in some embodiments, the present disclosure teaches creating a ladder of promoters with expression peaks spread throughout the different stages of a fermentation (*see e.g.*, Figure 28). In other embodiments, the present disclosure teaches creating a ladder of promoters with different expression peak dynamics in response to a specific stimulus (*see e.g.*, Figure 29). Persons skilled in the art will recognize that the regulatory promoter ladders of the present disclosure can be representative of any one or more regulatory profiles.

[0422] In some embodiments, the promoter ladders of the present disclosure are designed to perturb gene expression in a predictable manner across a continuous range of responses. In some embodiments, the continuous nature of a promoter ladder confers strain improvement programs with additional predictive power. For example, in some embodiments, swapping promoters or

termination sequences of a selected metabolic pathway can produce a host cell performance curve, which identifies the most optimum expression ratio or profile; producing a strain in which the targeted gene is no longer a limiting factor for a particular reaction or genetic cascade, while also avoiding unnecessary over expression or misexpression under inappropriate circumstances. In some embodiments, promoter ladders are created by: identifying natural, native, or wild-type promoters exhibiting the desired profiles. Examples of native promoters for use in the methods provided herein can be found in Table 1.4. In other embodiments, the promoter ladders are created by mutating naturally occurring promoters to derive multiple mutated promoter sequences. Each of these mutated promoters is tested for effect on target gene expression. In some embodiments, the edited promoters are tested for expression activity across a variety of conditions, such that each promoter variant's activity is documented/characterized/annotated and stored in a database. The resulting edited promoter variants are subsequently organized into promoter ladders arranged based on the strength of their expression (*e.g.*, with highly expressing variants near the top, and attenuated expression near the bottom, therefore leading to the term "ladder"). Examples of synthetic promoters for use in the methods provided herein can be found in Table 1.4.

[0423] In some embodiments, the present disclosure teaches promoter ladders that are a combination of identified naturally occurring promoters and mutated variant promoters.

[0424] In some embodiments, the present disclosure teaches methods of identifying natural, native, or wild-type promoters that satisfied both of the following criteria: 1) represented a ladder of constitutive promoters; and 2) could be encoded by short DNA sequences, ideally less than 100 base pairs. In some embodiments, constitutive promoters of the present disclosure exhibit constant gene expression across two selected growth conditions (typically compared among conditions experienced during industrial cultivation). An example of examining gene expression using different promoters provided herein can be found in Example 12. In some embodiments, the promoters of the present disclosure will consist of a ~60 base pair core promoter, and a 5' UTR between 26- and 40 base pairs in length.

[0425] Native promoters for inclusion in promoter ladders for use in the PROSWP methods provided herein can be selected based on said native promoter showing minimal variation in an associated gene's expression. Further, the native promoters can be 60-90 bps in length, and can consist of sequence that is located 50 bp in front of a putative transcription start site, and, optionally, the sequence up to but not including a putative start codon. Examples of native

promoters for use in the methods provided herein can be found in Table 1.4. In particular, the native promoters for use in the methods provided herein can be selected from nucleic acid SEQ ID NOs 71-131 from Table 1.4.

[0426] In some embodiments, one or more of the aforementioned identified naturally occurring promoter sequences are chosen for gene editing. In some embodiments, the natural promoters are edited *via* any of the mutation methods described *supra*. In other embodiments, the promoters of the present disclosure are edited by synthesizing new promoter variants with the desired sequence.

[0427] Synthetic promoters for inclusion in promoter ladders for use in the PROSWP methods provided herein can be chimeric sequences 60-90 bps in length. The synthetic promoter libraries for use herein can comprise a set or plurality of synthetic promoters that can be designed and constructed such that they are likely to express constitutively and/or represent a range of expression strengths in comparison to one another. Further, the synthetic promoters can be designed and constructed such that they are unlikely to bind any regulatory elements present in *E. coli* and therefore drive gene expression constitutively.

[0428] To achieve these design goals, the chimeric synthetic promoters can comprise all or a combination of the elements found in Table 1.5. In particular, relative to a transcription start site, the synthetic promoters can comprise or consist of a distal region, a -35 region, a core region, a -10 region and a 5'UTR/ribosomal binding site (RBS) region, as shown in **Figure 54**. The distal region can be located just upstream of the -35 region, while the core region can be located between the -35 and -10 regions as shown **Figure 54**. Both the distal and the core regions can be important for binding regulatory elements (*see* Cox et al., Mol Syst Biol. 2007; 3: 145). Since the lambda phage p_R promoter is expected to drive expression constitutively, distal and core regions from this promoter can be used in the design strategy. The core region of the lambda phage p_L promoter also be included for the same reason, as well as to add additional variety to the library.

[0429] The -35 and -10 regions can be included because they are known to be particularly important in prokaryotes for binding the RNA polymerase and can therefore be critical for modulating the degree of expression. In one embodiment, the -35 and -10 regions from the lambda phage p_R promoter and p_L promoter are used. The -35 and -10 regions from p_R and p_L can be used since they are expected to drive strong expression. Additionally, -35 and -10 regions found in many native *E. coli* promoters can be used, whereby said -35 and -10 regions represent small variations to p_R and p_L and can be expected to decrease the promoter strength in comparison with p_R and p_L .

The variable 6 bp sequence constituting the -35 and -10 regions can be selected from the -35 and -10 sequences found in Table 1.5.

[0430] In addition to the above elements, the chimeric synthetic promoters can comprise a 5' untranslated region (5'-UTR) that includes a ribosome binding site (RBS), which can be particularly important in prokaryotes for binding the ribosome and thus be critical to modulating the degree of protein expression. In one embodiment, the 5'-UTR/RBS from the native *E. coli* gene *acs* can be used to add additional variety to the library. In another embodiment, the 5'UTR/RBS from the lambda phage p_R promoter can be used.

[0431] Examples of synthetic promoters for use in the methods provided herein can be found in Table 1.4. In particular, the synthetic promoters for use in the methods provided herein can be selected from nucleic acid SEQ ID NOs 132-207 from Table 1.4.

[0432] The entire disclosure of U.S. Patent Application No. 62/264,232, filed on December 07, 2015, is hereby incorporated by reference in its entirety for all purposes

[0433] A non-exhaustive list of the promoters of the present disclosure is provided in the below Table 1 and/or Table 1.4. Each of the promoter sequences can be referred to as a heterologous promoter or heterologous promoter polynucleotide.

Table 1. Selected promoter sequences of the present disclosure.

SEQ ID No.	Promoter Short Name	Promoter Name
1	P1	Pcg0007_lib_39
2	P2	Pcg0007
3	P3	Pcg1860
4	P4	Pcg0755
5	P5	Pcg0007_265
6	P6	Pcg3381
7	P7	Pcg0007_119

8	P8	Pcg3121
---	----	---------

Table 1.4 Additional promoter sequences of the present disclosure.

Promoter name	SEQ ID NO.	Type*
b0904_promoter	71	Native
b2405_promoter	72	Native
b0096_promoter	73	Native
b0576_promoter	74	Native
b2017_promoter	75	Native
b1278_promoter	76	Native
b4255_promoter	77	Native
b0786_promoter	78	Native
b0605_promoter	79	Native
b1824_promoter	80	Native
b1061_promoter	81	Native
b0313_promoter	82	Native
b0814_promoter	83	Native
b4133_promoter	84	Native
b4268_promoter	85	Native
b0345_promoter	86	Native
b2096_promoter	87	Native
b1277_promoter	88	Native
b1646_promoter	89	Native
b4177_promoter	90	Native
b0369_promoter	91	Native
b1920_promoter	92	Native
b3742_promoter	93	Native
b3929_promoter	94	Native
b3743_promoter	95	Native
b1613_promoter	96	Native

b1749_promoter	97	Native
b2478_promoter	98	Native
b0031_promoter	99	Native
b2414_promoter	100	Native
b1183_promoter	101	Native
b0159_promoter	102	Native
b2837_promoter	103	Native
b3237_promoter	104	Native
b3778_promoter	105	Native
b2349_promoter	106	Native
b1434_promoter	107	Native
b3617_promoter	108	Native
b0237_promoter	109	Native
b4063_promoter	110	Native
b0564_promoter	111	Native
b0019_promoter	112	Native
b2375_promoter	113	Native
b1187_promoter	114	Native
b2388_promoter	115	Native
b1051_promoter	116	Native
b4241_promoter	117	Native
b4054_promoter	118	Native
b2425_promoter	119	Native
b0995_promoter	120	Native
b1399_promoter	121	Native
b3298_promoter	122	Native
b2114_promoter	123	Native
b2779_promoter	124	Native
b1114_promoter	125	Native
b3730_promoter	126	Native
b3025_promoter	127	Native
b0850_promoter	128	Native

b2365_promoter	129	Native
b4117_promoter	130	Native
b2213_promoter	131	Native
pMB029_promoter	132	Synthetic
pMB023_promoter	133	Synthetic
pMB025_promoter	134	Synthetic
pMB019_promoter	135	Synthetic
pMB008_promoter	136	Synthetic
pMB020_promoter	137	Synthetic
pMB022_promoter	138	Synthetic
pMB089_promoter	139	Synthetic
pMB001_promoter	140	Synthetic
pMB051_promoter	141	Synthetic
pMB070_promoter	142	Synthetic
pMB074_promoter	143	Synthetic
pMB046_promoter	144	Synthetic
pMB071_promoter	145	Synthetic
pMB013_promoter	146	Synthetic
pMB080_promoter	147	Synthetic
pMB038_promoter	148	Synthetic
pMB060_promoter	149	Synthetic
pMB064_promoter	150	Synthetic
pMB058_promoter	151	Synthetic
pMB085_promoter	152	Synthetic
pMB081_promoter	153	Synthetic
pMB091_promoter	154	Synthetic
pMB027_promoter	155	Synthetic
pMB048_promoter	156	Synthetic
pMB055_promoter	157	Synthetic
pMB006_promoter	158	Synthetic
pMB012_promoter	159	Synthetic
pMB014_promoter	160	Synthetic

pMB028_promoter	161	Synthetic
pMB059_promoter	162	Synthetic
pMB061_promoter	163	Synthetic
pMB043_promoter	164	Synthetic
pMB066_promoter	165	Synthetic
pMB079_promoter	166	Synthetic
pMB032_promoter	167	Synthetic
pMB068_promoter	168	Synthetic
pMB082_promoter	169	Synthetic
pMB030_promoter	170	Synthetic
pMB067_promoter	171	Synthetic
pMB050_promoter	172	Synthetic
pMB069_promoter	173	Synthetic
pMB017_promoter	174	Synthetic
pMB039_promoter	175	Synthetic
pMB011_promoter	176	Synthetic
pMB072_promoter	177	Synthetic
pMB016_promoter	178	Synthetic
pMB077_promoter	179	Synthetic
pMB047_promoter	180	Synthetic
pMB052_promoter	181	Synthetic
pMB090_promoter	182	Synthetic
pMB035_promoter	183	Synthetic
pMB073_promoter	184	Synthetic
pMB004_promoter	185	Synthetic
pMB054_promoter	186	Synthetic
pMB024_promoter	187	Synthetic
pMB007_promoter	188	Synthetic
pMB005_promoter	189	Synthetic
pMB003_promoter	190	Synthetic
pMB088_promoter	191	Synthetic
pMB065_promoter	192	Synthetic

pMB037_promoter	193	Synthetic
pMB009_promoter	194	Synthetic
pMB041_promoter	195	Synthetic
pMB036_promoter	196	Synthetic
pMB049_promoter	197	Synthetic
pMB044_promoter	198	Synthetic
pMB042_promoter	199	Synthetic
pMB086_promoter	200	Synthetic
pMB053_promoter	201	Synthetic
pMB057_promoter	202	Synthetic
pMB018_promoter	203	Synthetic
pMB002_promoter	204	Synthetic
pMB015_promoter	205	Synthetic
pMB087_promoter	206	Synthetic
pMB063_promoter	207	Synthetic

*Native promoters from Escherichia coli

Table 1.5. Sequence Parts used in combinatorial synthetic promoter-5'UTR library

Part name	Part sequences	Origin
distal	ACCGTGCGTG (SEQ ID NO. 208)	phage λ , P_R promoter
-35	TTTACA	variation
	TTGACT	phage λ , P_L promoter
	TTGACA	phage λ , P_R promoter
	TAGGCT	variation
	TAGACT	variation
Core	ATTTTACCTCTGGCGGT (SEQ ID NO. 209)	phage λ , P_R promoter
	TAAATACCACTGGCGGT (SEQ ID NO. 210)	phage λ , P_L promoter
-10	GATAAT	phage λ , P_L promoter
	GATACT	phage λ , P_R promoter
	TAGAGT	variation
	TATAAT	variation
	TATTTT	variation

5'- UTR/RBS	GGTTGCATGTACTAAGGAGGTTGT (SEQ ID NO. 211)	phage λ , P_R promoter
	TAACATCCTACAAGGAGAACAAAAGC (SEQ ID NO. 212)	<i>E. coli</i> , promoter of <i>acs</i> gene

[0434] In some embodiments, the promoters of the present disclosure exhibit at least 100%, 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 79%, 78%, 77%, 76%, or 75% sequence identity with a promoter from Table 1 and/or Table 1.4.

Bicistronic Regulatory Element Design

[0435] One of the barriers to efficient and scalable HTP genetic design is the lack of standard parts that can be reused reliably in novel combinations. Many examples within *E. coli*, highlight how seemingly simple genetic functions behave differently in different settings. For example, in some embodiments, a prokaryotic ribosome-binding site (RBS) element that initiates translation for one coding sequence might not function at all with another coding sequence (*see* Salis, H.M., et al. “Automated design of synthetic ribosome binding sites to control protein expression” Nat. Biotechnol. Vol 27, 946–950 (2009)). If the genetic elements that encode control of central cellular processes such as transcription and translation cannot be reliably reused, then there is little chance that higher-order objects encoded from such basic elements will be reliable in larger-scale systems. In some embodiments, the methods of the present disclosure overcome these aforementioned challenges via the use of bicistronic design regulatory sequences.

[0436] Bicistronic designs of the present disclosure can, in some embodiments, greatly reduce the context dependent variability in expression strength of a given promoter for a variety of coding genes (Mutalik, et al. “Precise and reliable gene expression via standard transcription and translation initiation elements” Nat. Biotechnol. Vol 10 (4) pp 354-368 (2013)). In some embodiments, the present disclosure teaches that a bicistronic design (BCD) is a nucleotide sequence in which a promoter drives the expression of two coding sequences, where the first coding sequence (Cistron 1) terminates and the second coding sequences initiates at the same nucleotide base (Cistron 2/target gene). This strategy provides a means to avoid variability in expression strength of the second coding sequences due to unpredictable interactions between the promoter and the second coding sequence.

[0437] In some embodiment, the promoters of the present disclosure are compound regulatory sequences following the bicistronic design. That is, in some embodiments, the promoters in the promoter ladders of the present disclosure are larger regulatory sequences comprising **i)** a promoter operably linked to **ii)** a first ribosome binding site (SD1), which is operably linked to **iii)** a first cistronic sequence (Cis1), wherein the Cis1 overlaps with **iv)** a second ribosome binding site (SD2), that is then operably linked to **v)** a target gene coding sequence (Cis2) (*See* Figure 43). In some embodiments, the present disclosure refers to the combination of elements i)-iv) as a “bicistronic design” or “bicistronic design regulatory sequence” (BCDs).

[0438] In some embodiments, the BCDs of the present disclosure can be operably linked to any target gene. Thus in some embodiments, the BCDs of the present disclosure can be used in place of traditional promoters. In some embodiments, the present disclosure teaches that using BCDs in the PRO swap toolbox increases the consistency with which expressed transcripts are translated. Without wishing to be bound by any one theory, the present inventors believe that the presence of an SD1 and Cis1 leader sequences operably linked to the target gene recruit active ribosomal complexes, which are then able to consistently reinitiate translation of the target gene via the SD2 ribosomal binding site.

[0439] A collection of promoters and bicistronic design elements have been reported that can be used for HTP genome engineering (*see* Mutalik, et al. “Precise and reliable gene expression via standard transcription and translation initiation elements” *Nat. Biotechnol.* Vol 10 (4) pp 354-368 (2013)). However, these reported sequences all contain identical DNA sequences in the first 35 nt of the 48nt of the regulatory bicistronic design sequence (*see* Mutalik state-of-the-art sequence in Figure 43).

[0440] In some embodiments, the present disclosure teaches that the BCD of Mutalik et al., could not be used to effectively engineer multiple target genes in a single organism. That is, in some embodiments, the present disclosure teaches against the multiple integration of the Mutalik BCD into the genome of a host cell. Without wishing to be bound by any one theory, the present inventors believe that repeated use of the Mutalik et al., BCD would result in increasing rates of undesired homologous recombination (HR) triggered by the presence of highly homologous sequences throughout the genome.

[0441] In some embodiments, the present disclosure solves this problem by describing novel BCDs with non-identical nucleotide sequences. These novel BCDs can be used for HTP genome

engineering in *E. coli* to provide predictable changes in expression of multiple genes within a single genome, independent of the coding sequences of these genes, without inducing undesirable homologous recombination.

[0442] In some embodiments, the present disclosure teaches methods of expressing two target gene proteins in a host organism at relatively similar levels. Thus in some embodiments, the present disclosure teaches expression of two or more target gene proteins within 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, or 3-fold of each other.

[0443] In some embodiments, the present disclosure teaches methods of expressing two target gene proteins in a host organism at similar levels, while reducing the risk of undesirable homologous recombination (HR) events triggered by the use of identical regulatory sequences. Thus in some embodiments, the present disclosure teaches methods of varying BCD sequence in a way that maintains expression levels, while reducing the risk of HR. That is, in some embodiments, the present disclosure teaches expression of two or more genes encoding for proteins via BCDs that are identical save for their Cis1 sequences.

BCD Promoters

[0444] In some embodiments, the BCDs of the present disclosure comprise a promoter sequence. In some embodiments the promoters comprised in BCDs can be any promoter capable of expressing in the host cell. Thus, in some embodiments, the promoters can be any promoter disclosed in the specification. In some embodiments, the promoter can be any promoter known to function in *E. coli*. In other embodiments, the promoters can be any promoter disclosed in Table 1 and/or Table 1.4.

First and Second Ribosomal Binding Site (SD1 and SD2)

[0445] In some embodiments, the BCDs of the present disclosure comprise a first and second ribosomal binding site, referred to as SD1 and SD2, respectively. In some embodiments, the sequences of SD1 and SD2 can be the same. In other embodiments the sequences of SD1 and SD2 can be different.

[0446] In some embodiments, the SD sequences can be any known ribosomal binding site functional in the host under the HTP genomic engineering. In other embodiments, the present disclosure teaches an SD sequence of NNNGGANN, wherein N refers to any nucleotide. In other embodiments, the present disclosure teaches SD sequences selected from the sequences disclosed in Table 1.1.

Table 1.1- Non-limiting list of Ribosomal Binding Sites, amenable for SD1 and SD2 use.

SD #	Sequence	SD #	Sequence	SD #	Sequence
1	TGCGGAGGG	50	AGTGGAACC	99	GGAGGAGAC
2	CACGGAGGC	51	GCCGGAGTG	100	TAGGGAACG
3	TGGGGAGGG	52	GCGGGACAG	101	TGAGGACGT
4	AGGGGAGGC	53	CCGGGATAA	102	ATCGGAGGT
5	GGGGGAGGG	54	CCAGGAACG	103	CCGGGAGGG
6	TATGGAGGT	55	GACGGAGCA	104	ACGGGACGG
7	GACGGAGCG	56	AGCGGATTG	105	GTGGGAGAG
8	AGGGGAGTC	57	ATGGGATAT	106	CGAGGATAG
9	ACAGGAGGC	58	CTGGGAAGT	107	TGGGGAGCC
10	GGGGGAGCG	59	TGGGGAGCC	108	CGTGGAGTG
11	ACAGGAGTC	60	GTGGGACAT	109	TGGGGACTG
12	ACAGGAGAC	61	GCCGGAGCG	110	GGTGGGAAGC
13	TGGGGAGGC	62	AATGGAGGC	111	GATGGACGG
14	TGGGGAGAT	63	GGAGGATTG	112	AATGGAGGT
15	GGGGGAGAA	64	GATGGACTC	113	GGAGGATCC
16	AGGGGAGGC	65	GAGGGATGG	114	ACAGGATCT
17	AGAGGAGTC	66	GAAGGACAG	115	CCCGGACAG

SD #	Sequence	SD #	Sequence	SD #	Sequence
18	AGGGGATAT	67	TAGGGAAGG	116	ACGGGAAAC
19	CGTGGAGTG	68	TGGGGAACC	117	AGTGGACCG
20	GGGGGAAGG	69	TTAGGAGTC	118	CATGGATCA
21	AGGGGAATC	70	GGAGGAGGA	119	CTGGGATGT
22	CCCGGAGGT	71	CCGGGATCT	120	TTAGGATGG
23	AGGGGAGGG	72	GGGGGATGA	121	GTAGGATTC
24	TTTGGAGTC	73	CTGGGAGTG	122	TTTGGAGTT
25	AGGGGACAC	74	CAAGGAACC	123	CGCGGACTC
26	CGGGGAGAC	75	GCTGGAGGC	124	CTCGGACAG
27	GGGGGAGGG	76	ATGGGACCT	125	CAAGGAGTC
28	AAGGAGATC	77	GGAGGAGGG	126	CTTGGACAG
29	TAAGGAGGT	78	CTGGGATGC	127	GGCGGATCG
30	GGGGGAGTC	79	ACAGGATAC	128	GGGGGACAG
31	GGAGGATCG	80	GAGGGAAGG	129	ATTGGATGG
32	CTGGGATCG	81	AGTGGATCT	130	CCTGGATAT
33	ATCGGACCG	82	CCGGGAGTT	131	CTCGGATAC
34	GGGGGAGTG	83	GGTGGAGGC	132	GGGGGAGCC

SD #	Sequence	SD #	Sequence	SD #	Sequence
35	TAGGGAGCA	84	CCAGGAAGA	133	TGGGGATTG
36	GGTGGAGGG	85	GGGGGATTT	134	CCAGGATCA
37	GCCGGAGGT	86	CGCGGAGTA	135	CGGGGACGG
38	TGAGGAAAG	87	CCAGGAATC	136	TCAGGACAA
39	CTTGGAGGG	88	ATTGGAGTG	137	GTAGGATGG
40	TGAGGAGAT	89	CTAGGAAGT	138	TCCGGACTA
41	CGTGGAGGG	90	GAAGGATAG	139	GTCGGATCA
42	GCGGGAGGG	91	AAAGGACAC	140	TGCGGAGTT
43	GGGGGATAG	92	CCGGGACGT	141	GGTGGACGG
44	GGGGGAGCG	93	ATGGGAGTG	142	CTTGGACGA
45	CCGGGAGCA	94	AGCGGACAG	143	CTGGGACTT
46	GCGGGAGTA	95	GGCGGATCT	144	TATGGAGTA
47	GTAGGACCG	96	ATAGGAGGG	145	ACAGGAGGC
48	CCGGGAAGG	97	TATGGAGGG		
49	GACGGAGTC	98	GGTGGACTC		

[0447] It should be clear that the epistatic mapping procedure described herein does not depend on which model is used by the analysis equipment 214. Given such a predictive model, it is

possible to score and rank all hypothetical strains accessible to the mutation library *via* combinatorial consolidation.

[0448] In some embodiments, the present disclosure teaches that varying individual SD sequences in the BCD will affect the overall expression of the BCD. Some SD sequences may serve to increase or decrease overall expression potential of the BCD. It is expected however, that each BCD will exhibit consistent expression results when combined over different target gene Cis2 sequences.

[0449] In some embodiments, the SD2 sequence is entirely embedded within the coding sequence of the first cistronic sequence. That is, in some embodiments, the SD2 sequence is integrated into the coding sequences of Cis2. Without wishing to be bound by any one theory, the present inventors believe that BCD arrangements in which the ribosomal binding site of the target gene (SD2) is entirely embedded in the coding sequence of the upstream gene (Cis1) results in the coupling of the translations of the Cis1 and Cis2 peptides. More specifically, the present inventors hypothesize that the intrinsic helicase activity of ribosomes arriving at the stop codon of an upstream Cis1 sequence eliminate inhibitory RNA structures that would otherwise disrupt translation initiation of the downstream Cis2 target gene.

First Cistronic Sequence (Cis1)

[0450] In some embodiments, the first cistronic sequence Cis1 of the present disclosure can be any sequence encoding a continuous peptide. For example, in some embodiments, the Cis1 sequence encodes for a peptide that is 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, or more amino acids long, including any ranges and subranges therein. In some embodiments, Cis1 does not need to encode a functional peptide.

[0451] In some embodiments, the Cis1 sequence encodes a 16-amino-acid leader peptide. In some embodiments the Cis1 nucleotide sequence is:

5'-ATGAAAGCAATTTTCGTACTGAAACATCTTAATCATGCACAGGAGACTTTCTAA-
3' (SEQ ID No. 17).

[0452] In other embodiments, the present disclosure teaches that the Cis1 sequence can be 5'-ATGNN-
- 3', where N can be any nucleic acid, so long as Cis1 encodes for a peptide

[0453] In some embodiments, the present disclosure teaches that the stop codon of Cis1 and the start codon of Cis2 must be in close proximity or overlapping. For example, in some embodiments, the stop codon of Cis1 must be within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, or 50 nucleotides of the start codon of Cis2, including all ranges and subranges therein.

[0454] In some embodiments, the Cis1 sequence overlaps with the Cis2 sequence by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100 or more nucleotides, including all ranges and subranges therein. In other embodiments, the BCD of the present disclosure is designed such that the Cis1 sequence overlaps by 1 nucleotide with the Cis2 target gene coding sequence, such that the last few nucleotides encode for both a stop and start codon via a -1 frame shift (*see* Figure 43). In some embodiments, the Cis1 and Cis 2 sequences must be coded on different open reading frames so as to prevent the formation of a chimeric protein combining the sequences of Cis1 and Cis2.

[0455] In some embodiments, the present disclosure teaches that the start codon of the Cis1 sequence can be any functional start codon. In some embodiments, the present disclosure teaches that prokaryotes use ATG (AUG) at the most common start codon, followed by GTG (GUG) and TTG (UUG).

[0456] In some embodiments, the present disclosure teaches that the Cis1 sequence does not have any premature stop codons. In other embodiments, the present disclosure teaches that rare codons in the Cis1 sequence can reduce the translation efficiency of Cis2. Thus, in some embodiments, Cis1 will encode for a peptide without any rare codons to achieve maximum expression. In other embodiments, Cis1 will encode for a peptide with one or more rare codons in order to modulate the expression of Cis2.

[0457] In other embodiments, the present disclosure teaches that multiple codons repeats in the Cis1 sequence can reduce the translation efficiency of Cis2. Thus, in some embodiments, Cis1 will encode for a peptide without any codon repeats to achieve maximum expression. In other embodiments, Cis1 will encode for a peptide with one or more codon repeats in order to modulate the expression of Cis2.

Second Cistronic Sequence (Cis2- Target Gene)

[0458] In some embodiments, the present disclosure teaches that the BCDs of the present disclosure are operably linked to a Cis2 target gene sequence, in much the same way in which the promoters of the PRO-swap libraries are operably linked to target sequences. That is, in some embodiments, the BCDs of the present disclosure can take the place of traditional promoters in the PRO-swap libraries and methods of the present disclosure. The Cis2 sequences, can in some embodiments, be any sequence of interest.

[0459] The present disclosure teaches that, in some embodiments, target genes encoding for a polypeptide will be more effectively regulated by BCDs than by a promoter. That is, in some embodiments, BCDs will not modulate the expression of non-coding RNA more than would be possible by a promoter.

Terminator Ladders

[0460] In some embodiments, the present disclosure teaches methods of improving genetically engineered host strains by providing one or more transcriptional termination sequences at a position 3' to the end of the RNA encoding element. In some embodiments, the present disclosure teaches that the addition of termination sequences improves the efficiency of RNA transcription of a selected gene in the genetically engineered host. In other embodiments, the present disclosure teaches that the addition of termination sequences reduces the efficiency of RNA transcription of a selected gene in the genetically engineered host. Thus in some embodiments, the terminator ladders of the present disclosure comprises a series of terminator sequences exhibiting a range of transcription efficiencies (e.g., one weak terminator, one average terminator, and one strong promoter).

[0461] A transcriptional termination sequence may be any nucleotide sequence, which when placed transcriptionally downstream of a nucleotide sequence encoding an open reading frame, causes the end of transcription of the open reading frame. Such sequences are known in the art and may be of prokaryotic, eukaryotic or phage origin. Examples of terminator sequences include, but are not limited to, PTH-terminator, pET-T7 terminator, T3-T ϕ terminator, pBR322-P4 terminator, vesicular stomatitis virus terminator, rrnB-T1 terminator, rrnC terminator, TTadc transcriptional terminator, and yeast-recognized termination sequences, such as Mata α (α -factor) transcription terminator, native α -factor transcription termination sequence, ADR1 transcription termination sequence, ADH2 transcription termination sequence, and GAPD transcription

termination sequence. A non-exhaustive listing of transcriptional terminator sequences may be found in the iGEM registry, which is available at: <http://partsregistry.org/Terminators/Catalog>.

[0462] In some embodiments, transcriptional termination sequences may be polymerase-specific or nonspecific, however, transcriptional terminators selected for use in the present embodiments should form a 'functional combination' with the selected promoter, meaning that the terminator sequence should be capable of terminating transcription by the type of RNA polymerase initiating at the promoter. For example, in some embodiments, the present disclosure teaches a eukaryotic RNA pol II promoter and eukaryotic RNA pol II terminators, a T7 promoter and T7 terminators, a T3 promoter and T3 terminators, a yeast-recognized promoter and yeast-recognized termination sequences, etc., would generally form a functional combination. The identity of the transcriptional termination sequences used may also be selected based on the efficiency with which transcription is terminated from a given promoter. For example, a heterologous transcriptional terminator sequence may be provided transcriptionally downstream of the RNA encoding element to achieve a termination efficiency of at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% from a given promoter.

[0463] In some embodiments, efficiency of RNA transcription from the engineered expression construct can be improved by providing nucleic acid sequence forms a secondary structure comprising two or more hairpins at a position 3' to the end of the RNA encoding element. Not wishing to be bound by a particular theory, the secondary structure destabilizes the transcription elongation complex and leads to the polymerase becoming dissociated from the DNA template, thereby minimizing unproductive transcription of non-functional sequence and increasing transcription of the desired RNA. Accordingly, a termination sequence may be provided that forms a secondary structure comprising two or more adjacent hairpins. Generally, a hairpin can be formed by a palindromic nucleotide sequence that can fold back on itself to form a paired stem region whose arms are connected by a single stranded loop. In some embodiments, the termination sequence comprises 2, 3, 4, 5, 6, 7, 8, 9, 10 or more adjacent hairpins. In some embodiments, the adjacent hairpins are separated by 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, or 15 unpaired nucleotides. In some embodiments, a hairpin stem comprises 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more base pairs in length. In certain embodiments, a hairpin stem is 12 to 30 base pairs in length. In certain embodiments, the

termination sequence comprises two or more medium-sized hairpins having stem region comprising about 9 to 25 base pairs. In some embodiments, the hairpin comprises a loop-forming region of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides. In some embodiments, the loop-forming region comprises 4-8 nucleotides. Not wishing to be bound by a particular theory, stability of the secondary structure can be correlated with termination efficiency. Hairpin stability is determined by its length, the number of mismatches or bulges it contains and the base composition of the paired region. Pairings between guanine and cytosine have three hydrogen bonds and are more stable compared to adenine-thymine pairings, which have only two. The G/C content of a hairpin-forming palindromic nucleotide sequence can be at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90% or more. In some embodiments, the G/C content of a hairpin-forming palindromic nucleotide sequence is at least 80%. In some embodiments, the termination sequence is derived from one or more transcriptional terminator sequences of prokaryotic, eukaryotic or phage origin. In some embodiments, a nucleotide sequence encoding a series of 4, 5, 6, 7, 8, 9, 10 or more adenines (A) are provided 3' to the termination sequence.

[0464] In some embodiments, the present disclosure teaches the use of a series of tandem termination sequences. In some embodiments, the first transcriptional terminator sequence of a series of 2, 3, 4, 5, 6, 7, or more may be placed directly 3' to the final nucleotide of the dsRNA encoding element or at a distance of at least 1-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50, 50-100, 100-150, 150-200, 200-300, 300-400, 400-500, 500-1,000 or more nucleotides 3' to the final nucleotide of the dsRNA encoding element. The number of nucleotides between tandem transcriptional terminator sequences may be varied, for example, transcriptional terminator sequences may be separated by 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50 or more nucleotides. In some embodiments, the transcriptional terminator sequences may be selected based on their predicted secondary structure as determined by a structure prediction algorithm. Structural prediction programs are well known in the art and include, for example, CLC Main Workbench.

[0465] Persons having skill in the art will recognize that the methods of the present disclosure are compatible with any termination sequence. In some embodiments, the present disclosure teaches use of annotated *Corynebacterium glutamicum* terminators as disclosed in from Pfeifer-Sancar et al. 2013. "Comprehensive analysis of the *Corynebacterium glutamicum* transcriptome using an improved RNAseq technique" Pfeifer-Sancar et al. BMC Genomics 2013, 14:888). In other

embodiments, the present disclosure teaches use of transcriptional terminator sequences found in the iGEM registry, which is available at: <http://partsregistry.org/Terminators/Catalog>. A non-exhaustive listing of transcriptional terminator sequences of the present disclosure is provided in Table 1.2 below.

Table 1.2. Non-exhaustive list of termination sequences of the present disclosure.

<i>E. coli</i>			
Name	Description	Direction	Length
BBa_B0010	T1 from <i>E. coli</i> rrnB	Forward	80
BBa_B0012	TE from coliphage T7	Forward	41
BBa_B0013	TE from coliphage T7 (+/-)	Forward	47
BBa_B0015	double terminator (B0010-B0012)	Forward	129
BBa_B0017	double terminator (B0010-B0010)	Forward	168
BBa_B0053	Terminator (His)	Forward	72
BBa_B0055	-- No description --		78
BBa_B1002	Terminator (artificial, small, %T _~ =85%)	Forward	34
BBa_B1003	Terminator (artificial, small, %T _~ =80)	Forward	34
BBa_B1004	Terminator (artificial, small, %T _~ =55)	Forward	34
BBa_B1005	Terminator (artificial, small, %T _~ =25%)	Forward	34
BBa_B1006	Terminator (artificial, large, %T _~ >90)	Forward	39

BBa_B1010	Terninator (artificial, large, %T~<10)	Forward	40
BBa_I11013	Modification of biobricks part BBa_B0015		129
BBa_I51003	-- No description --		110
BBa_J61048	[rnpB-T1] Terminator	Forward	113
BBa_K1392970	Terminator+Tetr Promoter+T4 Endolysin		623
BBa_K1486001	Arabinose promoter + CpxR	Forward	1924
BBa_K1486005	Arabinose promoter + sfGFP-CpxR [Cterm]	Forward	2668
BBa_K1486009	CxpR & Split IFP1.4 [Nterm + Nterm]	Forward	3726
BBa_K780000	Terminator for Bacillus subtilis		54
BBa_K864501	T22, P22 late terminator	Forward	42
BBa_K864600	T0 (21 imm) transcriptional terminator	Forward	52
BBa_K864601	Lambda t1 transcriptional terminator	Forward	
BBa_B0011	LuxICDABEG (+/-)	Bidirectional	46
BBa_B0014	double terminator (B0012-B0011)	Bidirectional	95
BBa_B0021	LuxICDABEG (+/-), reversed	Bidirectional	46

BBa_B0024	double terminator (B0012-B0011), reversed	Bidirectional	95
BBa_B0050	Terminator (pBR322, +/-)	Bidirectional	33
BBa_B0051	Terminator (yciA/tonA, +/-)	Bidirectional	35
BBa_B1001	Terminator (artificial, small, %T _~ =90)	Bidirectional	34
BBa_B1007	Terminator (artificial, large, %T _~ =80)	Bidirectional	40
BBa_B1008	Terminator (artificial, large, %T _~ =70)	Bidirectional	40
BBa_B1009	Terminator (artificial, large, %T _~ =40%)	Bidirectional	40
BBa_K187025	terminator in pAB, BioBytes plasmid		60
BBa_K259006	GFP-Terminator	Bidirectional	823
BBa_B0020	Terminator (Reverse B0010)	Reverse	82
BBa_B0022	TE from coliphageT7, reversed	Reverse	41
BBa_B0023	TE from coliphage T7, reversed	Reverse	47
BBa_B0025	double terminator (B0015), reversed	Reverse	129
BBa_B0052	Terminator (rrnC)	Forward	41
BBa_B0060	Terminator (Reverse B0050)	Bidirectional	33
BBa_B0061	Terminator (Reverse B0051)	Bidirectional	35
BBa_B0063	Terminator (Reverse B0053)	Reverse	72

Spy	Terminator (SEQ ID NO. 225)			90	
pheA	Terminator (SEQ ID NO. 226)			51	
osmE	Terminator (SEQ ID NO. 227)			42	
rpoH	Terminator (SEQ ID NO. 228)			41	
vibE	Terminator (SEQ ID NO. 229)			71	
Thrl_ABC	Terminator (SEQ ID NO. 230)			57	
<i>Corynebacterium</i>					
Terminator	Terminator Start	Terminator End	strand	Transcript End	DNA Sequence
cg0001 T1	1628	1647	+	loop	SEQ ID NO. 9
cg0007 T2	7504	7529	+	stem 1	SEQ ID NO. 10
cg0371 T3	322229	322252	+	stem 1	SEQ ID NO. 11
cg0480 T4	421697	421720	-	stem 1	SEQ ID NO. 12
cg0494 T5	436587	436608	+	loop	SEQ ID NO. 13
cg0564 T6	499895	499917	+	stem 1	SEQ ID NO. 14

cg0610 T7	541016	541039	+	stem 2	SEQ ID NO. 15
cg0695 T8	613847	613868	-	loop	SEQ ID NO. 16

Protein Solubility Tag Ladders

[0466] In some embodiments, the present disclosure teaches methods of improving genetically engineered host strains by providing one or more protein solubility tag sequences operably linked with a target protein derived from a target gene. The solubility tags can be fusion partners operably linked with the target protein on either the N-terminus or the C-terminus of the target protein. In some embodiments, the present disclosure teaches that the addition of solubility tag sequences improves the solubility of a protein translated from a selected gene in the genetically engineered host. In other embodiments, the solubility tags may also be used to assist in the purification of the target protein.

[0467] Effective tags for use in protein solubility tag ladders of the present disclosure can be any solubility tags known in the art that form independent, well-folded domains, highly soluble domains. These domains may contribute to the solubility of their target protein through an additive effect, or when used as an N-terminal tag may quickly fold after emerging from the ribosome and sterically block the emerging amino acid chain of the target protein from interacting with other cellular components which may cause misfolding. Further, solubility tags for inclusion in solubility tag ladders may have properties in common such as being small, tightly-folded domains or being leader sequences of proteins known to be highly soluble. The protein solubility tag sequences can be any such tags known in the art such as, for example, the any of the tags found in Costa et al., Front Microbiol. 2014; 5: 63, the contents of which are hereby incorporated by reference in its entirety. In one embodiment, solubility tag sequences include the tags found in Table 17.

[0468] In one embodiment, the protein solubility tag is a fusion partner. The fusion partner coding gene can be present in any of the vectors (e.g., shuttle vectors) provided herein such that integration of the gene for a target protein in the vector operably links the fusion partner coding gene to the

target gene. *E. coli* expression vectors comprising solubility tags for use herein can comprise a protease recognition sequence between the solubility tag fusion partner coding gene and the target protein coding gene that can allow for the tag removal as necessary. The choice of a fusion partner for use in the solubility swap methods provided herein can depend on:

[0469] (i) Purpose of the fusion: is it for solubility improvement or for affinity purification? A variety of fusion tags that render different purposes are available, and systems containing both solubility and affinity tags like, for instance, the dual hexahistidine (His6)-MBP tag, can be designed in order to get a rapid “in one step” protein production. Some protein tags can also function in both affinity and solubility roles, as for instance, the MBP or glutathione-S-transferase (GST; Esposito and Chatterjee, *Curr Opin Biotechnol.* 2006 Aug;17(4):353-8. Epub 2006 Jun 15).

[0470] (ii) Amino acid composition and size: target proteins may require larger or smaller tags depending on their application. Larger tags can present a major diversity in the amino acid content, and may impose a metabolic burden in the host cell different from that imposed by small tags.

[0471] (iii) Required production levels: structural studies may require higher protein production levels that can be rapidly achieved with a larger fusion tag, which has strong translational initiation signals, whereas the study of physiological interactions may demand for lower production levels and small tags.

[0472] (iv) Tag location: fusion partners can promote different effects when located at the N-terminus or C-terminus of the target protein. N-terminal tags may often be advantageous over C-terminal tags because: (1) they provide a reliable context for efficient translation initiation, in which fusion proteins take advantage of efficient translation initiation sites on the tag; (2) they can be removed leaving none or few additional residues at the native N-terminal sequence of the target protein, since most of endoproteases cleave at or near the C-terminus of their recognition sites.

Table 17. Non-exhaustive list of protein solubility tag sequences of the present disclosure.

Solubility Tag Name (Protein Solubility Tag #)	Description	Organism	Nucleic Acid SEQ ID NO.	Amino Acid SEQ ID NO.	Size (aa)
GB1 (PST1)	IgG domain B1 of Protein G	Streptococcus sp.	231	235	56

FH8 (PST2)	Fasciola hepatica 8-kDa antigen	F. hepatica	232	236	69
Ubiquitin (PST3)			233	237	
SUMO (PST4)	Small ubiquitin modified	Homo sapiens	234	238	~100

Protein Degradation Tag Ladders

[0473] In some embodiments, the present disclosure teaches methods of improving genetically engineered host strains by providing one or more protein degradation tag sequences operably linked with a target protein derived from a target gene. The addition of a degradation tag sequence using the methods provided herein can mark the target protein for degradation. Marking the target protein for degradation may reduce or modulate the target protein abundance within a cell. By reducing or modulating the target protein levels or abundance in the cell, the addition of degradation tag sequences to a target protein may ultimately impact the overall phenotype of resultant strains.

[0474] Effective tags for use in protein degradation tag ladders of the present disclosure can be any degradation tags known in the art that are part of a known degradation pathway in the host organism (e.g., *E. coli*). For example, known degradation pathways in *E. coli* can include the clpXP/clpAP system, the HflB system, the ftsH System and the lon system. Accordingly, degradation tags for use in the degradation tag swap methods provided herein can include any tags known to function in any of these *E. coli* protein degradation systems. In some cases, the degradation tags can be mutated in such a manner as to confer the ability of the resultant mutant tag to have its activity tuned. For example, the ssrA class of tags can be mutated such that the mutated ssrA degradation tags mark a tagged protein for degradation via the ClpXP degradation pathway with different degrees of efficiency. In one example, the ssrA tags can contain a single amino acid mutations in the last three residues of the AANDENYALAA consensus sequence such that target proteins comprising a C-terminal mutated ssrA tag can be degraded at varying levels of efficiency by certain intracellular tail specific proteases (e.g., Tsp protease) depending on which amino acid has been mutated in the last three residues of the ssrA tag consensus sequence (see Keiler K C, Sauer R T. Sequence determinants of C-terminal substrate recognition by the Tsp

protease. J Biol Chem. 1996;271:2589–2593, the contents of which are hereby incorporated by reference in their entirety). Accordingly, using the degradation tag swap methods of the present disclosure, it is possible to obtain strains of host cells possessing target proteins of varying stability by constructing variants carrying C-terminal peptide tags with minor alterations in the Tsp consensus sequence. Examples of mutant *ssrA* tags for use in the methods herein can comprise amino acid SEQ ID NO: 248, 249 or 250.

[0475] Another example of mutated *ssrA* tags for use in the methods provided herein can be the DAS tags found in McGinness et al., “Engineering Controllable Protein Degradation” Mol. Cell, Vol 22(5), June 2006, the contents of which are hereby incorporated by reference in its entirety. In the DAS tags, two residues in the *ssrA* tag were replaced resulting in mutated *ssrA* tags that exhibited weakened ClpX binding without diminishing SspB recognition. As such, target proteins bearing the DAS tags can be degraded efficiently by ClpXP only when SspB is present, allowing intracellular degradation to be regulated by controlling SspB levels.

[0476] A non-exhaustive list of protein degradation tag sequence of the present disclosure can be found in Table 18.

Table 18. Non-exhaustive list of protein degradation tag sequences of the present disclosure.

Degradation Tag Name (Protein Degradation Tag #)	Description	Organism	Nucleic Acid SEQ ID NO.	Amino Acid SEQ ID NO.	Source
<i>ssrA</i> _LAA (PDT1)	native	<i>E. coli</i>	239	247	Andersen et al., Appl Environ Microbiol. 1998 Jun; 64(6): 2240–2246.
<i>ssrA</i> _LVA (PDT2)	mutant	<i>E. coli</i>	240	248	Andersen et al., Appl Environ Microbiol. 1998 Jun; 64(6): 2240–2246.
<i>ssrA</i> _AAV (PDT3)	mutant	<i>E. coli</i>	241	249	Andersen et al., Appl Environ Microbiol. 1998 Jun; 64(6): 2240–2246.

ssrA_ASV (PDT4)	mutant	<i>E. coli</i>	242	250	Andersen et al., Appl Environ Microbiol. 1998 Jun; 64(6): 2240–2246.
ftsH-cII89-97 (PDT5)	native	<i>E. coli</i>	243	251	Kobiler O , Koby S , Teff D , Court D , Oppenheim AB. PNAS. 23 Oct 2002, 99(23):14964-14969.
cII08 (PDT6)	native	<i>E. coli</i>	244	252	Herman et al., Genes Dev. 1998 May 1; 12(9): 1348–1355.
sul20 (PDT7)	native	<i>E. coli</i>	245	253	Wohlever et al., Protein Eng Des Sel. 2013 Apr; 26(4): 299–305.
β20 (PDT8)	native	<i>E. coli</i>	246	254	Wohlever et al., Protein Eng Des Sel. 2013 Apr; 26(4): 299–305.

[0477] The degradation tags can be fusion partners operably linked with the target protein on either the N-terminus or the C-terminus of the target protein. Accordingly, the fusion partner coding gene can be present in any of the vectors (e.g., shuttle vectors) provided herein such that integration of the gene for a target protein in the vector operably links the fusion partner coding gene to the target gene such that translation of the construct generates a fusion protein with the degradation tag present on either the N-terminus or C-terminus of the target protein as desired. In one embodiment, placement of the degradation tags (or mutants thereof) at the N-terminus or C-terminus of a target protein can depend on the tag used. For example, degradation tags (or mutants thereof) associated with the clpXP/clpAP system, the HflB system, the ftsH System or the sul20 tag of the lon system can be operably linked to a target protein at the C-terminus, while β20 degradation tags (or mutants thereof) of the the lon system can be operably linked to a target protein at the N-terminus or internally. In one embodiment, the degradation tag is the N-degron tag (Ntag) for *E. coli* as found in Sekar K, Gentile AM, Bostick JW, Tyo KEJ (2016) N-Terminal-Based Targeted, Inducible Protein Degradation in Escherichia coli. PLoS ONE 11(2): e0149746, which is hereby incorporated by reference in its entirety. The Ntag can be placed on the N-terminus of a target protein of interest and can serve to mark the target protein of interest for degradation in the *E. coli*

host cell via the clpXP/clpAP system. In another embodiment, the degradation tag is the RepA tag which can be located at the N-terminus of a target protein as described in Butz et al., *Biochemistry*, 2011, 50 (40), pp 8594–8602, which is hereby incorporated by reference in its entirety. The N-terminal RepA tag can serve to mark the target protein of interest for degradation in the *E. coli* host cell via the clpXP/clpAP system.

Hypothesis-driven Diversity Pools and Hill Climbing

[0478] The present disclosure teaches that the HTP genomic engineering methods of the present disclosure do not require prior genetic knowledge in order to achieve significant gains in host cell performance. Indeed, the present disclosure teaches methods of generating diversity pools (e.g., Figure 1) via several functionally agnostic approaches, including random mutagenesis, and identification of genetic diversity among pre-existing host cell variants (e.g., such as the comparison between a wild type host cell and an industrial variant).

[0479] In some embodiments however, the present disclosure also teaches hypothesis-driven methods of designing genetic diversity mutations that will be used for downstream HTP engineering. That is, in some embodiments, the present disclosure teaches the directed design of selected mutations. In some embodiments, the directed mutations are incorporated into the engineering libraries of the present disclosure (e.g., SNP swap, PRO swap, STOP swap, SOLUBILITY TAG swap, or DEGRADATION TAG swap).

[0480] In some embodiments, the present disclosure teaches the creation of directed mutations based on gene annotation, hypothesized (or confirmed) gene function, or location within a genome. The diversity pools of the present disclosure may include mutations in genes hypothesized to be involved in a specific metabolic or genetic pathway associated in the literature with increased performance of a host cell. In other embodiments, the diversity pool of the present disclosure may also include mutations to genes present in an operon associated with improved host performance. In yet other embodiments, the diversity pool of the present disclosure may also include mutations to genes based on algorithmic predicted function, or other gene annotation.

[0481] In some embodiments, the present disclosure teaches a “shell” based approach for prioritizing the targets of hypothesis-driven mutations. The shell metaphor for target prioritization is based on the hypothesis that only a handful of primary genes are responsible for most of a particular aspect of a host cell’s performance (e.g., production of a single biomolecule). These primary genes are located at the core of the shell, followed by secondary effect genes in the second

layer, tertiary effects in the third shell, and... etc. For example, in one embodiment the core of the shell might comprise genes encoding critical biosynthetic enzymes within a selected metabolic pathway (*e.g.*, production of citric acid). Genes located on the second shell might comprise genes encoding for other enzymes within the biosynthetic pathway responsible for product diversion or feedback signaling. Third tier genes under this illustrative metaphor would likely comprise regulatory genes responsible for modulating expression of the biosynthetic pathway, or for regulating general carbon flux within the host cell.

[0482] The present disclosure also teaches “hill climb” methods for optimizing performance gains from every identified mutation. In some embodiments, the present disclosure teaches that random, natural, or hypothesis-driven mutations in HTP diversity libraries can result in the identification of genes associated with host cell performance. For example, the present methods may identify one or more beneficial SNPs located on, or near, a gene coding sequence. This gene might be associated with host cell performance, and its identification can be analogized to the discovery of a performance “hill” in the combinatorial genetic mutation space of an organism.

[0483] In some embodiments, the present disclosure teaches methods of exploring the combinatorial space around the identified hill embodied in the SNP mutation. That is, in some embodiments, the present disclosure teaches the perturbation of the identified gene and associated regulatory sequences in order to optimize performance gains obtained from that gene node (*i.e.*, hill climbing). Thus, according to the methods of the present disclosure, a gene might first be identified in a diversity library sourced from random mutagenesis, but might be later improved for use in the strain improvement program through the directed mutation of another sequence within the same gene.

[0484] The concept of hill climbing can also be expanded beyond the exploration of the combinatorial space surrounding a single gene sequence. In some embodiments, a mutation in a specific gene might reveal the importance of a particular metabolic or genetic pathway to host cell performance. For example, in some embodiments, the discovery that a mutation in a single RNA degradation gene resulted in significant host performance gains could be used as a basis for mutating related RNA degradation genes as a means for extracting additional performance gains from the host organism. Persons having skill in the art will recognize variants of the above described shell and hill climb approaches to directed genetic design.

Biosynthetic Pathway Scaffolding

[0485] In some embodiments, the present disclosure teaches that the productivity of some bioindustrial processes is limited by the random diffusion of substrates, intermediates, and biosynthetic enzymes within a host cell. In some embodiments, the present disclosure teaches that productivity of host cell cultures can be increased by co-localizing biosynthetic enzymes in a pathway. Thus, in some embodiments, the present disclosure teaches tethering biosynthetic enzymes to a scaffold, such as a DNA or protein scaffold.

[0486] In some embodiments, co-localization is achieved by recombinant fusions of DNA binding domains to the biosynthetic enzymes in the pathway, which then bind to a DNA scaffold region, thus constraining the pathway enzymes close to one another in the cell. In other embodiments, co-localization is achieved by recombinant fusions of protein binding domains to the biosynthetic enzymes in the pathway, which then bind to a protein scaffold region, thus constraining the pathway enzymes close to one another in the cell. In some embodiments, co-localization increases the rate of production and decreases the concentration of pathway intermediates in the cell (*see* Figure 44).

[0487] In some embodiments, the present disclosure teaches a high-throughput method for engineering the genome of *Escherichia coli*, wherein nucleotide sequences encoding DNA binding or protein binding domains are inserted into genes encoding enzymes in a biosynthetic pathway, and a DNA scaffold plasmid or a scaffold protein is introduced into the cell. In accordance with one aspect of the invention, it is believed that the DNA or protein binding domains tethered to the biosynthetic genes will localize the recombinant pathway enzymes together onto the scaffold plasmid or peptide, thus leading to improved productivity of the target product.

[0488] In some embodiments, this invention solves the problem of diffusion-limited productivity of small molecules in *E. coli* cells with genomes engineered via high-throughput methods. Currently, the only reported examples of DNA scaffolding to localize biosynthetic enzymes have been low throughput processes in which the recombinant pathway enzymes are encoded on plasmids (Lee, et al. "Improved Production of L-Threonine in *Escherichia coli* by Use of a DNA Scaffold System" *App. And Environ. Microbiol.* Vol 79(3), pp. 774-782 (2013)). In some embodiments, the current invention provides a means to incorporate DNA binding domains into pathway enzymes that are chromosomally-encoded, in a high-throughput manner.

[0489] In some embodiments, the present disclosure teaches chimeric biosynthetic enzymes and scaffold DNAs and proteins. The various aspects of this technology are discussed in more detail below.

DNA Binding Chimeric Proteins

[0490] In some embodiments, the present disclosure teaches chimeric proteins comprising selected biosynthetic enzymes that are tethered to DNA binding domain. In accordance with these embodiments, it is expected that the chimeric biosynthetic enzymes will be recruited to a DNA scaffold by their DNA binding domains, thereby concentrating the various biosynthetic activities to an area of the host cell.

[0491] In some embodiments, the biosynthetic enzymes and the DNA binding domains are covalently linked. In some embodiments, the biosynthetic enzymes are translationally fused to the DNA binding domains. Thus, in some embodiments the chimeric biosynthetic enzymes are formed by coupling the DNA binding domain to the amino terminus, the carboxy terminus, or to an internal site within the biosynthetic pathway protein. Persons having skill in the art will recognize the need to ensure that the addition of the DNA binding domain does not substantially reduce the activity of the biosynthetic enzyme.

[0492] In some embodiments of the present disclosure, the biosynthetic enzyme is coupled to its DNA-binding domain via a short polypeptide linker sequence. Suitable linkers include peptides of between about 6 and about 40 amino acids in length. Preferred linker sequences include glycine-rich (e.g. G3-5), serine-rich (e.g. GSG, GSGS (SEQ ID NO. 18), GSGSG (SEQ ID NO. 19), GSNG (SEQ ID NO. 20), or alanine rich (e.g., TSAAA (SEQ ID NO. 21)) linker sequences. Other exemplary linker sequences have a combination of glycine, alanine, proline and methionine residues such as AAAGGM (SEQ ID NO. 22); AAAGGMPPAAAGGM (SEQ ID NO. 23); AAAGGM (SEQ ID NO. 24); and PPAAAGGMM (SEQ ID NO. 25). Linkers may have virtually any sequence that results in a generally flexible chimeric biological pathway protein.

[0493] In some embodiments, the methods of the present disclosure are compatible with any DNA binding domain capable of function in cis with the biosynthetic enzyme. In some embodiments, the DNA binding domains are preferably exogenous to the host organism. In other embodiments, the present disclosure teaches selection of DNA binding domains which are sufficiently selective to avoid excessive binding outside of the designed scaffold DNA.

[0494] Various DNA-binding domains are known in the art along with their corresponding nucleotide recognition sites in DNA (i.e., DNA binding sites) and are suitable for use in the system and methods of the present invention. For example, in one embodiment of the present invention, the DNA binding portion of a chimeric biological pathway protein comprises a leucine zipper DNA binding domain wherein the scaffold comprises the corresponding leucine zipper DNA binding sequence. In another embodiment of the present invention, the DNA binding portion of a chimeric biological pathway protein comprises a helix-loop-helix DNA binding domain wherein the scaffold comprises the corresponding helix-loop-helix DNA binding sequence. In another embodiment, the DNA binding portion of a chimeric biological pathway protein comprises a winged helix DNA binding domain wherein the scaffold comprises the corresponding winged helix DNA binding sequence. In another embodiment, the DNA binding portion of a chimeric biological pathway protein comprises a winged helix-turn-helix DNA binding domain wherein the scaffold comprises the corresponding winged helix-turn-helix DNA binding sequence. In another embodiment, the DNA binding portion of a chimeric biological pathway protein comprises a helix-turn-helix DNA binding wherein the scaffold comprises the corresponding helix-turn-helix DNA binding sequence. In another embodiment, the DNA binding portion of the chimeric biological pathway protein comprises a HMG-box DNA binding domain wherein the scaffold comprises the corresponding HMG-box DNA binding sequence. In another embodiment, the DNA binding portion of the chimeric biological pathway protein comprises a custom designed TALE DNA binding domain wherein the scaffold comprises the corresponding designed TALE DNA binding sequence. In another embodiment of the present invention, the DNA binding portion of a chimeric biological pathway protein comprises a zinc finger DNA binding domain wherein the scaffold comprises the corresponding zinc finger DNA binding sequence.

[0495] Exemplary zinc finger DNA binding domain sequences and corresponding DNA binding sites are provided in Table 1.3. Other zinc finger DNA binding domains and their corresponding target DNA binding sequences known in the art are also suitable for use in the present invention (see e.g., Greisman H A and Pabo C O, "A General Strategy for Selecting High-Affinity Zinc Finger Proteins for Diverse DNA Target Sites," *Science* 275:657-661 (1997), Rebar E J and Pabo C O, "Zinc Finger Phage: Affinity Selection of Fingers with New DNA-Binding Specificities," *Science* 263:671-673 (1994); Maeder et al., "Rapid "Open-Source" Engineering of

Customized Zinc-Finger Nucleases for Highly Efficient Gene Modification,” *Mol. Cell.* 31:294-301 (2008), Sander et al., “Selection-Free Zinc-Finger-Nuclease Engineering by Context-Dependent Assembly (CoDA),” *Nat. Methods* 8:67-69 (2011), U.S. Pat. No. 5,5789,538 to Rebar, U.S. Pat. No. 6,410,248 to Greisman, U.S. Pat. No. 7,605,140 to Rebar, U.S. Pat. No. 6,140,081 to Barbas, U.S. Pat. No. 7,067,617 to Barbas, U.S. Pat. No. 6,205,404 to Michaels, and U.S. Patent Application Publication No. 20070178454 to Joung, each of which are hereby incorporated by reference in their entirety).

[0496] Methods for optimizing the DNA binding specificities of zinc finger domains and methods of engineering synthetic DNA binding sites are also known in the art and can be utilized in the present invention to generate new zinc finger binding partners (see e.g., Bulyk et al., “Exploring the DNA-binding Specificities of Zinc Fingers with DNA Microarrays,” *Proc. Nat'l Acad. Sci. U.S.A* 98(13): 7158-63 (2001) and “Hurt et al., “Highly Specific Zinc Finger Proteins Obtained by Directed Domain Shuffling and Cell-based Selection,” *Proc. Nat'l Acad. Sci. U.S.A.* 100(21): 12271-6 (2003), U.S. Pat. No. 5,5789,538 to Rebar, U.S. Pat. No. 6,410,248 to Greisman, U.S. Pat. No. 7,605,140 to Rebar, U.S. Pat. No. 6,140,081 to Barbas, U.S. Pat. No. 7,067,617 to Barbas, U.S. Pat. No. 6,205,404 to Michaels, and U.S. Patent Application Publication No. 20070178454 to Joung, each of which are hereby incorporated by reference in its entirety).

Table 1.3 - non-limiting list of DNA binding domains.

Zinc Finger	DNA Binding Domain Sequence	DNA Binding Sequence (5'→3')
Zif268	PGEKPYACPVESCDRRFSRSDDELTRHIRIHTGQKPFQC RICMRNFSRSDHLTTHIRTHITGEKPFACDICGRKFARS DERKRHTKIHT (SEQ ID NO. 26)	GCGTGGGCG GCG GGG GCG
PBSII	PGEKPYACPECGKSFSQRANLRAHQRTHTGEKPYKC PECGKSFSRSDHLTTHQRTHTGEKPYKCPECGKSFSR SDVLVRHQRTHT (SEQ ID NO. 27)	GTGTGGAAA

Zinc Finger	DNA Binding Domain Sequence	DNA Binding Sequence (5'→3')
ZFa	PGERPFQCRICMRNFSDSPTLRRHTRTHTGEKPFQCRI CMRNFSVRHNLTRHLRTHHTGEKPFQCRICMRNFSDRT SLARHLKTH (SEQ ID NO. 28)	GTCGATGCC
ZFb	PGERPFQCRICMRNFSSKDLHRHTRTHTGEKPFQCR ICMRNFSLSQTLKRHLRTHHTGEKPFQCRICMRNFSL DMLARHLKTH (SEQ ID NO. 29)	GCGGCTGGG
ZFc	PGERPFQCRICMRNFSSPSKLIRHTRTHTGEKPFQCRI MRNFSDGSNLARHLRTHHTGEKPFQCRICMRNFSD NLPRHLKTH (SEQ ID NO. 30)	GAGGACGGC
Tyr123	EKPYKCPECGKSFSDRSNLTRHQRTHTGEKPYKCPEC GKSFSSTSNLARHQRTHTGEKPFKCPECGKSFSRSDA LTRHQRTHT (SEQ ID NO. 31)	GTGGATGAC
Tyr456	EKPYKCPECGKSFSSQSSNLARHQRTHTGEKPYKCPEC GKSFSRSDHLTKHQRTHTGEKPFKCPECGKSFSSN LARHQRTHT (SEQ ID NO. 32)	GAAGGGGAA
Blues	ASDDRPYACPVESCDRRFSRRDVLNMNHIRIHTGQKPF QCRICMRNFSDHLTTHIRTHHTGEKPFACDICGRKFA NRDTLTRHSKIHLRQNDLE (SEQ ID NO. 33)	GTTTGGATG
Jazz	ASDDRPYACPVESCDRRFSRDELTRHIRIHTGQKPFQ CRICMRNFSSRDVLRHNRTHHTGEKPFACDICGRKFA SRDVLRRHNRIHLRQNDLE (SEQ ID NO. 34)	GCTGCTGCG

Zinc Finger	DNA Binding Domain Sequence	DNA Binding Sequence (5'→3')
Bagly	EFMTGDRPYACPVESCDRRFSRSDELTRHIRIHTGQKP FQCRICMRNFSSRDVLRRIHNRTHHTGEKPFACDICGRK FASRDVLRRIHNRHILRQGRSHVCAECGKAFVSSKLLK RHQLVHTGEKPFQLE (SEQ ID NO. 35)	CGGGCTGCTGC (SEQ ID NO. 36)
Gli1	KREPESVYETDCRWDGCSQEFDSQEQLVHHINSEHIIH GERKEFVCHWGGCSRELRPFKAQYMLVVHMRRHTG EKPHKCTFEGCRKSYSRLENLKTHLRSHTGEKPYMCE HEGCSKAFSNASDRAKHQNRTHSNEKPYVCKLPGCT KRYTDPSSLRKHVKTVHGPDAHVTKRHRGD (SEQ ID NO. 37)	GACCACCCAAG ACGA (SEQ ID NO. 38)
HIVC	PFQCRICMRNFSLRTDLDRHTRTHHTGEKPFQCRICMR NFSLSQTLRRHLRTHHTGEKPFQCRICMRNFSLSNLG RHLKTHHTGEK (SEQ ID NO. 39)	GATGCTGCA
B3	AQAALPKEKPYACPECGKSFSDPGNLVRHQRTHTG EKPYKCPECGKSFSRSDKLVRHQRTHTGEKPYKCPEC GKSFSQSSHLVRHQRTHTGKKTSGQAG (SEQ ID NO. 40)	GACGGGGG
N1	AQAALPKEKPYACPECGKSFSQSSSLVRHQRTHTGE KPYKCPECGKSFSQSSNLVRHQRTHTGEKPYKCPEC KFSRSDKLVRHQRTHTGKKTSGQAG (SEQ ID NO. 41)	GTAGAAGGG

Zinc Finger	DNA Binding Domain Sequence	DNA Binding Sequence (5'→3')
Sp-1	PGKKKQHICHIQGCQKQVYGKTSHLRAHLRWHTGERP FMCTWSYCGKRFRSDELQRHKRTHITGEKKFACPEC PKRFMRSDHLSKHIKTHQNKKG (SEQ ID NO. 42) PGKKKQHACPECQKSFSSHLRAHQRTHTGERPYK CPECQKSFSSRDELQRHQRTHTGEKPYKCPECQKSFSS RSDHLSKHQRTHTQNKKG (SEQ ID NO. 43)	GGGGCGGGG

Nucleic Acid Scaffold Sequence

[0497] In some embodiments, the present disclosure teaches a DNA scaffold comprising one or more of the DNA binding sequences corresponding to the DNA binding domains contained within the chimeric biosynthetic enzymes. In some embodiments the DNA scaffold is an extrachromosomal plasmid or other vector. In other embodiments, the DNA scaffold is encoded within the genome of a host cell.

[0498] Suitable nucleic acid vectors include, without limitation, plasmids, baculovirus vectors, bacteriophage vectors, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral vectors (for example, viral vectors based on vaccinia virus, poliovirus, adenovirus, adeno-associated virus, SV40, herpes simplex virus, and the like), artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and other vectors. In some embodiments of the present invention, vectors suitable for use in prokaryotic host cells are preferred. Accordingly, exemplary vectors for use in prokaryotes such as *Escherichia coli* include, but are not limited to, pACYC184, pBeloBac11, pBR332, pBAD33, pBBR1MCS and its derivatives, pSC101, SuperCos (cosmid), pWE15 (cosmid), pTrc99A, pBAD24, vectors containing a ColE1 origin of replication and its derivatives, pUC, pBluescript, pGEM, Ori_Plsmid27 (SEQ ID NO. 213), vector backbone 1 (SEQ ID NO. 214), vector backbone 2 (SEQ ID NO. 215), vector backbone 3 (SEQ ID NO. 216), vector backbone 4 (SEQ ID NO. 217), and pTZ vectors.

[0499] In some embodiments, the present disclosure teaches that a nucleic acid scaffold subunit may comprise 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, or more different DNA binding sites coupled

together to facilitate the binding and immobilization of 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, or more different biosynthetic pathway proteins. In some embodiments, the present disclosure teaches that the DNA scaffolds have a single DNA binding site for each corresponding chimeric biosynthetic protein.

[0500] In other embodiments, the nucleic acid scaffold may comprise two or more copies of the same DNA binding site. This architecture allows for optimizing the biological protein stoichiometry to be achieved. In accordance with this embodiment of the present invention, the same DNA binding sites may be coupled together to create enzymatic centers for a particular chemical reaction. Thus in some embodiments, the DNA scaffold comprises groups of multiple DNA binding sites, each group corresponding to a specific chimeric biosynthetic gene/enzyme.

[0501] In some embodiments of the present invention, the method of assembling a synthetic biological pathway involves immobilizing at least a first chimeric biosynthetic gene (e.g., enzyme) and a second chimeric biosynthetic enzyme onto nucleic acid scaffold. The first chimeric biosynthetic enzyme produces a first product that is a substrate for the second chimeric biological pathway protein. The second chimeric biosynthetic enzyme is immobilized onto the scaffold construct such that it is positioned adjacent to or very close to the first chimeric biosynthetic enzyme. In this way, the effective concentration of the first product is high, and the second chimeric biosynthetic enzyme can act efficiently on the first product. As an example, a synthetic nucleic acid scaffold has immobilized thereon, in order from 3'→5' or 5'→3' of the scaffold construct a) the first chimeric biosynthetic enzyme, and b) the second chimeric biosynthetic enzyme to form a "scaffold subunit." The scaffold subunit can be repeated two or more times within the synthetic nucleic acid scaffold.

[0502] In accordance with this and all aspects of the present invention, two or more copies (e.g., two, three, four, five, six, seven, eight, nine, ten, or more molecules) of each chimeric biosynthetic enzyme can be immobilized onto a scaffold subunit. For example, in some embodiments, a scaffold subunit has immobilized thereon, a) one molecule (copy) of the first chimeric biosynthetic enzyme and b) one molecule of the second chimeric biosynthetic enzyme. In other embodiments, a scaffold subunit has immobilized thereon, a) one molecule of the first chimeric biosynthetic enzyme and b) two or more molecules (e.g., two, three, four, five, six, or more molecules) of the second chimeric biosynthetic enzyme. Accordingly, the ratio of any given protein in a biological pathway to any other protein in the pathway can be varied. By way of

example only, the ratio of a first chimeric biological pathway protein to a second chimeric biological pathway protein can be varied from about 0.1:10 to about 10:0.1, e.g., from about 0.1:10 to about 0.5:10, from about 0.5:10 to about 1.0:10, from about 1.0:10 to about 2:10, from about 2:10 to about 5:10, from about 5:10 to about 7:10, from about 7:10 to about 10:10, from about 10:7 to about 10:5, from about 10:5 to about 10:2, from about 10:2 to about 10:1, from about 10:1 to about 10:0.5, or from about 10:0.5 to about 10:1.

[0503] In some embodiments, at least three chimeric biosynthetic enzymes are immobilized onto the synthetic nucleic acid scaffold to comprise a scaffold subunit. In accordance with this embodiment of the present invention, the first chimeric biosynthetic enzyme produces a first product that is a substrate for the second chimeric biosynthetic enzyme, and the second chimeric biological pathway protein produces a second product that is a substrate for the third chimeric biosynthetic enzyme. In these embodiments, a scaffold subunit has immobilized thereon, in order from 3'→5' or 5'→3' of the scaffold a) the first chimeric biosynthetic enzyme, b) the second chimeric biosynthetic enzyme, and c) the third biosynthetic enzyme. The scaffold unit can be repeated two or more times in the nucleic acid construct as described supra.

[0504] In another embodiment of the present invention, at least four chimeric biosynthetic enzymes are immobilized onto the nucleic acid scaffold. In another embodiment of the present invention, at least five chimeric biosynthetic enzymes are immobilized onto the nucleic acid scaffold. It will be apparent from these examples that a sixth, seventh, eighth, ninth, tenth, etc., chimeric biosynthetic enzymes can be immobilized onto the nucleic acid scaffold, that the chimeric proteins are immobilized spatially in the order in which they function in a pathway, and that each protein can be immobilized onto the scaffold in one two, three, four, five, six, seven, eight, nine, ten, or more copies (or molecules).

[0505] In some embodiments, the present disclosure teaches spacing of each DNA binding site within the scaffold nucleic acid. In accordance with this aspect of the present invention, the two or more DNA binding sites are located adjacent to each other within a scaffold subunit, coupled to each other in tandem or separated by at least one spacer nucleotide. The two or more DNA binding sites may be separated from each other by 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, or more spacer nucleotides. The spacing between different DNA binding sites can vary within one scaffold unit (i.e., the spacing between a first and second DNA binding site may differ from the spacing between the second and third DNA binding site). Optimal spacing

between different DNA binding sites within a scaffold subunit will vary depending on the biosynthetic enzyme requirements and the biological pathway being reconstructed, and should be optimized to achieve optimal biological pathway productivity.

Peptide Scaffolding

[0506] In some embodiments, the scaffolding methods of the present disclosure can also be applied to protein/structural scaffolds within the cell. In some embodiments, the present disclosure teaches application of methods disclosed in U.S. Published Patent Application No. 20110008829, which is hereby incorporated by reference in its entirety

Protein-Binding Chimeric Proteins

[0507] In some embodiments, the present disclosure teaches chimeric proteins comprising selected biosynthetic enzymes that are tethered to one or more protein binding domain(s) capable of binding to a recruitment peptide. In accordance with these embodiments, it is expected that the chimeric biosynthetic enzymes will be recruited to a scaffold peptide by interacting with recruitment peptides contained within the scaffold peptide.

[0508] In some embodiments, the biosynthetic enzymes and the protein binding domains are covalently linked. In some embodiments, the biosynthetic enzymes are translationally fused to the protein binding domains. Thus, in some embodiments the chimeric biosynthetic enzymes are formed by coupling the protein binding domain to the amino terminus, the carboxy terminus, or to an internal site within the biosynthetic pathway protein. Persons having skill in the art will recognize the need to ensure that the addition of the protein binding domain does not substantially reduce the activity of the biosynthetic enzyme.

[0509] In some embodiments of the present disclosure, the biosynthetic enzyme is coupled to its protein-binding domain via a short polypeptide linker sequence as described in earlier portions of this disclosure.

[0510] Various protein-binding domains (PBD) are known in the art along with their corresponding recruitment peptides sequences and are suitable for use in the system and methods of the present invention. A non-limiting illustrative discussion of suitable PBD follows.

[0511] SH3

[0512] Suitable PBDs include SH3 domains. SH3 domains include Class I SH3 domains; Class II SH3 domains; and unconventional SH3 domains. Amino acid sequences of SH3 domains are known in the art. See, for example, amino acids 136-189 of the amino acid sequence provided in

GenBank Accession No. NP.sub.--058431 (Homo sapiens Crk protein); amino acids 136-189 of the amino acid sequence provided in GenBank Accession No. AAH31149 (Mus musculus Crk protein); and amino acids 4-77 of the amino acid sequence provided in GenBank Accession No. P27986 (Homo sapiens p85 subunit of phosphatidylinositol 3-kinase).

[0513] In some embodiments, an SH3 domain is a Class I SH3 domain and comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: EGYQYRA LYDYKKEREE DIDLHLGDIL TVNKGSLVAL GFSDGQEARP EEIGWLNGYN ETTGERGDFP GTYVEYI (SEQ ID NO. 44), including all ranges and subranges there between.

[0514] In some embodiments, an SH3 domain is a Class II SH3 domain and comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: YVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRG MIPVPYVEK (SEQ ID NO. 45). As one non-limiting example, an SH3 domain comprises the amino acid sequence: MAEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRG MIPVPYVEKY (SEQ ID NO. 46), including all ranges and subranges there between.

[0515] An SH3 domain binds proline-rich peptides that form a left-handed poly-proline type II helix, where such peptides comprise the minimal consensus sequence Pro-X-X-Pro. In some embodiments, each Pro is preceded by an aliphatic residue. In some embodiments, a recruitment peptide is an SH3 domain ligand. An SH3 domain binds proline-rich peptides that form a left-handed poly-proline type II helix, where such peptides comprise the minimal consensus sequence Pro-X-X-Pro. In some embodiments, each Pro is preceded by an aliphatic residue. Exemplary, non-limiting examples of amino acid sequences of peptides comprising SH3 domain ligands include: RPLPVAP (SEQ ID NO. 47; bound by a Class I SH3 domain); PPPALPPKRRRPG (SEQ ID NO. 48); and PPPALPPKKR (SEQ ID NO. 49; bound by a Class II SH3 domain).

PDZ

[0516] Suitable PBD include PDZ domains. Amino acid sequences of PDZ domains are known in the art. See, for example, amino acids 108-191, amino acids 201-287, and amino acids 354-434 of the amino acid sequence provided in Gen Bank Accession No. AAC52113 (Homo sapiens post-

synaptic density protein 95); and amino acids 80-161 of the amino acid sequence provided in GenBank Accession No. NP__033254 (Mus musculus syntrophin).

[0517] In some embodiments, a suitable PDZ domain comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: EITLERGNSGLGFSIAGGTDNPHIGDDPSIFIT KIIPGGAAAQDGRRLRVNDSILFVNEVDVREVTHTSAAVEALKEAGSIVRLYV (SEQ ID NO. 50), including all ranges and subranges there between.

[0518] In some embodiments, a suitable PDZ domain comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: VMEIKLIKGPKGLGFSIAGGVGNQHIPGDN SIYVTKIIEGGAAHKDGRLQ IGDKILAVNSVGLDVMHEDAVAALKNTYDVVYLKVA (SEQ ID NO.51), including all ranges and subranges there between.

[0519] In some embodiments, a suitable PDZ domain comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: RIVHRGSTGLGFNIVGGEDGEGIFISFILAGGPA DLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQ (SEQ ID NO. 52), including all ranges and subranges there between.

[0520] In some embodiments, a suitable PDZ domain comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: RRVTVRKADAGGLGISIKGGRENKMPILISK IFKGLAADQTEALFVGDAILSVNGEDLSSATHDEAVQALKKKTGKEVVLEVK (SEQ ID NO. 53), including all ranges and sub ranges there between. For example, a PDZ domain can comprise the amino acid sequence MLQRRRVTVRKADAGGLGISIKGGRENKMPILISKIFKGLAADQTEALFVGDAILSVNGE DLSS ATHDEAVQALKKKTGKEVVLEVKYMKEVSPYFKGS (SEQ ID NO.54).

[0521] In some embodiments, a recruitment peptide is a PDZ domain ligand. A PDZ domain binds to the C-terminal 4-5 residues of target proteins. In some embodiments, a consensus PDZ domain

ligand comprises a hydrophobic residue, e.g., Val or Ile, at the carboxyl terminus. Exemplary, non-limiting examples of amino acid sequences of peptides comprising PDZ domain ligands include: IESDV (SEQ ID NO. 55); VKESLV (SEQ ID NO. 56); GVKESLV (SEQ ID NO. 57); GVKQSL (SEQ ID NO. 58); GVKESGA (SEQ ID NO. 59); YVKESLV (SEQ ID NO. 60); and VETDV (SEQ ID NO. 61).

GBD

[0522] Suitable PBD include GTPase-binding domains (GBD), also referred to in the art as CRIB (Cdc42/Rac-interactive binding) motifs. In some embodiments, a GBD binds a Cdc42p-like and/or a Rho-like small GTPase. Amino acid sequences of GBD are known in the art. See, e.g., amino acids 198-240 of the amino acid sequence provided in GenBank Accession No. NP.sub.--001103835 (*Rattus norvegicus* Wiskott-Aldrich syndrome-like protein (WASP)); amino acids 69-112 of the amino acid sequence provided in GenBank Accession No. Q13177 (*Homo sapiens* PAK-2); and amino acids 70-105 of the amino acid sequence provided in GenBank Accession No. P35465 (*Rattus norvegicus* PAK-1). See also the amino acid sequences PAK (75-111), ACK (504-549), and WASP (232-274), presented in FIG. 3A of Garrard et al. (2003) EMBO J. 22:1125. See also the amino acid sequences ACK (505-531), WASP (236-258), PAK1 (70-94), PAK2 (71-91), PAK-4 (6-30), presented in FIG. 1A of Bishop and Hall (2000) Biochem. J. 348:241.

[0523] In some embodiments, a suitable GBD comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: ADI GTPSNFQHIG HVGWDPNTGF DLNNLDPELK NLFDMCGISE (SEQ ID NO. 62), and all ranges and sub ranges there between.

[0524] In some embodiments, a suitable GBD comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: KERPEISLPSDFEHTIHVGFDVAVTGEFTGMPEQWAR (SEQ ID NO. 63), and all ranges and sub ranges there between.

[0525] In some embodiments, a suitable GBD comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence:

MTKADIGTPSNFQHIGHVGWDPNTGFDLNLDPELKNLFDMCGISEAQLKDRETSKVIY
DFIEK TGGVEAVKNELRRQAP (SEQ ID NO. 64), and all ranges and subranges there between.

[0526] In some embodiments, an recruitment peptide is a GBD ligand. An exemplary, non-limiting GBD ligand comprises the amino acid sequence LVGALMHVMQKRSRAIHSSDEGEDQAGDEDED (SEQ ID NO. 65).

Leucine Zipper Peptides

[0527] Suitable PBD include leucine zipper peptides. In some embodiments, leucine zipper peptides are peptides that interact via a coiled-coil domain. Amino acid sequences of leucine zipper domains are known in the art. Leucine zipper peptides include an EE12RR345L leucine zipper peptide; an RR12EE354L leucine zipper peptide; and the like.

[0528] An example of an amino acid sequence of a leucine zipper peptide is an EE12RR345L leucine zipper of the amino acid sequence: LEIEAAFLERENTALETRVAELRQRVQRLR NRVSQYRTRYGPLGGGK (SEQ ID NO. 66).

[0529] In some embodiments, a leucine zipper peptide comprises an amino acid sequence having at least about 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% amino acid sequence similarity to amino acid sequence: LEIEAA FLERENTALETRVAELRQRVQRLRNRVSQYRTRYGPLGGGK (SEQ ID NO. 67), and all ranges and subranges there between. Such a leucine zipper peptide can serve as a PBD or as a recruitment peptide.

[0530] Another non-limiting example of an amino acid sequence of a leucine zipper peptide is an RR12EE345L leucine zipper peptide of the amino acid sequence: LEIRAAFLRQRNTALRT EVAELEQEVQRLENEVSQYETRYGPLGGGK (SEQ ID NO. 68).

[0531] The descriptions above have described the production of chimeric biosynthetic proteins comprising a protein binding domain designed to target (bind to) one or more recruitment peptides located on a scaffold polypeptide. Persons having skill in the art will recognize other compatible variations on this arrangement. For example, in some embodiments, the present disclosure teaches the production of chimeric biosynthetic proteins comprising recruitment peptides targeted by protein binding domains located on a scaffold polypeptide. In other embodiments, the protein binding domain and recruitment peptides are each incorporated into two or more chimeric biosynthetic proteins, such that the chimeric proteins form complexes (e.g., dimers or

heterodimers). One illustrative example of complex formation is the use of compatible leucine zipper domains placed on chimeric biosynthetic proteins, such that two or more chimeric biosynthetic proteins are able to form a complex via the leucine zipper domains.

Scaffold Polypeptide

[0532] In some embodiments, the present disclosure teaches a scaffold polypeptide that organizes biosynthetic pathway enzymes into a functional complex. In some embodiments, the scaffold polypeptides of the present disclosure comprise two or more recruitment peptides. That is, in some embodiments, the scaffold polypeptides of the present disclosure are capable of recruiting two or more chimeric biosynthetic proteins.

[0533] In some embodiments, the scaffold polypeptide of the present disclosure is an exogenous peptide introduced into a host cell (e.g., by the transformation of a DNA sequence encoding for the scaffold polypeptide, or direct introduction of the peptide). In other embodiments, the scaffold polypeptide is a naturally occurring structure within the host cell. That is, in some embodiments, the scaffold polypeptide is an organelle or membrane (e.g., the endoplasmic reticulum, or the golgi apparatus). Thus, in some embodiments, the scaffold polypeptide of the present disclosure includes host-cell structures composed of more than one peptide sequence.

[0534] In some embodiments, the recruitment peptide sequences within the scaffold polypeptide are organized so as to optimize target biosynthetic pathways whose enzymes are being recruited. In some embodiments, the organization of the scaffold polypeptide is similar to that described for DNA scaffolds above. Thus, in some embodiments, scaffold polypeptides contain groupings of recruitment peptides to regulate the order and ratios of various chimeric biosynthetic proteins.

Cell Culture and Fermentation

[0535] Cells of the present disclosure can be cultured in conventional nutrient media modified as appropriate for any desired biosynthetic reactions or selections. In some embodiments, the present disclosure teaches culture in inducing media for activating promoters. In some embodiments, the present disclosure teaches media with selection agents, including selection agents of transformants (e.g., antibiotics), or selection of organisms suited to grow under inhibiting conditions (e.g., high ethanol conditions). In some embodiments, the present disclosure teaches growing cell cultures in media optimized for cell growth. In other embodiments, the present disclosure teaches growing cell cultures in media optimized for product yield. In some embodiments, the present disclosure

teaches growing cultures in media capable of inducing cell growth and also contains the necessary precursors for final product production (e.g., high levels of sugars for ethanol production).

[0536] Culture conditions, such as temperature, pH and the like, are those suitable for use with the host cell selected for expression, and will be apparent to those skilled in the art. As noted, many references are available for the culture and production of many cells, including cells of bacterial, plant, animal (including mammalian) and archaeobacterial origin. See e.g., Sambrook, Ausubel (all supra), as well as Berger, *Guide to Molecular Cloning Techniques, Methods in Enzymology* volume 152 Academic Press, Inc., San Diego, CA; and Freshney (1994) *Culture of Animal Cells, a Manual of Basic Technique*, third edition, Wiley-Liss, New York and the references cited therein; Doyle and Griffiths (1997) *Mammalian Cell Culture: Essential Techniques* John Wiley and Sons, NY; Humason (1979) *Animal Tissue Techniques*, fourth edition W.H. Freeman and Company; and Ricciardelle *et al.*, (1989) *In Vitro Cell Dev. Biol.* 25:1016-1024, all of which are incorporated herein by reference. For plant cell culture and regeneration, Payne *et al.* (1992) *Plant Cell and Tissue Culture in Liquid Systems* John Wiley & Sons, Inc. New York, N.Y.; Gamborg and Phillips (eds) (1995) *Plant Cell, Tissue and Organ Culture; Fundamental Methods Springer Lab Manual*, Springer-Verlag (Berlin Heidelberg N.Y.); Jones, ed. (1984) *Plant Gene Transfer and Expression Protocols*, Humana Press, Totowa, N.J. and *Plant Molecular Biology* (1993) R. R. D. Croy, Ed. Bios Scientific Publishers, Oxford, U.K. ISBN 0 12 198370 6, all of which are incorporated herein by reference. Cell culture media in general are set forth in Atlas and Parks (eds.) *The Handbook of Microbiological Media* (1993) CRC Press, Boca Raton, Fla., which is incorporated herein by reference. Additional information for cell culture is found in available commercial literature such as the *Life Science Research Cell Culture Catalogue* from Sigma-Aldrich, Inc (St Louis, Mo.) (“Sigma-LSRCCC”) and, for example, *The Plant Culture Catalogue* and supplement also from Sigma-Aldrich, Inc (St Louis, Mo.) (“Sigma-PCCS”), all of which are incorporated herein by reference.

[0537] The culture medium to be used must in a suitable manner satisfy the demands of the respective strains. Descriptions of culture media for various microorganisms are present in the “Manual of Methods for General Bacteriology” of the American Society for Bacteriology (Washington D.C., USA, 1981).

[0538] The present disclosure furthermore provides a process for fermentative preparation of a product of interest, comprising the steps of: a) culturing a microorganism according to the present

disclosure in a suitable medium, resulting in a fermentation broth; and b) concentrating the product of interest in the fermentation broth of a) and/or in the cells of the microorganism.

[0539] In some embodiments, the present disclosure teaches that the microorganisms produced may be cultured continuously—as described, for example, in WO 05/021772—or discontinuously in a batch process (batch cultivation) or in a fed-batch or repeated fed-batch process for the purpose of producing the desired organic-chemical compound. A summary of a general nature about known cultivation methods is available in the textbook by Chmiel (*Bioprozeßtechnik. 1: Einführung in die Bioverfahrenstechnik* (Gustav Fischer Verlag, Stuttgart, 1991)) or in the textbook by Storhas (*Bioreaktoren and periphere Einrichtungen* (Vieweg Verlag, Braunschweig/Wiesbaden, 1994)).

[0540] In some embodiments, the cells of the present disclosure are grown under batch or continuous fermentations conditions.

[0541] Classical batch fermentation is a closed system, wherein the compositions of the medium is set at the beginning of the fermentation and is not subject to artificial alternations during the fermentation. A variation of the batch system is a fed-batch fermentation which also finds use in the present disclosure. In this variation, the substrate is added in increments as the fermentation progresses. Fed-batch systems are useful when catabolite repression is likely to inhibit the metabolism of the cells and where it is desirable to have limited amounts of substrate in the medium. Batch and fed-batch fermentations are common and well known in the art.

[0542] Continuous fermentation is a system where a defined fermentation medium is added continuously to a bioreactor and an equal amount of conditioned medium is removed simultaneously for processing and harvesting of desired biomolecule products of interest. In some embodiments, continuous fermentation generally maintains the cultures at a constant high density where cells are primarily in log phase growth. In some embodiments, continuous fermentation generally maintains the cultures at a stationary or late log/stationary, phase growth. Continuous fermentation systems strive to maintain steady state growth conditions.

[0543] Methods for modulating nutrients and growth factors for continuous fermentation processes as well as techniques for maximizing the rate of product formation are well known in the art of industrial microbiology.

[0544] For example, a non-limiting list of carbon sources for the cultures of the present disclosure include, sugars and carbohydrates such as, for example, glucose, sucrose, lactose, fructose, maltose, molasses, sucrose-containing solutions from sugar beet or sugar cane processing, starch,

starch hydrolysate, and cellulose; oils and fats such as, for example, soybean oil, sunflower oil, groundnut oil and coconut fat; fatty acids such as, for example, palmitic acid, stearic acid, and linoleic acid; alcohols such as, for example, glycerol, methanol, and ethanol; and organic acids such as, for example, acetic acid or lactic acid.

[0545] A non-limiting list of the nitrogen sources for the cultures of the present disclosure include, organic nitrogen-containing compounds such as peptones, yeast extract, meat extract, malt extract, corn steep liquor, soybean flour, and urea; or inorganic compounds such as ammonium sulfate, ammonium chloride, ammonium phosphate, ammonium carbonate, and ammonium nitrate. The nitrogen sources can be used individually or as a mixture.

[0546] A non-limiting list of the possible phosphorus sources for the cultures of the present disclosure include, phosphoric acid, potassium dihydrogen phosphate or dipotassium hydrogen phosphate or the corresponding sodium-containing salts.

[0547] The culture medium may additionally comprise salts, for example in the form of chlorides or sulfates of metals such as, for example, sodium, potassium, magnesium, calcium and iron, such as, for example, magnesium sulfate or iron sulfate, which are necessary for growth.

[0548] Finally, essential growth factors such as amino acids, for example homoserine and vitamins, for example thiamine, biotin or pantothenic acid, may be employed in addition to the abovementioned substances.

[0549] In some embodiments, the pH of the culture can be controlled by any acid or base, or buffer salt, including, but not limited to sodium hydroxide, potassium hydroxide, ammonia, or aqueous ammonia; or acidic compounds such as phosphoric acid or sulfuric acid in a suitable manner. In some embodiments, the pH is generally adjusted to a value of from 6.0 to 8.5, preferably 6.5 to 8.

[0550] In some embodiments, the cultures of the present disclosure may include an anti-foaming agent such as, for example, fatty acid polyglycol esters. In some embodiments the cultures of the present disclosure are modified to stabilize the plasmids of the cultures by adding suitable selective substances such as, for example, antibiotics.

[0551] In some embodiments, the culture is carried out under aerobic conditions. In order to maintain these conditions, oxygen or oxygen-containing gas mixtures such as, for example, air are introduced into the culture. It is likewise possible to use liquids enriched with hydrogen peroxide. The fermentation is carried out, where appropriate, at elevated pressure, for example at an elevated pressure of from 0.03 to 0.2 MPa. The temperature of the culture is normally from 20°C to 45°C

and preferably from 25°C to 40°C, particularly preferably from 30°C to 37°C. In batch or fed-batch processes, the cultivation is preferably continued until an amount of the desired product of interest (*e.g.* an organic-chemical compound) sufficient for being recovered has formed. This aim can normally be achieved within 10 hours to 160 hours. In continuous processes, longer cultivation times are possible. The activity of the microorganisms results in a concentration (accumulation) of the product of interest in the fermentation medium and/or in the cells of said microorganisms.

[0552] In some embodiments, the culture is carried out under anaerobic conditions.

Screening

[0553] In some embodiments, the present disclosure teaches high-throughput initial screenings. In other embodiments, the present disclosure also teaches robust tank-based validations of performance data (*see* Figure 6B).

[0554] In some embodiments, the high-throughput screening process is designed to predict performance of strains in bioreactors. As previously described, culture conditions are selected to be suitable for the organism and reflective of bioreactor conditions. Individual colonies are picked and transferred into 96 well plates and incubated for a suitable amount of time. Cells are subsequently transferred to new 96 well plates for additional seed cultures, or to production cultures. Cultures are incubated for varying lengths of time, where multiple measurements may be made. These may include measurements of product, biomass or other characteristics that predict performance of strains in bioreactors. High-throughput culture results are used to predict bioreactor performance.

[0555] In some embodiments, the tank-based performance validation is used to confirm performance of strains isolated by high-throughput screening. Candidate strains are screened using bench scale fermentation reactors (*e.g.*, reactors disclosed in Table 5 of the present disclosure) for relevant strain performance characteristics such as productivity or yield.

Product Recovery and Quantification

[0556] Methods for screening for the production of products of interest are known to those of skill in the art and are discussed throughout the present specification. Such methods may be employed when screening the strains of the disclosure.

[0557] In some embodiments, the present disclosure teaches methods of improving strains designed to produce non-secreted intracellular products. For example, the present disclosure teaches methods of improving the robustness, yield, efficiency, or overall desirability of cell

cultures producing intracellular enzymes, oils, pharmaceuticals, or other valuable small molecules or peptides. The recovery or isolation of non-secreted intracellular products can be achieved by lysis and recovery techniques that are well known in the art, including those described herein.

[0558] For example, in some embodiments, cells of the present disclosure can be harvested by centrifugation, filtration, settling, or other method. Harvested cells are then disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, or other methods, which are well known to those skilled in the art.

[0559] The resulting product of interest, *e.g.* a polypeptide, may be recovered/isolated and optionally purified by any of a number of methods known in the art. For example, a product polypeptide may be isolated from the nutrient medium by conventional procedures including, but not limited to: centrifugation, filtration, extraction, spray-drying, evaporation, chromatography (*e.g.*, ion exchange, affinity, hydrophobic interaction, chromatofocusing, and size exclusion), or precipitation. Finally, high performance liquid chromatography (HPLC) can be employed in the final purification steps. (*See* for example Purification of intracellular protein as described in Parry *et al.*, 2001, *Biochem. J.*353:117, and Hong *et al.*, 2007, *Appl. Microbiol. Biotechnol.* 73:1331, both incorporated herein by reference).

[0560] In addition to the references noted *supra*, a variety of purification methods are well known in the art, including, for example, those set forth in: Sandana (1997) *Bioseparation of Proteins*, Academic Press, Inc.; Bollag *et al.* (1996) *Protein Methods*, 2nd Edition, Wiley-Liss, NY; Walker (1996) *The Protein Protocols Handbook* Humana Press, NJ; Harris and Angal (1990) *Protein Purification Applications: A Practical Approach*, IRL Press at Oxford, Oxford, England; Harris and Angal *Protein Purification Methods: A Practical Approach*, IRL Press at Oxford, Oxford, England; Scopes (1993) *Protein Purification: Principles and Practice* 3rd Edition, Springer Verlag, NY; Janson and Ryden (1998) *Protein Purification: Principles, High Resolution Methods and Applications*, Second Edition, Wiley-VCH, NY; and Walker (1998) *Protein Protocols on CD-ROM*, Humana Press, NJ, all of which are incorporated herein by reference.

[0561] In some embodiments, the present disclosure teaches the methods of improving strains designed to produce secreted products. For example, the present disclosure teaches methods of improving the robustness, yield, efficiency, or overall desirability of cell cultures producing valuable small molecules or peptides.

[0562] In some embodiments, immunological methods may be used to detect and/or purify secreted or non-secreted products produced by the cells of the present disclosure. In one example approach, antibody raised against a product molecule (*e.g.*, against an insulin polypeptide or an immunogenic fragment thereof) using conventional methods is immobilized on beads, mixed with cell culture media under conditions in which the endoglucanase is bound, and precipitated. In some embodiments, the present disclosure teaches the use of enzyme-linked immunosorbent assays (ELISA).

[0563] In other related embodiments, immunochromatography is used, as disclosed in U.S. Pat. No. 5,591,645, U.S. Pat. No. 4,855,240, U.S. Pat. No. 4,435,504, U.S. Pat. No. 4,980,298, and Se-Hwan Paek, *et al.*, “Development of rapid One-Step Immunochromatographic assay, Methods”, 22, 53-60, 2000), each of which are incorporated by reference herein. A general immunochromatography detects a specimen by using two antibodies. A first antibody exists in a test solution or at a portion at an end of a test piece in an approximately rectangular shape made from a porous membrane, where the test solution is dropped. This antibody is labeled with latex particles or gold colloidal particles (this antibody will be called as a labeled antibody hereinafter). When the dropped test solution includes a specimen to be detected, the labeled antibody recognizes the specimen so as to be bonded with the specimen. A complex of the specimen and labeled antibody flows by capillarity toward an absorber, which is made from a filter paper and attached to an end opposite to the end having included the labeled antibody. During the flow, the complex of the specimen and labeled antibody is recognized and caught by a second antibody (it will be called as a tapping antibody hereinafter) existing at the middle of the porous membrane and, as a result of this, the complex appears at a detection part on the porous membrane as a visible signal and is detected.

[0564] In some embodiments, the screening methods of the present disclosure are based on photometric detection techniques (absorption, fluorescence). For example, in some embodiments, detection may be based on the presence of a fluorophore detector such as GFP bound to an antibody. In other embodiments, the photometric detection may be based on the accumulation on the desired product from the cell culture. In some embodiments, the product may be detectable via UV of the culture or extracts from said culture.

[0565] Persons having skill in the art will recognize that the methods of the present disclosure are compatible with host cells producing any desirable biomolecule product of interest. Table 2 below

presents a non-limiting list of the product categories, biomolecules, and host cells, included within the scope of the present disclosure. These examples are provided for illustrative purposes, and are not meant to limit the applicability of the presently disclosed technology in any way.

Table 2. – A non-limiting list of the host cells and products of interest of the present disclosure.

Product category	Products	Host category	Hosts
Amino acids	Lysine	Bacteria	<i>Escherichia coli</i>
Amino acids	Methionine	Bacteria	<i>Escherichia coli</i>
Amino acids	MSG	Bacteria	<i>Escherichia coli</i>
Amino acids	Threonine	Bacteria	<i>Escherichia coli</i>
Amino acids	Threonine	Bacteria	<i>Escherichia coli</i>
Amino acids	Tryptophan	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Agarwood	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Ambrox	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Nootkatone	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Patchouli oil	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Saffron	Bacteria	<i>Escherichia coli</i>

Product category	Products	Host category	Hosts
Flavor & Fragrance	Sandalwood oil	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Valencene	Bacteria	<i>Escherichia coli</i>
Flavor & Fragrance	Vanillin	Bacteria	<i>Escherichia coli</i>
Food	CoQ10/Ubiquinol	Bacteria	<i>Escherichia coli</i>
Food	Omega 3 fatty acids	Bacteria	<i>Escherichia coli</i>
Food	Omega 6 fatty acids	Bacteria	<i>Escherichia coli</i>
Food	Vitamin B12	Bacteria	<i>Escherichia coli</i>
Food	Vitamin B2	Bacteria	<i>Escherichia coli</i>
Food	Vitamin B2	Bacteria	<i>Escherichia coli</i>
Food	Erythritol	Bacteria	<i>Escherichia coli</i>
Food	Erythritol	Bacteria	<i>Escherichia coli</i>
Food	Erythritol	Bacteria	<i>Escherichia coli</i>
Food	Steviol glycosides	Bacteria	<i>Escherichia coli</i>
Hydrocolloids	Diutan gum	Bacteria	<i>Escherichia coli</i>
Hydrocolloids	Gellan gum	Bacteria	<i>Escherichia coli</i>

Product category	Products	Host category	Hosts
Hydrocolloids	Xanthan gum	Bacteria	<i>Escherichia coli</i>
Intermediates	1,3-PDO	Bacteria	<i>Escherichia coli</i>
Intermediates	1,4-BDO	Bacteria	<i>Escherichia coli</i>
Intermediates	Butadiene	Bacteria	<i>Escherichia coli</i>
Intermediates	n-butanol	Bacteria	<i>Escherichia coli</i>
Organic acids	Citric acid	Bacteria	<i>Escherichia coli</i>
Organic acids	Citric acid	Bacteria	<i>Escherichia coli</i>
Organic acids	Gluconic acid	Bacteria	<i>Escherichia coli</i>
Organic acids	Itaconic acid	Bacteria	<i>Escherichia coli</i>
Organic acids	Lactic acid	Bacteria	<i>Escherichia coli</i>
Organic acids	Lactic acid	Bacteria	<i>Escherichia coli</i>
Organic acids	LCDAAs - DDDA	Bacteria	<i>Escherichia coli</i>
Polyketides/Ag	Spinosad	Bacteria	<i>Escherichia coli</i>
Polyketides/Ag	Spinetoram	Bacteria	<i>Escherichia coli</i>

Heterologous Genes Of Interest

[0566] In one embodiment, provided herein are methods for expressing heterologous genes in a microbial host cell. The heterologous gene can be introduced into the microbial host cell using the methods provided herein and/or known in the art such that the microbial host cell uses the heterologous gene to produce a product of interest. In one embodiment, the microbial host cell is a strain of *E. coli*. The strain of *E. coli* can be any strain of *E. coli* known in the art and/or provided herein. The heterologous gene can be a wild-type version of said gene or a mutant thereof. The

heterologous gene can be operably linked to a promoter, terminator, protein solubility tag, protein degradation tag, or any combination thereof. Operably linking the heterologous gene to the promoter, terminator, protein solubility tag or protein degradation tag can be accomplished using the promoter swap, terminator swap, solubility tag swap and/or degradation swap methods provided throughout this disclosure.

[0567] In one embodiment, the heterologous gene is operably linked to a promoter selected from Table 1. In one embodiment, the heterologous gene is operably linked to a 60-90 bp chimeric synthetic promoter sequence, wherein the chimeric synthetic promoter consists of a distal portion of the lambda phage p_R promoter, variable -35 and -10 regions of the lambda phage p_L and p_R promoters, core portions of the the lambda phage p_L and p_R promoters and either a 5' UTR/Ribosomal Binding Site (RBS) portion of the lambda phage p_R promoter or a 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene. The nucleic acid sequences of the distal portion of the lambda phage p_R promoter, variable -35 and -10 regions of the lambda phage p_L and p_R promoters, core portions of the lambda phage p_L and p_R promoters and either a 5' UTR/Ribosomal Binding Site (RBS) portion of the lambda phage p_R promoter or a 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene for use in the chimeric synthetic promoter can be selected from the nucleic acid sequences found in Table 1.5. In one embodiment, the heterologous gene can be operably linked to a chimeric synthetic promoter that has a nucleic acid sequence selected from SEQ ID NOs. 132-207 found in Table 1.4.

[0568] In one embodiment, the heterologous gene is operably linked to a terminator selected from Table 1.2. In another embodiment, the heterologous gene is operably linked to a terminator sequence selected from Table 19.

[0569] In one embodiment, the heterologous gene is operably linked to a solubility tag selected from Table 17.

[0570] In one embodiment, the heterologous gene is operably linked to a degradation tag sequence selected from Table 18.

[0571] Further to the above embodiments, the heterologous gene can be any of the genes required to generate the products of interest found in Table 2 or any gene known in the art that can be expressed as a heterologous gene in the microbial host cell (e.g., *E. coli*) to produce a product of interest.

[0572] In one embodiment, the heterologous gene is a gene that is part of the lysine biosynthetic pathway as illustrated in Figure 19. Further to this embodiment, the heterologous gene can be selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene, the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene. In one embodiment, a heterologous gene that is part of the lysine pathway as provided herein is operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.

[0573] In one embodiment, the heterologous gene is a gene that is part of the lycopene biosynthetic pathway as illustrated, for example, in Figure 59. Further to this embodiment, the heterologous gene can be selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA* gene, the *ispB* gene, the *crtE* gene, the *crtB* gene, the *crtI* gene, the *crtY* gene, the *ymgA* gene, the *dxr* gene, the *elbA* gene, the *gdhA* gene, the *appY* gene, the *elbB* gene, or the *ymgB* gene. In one embodiment, a heterologous gene that is part of the lycopene pathway as provided herein is operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.

[0574] In one embodiment, the heterologous gene is a gene that encodes a biopharmaceutical or a gene in pathway for generating a biopharmaceutical. In one embodiment, the microbial host cell is *E. coli* and the biopharmaceutical is any biopharmaceutical that has been shown to be produced in *E. coli*. The biopharmaceutical can be selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone). In one embodiment, a heterologous gene that encodes a biopharmaceutical or a gene in a pathway that generates a biopharmaceutical as provided herein is operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.

Selection Criteria and Goals

[0575] The selection criteria applied to the methods of the present disclosure will vary with the specific goals of the strain improvement program. The present disclosure may be adapted to meet any program goals. For example, in some embodiments, the program goal may be to maximize single batch yields of reactions with no immediate time limits. In other embodiments, the program goal may be to rebalance biosynthetic yields to produce a specific product, or to produce a particular ratio of products. In other embodiments, the program goal may be to modify the chemical structure of a product, such as lengthening the carbon chain of a polymer. In some embodiments, the program goal may be to improve performance characteristics such as yield, titer, productivity, by-product elimination, tolerance to process excursions, optimal growth temperature and growth rate. In some embodiments, the program goal is improved host performance as measured by volumetric productivity, specific productivity, yield or titre, of a product of interest produced by a microbe.

[0576] In other embodiments, the program goal may be to optimize synthesis efficiency of a commercial strain in terms of final product yield per quantity of inputs (*e.g.*, total amount of ethanol produced per pound of sucrose). In other embodiments, the program goal may be to optimize synthesis speed, as measured for example in terms of batch completion rates, or yield rates in continuous culturing systems. In other embodiments, the program goal may be to increase strain resistance to a particular phage, or otherwise increase strain vigor/robustness under culture conditions.

[0577] In some embodiments, strain improvement projects may be subject to more than one goal. In some embodiments, the goal of the strain project may hinge on quality, reliability, or overall profitability. In some embodiments, the present disclosure teaches methods of associated selected mutations or groups of mutations with one or more of the strain properties described above.

[0578] Persons having ordinary skill in the art will recognize how to tailor strain selection criteria to meet the particular project goal. For example, selections of a strain's single batch max yield at reaction saturation may be appropriate for identifying strains with high single batch yields. Selection based on consistency in yield across a range of temperatures and conditions may be appropriate for identifying strains with increased robustness and reliability.

[0579] In some embodiments, the selection criteria for the initial high-throughput phase and the tank-based validation will be identical. In other embodiments, tank-based selection may operate under additional and/or different selection criteria. For example, in some embodiments, high-

throughput strain selection might be based on single batch reaction completion yields, while tank-based selection may be expanded to include selections based on yields for reaction speed.

Sequencing

[0580] In some embodiments, the present disclosure teaches whole-genome sequencing of the organisms described herein. In other embodiments, the present disclosure also teaches sequencing of plasmids, PCR products, and other oligos as quality controls to the methods of the present disclosure. Sequencing methods for large and small projects are well known to those in the art.

[0581] In some embodiments, any high-throughput technique for sequencing nucleic acids can be used in the methods of the disclosure. In some embodiments, the present disclosure teaches whole genome sequencing. In other embodiments, the present disclosure teaches amplicon sequencing ultra deep sequencing to identify genetic variations. In some embodiments, the present disclosure also teaches novel methods for library preparation, including tagmentation (see WO/2016/073690). DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary; sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing; 454 sequencing; allele specific hybridization to a library of labeled oligonucleotide probes; sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation; real time monitoring of the incorporation of labeled nucleotides during a polymerization step; polony sequencing; and SOLiD sequencing.

[0582] In one aspect of the disclosure, high-throughput methods of sequencing are employed that comprise a step of spatially isolating individual molecules on a solid surface where they are sequenced in parallel. Such solid surfaces may include nonporous surfaces (such as in Solexa sequencing, *e.g.* Bentley et al, Nature, 456: 53-59 (2008) or Complete Genomics sequencing, *e.g.* Drmanac et al, Science, 327: 78-81 (2010)), arrays of wells, which may include bead- or particle-bound templates (such as with 454, *e.g.* Margulies et al, Nature, 437: 376-380 (2005) or Ion Torrent sequencing, U.S. patent publication 2010/0137143 or 2010/0304982), micromachined membranes (such as with SMRT sequencing, *e.g.* Eid et al, Science, 323: 133-138 (2009)), or bead arrays (as with SOLiD sequencing or polony sequencing, *e.g.* Kim et al, Science, 316: 1481-1414 (2007)).

[0583] In another embodiment, the methods of the present disclosure comprise amplifying the isolated molecules either before or after they are spatially isolated on a solid surface. Prior

amplification may comprise emulsion-based amplification, such as emulsion PCR, or rolling circle amplification. Also taught is Solexa-based sequencing where individual template molecules are spatially isolated on a solid surface, after which they are amplified in parallel by bridge PCR to form separate clonal populations, or clusters, and then sequenced, as described in Bentley et al (cited above) and in manufacturer's instructions (e.g. TruSeq™ Sample Preparation Kit and Data Sheet, Illumina, Inc., San Diego, Calif., 2010); and further in the following references: U.S. Pat. Nos. 6,090,592; 6,300,070; 7,115,400; and EP0972081B1; which are incorporated by reference.

[0584] In one embodiment, individual molecules disposed and amplified on a solid surface form clusters in a density of at least 10^5 clusters per cm^2 ; or in a density of at least 5×10^5 per cm^2 ; or in a density of at least 10^6 clusters per cm^2 . In one embodiment, sequencing chemistries are employed having relatively high error rates. In such embodiments, the average quality scores produced by such chemistries are monotonically declining functions of sequence read lengths. In one embodiment, such decline corresponds to 0.5 percent of sequence reads have at least one error in positions 1-75; 1 percent of sequence reads have at least one error in positions 76-100; and 2 percent of sequence reads have at least one error in positions 101-125.

Computational Analysis and Prediction of Effects of Genome-Wide Genetic Design Criteria

[0585] In some embodiments, the present disclosure teaches methods of predicting the effects of particular genetic alterations being incorporated into a given host strain. In further aspects, the disclosure provides methods for generating proposed genetic alterations that should be incorporated into a given host strain, in order for said host to possess a particular phenotypic trait or strain parameter. In given aspects, the disclosure provides predictive models that can be utilized to design novel host strains.

[0586] In some embodiments, the present disclosure teaches methods of analyzing the performance results of each round of screening and methods for generating new proposed genome-wide sequence modifications predicted to enhance strain performance in the following round of screening.

[0587] In some embodiments, the present disclosure teaches that the system generates proposed sequence modifications to host strains based on previous screening results. In some embodiments, the recommendations of the present system are based on the results from the immediately preceding screening. In other embodiments, the recommendations of the present system are based on the cumulative results of one or more of the preceding screenings.

[0588] In some embodiments, the recommendations of the present system are based on previously developed HTP genetic design libraries. For example, in some embodiments, the present system is designed to save results from previous screenings, and apply those results to a different project, in the same or different host organisms.

[0589] In other embodiments, the recommendations of the present system are based on scientific insights. For example, in some embodiments, the recommendations are based on known properties of genes (from sources such as annotated gene databases and the relevant literature), codon optimization, transcriptional slippage, uORFs, or other hypothesis driven sequence and host optimizations.

[0590] In some embodiments, the proposed sequence modifications to a host strain recommended by the system, or predictive model, are carried out by the utilization of one or more of the disclosed molecular tools sets comprising: (1) Promoter swaps, (2) SNP swaps, (3) Start/Stop codon exchanges, (4) Sequence optimization, (5) Stop swaps, (6) Solubility Tag swaps, (7) Degradation Tag swaps and (8) Epistasis mapping.

[0591] The HTP genetic engineering platform described herein is agnostic with respect to any particular microbe or phenotypic trait (*e.g.* production of a particular compound). That is, the platform and methods taught herein can be utilized with *any* host cell to engineer said host cell to have *any* desired phenotypic trait. Furthermore, the lessons learned from a given HTP genetic engineering process used to create one novel host cell, can be applied to any number of other host cells, as a result of the storage, characterization, and analysis of a myriad of process parameters that occurs during the taught methods.

[0592] As alluded to in the epistatic mapping section, it is possible to estimate the performance (a.k.a. score) of a hypothetical strain obtained by consolidating a collection of mutations from a HTP genetic design library into a particular background *via* some preferred predictive model. Given such a predictive model, it is possible to score and rank all hypothetical strains accessible to the mutation library *via* combinatorial consolidation. The below section outlines particular models utilized in the present HTP platform.

Predictive Strain Design

[0593] Described herein is an approach for predictive strain design, including: methods of describing genetic changes and strain performance, predicting strain performance based on the composition of changes in the strain, recommending candidate designs with high predicted

performance, and filtering predictions to optimize for second-order considerations, *e.g.* similarity to existing strains, epistasis, or confidence in predictions.

Inputs to Strain Design Model

[0594] In one embodiment, for the sake of ease of illustration, input data may comprise two components: (1) sets of genetic changes and (2) relative strain performance. Those skilled in the art will recognize that this model can be readily extended to consider a wide variety of inputs, while keeping in mind the countervailing consideration of overfitting. In addition to genetic changes, some of the input parameters (independent variables) that can be adjusted are cell types (genus, species, strain, phylogenetic characterization, etc.) and process parameters (*e.g.*, environmental conditions, handling equipment, modification techniques, etc.) under which fermentation is conducted with the cells.

[0595] The sets of genetic changes can come from the previously discussed collections of genetic perturbations termed HTP genetic design libraries. The relative strain performance can be assessed based upon any given parameter or phenotypic trait of interest (*e.g.* production of a compound, small molecule, or product of interest).

[0596] Cell types can be specified in general categories such as prokaryotic and eukaryotic systems, genus, species, strain, tissue cultures (*vs.* disperse cells), etc. Process parameters that can be adjusted include temperature, pressure, reactor configuration, and medium composition. Examples of reactor configuration include the volume of the reactor, whether the process is a batch or continuous, and, if continuous, the volumetric flow rate, etc. One can also specify the support structure, if any, on which the cells reside. Examples of medium composition include the concentrations of electrolytes, nutrients, waste products, acids, pH, and the like.

Sets of Genetic Changes From Selected HTP Genetic Design Libraries to be Utilized in the Initial Linear Regression Model that Subsequently is Used to Create the Predictive Strain Design Model

[0597] An example, a set of entries from a table of genetic changes in *Corynebacterium* is shown below in Table 3. Each row indicates a genetic change in strain 7000051473, as well as metadata about the mechanism of change, *e.g.* promoter swap or SNP swap. *aceE*, *zwf*, and *pyc* are all related to the citric acid cycle.

[0598] In this case strain 7000051473 has a total of 7 changes. “Last change” means the change in this strain represents the most recent modification in this strain lineage. Thus, comparing this

strain's performance to the performance of its parent represents a data point concerning the performance of the "last change" mutation.

Table 3 - Strain design entry table for strain 7000051473

strain	name	library	change	from	to	last_change
7000051473	dlc19_42	proswp	pcg3121	cg1144	pcg3121_cg1144	1
7000051473	dlc19_42	scswp	acee atg>ttg	ttg	acee_atg	0
7000051473	dlc19_42	snpswp	dss_033	NA	na	0
7000051473	dlc19_42	snpswp	dss_084	NA	t	0
7000051473	dlc19_42	snpswp	dss_316	NA	na	0
7000051473	dlc19_42	proswp	pcg0007_39	zwf	pcg0007_39_zwf	0
7000051473	dlc19_42	proswp	pcg1860	pyc	pcg1860_pyc	0

Built Strain Performance Assessment

[0599] The goal of the taught model is to predict strain performance based on the composition of genetic changes introduced to the strain. To construct a standard for comparison, strain performance is computed relative to a common reference strain, by first calculating the median performance per strain, per assay plate. Relative performance is then computed as the difference in average performance between an engineered strain and the common reference strain within the same plate. Restricting the calculations to within-plate comparisons ensures that the samples under consideration all received the same experimental conditions.

[0600] Figure 23 shows the distribution of relative strain performances of *Corynebacterium* for the input data under consideration. A relative performance of zero indicates that the engineered strain performed equally well to the in-plate base or "reference" strain. Of interest is the ability of the predictive model to identify the strains that are likely to perform significantly above zero. Further, and more generally, of interest is whether any given strain outperforms its parent by some criteria. In practice, the criteria can be a product titer meeting or exceeding some threshold above

the parent level, though having a statistically significant difference from the parent in the desired direction could also be used instead or in addition. The role of the base or “reference” strain is simply to serve as an added normalization factor for making comparisons within or between plates.

[0601] A concept to keep in mind is that of differences between: parent strain and reference strain. The parent strain is the background that was used for a current round of mutagenesis. The reference strain is a control strain run in every plate to facilitate comparisons, especially between plates, and is typically the “base strain” as referenced above. But since the base strain (*e.g.*, the wild-type or industrial strain being used to benchmark overall performance) is not necessarily a “base” in the sense of being a mutagenesis target in a given round of strain improvement, a more descriptive term is “reference strain.”

[0602] In summary, a base/reference strain is used to benchmark the performance of built strains, generally, while the parent strain is used to benchmark the performance of a specific genetic change in the relevant genetic background.

Ranking the Performance of Built Strains with Linear Regression

[0603] The goal of the disclosed model is to rank the performance of built strains, by describing relative strain performance, as a function of the composition of genetic changes introduced into the built strains. As discussed throughout the disclosure, the various HTP genetic design libraries provide the repertoire of possible genetic changes (*e.g.*, genetic perturbations/alterations) that are introduced into the engineered strains. Linear regression is the basis for the currently described exemplary predictive model.

[0604] The below table (*i.e.*, Table 4) contains example input for regression-based modeling. The strain performances are ranked relative to a common base strain, as a function of the composition of the genetic changes contained in the strain.

[0605] Each column heading represents a genetic change, a “1” represents the presence of the change, whereas a “0” represents the absence of a change. “DSS” refers to SNP swaps from a particular library (first 3 columns after relative `_perf`). The last 3 columns are promoter swaps, where the `pcgXXXX` denotes the particular promoter, and the last 3 letters represent the gene the promoter is being applied to. The genes are related to central metabolism. The promoters are from *Corynebacterium glutamicum* (hence the “cg” notation). Further information on the utilized promoters can be found in Table 1, listing promoters P1-P8, and the sequence listing of the present application. Further, detailed information on each promoter P1-P8 can be found in U.S. Provisional

Application No. 62/264,232, filed on December 07, 2015, and entitled “Promoters from *Corynebacterium glutamicum*,” which is incorporated herein by reference. For ease of reference, in the below table, pcg3121 = P8; pcg0755 = P4; and pcg1860 = P3.

Table 4 - Summary of genetic changes and their effect on relative performance.

relative_perf	dss_033	dss_034	dss_056	pcg3121_pgi	pcg0755_zwf	pcg1860_pyc
0.1358908	0	0	0	0	0	1
-1.8946985	1	0	0	1	0	1
-0.0222045	0	0	0	1	0	0
0.6342183	1	0	1	0	0	0
-0.0803285	1	1	0	0	0	0
2.6468117	0	0	0	1	0	0

Linear Regression to Characterize Built Strains

[0606] Linear regression is an attractive method for the described HTP genomic engineering platform, because of the ease of implementation and interpretation. The resulting regression coefficients can be interpreted as the average increase or decrease in relative strain performance attributable to the presence of each genetic change.

[0607] For example, as seen in Figure 24, this technique allows us to conclude that changing the pgi promoter to pcg3121 improves relative strain performance by approximately 5 units on average and is thus a potentially highly desirable change, in the absence of any negative epistatic interactions (note: the input is a unit-less normalized value).

[0608] The taught method therefore uses linear regression models to describe/characterize and rank built strains, which have various genetic perturbations introduced into their genomes from the various taught libraries.

Predictive Design Modeling

[0609] The linear regression model described above, which utilized data from constructed strains, can be used to make performance predictions for strains that haven't yet been built.

[0610] The procedure can be summarized as follows: generate *in silico* all possible configurations of genetic changes → use the regression model to predict relative strain performance → order the candidate strain designs by performance. Thus, by utilizing the regression model to predict the performance of as-yet-unbuilt strains, the method allows for the production of higher performing strains, while simultaneously conducting fewer experiments.

Generate Configurations

[0611] When constructing a model to predict performance of as-yet-unbuilt strains, the first step is to produce a sequence of design candidates. This is done by fixing the total number of genetic changes in the strain, and then defining all possible combinations of genetic changes. For example, one can set the total number of potential genetic changes/perturbations to 29 (e.g. 29 possible SNPs, or 29 different promoters, or any combination thereof as long as the universe of genetic perturbations is 29) and then decide to design all possible 3-member combinations of the 29 potential genetic changes, which will result in 3,654 candidate strain designs.

[0612] To provide context to the aforementioned 3,654 candidate strains, consider that one can calculate the number of non-redundant groupings of size r from n possible members using $n! / ((n - r)! * r!)$. If $r = 3$, $n = 29$ gives 3,654. Thus, if one designs all possible 3-member combinations of 29 potential changes the results is 3,654 candidate strains. The 29 potential genetic changes are present in the x-axis of Figure 25.

Predict Performance of New Strain Designs

[0613] Using the linear regression constructed above with the combinatorial configurations as input, one can then predict the expected relative performance of each candidate design. Figure 25 summarizes the composition of changes for the top 100 predicted strain designs for *Corynebacterium*. The x-axis lists the pool of potential genetic changes (29 possible genetic changes), and the y-axis shows the rank order. Black cells indicate the presence of a particular change in the candidate design, while white cells indicate the absence of that change. In this particular example, all of the top 100 designs contain the changes *pcg3121_pgi*, *pcg1860_pyc*, *dss_339*, and *pcg0007_39_lysa*. Additionally, the top candidate design contains the changes *dss_034*, *dss_009*.

[0614] Predictive accuracy should increase over time as new observations are used to iteratively retrain and refit the model. Results from a study by the inventors illustrate the methods by which the predictive model can be iteratively retrained and improved. Figure 46 compares model

predictions with observed measurement values. The quality of model predictions can be assessed through several methods, including a correlation coefficient indicating the strength of association between the predicted and observed values, or the root-mean-square error, which is a measure of the average model error. Using a chosen metric for model evaluation, the system may define rules for when the model should be retrained.

[0615] A couple of unstated assumptions to the above model include: (1) there are no epistatic interactions; and (2) the genetic changes/perturbations utilized to build the predictive model (e.g. from built strain data as illustrated in Figure 24, or whatever data set is used as the reference to construct the model) were all made in the same *Corynebacterium* background, as the proposed combinations of genetic changes (e.g. as illustrated in Figure 25).

Filtering for Second-order Features

[0616] The above illustrative example focused on linear regression predictions based on predicted host cell performance. In some embodiments, the present linear regression methods can also be applied to non-biomolecule factors, such as saturation biomass, resistance, or other measurable host cell features. Thus the methods of the present disclosure also teach in considering other features outside of predicted performance when prioritizing the candidates to build. Assuming there is additional relevant data, nonlinear terms are also included in the regression model.

Closeness with Existing Strains

[0617] Predicted strains that are similar to ones that have already been built could result in time and cost savings despite not being a top predicted candidate

Diversity of Changes

[0618] When constructing the aforementioned models, one cannot be certain that genetic changes will truly be additive (as assumed by linear regression and mentioned as an assumption above) due to the presence of epistatic interactions. Therefore, knowledge of genetic change dissimilarity can be used to increase the likelihood of positive additivity. If one knows, for example, that the changes dss_034 and dss_009 (which are SNP swaps) from the top ranked strain above are on the same metabolic pathway and have similar performance characteristics, then that information could be used to select another top ranking strain with a dissimilar composition of changes. As described in the section above concerning epistasis mapping, the predicted best genetic changes may be filtered to restrict selection to mutations with sufficiently dissimilar response profiles. Alternatively, the

linear regression may be a weighted least squares regression using the similarity matrix to weight predictions.

Diversity of Predicted Performance

[0619] Finally, one may choose to design strains with middling or poor predicted performance, in order to validate and subsequently improve the predictive models.

Iterative strain design optimization

[0620] As described for the example above, all of the top 100 strain designs contain the changes pcg3121_pgi, pcg1860_pyc, dss_339, and pcg0007_39_lysa. Additionally, the top candidate strain design contains the changes dss_034, dss_009.

[0621] In embodiments, the order placement engine 208 places a factory order to the factory 210 to manufacture microbial strains incorporating the top candidate mutations. In feedback-loop fashion, the results may be analyzed by the analysis equipment 214 to determine which microbes exhibit desired phenotypic properties (314). During the analysis phase, the modified strain cultures are evaluated to determine their performance, i.e., their expression of desired phenotypic properties, including the ability to be produced at industrial scale. For example, the analysis phase uses, among other things, image data of plates to measure microbial colony growth as an indicator of colony health. The analysis equipment 214 is used to correlate genetic changes with phenotypic performance, and save the resulting genotype-phenotype correlation data in libraries, which may be stored in library 206, to inform future microbial production.

[0622] In particular, the candidate changes that actually result in sufficiently high measured performance may be added as rows in the database to tables such as Table 4 above. In this manner, the best performing mutations are added to the predictive strain design model in a supervised machine learning fashion.

[0623] LIMS iterates the design/build/test/analyze cycle based on the correlations developed from previous factory runs. During a subsequent cycle, the analysis equipment 214 alone, or in conjunction with human operators, may select the best candidates as base strains for input back into input interface 202, using the correlation data to fine tune genetic modifications to achieve better phenotypic performance with finer granularity. In this manner, the laboratory information management system of embodiments of the disclosure implements a quality improvement feedback loop.

[0624] In sum, with reference to the flowchart of Figure 33 the iterative predictive strain design workflow may be described as follows:

- Generate a training set of input and output variables, e.g., genetic changes as inputs and performance features as outputs (3302). Generation may be performed by the analysis equipment 214 based upon previous genetic changes and the corresponding measured performance of the microbial strains incorporating those genetic changes.
- Develop an initial model (e.g., linear regression model) based upon training set (3304). This may be performed by the analysis equipment 214.
- Generate design candidate strains (3306)
 - In one embodiment, the analysis equipment 214 may fix the number of genetic changes to be made to a background strain, in the form of combinations of changes. To represent these changes, the analysis equipment 214 may provide to the interpreter 204 one or more DNA specification expressions representing those combinations of changes. (These genetic changes or the microbial strains incorporating those changes may be referred to as “test inputs.”) The interpreter 204 interprets the one or more DNA specifications, and the execution engine 207 executes the DNA specifications to populate the DNA specification with resolved outputs representing the individual candidate design strains for those changes.
- Based upon the model, the analysis equipment 214 predicts expected performance of each candidate design strain (3308).
- The analysis equipment 214 selects a limited number of candidate designs, e.g., 100, with highest predicted performance (3310).
 - As described elsewhere herein with respect to epistasis mapping, the analysis equipment 214 may account for second-order effects such as epistasis, by, e.g., filtering top designs for epistatic effects, or factoring epistasis into the predictive model.
- Build the filtered candidate strains (at the factory 210) based on the factory order generated by the order placement engine 208 (3312).
- The analysis equipment 214 measures the actual performance of the selected strains, selects a limited number of those selected strains based upon their superior actual performance (3314),

and adds the design changes and their resulting performance to the predictive model (3316). In the linear regression example, add the sets of design changes and their associated performance as new rows in Table 4.

- The analysis equipment 214 then iterates back to generation of new design candidate strains (3306), and continues iterating until a stop condition is satisfied. The stop condition may comprise, for example, the measured performance of at least one microbial strain satisfying a performance metric, such as yield, growth rate, or titer.

[0625] In the example above, the iterative optimization of strain design employs feedback and linear regression to implement machine learning. In general, machine learning may be described as the optimization of performance criteria, e.g., parameters, techniques or other features, in the performance of an informational task (such as classification or regression) using a limited number of examples of labeled data, and then performing the same task on unknown data. In supervised machine learning such as that of the linear regression example above, the machine (e.g., a computing device) learns, for example, by identifying patterns, categories, statistical relationships, or other attributes, exhibited by training data. The result of the learning is then used to predict whether new data will exhibit the same patterns, categories, statistical relationships or other attributes.

[0626] Embodiments of the disclosure may employ other supervised machine learning techniques when training data is available. In the absence of training data, embodiments may employ unsupervised machine learning. Alternatively, embodiments may employ semi-supervised machine learning, using a small amount of labeled data and a large amount of unlabeled data. Embodiments may also employ feature selection to select the subset of the most relevant features to optimize performance of the machine learning model. Depending upon the type of machine learning approach selected, as alternatives or in addition to linear regression, embodiments may employ for example, logistic regression, neural networks, support vector machines (SVMs), decision trees, hidden Markov models, Bayesian networks, Gram Schmidt, reinforcement-based learning, cluster-based learning including hierarchical clustering, genetic algorithms, and any other suitable learning machines known in the art. In particular, embodiments may employ logistic regression to provide probabilities of classification (e.g., classification of genes into different functional groups) along with the classifications themselves. See, e.g., Shevade, A simple and

efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics*, Vol. 19, No. 17 2003, pp. 2246-2253, Leng, et al., Classification using functional data analysis for temporal gene expression data, *Bioinformatics*, Vol. 22, No. 1, Oxford University Press (2006), pp. 68-76, all of which are incorporated by reference in their entirety herein.

[0627] Embodiments may employ graphics processing unit (GPU) accelerated architectures that have found increasing popularity in performing machine learning tasks, particularly in the form known as deep neural networks (DNN). Embodiments of the disclosure may employ GPU-based machine learning, such as that described in GPU-Based Deep Learning Inference: A Performance and Power Analysis, NVidia Whitepaper, November 2015, Dahl, et al., Multi-task Neural Networks for QSAR Predictions, Dept. of Computer Science, Univ. of Toronto, June 2014 (arXiv:1406.1231 [stat.ML]), all of which are incorporated by reference in their entirety herein. Machine learning techniques applicable to embodiments of the disclosure may also be found in, among other references, Libbrecht, et al., Machine learning applications in genetics and genomics, *Nature Reviews: Genetics*, Vol. 16, June 2015, Kashyap, et al., Big Data Analytics in Bioinformatics: A Machine Learning Perspective, *Journal of Latex Class Files*, Vol. 13, No. 9, Sept. 2014, Prompramote, et al., Machine Learning in Bioinformatics, Chapter 5 of *Bioinformatics Technologies*, pp. 117-153, Springer Berlin Heidelberg 2005, all of which are incorporated by reference in their entirety herein.

Iterative Predictive Strain Design: Example

[0628] The following provides an example application of the iterative predictive strain design workflow outlined above.

[0629] An initial set of training inputs and output variables was prepared. This set comprised 1864 unique engineered strains with defined genetic composition. Each strain contained between 5 and 15 engineered changes. A total of 336 unique genetic changes were present in the training.

[0630] An initial predictive computer model was developed. The implementation used a generalized linear model (Kernel Ridge Regression with 4th order polynomial kernel). The implementation models two distinct phenotypes (*yield* and *productivity*). These phenotypes were combined as weighted sum to obtain a single score for ranking, as shown below. Various model parameters, e.g. regularization factor, were tuned via k-fold cross validation over the designated training data.

[0631] The implementation does not incorporate any explicit analysis of interaction effects as described in the Epistasis Mapping section above. However, as those skilled in the art would understand, the implemented generalized linear model may capture interaction effects implicitly through the second, third and fourth order terms of the kernel.

[0632] The model was trained against the training set. The fitted model has an R^2 value (*coefficient of determination*) of 0.52 with respect to *yield* and an R^2 value of 0.67 with respect to *productivity*. Figure 46 demonstrates a significant quality fitting of the *yield* model to the training data.

[0633] Candidate strains were generated. This example includes a serial build constraint associated with the introduction of new genetic changes to a parent strain (in this example, only one new mutation was engineered into a strain at a time). Here, candidates are not considered simply as a function of the desired number of changes. Instead, the analysis equipment 214 selected, as a starting point, a collection of previously designed strains known to have high performance metrics (“seed strains”). The analysis equipment 214 individually applied genetic changes to each of the seed strains. The introduced genetic changes did not include those already present in the seed strain. For various technical, biological or other reasons, certain mutations were explicitly required, e.g., *opca_4*, or explicitly excluded, e.g., *dss_422*. Using 166 available seed strains and the 336 changes characterized by the model, 6239 novel candidate strains were designed.

[0634] Based upon the model, the analysis equipment 214 predicted the performance of candidate strain designs. The analysis equipment 214 ranked candidates from “best” to “worst” based on predicted performance with respect to two phenotypes of interest (*yield* and *productivity*). Specifically, the analysis equipment 214 used a weighted sum to score a candidate strain:

[0635] $\text{Score} = 0.8 * \text{yield} / \text{max}(\text{yields}) + 0.2 * \text{prod} / \text{max}(\text{prods})$,

where *yield* represents predicted yield for the candidate strain,

max(yields) represents the maximum yield over all candidate strains,

prod represents productivity for the candidate strain, and

max(prods) represents the maximum yield over all candidate strains.

[0636] The analysis equipment 214 generated a final set of recommendations from the ranked list of candidates by imposing both capacity constraints and operational constraints. In this example, the capacity limit was set at 48 computer-generated candidate design strains. Due to operational

constraints, in this example only one seed strain was used per column of a 96-well plate. This means that after a seed strain was chosen, up to 8 changes to that strain could be built, but only 6 seed strains could be chosen in any given week.

[0637] The trained model (described above) was used to predict the expected performance (for yield and productivity) of each candidate strain. The analysis equipment 214 ranked the candidate strains using the scoring function given above. Capacity and operational constraints were applied to yield a filtered set of 48 candidate strains. This set of filtered candidate strains is depicted in Figure 47.

[0638] Filtered candidate strains were built (at the factory 210) based on a factory order generated by the order placement engine 208 (3312). The order was based upon DNA specifications corresponding to the candidate strains.

[0639] In practice, the build process has an expected failure rate whereby a random set of strains is not built. For this build cycle, roughly 20% of the candidate strains failed build, resulting in 37 built strains.

[0640] The analysis equipment 214 was used to measure the actual *yield* and *productivity* performance of the selected strains. The analysis equipment 214 evaluated the model and recommended strains based on three criteria: model accuracy; improvement in strain performance; and equivalence (or improvement) to human expert-generated designs.

[0641] The *yield* and *productivity* phenotypes were measured for recommended strains and compared to the values predicted by the model. As shown in Figure 48, the model demonstrates useful predictive utility. In particular, the predicted *yield* values for the recommended strains have a Pearson-r correlation coefficient of 0.59 with the corresponding observations.

[0642] Next, the analysis equipment 214 computed percentage performance change from the parent strain for each of the recommended strains. This data is shown in Figure 49 (in light gray). The inventors found that many of the predicted strains in fact exhibited the expected performance gains with respect to their immediate parents. In particular, the best predicted strain showed a 6% improvement in yield with respect to its immediate parent.

[0643] In parallel with the model-based strain design process described above, a collection of 48 strains was independently designed by a human expert. Of these strains, 37 were successfully built and tested. This data demonstrated that the model-based strain designs performed comparably to strains designed by human experts. These experts are highly-skilled (e.g., Ph.D.-level) scientists

employed or otherwise engaged by the assignee of the present invention, and familiar with the embodiments of this disclosure. To compare the two methods, the inventors first inspected the performance distributions of each group (Figure 50). In this experiment, the mean yield of model-based strains showed a 1% increase with respect to human expert generated designs.

[0644] The inventors then compared human expert-designed and computer-model-designed strains grouped by background, *i.e.*, new strains with the same parent (Figure 51). Again, the inventors found that computer-generated designs perform comparably to, and in some cases better than, the human expert-generated designs, and further tend to produce less variability. Finally, the inventors compared the percentage change with respect to the parent strains of the human expert and model-designed strains (Figure 49). Again, these populations showed comparable gains.

[0645] See Table 4.1 for tabulated summary statistics.

Table 4.1. Measured performance statistics for strains designed by the predictive model and by a human expert reference.

		Yield [AU]	Yield change from parent [%]	Productivity [AU]	Productivity change from parent [%]
design method					
computer model	count	37	37	37	37
	mean	1.058068108	0.3578340	0.737928919	-2.5428848
	std	0.017811031	1.8293665	0.083619804	9.6743873
	min	1.015310000	-4.5346677	0.572780000	-23.3626353
	median	1.058710000	0.005007939	0.766870000	-1.1824159

	max	1.093510000	6.0097309	0.872790000	26.6124119
Human expert	count	37	37	37	37
	mean	1.038804595	-0.0005237	0.748320811	-1.6126436
	std	0.032053625	1.9227716	0.120527468	9.8530758
	min	0.964910000	-3.1043233	0.535980000	-21.4589256
	median	1.045530000	0.0449168	0.760300000	-1.9241048
	max	1.094790000	7.8487174	0.984110000	21.7335193

[0646] At the conclusion of each round of the prediction → build → test cycle, the inventors were interested in evaluating the quality of the model predictions and iteratively incorporating new data into the previous model. For the former—model evaluation—the inventors focused on measuring predictive accuracy by comparing model predictions with experimental measurements. Predictive accuracy can be assessed through several methods, including a correlation coefficient indicating the strength of association between the predicted and observed values, or the root-mean-square error, which is a measure of the average model error.

[0647] Over many rounds of experimentation, model predictions may drift, and new genetic changes may be added to the training inputs to improve predictive accuracy. For this example, design changes and their resulting performance were added to the predictive model (3316).

Genomic design and engineering as a service

[0648] In embodiments of the disclosure, the LIMS system software 3210 of Figure 32 may be implemented in a cloud computing system 3202 of Figure 32, to enable multiple users to design and build microbial strains according to embodiments of the present disclosure. Figure 32 illustrates a cloud computing environment 3204 according to embodiments of the present

disclosure. Client computers 3206, such as those illustrated in Figure 32, access the LIMS system via a network 3208, such as the Internet. In embodiments, the LIMS system application software 3210 resides in the cloud computing system 3202. The LIMS system may employ one or more computing systems using one or more processors, of the type illustrated in Figure 32. The cloud computing system itself includes a network interface 3212 to interface the LIMS system applications 3210 to the client computers 3206 via the network 3208. The network interface 3212 may include an application programming interface (API) to enable client applications at the client computers 3206 to access the LIMS system software 3210. In particular, through the API, client computers 3206 may access components of the LIMS system 200, including without limitation the software running the input interface 202, the interpreter 204, the execution engine 207, the order placement engine 208, the factory 210, as well as test equipment 212 and analysis equipment 214. A software as a service (SaaS) software module 3214 offers the LIMS system software 3210 as a service to the client computers 3206. A cloud management module 3216 manages access to the LIMS system 3210 by the client computers 3206. The cloud management module 3216 may enable a cloud architecture that employs multitenant applications, virtualization or other architectures known in the art to serve multiple users.

Genomic Automation

[0649] Automation of the methods of the present disclosure enables high-throughput phenotypic screening and identification of target products from multiple test strain variants simultaneously.

[0650] The aforementioned genomic engineering predictive modeling platform is premised upon the fact that hundreds and thousands of mutant strains are constructed in a high-throughput fashion. The robotic and computer systems described below are the structural mechanisms by which such a high-throughput process can be carried out.

[0651] In some embodiments, the present disclosure teaches methods of improving host cell productivities, or rehabilitating industrial strains. As part of this process, the present disclosure teaches methods of assembling DNA, building new strains, screening cultures in plates, and screening cultures in models for tank fermentation. In some embodiments, the present disclosure teaches that one or more of the aforementioned methods of creating and testing new host strains is aided by automated robotics.

[0652] In some embodiments, the present disclosure teaches a high-throughput strain engineering platform as depicted in Figure 6A-B or Figure 26.

HTP Robotic Systems

[0653] In some embodiments, the automated methods of the disclosure comprise a robotic system. The systems outlined herein are generally directed to the use of 96- or 384-well microtiter plates, but as will be appreciated by those in the art, any number of different plates or configurations may be used. In addition, any or all of the steps outlined herein may be automated; thus, for example, the systems may be completely or partially automated.

[0654] In some embodiments, the automated systems of the present disclosure comprise one or more work modules. For example, in some embodiments, the automated system of the present disclosure comprises a DNA synthesis module, a vector cloning module, a strain transformation module, a screening module, and a sequencing module (see Figure 7).

[0655] As will be appreciated by those in the art, an automated system can include a wide variety of components, including, but not limited to: liquid handlers; one or more robotic arms; plate handlers for the positioning of microplates; plate sealers, plate piercers, automated lid handlers to remove and replace lids for wells on non-cross contamination plates; disposable tip assemblies for sample distribution with disposable tips; washable tip assemblies for sample distribution; 96 well loading blocks; integrated thermal cyclers; cooled reagent racks; microtiter plate pipette positions (optionally cooled); stacking towers for plates and tips; magnetic bead processing stations; filtrations systems; plate shakers; barcode readers and applicators; and computer systems.

[0656] In some embodiments, the robotic systems of the present disclosure include automated liquid and particle handling enabling high-throughput pipetting to perform all the steps in the process of gene targeting and recombination applications. This includes liquid and particle manipulations such as aspiration, dispensing, mixing, diluting, washing, accurate volumetric transfers; retrieving and discarding of pipette tips; and repetitive pipetting of identical volumes for multiple deliveries from a single sample aspiration. These manipulations are cross-contamination-free liquid, particle, cell, and organism transfers. The instruments perform automated replication of microplate samples to filters, membranes, and/or daughter plates, high-density transfers, full-plate serial dilutions, and high capacity operation.

[0657] In some embodiments, the customized automated liquid handling system of the disclosure is a TECAN machine (*e.g.* a customized TECAN Freedom Evo).

[0658] In some embodiments, the automated systems of the present disclosure are compatible with platforms for multi-well plates, deep-well plates, square well plates, reagent troughs, test tubes,

mini tubes, microfuge tubes, cryovials, filters, micro array chips, optic fibers, beads, agarose and acrylamide gels, and other solid-phase matrices or platforms are accommodated on an upgradeable modular deck. In some embodiments, the automated systems of the present disclosure contain at least one modular deck for multi-position work surfaces for placing source and output samples, reagents, sample and reagent dilution, assay plates, sample and reagent reservoirs, pipette tips, and an active tip-washing station.

[0659] In some embodiments, the automated systems of the present disclosure include high-throughput electroporation systems. In some embodiments, the high-throughput electroporation systems are capable of transforming cells in 96 or 384- well plates. In some embodiments, the high-throughput electroporation systems include VWR® High-throughput Electroporation Systems, BTX™, Bio-Rad® Gene Pulser MXcell™ or other multi-well electroporation system.

[0660] In some embodiments, the integrated thermal cycler and/or thermal regulators are used for stabilizing the temperature of heat exchangers such as controlled blocks or platforms to provide accurate temperature control of incubating samples from 0°C to 100°C.

[0661] In some embodiments, the automated systems of the present disclosure are compatible with interchangeable machine-heads (single or multi-channel) with single or multiple magnetic probes, affinity probes, replicators or pipettors, capable of robotically manipulating liquid, particles, cells, and multi-cellular organisms. Multi-well or multi-tube magnetic separators and filtration stations manipulate liquid, particles, cells, and organisms in single or multiple sample formats.

[0662] In some embodiments, the automated systems of the present disclosure are compatible with camera vision and/or spectrometer systems. Thus, in some embodiments, the automated systems of the present disclosure are capable of detecting and logging color and absorption changes in ongoing cellular cultures.

[0663] In some embodiments, the automated system of the present disclosure is designed to be flexible and adaptable with multiple hardware add-ons to allow the system to carry out multiple applications. The software program modules allow creation, modification, and running of methods. The system's diagnostic modules allow setup, instrument alignment, and motor operations. The customized tools, labware, and liquid and particle transfer patterns allow different applications to be programmed and performed. The database allows method and parameter storage. Robotic and computer interfaces allow communication between instruments.

[0664] Thus, in some embodiments, the present disclosure teaches a high-throughput strain engineering platform, as depicted in Figure 26.

[0665] Persons having skill in the art will recognize the various robotic platforms capable of carrying out the HTP engineering methods of the present disclosure. Table 5 below provides a non-exclusive list of scientific equipment capable of carrying out each step of the HTP engineering steps of the present disclosure as described in Figure 26.

Table 5 - Non-exclusive list of Scientific Equipment Compatible with the HTP engineering methods of the present disclosure.

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
Acquire and build DNA pieces	liquid handlers	Hitpicking (combining by transferring) primers/templates for PCR amplification of DNA parts	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Thermal cyclers	PCR amplification of DNA parts	Inheco Cyclor, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents
QC DNA parts	Fragment analyzers (capillary electrophoresis)	gel electrophoresis to confirm PCR products of appropriate size	Agilent Bioanalyzer, AATI Fragment Analyzer, or equivalents
	Sequencer (sanger: Beckman)	Verifying sequence of parts/templates	Beckman Ceq-8000, Beckman GenomeLab™, or equivalents

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	NGS (next generation sequencing) instrument	Verifying sequence of parts/templates	Illumina MiSeq series sequences, illumina Hi-Seq, Ion torrent, pac bio or other equivalents
	nanodrop/plate reader	assessing concentration of DNA samples	Molecular Devices SpectraMax M5, Tecan M1000, or equivalents.
Generate DNA assembly	liquid handlers	Hitpicking (combining by transferring) DNA parts for assembly along with cloning vector, addition of reagents for assembly reaction/process	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
QC DNA assembly	Colony pickers	for inoculating colonies in liquid media	Scirobotics Pickolo, Molecular Devices QPix 420
	liquid handlers	Hitpicking primers/templates, diluting samples	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Fragment analyzers (capillary electrophoresis)	gel electrophoresis to confirm assembled products of appropriate size	Agilent Bioanalyzer, AATI Fragment Analyzer

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	Sequencer (sanger: Beckman)	Verifying sequence of assembled plasmids	ABI3730 Thermo Fisher, Beckman Ceq-8000, Beckman GenomeLab™, or equivalents
	NGS (next generation sequencing) instrument	Verifying sequence of assembled plasmids	Illumina MiSeq series sequences, illumina Hi-Seq, Ion torrent, pac bio or other equivalents
Prepare base strain and DNA assembly	centrifuge	spinning / pelleting cells	Beckman Avanti floor centrifuge, Hettich Centrifuge
Transform DNA into base strain	Electroporators	electroporative transformation of cells	BTX Gemini X2, BIO-RAD MicroPulser Electroporator
	Ballistic transformation	ballistic transformation of cells	BIO-RAD PDS1000
	Incubators, thermal cyclers	for chemical transformation/heat shock	Inheco Cyclor, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	Liquid handlers	for combining DNA, cells, buffer	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
Integrate DNA into genome of base strain	Colony pickers	for inoculating colonies in liquid media	Scirobotics Pickolo, Molecular Devices QPix 420
	Liquid handlers	For transferring cells onto Agar, transferring from culture plates to different culture plates (inoculation into other selective media)	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Platform shaker-incubators	incubation with shaking of microtiter plate cultures	Kuhner Shaker ISF4-X, Inforsht Multitron Pro
QC transformed strain	Colony pickers	for inoculating colonies in liquid media	Scirobotics Pickolo, Molecular Devices QPix 420
	liquid handlers	Hitpicking primers/templates, diluting samples	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	Thermal cyclers	cPCR verification of strains	Inheco Cyclers, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents
	Fragment analyzers (capillary electrophoresis)	gel electrophoresis to confirm cPCR products of appropriate size	Infors-ht Multitron Pro, Kuhner Shaker ISF4-X
	Sequencer (sanger: Beckman)	Sequence verification of introduced modification	Beckman Ceq-8000, Beckman GenomeLab™, or equivalents
	NGS (next generation sequencing) instrument	Sequence verification of introduced modification	Illumina MiSeq series sequences, illumina Hi-Seq, Ion torrent, pac bio or other equivalents
Select and consolidate QC'd strains into test plate	Liquid handlers	For transferring from culture plates to different culture plates (inoculation into production media)	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Colony pickers	for inoculating colonies in liquid media	Scirobotics Pickolo, Molecular Devices QPix 420
	Platform shaker-incubators	incubation with shaking of microtiter plate cultures	Kuhner Shaker ISF4-X, Infors-ht Multitron Pro

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
Culture strains in seed plates	Liquid handlers	For transferring from culture plates to different culture plates (inoculation into production media)	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Platform shaker-incubators	incubation with shaking of microtiter plate cultures	Kuhner Shaker ISF4-X, Infors-ht Multitron Pro
	liquid dispensers	Dispense liquid culture media into microtiter plates	Well mate (Thermo), Benchcel2R (velocity 11), plateloc (velocity 11)
	microplate labeler	apply barcoders to plates	Microplate labeler (a2+ cab - agilent), benchcell 6R (velocity11)
Generate product from strain	Liquid handlers	For transferring from culture plates to different culture plates (inoculation into production media)	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	Platform shaker-incubators	incubation with shaking of microtiter plate cultures	Kuhner Shaker ISF4-X, Infors-ht Multitron Pro

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	liquid dispensers	Dispense liquid culture media into multiple microtiter plates and seal plates	well mate (Thermo), Benchcel2R (velocity 11), plateloc (velocity 11)
	microplate labeler	Apply barcodes to plates	microplate labeler (a2+ cab - agilent), benchcell 6R (velocity11)
Evaluate performance	Liquid handlers	For processing culture broth for downstream analytical	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	UHPLC, HPLC	quantitative analysis of precursor and target compounds	Agilent 1290 Series UHPLC and 1200 Series HPLC with UV and RI detectors, or equivalent; also any LC/MS
	LC/MS	highly specific analysis of precursor and target compounds as well as side and degradation products	Agilent 6490 QQQ and 6550 QTOF coupled to 1290 Series UHPLC
	Spectrophotometer	Quantification of different compounds using spectrophotometer based assays	Tecan M1000, spectramax M5, Genesys 10S

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
Culture strains in flasks	Fermenters:	incubation with shaking	Sartorius, DASGIPs (Eppendorf), BIO-FLOs (Sartorius-stedim). Applikon
	Platform shakers		innova 4900, or any equivalent
Generate product from strain	Fermenters: DASGIPs (Eppendorf), BIO-FLOs (Sartorius-stedim)		
Evaluate performance	Liquid handlers	For transferring from culture plates to different culture plates (inoculation into production media)	Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents
	UHPLC, HPLC	quantitative analysis of precursor and target compounds	Agilent 1290 Series UHPLC and 1200 Series HPLC with UV and RI detectors, or equivalent; also any LC/MS
	LC/MS	highly specific analysis of precursor and target compounds as well as side and degradation products	Agilent 6490 QQQ and 6550 QTOF coupled to 1290 Series UHPLC

	Equipment Type	Operation(s) performed	Compatible Equipment Make/Model/Configuration
	Flow cytometer	Characterize strain performance (measure viability)	BD Accuri, Millipore Guava
	Spectrophotometer	Characterize strain performance (measure biomass)	Tecan M1000, Spectramax M5, or other equivalents

Computer System Hardware

[0666] Figure 34 illustrates an example of a computer system 800 that may be used to execute program code stored in a non-transitory computer readable medium (e.g., memory) in accordance with embodiments of the disclosure. The computer system includes an input/output subsystem 802, which may be used to interface with human users and/or other computer systems depending upon the application. The I/O subsystem 802 may include, e.g., a keyboard, mouse, graphical user interface, touchscreen, or other interfaces for input, and, e.g., an LED or other flat screen display, or other interfaces for output, including application program interfaces (APIs). Other elements of embodiments of the disclosure, such as the components of the LIMS system, may be implemented with a computer system like that of computer system 800.

[0667] Program code may be stored in non-transitory media such as persistent storage in secondary memory 810 or main memory 808 or both. Main memory 808 may include volatile memory such as random access memory (RAM) or non-volatile memory such as read only memory (ROM), as well as different levels of cache memory for faster access to instructions and data. Secondary memory may include persistent storage such as solid state drives, hard disk drives or optical disks. One or more processors 804 reads program code from one or more non-transitory media and executes the code to enable the computer system to accomplish the methods performed by the embodiments herein. Those skilled in the art will understand that the processor(s) may ingest source code, and interpret or compile the source code into machine code that is understandable at the hardware gate level of the processor(s) 804. The processor(s) 804 may include graphics

processing units (GPUs) for handling computationally intensive tasks. Particularly in machine learning, one or more CPUs 804 may offload the processing of large quantities of data to one or more GPUs 804.

[0668] The processor(s) 804 may communicate with external networks via one or more communications interfaces 807, such as a network interface card, WiFi transceiver, etc. A bus 805 communicatively couples the I/O subsystem 802, the processor(s) 804, peripheral devices 806, communications interfaces 807, memory 808, and persistent storage 810. Embodiments of the disclosure are not limited to this representative architecture. Alternative embodiments may employ different arrangements and types of components, e.g., separate buses for input-output components and memory subsystems.

[0669] Those skilled in the art will understand that some or all of the elements of embodiments of the disclosure, and their accompanying operations, may be implemented wholly or partially by one or more computer systems including one or more processors and one or more memory systems like those of computer system 800. In particular, the elements of the LIMS system 200 and any robotics and other automated systems or devices described herein may be computer-implemented. Some elements and functionality may be implemented locally and others may be implemented in a distributed fashion over a network through different servers, e.g., in client-server fashion, for example. In particular, server-side operations may be made available to multiple clients in a software as a service (SaaS) fashion, as shown in Figure 32.

[0670] The term component in this context refers broadly to software, hardware, or firmware (or any combination thereof) component. Components are typically functional components that can generate useful data or other output using specified input(s). A component may or may not be self-contained. An application program (also called an “application”) may include one or more components, or a component can include one or more application programs.

[0671] Some embodiments include some, all, or none of the components along with other modules or application components. Still yet, various embodiments may incorporate two or more of these components into a single module and/or associate a portion of the functionality of one or more of these components with a different component.

[0672] The term “memory” can be any device or mechanism used for storing information. In accordance with some embodiments of the present disclosure, memory is intended to encompass any type of, but is not limited to: volatile memory, nonvolatile memory, and dynamic memory.

For example, memory can be random access memory, memory storage devices, optical memory devices, magnetic media, floppy disks, magnetic tapes, hard drives, SIMMs, SDRAM, DIMMs, RDRAM, DDR RAM, SODIMMS, erasable programmable read-only memories (EPROMs), electrically erasable programmable read-only memories (EEPROMs), compact disks, DVDs, and/or the like. In accordance with some embodiments, memory may include one or more disk drives, flash drives, databases, local cache memories, processor cache memories, relational databases, flat databases, servers, cloud based platforms, and/or the like. In addition, those of ordinary skill in the art will appreciate many additional devices and techniques for storing information can be used as memory.

[0673] Memory may be used to store instructions for running one or more applications or modules on a processor. For example, memory could be used in some embodiments to house all or some of the instructions needed to execute the functionality of one or more of the modules and/or applications disclosed in this application.

HTP Microbial Strain Engineering Based Upon Genetic Design Predictions: An Example Workflow

[0674] In some embodiments, the present disclosure teaches the directed engineering of new host organisms based on the recommendations of the computational analysis systems of the present disclosure.

[0675] In some embodiments, the present disclosure is compatible with all genetic design and cloning methods. That is, in some embodiments, the present disclosure teaches the use of traditional cloning techniques such as polymerase chain reaction, restriction enzyme digestions, ligation, homologous recombination, RT PCR, and others generally known in the art and are disclosed in for example: Sambrook *et al.* (2001) *Molecular Cloning: A Laboratory Manual* (3rd ed., Cold Spring Harbor Laboratory Press, Plainview, New York), incorporated herein by reference.

[0676] In some embodiments, the cloned sequences can include possibilities from any of the HTP genetic design libraries taught herein, for example: promoters from a promoter swap library, SNPs from a SNP swap library, start or stop codons from a start/stop codon exchange library, terminators from a STOP swap library, protein solubility tags from a SOLUBILITY TAG swap library, protein degradation tags from a DEGRADATION TAG swap library or sequence optimizations from a sequence optimization library.

[0677] Further, the exact sequence combinations that should be included in a particular construct can be informed by the epistatic mapping function.

[0678] In other embodiments, the cloned sequences can also include sequences based on rational design (hypothesis-driven) and/or sequences based on other sources, such as scientific publications.

[0679] In some embodiments, the present disclosure teaches methods of directed engineering, including the steps of i) generating custom-made SNP-specific DNA, ii) assembling SNP-specific plasmids, iii) transforming target host cells with SNP-specific DNA, and iv) looping out any selection markers (*See* Figure 2).

[0680] Figure 6A depicts the general workflow of the strain engineering methods of the present disclosure, including acquiring and assembling DNA, assembling vectors, transforming host cells and removing selection markers.

Build Specific DNA Oligonucleotides

[0681] In some embodiments, the present disclosure teaches inserting and/or replacing and/or altering and/or deleting a DNA segment of the host cell organism. In some aspects, the methods taught herein involve building an oligonucleotide of interest (*i.e.* a target DNA segment), that will be incorporated into the genome of a host organism. In some embodiments, the target DNA segments of the present disclosure can be obtained *via* any method known in the art, including: copying or cutting from a known template, mutation, or DNA synthesis. In some embodiments, the present disclosure is compatible with commercially available gene synthesis products for producing target DNA sequences (*e.g.*, GeneArt™, GeneMaker™, GenScript™, Anagen™, Blue Heron™, Entelechon™, GeNOsys, Inc., or Qiagen™).

[0682] In some embodiments, the target DNA segment is designed to incorporate a SNP into a selected DNA region of the host organism (*e.g.*, adding a beneficial SNP). In other embodiments, the DNA segment is designed to remove a SNP from the DNA of the host organisms (*e.g.*, removing a detrimental or neutral SNP).

[0683] In some embodiments, the oligonucleotides used in the inventive methods can be synthesized using any of the methods of enzymatic or chemical synthesis known in the art. The oligonucleotides may be synthesized on solid supports such as controlled pore glass (CPG), polystyrene beads, or membranes composed of thermoplastic polymers that may contain CPG. Oligonucleotides can also be synthesized on arrays, on a parallel microscale using microfluidics

(Tian *et al.*, *Mol. BioSyst.*, 5, 714-722 (2009)), or known technologies that offer combinations of both (see Jacobsen *et al.*, U.S. Pat. App. No. 2011/0172127).

[0684] Synthesis on arrays or through microfluidics offers an advantage over conventional solid support synthesis by reducing costs through lower reagent use. The scale required for gene synthesis is low, so the scale of oligonucleotide product synthesized from arrays or through microfluidics is acceptable. However, the synthesized oligonucleotides are of lesser quality than when using solid support synthesis (See Tian *infra.*; see also Staehler *et al.*, U.S. Pat. App. No. 2010/0216648).

[0685] A great number of advances have been achieved in the traditional four-step phosphoramidite chemistry since it was first described in the 1980s (see for example, Sierzchala, *et al. J. Am. Chem. Soc.*, 125, 13427-13441 (2003) using peroxy anion deprotection; Hayakawa *et al.*, U.S. Pat. No. 6,040,439 for alternative protecting groups; Azhayev *et al.*, *Tetrahedron* 57, 4977-4986 (2001) for universal supports; Kozlov *et al.*, *Nucleosides, Nucleotides, and Nucleic Acids*, 24 (5-7), 1037-1041 (2005) for improved synthesis of longer oligonucleotides through the use of large-pore CPG; and Damha *et al.*, *NAR*, 18, 3813-3821 (1990) for improved derivatization).

[0686] Regardless of the type of synthesis, the resulting oligonucleotides may then form the smaller building blocks for longer oligonucleotides. In some embodiments, smaller oligonucleotides can be joined together using protocols known in the art, such as polymerase chain assembly (PCA), ligase chain reaction (LCR), and thermodynamically balanced inside-out synthesis (TBIO) (see Czar *et al.* *Trends in Biotechnology*, 27, 63-71 (2009)). In PCA, oligonucleotides spanning the entire length of the desired longer product are annealed and extended in multiple cycles (typically about 55 cycles) to eventually achieve full-length product. LCR uses ligase enzyme to join two oligonucleotides that are both annealed to a third oligonucleotide. TBIO synthesis starts at the center of the desired product and is progressively extended in both directions by using overlapping oligonucleotides that are homologous to the forward strand at the 5' end of the gene and against the reverse strand at the 3' end of the gene.

[0687] Another method of synthesizing a larger double stranded DNA fragment is to combine smaller oligonucleotides through top-strand PCR (TSP). In this method, a plurality of oligonucleotides spans the entire length of a desired product and contain overlapping regions to the adjacent oligonucleotide(s). Amplification can be performed with universal forward and

reverse primers, and through multiple cycles of amplification a full-length double stranded DNA product is formed. This product can then undergo optional error correction and further amplification that results in the desired double stranded DNA fragment end product.

[0688] In one method of TSP, the set of smaller oligonucleotides that will be combined to form the full-length desired product are between 40-200 bases long and overlap each other by at least about 15-20 bases. For practical purposes, the overlap region should be at a minimum long enough to ensure specific annealing of oligonucleotides and have a high enough melting temperature (T_m) to anneal at the reaction temperature employed. The overlap can extend to the point where a given oligonucleotide is completely overlapped by adjacent oligonucleotides. The amount of overlap does not seem to have any effect on the quality of the final product. The first and last oligonucleotide building block in the assembly should contain binding sites for forward and reverse amplification primers. In one embodiment, the terminal end sequence of the first and last oligonucleotide contain the same sequence of complementarity to allow for the use of universal primers.

Assembling/Cloning Custom Plasmids

[0689] In some embodiments, the present disclosure teaches methods for constructing vectors capable of inserting desired target DNA sections (*e.g.* containing a particular SNP) into the genome of host organisms. In some embodiments, the present disclosure teaches methods of cloning vectors comprising the target DNA, homology arms, and at least one selection marker (*see* Figure 3).

[0690] In some embodiments, the present disclosure is compatible with any vector suited for transformation into the host organism. In some embodiments, the present disclosure teaches use of shuttle vectors compatible with a host cell. In one embodiment, a shuttle vector for use in the methods provided herein is a shuttle vector compatible with an *E. coli* and/or *Corynebacterium* host cell. Shuttle vectors for use in the methods provided herein can comprise markers for selection and/or counter-selection as described herein. The markers can be any markers known in the art and/or provided herein. The shuttle vectors can further comprise any regulatory sequence(s) and/or sequences useful in the assembly of said shuttle vectors as known in the art. The shuttle vectors can further comprise any origins of replication that may be needed for propagation in a host cell as provided herein such as, for example, *E. coli* or *C. glutamicum*. The regulatory sequence can be any regulatory sequence known in the art or provided herein such as, for example, a promoter,

start, stop, insulator, signal, secretion and/or termination sequence used by the genetic machinery of the host cell. In certain instances, the target DNA can be inserted into vectors, constructs or plasmids obtainable from any repository or catalogue product, such as a commercial vector (*see e.g.*, DNA2.0 custom or GATEWAY® vectors). In certain instances, the target DNA can be inserted into vectors, constructs or plasmids obtainable from any repository or catalogue product, such as a commercial vector (*see e.g.*, DNA2.0 custom or GATEWAY® vectors).

[0691] In some embodiments, the assembly/cloning methods of the present disclosure may employ at least one of the following assembly strategies: **i)** type II conventional cloning, **ii)** type II S-mediated or “Golden Gate” cloning (*see, e.g.*, Engler, C., R. Kandzia, and S. Marillonnet. 2008 “A one pot, one step, precision cloning method with high-throughput capability”. PLoS One 3:e3647; Kotera, I., and T. Nagai. 2008 “A high-throughput and single-tube recombination of crude PCR products using a DNA polymerase inhibitor and type IIS restriction enzyme.” J Biotechnol 137:1-7.; Weber, E., R. Gruetzner, S. Werner, C. Engler, and S. Marillonnet. 2011 Assembly of Designer TAL Effectors by Golden Gate Cloning. PloS One 6:e19722), **iii)** GATEWAY® recombination, **iv)** TOPO® cloning, exonuclease-mediated assembly (Aslanidis and de Jong 1990. “Ligation-independent cloning of PCR products (LIC-PCR).” Nucleic Acids Research, Vol. 18, No. 20 6069), **v)** homologous recombination, **vi)** non-homologous end joining, **vii)** Gibson assembly (Gibson et al., 2009 “Enzymatic assembly of DNA molecules up to several hundred kilobases” Nature Methods 6, 343-345) or a combination thereof. Modular type IIS based assembly strategies are disclosed in PCT Publication WO 2011/154147, the disclosure of which is incorporated herein by reference.

[0692] In some embodiments, the present disclosure teaches cloning vectors with at least one selection marker. Various selection marker genes are known in the art often encoding antibiotic resistance function for selection in prokaryotic (*e.g.*, against ampicillin, kanamycin, tetracycline, chloramphenicol, zeocin, spectinomycin/streptomycin) or eukaryotic cells (*e.g.* geneticin, neomycin, hygromycin, puromycin, blasticidin, zeocin) under selective pressure. Other marker systems allow for screening and identification of wanted or unwanted cells such as the well-known blue/white screening system used in bacteria to select positive clones in the presence of X-gal or fluorescent reporters such as green or red fluorescent proteins expressed in successfully transduced host cells. Another class of selection markers most of which are only functional in prokaryotic systems relates to counter selectable marker genes often also referred to as “death genes” which

express toxic gene products that kill producer cells. Examples of such genes include *sacB*, *rpsL(strA)*, *tetAR*, *pheS*, *thyA*, *gata-1*, or *ccdB*, the function of which is described in (Reyrat *et al.* 1998 “Counterselectable Markers: Untapped Tools for Bacterial Genetics and Pathogenesis.” *Infect Immun.* 66(9): 4011-4017).

DNA Vector Assembly, Amplification, and Genome Editing

[0693] In some embodiments, the present disclosure teaches HTP genomic engineering steps specific for *E. coli*. In some embodiments, the present disclosure thus teaches methods of constructing and amplifying constructs in *E. coli*, as well as methods of engineering *E. coli*.

[0694] In some embodiments, the DNA vectors of the present disclosure comprise **i)** a conditional replication ori (R6K), **ii)** an antibiotic resistance gene, **iii)** one or more counter-selection gene(s) (e.g. *sacB* and/or *PheS*) and **iv)** a replication ori for *S. cerevisiae*.

[0695] In some embodiments, the present disclosure teaches methods of assembling DNA constructs in auxotrophic *S. cerevisiae*. Thus in, some embodiments, the vectors of the present disclosure comprise a replication origin for *S. cerevisiae*. This permits the vector to replicate in *S. cerevisiae* during assembly.

[0696] In some embodiments, the present disclosure teaches methods of propagating assembled DNA occurs in *E. coli* containing *pir* protein. Thus, in some embodiments, the vectors of the present disclosure comprise an R6K origin of replication. In some embodiments, the R6K replication origin is conditional on the presence of the *pir* protein. That is, in some embodiments, the presently disclosed vectors comprising the R6K replication origin will only be amplified in host cells comprising the *pir* gene. This allows researchers to amplify the vectors of the present disclosure during the vector construction and amplification steps, while also preventing extra chromosomal expression of the vectors during the host cell engineering steps.

[0697] In some embodiments, the vectors of the present disclosure comprise a *PheS* gene. *Escherichia coli* phenylalanyl-tRNA synthetase (*PheS*) can be useful as a counterselection marker, since its A294G variant misincorporates 4-chloro-phenylalanine (4CP) into cellular proteins during translation, thereby causing cell death. In some embodiments, the *PheS* gene is designed to temporarily incorporate into the genome of the host cell. In some embodiments, the present disclosure teaches counterselection methods comprising growing the host cells in minimal media with 4CP. Cells which still comprise the vector will incorporate 4CP into proteins and die. Cells which have looped out the *PheS* sequence survive.

[0698] In some embodiments, the present disclosure teaches methods of constructing, assembling and integrating vectors into host cells. In some embodiments, the present disclosure teaches that the homology arms (homL and homR) are amplified by PCR. In some embodiments, the desired genetic change (black bar between homL and homR) are present in the reverse primer of homL and the forward primer of homR (*see* Figure 3). In some embodiments, the forward primer of homL and the reverse primer of homR have sequence homology to the backbone plasmid. Further illustrations of the loop-in and loop-out process are depicted in Figure 45.

[0699] In some embodiments, the vectors of the present disclosure comprise one or more insulator sequences. The insulator sequence can be any insulator sequence known in the art. In one embodiment, the insulator nucleic acid sequence is the insulator1 sequence (SEQ ID NO. 218), insulator 2 sequence (SEQ ID NO. 219) or both the insulator 1 and 2 sequences provided herein. In one embodiment, the vectors of the present disclosure comprise insulator sequences flanking homology arms (homL and homR). In one embodiment, the vectors of the present disclosure comprise insulator sequences flanking homology arms (homL and homR) and terminator sequence(s). The insulator sequences can be generated to be free of restriction endonuclease sequences.

[0700] In some embodiments, the vectors of the present disclosure comprise a combination of elements provided herein. In some cases, a vector for use in the methods provided herein can comprise an R6K origin of replication, a SacB gene, a PheS gene as a counterselection marker and a URA3 yeast auxotrophic marker such as, for example, vector 1 (*see* Figure 55), which has the nucleic acid sequence of SEQ ID NO. 214. In some cases, a vector for use in the methods provided herein can be an altered version of vector 1, such as, for example, vector 2 (*see* Figure 56), which has the nucleic acid sequence of SEQ ID NO. 215. In addition to the elements previously recited for vector 1, vector 2 can further comprise the elements found in Table 15. Additional vectors for use in the methods provided herein include vector 3 (nucleic acid SEQ ID NO. 216; Figure 57) and vector 4 (nucleic acid SEQ ID NO. 217; Figure 58). Both vectors 3 and 4 were built off of the vector 2 background. However, in vector 3, the promoter sequence of sacB was replaced with a promoter containing the P2-MCD2 promoter (*see* Mutalik et al, Nat Methods. 2013 Apr;10(4):354-60) and a codon optimized version of the native pheS gene containing the T251A/A294G mutations (*see* Miyazaki, K. Biotechniques. 2015 Feb 1;58(2):86-8), while vector 4 comprises the vector 3 background with the URA3 selection marker being replaced with the TRP1 marker.

Table 15. Select Sequence Elements of Vector 2

Element name	Part sequence
Insulator1	TATTCACACGCAATCAACAGGCAGGATAATCGCTGGTAAGGTCAGTGC TTTCTTCAGGTAGTAGAGATAACAATAGTTCCCAACGATAGGTGGCAGA TTTCACTTTACAGACCGACTGGTTCAGAAGCGTAGATAATAGCCCGTGT TTTCCAATAAGGGATAGTGTAGGTAAGTCAACTCCTCCGTCAGAGCCA ACCGTTT (SEQ ID NO. 218)
Insulator2	GCACTAGGACTTGCCGCGGATACTGCCCCATTACATGAATTGCAGCCT CAGGGACGTCAGTAGATCATGGAGGTAGGGCATATGTCCTCTGTTGTT AAAATGTGAGTTCTCAACGAAGCACGAATCGGTCAGAACCTACACTAA GGAGATTTGGTAGGTGCACGGTTTCTGTTCGCATAGACCAGTTCATTTCA GATGTCT (SEQ ID NO. 219)
T1	GGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTT TTATCTGTTGTTTGTTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCC GCCCTAGA (SEQ ID NO. 220)
B0015	CCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTC GTTTTATCTGTTGTTTGTTCGGTGAACGCTCTCTACTAGAGTCACACTGG CTCACCTTCGGGTGGGCCTTCTGCGTTTATA (SEQ ID NO. 221)
SacB promoter	TTGACAATTAATCATCCGGCTCGTAATTTATGTGGATCTTAATCATGCT AAGGAGGTTTTCTAATG (SEQ ID NO. 222)
PheS promoter & sequence	GGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGC AGGTCACACAGGAAAGTACTAGATGTCGCATCTTGCAGAATTAGTAGC TTCAGCGAAGGCCGCGATTTCTCAGGCGAGTGACGTCGCAGCACTGGA TAATGTACGTGTTGAGTACCTGGGAAAGAAGGGACACCTTACTCTTCA AATGACAACCCTGCGCGAACTGCCGCCGGAGGAACGCCCCGCAGCAG GAGCGGTAATCAATGAGGCAAAGGAGCAAGTACAACAGGCACTGAAC GCCCGTAAGGCTGAGTTGGAATCCGCCGCATTAACGCGCGCCTTGCT GCGGAAACCATTGATGTCTCGCTGCCCGGGCGCCGCATTGAGAATGGA GGCTTACACCCAGTACTCGTACCATCGACCGTATCGAATCTTTCTTTG GCGAACTTGGCTTCACTGTGGCAACTGGACCGGAGATTGAGGACGACT ACCACAATTTTCGATGCCTTGAACATTCCCGGTCATCATCCTGCACGCGC CGATCATGATACATTCTGGTTTGATACCACCCGTTTGCTTCGTACCCAG ACAAGCGGTGTCCAAATCCGTACGATGAAGGCTCAGCAACCACCGATC CGTATCATTGCTCCAGGGCGCGTGTACCGTAACGATTATGACCAGACA CATACACCGATGTTTCACCAAATGGAAGGGTTGATTGTGGATACGAAT ATCTCTTTCACGAATCTGAAGGGCACCTTACATGATTTCTTACGCAACT TTTTTCGAGGAGGACCTTCAAATTCGCTTTCGTCCATCGTACTTCCCTTTT GCAGAACCTTCGGCTGAAGTGGATGTAATGGGGAAAAACGGTAAGTG GCTGGAGGTTTTAGGTTGCGGGATGGTTCATCCAAATGTGCTTCGCAAC GTCGGCATCGACCCCGAAGTCTACAGTGGATTTCGGATTTCGGGATGGGA ATGGAACGTCTGACTATGCTTTCGTTACGGCGTAACGGATTTGCGCTCCT TTTTTGAGAACGATCTTCGTTTTCTGAAGCAATTCAAATAA (SEQ ID NO. 223)

[0701] A general workflow associated with the DNA assembly according to one embodiment of the present disclosure is shown in Figure 35.

[0702] In some embodiments, DNA assembly among the backbone and the inserts (homL and homR) is performed in yeast via yeast gap repair recombination. In some embodiments, auxotrophic marker (TRP or URA) is present in the backbone plasmid for the selection of assembled DNA in minimal media. The assembled plasmids are then extracted out of yeast culture.

[0703] In some embodiments, the extracted plasmids are next transformed into an *E. coli* strain containing pir gene to propagate the desired plasmids for the following transformation of the strain of interest. The transformants are selected for resistance to the given antibiotics. The transformants will be picked for sequencing to select the correctly assembled plasmid of interest.

[0704] In some embodiments correctly assembled plasmids are transformed into host cells for genome engineering via electroporation. Since the target host cells do not contain pir protein, the colonies formed on the antibiotic selective media are expected to have integration of the plasmid at the desired locus in the genome. The correct integration of the plasmid can be verified by PCR with the primer outside the homL and homR, respectively, and the primer binding inside the plasmid.

[0705] In some embodiments, the present disclosure teaches loop-out methods of removing the backbone of the plasmid from the genome of the host. Thus, in some embodiments, the present disclosure teaches that subsequent selection on counter-selective media (sucrose) and/or 4CP can be used to isolate the clones that do not contain the backbone part of the plasmid. Thus, in some embodiments, isolated host cells comprising the correct integrant are inoculated in LB media and the culture is plated on LB agar plate containing sucrose and 4-p-chloro-phenylalanine (LB+suc+4CP). Due to the sensitivity of cells expressing sacB gene to sucrose, and the sensitivity of cells expressing PheS to 4CP, the colonies formed onto LB+suc+4CP agar plate are expected to have either the mutant or the wild-type of the gene of interest. PCR amplification of the target nucleotide(s) and sequencing of the PCR product allows us to isolate the new clones with the desired genome modification.

[0706] In some embodiments, the resulting clones are sequenced to find the clones with the desired nucleotide change(s). In some embodiments, all the process above can be performed with the liquid handler.

Transformation of Host Cells

[0707] In some embodiments, the vectors of the present disclosure may be introduced into the host cells using any of a variety of techniques, including transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer (see Christie, P.J., and Gordon, J.E., 2014 “The Agrobacterium Ti Plasmids” *Microbiol Spectr.* 2014; 2(6); 10.1128). Particular methods include calcium phosphate transfection, DEAE-Dextran mediated transfection, lipofection, or electroporation (Davis, L., Dibner, M., Battey, I., 1986 “Basic Methods in Molecular Biology”). Other methods of transformation include for example, lithium acetate transformation and electroporation See, e.g., Gietz et al., *Nucleic Acids Res.* 27:69-74 (1992); Ito et al., *J. Bacteriol.* 153:163-168 (1983); and Becker and Guarente, *Methods in Enzymology* 194:182-187 (1991). In some embodiments, transformed host cells are referred to as recombinant host strains.

[0708] In some embodiments, the present disclosure teaches high-throughput transformation of cells using the 96-well plate robotics platform and liquid handling machines of the present disclosure.

[0709] In some embodiments, the present disclosure teaches screening transformed cells with one or more selection markers as described above. In one such embodiment, cells transformed with a vector comprising a kanamycin resistance marker (KanR) are plated on media containing effective amounts of the kanamycin antibiotic. Colony forming units visible on kanamycin-laced media are presumed to have incorporated the vector cassette into their genome. Insertion of the desired sequences can be confirmed via PCR, restriction enzyme analysis, and/or sequencing of the relevant insertion site.

Looping Out of Selected Sequences

[0710] In some embodiments, the present disclosure teaches methods of looping out selected regions of DNA from the host organisms. The looping out method can be as described in Nakashima et al. 2014 “Bacterial Cellular Engineering by Genome Editing and Gene Silencing.” *Int. J. Mol. Sci.* 15(2), 2773-2793. In some embodiments, the present disclosure teaches looping out selection markers from positive transformants. Looping out deletion techniques are known in the art, and are described in (Tear et al. 2014 “Excision of Unstable Artificial Gene-Specific inverted Repeats Mediates Scar-Free Gene Deletions in Escherichia coli.” *Appl. Biochem. Biotech.* 175:1858-1867). The looping out methods used in the methods provided herein can be performed using single-crossover homologous recombination or double-crossover homologous

recombination. In one embodiment, looping out of selected regions as described herein can entail using single-crossover homologous recombination as described herein.

[0711] First, loop out vectors are inserted into selected target regions within the genome of the host organism (*e.g.*, via homologous recombination, CRISPR, lambda red-mediated recombineering or other gene editing technique). In one embodiment, single-crossover homologous recombination is used between a circular plasmid or vector and the host cell genome in order to loop-in the circular plasmid or vector such as depicted in Figure 3. The inserted vector can be designed with a sequence which is a direct repeat of an existing or introduced nearby host sequence, such that the direct repeats flank the region of DNA slated for looping and deletion. Once inserted, cells containing the loop out plasmid or vector can be counter selected for deletion of the selection region (*e.g.*, see Figure 4; lack of resistance to the selection gene). Further illustrations of the loop-in and loop-out process are depicted in Figure 45.

[0712] Persons having skill in the art will recognize that the description of the loopout procedure represents but one illustrative method for deleting unwanted regions from a genome. Indeed the methods of the present disclosure are compatible with any method for genome deletions, including but not limited to gene editing via lambda red, CRISPR, TALENS, FOK, or other endonucleases. Persons skilled in the art will also recognize the ability to replace unwanted regions of the genome via homologous recombination techniques.

Lambda RED Mediated Gene Editing

[0713] As provided herein, gene editing as described herein can be performed using Lambda Red-mediated homologous recombination as described by Datsenko and Wanner, PNAS USA 97:6640-6645 (2000), the contents of which are hereby incorporated by reference in their entirety.

[0714] The lambda red system is derived from the lambda red bacteriophage and its use as a genetic engineering tool can be called recombineering - short for homologous recombination-mediated genetic engineering. It can be used to make an assortment of modifications: insertion and deletion of selectable and non-selectable sequences, point mutations or other small base pair changes, and the addition of protein tags. It also has the flexibility to modify the *E. coli* chromosome, plasmid DNA or BAC DNA. To use the lambda red recombineering system to modify target DNA, a linear donor DNA substrate (either dsDNA or ssDNA – see below) can be electroporated into *E. coli* expressing the lambda red enzymes. These enzymes then catalyze the homologous recombination of the substrate with the target DNA sequence. This means cloning

occurs in vivo, as compared to restriction enzyme cloning where the genetic changes occur in a test tube. The donor DNA substrate only requires ~50 nucleotides of homology to the target site for recombination.

[0715] The lambda red recombineering system has three components: 1) Exo, 2) Beta, and 3) Gam. All three are required for recombineering with a dsDNA substrate; however, only Beta is required when generating a modification with an ssDNA substrate.

[0716] Gam: Gam prevents both the endogenous RecBCD and SbcCD nucleases from digesting linear DNA introduced into a E. coli host cell.

[0717] Exo: Exo is a 5'→3' dsDNA-dependent exonuclease. Exo can degrade linear dsDNA starting from the 5' end and generate 2 possible products: 1) a partially dsDNA duplex with single-stranded 3' overhangs or 2) if the dsDNA was short enough, a ssDNA whose entire complementary strand was degraded.

[0718] Beta: Beta can protect the ssDNA created by Exo and promote its annealing to a complementary ssDNA target in the cell. Only Beta expression is required for recombineering with an ssDNA oligo substrate.

[0719] For use herein, a lambda red recombineering method can entail designing and generating substrate DNA; expressing lambda red recombination genes; transforming (e.g., electroporating) substrate DNA; growing transformants; and selecting and confirming recombinant clones.

Substrate DNA Design and Generation

[0720] Whether a linear dsDNA or ssDNA substrate is used can depend on the goal of the experiment. dsDNA substrate may be best for insertions or deletions greater than approximately 20 nucleotides, while ssDNA substrate may be best for point mutations or changes of only a few base pairs.

dsDNA Substrate

[0721] dsDNA inserts can be made by PCR using primers that amplify the DNA sequence of interest and flank it with 50 base pairs of homology to the targeted insert site. The primers can be ~70 nucleotides long (20 nucleotides that anneal to the DNA sequence of interest and 50 nucleotides of homology to regions flanking the target site). The dsDNA inserts can include: large insertions or deletions, including selectable DNA fragments, such as antibiotic resistance genes, as well as non-selectable DNA fragments, such as gene replacements and tags.

ssDNA Substrate

[0722] ssDNA substrates can be synthetic oligonucleotides or short PCR products. Either way, the substrate should be ~70-100 nucleotides long with the desired alteration(s) located in the center of the sequence. Since lambda red has a higher recombination frequency when the lagging strand of DNA is targeted, it's best to determine the direction of replication through a target region of interest and design an oligo that is complementary to the lagging strand. In some cases, oligos that target both strands are designed. One of the two oligos will recombine with a higher efficiency than the other which can aid in identifying the lagging strand.

[0723] ssDNA substrate can be more efficient than dsDNA with a recombination frequency between 0.1% to 1%, and can be increased to as high as 25-50% by designing oligos that avoid activating the methyl -directed mismatch repair (MMR) system. MMR's job is to correct DNA mismatches that occur during DNA replication. Activation of MMR can be avoided by: 1) using a strain of bacteria that has key MMR proteins knocked out or 2) specially design ssDNA oligos to avoid MMR: 1) *E. coli* with inactivated MMR: Using *E. coli* with inactive MMR is definitely the easier of the two options, but these cells are prone to mutations and can have more unintended changes to their genomes. 2) Designing ssDNA oligos that avoid MMR activation: In one embodiment, a C/C mismatch at or within 6 base pairs of the edit site is introduced. In another embodiment, the desired change is flanked with 4-5 silent changes in the wobble codons, i.e. make changes to the third base pair of the adjacent 4-5 codons that alter the nucleotide sequence but not the amino acid sequence of the translated protein. These changes can be 5' or 3' of the desired change.

Expression of Lambda Red Recombination Genes

[0724] The lambda red recombineering system can be expressed in a host cell by: 1) From a bacteria with integrated defective prophage 2) from a plasmid, 3) from mini- λ or 4) from the lambda red phage itself. Controlling expression of Red proteins is critical for minimizing the toxic effects of Gam expression and to limit spontaneous mutations that occur when Red is constitutively expressed. Which recombination system you use depends on what type of DNA you want to edit; however, BAC DNA can be modified with any of the below described approaches.

[0725] Bacterial Strain with Integrated Defective Prophage:

[0726] A number of *E. coli* strains exist that stably express lambda red recombineering genes due to integration of a defective lambda red phage. One such strain is DY380, which is derived from

the DH10B *E. coli* strain. Several other bacterial strains commonly used for recombineering can be found in Thomason et al (Recombineering: Genetic Engineering in Bacteria Using Homologous Recombination. *Current Protocols in Molecular Biology*. 106:V:1.16:1.16.1–1.16.39) and Sharan et al (Recombineering: A Homologous Recombination-Based Method of Genetic Engineering. *Nature protocols*. 2009;4(2):206-223).

[0727] In some of these strains, expression of *exo*, *beta* and *gam* is tightly regulated by the endogenous phage promoter pL and repressor CI. For recombineering purposes, a temperature sensitive version of the repressor gene, CI857, is used. This mutant repressor prevents expression of the recombination genes at low temperatures (30-34°C). Shifting bacteria to 42°C for 15 minutes quickly inactivates the repressor and allows for expression of the recombination genes. After this, lowering the temperature allows the repressor to re-nature and again repress expression of *exo*, *beta* and *gam*. A major advantage of using this method for lambda red expression is that it doesn't require antibiotic selection to maintain expression of the recombineering system. This setup can be also be used to modify chromosomal genes. After the initial editing event, the defective prophage can be removed from the chromosome of the host *E. coli* by a second lambda red recombination event. Alternatively, if the modified allele is selectable, it can be transferred to a different genetic background via P1 transduction.

[0728] Plasmid:

[0729] Expressing lambda red genes from a plasmid allows for a mobile recombineering system, but tight regulation of expression is required for a successful experiment. Promoters commonly used to control expression of Red include the IPTG-inducible lac promoter, the arabinose-inducible pBAD promoter and the endogenous phage pL promoter. Plasmids that also express the repressors associated with these promoters (*lacI*, *araC*, *ci857*) can be used in some cases in order to limit leaky expression of the Red system. Using a plasmid to express the lambda recombineering system can be used for editing bacterial chromosomal DNA because it is easy to remove the recombineering system once recombinant clones are generated. A simple way to do this is to express lambda red genes from a plasmid with a heat sensitive origin of replication. Once the recombineering system is no longer needed, bacteria can be "cured" by growing them at 42°C.

[0730] Mini-λ:

[0731] A hybrid between using a plasmid and stable integration of a defective prophage is to use mini-λ, a defective non-replicating, circular piece of phage DNA, that when introduced into

bacteria, integrates into the genome. Mini- λ uses the endogenous red promoter pL and the cI857 repressor to regulate expression. An antibiotic can be used to select for positive clones but, because mini- λ stably integrates, drug selection is not required for maintenance. A temperature shift to 42°C not only allows for activation of the red genes needed for recombineering, but also leads to expression of the int and xis genes which are responsible for excision of mini- λ from the host's chromosome. After this, mini- λ can readily be purified from bacteria just like a plasmid.

[0732] Phage:

[0733] Another option for expressing the Red system can be to use a lambda red phage, λ TetR, that carries the tetracycline resistance gene and the lambda red repressor cI857. Once introduced, the prophage is stable and no longer requires drug selection. One drawback to this approach is that it requires the generation of bacteriophage, which is not a common molecular biology technique. However, an advantage of this method is that you can stably integrate the Red system into yoa strain of interest and P1 transduction could be used to move the modification into a different background, if needed. This approach is best suited for modifying plasmids or BACs because it results in stable integration of the phage into the *E. coli* genome.

Selection and Confirmation of Recombinant Clones

[0734] If a antibiotic resistance gene has been inserted, recombinants can first be selected via antibiotic resistance, but all clones should be further tested to confirm the presence of the desired modification. Colony PCR can be used to screen for positive clones in most cases, and restriction enzyme digest can be used to screen plasmids for the appropriate mutations. Point mutations and other subtle changes can be confirmed by sequencing, which can also be used for confirmation for all clones, regardless of what type of DNA is being targeted for modification: the *E. coli* chromosome, a plasmid, or a BAC.

CRISPR Mediated Gene Editing

[0735] In one aspect provided herein, the genome of a host cell can be modified by CRISPR. An exemplary embodiment for utilizing the CRISPR/Cas9 system for gene editing in *E. coli* can be found in Examples 18 and 19.

[0736] The CRISPR/Cas system is a prokaryotic immune system that confers resistance to foreign genetic elements such as those present within plasmids and phages and that provides a form of acquired immunity. CRISPR stands for Clustered Regularly Interspaced Short Palindromic

Repeat, and cas stands for CRISPR-associated system, and refers to the small cas genes associated with the CRISPR complex.

[0737] CRISPR-Cas systems are most broadly characterized as either Class 1 or Class 2 systems. The main distinguishing feature between these two systems is the nature of the Cas-effector module. Class 1 systems require assembly of multiple Cas proteins in a complex (referred to as a “Cascade complex”) to mediate interference, while Class 2 systems use a large single Cas enzyme to mediate interference. Each of the Class 1 and Class 2 systems are further divided into multiple CRISPR-Cas types based on the presence of a specific Cas protein. For example, the Class 1 system is divided into the following three types: Type I systems, which contain the Cas3 protein; Type III systems, which contain the Cas10 protein; and the putative Type IV systems, which contain the Csf1 protein, a Cas8-like protein. Class 2 systems are generally less common than Class 1 systems and are further divided into the following three types: Type II systems, which contain the Cas9 protein; Type V systems, which contain Cas12a protein (previously known as Cpf1, and referred to as Cpf1 herein), Cas12b (previously known as C2c1), Cas12c (previously known as C2c3), Cas12d (previously known as CasY), and Cas12e (previously known as CasX); and Type VI systems, which contain Cas13a (previously known as C2c2), Cas13b, and Cas13c. Pyzocha *et al.*, ACS Chemical Biology, Vol. 13 (2), pgs. 347-356. In one embodiment, the CRISPR-Cas system for use in the methods provided herein is a Class 2 system. In one embodiment, the CRISPR-Cas system for use in the methods provided herein is a Type II, Type V or Type VI Class 2 system. In one embodiment, the CRISPR-Cas system for use in the methods provided herein is selected from Cas9, Cas12a, Cas12b, Cas12c, Cas12d, Cas12e, Cas13a, Cas13b, Cas13c or homologs, orthologs or paralogs thereof

[0738] CRISPR systems used in methods disclosed herein comprise a Cas effector module comprising one or more nucleic acid guided CRISPR-associated (Cas) nucleases, referred to herein as Cas effector proteins. In some embodiments, the Cas proteins can comprise one or multiple nuclease domains. A Cas effector protein can target single stranded or double stranded nucleic acid molecules (*e.g.* DNA or RNA nucleic acids) and can generate double strand or single strand breaks. In some embodiments, the Cas effector proteins are wild-type or naturally occurring Cas proteins. In some embodiments, the Cas effector proteins are mutant Cas proteins, wherein one or more mutations, insertions, or deletions are made in a WT or naturally occurring Cas protein (*e.g.*,

a parental Cas protein) to produce a Cas protein with one or more altered characteristics compared to the parental Cas protein.

[0739] In some instances, the Cas protein is a wild-type (WT) nuclease. Non-limiting examples of suitable Cas proteins for use in the present disclosure include C2c1, C2c2, C2c3, Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Cpf1, Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm1, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx100, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4, MAD1-20, SmCsm1, homologues thereof, orthologues thereof, variants thereof, mutants thereof, or modified versions thereof. Suitable nucleic acid guided nucleases (e.g., Cas 9) can be from an organism from a genus, which includes but is not limited to: *Thiomicrospira*, *Succinivibrio*, *Candidatus*, *Porphyromonas*, *Acidomonococcus*, *Prevotella*, *Smithella*, *Moraxella*, *Synergistes*, *Francisella*, *Leptospira*, *Catenibacterium*, *Kandleria*, *Clostridium*, *Dorea*, *Coprococcus*, *Enterococcus*, *Fructobacillus*, *Weissella*, *Pediococcus*, *Corynebacter*, *Sutterella*, *Legionella*, *Treponema*, *Roseburia*, *Filifactor*, *Eubacterium*, *Streptococcus*, *Lactobacillus*, *Mycoplasma*, *Bacteroides*, *Flaviivola*, *Flavobacterium*, *Sphaerochaeta*, *Azospirillum*, *Gluconacetobacter*, *Neisseria*, *Roseburia*, *Parvibaculum*, *Staphylococcus*, *Nitratifactor*, *Mycoplasma*, *Alicyclobacillus*, *Brevibacillus*, *Bacillus*, *Bacteroidetes*, *Brevibacillus*, *Carnobacterium*, *Clostridiaridium*, *Clostridium*, *Desulfonatrum*, *Desulfovibrio*, *Helcococcus*, *Leptotrichia*, *Listeria*, *Methanomethyophilus*, *Methylobacterium*, *Opitutaceae*, *Paludibacter*, *Rhodobacter*, *Sphaerochaeta*, *Tuberibacillus*, and *Campylobacter*. Species of organism of such a genus can be as otherwise herein discussed.

[0740] Suitable nucleic acid guided nucleases (e.g., Cas9) can be from an organism from a phylum, which includes but is not limited to: Firmicute, Actinobacteria, Bacteroidetes, Proteobacteria, Spirochates, and Tenericutes. Suitable nucleic acid guided nucleases can be from an organism from a class, which includes but is not limited to: Erysipelotrichia, Clostridia, Bacilli, Actinobacteria, Bacteroidetes, Flavobacteria, Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, Spirochaetes, and Mollicutes. Suitable nucleic acid guided nucleases can be from an organism from an order, which includes but is not limited to: Clostridiales, Lactobacillales, Actinomycetales, Bacteroidales, Flavobacteriales, Rhizobiales, Rhodospirillales, Burkholderiales, Neisseriales, Legionellales, Nautiliales, Campylobacterales, Spirochaetales, Mycoplasmatales, and Thiotrichales. Suitable nucleic acid

guided nucleases can be from an organism from within a family, which includes but is not limited to: Lachnospiraceae, Enterococcaceae, Leuconostocaceae, Lactobacillaceae, Streptococcaceae, Peptostreptococcaceae, Staphylococcaceae, Eubacteriaceae, Corynebacterineae, Bacteroidaceae, Flavobacterium, Cryomorphaceae, Rhodobiaceae, Rhodospirillaceae, Acetobacteraceae, Sutterellaceae, Neisseriaceae, Legionellaceae, Nautiliaceae, Campylobacteraceae, Spirochaetaceae, Mycoplasmataceae, and Francisellaceae.

[0741] Other nucleic acid guided nucleases (*e.g.*, Cas9) suitable for use in the methods, systems, and compositions of the present disclosure include those derived from an organism such as, but not limited to: *Thiomicrospira sp.* XS5, *Eubacterium rectale*, *Succinivibrio dextrinosolvens*, *Candidatus Methanoplasma termitum*, *Candidatus Methanomethylophilus alvus*, *Porphyromonas crevioricanis*, *Flavobacterium branchiophilum*, *Acidomonococcus sp.*, *Lachnospiraceae bacterium* COE1, *Prevotella brevis* ATCC 19188, *Smithella sp.* SCADC, *Moraxella bovoculi*, *Synergistes jonesii*, *Bacteroidetes* oral taxon 274, *Francisella tularensis*, *Leptospira inadai* serovar Lyme str. 10, *Acidomonococcus sp.* crystal structure (5B43) *S. mutans*, *S. agalactiae*, *S. equisimilis*, *S. sanguinis*, *S. pneumoniae*; *C. jejuni*, *C. coli*; *N. salsuginis*, *N. tergaricus*; *S. auricularis*, *S. carnosus*; *N. meningitidis*, *N. gonorrhoeae*; *L. monocytogenes*, *L. ivanovii*; *C. botulinum*, *C. difficile*, *C. tetani*, *C. sordellii*, *Francisella tularensis* I, *Prevotella albensis*, *Lachnospiraceae bacterium* MC2017 1, *Butyrivibrio proteoclasticus*, *Peregrinibacteria bacterium* GW2011_GWA2_33_10, *Parcubacteria bacterium* GW2011_GWC2_44_17, *Smithella sp.* SCADC, *Microgenomates*, *Acidaminococcus sp.* BV3L6, *Lachnospiraceae bacterium* MA2020, *Candidatus Methanoplasma termitum*, *Eubacterium eligens*, *Moraxella bovoculi* 237, *Leptospira inadai*, *Lachnospiraceae bacterium* ND2006, *Porphyromonas crevioricanis* 3, *Prevotella disiens*, *Porphyromonas macacae*, *Catenibacterium sp.* CAG:290, *Kandleria vitulina*, *Clostridiales bacterium* KA00274, *Lachnospiraceae bacterium* 3-2, *Dorea longicatena*, *Coprococcus catus* GD/7, *Enterococcus columbae* DSM 7374, *Fructobacillus sp.* EFB-N1, *Weissella halotolerans*, *Pediococcus acidilactici*, *Lactobacillus curvatus*, *Streptococcus pyogenes*, *Lactobacillus versmoldensis*, and *Filifactor alocis* ATCC 35896. See, U.S. Pat. Nos. 8,697,359; 8,771,945; 8,795,965; 8,865,406; 8,871,445; 8,889,356; 8,895,308; 8,906,616; 8,932,814; 8,945,839; 8,993,233; 8,999,641; 9,822,372; 9,840,713; U.S. Pat. App. No. 13/842,859 (US 2014/0068797 A1); 9,260,723; 9,023,649; 9,834,791; 9,637,739; U.S. Pat. App. No. 14/683,443 (US 2015/0240261 A1); U.S. Pat. App. No. 14/743,764 (US 2015/0291961 A1);

9,790,490; 9,688,972; 9,580,701; 9,745,562; 9,816,081; 9,677,090; 9,738,687; U.S. App. No. 15/632,222 (US 2017/0369879 A1); U.S. App. No. 15/631,989; U.S. App. No. 15/632,001; and U.S. Pat. No. 9,896,696, each of which is herein incorporated by reference.

[0742] In some embodiments, a Cas effector protein comprises one or more of the following activities:

[0743] a nickase activity, i.e., the ability to cleave a single strand of a nucleic acid molecule;

[0744] a double stranded nuclease activity, i.e., the ability to cleave both strands of a double stranded nucleic acid and create a double stranded break;

[0745] an endonuclease activity;

[0746] an exonuclease activity; and/or

[0747] a helicase activity, i.e., the ability to unwind the helical structure of a double stranded nucleic acid.

[0748] In aspects of the disclosure the term “guide nucleic acid” refers to a polynucleotide comprising 1) a guide sequence capable of hybridizing to a target sequence (referred to herein as a “targeting segment”) and 2) a scaffold sequence capable of interacting with (either alone or in combination with a tracrRNA molecule) a nucleic acid guided nuclease as described herein (referred to herein as a “scaffold segment”). A guide nucleic acid can be DNA. A guide nucleic acid can be RNA. A guide nucleic acid can comprise both DNA and RNA. A guide nucleic acid can comprise modified non-naturally occurring nucleotides. In cases where the guide nucleic acid comprises RNA, the RNA guide nucleic acid can be encoded by a DNA sequence on a polynucleotide molecule such as a plasmid, linear construct, or editing cassette as disclosed herein.

[0749] In some embodiments, the guide nucleic acids described herein are RNA guide nucleic acids (“guide RNAs” or “gRNAs”) and comprise a targeting segment and a scaffold segment. In some embodiments, the scaffold segment of a gRNA is comprised in one RNA molecule and the targeting segment is comprised in another separate RNA molecule. Such embodiments are referred to herein as “double-molecule gRNAs” or “two-molecule gRNA” or “dual gRNAs.” In some embodiments, the gRNA is a single RNA molecule and is referred to herein as a “single-guide RNA” or an “sgRNA.” The term “guide RNA” or “gRNA” is inclusive, referring both to two-molecule guide RNAs and sgRNAs.

[0750] The DNA-targeting segment of a gRNA comprises a nucleotide sequence that is complementary to a sequence in a target nucleic acid sequence. As such, the targeting segment of

a gRNA interacts with a target nucleic acid in a sequence-specific manner via hybridization (i.e., base pairing), and the nucleotide sequence of the targeting segment determines the location within the target DNA that the gRNA will bind. The degree of complementarity between a guide sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is about or more than about 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or more. Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences. In some embodiments, a guide sequence is about or more than about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, 75, or more nucleotides in length. In some embodiments, a guide sequence is less than about 75, 50, 45, 40, 35, 30, 25, 20 nucleotides in length. In aspects, the guide sequence is 10-30 nucleotides long. The guide sequence can be 15-20 nucleotides in length. The guide sequence can be 15 nucleotides in length. The guide sequence can be 16 nucleotides in length. The guide sequence can be 17 nucleotides in length. The guide sequence can be 18 nucleotides in length. The guide sequence can be 19 nucleotides in length. The guide sequence can be 20 nucleotides in length.

[0751] The scaffold segment of a guide RNA interacts with a one or more Cas effector proteins to form a ribonucleoprotein complex (referred to herein as a CRISPR-RNP or a RNP-complex). The guide RNA directs the bound polypeptide to a specific nucleotide sequence within a target nucleic acid sequence via the above-described targeting segment. The scaffold segment of a guide RNA comprises two stretches of nucleotides that are complementary to one another and which form a double stranded RNA duplex. Sufficient sequence within the scaffold sequence to promote formation of a targetable nuclease complex may include a degree of complementarity along the length of two sequence regions within the scaffold sequence, such as one or two sequence regions involved in forming a secondary structure. In some cases, the one or two sequence regions are comprised or encoded on the same polynucleotide. In some cases, the one or two sequence regions are comprised or encoded on separate polynucleotides. Optimal alignment may be determined by any suitable alignment algorithm, and may further account for secondary structures, such as self-complementarity within either the one or two sequence regions. In some embodiments, the degree of complementarity between the one or two sequence regions along the length of the shorter of the two when optimally aligned is about or more than about 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97.5%, 99%, or higher. In some embodiments, at least one of the two sequence regions

is about or more than about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 40, 50, or more nucleotides in length.

[0752] A scaffold sequence of a subject gRNA can comprise a secondary structure. A secondary structure can comprise a pseudoknot region or stem-loop structure. In some examples, the compatibility of a guide nucleic acid and nucleic acid guided nuclease is at least partially determined by sequence within or adjacent to the secondary structure region of the guide RNA. In some cases, binding kinetics of a guide nucleic acid to a nucleic acid guided nuclease is determined in part by secondary structures within the scaffold sequence. In some cases, binding kinetics of a guide nucleic acid to a nucleic acid guided nuclease is determined in part by nucleic acid sequence with the scaffold sequence.

[0753] A compatible scaffold sequence for a gRNA-Cas effector protein combination can be found by scanning sequences adjacent to a native Cas nuclease loci. In other words, native Cas nucleases can be encoded on a genome within proximity to a corresponding compatible guide nucleic acid or scaffold sequence.

[0754] Nucleic acid guided nucleases can be compatible with guide nucleic acids that are not found within the nucleases endogenous host. Such orthogonal guide nucleic acids can be determined by empirical testing. Orthogonal guide nucleic acids can come from different bacterial species or be synthetic or otherwise engineered to be non-naturally occurring. Orthogonal guide nucleic acids that are compatible with a common nucleic acid-guided nuclease can comprise one or more common features. Common features can include sequence outside a pseudoknot region. Common features can include a pseudoknot region. Common features can include a primary sequence or secondary structure

[0755] A guide nucleic acid can be engineered to target a desired target sequence by altering the guide sequence such that the guide sequence is complementary to the target sequence, thereby allowing hybridization between the guide sequence and the target sequence. A guide nucleic acid with an engineered guide sequence can be referred to as an engineered guide nucleic acid. Engineered guide nucleic acids are often non-naturally occurring and are not found in nature.

[0756] In some embodiments, the present disclosure provides a polynucleotide encoding a gRNA. In some embodiments, a gRNA-encoding nucleic acid is comprised in an expression vector, *e.g.*, a recombinant expression vector. In some embodiments, the present disclosure provides a polynucleotide encoding a site-directed modifying polypeptide. In some embodiments, the

polynucleotide encoding a site-directed modifying polypeptide is comprised in an expression vector, *e.g.*, a recombinant expression vector.

EXAMPLES

[0757] The following examples are given for the purpose of illustrating various embodiments of the disclosure and are not meant to limit the present disclosure in any fashion. Changes therein and other uses which are encompassed within the spirit of the disclosure, as defined by the scope of the claims, will be recognized by those skilled in the art.

[0758] In particular, Examples 1-9 are demonstrations of the HTP genomic engineering platform in *Corynebacterium*. However, similar procedures have been customized for *E. coli* and are being successfully carried out by the inventors.

[0759] A brief table of contents is provided below solely for the purpose of assisting the reader. Nothing in this table of contents is meant to limit the scope of the examples or disclosure of the application.

Table 5.1 - Table of Contents For Example Section

Example #	Title	Brief Description
1	HTP Transformation of <i>Corynebacterium</i> & Demonstration of SNP Library Creation	Describes embodiments of the high-throughput genetic engineering methods of the present disclosure.
2	HTP Genomic Engineering – Implementation of a SNP Library to Rehabilitate/Improve an Industrial Microbial Strain	Describes approaches for rehabilitating industrial organisms through SNP swap methods of the present disclosure.
3	HTP Genomic Engineering – Implementation of a SNP Swap Library to Improve Strain Performance in Lysine Production in <i>Corynebacterium</i> .	Describes an implementation of SNP swap techniques for improving the performance of <i>Corynebacterium</i> strain producing lysine . Also discloses selected

Example #	Title	Brief Description
		second and third order mutation consolidations.
4	HTP Genomic Engineering – Implementation of a Promoter Swap Library to Improve an Industrial Microbial Strain	Describes methods for improving the strain performance of host organisms through PRO swap genetic design libraries of the present disclosure.
5	HTP Genomic Engineering – Implementation of a PRO Swap Library to Improve Strain Performance for Lysine Production	Describes an implementation of PRO swap techniques for improving the performance of <i>Corynebacterium</i> strain producing lysine.
6	Epistasis Mapping- An Algorithmic Tool for Predicting Beneficial Mutation Consolidations	Describes an embodiment of the automated tools/algorithms of the present disclosure for predicting beneficial gene mutation consolidations.
7	HTP Genomic Engineering – PRO Swap Mutation Consolidation and Multi-Factor Combinatorial Testing	Describes and illustrates the ability of the HTP methods of the present disclosure to effectively explore the large solution space created by the combinatorial consolidation of multiple gene/genetic design library combinations.

Example #	Title	Brief Description
8	HTP Genomic Engineering – Implementation of a Terminator Library to Improve an Industrial Host Strain	Describes and illustrates an application of the STOP swap genetic design libraries of the present disclosure.
9	Comparing HTP Toolsets vs. Traditional UV Mutations	Provides experimental results comparing the HTP genetic design methods of the present disclosure vs. traditional mutational strain improvement programs.
10	HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance <i>Escherichia coli</i>	Describes embodiments of using a transposon mutagenesis library with the high-throughput genetic engineering methods of the present disclosure, as applied to <i>E. coli</i> cells.
11	HTP Genomic Engineering – Generation of Vector Backbones for use in HTP Genomic Engineering in <i>Escherichia coli</i>	Describes embodiments of generating vector backbones for use in the high-throughput genetic engineering methods of the present disclosure, as applied to <i>E. coli</i> cells.
12	HTP Genomic Engineering – Generation and testing of an additional Promoter Swap Library for use in improving an Industrial Microbial Strain	Describes methods for designing, generating and testing PRO swap genetic design libraries of the present disclosure.

Example #	Title	Brief Description
13	HTP Genomic Engineering –Testing integration of Promoter Swap Library of Table 1.4 into the genome of <i>E. coli</i> using vector 2 backbone	Describes proof of concept of the use of the vector 2 backbone from Example 11 in combination with promoters from the promoter library of Table 1.4 to drive integration of a single copy of a heterologous promoter-gene construct into the genome of <i>E. coli</i>
14	HTP Genomic Engineering – Implementation of a PROSWP methods using promoter library derived from Table 1.4.	Describes methods for improving the strain performance of host organisms through PRO swap genetic design libraries derived from Table 1.4 of the present disclosure.
15	HTP Genomic Engineering – Implementation of a TERMINATOR Swap Library to Improve Strain Performance for Lycopene Production	Describes methods for improving the strain performance of host organisms through terminator swap genetic design libraries of the present disclosure.
16	HTP Genomic Engineering – Implementation of a TERMINATOR Swap Library or PRO Swap Library in combination with either a SOLUBILITY TAG swap Library or DEGRADATION Tag swap Library to Improve Strain Performance for Lycopene Production	Describes methods for improving the strain performance of host organisms through terminator swap, solubility tag swap and degradation tag swap genetic design libraries of the present disclosure.

Example 1: HTP Transformation of *Corynebacterium* & Demonstration of SNP Library Creation

[0760] This example illustrates embodiments of the HTP genetic engineering methods of the present disclosure. Host cells are transformed with a variety of SNP sequences of different sizes, all targeting different areas of the genome. The results demonstrate that the methods of the present disclosure are able to generate rapid genetic changes of any kind, across the entire genome of a host cell.

A. Cloning of Transformation Vectors

[0761] A variety of SNPs were chosen at random from *Corynebacterium glutamicum* (ATCC21300) and were cloned into *Corynebacterium* cloning vectors using yeast homologous recombination cloning techniques to assemble a vector in which each SNP was flanked by direct repeat regions, as described *supra* in the “Assembling/Cloning Custom Plasmids” section, and as illustrated in Figure 3.

[0762] The SNP cassettes for this example were designed to include a range of homology direct repeat arm lengths ranging from 0.5Kb, 1Kb, 2Kb, and 5Kb. Moreover, SNP cassettes were designed for homologous recombination targeted to various distinct regions of the genome, as described in more detail below.

[0763] The *C. glutamicum* genome is 3,282,708 bp in size (see Figure 9). The genome was arbitrarily divided into 24 equal-sized genetic regions, and SNP cassettes were designed to target each of the 24 regions. Thus, a total of 96 distinct plasmids were cloned for this Example (4 different insert sizes x 24 distinct genomic regions).

[0764] Each DNA insert was produced by PCR amplification of homologous regions using commercially sourced oligos and the host strain genomic DNA described above as template. The SNP to be introduced into the genome was encoded in the oligo tails. PCR fragments were assembled into the vector backbone using homologous recombination in yeast.

[0765] Cloning of each SNP and homology arm into the vector was conducted according to the HTP engineering workflow described in Figure 6A-B, Figure 3, and Table 5.

B. Transformation of Assembled Clones into *E. coli*

[0766] Vectors were initially transformed into *E. coli* using standard heat shock transformation techniques in order to identify correctly assembled clones, and to amplify vector DNA for *Corynebacterium* transformation.

[0767] For example, transformed *E. coli* bacteria were tested for assembly success. Four colonies from each *E. coli* transformation plate were cultured and tested for correct assembly via PCR. This process was repeated for each of the 24 transformation locations and for each of the 4 different insert sizes (i.e., for all 96 transformants of this example). Results from this experiment were represented as the number of correct colonies identified out of the four colonies that were tested for each treatment (insert size and genomic location) (see Figure 12). Longer 5kb inserts exhibited a decrease in assembly efficiency compared to shorter counterparts (n=96).

C. Transformation of Assembled Clones into *Corynebacterium*

[0768] Validated clones were transformed into *Corynebacterium glutamicum* host cells via electroporation. For each transformation, the number of Colony Forming Units (CFUs) per μg of DNA was determined as a function of the insert size (see Figure 13). *Coryne* genome integration was also analyzed as a function of homology arm length, and the results showed that shorter arms had a lower efficiency (see Figure 13).

[0769] Genomic integration efficiency was also analyzed with respect to the targeted genome location in *C. glutamicum* transformants. Genomic positions 1 and 2 exhibited slightly lowered integration efficiency compared to the rest of the genome (see Figure 10).

D. Looping Out Selection Markers

[0770] Cultures of *Corynebacterium* identified as having successful integrations of the insert cassette were cultured on media containing 5% sucrose to counter select for loop outs of the *sacb* selection gene. Sucrose resistance frequency for various homology direct repeat arms did not vary significantly with arm length (see Figure 14). These results suggested that loopout efficiencies remained steady across homology arm lengths of .5 kb to 5kb.

[0771] In order to further validate loop out events, colonies exhibiting sucrose resistance were cultured and analyzed via sequencing.

[0772] The results for the sequencing of the insert genomic regions are summarized in Table 6 below.

Table 6 – Loop-out Validation Frequency

Outcome	Frequency (sampling error 95% confidence)
Successful Loop out	13% (9%/20%)
Loop Still present	42% (34%/50%)
Mixed read	44% (36%/52%)

[0773] Sequencing results showed a 10-20% efficiency in loop outs. Actual loop-out probably is somewhat dependent on insert sequence. However, picking 10-20 sucrose-resistant colonies leads to high success rates.

E. Summary

[0774] Table 7 below provides a quantitative assessment of the efficiencies of the HTP genome engineering methods of the present invention. Construct assembly rates for yeast homology methodologies yielded expected DNA constructs in nearly 9 out of 10 tested colonies. *Coryne* transformations of SNP constructs with 2kb homology arms yielded an average of 51 colony forming units per micro gram of DNA (CFU/ μ g), with 98% of said colonies exhibiting correctly integrated SNP inserts (targeting efficiency). Loop out efficiencies remained at .2% of cells becoming resistant when exposed to sucrose, with 13% of these exhibiting correctly looped out sequences.

Table 7- Summary Results for *Corynebacterium glutamicum* Strain Engineering

QC Step	Results for 2 kb Homology Arms
Construct Assembly Success	87%
Coryne Transformation efficiency	51 CFU/ μ g DNA (+/- 15)
Targeting efficiency	98%

QC Step	Results for 2 kb Homology Arms
Loop out Efficiency	0.2% (+/- 0.03%)

Example 2: HTP Genomic Engineering – Implementation of a SNP Library to Rehabilitate/Improve an Industrial Microbial Strain

[0775] This example illustrates several aspects of the SNP swap libraries of the HTP strain improvement programs of the present disclosure. Specifically, the example illustrates several envisioned approaches for rehabilitating currently existing industrial strains. This example describes the wave up and wave down approaches to exploring the phenotypic solution space created by the multiple genetic differences that may be present between “base,” “intermediate,” and industrial strains.

A. Identification of SNPs in Diversity Pool

[0776] An exemplary strain improvement program using the methods of the present disclosure was conducted on an industrial production microbial strain, herein referred to as “C.” The diversity pool strains for this program are represented by A, B, and C. Strain A represented the original production host strain, prior to any mutagenesis. Strain C represented the current industrial strain, which has undergone many years of mutagenesis and selection *via* traditional strain improvement programs. Strain B represented a “middle ground” strain, which had undergone some mutagenesis, and had been the predecessor of strain C. (*see* Figure 17A).

[0777] Strains A, B, and C were sequenced and their genomes were analyzed for genetic differences between strains. A total of 332 non-synonymous SNPs were identified. Of these, 133 SNPs were unique to C, 153 were additionally shared by B and C, and 46 were unique to strain B (*see* Figure 17B). These SNPs will be used as the diversity pool for downstream strain improvement cycles.

B. SNP Swapping Analysis

[0778] SNPs identified from the diversity pool in Part A of Example 2 will be analyzed to determine their effect on host cell performance. The initial “learning” round of the strain performance will be broken down into six steps as described below, and diagramed in Figure 18.

[0779] *First*, all the SNPs from C will be individually and/or combinatorially cloned into the base A strain. This will represent a minimum of 286 individual transformants. The purpose of these transformants will be to identify beneficial SNPs.

[0780] *Second*, all the SNPs from C will be individually and/or combinatorially removed from the commercial strain C. This will represent a minimum of 286 individual transformants. The purpose of these transformants will be to identify neutral and detrimental SNPs. Additional optional steps 3-6 are also described below. The first and second steps of adding and subtracting SNPS from two genetic time points (base strain A, and industrial strain C) is herein referred to as “wave,” which comprises a “wave up” (addition of SNPs to a base strain, first step), and a “wave down” (removal of SNPs from the industrial strain, second step). The wave concept extends to further additions/subtractions of SNPS.

[0781] *Third*, all the SNPs from B will be individually and/or combinatorially cloned into the base A strain. This will represent a minimum of 199 individual transformants. The purpose of these transformants will be to identify beneficial SNPs. Several of the transformants will also serve as validation data for transformants produced in the first step.

[0782] *Fourth*, all the SNPs from B will be individually and/or combinatorially removed from the commercial strain B. This will represent a minimum of 199 individual transformants. The purpose of these transformants will be to identify neutral and detrimental SNPs. Several of the transformants will also serve as validation data for transformants produced in the second step.

[0783] *Fifth*, all the SNPs *unique* to C (i.e., not also present in B) will be individually and/or combinatorially cloned into the commercial B strain. This will represent a minimum of 46 individual transformants. The purpose of these transformants will be to identify beneficial SNPs. Several of the transformants will also serve as validation data for transformants produced in the first and third steps.

[0784] *Sixth*, all the SNPs *unique* to C will be individually and/or combinatorially removed from the commercial strain C. This will represent a minimum of 46 individual transformants. The purpose of these transformants will be to identify neutral and detrimental SNPs. Several of the

transformants will also serve as validation data for transformants produced in the second and fourth steps.

[0785] Data collected from each of these steps is used to classify each SNP as *prima facie* beneficial, neutral, or detrimental.

C. Utilization of Epistatic Mapping to Determine Beneficial SNP Combinations

[0786] Beneficial SNPs identified in Part B of Example 2 will be analyzed *via* the epistasis mapping methods of the present disclosure, in order to identify SNPs that are likely to improve host performance when combined.

[0787] New engineered strain variants will be created using the engineering methods of Example 1 to test SNP combinations according to epistasis mapping predictions. SNPs consolidation may take place sequentially, or may alternatively take place across multiple branches such that more than one improved strain may exist with a subset of beneficial SNPs. SNP consolidation will continue over multiple strain improvement rounds, until a final strain is produced containing the optimum combination of beneficial SNPs, without any of the neutral or detrimental SNP baggage

Example 3: HTP Genomic Engineering – Implementation of a SNP Swap Library to Improve Strain Performance in Lysine Production in *Corynebacterium*

[0788] This example provides an illustrative implementation of a portion of the SNP Swap HTP design strain improvement program of Example 2 with the goal of producing yield and productivity improvements of lysine production in *Corynebacterium*.

[0789] Section B of this example further illustrates the mutation consolidation steps of the HTP strain improvement program of the present disclosure. The example thus provides experimental results for a first, second, and third round consolidation of the HTP strain improvement methods of the present disclosure.

[0790] Mutations for the second and third round consolidations are derived from separate genetic library swaps. These results thus also illustrate the ability for the HTP strain programs to be carried out multi-branch parallel tracks, and the “memory” of beneficial mutations that can be embedded into meta data associated with the various forms of the genetic design libraries of the present disclosure.

[0791] As described above, the genomes of a provided base reference strain (Strain A), and a second “engineered” strain (Strain C) were sequenced, and all genetic differences were identified.

The base strain was a *Corynebacterium glutamicum* variant that had not undergone UV mutagenesis. The engineered strain was also a *C. glutamicum* strain that had been produced from the base strain after several rounds of traditional mutation improvement programs. This Example provides the SNP Swap results for 186 distinct non-synonymous SNP differences identified between strains A and C.

A. HTP engineering and High-Throughput Screening

[0792] Each of the 186 identified SNPs were individually added back into the base strain, according to the cloning and transformation methods of the present disclosure. Each newly created strain comprising a single SNP was tested for lysine yield in small scale cultures designed to assess product titer performance. Small scale cultures were conducted using media from industrial scale cultures. Product titer was optically measured at carbon exhaustion (*i.e.*, representative of single batch yield) with a standard colorimetric assay. Briefly, a concentrated assay mixture was prepared and was added to fermentation samples such that final concentrations of reagents were 160 mM sodium phosphate buffer, 0.2 mM Amplex Red, 0.2 U/mL Horseradish Peroxidase and 0.005 U/mL of lysine oxidase. Reactions were allowed to proceed to an end point and optical density measured using a Tecan M1000 plate spectrophotometer at a 560nm wavelength. The results of the experiment are summarized in Table 8 below, and depicted in Figure 38.

Table 8 - Summary Results for SNP Swap Strain Engineering for Lysine Production

SNP	N	Mean Lysine Yield (change in A ₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_033	4	0.1062	0.00888	11.54348	2.895652
DSS_311	2	0.03603	0.01256	3.916304	4.095652
DSS_350	1	0.03178	0.01777	3.454348	5.794565
DSS_056	3	0.02684	0.01026	2.917391	3.345652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_014	4	0.02666	0.00888	2.897826	2.895652
DSS_338	3	0.02631	0.01026	2.859783	3.345652
DSS_128	1	0.02584	0.01777	2.808696	5.794565
DSS_038	4	0.02467	0.00888	2.681522	2.895652
DSS_066	4	0.02276	0.00888	2.473913	2.895652
DSS_108	2	0.02216	0.01256	2.408696	4.095652
DSS_078	4	0.02169	0.00888	2.357609	2.895652
DSS_017	3	0.02102	0.01026	2.284783	3.345652
DSS_120	3	0.01996	0.01026	2.169565	3.345652
DSS_064	4	0.01889	0.00888	2.053261	2.895652
DSS_380	4	0.01888	0.00888	2.052174	2.895652
DSS_105	3	0.0184	0.01026	2	3.345652
DSS_407	1	0.01831	0.01777	1.990217	5.794565
DSS_018	2	0.01825	0.01256	1.983696	4.095652
DSS_408	3	0.01792	0.01026	1.947826	3.345652
DSS_417	3	0.01725	0.01026	1.875	3.345652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_130	3	0.01724	0.01026	1.873913	3.345652
DSS_113	4	0.0172	0.00888	1.869565	2.895652
DSS_355	3	0.01713	0.01026	1.861957	3.345652
DSS_121	3	0.01635	0.01026	1.777174	3.345652
DSS_097	2	0.0162	0.01256	1.76087	4.095652
DSS_107	3	0.01604	0.01026	1.743478	3.345652
DSS_110	2	0.01524	0.01256	1.656522	4.095652
DSS_306	4	0.01501	0.00888	1.631522	2.895652
DSS_316	1	0.01469	0.01777	1.596739	5.794565
DSS_325	4	0.01436	0.00888	1.56087	2.895652
DSS_016	4	0.01416	0.00888	1.53913	2.895652
DSS_324	4	0.01402	0.00888	1.523913	2.895652
DSS_297	4	0.01391	0.00888	1.511957	2.895652
DSS_118	2	0.01371	0.01256	1.490217	4.095652
DSS_100	2	0.01326	0.01256	1.441304	4.095652
DSS_019	1	0.01277	0.01777	1.388043	5.794565

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_131	3	0.01269	0.01026	1.379348	3.345652
DSS_394	4	0.01219	0.00888	1.325	2.895652
DSS_385	3	0.01192	0.01026	1.295652	3.345652
DSS_395	1	0.01162	0.01777	1.263043	5.794565
DSS_287	4	0.01117	0.00888	1.21413	2.895652
DSS_418	2	0.01087	0.01256	1.181522	4.095652
DSS_290	3	0.01059	0.01026	1.151087	3.345652
DSS_314	2	0.01036	0.01256	1.126087	4.095652
DSS_073	4	0.00986	0.00888	1.071739	2.895652
DSS_040	4	0.00979	0.00888	1.06413	2.895652
DSS_037	4	0.00977	0.00888	1.061957	2.895652
DSS_341	1	0.00977	0.01777	1.061957	5.794565
DSS_302	4	0.00939	0.00888	1.020652	2.895652
DSS_104	4	0.00937	0.00888	1.018478	2.895652
DSS_273	2	0.00915	0.01256	0.994565	4.095652
DSS_322	4	0.00906	0.00888	0.984783	2.895652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_271	3	0.00901	0.01026	0.979348	3.345652
DSS_334	2	0.00898	0.01256	0.976087	4.095652
DSS_353	4	0.00864	0.00888	0.93913	2.895652
DSS_391	4	0.00764	0.00888	0.830435	2.895652
DSS_372	1	0.00737	0.01777	0.801087	5.794565
DSS_007	1	0.00729	0.01777	0.792391	5.794565
DSS_333	2	0.0072	0.01256	0.782609	4.095652
DSS_402	4	0.00718	0.00888	0.780435	2.895652
DSS_084	1	0.0069	0.01777	0.75	5.794565
DSS_103	3	0.00676	0.01026	0.734783	3.345652
DSS_362	1	0.00635	0.01777	0.690217	5.794565
DSS_012	2	0.00595	0.01256	0.646739	4.095652
DSS_396	2	0.00574	0.01256	0.623913	4.095652
DSS_133	3	0.00534	0.01026	0.580435	3.345652
DSS_065	3	0.00485	0.01026	0.527174	3.345652
DSS_284	2	0.00478	0.01256	0.519565	4.095652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_301	3	0.00465	0.01026	0.505435	3.345652
DSS_281	4	0.00461	0.00888	0.501087	2.895652
DSS_405	2	0.00449	0.01256	0.488043	4.095652
DSS_361	3	0.00438	0.01026	0.476087	3.345652
DSS_342	4	0.00434	0.00888	0.471739	2.895652
DSS_053	3	0.00422	0.01026	0.458696	3.345652
DSS_074	4	0.00422	0.00888	0.458696	2.895652
DSS_079	4	0.00375	0.00888	0.407609	2.895652
DSS_381	3	0.0036	0.01026	0.391304	3.345652
DSS_294	1	0.00336	0.01777	0.365217	5.794565
DSS_313	2	0.00332	0.01256	0.36087	4.095652
DSS_388	2	0.00305	0.01256	0.331522	4.095652
DSS_392	4	0.00287	0.00888	0.311957	2.895652
DSS_319	4	0.00282	0.00888	0.306522	2.895652
DSS_310	4	0.00263	0.00888	0.28587	2.895652
DSS_344	3	0.00259	0.01026	0.281522	3.345652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_025	4	0.00219	0.00888	0.238043	2.895652
DSS_412	1	0.00204	0.01777	0.221739	5.794565
DSS_300	3	0.00188	0.01026	0.204348	3.345652
DSS_299	2	0.00185	0.01256	0.201087	4.095652
DSS_343	4	0.00184	0.00888	0.2	2.895652
DSS_330	3	0.00153	0.01026	0.166304	3.345652
DSS_416	4	0.00128	0.00888	0.13913	2.895652
DSS_034	3	0.00128	0.01026	0.13913	3.345652
DSS_291	2	0.00102	0.01256	0.11087	4.095652
DSS_115	4	0.00063	0.00888	0.068478	2.895652
DSS_288	4	0.00044	0.00888	0.047826	2.895652
DSS_309	4	0.00008	0.00888	0.008696	2.895652
DSS_125	3	0	0.01026	0	3.345652
DSS_358	3	-0.00015	0.01026	-0.0163	3.345652
DSS_099	2	-0.00015	0.01256	-0.0163	4.095652
DSS_111	4	-0.00017	0.00888	-0.01848	2.895652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_359	3	-0.00022	0.01026	-0.02391	3.345652
DSS_015	4	-0.00043	0.00888	-0.04674	2.895652
DSS_060	3	-0.0007	0.01026	-0.07609	3.345652
DSS_098	2	-0.00088	0.01256	-0.09565	4.095652
DSS_379	4	-0.00089	0.00888	-0.09674	2.895652
DSS_356	4	-0.0009	0.00888	-0.09783	2.895652
DSS_278	4	-0.00095	0.00888	-0.10326	2.895652
DSS_368	4	-0.001	0.00888	-0.1087	2.895652
DSS_351	1	-0.0015	0.01777	-0.16304	5.794565
DSS_296	1	-0.0015	0.01777	-0.16304	5.794565
DSS_119	3	-0.00156	0.01026	-0.16957	3.345652
DSS_307	3	-0.00163	0.01026	-0.17717	3.345652
DSS_077	4	-0.00167	0.00888	-0.18152	2.895652
DSS_030	3	-0.00188	0.01026	-0.20435	3.345652
DSS_370	2	-0.00189	0.01256	-0.20543	4.095652
DSS_375	2	-0.00212	0.01256	-0.23043	4.095652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_280	3	-0.00215	0.01026	-0.2337	3.345652
DSS_345	4	-0.00225	0.00888	-0.24457	2.895652
DSS_419	1	-0.00234	0.01777	-0.25435	5.794565
DSS_298	2	-0.00249	0.01256	-0.27065	4.095652
DSS_367	3	-0.0026	0.01026	-0.28261	3.345652
DSS_072	3	-0.00268	0.01026	-0.2913	3.345652
DSS_366	4	-0.00272	0.00888	-0.29565	2.895652
DSS_063	4	-0.00283	0.00888	-0.30761	2.895652
DSS_092	3	-0.00292	0.01026	-0.31739	3.345652
DSS_347	4	-0.0033	0.00888	-0.3587	2.895652
DSS_114	4	-0.0034	0.00888	-0.36957	2.895652
DSS_303	3	-0.00396	0.01026	-0.43043	3.345652
DSS_276	4	-0.00418	0.00888	-0.45435	2.895652
DSS_083	1	-0.00446	0.01777	-0.48478	5.794565
DSS_031	2	-0.00456	0.01256	-0.49565	4.095652
DSS_328	3	-0.00463	0.01026	-0.50326	3.345652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_039	4	-0.00475	0.00888	-0.5163	2.895652
DSS_331	4	-0.00475	0.00888	-0.5163	2.895652
DSS_117	4	-0.00485	0.00888	-0.52717	2.895652
DSS_382	4	-0.00506	0.00888	-0.55	2.895652
DSS_323	4	-0.00507	0.00888	-0.55109	2.895652
DSS_041	2	-0.00527	0.01256	-0.57283	4.095652
DSS_069	4	-0.00534	0.00888	-0.58043	2.895652
DSS_308	3	-0.00534	0.01026	-0.58043	3.345652
DSS_365	3	-0.00536	0.01026	-0.58261	3.345652
DSS_403	3	-0.00594	0.01026	-0.64565	3.345652
DSS_376	1	-0.00648	0.01777	-0.70435	5.794565
DSS_293	3	-0.00652	0.01026	-0.7087	3.345652
DSS_286	1	-0.00672	0.01777	-0.73043	5.794565
BS.2C	139	-0.00694	0.00151	-0.75435	0.492391
DSS_410	1	-0.00724	0.01777	-0.78696	5.794565
DSS_312	2	-0.00725	0.01256	-0.78804	4.095652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_336	1	-0.00747	0.01777	-0.81196	5.794565
DSS_327	2	-0.00748	0.01256	-0.81304	4.095652
DSS_127	4	-0.00801	0.00888	-0.87065	2.895652
DSS_332	3	-0.0085	0.01026	-0.92391	3.345652
DSS_054	2	-0.00887	0.01256	-0.96413	4.095652
DSS_024	2	-0.00902	0.01256	-0.98043	4.095652
DSS_106	3	-0.0096	0.01026	-1.04348	3.345652
DSS_400	4	-0.00964	0.00888	-1.04783	2.895652
DSS_346	3	-0.00976	0.01026	-1.06087	3.345652
DSS_320	1	-0.01063	0.01777	-1.15543	5.794565
DSS_275	4	-0.01066	0.00888	-1.1587	2.895652
DSS_371	3	-0.01111	0.01026	-1.20761	3.345652
DSS_277	1	-0.01315	0.01777	-1.42935	5.794565
DSS_282	3	-0.01326	0.01026	-1.4413	3.345652
DSS_393	3	-0.01379	0.01026	-1.49891	3.345652
DSS_378	3	-0.01461	0.01026	-1.58804	3.345652

SNP	N	Mean Lysine Yield (change in A₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_289	3	-0.01563	0.01026	-1.69891	3.345652
DSS_317	1	-0.01565	0.01777	-1.70109	5.794565
DSS_062	4	-0.01626	0.00888	-1.76739	2.895652
DSS_340	1	-0.01657	0.01777	-1.80109	5.794565
DSS_109	2	-0.01706	0.01256	-1.85435	4.095652
DSS_011	2	-0.0178	0.01256	-1.93478	4.095652
DSS_089	4	-0.01844	0.00888	-2.00435	2.895652
DSS_059	1	-0.01848	0.01777	-2.0087	5.794565
DSS_112	2	-0.01959	0.01256	-2.12935	4.095652
DSS_043	2	-0.0213	0.01256	-2.31522	4.095652
DSS_413	1	-0.02217	0.01777	-2.40978	5.794565
DSS_305	4	-0.0227	0.00888	-2.46739	2.895652
DSS_045	4	-0.02289	0.00888	-2.48804	2.895652
DSS_082	2	-0.0231	0.01256	-2.51087	4.095652
DSS_272	1	-0.02311	0.01777	-2.51196	5.794565
DSS_390	4	-0.02319	0.00888	-2.52065	2.895652

SNP	N	Mean Lysine Yield (change in A ₅₆₀ compared to reference strain)	Std Error	% Change over Reference	% Change error
DSS_010	3	-0.02424	0.01026	-2.63478	3.345652
DSS_357	2	-0.02525	0.01256	-2.74457	4.095652
DSS_085	4	-0.03062	0.00888	-3.32826	2.895652
DSS_044	3	-0.04088	0.01026	-4.44348	3.345652
DSS_315	2	-0.0501	0.01256	-5.44565	4.095652
DSS_080	2	-0.13519	0.01256	-14.6946	4.095652

B. Second Round HTP engineering and High-Throughput Screening - Consolidation of SNP swap Library with Selected PRO swap Hits

[0793] One of the strengths of the HTP methods of the present disclosure is their ability to store HTP genetic design libraries together with information associated with each SNP/Promoter/Terminator/Start Codon's effects on host cell phenotypes. The present inventors had previously conducted a promoter swap experiment that had identified several zwf promoter swaps in *C. glutamicum* with positive effects on biosynthetic yields (*see e.g.*, results for target "N" in Figure 22).

[0794] The present inventors modified the base strain A of this Example to also include one of the previously identified zwf promoter swaps from Example 5. The top 176 SNPs identified from the initial screen described above in Table 8 were re-introduced into this new base strain to create a new SNP swap genetic design microbial library. As with the previous step, each newly created strain comprising a single SNP was tested for lysine yield. Selected SNP mutant strains were also tested for a productivity proxy, by measuring lysine production at 24 hours using the colorimetric

method described supra. The results from this step are summarized in Table 9 below, and are depicted in Figure 39.

Table 9- Second Round Screening for SNP Swap Strain Engineering for Lysine Production

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000006318	BS2C_P00 07_39zwf	20	2	0.49	0.82	0.00	0.02
7000008538	DSS_002	4	2	0.53	0.78	0.01	0.02
7000008539	DSS_003	4		0.56		0.01	
7000008541	DSS_005	4		0.27		0.01	
7000008542	DSS_006	4		0.49		0.01	
7000008547	DSS_011	4		0.55		0.01	
7000008548	DSS_012	4		0.58		0.01	
7000008549	DSS_013	4		0.56		0.01	
7000008550	DSS_014	4		0.52		0.01	
7000008551	DSS_015	4		0.54		0.01	
7000008552	DSS_016	4	2	0.50	0.84	0.01	0.02
7000008553	DSS_017	4		0.44		0.01	
7000008555	DSS_019	4	4	0.46	0.84	0.01	0.01
7000008557	DSS_021	4	4	0.46	0.86	0.01	0.01

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008559	DSS_023	4	2	0.55	0.86	0.01	0.02
7000008561	DSS_025	4		0.54		0.01	
7000008562	DSS_026	2		0.46		0.01	
7000008564	DSS_028	4		0.51		0.01	
7000008565	DSS_029	4	4	0.48	0.87	0.01	0.01
7000008566	DSS_030	4	4	0.47	0.85	0.01	0.01
7000008567	DSS_031	4		0.56		0.01	
7000008569	DSS_033	4	4	0.46	0.86	0.01	0.01
7000008570	DSS_034	2	2	0.53	0.85	0.01	0.02
7000008573	DSS_037	4		0.54		0.01	
7000008574	DSS_038	4		0.53		0.01	
7000008575	DSS_039	4		0.55		0.01	
7000008576	DSS_040	4		0.57		0.01	
7000008577	DSS_041	4		0.45		0.01	
7000008578	DSS_042	4	4	0.52	0.87	0.01	0.01
7000008579	DSS_043	4	4	0.45	0.87	0.01	0.01

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008580	DSS_044	4	2	0.50	0.85	0.01	0.02
7000008581	DSS_045	4		0.47		0.01	
7000008582	DSS_046	4	2	0.61	0.85	0.01	0.02
7000008583	DSS_047	4	2	0.61	0.82	0.01	0.02
7000008586	DSS_050	4		0.57		0.01	
7000008587	DSS_051	4		0.56		0.01	
7000008588	DSS_052	4	2	0.49	0.85	0.01	0.02
7000008589	DSS_053	4	4	0.45	0.85	0.01	0.01
7000008590	DSS_054	4	4	0.45	0.88	0.01	0.01
7000008592	DSS_056	4		0.42		0.01	
7000008596	DSS_060	4	2	0.55	0.87	0.01	0.02
7000008597	DSS_061	4	2	0.37	0.86	0.01	0.02
7000008598	DSS_062	4	4	0.45	0.87	0.01	0.01
7000008601	DSS_065	4	4	0.47	0.88	0.01	0.01
7000008602	DSS_066	4		0.47		0.01	
7000008604	DSS_068		2		0.51		0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008605	DSS_069	4	4	0.47	0.88	0.01	0.01
7000008606	DSS_070	4		0.55		0.01	
7000008607	DSS_071	4	2	0.56	0.84	0.01	0.02
7000008608	DSS_072	4	2	0.54	0.83	0.01	0.02
7000008609	DSS_073	4	2	0.47	0.84	0.01	0.02
7000008610	DSS_074	4	2	0.51	0.83	0.01	0.02
7000008612	DSS_076	4	4	0.48	0.76	0.01	0.01
7000008613	DSS_077	4	4	0.46	0.87	0.01	0.01
7000008614	DSS_078	4	2	0.44	0.87	0.01	0.02
7000008615	DSS_079	4	2	0.47	0.90	0.01	0.02
7000008616	DSS_080	4	2	0.48	0.81	0.01	0.02
7000008619	DSS_083	4	2	0.59	0.86	0.01	0.02
7000008620	DSS_084	4	2	0.70	0.89	0.01	0.02
7000008621	DSS_085	4	4	0.49	0.89	0.01	0.01
7000008622	DSS_086	4	2	0.48	0.82	0.01	0.02
7000008624	DSS_088	4	2	0.47	0.88	0.01	0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008625	DSS_089	4	4	0.45	0.89	0.01	0.01
7000008626	DSS_090	4	4	0.47	0.87	0.01	0.01
7000008627	DSS_091	4		0.46		0.01	
7000008629	DSS_093	4	4	0.50	0.87	0.01	0.01
7000008630	DSS_094	4	2	0.57	0.86	0.01	0.02
7000008634	DSS_098	4	2	0.53	0.85	0.01	0.02
7000008636	DSS_100	4		0.52		0.01	
7000008637	DSS_101	4	2	0.49	0.85	0.01	0.02
7000008640	DSS_104	4	2	0.51	0.84	0.01	0.02
7000008645	DSS_109	4		0.51		0.01	
7000008646	DSS_110	4	2	0.57	0.86	0.01	0.02
7000008648	DSS_112	4	2	0.54	0.86	0.01	0.02
7000008651	DSS_115	4		0.49		0.01	
7000008652	DSS_116	4	2	0.52	0.82	0.01	0.02
7000008653	DSS_117	4	2	0.50	0.84	0.01	0.02
7000008657	DSS_121	4	2	0.78	0.88	0.01	0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008659	DSS_123	4		0.54		0.01	
7000008663	DSS_127	4		0.58		0.01	
7000008665	DSS_129	4		0.48		0.01	
7000008666	DSS_130	4		0.56		0.01	
7000008669	DSS_133	4		0.50		0.01	
7000008670	DSS_271	4	2	0.52	0.86	0.01	0.02
7000008672	DSS_273	4		0.56		0.01	
7000008677	DSS_278	2		0.46		0.01	
7000008678	DSS_279	4		0.55		0.01	
7000008681	DSS_282	4		0.51		0.01	
7000008683	DSS_284	4		0.59		0.01	
7000008684	DSS_285	4		0.51		0.01	
7000008685	DSS_286	4		0.56		0.01	
7000008687	DSS_288	4		0.46		0.01	
7000008688	DSS_289	4		0.57		0.01	
7000008689	DSS_290	4		0.47		0.01	

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008693	DSS_294	4	2	0.52	0.63	0.01	0.02
7000008696	DSS_297	4	2	0.52	0.86	0.01	0.02
7000008697	DSS_298	4		0.58		0.01	
7000008699	DSS_300	4		0.48		0.01	
7000008700	DSS_301	4		0.58		0.01	
7000008701	DSS_302	4		0.47		0.01	
7000008702	DSS_303	3		0.46		0.01	
7000008703	DSS_304	3		0.48		0.01	
7000008705	DSS_306	4	2	0.53	0.80	0.01	0.02
7000008708	DSS_309	4		0.56		0.01	
7000008709	DSS_310	4		0.56		0.01	
7000008711	DSS_312	4		0.55		0.01	
7000008712	DSS_313	4		0.51		0.01	
7000008718	DSS_319	4	2	0.50	0.82	0.01	0.02
7000008720	DSS_321	4		0.56		0.01	
7000008722	DSS_323	2	2	0.48	0.85	0.01	0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A ₅₆₀)	Mean 96hr (A ₅₆₀)	Std Error 24hr	Std Error 96hr
7000008723	DSS_324	4		0.55		0.01	
7000008724	DSS_325	4		0.50		0.01	
7000008725	DSS_326	3		0.46		0.01	
7000008726	DSS_327	3		0.47		0.01	
7000008730	DSS_331	4		0.56		0.01	
7000008731	DSS_332	4	4	0.47	0.89	0.01	0.01
7000008732	DSS_333	4	4	0.47	0.87	0.01	0.01
7000008733	DSS_334	4		0.45		0.01	
7000008734	DSS_335	2		0.47		0.01	
7000008735	DSS_336	4		0.47		0.01	
7000008739	DSS_340	4		0.46		0.01	
7000008740	DSS_341	4	2	0.46	0.89	0.01	0.02
7000008741	DSS_342	4		0.56		0.01	
7000008742	DSS_343	4		0.55		0.01	
7000008743	DSS_344	4	4	0.48	0.87	0.01	0.01
7000008746	DSS_347	4	4	0.48	0.85	0.01	0.01

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008747	DSS_348	4	4	0.46	0.86	0.01	0.01
7000008749	DSS_350	4	2	0.29	0.74	0.01	0.02
7000008752	DSS_353	4	2	0.46	0.85	0.01	0.02
7000008753	DSS_354	4	4	0.45	0.87	0.01	0.01
7000008755	DSS_356	4	4	0.46	0.86	0.01	0.01
7000008756	DSS_357	4	4	0.46	0.86	0.01	0.01
7000008758	DSS_359	2	2	0.45	0.85	0.01	0.02
7000008760	DSS_361	4	2	0.46	0.84	0.01	0.02
7000008761	DSS_362	4		0.44		0.01	
7000008763	DSS_364	4		0.44		0.01	
7000008764	DSS_365	4		0.46		0.01	
7000008765	DSS_366	4		0.55		0.01	
7000008766	DSS_367	4		0.55		0.01	
7000008767	DSS_368	4	2	0.44	0.86	0.01	0.02
7000008770	DSS_371	4	2	0.47	0.88	0.01	0.02
7000008771	DSS_372	4	2	0.46	0.83	0.01	0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A₅₆₀)	Mean 96hr (A₅₆₀)	Std Error 24hr	Std Error 96hr
7000008772	DSS_373	4	2	0.46	0.88	0.01	0.02
7000008774	DSS_375	4		0.45		0.01	
7000008776	DSS_377	4		0.45		0.01	
7000008777	DSS_378	4		0.57		0.01	
7000008778	DSS_379	4		0.54		0.01	
7000008779	DSS_380	4	2	0.46	0.87	0.01	0.02
7000008781	DSS_382	4	2	0.46	0.84	0.01	0.02
7000008782	DSS_383	4		0.48		0.01	
7000008783	DSS_384	4	2	0.47	0.82	0.01	0.02
7000008784	DSS_385	4	2	0.46	0.83	0.01	0.02
7000008786	DSS_387	3		0.43		0.01	
7000008787	DSS_388	3		0.47		0.01	
7000008788	DSS_389	4	2	0.46	0.89	0.01	0.02
7000008790	DSS_391	4		0.57		0.01	
7000008791	DSS_392	4		0.44		0.01	
7000008795	DSS_396	4	2	0.46	0.82	0.01	0.02

Strain ID	SNP	N for 24hr	N for 96hr	Mean 24hr (A ₅₆₀)	Mean 96hr (A ₅₆₀)	Std Error 24hr	Std Error 96hr
7000008799	DSS_400	4		0.47		0.01	
7000008800	DSS_401	4	2	0.46	0.86	0.01	0.02
7000008801	DSS_402	4		0.54		0.01	
7000008805	DSS_406	4	2	0.47	0.85	0.01	0.02
7000008807	DSS_408	4		0.45		0.01	
7000008810	DSS_411	4	2	0.46	0.87	0.01	0.02
7000008812	DSS_413	3		0.47		0.01	
7000008813	DSS_414	4	2	0.45	0.84	0.01	0.02
7000008815	DSS_416	4	2	0.45	0.87	0.01	0.02
7000008816	DSS_417	4		0.46		0.01	
7000008818	DSS_419	4	2	0.47	0.84	0.01	0.02
7000008820	DSS_421	4	2	0.45	0.79	0.01	0.02
7000008821	DSS_422	4		0.44		0.01	

[0795] The results from this second round of SNP swap identified several SNPs capable of increasing base strain yield *and* productivity of lysine in a base strain comprising the zwf promoter swap mutation (*see e.g.*, SNP 084 and SNP 121 on the upper right hand corner of Figure 39).

C. Tank Culture Validation

[0796] Strains containing top SNPs identified during the HTP steps above were cultured into medium sized test fermentation tanks. Briefly, small 100ml cultures of each strain were grown over night, and were then used to inoculate 5 liter cultures in the test fermentation tanks with equal amounts of inoculate. The inoculate was normalized to contain the same cellular density following an OD600 measurement.

[0797] The resulting tank cultures were allowed to proceed for 3 days before harvest. Yield and productivity measurements were calculated from substrate and product titers in samples taken from the tank at various points throughout the fermentation. Samples were analyzed for particular small molecule concentrations by high pressure liquid chromatography using the appropriate standards. Results for this experiment are summarized in Table 10 below, and depicted in Figure 40.

Table 10- Tank Validation of SNP Swap Microbes

Strain	N	Mean Yield (%)(g lysine produced / g glucose consumed)	Std Error	Mean Productivity (g/L/h)	Std Error
base strain	1	41.1502	0.5940 1	3.29377	0.2450 8
base strain + zwf	7	48.2952	0.2245 1	2.73474	0.1000 5
base strain + zwf + SNP121	2	50.325	0.4200 3	4.51397	0.1733
base strain + zwf + pyc + lysA	5	52.191	0.2656 5	4.15269	0.1225 4

[0798] As predicted by the small scale high-throughput cultures, larger tank cultures for strains comprising the combined zwf promoter swap and SNP 121 exhibited significant increases in yield

and productivity over the base reference strain. Productivity of this strain for example, jumped to 4.5 g/L/h compared to the 3.29 g/L/h productivity of the base strain (a 37.0% increase in productivity in only 2 rounds of SNP Swap).

Example 4: HTP Genomic Engineering – Implementation of a Promoter Swap Library to Improve an Industrial Microbial Strain

[0799] Previous examples have demonstrated the power of the HTP strain improvement programs of the present disclosure for rehabilitating industrial strains. Examples 2 and 3 described the implementation of SNP swap techniques and libraries exploring the existing genetic diversity within various base, intermediate, and industrial strains

[0800] This example illustrates embodiments of the HTP strain improvement programs using the PRO swap techniques of the present disclosure. Unlike Example 3, this example teaches methods for the *de-novo* generation of mutations via PRO swap library generation.

A. Identification of a Target for Promoter Swapping

[0801] As aforementioned, promoter swapping is a multi-step process that comprises a step of: Selecting a set of “n” genes to target.

[0802] In this example, the inventors have identified a group of 23 potential pathway genes to modulate via the promoter ladder methods of the present disclosure (19 genes to overexpress and 4+ diverting genes to downregulate, in an exemplary metabolic pathway producing the molecule lysine). (See, Figure 19).

B. Creation of Promoter Ladder

[0803] Another step in the implementation of a promoter swap process is the selection of a set of “x” promoters to act as a “ladder”. Ideally these promoters have been shown to lead to highly variable expression across multiple genomic loci, but the only requirement is that they perturb gene expression in some way.

[0804] These promoter ladders, in particular embodiments, are created by: identifying natural, native, or wild-type promoters associated with the target gene of interest and then mutating said promoter to derive multiple mutated promoter sequences. Each of these mutated promoters is tested for effect on target gene expression. In some embodiments, the edited promoters are tested for expression activity across a variety of conditions, such that each promoter variant’s activity is documented/characterized/annotated and stored in a database. The resulting edited promoter

variants are subsequently organized into “ladders” arranged based on the strength of their expression (*e.g.*, with highly expressing variants near the top, and attenuated expression near the bottom, therefore leading to the term “ladder”).

[0805] In the present exemplary embodiment, the inventors have created promoter ladder:ORF combinations for each of the target genes identified in Figure 19.

C. Associating Promoters from the Ladder with Target Genes

[0806] Another step in the implementation of a promoter swap process is the HTP engineering of various strains that comprise a given promoter from the promoter ladder associated with a particular target gene.

[0807] If a native promoter exists in front of target gene *n* and its sequence is known, then replacement of the native promoter with each of the *x* promoters in the ladder can be carried out. When the native promoter does not exist or its sequence is unknown, then insertion of each of the *x* promoters in the ladder in front of gene *n* can be carried out. In this way a library of strains is constructed, wherein each member of the library is an instance of *x* promoter operably linked to *n* target, in an otherwise identical genetic context (*see e.g.*, Figure 20).

D. HTP Screening of the Strains

[0808] A final step in the promoter swap process is the HTP screening of the strains in the aforementioned library. Each of the derived strains represents an instance of *x* promoter linked to *n* target, in an otherwise identical genetic background.

[0809] By implementing a HTP screening of each strain, in a scenario where their performance against one or more metrics is characterized, the inventors are able to determine what promoter/target gene association is most beneficial for a given metric (*e.g.* optimization of production of a molecule of interest). *See*, Figure 20 (promoters P1-P8 effect on gene of interest).

[0810] In the exemplary embodiment illustrated in Figures 19-22, the inventors have utilized the promoter swap process to optimize the production of lysine. **An application of the Pro SWAP methods described above is described in Example 5, below.**

Example 5: HTP Genomic Engineering – Implementation of a PRO Swap Library to Improve Strain Performance for Lysine Production.

[0811] The section below provides an illustrative implementation of the PRO swap HTP design strain improvement program tools of the present disclosure, as described in Example 4. In this

example, a *Corynebacterium* strain was subjected to the PRO swap methods of the present disclosure in order to increase host cell yield of lysine.

A. Promoter Swap

[0812] Promoter Swaps were conducted as described in Example 4. Selected genes from the Lysine biosynthetic pathway in Figure 19 were targeted for promoter swaps using promoters P1-P8.

B. HTP engineering and High-Throughput Screening

[0813] HTP engineering of the promoter swaps was conducted as described in Example 1 and 3. HTP screening of the resulting promoter swap strains was conducted as described in Example 3. In total 145 PRO swaps were conducted. The results of the experiment are summarized in Table 11 below, and are depicted in Figure 41.

Table 11- HTP Screening of Lysine PRO Swap Libraries

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007713	Pcg1860-asd	8	0.84595	0.00689	3.927615
7000007736	Pcg0755-asd	4	0.84036	0.00974	3.240866
7000007805	Pcg0007_119-asd	8	0.82493	0.00689	1.345242
7000007828	Pcg3121-asd	8	0.8246	0.00689	1.3047
7000007759	Pcg0007_265-asd	8	0.81155	0.00689	-0.29853
7000007782	Pcg3381-asd	8	0.8102	0.00689	-0.46438
7000007712	Pcg1860-ask	8	0.83958	0.00689	3.14504
7000007735	Pcg0755-ask	8	0.81673	0.00689	0.337846
7000007827	Pcg3121-ask	8	0.81498	0.00689	0.122853
7000007804	Pcg0007_119-ask	8	0.81492	0.00689	0.115482

Strain	promoter-target	N	Mean (A₅₆₀)	Std Error	% Yield Change From Base
7000007758	Pcg0007_265-ask	8	0.80381	0.00689	-1.24942
7000007781	Pcg3381-ask	8	0.80343	0.00689	-1.2961
7000007780	Pcg3381-aspB	8	0.84072	0.00689	3.285093
7000007803	Pcg0007_119- aspB	8	0.82106	0.00689	0.8698
7000007809	Pcg0007_119- cg0931	8	0.83446	0.00689	2.516032
7000007717	Pcg1860-cg0931	4	0.83129	0.00974	2.126588
7000007763	Pcg0007_265- cg0931	4	0.82628	0.00974	1.511094
7000007671	Pcg0007_39- cg0931	8	0.82554	0.00689	1.420182
7000007740	Pcg0755-cg0931	8	0.81921	0.00689	0.642522
7000007694	Pcg0007-cg0931	8	0.80444	0.00689	-1.17202
7000007691	Pcg0007-dapA	8	0.8299	0.00689	1.955822
7000007783	Pcg3381-dapA	8	0.80951	0.00689	-0.54915
7000007760	Pcg0007_265- dapA	8	0.76147	0.00689	-6.45102

Strain	promoter-target	N	Mean (A₅₆₀)	Std Error	% Yield Change From Base
7000007806	Pcg0007_119- dapA	8	0.35394	0.00689	-56.5174
7000007761	Pcg0007_265- dapB	8	0.84157	0.00689	3.389518
7000007738	Pcg0755-dapB	4	0.84082	0.00974	3.297378
7000007692	Pcg0007-dapB	8	0.83088	0.00689	2.076218
7000007784	Pcg3381-dapB	8	0.82474	0.00689	1.3219
7000007715	Pcg1860-dapB	8	0.82232	0.00689	1.024595
7000007830	Pcg3121-dapB	8	0.81236	0.00689	-0.19902
7000007807	Pcg0007_119- dapB	4	0.69622	0.00974	-14.4672
7000007762	Pcg0007_265- dapD	8	0.84468	0.00689	3.771591
7000007808	Pcg0007_119- dapD	8	0.83869	0.00689	3.035701
7000007785	Pcg3381-dapD	8	0.83397	0.00689	2.455834
7000007670	Pcg0007_39- dapD	8	0.81698	0.00689	0.368559
7000007831	Pcg3121-dapD	4	0.8155	0.00974	0.186737

Strain	promoter-target	N	Mean (A₅₆₀)	Std Error	% Yield Change From Base
7000007693	Pcg0007-dapD	8	0.8117	0.00689	-0.28011
7000007716	Pcg1860-dapD	8	0.79044	0.00689	-2.89196
7000007739	Pcg0755-dapD	8	0.78694	0.00689	-3.32195
7000007787	Pcg3381-dapE	8	0.83814	0.00689	2.968132
7000007833	Pcg3121-dapE	8	0.83721	0.00689	2.853878
7000007741	Pcg0755-dapE	8	0.83263	0.00689	2.291211
7000007810	Pcg0007_119- dapE	8	0.83169	0.00689	2.175729
7000007718	Pcg1860-dapE	8	0.81855	0.00689	0.561439
7000007672	Pcg0007_39- dapE	8	0.80932	0.00689	-0.5725
7000007765	Pcg0007_265- dapF	8	0.8327	0.00689	2.299811
7000007788	Pcg3381-dapF	8	0.82942	0.00689	1.896853
7000007811	Pcg0007_119- dapF	8	0.82926	0.00689	1.877196
7000007696	Pcg0007-dapF	8	0.82099	0.00689	0.861201
7000007719	Pcg1860-dapF	8	0.82067	0.00689	0.821888

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007673	Pcg0007_39-dapF	8	0.82062	0.00689	0.815745
7000007789	Pcg3381-ddh	8	0.84817	0.00689	4.200349
7000007835	Pcg3121-ddh	8	0.82141	0.00689	0.912799
7000007812	Pcg0007_119-ddh	8	0.82093	0.00689	0.853829
7000007674	Pcg0007_39-ddh	8	0.81494	0.00689	0.117939
7000007720	Pcg1860-ddh	8	0.81473	0.00689	0.09214
7000007766	Pcg0007_265-ddh	8	0.81427	0.00689	0.035627
7000007743	Pcg0755-ddh	8	0.80655	0.00689	-0.9128
7000007697	Pcg0007-ddh	8	0.80621	0.00689	-0.95457
7000007779	Pcg3381-fbp	8	0.85321	0.00689	4.819529
7000007802	Pcg0007_119-fbp	4	0.81425	0.00974	0.03317
7000007710	Pcg1860-fbp	4	0.40253	0.00974	-50.5479
7000007687	Pcg0007-fbp	8	0.14881	0.00689	-81.7182
7000007825	Pcg3121-fbp	4	0.12471	0.00974	-84.679
7000007733	Pcg0755-fbp	4	0.08217	0.00974	-89.9052

Strain	promoter-target	N	Mean (A₅₆₀)	Std Error	% Yield Change From Base
7000007746	Pcg0755-hom	8	0.81925	0.00689	0.647436
7000007792	Pcg3381-hom	4	0.77674	0.00974	-4.57505
7000007723	Pcg1860-hom	8	0.71034	0.00689	-12.7325
7000007838	Pcg3121-hom	8	0.559	0.00689	-31.3251
7000007800	Pcg0007_119-icd	8	0.83236	0.00689	2.258041
7000007823	Pcg3121-icd	8	0.83155	0.00689	2.15853
7000007777	Pcg3381-icd	8	0.82844	0.00689	1.776456
7000007708	Pcg1860-icd	8	0.82384	0.00689	1.211332
7000007662	Pcg0007_39-icd	12	0.82008	0.00562	0.749404
7000007685	Pcg0007-icd	8	0.81257	0.00689	-0.17322
7000007754	Pcg0007_265-icd	4	0.81172	0.00974	-0.27765
7000007698	Pcg0007-lysA	4	0.8504	0.00974	4.474311
7000007675	Pcg0007_39- lysA	8	0.84414	0.00689	3.705251
7000007836	Pcg3121-lysA	4	0.83545	0.00974	2.637657
7000007767	Pcg0007_265- lysA	8	0.83249	0.00689	2.274012

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007813	Pcg0007_119-lysA	8	0.83096	0.00689	2.086046
7000007790	Pcg3381-lysA	8	0.8118	0.00689	-0.26782
7000007676	Pcg0007_39-lysE	8	0.84394	0.00689	3.68068
7000007699	Pcg0007-lysE	4	0.83393	0.00974	2.45092
7000007768	Pcg0007_265-lysE	8	0.83338	0.00689	2.383351
7000007837	Pcg3121-lysE	4	0.83199	0.00974	2.212585
7000007791	Pcg3381-lysE	8	0.81476	0.00689	0.095825
7000007814	Pcg0007_119-lysE	8	0.81315	0.00689	-0.10197
7000007775	Pcg3381-odx	8	0.82237	0.00689	1.030738
7000007752	Pcg0007_265-odx	8	0.81118	0.00689	-0.34399
7000007729	Pcg0755-odx	8	0.81103	0.00689	-0.36242
7000007683	Pcg0007-odx	8	0.80507	0.00689	-1.09462
7000007706	Pcg1860-odx	4	0.79332	0.00974	-2.53815
7000007660	Pcg0007_39-odx	8	0.79149	0.00689	-2.76297

Strain	promoter-target	N	Mean (A₅₆₀)	Std Error	% Yield Change From Base
7000007798	Pcg0007_119- odx	8	0.77075	0.00689	-5.31094
7000007821	Pcg3121-odx	4	0.74788	0.00974	-8.12059
7000007822	Pcg3121-pck	8	0.85544	0.00689	5.093491
7000007776	Pcg3381-pck	8	0.8419	0.00689	3.43006
7000007799	Pcg0007_119- pck	8	0.83851	0.00689	3.013588
7000007753	Pcg0007_265- pck	8	0.82738	0.00689	1.646232
7000007730	Pcg0755-pck	4	0.81785	0.00974	0.475442
7000007661	Pcg0007_39-pck	8	0.80976	0.00689	-0.51844
7000007684	Pcg0007-pck	8	0.79007	0.00689	-2.93742
7000007707	Pcg1860-pck	8	0.71566	0.00689	-12.0789
7000007840	Pcg3121-pgi	4	1.01046	0.00974	24.13819
7000007817	Pcg0007_119-pgi	7	0.99238	0.00736	21.917
7000007794	Pcg3381-pgi	7	0.99008	0.00736	21.63444
7000007771	Pcg0007_265-pgi	8	0.94665	0.00689	16.29893
7000007725	Pcg1860-pgi	8	0.85515	0.00689	5.057864

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007702	Pcg0007-pgi	4	0.8056	0.00974	-1.02951
7000007658	Pcg0007_39-ppc	4	0.85221	0.00974	4.696676
7000007750	Pcg0007_265-ppc	8	0.84486	0.00689	3.793705
7000007727	Pcg0755-ppc	8	0.84166	0.00689	3.400575
7000007773	Pcg3381-ppc	4	0.82883	0.00974	1.824369
7000007796	Pcg0007_119-ppc	8	0.82433	0.00689	1.27153
7000007704	Pcg1860-ppc	8	0.81736	0.00689	0.415244
7000007819	Pcg3121-ppc	8	0.79898	0.00689	-1.8428
7000007732	Pcg0755-ptsG	8	0.84055	0.00689	3.264208
7000007709	Pcg1860-ptsG	8	0.81075	0.00689	-0.39682
7000007663	Pcg0007_39-ptsG	8	0.80065	0.00689	-1.63763
7000007778	Pcg3381-ptsG	8	0.23419	0.00689	-71.229
7000007801	Pcg0007_119-ptsG	8	0.17295	0.00689	-78.7525
7000007824	Pcg3121-ptsG	8	0.16035	0.00689	-80.3005
7000007705	Pcg1860-pyc	8	0.85143	0.00689	4.60085

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007728	Pcg0755-pyc	8	0.79803	0.00689	-1.95951
7000007659	Pcg0007_39-pyc	8	0.75539	0.00689	-7.19797
7000007751	Pcg0007_265-pyc	8	0.73664	0.00689	-9.50146
7000007682	Pcg0007-pyc	4	0.73142	0.00974	-10.1428
7000007774	Pcg3381-pyc	4	0.66667	0.00974	-18.0975
7000007797	Pcg0007_119-pyc	4	0.52498	0.00974	-35.5046
7000007820	Pcg3121-pyc	8	0.52235	0.00689	-35.8277
7000007841	Pcg3121-tkt	8	0.82565	0.00689	1.433696
7000007818	Pcg0007_119-tkt	8	0.81674	0.00689	0.339075
7000007749	Pcg0755-tkt	8	0.81496	0.00689	0.120396
7000007703	Pcg0007-tkt	4	0.76763	0.00974	-5.69424
7000007795	Pcg3381-tkt	8	0.72213	0.00689	-11.2841
7000007772	Pcg0007_265-tkt	8	0.68884	0.00689	-15.3738
7000007701	Pcg0007-zwf	4	0.95061	0.00974	16.78542
7000007747	Pcg0755-zwf	8	0.92595	0.00689	13.75587

Strain	promoter-target	N	Mean (A ₅₆₀)	Std Error	% Yield Change From Base
7000007770	Pcg0007_265-zwf	8	0.9029	0.00689	10.9241
7000007724	Pcg1860-zwf	8	0.79309	0.00689	-2.5664
7000007839	Pcg3121-zwf	4	0.13379	0.00974	-83.5635
7000000017	-	11 6	0.92115	0.00181	13.16617
7000006284	-	12 8	0.81398	0.00172	0
7000005754	-	64	0.79489	0.00243	-2.34527

[0814] When visualized, the results of the promoter swap library screening serve to identify gene targets that are most closely correlated with the performance metric being measured. In this case, gene targets *pgi*, *zwf*, *ppc*, *pck*, *fbp*, and *ddh* were identified as genes for which promoter swaps produce large gains in yield over base strains.

[0815] Selected strains from Table 11 were re-cultured in small plates and tested for lysine yield as describe above. The results from this secondary screening are provided in Figure 22.

Example 6: Epistasis Mapping- An Algorithmic Tool for Predicting Beneficial Mutation Consolidations

[0816] This example describes an embodiment of the predictive modeling techniques utilized as part of the HTP strain improvement program of the present disclosure. After an initial identification of potentially beneficial mutations (through the use of genetic design libraries as described above), the present disclosure teaches methods of consolidating beneficial mutations in second, third, fourth, and additional subsequent rounds of HTP strain improvement. In some embodiments, the present disclosure teaches that mutation consolidations may be based on the individual performance of each of said mutations. In other embodiments, the present disclosure

teaches methods for predicting the likelihood that two or more mutations will exhibit additive or synergistic effects if consolidated into a single host cell. The example below illustrates an embodiment of the predicting tools of the present disclosure.

[0817] Selected mutations from the SNP swap and promoter swapping (PRO swap) libraries of Examples 3 and 5 were analyzed to identify SNP/PRO swap combinations that would be most likely to lead to strain host performance improvements.

[0818] SNP swapping library sequences were compared to each other using a cosine similarity matrix, as described in the “Epistasis Mapping” section of the present disclosure. The results of the analysis yielded functional similarity scores for each SNP/ PRO swap combination. A visual representation of the functional similarities among all SNPs/ PRO swaps is depicted in a heat map in Figure 15. The resulting functional similarity scores were also used to develop a dendrogram depicting the similarity distance between each of the SNPs/PRO swaps (Figure 16A).

[0819] Mutations from the same or similar functional group (*i.e.*, SNPs/PRO swaps with high functional similarity) are more likely to operate by the same mechanism, and are thus more likely to exhibit negative or neutral epistasis on overall host performance when combined. In contrast, mutations from different functional groups would be more likely to operate by independent mechanisms, and thus more likely to produce beneficial additive or combinatorial effects on host performance.

[0820] In order to illustrate the effects of biological pathways on epistasis, SNPs and PRO swaps exhibiting various functional similarities were combined and tested on host strains. Three SNP/PRO swap combinations were engineered into the genome of *Corynebacterium glutamicum* as described in Example 1: **i)** Pcg0007::zwf PRO swap + Pcg1860::pyc PRO swap, **ii)** Pcg0007::zwf PRO swap + SNP 309, and **iv)** Pcg0007::zwf PRO swap + Pcg0007::lysA PRO swap (see Figure 15 and 16A for functional similarity relationships).

[0821] The performance of each of the host cells containing the SNP/PRO swap combinations was tested as described in Example 3, and was compared to that of a control host cell containing only zwf PRO swap. Tables 12 and 13 below summarize the results of host cell yield (96hr measurements) and productivity (24hr measurements) of each of the strains.

Table 12- Lysine Accumulation for Epistasis Mapping Experiment at 24 hours.

SNP/ PRO swap	Mean Lysine (A₅₆₀)	StDev
6318 (zwf)	0.51	0.03
8126 (zwf+ lysA)	0.88	0.06
8156 (zwf+ pyc)	0.53	0.01
8708 (zwf+ SNP 309)	0.56	0.00

Table 13- Lysine Accumulation for Epistasis Mapping Experiment at 96 hours.

SNP/ PRO swap	Mean Lysine (A₅₆₀)	StDev
6318 (zwf)	0.83	0.01
8126 (zwf + lysA)	0.94	0.02
8156 (zwf+ pyc)	0.83	0.06

[0822] Host yield performance results for each SNP/PRO swap combination are also depicted in Figure 16B. Host strains combining SNPs/PRO swaps exhibiting lower functional similarity outperformed strains in which the combined SNPs had exhibited higher functional similarity at both 24, and 96 hour measurements.

[0823] Thus, the epistatic mapping procedure is useful for predicting/programming/informing effective and/or positive consolidations of designed genetic changes. The analytical insight from

the epistatic mapping procedure allows for the creation of predictive rule sets that can guide subsequent rounds of microbial strain development. The predictive insight gained from the epistatic library may be used across microbial types and target molecule types.

Example 7: HTP Genomic Engineering –Pro Swap Mutation Consolidation and Multi-Factor Combinatorial Testing

[0824] Previous examples have illustrated methods for consolidating a small number of pre-selected PRO swap mutations with SNP swap libraries (Example 3). Other examples have illustrated the epistatic methods for selecting mutation consolidations that are most likely to yield additive or synergistic beneficial host cell properties (Example 6). This example illustrates the ability of the HTP methods of the present disclosure to effectively explore the large solution space created by the combinatorial consolidation of multiple gene/genetic design library combinations (e.g., PRO swap library x SNP Library or combinations within a PRO swap library).

[0825] In this illustrative application of the HTP strain improvement methods of the present disclosure, promoter swaps identified as having a positive effect on host performance in Example 5 are consolidated in second order combinations with the original PRO swap library. The decision to consolidate PRO swap mutations was based on each mutation's overall effect on yield or productivity, and the likelihood that the combination of the two mutations would produce an additive or synergistic effect.

[0826] For example, applicants refer to their choice of combining Pcg0007::zwf and Pcg0007::lysA, based on the epistasis mapping results of Example 6.

A. Consolidation Round for PRO Swap Strain Engineering

[0827] Strains were transformed as described in previous Example 1. Briefly, strains already containing one desired PRO swap mutation were once again transformed with the second desired PRO swap mutation. In total, the 145 tested PRO swaps from Example 5 were consolidated into 53 second round consolidation strains, each comprising two PRO swap mutations expected to exhibit beneficial additive or synergistic effects.

[0828] The resulting second round strains were once again screened as described in Example 3. Results from this experiment are summarized in Table 14 below, and depicted in Figure 11.

Table 14- HTP Screening of Second Round Consolidated Lysine PRO Swap Libraries

Strain ID	Number	PRO Swap 1	PRO Swap 2	Mean Yield (A₅₆₀)	Std Dev
7000008489	4	Pcg0007-lysA	Pcg3121-pgi	1.17333	0.020121
7000008530	8	Pcg1860-pyc	Pcg0007-zwf	1.13144	0.030023
7000008491	7	Pcg0007-lysA	Pcg0007-zwf	1.09836	0.028609
7000008504	8	Pcg3121-pck	Pcg0007-zwf	1.09832	0.021939
7000008517	8	Pcg0007_39- ppc	Pcg0007-zwf	1.09502	0.030777
7000008502	4	Pcg3121-pck	Pcg3121-pgi	1.09366	0.075854
7000008478	4	Pcg3381-ddh	Pcg0007-zwf	1.08893	0.025505
7000008465	4	Pcg0007_265- dapB	Pcg0007-zwf	1.08617	0.025231
7000008535	8	Pcg0007-zwf	Pcg3121-pgi	1.06261	0.019757
7000008476	6	Pcg3381-ddh	Pcg3121-pgi	1.04808	0.084307
7000008510	8	Pcg3121-pgi	Pcg1860-pyc	1.04112	0.021087
7000008525	8	Pcg1860-pyc	Pcg0007_265- dapB	1.0319	0.034045
7000008527	8	Pcg1860-pyc	Pcg0007-lysA	1.02278	0.043549
7000008452	5	Pcg1860-asd	Pcg0007-zwf	1.02029	0.051663

Strain ID	Number	PRO Swap 1	PRO Swap 2	Mean Yield (A₅₆₀)	Std Dev
7000008463	4	Pcg0007_265-dapB	Pcg3121-pgi	1.00511	0.031604
7000008524	8	Pcg1860-pyc	Pcg1860-asd	1.00092	0.026355
7000008458	4	Pcg3381-aspB	Pcg1860-pyc	1.00043	0.020083
7000008484	8	Pcg3381-fbp	Pcg1860-pyc	0.99686	0.061364
7000008474	8	Pcg3381-ddh	Pcg3381-fbp	0.99628	0.019733
7000008522	8	Pcg0755-ptsG	Pcg3121-pgi	0.99298	0.066021
7000008528	8	Pcg1860-pyc	Pcg3121-pck	0.99129	0.021561
7000008450	4	Pcg1860-asd	Pcg3121-pgi	0.98262	0.003107
7000008448	8	Pcg1860-asd	Pcg3381-fbp	0.97814	0.022285
7000008494	8	Pcg0007_39-lysE	Pcg3381-fbp	0.97407	0.027018
7000008481	8	Pcg3381-fbp	Pcg0007-lysA	0.9694	0.029315
7000008497	8	Pcg0007_39-lysE	Pcg1860-pyc	0.9678	0.028569
7000008507	8	Pcg3121-pgi	Pcg3381-fbp	0.96358	0.035078
7000008501	8	Pcg3121-pck	Pcg0007-lysA	0.96144	0.018665

Strain ID	Number	PRO Swap 1	PRO Swap 2	Mean Yield (A₅₆₀)	Std Dev
7000008486	8	Pcg0007-lysA	Pcg0007_265-dapB	0.94523	0.017578
7000008459	8	Pcg0007_265-dapB	Pcg1860-asd	0.94462	0.023847
7000008506	2	Pcg3121-pgi	Pcg0007_265-dapD	0.94345	0.014014
7000008487	8	Pcg0007-lysA	Pcg3381-ddh	0.94249	0.009684
7000008498	8	Pcg3121-pck	Pcg1860-asd	0.94154	0.016802
7000008485	8	Pcg0007-lysA	Pcg1860-asd	0.94135	0.013578
7000008499	8	Pcg3121-pck	Pcg0007_265-dapB	0.93805	0.013317
7000008472	8	Pcg3381-ddh	Pcg1860-asd	0.93716	0.012472
7000008511	8	Pcg0007_39-ppc	Pcg1860-asd	0.93673	0.015697
7000008514	8	Pcg0007_39-ppc	Pcg0007-lysA	0.93668	0.027204
7000008473	8	Pcg3381-ddh	Pcg0007_265-dapB	0.93582	0.030377

Strain ID	Number	PRO Swap 1	PRO Swap 2	Mean Yield (A₅₆₀)	Std Dev
7000008461	7	Pcg0007_265-dapB	Pcg3381-fbp	0.93498	0.037862
7000008512	8	Pcg0007_39-ppc	Pcg0007_265-dapB	0.93033	0.017521
7000008456	8	Pcg3381-aspB	Pcg3121-pck	0.92544	0.020075
7000008460	8	Pcg0007_265-dapB	Pcg0007_265-dapD	0.91723	0.009508
7000008492	8	Pcg0007_39-lysE	Pcg3381-aspB	0.91165	0.012988
7000008493	8	Pcg0007_39-lysE	Pcg0007_265-dapD	0.90609	0.031968
7000008453	8	Pcg3381-aspB	Pcg0007_265-dapB	0.90338	0.013228
7000008447	8	Pcg1860-asd	Pcg0007_265-dapD	0.89886	0.028896
7000008455	8	Pcg3381-aspB	Pcg0007-lysA	0.89531	0.027108
7000008454	6	Pcg3381-aspB	Pcg3381-ddh	0.87816	0.025807
7000008523	8	Pcg0755-ptsG	Pcg1860-pyc	0.87693	0.030322
7000008520	8	Pcg0755-ptsG	Pcg3381-fbp	0.87656	0.018452

Strain ID	Number	PRO Swap 1	PRO Swap 2	Mean Yield (A ₅₆₀)	Std Dev
7000008533	4	Pcg0007-zwf	Pcg3381-fbp	0.84584	0.017012
7000008519	8	Pcg0755-ptsG	Pcg0007_265-dapD	0.84196	0.025747

[0829] As predicted by the epistasis model, the second round PRO swap strain comprising the Pcg0007:: zwf and Pcg0007:: lysA mutations exhibited one of the highest yield improvements, with a nearly 30% improvement in yield over Pcg0007::lysA alone, and a 35.5% improvement over the base strain (see circled data point on Figure 11).

[0830] The HTP methods for exploring solution space of single and double consolidated mutations, can also be applied to third, fourth, and subsequent mutation consolidations. Attention is also drawn, for example, to the disclosed 3-change consolidation strain corresponding to zwf, pyc, and lysa that was made from amongst the top hits of identified in the 2 change consolidations as shown in Table 14 above, and as identified by the epistatic methods of the present disclosure. This 3-change consolidation strain was further validated in tanks as being significantly improved as compared to the parent or parent + zwf (*see* Table 10 *supra*, and Figure 40).

Example 8: HTP Genomic Engineering – Implementation of a Terminator Library to Improve an Industrial Host Strain

[0831] The present example applies the HTP methods of the present disclosure to additional HTP genetic design libraries, including STOP swap. The example further illustrates the ability of the present disclosure to combine elements from basic genetic design libraries (e.g., PRO swap, SNP swap, STOP swap, etc.) to create more complex genetic design libraries (e.g., PRO-STOP swap libraries, incorporating both a promoter and a terminator). In some embodiments, the present disclosure teaches any and all possible genetic design libraries, including those derived from combining any of the previously disclosed genetic design libraries.

[0832] In this example, a small scale experiment was conducted to demonstrate the effect of the STOP swap methods of the present invention on gene expression. Terminators T1-T8 of the present disclosure were paired with one of two native *Corynebacterium glutamicum* promoters as described below, and were analyzed for their ability to impact expression of a fluorescent protein.

A. Assembly of DNA constructs

[0833] Terminators T1-T8 were paired with one of two native *Corynebacterium glutamicum* promoters (e.g., Pcg0007 or Pcg0047) expressing a yellow fluorescence protein (YFP). To facilitate DNA amplification and assembly, the final promoter-YFP-terminator sequence was synthesized in two portions; the first portion encoded (from 5' to 3') **i)** the vector homology arm, **ii)** the selected promoter, **iii)** and 2/3 of the YFP gene. The second portion encoded (from 5' to 3') **iv)** the next 2/3 of the YFP gene, **v)** the selected terminator, and **vi)** the second vector homology arm. Each portion was amplified using synthetic oligonucleotides and gel purified. Gel purified amplicons were assembled with a vector backbone using yeast homologous recombination.

B. Transformation of Assembled Clones into *E. coli*

[0834] Vectors containing the Promoter-YFP-terminator sequences were each individually transformed into *E. coli* in order to identify correctly assembled clones, and to amplify vector DNA for *Corynebacterium* transformation. Correctly assembled vectors were confirmed by restriction enzyme digest and Sanger sequencing. Positive clones were stored at -20°C for future use.

C. Transformation of Assembled Clones into *Corynebacterium*

[0835] Verified vector clones were individually transformed into *Corynebacterium glutamicum* host cells via electroporation. Each vector was designed to integrate into a neutral integration site within the *Corynebacterium glutamicum* genome that was empirically determined to permit expression of heterologous yellow fluorescence protein but not be detrimental to the host cell. To facilitate integration, the expression vector further comprised about 2 kbp of sequence homologous (i.e., homology arm) to the desired integration site whereby each gene cassette described above was inserted downstream of the homology arm. Integration into the genome occurred by single-crossover integration. Transformed *Corynebacterium* were then tested for correct integration via PCR. This process was repeated for each of the transformations conducted for each gene construct.

D. Evaluation of Individual terminator constructs in *Corynebacterium*

[0836] The phenotype of each *Corynebacterium* transformant containing promoter-YFP-terminator constructs was then tested in two media types (brain heart infusion-BHI and HTP test media) at two time points in order to evaluate expression. Briefly, between four and six PCR-confirmed transformants were chosen and cultivated in selective media in a 96-well format. The initial cultures were then split into selective BHI media or selective seed media. At 48 hours, cultures in seed media were inoculated into selective HTP test media or BHI media and analyzed at two time points representing different portions of the growth curve. Time points for HTP test media cultures were 48 and 96 hours after inoculation. Cultures in the selective BHI media were analyzed at 48 and 72 hours after inoculation.

[0837] Analysis of the cultures was performed using a benchtop flow cytometer. Briefly, cultures were diluted 1:100 in 200 μ l of phosphate buffered saline (PBS). For each culture, between 3000 and 5000 individual events (i.e., cells) were analyzed for yellow fluorescence. The benchtop flow cytometer plots a histogram of yellow fluorescence of each “event” and calculates the median fluorescence within each well. Figure 36 depicts the mean of the median fluorescence for each construct (across the 4-6 biological replicates). Error bars indicate the 95% confidence interval of each data point. Conditions A-D each refer to a single media and a single time point. Thus conditions A and B represent the two time points for the BHI media, while the C and D points represent the two time points for the HTP test media. Note that the arbitrary units (e.g., AU) represent the median fluorescence recorded by the benchtop flow cytometer.

[0838] The results show that terminators 1-8 of the STOP swap genetic design library result in a continuous range of YFP expression. These terminators thus form a terminator ladder that can be implemented into future genetic design libraries, according to the HTP methods of the present disclosure.

Example 9: Comparing HTP Toolsets vs. Traditional UV Mutations.

[0839] This example demonstrates the benefits of the HTP genetic design libraries of the present disclosure over traditional mutational strain improvement programs. The experiments in this portion of the specification quantify the improved magnitude and speed of the phenotypical improvements achieved through the HTP methods of the present disclosure over traditional UV mutagenesis.

[0840] The present disclosure teaches new methods for accelerating the strain improvement programs of host cells. In some embodiments, the HTP strain improvement program of the present disclosure relies on the ability of the HTP toolsets to generate and identify genetic perturbations. The present inventors attempted to quantify the benefits of the HTP tool sets by conducting a small parallel track strain improvement program comparing the promoter swap techniques of the present disclosure against traditional UV mutations approaches.

[0841] A base reference strain producing a biochemical metabolite of interest was chosen as the starting point for both UV and promoter swap genetic perturbations.

A. UV mutations

[0842] Cultures of the base strain were grown in BHI medium in cultures that were OD normalized to OD₆₀₀ of 10. This culture was aliquoted into a sterile petri dish and agitated using a small magnetic stirrer bar. A UV trans illuminator at 254 nm wavelength was then inverted over the culture and aliquots taken at 5 and 9 minutes of UV exposure. These samples were serially diluted 10-fold and each dilution plated onto BHI medium Q-trays. From these Q-trays, approximately 2500 colonies from each UV exposure point were picked using an automated colony picking apparatus and the performance evaluated as below.

B. Promoter Swap

[0843] PRO swap constructs were generated in the base strain for 15 gene targets using either all or a subset of promoters selected from P1, P3, P4 and P8 described in Table 1. The final step in the biosynthesis of the product of interest is catalyzed by an O-methyltransferase enzyme that utilizes the potentially rate limiting cofactor S-adenosylmethionine. Gene targets for PRO swaps were therefore selected on the basis that they are directly involved in the biosynthesis of this cofactor or upstream metabolites.

C. UV and Promoter Swap Library Evaluation

[0844] The phenotype of each *Corynebacterium* strain developed for this example was tested for its ability to produce a selected biomolecule. Briefly, between four and six sequence confirmed colonies from each PRO swap strain, and single colonies for each UV strain were chosen and propagated in selective media in a 96-well format in production liquid media.

[0845] After biomass propagation in 96-well microwell plates, cell mass was added to fermentation media containing substrate in 96-well microwell plates and bioconversion was allowed to proceed for 24 hrs. Titer of product were determined for each strain using high-

performance liquid chromatography from samples taken at 24 hrs. The titer results for each genetic perturbation (UV and PRO swap) was analyzed. Results for each replicate was averaged and assigned to represent the overall performance of said strain. Strains were then binned into categories based on each mutation's effect on measured yield expressed as a ratio over the yield of the base strain.

[0846] Figure 37 summarizes the results of this experiment, which are presented as the number of strains for each strain improvement technique that produced: i) no change in yield, ii) a 1.2 to 1.4 fold improvement to yield, iii) a 1.4 to 1.6 fold improvement to yield, iv) a 1.6 to 1.8 fold improvement to yield, or v) a 1.8 to 2 fold improvement to yield.

[0847] The results are illustrative of the benefits of the HTP toolsets of the present disclosure over traditional UV mutagenesis approaches. For example, the results of Figure 37 demonstrate that the PRO swap strains exhibited a higher rate of positive changes in yield, and were therefore more likely to provide mutations that could significantly improve the strain. Most striking, was the high incidence of high improvement strains showing 1.6, 1.8 and 2 fold increases in the PRO swap library, with little to no identified improvements in the UV library.

[0848] The results are also important because they highlight the accelerated rate of improvement of the PRO swap methods of the present disclosure. Indeed, results for the PRO swap library were based on less than 100 promoter::gene perturbations, whereas UV mutation results included the screening of over 4,000 distinct mutant strains. Thus the methods of the present disclosure drastically reduce the number of mutants that must be screened before identifying genetic perturbations capable of conferring strains with high gains in performance.

Example 10 – HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance *Escherichia coli*

[0849] Previous examples illustrate applications of HTP strain improvement programs on *Corynebacterium*. This example demonstrates the applicability of the same techniques to *E. coli* cells.

[0850] This example describes the application of transposon mutagenesis to generate *Escherichia coli* random strain libraries for the purpose of strain improvement. These strain libraries can be screened against a desired phenotype such as yield of tryptophan to identify variants with improved performance.

[0851] The present disclosure describes a method for generating a library of mutants through the application of the EZ-Tn5 transposon system (Epicenter Bio) in *Escherichia coli*. The EZ-Tn5 Transposase is incubated with payload DNA flanked by mosaic element sequences. Upon incubation, the Ez-Tn5 Transposase complexes with the DNA to form a transposome. The DNA/protein transposome complex is then transformed into *Escherichia coli* through electroporation. The EZ-Tn5 Transposase catalyzes the random integration of the payload DNA into the *Escherichia coli* genome thus giving rise to a random library of strain variants.

[0852] The specific sequence of the payload DNA can further be varied to bias toward either loss of function (LoF) or gain of function (GoF) effects of transposon insertion into the target genome.

[0853] LoF can be accomplished through inclusion of an antibiotic selection marker in the DNA payload. The antibiotic marker allows for the selection of cells with a productive transposon insertion. The insertion of the DNA payload may disrupt the function of DNA into which is has inserted in various ways including but not limited disruption of an open reading frame that prevents translation of the disrupted gene.

[0854] GoF can be accomplished through the inclusion of an antibiotic marker and a strong promoter in the DNA payload. The antibiotic marker allows for the selection of cells with a productive transposon insertion. The insertion of the DNA payload may increase the expression in genes proximal to the insertion site through the action of the strong promoter.

[0855] Either LoF of GoF DNA payloads may further contain a counterselection marker in addition to a selection marker to enable marker recycling and thus further rounds of engineering.

[0856] The library of strain variants generated through the transposon mutagenesis method described above can be screened against a desired phenotype. Strains can be cultivated and tested in high throughput to identify strains with an improved desired phenotype relative to the parent strain.

[0857] The improved strain variants can be subjected to additional rounds of cyclical engineering to further improve the desired phenotype such as yield of tryptophan. The additional rounds of engineering may consist of transposon mutagenesis or other library types such as SNPSWP, PROSWP, or random mutagenesis. The improved strains may also be consolidated with other strain variants exhibiting an improved phenotype to produce a further improved strain through the additive effect of distinct beneficial mutations.

[0858] The methods described herein reduce the cost for building high quality libraries for screening in cyclical engineering. Application of transposon mutagenesis to *Escherichia coli*

enables the production of thousands of genome wide of LoF or GoF mutants in a single reaction. An alternative method is to construct thousands of assigned plasmids to engineer strains through single crossover homologous recombination (SCHR). Another alternative method is to construct thousands of assigned linear fragments to engineer strains through lambda red recombineering. Both methods are costly because they require generating a unique DNA fragments for each mutant that contains the intended payload DNA and sequence homology that directs recombination to a specific location on the target genome. Transposon mutagenesis uses a single DNA payload and diversity is generated through random integration into the target genome.

Example 11– HTP Genomic Engineering – Generation of Vector Backbones for use in HTP Genomic Engineering in *Escherichia coli*

[0859] This example describes the generation of vectors for use in the HTP genomic engineering for recombineering in *Escherichia coli* such that said vectors confer efficient transformation and plasmid integration.

[0860] Vector 1 (nucleic acid SEQ ID NO. 214) was generated and comprises an R6K origin of replication, a SacB gene, a PheS gene as a counterselection marker and a URA3 yeast auxotrophic marker. In order to improve the efficiency and single-crossover homologous recombination, the backbone of vector 1 was modified to contain the elements in Table 15, which resulted in vector 2 (nucleic acid SEQ ID NO. 215). The plasmid map shown in Figure 55 shows the general components of vector 1. In vector 2, random insulator sequences Insulator1 (SEQ ID NO. 218) and Insulator2 (SEQ ID NO. 219) were added flanking the homology arms, and terminator sequences T1 (nucleic acid SEQ ID NO. 220; see Orosz et al., Eur J Biochem. 1991 Nov 1;201(3):653-9) and B0015 (nucleic acid SEQ ID NO. 221) were added to eliminate transcriptional read-through at the site of genomic insertion. The plasmid map shown in Figure 56 shows the general components of vector 2.

[0861] The utility of the vectors or plasmids was tested in knock-out experiments. In summary, *E. coli* was inoculated into LB broth and grown for 8 hours at 37C with shaking. Subsequently, an aliquot of the overnight culture was then used to inoculate a larger volume of LB broth and grown for 16 hours at 18⁰C with shaking. For transformations, 100-400 ng of test plasmid was added to competent cells and transformation was carried out by electroporation. Cells were recovered in SOC media with incubation at 37C for 3 hours before plating on LB-agar with kanamycin. The plate was incubated at 37C to grow colonies with test plasmid.

[0862] The gene target to be knocked out was the *E. coli* aroA gene. As such, test plasmids of “aroA-KO in version 1” (i.e., vector 1) and “aroA-KO in version 2” (i.e., vector 2) were constructed by the insertion of homology arms to the *E. coli* aroA gene into the backbones of vector 1 (version 1) and vector 2 (version 2), respectively, such that the homology arms flanked a kanamycin resistance gene to allow single-crossover homology recombination in the *E. coli* host cell. Transformation of these test plasmids and selection on kanamycin verified that “aroA-KO in version 2” showed improved efficiency of transformation and plasmid integration (**Figure 53**).

[0863] Further modification of the vector 1 backbone in vector 2 allowed efficient counter-selection in media containing sucrose and 4-chlorophenylalanine by adding the PheS sequence (Table 15). It should be noted that the PheS promoter sequence in vector 2 consists of the phage λ PL promoter identified by Kincaide and deHaseth (*see* Kinacade and deHaseth Gene. 1991 Jan 2;97(1):7-12) immediately followed by an RBS sequence called B0032 which came from iGEM. Further, in vector 2, the promoter sequence of sacB gene was replaced with a promoter containing the P5-MCD2 promoter (Mutalik et al, Nat Methods. 2013 Apr;10(4):354-60) and an additional ATG. This modification allowed efficient counter-selection in sucrose with strains integrated with the version 2 backbone. To generate vector 3, the promoter sequence and CDS of the *C. glutamicum* pheS* gene in backbone vector 1 were replaced with a new promoter sequence and CDS, specifically a codon-optimized version of the native *E. coli* pheS containing the requisite mutations (T251A/ A294G, *see* Miyazaki, K. Biotechniques. 2015 Feb 1;58(2):86-8) (Table 15). This modification allowed improved counter-selection in 4 chlorophenylalanine with strains integrated with the vector 3 backbone. The plasmid map shown in Figure 57 shows the general components of vector 3.

[0864] The backbone used for HTP genomic engineering may contain various yeast selection markers for use in plasmid assembly. In the present disclosure, modification of the vector 3 backbone replaced the URA3 yeast selection marker with a TRP1 marker to give vector 4. The plasmid map shown in Figure 58 shows the general components of vector 4.

Example 12 – HTP Genomic Engineering – Generation and testing of an additional Promoter Swap Library for use in improving an Industrial Microbial Strain

[0865] This example describes the generation of an additional PROSWP library for later use in the HTP genomic engineering methods provided herein for genetically engineering microbial host cells (e.g., *Escherichia coli*) in an effort to improve industrial strain performance.

[0866] In this example, a number of native *E. coli* promoters and synthetic promoters were compiled to generate the promoter swap library found in Table 1.4. For the native promoters, a set of promoter sequences 60-90 bp in length was selected from the genome of an *E. coli* K-12 strain (i.e., *E. coli* W3110). In particular, promoters were selected that showed minimal variation in the associated gene's expression, according to microarray-based expression data across multiple growth conditions (Lewis et al., Mol Syst Biol. 2010; 6: 390). The native promoter sequences consisted of 50bp in front of putative transcription start sites as well as the sequence up to but not including the putative start codon (see Table 1.4). Additionally, a set of chimeric synthetic promoter sequences was created consisting of portions of the known lambda phage promoters *p_L* and *p_R*, the promoter in front of *E. coli* gene *acs*, and variable 6bp sequences constituting the -35 and -10 regions (Figure 54, Table 1.5). Each of the synthetic promoters were 60-90 bp in length.

[0867] In order to test the ability of each of the promoters found in Table 1.4 to drive expression of a gene operably linked thereto, a set of low-copy replicating plasmids was constructed, each containing the RFP gene driven by one of the promoters listed in Table 1.4. The low-copy replicating plasmid of choice was a plasmid called Ori_Plsmid27, which has nucleic sequence SEQ ID NO. 213. The vector was chosen because a replicating plasmid was desired in order to construct and evaluate the promoter library as rapidly as possible and a low copy plasmid such as Ori_Plsmid27 would more closely approximate the scenario in which only a single copy is integrated into the genome. Ori_Plsmid27 is low-copy because it possesses the *E. coli* origin of replication p15A. The p15A origin of replication typically results in approximately 10 copies of the plasmid in each cell. This is "low-copy" in comparison to other common plasmids which may maintain >20 or even several hundred plasmid copies per cell.

[0868] The plasmids were constructed using standard molecular biology techniques. Specifically, forward PCR primers were purchased consisting of a sequence to anneal to the RFP gene; the promoter sequence to be introduced; and a sequence overlapping with Ori_Plsmid27. A single reverse PCR primer was obtained consisting of a sequence to anneal to the ECK120033737 terminator (a native *E. coli* terminator) and a sequence overlapping with Ori_Plsmid27. The RFP gene was amplified by PCR with the forward primers and reverse primer to generate a set of PCR amplicons, each containing the RFP gene and one of the promoters listed in Table 1.4. The plasmids were constructed by digesting Ori_Plsmid27 with XhoI restriction enzyme and inserting the corresponding PCR amplicon using commercial DNA assembly enzyme mix. As a negative

control , a construct comprising the *C. glutamicum* Tsod terminator (nucleic acid SEQ ID NO. 224 in Table 16) placed upstream of the RFP gene was generated.

[0869] The plasmids were transformed by electroporation into *E. coli* W3110. For each promoter to be evaluated, four colonies were picked and inoculated into 1 mL LB broth containing 25 µg/mL kanamycin in a 96-well culture plate. Cultures were grown at 37°C overnight with shaking at 1000 rpm. 10 µL of culture were used to inoculate 1 mL Media 1 (a rich medium containing glucose, yeast extract, salt, and phosphate buffer) containing 25 µg/mL kanamycin in a 96-well culture plate. Cultures were grown at 37°C for 24 hrs with shaking at 1000 rpm. The cultures were diluted in water in a black wall clear bottom 96-well plate and two measurements were taken on a spectrophotometer: OD₆₀₀ (optical density at 600 nm) and fluorescence (554 nm excitation, 590 nm emission). 10 µL of the cultures in Media 1 were used to inoculate 1 mL Media 2 (a rich media containing higher glucose than Media 1 but only a small amount of yeast extract, instead containing ammonium sulfate as nitrogen source along with trace elements) containing 25 µg/mL kanamycin in a 96-well culture plate. Cultures in Media 2 were likewise grown at 37°C overnight and measured after 24 hrs.

[0870] OD₆₀₀ measurements were corrected by subtracting the value of blank wells (wells containing only media). Table 16 shows the fluorescence measurements normalized for the corrected OD₆₀₀. As can be seen in Table 16, the resulting strains were effectively grown in two different media and enabled fluorescent protein expression over a ~5000-fold range.

Table 16. RFP Expression Levels for promoter-RFP constructs in (2) different media.

Promoter name	SEQ ID NO.	Type*	Fluorescence in media 1	Fluorescence in media 2
Empty vector		Control	0.59310262	1.038945766
No promoter		Control	0.372576311	0.759290158
Terminator	224	Control	0.41480568	0.638974345
b0904_promoter	71	Native	797.7465067	5344.268555
b2405_promoter	72	Native	181.125065	918.6698873
b0096_promoter	73	Native	124.2236508	404.6536222
b0576_promoter	74	Native	22.74750166	231.1578591
b2017_promoter	75	Native	43.07006732	214.1483559

b1278_promoter	76	Native	24.51965879	167.5502652
b4255_promoter	77	Native	35.05859344	159.6253961
b0786_promoter	78	Native	34.42969151	136.7168488
b0605_promoter	79	Native	19.80356197	116.9009802
b1824_promoter	80	Native	24.81069692	111.2950744
b1061_promoter	81	Native	23.595542	104.971453
b0313_promoter	82	Native	20.68806067	75.34101875
b0814_promoter	83	Native	16.46706854	70.36934251
b4133_promoter	84	Native	18.1034447	67.11707926
b4268_promoter	85	Native	12.80314528	60.91103948
b0345_promoter	86	Native	9.57124889	56.38701501
b2096_promoter	87	Native	9.852378505	44.58545057
b1277_promoter	88	Native	7.821796935	43.50459312
b1646_promoter	89	Native	18.43768693	33.66381279
b4177_promoter	90	Native	12.51676981	30.71734482
b0369_promoter	91	Native	6.614285367	28.3955664
b1920_promoter	92	Native	6.196134846	28.12617522
b3742_promoter	93	Native	5.646106239	27.85277232
b3929_promoter	94	Native	9.120314333	27.26180409
b3743_promoter	95	Native	5.416134854	26.11889953
b1613_promoter	96	Native	6.653798388	24.49553744
b1749_promoter	97	Native	4.14048349	24.04592462
b2478_promoter	98	Native	5.387025634	20.49940077
b0031_promoter	99	Native	3.227415659	18.8898903
b2414_promoter	100	Native	5.054709408	17.84254467
b1183_promoter	101	Native	5.109245812	14.42155472
b0159_promoter	102	Native	3.662900763	14.16361794
b2837_promoter	103	Native	3.427953441	13.22830567
b3237_promoter	104	Native	3.238274954	11.27938911
b3778_promoter	105	Native	1.778132048	8.89448301
b2349_promoter	106	Native	1.918547188	8.854115649
b1434_promoter	107	Native	3.05979593	7.982575199

b3617_promoter	108	Native	1.631428047	7.872487968
b0237_promoter	109	Native	2.524629463	6.00416496
b4063_promoter	110	Native	1.75938278	5.338598994
b0564_promoter	111	Native	1.099810487	5.27271929
b0019_promoter	112	Native	1.407515599	5.174702317
b2375_promoter	113	Native	0.758754127	4.329319889
b1187_promoter	114	Native	1.042830698	4.119645144
b2388_promoter	115	Native	1.068375672	4.008318722
b1051_promoter	116	Native	2.595570861	3.770122087
b4241_promoter	117	Native	0.86986247	3.520447288
b4054_promoter	118	Native	1.166071321	3.009684331
b2425_promoter	119	Native	0.684355198	2.847357334
b0995_promoter	120	Native	0.783460564	2.701492023
b1399_promoter	121	Native	0.67920166	2.675743061
b3298_promoter	122	Native	0.693842895	2.248914103
b2114_promoter	123	Native	0.680005978	2.166569672
b2779_promoter	124	Native	0.516911046	1.708793442
b1114_promoter	125	Native	0.477678475	1.423398279
b3730_promoter	126	Native	0.527510171	1.40174991
b3025_promoter	127	Native	0.528069061	1.375386866
b0850_promoter	128	Native	0.580500872	1.365244168
b2365_promoter	129	Native	0.600099606	1.293746086
b4117_promoter	130	Native	0.465598493	1.292314074
b2213_promoter	131	Native	0.502863501	1.066458143
pMB029_promoter	132	Synthetic	3244.159758	6397.50355
pMB023_promoter	133	Synthetic	2716.103558	5324.355721
pMB025_promoter	134	Synthetic	2529.260134	5315.845766
pMB019_promoter	135	Synthetic	2632.199487	4730.881793
pMB008_promoter	136	Synthetic	3183.229982	4685.405472
pMB020_promoter	137	Synthetic	1719.381656	4331.834596
pMB022_promoter	138	Synthetic	1432.727416	4224.994617
pMB089_promoter	139	Synthetic	1588.496061	4173.815385

pMB001_promoter	140	Synthetic	1344.15431	3710.844718
pMB051_promoter	141	Synthetic	1485.624649	3614.713536
pMB070_promoter	142	Synthetic	1148.371596	3494.100223
pMB074_promoter	143	Synthetic	1372.63931	3423.518954
pMB046_promoter	144	Synthetic	1070.100822	3325.635875
pMB071_promoter	145	Synthetic	829.6337908	2753.919683
pMB013_promoter	146	Synthetic	649.3261186	2255.268459
pMB080_promoter	147	Synthetic	584.6548742	2064.645666
pMB038_promoter	148	Synthetic	580.7846278	2043.19218
pMB060_promoter	149	Synthetic	575.1167059	1916.349295
pMB064_promoter	150	Synthetic	376.0013091	1390.767111
pMB058_promoter	151	Synthetic	292.4875916	1093.376645
pMB085_promoter	152	Synthetic	288.8178401	1085.653214
pMB081_promoter	153	Synthetic	278.7856475	670.8190328
pMB091_promoter	154	Synthetic	296.1836771	655.9453316
pMB027_promoter	155	Synthetic	262.5866089	601.1559107
pMB048_promoter	156	Synthetic	235.0853436	567.7578783
pMB055_promoter	157	Synthetic	189.6983907	499.3358783
pMB006_promoter	158	Synthetic	205.2252242	497.2257006
pMB012_promoter	159	Synthetic	118.144338	480.5859474
pMB014_promoter	160	Synthetic	119.9861606	474.4330596
pMB028_promoter	161	Synthetic	204.8853226	454.9894988
pMB059_promoter	162	Synthetic	101.7399612	387.0993978
pMB061_promoter	163	Synthetic	128.6404482	381.8073111
pMB043_promoter	164	Synthetic	139.0031576	380.0980602
pMB066_promoter	165	Synthetic	95.64407537	328.5708555
pMB079_promoter	166	Synthetic	105.9768914	312.5115498
pMB032_promoter	167	Synthetic	89.71910699	311.1818243
pMB068_promoter	168	Synthetic	97.56292371	276.8593245
pMB082_promoter	169	Synthetic	84.20594634	262.0127742
pMB030_promoter	170	Synthetic	92.28874802	255.1052708
pMB067_promoter	171	Synthetic	49.74078463	243.9642233

pMB050_promoter	172	Synthetic	69.03704925	223.6129199
pMB069_promoter	173	Synthetic	66.09043115	213.5258967
pMB017_promoter	174	Synthetic	50.43740198	213.0316985
pMB039_promoter	175	Synthetic	52.58639745	209.9432421
pMB011_promoter	176	Synthetic	63.05192023	194.8104082
pMB072_promoter	177	Synthetic	47.21948722	165.6730417
pMB016_promoter	178	Synthetic	47.20169015	163.6018648
pMB077_promoter	179	Synthetic	43.98192292	158.5866563
pMB047_promoter	180	Synthetic	30.55918215	121.2753356
pMB052_promoter	181	Synthetic	34.86596074	114.4169675
pMB090_promoter	182	Synthetic	14.01229264	55.03412277
pMB035_promoter	183	Synthetic	14.7011564	50.95895734
pMB073_promoter	184	Synthetic	11.79654115	47.47608915
pMB004_promoter	185	Synthetic	8.701914436	33.9085268
pMB054_promoter	186	Synthetic	8.934971498	32.06510507
pMB024_promoter	187	Synthetic	6.760860897	27.59246488
pMB007_promoter	188	Synthetic	11.5557009	25.41738039
pMB005_promoter	189	Synthetic	5.354276069	23.77742451
pMB003_promoter	190	Synthetic	6.389447374	22.7246451
pMB088_promoter	191	Synthetic	5.873205603	22.42214302
pMB065_promoter	192	Synthetic	5.44173707	20.60442307
pMB037_promoter	193	Synthetic	4.809680789	19.26108105
pMB009_promoter	194	Synthetic	3.591881545	13.6498423
pMB041_promoter	195	Synthetic	4.78150005	13.03957332
pMB036_promoter	196	Synthetic	2.508122856	11.98560161
pMB049_promoter	197	Synthetic	2.916151498	11.02944113
pMB044_promoter	198	Synthetic	2.719080697	10.47251128
pMB042_promoter	199	Synthetic	2.661753833	8.883733663
pMB086_promoter	200	Synthetic	2.33212712	8.316649305
pMB053_promoter	201	Synthetic	2.336160735	7.262981543
pMB057_promoter	202	Synthetic	1.518801989	6.27420435
pMB018_promoter	203	Synthetic	1.261000991	4.714509042

pMB002_promoter	204	Synthetic	1.290587949	4.473711828
pMB015_promoter	205	Synthetic	0.994888783	4.328541658
pMB087_promoter	206	Synthetic	0.94621781	2.900396821
pMB063_promoter	207	Synthetic	0.425085793	1.459419227

*Native promoters from Escherichia coli

Example 13– HTP Genomic Engineering –Testing integration of Promoter Swap Library of Table 1.4 into the genome of E. coli using vector 2 backbone

[0871] This example describes a proof of concept of the use of the vector 2 backbone from Example 11 in combination with a subset of promoters from the promoter library of Table 1.4 to drive integration of a single copy of a heterologous promoter-gene construct into the genome of *E. coli*.

[0872] For this Example, a set of plasmids will be built to insert fluorescent genes RFP and GFP at two loci (*nupG* and *asl*) in *E. coli* with a subset of 14 promoters from the set in Table 1.4. This will allow those promoters to be evaluated as a single copy integrated into the genome, rather than on low-copy replicating plasmids (see Example 12).

[0873] The plasmids will comprise homology arms flanking the RFP or GFP genes in order to facilitate integration into the genome of *E. coli* via “loop-in” as provided throughout this disclosure. The resulting strains will be tested for fluorescence, which will demonstrate that this subset of 14 promoters from Table 1.4 has been tested using a vector backbone described in Example 11 and can be used to insert a heterologous gene into the genome of *E. coli* using the methods described in this disclosure.

Example 14- HTP Genomic Engineering – Implementation of a PROSWP methods using promoter library derived from of Table 1.4.

[0874] The section below provides an illustrative implementation of the PRO swap HTP design strain improvement program tools of the present disclosure, as described in Examples 4 and 5. In this example, an *E. coli* strain was subjected to the PRO swap methods of the present disclosure in order to modulate the expression of genes in the *E. coli* genome. This example builds upon the results of Examples 12 and 13 in that this example illustrates the use of a promoter library comprising promoters from Table 1.4 in PROSWP methods of the present disclosure.

A. Promoter Swap

[0875] Promoter Swaps will be conducted as described in Example 4. Genes from the *E. coli* genome will be subjected to promoter swaps using the promoter library described in Example 13, which comprises a subset of 14 promoters from Table 1.4. The subset of 14 promoters to be used in this Example, will be selected based on their effect on gene expression as determined in Examples 12 and 13.

B. HTP engineering and High-Throughput Screening

[0876] HTP engineering of the promoter swaps will be conducted as described in Example 1 and 3. HTP screening of the resulting promoter swap strains will be conducted as described in Example 3. In total, 14 PRO swaps will be conducted. Finally, the impact of these modifications on production of products of interest will be tested.

Example 15– HTP Genomic Engineering – Implementation of a TERMINATOR Swap Library to Improve Strain Performance for Lycopene Production

[0877] The section below provides an illustrative implementation of the TERMINATOR swap HTP design strain improvement program tools of the present disclosure. In this example, an *E. coli* strain was subjected to the TERMINATOR swap methods of the present disclosure in order to affect host cell yield of lycopene.

[0878] Terminator swaps targeting genes in the lycopene biosynthetic pathway shown in Figure 59 were conducted using the terminator swap methods present throughout this disclosure. Constructs were designed as described below and recombination was mediated with CRISPR/Cas9 system. The terminators used for the terminator swaps in this example were the terminators found in Table 19.

Table 19-Terminators used in this Example for targeting genes in the lycopene biosynthetic pathway

Name	Description	Length (bp)
Spy	Terminator (SEQ ID NO. 225)	90
pheA	Terminator (SEQ ID NO. 226)	51
osmE	Terminator (SEQ ID NO. 227)	42
rpoH	Terminator (SEQ ID NO. 228)	41
vibE	Terminator (SEQ ID NO. 229)	71

Thr1_ABC	Terminator (SEQ ID NO. 230)	57
----------	-----------------------------	----

Construct design

[0879] A 20 base guide RNA near the target insertion sequence and adjacent to an NGG PAM sequence was identified to cut the genome at the desired position. The sequence intended to be inserted into the genome was flanked on both ends by 90 bases of homology, such that the homology would direct native sequence to be deleted or retained as desired. It should be noted that while recombination was facilitated by the CRISPR/Cas9 system in this Example, all strains can also be built by traditional single- and double-crossover homologous recombination methods as well as the Lambda Red system as described throughout the present disclosure. As such, each of the terminator swap library types is agnostic to the construction/recombination method.

Inoculate seed culture

[0880] A colony of editing base strain (W3110 pKD46-cas9 pLYC4) from a petri plate was picked and inoculated into larger volume of LB clin100 cmp25 and grew culture at 30°C with shaking for ~16 hours.

Prepare competent cells and transform

[0881] A 1:10 dilution of the overnight culture was prepared and measured the OD600 was measured. LB clin100 cmp25 was inoculated to an OD600 of 0.05 and grow at 30°C with shaking for ~2 hours.

[0882] After 2 hours, the OD600 was measured periodically until the induction target OD was reached, and upon reaching the target 20% arabinose was added to a final concentration of 0.2%.

[0883] Centrifuge the culture for 5 minutes at 5,000 x G at 4°C. Pour off the supernatant and resuspend to a final volume equivalent to the original culture volume.

[0884] Repeat step 7 to wash the cells for a third time.

[0885] After a 3 washes, the cells were pelleted and resuspended in 10% glycerol to ~1/250th the original culture volume. Prepare a 1:500 dilution of the resuspended cells was prepared and resuspended to a desired OD600 with an adequate volume for 40 uL of cells per transformation.

[0886] In a Framestar PCR plate (or microfuge tubes) 40 uL of cells were mixed with 100 ng of guide RNA plasmid, and ~4 uL of purified PCR product repair template. If using oligos for the repair template, the oligos were added to a final concentration of 2 uM.

[0887] The cells were electroporated and immediately resuspended in LB and recovered in a deep well plate 1 hour at 30°C with shaking.

[0888] The recovered cells were diluted and plated on LB agar containing 100 µg/ml kanamycin and 25 µg/ml chloramphenicol and at 30°C for 24 to 36 hours. Colonies were screened by colony PCR, sequencing, or phenotype screening. pKD46-cas9 plasmid can be cured by growth at 37°C or above, and pCRISPR2 can be cured by growth on 10% sucrose.

HTP engineering and High-Throughput Screening

[0889] HTP engineering of the swap combinations was conducted as described in Example 1 and 3 with the exception that CRISPR/Cas 9 was used to facilitate homologous recombination of constructs into the *E. coli* genome. HTP screening of the resulting promoter swap/terminator swap strains, promoter swap/degradation tag swap strains, promoter swap/solubility tag swap strains and promoter swap/terminator swap/degradation tag swap/solubility tag swap strains was conducted as described in Example 3. The results of the experiments are depicted in Figures 60 and 61.

[0890] As shown in Figure 60, terminator swaps at lycopene pathway targets *idi* and *ymgA* using the terminator TyjBE demonstrated decreased strain performance relative to the control, thus highlighting the utility of these library types for identifying critical pathway targets. This conclusion was further supported by the results shown in Figure 61, where terminator swaps were conducted on multiple lycopene pathway targets.

Example 16— HTP Genomic Engineering – Implementation of a TERMINATOR Swap Library or PRO Swap Library in combination with either a SOLUBILITY TAG swap Library or DEGRADATION Tag swap Library to Improve Strain Performance for Lycopene Production

[0891] The section below provides an illustrative implementation of the SOLUBILITY TAG swap and TERMINATOR swap HTP design strain improvement program tools of the present disclosure as well as a PRO swap and DEGRADATION TAG swap design strain improvement program tools of the present disclosure. In this example, an *E. coli* strain was subjected to the PRO swap in combination with DEGRADATION TAG swap methods of the present disclosure as well as SOLUBILITY TAG swap and TERMINATOR swap methods of the present disclosure in order to affect host cell yield of lycopene.

Promoter Swap/Terminator Swap/Solubility Tag Swap/Degradation Tag Swap

[0892] The Terminator swap was conducted as described in Example 15, while the Solubility Tag swap and Promoter Swap in combination with the degradation tag swaps were essentially conducted as described in Examples 4 and 5. Constructs were designed as described below and recombination was mediated with CRISPR/Cas9 system. The bicistronic promoters used for the promoter swaps in this example were from Mutalik et al., Nat Methods. 2013 Apr;10(4):354-60 and can be found in Table 20. It should be noted that any of the promoters provided herein could be used in the methods described below.

Table 20-Promoters used for promoter swap combinations in this example.

Name	Length	SEQ ID NO.
P3_BCD1	133	255
P4_BCD22	121	256
P7_BCD19	132	257
P8_BCD15	121	258
P11_BCD17	121	259
P13_BCD8	121	260

Construct design

[0893] A 20 base guide RNA near the target insertion sequence and adjacent to an NGG PAM sequence was identified to cut the genome at the desired position. The sequence intended to be inserted into the genome was flanked on both ends by 90 bases of homology, such that the homology would direct native sequence to be deleted or retained as desired. It should be noted that while recombination was facilitated by the CRISPR/Cas9 system in this Example, all strains can also be built by traditional single- and double-crossover homologous recombination methods as well as the Lambda Red system as described throughout the present disclosure. As such, each of these library types (promoter swap, protein solubility tag swap, protein degradation tag swap, and terminator swap) alone or in combination are agnostic to the construction/recombination method.

Inoculate seed culture

[0894] A colony of editing base strain (W3110 pKD46-cas9 pLYC4) from a petri plate was picked and inoculated into larger volume of LB clin100 cmp25 and grew culture at 30⁰C with shaking for ~16 hours.

Prepare competent cells and transform

[0895] A 1:10 dilution of the overnight culture was prepared and measured the OD600 was measured. LB clin100 cmp25 was inoculated to an OD600 of 0.05 and grow at 30°C with shaking for ~2 hours.

[0896] After 2 hours, the OD600 was measured periodically until the induction target OD was reached, and upon reaching the target 20% arabinose was added to a final concentration of 0.2%.

[0897] Centrifuge the culture for 5 minutes at 5,000 x G at 4°C. Pour off the supernatant and resuspend to a final volume equivalent to the original culture volume.

[0898] Repeat step 7 to wash the cells for a third time.

[0899] After a 3 washes, the cells were pelleted and resuspended in 10% glycerol to ~1/250th the original culture volume. Prepare a 1:500 dilution of the resuspended cells was prepared and resuspended to a desired OD600 with an adequate volume for 40 uL of cells per transformation.

[0900] In a Framestar PCR plate (or microfuge tubes) 40 uL of cells were mixed with 100 ng of guide RNA plasmid, and ~4 uL of purified PCR product repair template. If using oligos for the repair template, the oligos were added to a final concentration of 2 uM.

[0901] The cells were electroporated and immediately resuspended in LB and recover edin a deep well plate 1 hour at 30°C with shaking.

[0902] The recovered cells were diluted and plated on LB agar clin100 kan50 cmp25 and at 30°C 24 to 36 hours. Colonies were screened by colony PCR, sequencing, or phenotype screening. pKD46-cas9 plasmid can be cured by growth at 37°C or above, and pCRISPR2 can be cured by growth on 10% sucrose.

HTP engineering and High-Throughput Screening

[0903] HTP engineering of the swap combinations was conducted as described in Example 1 and 3 with the exception that CRISPR/Cas 9 was used to facilitate homologous recombination of constructs into the *E. coli* genome. HTP screening of the resulting promoter swap/terminator swap strains, promoter swap/degradation tag swap strains, promoter swap/solubility tag swap strains and promoter swap/terminator swap/degradation tag swap/solubility tag swap strains was conducted as described in Example 3.

[0904] It should be noted that the promoter P3_BCD1 was used in all strains where modification at the dxs locus was under study, unless otherwise noted to be different, such as P4_BCD22. At any locus other than dxs, the native promoter sequence was used unless otherwise noted. This means that a strain described as ssrA_LAA at the dxs locus, for example, also contained P3_BCD1,

but a strain described as *ssrA_LAA* at the *gdhA* locus used the native promoter sequence. The full contents of the strains tested in Table 21 below.

Table 21-Contents of strains generated in Example 16.

Modifications at <i>dxs</i> locus						
	*Control -P	P4_BCD2 2	<i>ssrA_AS</i> V	<i>ssrA_LA</i> A	Tspy	TyjbE
Promoter	P3_BCD 1	P4_BCD2 2	P3_BCD1	P3_BCD1	P3_BCD 1	P3_BCD 1
Solubility tag	none	none	none	none	none	none
Degradation tag	none	none	<i>ssrA_AS</i> V	<i>ssrA_LAA</i>	none	none
Terminator	native sequence	native sequence	native sequence	native sequence	Tspy	TyjbE
Modifications at <i>gdhA</i> locus						
	*Control- P	FH8	GB1	P4_BCD2 2	Tspy	TyjbE
Promoter	native sequence	native sequence	native sequence	P4_BCD2 2	native sequence	native sequence
Solubility tag	none	FH8	GB1	none	none	none
Degradation tag	none	none	none	none	none	none
Terminator	native sequence	native sequence	native sequence	native sequence	Tspy	TyjbE

[0905] The results of the experiments are summarized are depicted in Figures 62 and 63.

[0906] As shown in Figure 62, The *ssrA_LAA* degradation tag demonstrates improved strain performance relative to the control. This is unexpected as this strain is a combination of a PROSWP with a degradation tag at a single pathway target. The initial PROSWP is expected to increase protein abundance, and the degradation tag is expected to decrease protein abundance, thus demonstrating the utility of combinations of library types for tuning optimal strain performance. As shown in Figure 63, the solubility tag FH8 demonstrates improved strain

performance relative to the control, but the GB1 solubility tag does not, thus demonstrating the necessity for evaluating libraries of each modification type.

[0907] Overall, what this Example demonstrates is that while components of the present invention may be useful individually for systematic strain improvement, they can also be used synergistically with other approaches. For example, after improving mRNA stability through terminator modification, a strong promoter may be inserted to further increase protein production beyond the level of either approach alone. Likewise, this new elevated protein production level may be further improved through a protein solubility tag in conjunction with other modifications. When employed together and with previous approaches, the components of the present invention may allow for more robust and effective strain improvement for the production of a target molecule.

Further Numbered Embodiments of the Disclosure

[0908] Other subject matter contemplated by the present disclosure is set out in the following numbered embodiments:

1. A high-throughput (HTP) method of genomic engineering to evolve an *E. coli* microbe to acquire a desired phenotype, comprising:
 - a. perturbing the genomes of an initial plurality of *E. coli* microbes having the same genomic strain background, to thereby create an initial HTP genetic design *E. coli* strain library comprising individual *E. coli* strains with unique genetic variations;
 - b. screening and selecting individual strains of the initial HTP genetic design *E. coli* strain library for the desired phenotype;
 - c. providing a subsequent plurality of *E. coli* microbes that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent HTP genetic design *E. coli* strain library;
 - d. screening and selecting individual *E. coli* strains of the subsequent HTP genetic design *E. coli* strain library for the desired phenotype; and
 - e. repeating steps c)-d) one or more times, in a linear or non-linear fashion, until an *E. coli* microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new HTP genetic design *E. coli* strain library comprising individual *E. coli* strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual *E. coli* strains of a preceding HTP genetic design *E. coli* strain library.

2. The HTP method of genomic engineering according to embodiment 1, wherein the initial HTP genetic design *E. coli* strain library comprises at least one library selected from the group consisting of: a promoter swap microbial strain library, SNP swap microbial strain library, start/stop codon microbial strain library, optimized sequence microbial strain library, a terminator swap microbial strain library, a protein solubility tag microbial strain library, a protein degradation tag microbial strain library and any combination thereof.

3. The HTP method of genomic engineering according to embodiment 1, wherein the initial HTP genetic design *E. coli* strain library comprises a promoter swap microbial strain library.
4. The HTP method of genomic engineering according to embodiment 1, wherein the initial HTP genetic design *E. coli* strain library comprises a promoter swap microbial strain library that contains at least one bicistronic design (BCD) regulatory sequence.
 - 4.1 The HTP method of genomic engineering according to embodiment 4, wherein said BCD regulatory sequence comprises in order:
 - a. a promoter operably linked to;
 - b. a first ribosomal binding site (SD1);
 - c. a first cistronic sequence (Cis1);
 - d. a second ribosome binding site (SD2);wherein said BCD sequence is operably linked to a target gene sequence (Cis2).
 - 4.2 The HTP method of genomic engineering according to embodiment 4.1, wherein SD1 and SD2 each comprise a sequence of NNNGGANNN.
 - 4.3 The HTP method of genomic engineering according to embodiment 4.1 or 4.2, wherein SD1 and SD2 are different.
 - 4.4 The HTP method of genomic engineering according to any one of embodiments 4.1-4.3, wherein Cis1 comprises a stop codon, and wherein Cis2 comprises a start codon, and wherein the Cis1 stop codon and the Cis2 start codon overlap by at least 1 nucleotide.
 - 4.5 The HTP method of genomic engineering according to any one of embodiments 4.1-4.4, wherein SD2 is entirely embedded within Cis1.

5. The HTP method of genomic engineering according to any one of embodiments 1-4.5, wherein the initial HTP genetic design *E. coli* strain library comprises a SNP swap microbial strain library.
6. The HTP method of genomic engineering according to any one of embodiments 1-5, wherein the initial HTP genetic design *E. coli* strain library comprises a microbial strain library that comprises:
 - a. at least one polynucleotide encoding for a chimeric biosynthetic enzyme, wherein said chimeric biosynthetic enzyme comprises:
 - i. an enzyme involved in a regulatory pathway in *E. coli*;
 - ii. translationally fused to a DNA binding domain capable of binding a DNA binding site; and
 - b. at least one DNA scaffold sequence that comprises the DNA binding site corresponding to the DNA binding domain of the chimeric biosynthetic enzyme.
- 6.1 The HTP method of genomic engineering according to any one of embodiments 1-5, wherein the initial HTP genetic design *E. coli* strain library comprises a microbial strain library that comprises:
 - a. at least one polynucleotide encoding for a chimeric biosynthetic enzyme, wherein said chimeric biosynthetic enzyme comprises:
 - i. an enzyme involved in a regulatory pathway in *E. coli*;
 - ii. translationally fused to a protein binding domain capable of binding a recruitment peptide; and
 - b. at least one protein scaffold sequence that comprises the recruitment peptide corresponding to the protein binding domain of the chimeric biosynthetic enzyme.
7. The HTP method of genomic engineering according to any one of embodiments 1-6.1, wherein the subsequent HTP genetic design *E. coli* strain library is a full combinatorial strain library derived from the genetic variations in the initial HTP genetic design *E. coli* strain library.

8. The HTP method of genomic engineering according to any one of embodiments 1-6.1, wherein the subsequent HTP genetic design *E. coli* strain library is a subset of a full combinatorial strain library derived from the genetic variations in the initial HTP genetic design *E. coli* strain library.
9. The HTP method of genomic engineering according to any one of embodiments 1-6.1, wherein the subsequent HTP genetic design *E. coli* strain library is a full combinatorial strain library derived from the genetic variations in a preceding HTP genetic design *E. coli* strain library.
10. The HTP method of genomic engineering according to any one of embodiments 1-6.1, wherein the subsequent HTP genetic design *E. coli* strain library is a subset of a full combinatorial strain library derived from the genetic variations in a preceding HTP genetic design *E. coli* strain library.
11. The HTP method of genomic engineering according to any one of embodiments 1-10, wherein perturbing the genome comprises utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, and any combination thereof.
12. The HTP method of genomic engineering according to any one of embodiments 1-11, wherein the initial plurality of *E. coli* microbes comprise unique genetic variations derived from an industrial production *E. coli* strain.
13. The HTP method of genomic engineering according to any one of embodiments 1-12, wherein the initial plurality of *E. coli* microbes comprise industrial production strain microbes denoted S₁Gen₁ and any number of subsequent microbial generations derived therefrom denoted S_nGen_n.
14. A method for generating a SNP swap *E. coli* strain library, comprising the steps of:

- a. providing a reference *E. coli* strain and a second *E. coli* strain, wherein the second *E. coli* strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, which are not present in the reference *E. coli* strain; and
 - b. perturbing the genome of either the reference *E. coli* strain, or the second *E. coli* strain, to thereby create an initial SNP swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations corresponds to a single genetic variation selected from the plurality of identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.
- 15.** The method for generating a SNP swap *E. coli* strain library according to embodiment 14, wherein the genome of the reference *E. coli* strain is perturbed to add one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the second *E. coli* strain.
- 16.** The method for generating a SNP swap *E. coli* strain library according to embodiments 14 or 15, wherein the genome of the second *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the reference *E. coli* strain.
- 17.** The method for generating a SNP swap *E. coli* strain library according to any one of embodiments 14-16, wherein the resultant plurality of individual *E. coli* strains with unique genetic variations, together comprise a full combinatorial library of all the identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.
- 18.** The method for generating a SNP swap *E. coli* strain library according to any one of embodiments 14-16, wherein the resultant plurality of individual *E. coli* strains with unique genetic variations, together comprise a subset of a full combinatorial library of all

the identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.

- 19.** A method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a parental lineage *E. coli* strain and a production *E. coli* strain derived therefrom, wherein the production *E. coli* strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, not present in the parental lineage strain;
 - b. perturbing the genome of either the parental lineage *E. coli* strain, or the production *E. coli* strain, to create an initial library of *E. coli* strains, wherein each strain in the initial library comprises a unique genetic variation from the plurality of identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain;
 - c. screening and selecting individual strains of the initial library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
 - d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent library of *E. coli* strains;
 - e. screening and selecting individual strains of the subsequent library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
 - f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of

genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

20. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the initial library of *E. coli* strains is a full combinatorial library comprising all of the identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain.
21. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the initial library of *E. coli* strains is a subset of a full combinatorial library comprising a subset of the identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain.
22. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the subsequent library of *E. coli* strains is a full combinatorial library of the initial library.
23. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the subsequent library of *E. coli* strains is a subset of a full combinatorial library of the initial library.
24. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the subsequent library of *E. coli* strains is a full combinatorial library of a preceding library.
25. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to embodiment 19, wherein the subsequent library of *E. coli* strains is a subset of a full combinatorial library of a preceding library.
26. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-25, wherein the genome of the

parental lineage *E. coli* strain is perturbed to add one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the production *E. coli* strain.

27. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-25, wherein the genome of the production *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the parental lineage *E. coli* strain.
28. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-27, wherein perturbing the genome comprises utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, and combinations thereof.
29. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-28, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
30. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-28, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
31. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19, 29, and 30 wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric

productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.

32. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19, 29, and 30 wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
33. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-32, wherein the identified genetic variations further comprise artificial promoter swap genetic variations from a promoter swap library.
34. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 19-33, further comprising:
 - engineering the genome of at least one microbial strain of either:
 - the initial library of *E. coli* strains, or
 - a subsequent library of *E. coli* strains,
 - to comprise one or more promoters from a promoter ladder operably linked to an endogenous *E. coli* target gene.
35. A method for generating a promoter swap *E. coli* strain library, comprising the steps of:
 - a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain; and
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of

individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

36. The method for generating a promoter swap *E. coli* strain library according to embodiment 35, wherein at least one of the plurality of promoters comprises a bicistronic design (BCD) regulatory sequence.
37. A promoter swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain;
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
 - c. screening and selecting individual *E. coli* strains of the initial promoter swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
 - d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent promoter swap *E. coli* strain library;
 - e. screening and selecting individual *E. coli* strains of the subsequent promoter swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and

- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new promoter swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 37.1** The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 37, wherein at least one of the plurality of promoters comprises a bicistronic design (BCD) regulatory sequence.
- 38.** The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiments 37 or 37.1, wherein the subsequent promoter swap *E. coli* strain library is a full combinatorial library of the initial promoter swap *E. coli* strain library.
- 39.** The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiments 37 or 37.1, wherein the subsequent promoter swap *E. coli* strain library is a subset of a full combinatorial library of the initial promoter swap *E. coli* strain library.
- 40.** The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiments 37 or 37.1, wherein the subsequent promoter swap *E. coli* strain library is a full combinatorial library of a preceding promoter swap *E. coli* strain library.
- 41.** The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiments 37 or 37.1, wherein the subsequent promoter swap *E. coli* strain library is a subset of a full combinatorial library of a preceding promoter swap *E. coli* strain library.

42. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 37-41, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent promoter swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
43. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one embodiment 37-41, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent promoter swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
44. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 37, 42 and 43, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
45. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 37, 42 and 43, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
46. A method for generating a terminator swap *E. coli* strain library, comprising the steps of:
- providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain; and

- b. engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the terminators from the terminator ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.
47. A terminator swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain;
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the terminators from the terminator ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
 - c. screening and selecting individual *E. coli* strains of the initial terminator swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
 - d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent terminator swap *E. coli* strain library;
 - e. screening and selecting individual *E. coli* strains of the subsequent terminator swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and

- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new terminator swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 48.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 47, wherein the subsequent terminator swap *E. coli* strain library is a full combinatorial library of the initial terminator swap *E. coli* strain library.
- 49.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 47, wherein the subsequent terminator swap *E. coli* strain library is a subset of a full combinatorial library of the initial terminator swap *E. coli* strain library.
- 50.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 47, wherein the subsequent terminator swap *E. coli* strain library is a full combinatorial library of a preceding terminator swap *E. coli* strain library.
- 51.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 47, wherein the subsequent terminator swap *E. coli* strain library is a subset of a full combinatorial library of a preceding terminator swap *E. coli* strain library.
- 52.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 47-51, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent terminator swap *E.*

coli strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

53. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 47-51, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent terminator swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
54. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 47-51, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
55. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 47-51, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
56. A system for colocalizing biosynthetic enzymes from a biosynthetic pathway in an *E. coli* host cell, said system comprising:
 - a. two or more chimeric enzyme proteins involved in an enzymatic reaction, each chimeric enzyme protein comprising an enzyme portion coupled to a DNA binding domain portion;
 - b. a DNA scaffold comprising
 - i. one or more subunits, each subunit comprising two or more different DNA binding sites separated by at least one nucleic acid spacer;

wherein the chimeric enzyme proteins are recruited to the DNA scaffold by their coupled DNA binding domain portions, each of which bind at least one DNA binding site in the DNA scaffold.

- 57.** The system of embodiment 56, wherein the DNA binding domain portions of the chimeric enzyme proteins comprise zinc finger DNA binding domains and the DNA binding sites of the DNA scaffold comprise corresponding zinc finger binding sequences.
- 58.** The system of embodiments 56 or 57, wherein the enzyme portion of each of the two or more chimeric enzyme proteins is coupled to its respective DNA binding domain portion via a polypeptide linker sequence.
- 59.** The system of any one of embodiments 56-58, wherein the enzyme portion of each of the two or more chimeric enzyme proteins is coupled to its respective DNA binding domain portion via its amino-terminus or its carboxy-terminus.
- 60.** The system of any one of embodiments 56-59, wherein the two or more chimeric enzyme proteins comprise enzymes of an amino acid biosynthetic pathway.
- 61.** A bicistronic design regulatory (BCD) sequence, said BCD sequence comprising in order:
- a promoter operably linked to;
 - a first ribosomal binding site (SD1);
 - a first cistronic sequence (Cis1);
 - a second ribosome binding site (SD2);
- wherein said BCD sequence is operably linked to a target gene sequence (Cis2).
- 62.** The BCD of embodiment 61, wherein SD1 and SD2 each comprise a sequence of NNNGGANNN.
- 63.** The BCD of embodiment 61 or 62, wherein SD1 and SD2 are different.

- 64.** The BCD of any one of embodiments 61-63, wherein Cis1 comprises a stop codon, and wherein Cis2 comprises a start codon, and wherein the Cis1 stop codon and the Cis2 start codon overlap by at least 1 nucleotide.
- 65.** The BCD of any one of embodiments 61-63, wherein SD2 is entirely embedded within Cis1.
- 66.** A method for expressing two target gene proteins in a host organism, said method comprising the steps of:
- a. introducing into the host organism a first polynucleotide encoding for a first target gene protein, wherein said first polynucleotide is operably linked to a first bicistronic design regulatory (BCD) sequence according to any one of embodiments 61-65;
 - b. introducing into the host organism a second polynucleotide encoding for a second target gene protein, wherein said second polynucleotide is operably linked to a second BCD according to any one of embodiments 61-65;
- wherein the first and second BCDs are identical except for their respective Cis1 sequences, and wherein the target gene proteins are expressed in the host organism at a first and second expression level, respectively.
- 67.** The method of embodiment 66, wherein the first expression level is within 1.5 fold of the second expression level.
- 68.** The method of embodiment 66 or 67, wherein the first and second polynucleotides experience a lower level of homologous recombination in the host cell compared to a control host cell in which the first and second polynucleotides were expressed by identical BCDs.
- 69.** A method for generating a protein solubility tag swap *E. coli* strain library, comprising the steps of:

- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different solubility profiles in the base *E. coli* strain; and
- b. engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

70. A protein solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:

- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different expression profiles in the base *E. coli* strain;
- b. engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
- c. screening and selecting individual *E. coli* strains of the initial solubility tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
- d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent solubility tag swap *E. coli* strain library;
- e. screening and selecting individual *E. coli* strains of the subsequent solubility tag swap *E. coli* strain library for phenotypic performance improvements over the

reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and

- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new solubility tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

71. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 70, wherein the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of the initial solubility tag swap *E. coli* strain library.

72. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 70, wherein the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial solubility tag swap *E. coli* strain library.

73. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 70, wherein the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.

74. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 70, wherein the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.

75. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 70-74, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
76. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 70-74, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
77. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 70, 75 and 76, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
78. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 70, 75 and 76, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
79. A method for generating a protein degradation tag swap *E. coli* strain library, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different solubility profiles in the base *E. coli* strain; and

b. engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

80. A protein degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:

- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different expression profiles in the base *E. coli* strain;
- b. engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
- c. screening and selecting individual *E. coli* strains of the initial degradation tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
- d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent degradation tag swap *E. coli* strain library;
- e. screening and selecting individual *E. coli* strains of the subsequent degradation tag swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and

- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new degradation tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 81.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 80, wherein the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of the initial degradation tag swap *E. coli* strain library.
- 82.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 80, wherein the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial degradation tag swap *E. coli* strain library.
- 83.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 80, wherein the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.
- 84.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to embodiment 80, wherein the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.
- 85.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 80-84, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent

degradation tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.

- 86.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 80-84, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent degradation tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
- 87.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 80, 85 and 86, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
- 88.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of embodiments 80, 85 and 86, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
- 89.** A chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions

of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of lambda phage p_R promoter.

90. The chimeric synthetic promoter of embodiment 89, wherein nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site (RBS) portion of the lambda phage p_R promoter are selected from the nucleic acid sequences found in Table 1.5.
91. A chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene.
92. The chimeric synthetic promoter of embodiment 91, wherein nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene are selected from the nucleic acid sequences found in Table 1.5.
93. The chimeric synthetic promoter of any of embodiments 89-90, wherein the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 132-152, 159-160, 162, 165, 174-175, 188, 190, 199-201 or 207.
94. The chimeric synthetic promoter of any of embodiments 91-92, wherein the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 153-158, 161, 163-164, 166-173, 176-187, 189, 191-198 or 202-206.
95. The chimeric synthetic promoter of any of embodiments 89-94, wherein the microbial host cell is *E. coli*.
96. The chimeric synthetic promoter of embodiment 95, wherein the heterologous gene encodes a protein product of interest found in Table 2.
97. The chimeric synthetic promoter of embodiment 95, wherein the heterologous gene is a gene that is part of a lysine biosynthetic pathway.

- 98.** The chimeric synthetic promoter of embodiment 97, wherein the heterologous gene is selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene, the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene.
- 99.** The chimeric synthetic promoter of embodiment 95, wherein the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.
- 100.** The chimeric synthetic promoter of embodiment 99, wherein the heterologous gene is selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA* gene, the *ispB* gene, the *crtE* gene, the *crtB* gene, the *crtI* gene, the *crtY* gene, the *ymgA* gene, the *dxr* gene, the *elbA* gene, the *gdhA* gene, the *appY* gene, the *elbB* gene, or the *ymgB* gene.
- 101.** The chimeric synthetic promoter of embodiment 95, wherein the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.
- 102.** The chimeric synthetic promoter of embodiment 101, wherein the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).
- 103.** A heterologous gene operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.
- 104.** The heterologous gene of embodiment 103, wherein the heterologous gene encodes a protein product of interest found in Table 2.
- 105.** The heterologous gene of embodiment 103, wherein the heterologous gene is a gene that is part of a lysine biosynthetic pathway.

- 106.** The heterologous gene of embodiment 105, wherein the heterologous gene is selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene, the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene.
- 107.** The heterologous gene of embodiment 103, wherein the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.
- 108.** The heterologous gene of embodiment 107, wherein the heterologous gene is selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA* gene, the *ispB* gene, the *crtE* gene, the *crtB* gene, the *crtI* gene, the *crtY* gene, the *ymgA* gene, the *dxr* gene, the *elbA* gene, the *gdhA* gene, the *appY* gene, the *elbB* gene, or the *ymgB* gene.
- 109.** The heterologous gene of embodiment 103, wherein the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.
- 110.** The heterologous gene of embodiment 109, wherein the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).

SEQUENCES OF THE DISCLOSURE WITH SEQ ID NO IDENTIFIERS

NAME (SHORT NAME)	SOURCE	NUCLEIC ACID SEQ ID NO.	AMINO ACID SEQ ID NO.	DESCRIPTION
Pcg0007_lib_39 (P1)		1		
Pcg0007 (P2)		2		
Pcg1860 (P3)		3		
Pcg0755 (P4)		4		
Pcg0007_265 (P5)		5		
Pcg3381 (P6)		6		
Pcg0007_119 (P7)		7		
Pcg3121 (P8)		8		
cg0001 (T1)		9		
cg0007 (T2)		10		
cg0371 (T3)		11		
cg0480 (T4)		12		
cg0494 (T5)		13		
cg0564 (T6)		14		
cg0610 (T7)		15		
cg0695 (T8)		16		
Cis1 nucleotide sequence		17		
Serine-rich linker		18		
Serine-rich linker		19		
Serine-rich linker		20		
Alanine-rich linker		21		
linker		22		

linker		23		
linker		24		
linker		25		
Zif268		26		DNA Binding Domain Sequence
PBSII		27		DNA Binding Domain Sequence
ZFa		28		DNA Binding Domain Sequence
ZFb		29		DNA Binding Domain Sequence
ZFc		30		DNA Binding Domain Sequence
Tyr123		31		DNA Binding Domain Sequence
Tyr456		32		DNA Binding Domain Sequence
Blues Zinc Finger		33		DNA Binding Domain Sequence
Jazz Zinc Finger		34		DNA Binding Domain Sequence
Bagly Zinc Finger		35		DNA Binding Domain Sequence

Bagly Zinc Finger Binding Site		36		DNA Binding Sequence (5'→3')
Gli1		37		DNA Binding Domain Sequence
Gli1 binding site		38		DNA Binding Sequence (5'→3')
HIVC zinc finger		39		DNA Binding Domain Sequence
B3 zinc finger		40		DNA Binding Domain Sequence
N1 zinc finger		41		DNA Binding Domain Sequence
Sp-1 first zinc finger		42		DNA Binding Domain Sequence
Sp-1 second finger		43		DNA Binding Domain Sequence
Class I SH3 protein binding domain			44	
Class II SH3 Protein Binding Site			45	
SH3 protein binding site			46	
SH3 recruitment peptide			47	
SH3 recruitment Peptide			48	
SH3 recruitment peptide			49	
PDZ protein binding domain			50	
PDZ protein binding domain			51	

PDZ protein binding domain			52	
PDZ protein binding domain			53	
PDZ protein binding domain			54	
PDZ recruitment peptide			55	
PDZ recruitment peptide			56	
PDZ recruitment peptide			57	
PDZ recruitment peptide			58	
PDZ recruitment peptide			59	
PDZ recruitment peptide			60	
PDZ recruitment peptide			61	
GBD protein binding domain			62	
GBD protein binding domain			63	
GBD protein binding domain			64	
GBD recruitment peptide			65	
Leucine Zipper binding domain			66	
Leucine Zipper binding domain			67	
Leucine Zipper binding domain			68	
Bicistronic Design (BCD) regulatory sequence			69	<i>see Figure 43</i>
Bicistronic Design (BCD) regulatory sequence			70	<i>see Figure 43</i>
b0904_promoter	<i>E. coli</i>	71		
b2405_promoter	<i>E. coli</i>	72		
b0096_promoter	<i>E. coli</i>	73		
b0576_promoter	<i>E. coli</i>	74		
b2017_promoter	<i>E. coli</i>	75		
b1278_promoter	<i>E. coli</i>	76		
b4255_promoter	<i>E. coli</i>	77		
b0786_promoter	<i>E. coli</i>	78		
b0605_promoter	<i>E. coli</i>	79		
b1824_promoter	<i>E. coli</i>	80		
b1061_promoter	<i>E. coli</i>	81		

b0313_promoter	<i>E. coli</i>	82		
b0814_promoter	<i>E. coli</i>	83		
b4133_promoter	<i>E. coli</i>	84		
b4268_promoter	<i>E. coli</i>	85		
b0345_promoter	<i>E. coli</i>	86		
b2096_promoter	<i>E. coli</i>	87		
b1277_promoter	<i>E. coli</i>	88		
b1646_promoter	<i>E. coli</i>	89		
b4177_promoter	<i>E. coli</i>	90		
b0369_promoter	<i>E. coli</i>	91		
b1920_promoter	<i>E. coli</i>	92		
b3742_promoter	<i>E. coli</i>	93		
b3929_promoter	<i>E. coli</i>	94		
b3743_promoter	<i>E. coli</i>	95		
b1613_promoter	<i>E. coli</i>	96		
b1749_promoter	<i>E. coli</i>	97		
b2478_promoter	<i>E. coli</i>	98		
b0031_promoter	<i>E. coli</i>	99		
b2414_promoter	<i>E. coli</i>	100		
b1183_promoter	<i>E. coli</i>	101		
b0159_promoter	<i>E. coli</i>	102		
b2837_promoter	<i>E. coli</i>	103		
b3237_promoter	<i>E. coli</i>	104		
b3778_promoter	<i>E. coli</i>	105		
b2349_promoter	<i>E. coli</i>	106		
b1434_promoter	<i>E. coli</i>	107		
b3617_promoter	<i>E. coli</i>	108		
b0237_promoter	<i>E. coli</i>	109		
b4063_promoter	<i>E. coli</i>	110		
b0564_promoter	<i>E. coli</i>	111		
b0019_promoter	<i>E. coli</i>	112		
b2375_promoter	<i>E. coli</i>	113		

b1187_promoter	<i>E. coli</i>	114		
b2388_promoter	<i>E. coli</i>	115		
b1051_promoter	<i>E. coli</i>	116		
b4241_promoter	<i>E. coli</i>	117		
b4054_promoter	<i>E. coli</i>	118		
b2425_promoter	<i>E. coli</i>	119		
b0995_promoter	<i>E. coli</i>	120		
b1399_promoter	<i>E. coli</i>	121		
b3298_promoter	<i>E. coli</i>	122		
b2114_promoter	<i>E. coli</i>	123		
b2779_promoter	<i>E. coli</i>	124		
b1114_promoter	<i>E. coli</i>	125		
b3730_promoter	<i>E. coli</i>	126		
b3025_promoter	<i>E. coli</i>	127		
b0850_promoter	<i>E. coli</i>	128		
b2365_promoter	<i>E. coli</i>	129		
b4117_promoter	<i>E. coli</i>	130		
b2213_promoter	<i>E. coli</i>	131		
pMB029_promoter	Synthetic	132		
pMB023_promoter	Synthetic	133		
pMB025_promoter	Synthetic	134		
pMB019_promoter	Synthetic	135		
pMB008_promoter	Synthetic	136		
pMB020_promoter	Synthetic	137		
pMB022_promoter	Synthetic	138		
pMB089_promoter	Synthetic	139		
pMB001_promoter	Synthetic	140		
pMB051_promoter	Synthetic	141		
pMB070_promoter	Synthetic	142		
pMB074_promoter	Synthetic	143		
pMB046_promoter	Synthetic	144		
pMB071_promoter	Synthetic	145		

pMB013_promoter	Synthetic	146		
pMB080_promoter	Synthetic	147		
pMB038_promoter	Synthetic	148		
pMB060_promoter	Synthetic	149		
pMB064_promoter	Synthetic	150		
pMB058_promoter	Synthetic	151		
pMB085_promoter	Synthetic	152		
pMB081_promoter	Synthetic	153		
pMB091_promoter	Synthetic	154		
pMB027_promoter	Synthetic	155		
pMB048_promoter	Synthetic	156		
pMB055_promoter	Synthetic	157		
pMB006_promoter	Synthetic	158		
pMB012_promoter	Synthetic	159		
pMB014_promoter	Synthetic	160		
pMB028_promoter	Synthetic	161		
pMB059_promoter	Synthetic	162		
pMB061_promoter	Synthetic	163		
pMB043_promoter	Synthetic	164		
pMB066_promoter	Synthetic	165		
pMB079_promoter	Synthetic	166		
pMB032_promoter	Synthetic	167		
pMB068_promoter	Synthetic	168		
pMB082_promoter	Synthetic	169		
pMB030_promoter	Synthetic	170		
pMB067_promoter	Synthetic	171		
pMB050_promoter	Synthetic	172		
pMB069_promoter	Synthetic	173		
pMB017_promoter	Synthetic	174		
pMB039_promoter	Synthetic	175		
pMB011_promoter	Synthetic	176		
pMB072_promoter	Synthetic	177		

pMB016_promoter	Synthetic	178		
pMB077_promoter	Synthetic	179		
pMB047_promoter	Synthetic	180		
pMB052_promoter	Synthetic	181		
pMB090_promoter	Synthetic	182		
pMB035_promoter	Synthetic	183		
pMB073_promoter	Synthetic	184		
pMB004_promoter	Synthetic	185		
pMB054_promoter	Synthetic	186		
pMB024_promoter	Synthetic	187		
pMB007_promoter	Synthetic	188		
pMB005_promoter	Synthetic	189		
pMB003_promoter	Synthetic	190		
pMB088_promoter	Synthetic	191		
pMB065_promoter	Synthetic	192		
pMB037_promoter	Synthetic	193		
pMB009_promoter	Synthetic	194		
pMB041_promoter	Synthetic	195		
pMB036_promoter	Synthetic	196		
pMB049_promoter	Synthetic	197		
pMB044_promoter	Synthetic	198		
pMB042_promoter	Synthetic	199		
pMB086_promoter	Synthetic	200		
pMB053_promoter	Synthetic	201		
pMB057_promoter	Synthetic	202		
pMB018_promoter	Synthetic	203		
pMB002_promoter	Synthetic	204		
pMB015_promoter	Synthetic	205		
pMB087_promoter	Synthetic	206		
pMB063_promoter	Synthetic	207		
Distal portion of synthetic promoter	Phage λ	208		P_R promoter

Core portion of synthetic promoter	Phage λ	209		P_R promoter
Distal portion of synthetic promoter	Phage λ	210		P_L promoter
5'UTR/RBS portion of synthetic promoter	Phage λ	211		P_R promoter
5'UTR/RBS portion of synthetic promoter	<i>E. coli</i>	212		promoter of <i>acs</i> gene
Ori_Plsm27		213		
vector backbone 1		214		
vector backbone 2		215		
vector backbone 3		216		
vector backbone 4		217		
Insulator 1		218		see Table 15
Insulator 2		219		see Table 15
T1		220		from Orosz et al., Eur J Biochem. 1991 Nov 1;201(3):653-9; see Table 15
B0015		221		see Table 15
SacB promoter		222		see Table 15
PheS promoter & sequence		223		see Table 15
Tsod terminator	<i>C. glutamicum</i>	224		
Spy	<i>E. coli</i>	225		Terminator Sequence
pheA	<i>E. coli</i>	226		Terminator Sequence
osmE	<i>E. coli</i>	227		Terminator Sequence
rpoH	<i>E. coli</i>	228		Terminator Sequence
vibE	<i>E. coli</i>	229		Terminator Sequence

Thr1_ABC	<i>E. coli</i>	230		Terminator Sequence
GB1 (PST1)	Streptococcus sp.	231	235	IgG domain B1 of Protein G
FH8 (PST2)	<i>F. hepatica</i>	232	236	Fasciola hepatica 8-kDa antigen
Ubiquitin (PST3)		233	237	
SUMO (PST4)	<i>Homo sapiens</i>	234	238	Small ubiquitin modified
ssrA_LAA (PDT1)	<i>E. coli</i>	239	247	native
ssrA_LVA (PDT2)	<i>E. coli</i>	240	248	mutant
ssrA_AAV (PDT3)	<i>E. coli</i>	241	249	mutant
ssrA_ASV (PDT4)	<i>E. coli</i>	242	250	mutant
ftsH-cII89-97 (PDT5)	<i>E. coli</i>	243	251	native
cI108 (PDT6)	<i>E. coli</i>	244	252	native
sul20 (PDT7)	<i>E. coli</i>	245	253	native
β 20 (PDT8)	<i>E. coli</i>	246	254	native
P3_BCD1	Artificial	255		
P4_BCD22	Artificial	256		
P7_BCD19	Artificial	257		
P8_BCD15	Artificial	258		
P11_BCD17	Artificial	259		
P13_BCD8	Artificial	260		

INCORPORATION BY REFERENCE

All references, articles, publications, patents, patent publications, and patent applications cited herein are incorporated by reference in their entireties for all purposes. However, mention of any reference, article, publication, patent, patent publication, and patent application cited herein is not, and should not be taken as an acknowledgment or any form of suggestion that they constitute valid prior art or form part of the common general knowledge in any country in the world.

In addition, the following particular applications are incorporated herein by reference: U.S. Application No. 15/396,230 (U.S. Pub. No. US 2017/0159045 A1) filed on December 30, 2016; PCT/US2016/065465 (WO 2017/100377 A1) filed on December 07, 2016; U.S. App. No. 15/140,296 (US 2017/0316353 A1) filed on April 27, 2016; PCT/US2017/029725 (WO 2017/189784 A1) filed on April 26, 2017; PCT/US2016/065464 (WO 2017/100376 A2) filed on December 07, 2016; U.S. Prov. App. No. 62/431,409 filed on December 07, 2016; U.S. Prov. App. No. 62/264,232 filed on December 07, 2015; and U.S. Prov. App. No. 62/368,786 filed on July 29, 2016.

CLAIMS

What is claimed is:

1. A high-throughput (HTP) method of genomic engineering to evolve an *E. coli* microbe to acquire a desired phenotype, comprising:
 - a. perturbing the genomes of an initial plurality of *E. coli* microbes having the same genomic strain background, to thereby create an initial HTP genetic design *E. coli* strain library comprising individual *E. coli* strains with unique genetic variations;
 - b. screening and selecting individual strains of the initial HTP genetic design *E. coli* strain library for the desired phenotype;
 - c. providing a subsequent plurality of *E. coli* microbes that each comprise a unique combination of genetic variation, said genetic variation selected from the genetic variation present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent HTP genetic design *E. coli* strain library;
 - d. screening and selecting individual *E. coli* strains of the subsequent HTP genetic design *E. coli* strain library for the desired phenotype; and
 - e. repeating steps c)-d) one or more times, in a linear or non-linear fashion, until an *E. coli* microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new HTP genetic design *E. coli* strain library comprising individual *E. coli* strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual *E. coli* strains of a preceding HTP genetic design *E. coli* strain library.

2. The HTP method of genomic engineering according to claim 1, wherein the initial HTP genetic design *E. coli* strain library comprises at least one library selected from the group consisting of: a promoter swap microbial strain library, SNP swap microbial strain library, start/stop codon microbial strain library, optimized sequence microbial strain library, a terminator swap microbial strain library, a protein solubility tag microbial strain library, a protein degradation tag microbial strain library and any combination thereof.

3. The HTP method of genomic engineering according to claim 1, wherein the initial HTP genetic design *E. coli* strain library comprises a promoter swap microbial strain library.
4. The HTP method of genomic engineering according to claim 1 or 2, wherein the initial HTP genetic design *E. coli* strain library comprises a promoter swap microbial strain library that contains at least one bicistronic design (BCD) regulatory sequence.
5. The HTP method of genomic engineering according to claim 1, wherein the initial HTP genetic design *E. coli* strain library comprises a SNP swap microbial strain library.
6. The HTP method of genomic engineering according to claim 1 or 2, wherein the initial HTP genetic design *E. coli* strain library comprises a microbial strain library that comprises:
 - a. at least one polynucleotide encoding for a chimeric biosynthetic enzyme, wherein said chimeric biosynthetic enzyme comprises an enzyme involved in a regulatory pathway in *E. coli* translationally fused to a DNA binding domain capable of binding a DNA binding site; and
 - b. at least one DNA scaffold sequence that comprises the DNA binding site.
7. The HTP method of genomic engineering according to claim 1, wherein the subsequent HTP genetic design *E. coli* strain library is a full combinatorial strain library derived from the genetic variations in the initial HTP genetic design *E. coli* strain library.
8. The HTP method of genomic engineering according to claim 1, wherein the subsequent HTP genetic design *E. coli* strain library is a subset of a full combinatorial strain library derived from the genetic variations in the initial HTP genetic design *E. coli* strain library.
9. The HTP method of genomic engineering according to claim 1, wherein the subsequent HTP genetic design *E. coli* strain library is a full combinatorial strain library derived from the genetic variations in a preceding HTP genetic design *E. coli* strain library.

10. The HTP method of genomic engineering according to claim 1, wherein the subsequent HTP genetic design *E. coli* strain library is a subset of a full combinatorial strain library derived from the genetic variations in a preceding HTP genetic design *E. coli* strain library.
11. The HTP method of genomic engineering according to claim 1, wherein perturbing the genome comprises utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, and any combination thereof.
12. The HTP method of genomic engineering according to claim 1, wherein the initial plurality of *E. coli* microbes comprise unique genetic variations derived from an industrial production *E. coli* strain.
13. The HTP method of genomic engineering according to claim 1, wherein the initial plurality of *E. coli* microbes comprise industrial production strain microbes denoted S_1Gen_1 and any number of subsequent microbial generations derived therefrom denoted S_nGen_n .
14. A method for generating a SNP swap *E. coli* strain library, comprising the steps of:
 - a. providing a reference *E. coli* strain and a second *E. coli* strain, wherein the second *E. coli* strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, which are not present in the reference *E. coli* strain; and
 - b. perturbing the genome of either the reference *E. coli* strain, or the second *E. coli* strain, to thereby create an initial SNP swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual strains, wherein each of said unique genetic variations corresponds to a single genetic variation selected from the plurality of identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.
15. The method for generating a SNP swap *E. coli* strain library according to claim 14, wherein the genome of the reference *E. coli* strain is perturbed to add one or more of the identified

single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the second *E. coli* strain.

- 16.** The method for generating a SNP swap *E. coli* strain library according to claim 14, wherein the genome of the second *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the reference *E. coli* strain.
- 17.** The method for generating a SNP swap *E. coli* strain library according to any one of claims 14-16, wherein the resultant plurality of individual *E. coli* strains with unique genetic variations, together comprise a full combinatorial library of all the identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.
- 18.** The method for generating a SNP swap *E. coli* strain library according to any one of claims 14-16, wherein the resultant plurality of individual *E. coli* strains with unique genetic variations, together comprise a subset of a full combinatorial library of all the identified genetic variations between the reference *E. coli* strain and the second *E. coli* strain.
- 19.** A method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
 - a. providing a parental lineage *E. coli* strain and a production *E. coli* strain derived therefrom, wherein the production *E. coli* strain comprises a plurality of identified genetic variations selected from single nucleotide polymorphisms, DNA insertions, and DNA deletions, not present in the parental lineage strain;
 - b. perturbing the genome of either the parental lineage *E. coli* strain, or the production *E. coli* strain, to create an initial library of *E. coli* strains, wherein each strain in the initial library comprises a unique genetic variation from the plurality of identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain;

- c. screening and selecting individual strains of the initial library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
 - d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent library of *E. coli* strains;
 - e. screening and selecting individual strains of the subsequent library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
 - f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 20.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the initial library of *E. coli* strains is a full combinatorial library comprising all of the identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain.
- 21.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the initial library of *E. coli* strains is a subset of a full combinatorial library comprising a subset of the identified genetic variations between the parental lineage *E. coli* strain and the production *E. coli* strain.

22. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the subsequent library of *E. coli* strains is a full combinatorial library of the initial library.
23. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the subsequent library of *E. coli* strains is a subset of a full combinatorial library of the initial library.
24. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the subsequent library of *E. coli* strains is a full combinatorial library of a preceding library.
25. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the subsequent library of *E. coli* strains is a subset of a full combinatorial library of a preceding library.
26. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the genome of the parental lineage *E. coli* strain is perturbed to add one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are found in the production *E. coli* strain.
27. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the genome of the production *E. coli* strain is perturbed to remove one or more of the identified single nucleotide polymorphisms, DNA insertions, or DNA deletions, which are not found in the parental lineage *E. coli* strain.
28. The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to any one of claims 19-25, wherein perturbing the genome comprises utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, and combinations thereof.

- 29.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
- 30.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
- 31.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
- 32.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
- 33.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, wherein the identified genetic variations further comprise artificial promoter swap genetic variations from a promoter swap library.

- 34.** The method for rehabilitating and improving the phenotypic performance of a production *E. coli* strain according to claim 19, further comprising:
- engineering the genome of at least one microbial strain of either:
 - the initial library of *E. coli* strains, or
 - a subsequent library of *E. coli* strains,
 - to comprise one or more promoters from a promoter ladder operably linked to an endogenous *E. coli* target gene.
- 35.** A method for generating a promoter swap *E. coli* strain library, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain; and
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.
- 36.** The method for generating a promoter swap *E. coli* strain library according to claim 35, wherein at least one of the plurality of promoters comprises a bicistronic design (BCD) regulatory sequence.
- 37.** A promoter swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a promoter ladder, wherein said promoter ladder comprises a plurality of promoters exhibiting different expression profiles in the base *E. coli* strain;
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial promoter swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of

individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the promoters from the promoter ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;

- c. screening and selecting individual *E. coli* strains of the initial promoter swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
- d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent promoter swap *E. coli* strain library;
- e. screening and selecting individual *E. coli* strains of the subsequent promoter swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new promoter swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

38. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the subsequent promoter swap *E. coli* strain library is a full combinatorial library of the initial promoter swap *E. coli* strain library.

39. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the subsequent promoter swap *E. coli* strain library is a subset of a full combinatorial library of the initial promoter swap *E. coli* strain library.

40. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the subsequent promoter swap *E. coli* strain library is a full combinatorial library of a preceding promoter swap *E. coli* strain library.
41. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the subsequent promoter swap *E. coli* strain library is a subset of a full combinatorial library of a preceding promoter swap *E. coli* strain library.
42. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 37-41, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent promoter swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
43. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 37-41, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent promoter swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
44. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
45. The promoter swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 37, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest

selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.

46. A method for generating a terminator swap *E. coli* strain library, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain; and
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the terminators from the terminator ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.
47. A terminator swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a terminator ladder, wherein said terminator ladder comprises a plurality of terminators exhibiting different expression profiles in the base *E. coli* strain;
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial terminator swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the terminators from the terminator ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
 - c. screening and selecting individual *E. coli* strains of the initial terminator swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;

- d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent terminator swap *E. coli* strain library;
 - e. screening and selecting individual *E. coli* strains of the subsequent terminator swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
 - f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new terminator swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 48.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the subsequent terminator swap *E. coli* strain library is a full combinatorial library of the initial terminator swap *E. coli* strain library.
- 49.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the subsequent terminator swap *E. coli* strain library is a subset of a full combinatorial library of the initial terminator swap *E. coli* strain library.
- 50.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the subsequent terminator swap *E. coli* strain library is a full combinatorial library of a preceding terminator swap *E. coli* strain library.
- 51.** The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the subsequent terminator swap *E. coli* strain

library is a subset of a full combinatorial library of a preceding terminator swap *E. coli* strain library.

52. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 47-51, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent terminator swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
53. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 47-51, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent terminator swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
54. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
55. The terminator swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 47, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
56. A system for colocalizing biosynthetic enzymes from a biosynthetic pathway in an *E. coli* host cell, said system comprising:

- a. two or more chimeric enzyme proteins involved in an enzymatic reaction, each chimeric enzyme protein comprising an enzyme portion coupled to a DNA binding domain portion; and
 - b. a DNA scaffold comprising
 - i. one or more subunits, each subunit comprising two or more different DNA binding sites separated by at least one nucleic acid spacer;wherein the chimeric enzyme proteins are recruited to the DNA scaffold by their coupled DNA binding domain portions, each of which bind at least one DNA binding site in the DNA scaffold.
- 57.** The system of claim 56, wherein the DNA binding domain portions of the chimeric enzyme proteins comprise zinc finger DNA binding domains and the DNA binding sites of the DNA scaffold comprise corresponding zinc finger binding sequences.
- 58.** The system of claim 56, wherein the enzyme portion of each of the two or more chimeric enzyme proteins is coupled to its respective DNA binding domain portion via a polypeptide linker sequence.
- 59.** The system of claim 56, wherein the enzyme portion of each of the two or more chimeric enzyme proteins is coupled to its respective DNA binding domain portion via its amino-terminus or its carboxy-terminus.
- 60.** The system of claim 56, wherein the two or more chimeric enzyme proteins comprise enzymes of an amino acid biosynthetic pathway.
- 61.** A bicistronic design regulatory (BCD) sequence, said BCD sequence comprising in order:
 - a. a promoter operably linked to;
 - b. a first ribosomal binding site (SD1);
 - c. a first cistronic sequence (Cis1);
 - d. a second ribosome binding site (SD2);wherein said BCD sequence is operably linked to a target gene sequence (Cis2).

- 62.** The BCD of claim 61, wherein SD1 and SD2 each comprise a sequence of NNNGGANNN.
- 63.** The BCD of claim 61, wherein SD1 and SD2 are different.
- 64.** The BCD of claim 61, wherein Cis1 comprises a stop codon, and wherein Cis2 comprises a start codon, and wherein the Cis1 stop codon and the Cis2 start codon overlap by at least 1 nucleotide.
- 65.** The BCD of claim 61, wherein SD2 is entirely embedded within Cis1.
- 66.** A method for expressing two target gene proteins in a host organism, said method comprising the steps of:
- a. introducing into the host organism a first polynucleotide encoding for a first target gene protein, wherein said first polynucleotide is operably linked to a first bicistronic design regulatory (BCD) sequence according to claim 61; and
 - b. introducing into the host organism a second polynucleotide encoding for a second target gene protein, wherein said second polynucleotide is operably linked to a second BCD according to claim 61;
- wherein the first and second BCDs are identical except for their respective Cis1 sequences, and wherein the target gene proteins are expressed in the host organism at a first and second expression level, respectively.
- 67.** The method of claim 66, wherein the first expression level is within 1.5 fold of the second expression level.
- 68.** The method of claim 66, wherein the first and second polynucleotides experience a lower level of homologous recombination in the host cell compared to a control host cell in which the first and second polynucleotides were expressed by identical BCDs.

69. A method for generating a protein solubility tag swap *E. coli* strain library, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different solubility profiles in the base *E. coli* strain; and
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.
70. A protein solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:
- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a solubility tag ladder, wherein said solubility tag ladder comprises a plurality of solubility tags exhibiting different expression profiles in the base *E. coli* strain;
 - b. engineering the genome of the base *E. coli* strain, to thereby create an initial solubility tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the solubility tags from the solubility tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
 - c. screening and selecting individual *E. coli* strains of the initial solubility tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
 - d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent solubility tag swap *E. coli* strain library;

- e. screening and selecting individual *E. coli* strains of the subsequent solubility tag swap *E. coli* strain library for phenotypic performance improvements over the reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
 - f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new solubility tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.
- 71.** The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 70, wherein the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of the initial solubility tag swap *E. coli* strain library.
- 72.** The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 70, wherein the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial solubility tag swap *E. coli* strain library.
- 73.** The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 70, wherein the subsequent solubility tag swap *E. coli* strain library is a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.
- 74.** The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 70, wherein the subsequent solubility tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding solubility tag swap *E. coli* strain library.

75. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 70-74, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
76. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 70-74, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent solubility tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
77. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 70-74, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
78. The solubility tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 70-74, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
79. A method for generating a protein degradation tag swap *E. coli* strain library, comprising the steps of:

- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different solubility profiles in the base *E. coli* strain; and
- b. engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain.

80. A protein degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain, comprising the steps of:

- a. providing a plurality of target genes endogenous to a base *E. coli* strain, and a degradation tag ladder, wherein said degradation tag ladder comprises a plurality of degradation tags exhibiting different expression profiles in the base *E. coli* strain;
- b. engineering the genome of the base *E. coli* strain, to thereby create an initial degradation tag swap *E. coli* strain library comprising a plurality of individual *E. coli* strains with unique genetic variations found within each strain of said plurality of individual *E. coli* strains, wherein each of said unique genetic variations comprises one or more of the degradation tags from the degradation tag ladder operably linked to one of the target genes endogenous to the base *E. coli* strain;
- c. screening and selecting individual *E. coli* strains of the initial degradation tag swap *E. coli* strain library for phenotypic performance improvements over a reference *E. coli* strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
- d. providing a subsequent plurality of *E. coli* microbes that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual *E. coli* strains screened in the preceding step, to thereby create a subsequent degradation tag swap *E. coli* strain library;
- e. screening and selecting individual *E. coli* strains of the subsequent degradation tag swap *E. coli* strain library for phenotypic performance improvements over the

reference *E. coli* strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and

- f. repeating steps d)-e) one or more times, in a linear or non-linear fashion, until an *E. coli* strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production *E. coli* strain, wherein each subsequent iteration creates a new degradation tag swap *E. coli* strain library of microbial strains, where each strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual *E. coli* strains of a preceding library.

- 81.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 80, wherein the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of the initial degradation tag swap *E. coli* strain library.
- 82.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 80, wherein the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of the initial degradation tag swap *E. coli* strain library.
- 83.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 80, wherein the subsequent degradation tag swap *E. coli* strain library is a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.
- 84.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to claim 80, wherein the subsequent degradation tag swap *E. coli* strain library is a subset of a full combinatorial library of a preceding degradation tag swap *E. coli* strain library.

- 85.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 80-84, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent degradation tag swap *E. coli* strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
- 86.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 80-84, wherein steps d)-e) are repeated until the phenotypic performance of an *E. coli* strain of a subsequent degradation tag swap *E. coli* strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production *E. coli* strain.
- 87.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 80-84, wherein the improved phenotypic performance of step f) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, and combinations thereof.
- 88.** The degradation tag swap method for improving the phenotypic performance of a production *E. coli* strain according to any one of claims 80-84, wherein the improved phenotypic performance of step f) is: increased or more efficient production of a product of interest, said product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.
- 89.** A chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions

of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of lambda phage p_R promoter.

90. The chimeric synthetic promoter of claim 89, wherein nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site (RBS) portion of the lambda phage p_R promoter are selected from the nucleic acid sequences found in Table 1.5.
91. A chimeric synthetic promoter operably linked to a heterologous gene for expression in a microbial host cell, wherein the chimeric synthetic promoter is 60-90 nucleotides in length and consists of a distal portion of lambda phage p_R promoter, variable -35 and -10 regions of lambda phage p_L and p_R promoters that are each six nucleotides in length, core portions of lambda phage p_L and p_R promoters and a 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene.
92. The chimeric synthetic promoter of claim 91, wherein nucleic acid sequences of the distal portion of the lambda phage p_R promoter, the variable -35 and -10 regions of the lambda phage p_L and p_R promoters, the core portions of the the lambda phage p_L and p_R promoters and the 5' UTR/Ribosomal Binding Site (RBS) portion of the promoter of the *E. coli acs* gene are selected from the nucleic acid sequences found in Table 1.5.
93. The chimeric synthetic promoter of any of claims 89-90, wherein the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 132-152, 159-160, 162, 165, 174-175, 188, 190, 199-201 or 207.
94. The chimeric synthetic promoter of any of claims 91-92, wherein the chimeric synthetic promoter consists of a nucleic acid sequence selected from SEQ ID NOs. 153-158, 161, 163-164, 166-173, 176-187, 189, 191-198 or 202-206.
95. The chimeric synthetic promoter of any of claims 89-92, wherein the microbial host cell is *E. coli*.
96. The chimeric synthetic promoter of claim 95, wherein the heterologous gene encodes a protein product of interest found in Table 2.
97. The chimeric synthetic promoter of claim 95, wherein the heterologous gene is a gene that is part of a lysine biosynthetic pathway.

- 98.** The chimeric synthetic promoter of claim 97, wherein the heterologous gene is selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene, the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene.
- 99.** The chimeric synthetic promoter of claim 95, wherein the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.
- 100.** The chimeric synthetic promoter of claim 99, wherein the heterologous gene is selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA* gene, the *ispB* gene, the *crtE* gene, the *crtB* gene, the *crtI* gene, the *crtY* gene, the *ymgA* gene, the *dxr* gene, the *elbA* gene, the *gdhA* gene, the *appY* gene, the *elbB* gene, or the *ymgB* gene.
- 101.** The chimeric synthetic promoter of claim 95, wherein the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.
- 102.** The chimeric synthetic promoter of claim 99, wherein the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).
- 103.** A heterologous gene operably linked to a chimeric synthetic promoter with a nucleic acid sequence selected from SEQ ID NOs. 132-207.
- 104.** The heterologous gene of claim 103, wherein the heterologous gene encodes a protein product of interest found in Table 2.
- 105.** The heterologous gene of claim 103, wherein the heterologous gene is a gene that is part of a lysine biosynthetic pathway.
- 106.** The heterologous gene of claim 105, wherein the heterologous gene is selected from the *asd* gene, the *ask* gene, the *hom* gene, the *dapA* gene, the *dapB* gene, the *dapD* gene,

the *ddh* gene, the *argD* gene, the *dapE* gene, the *dapF* gene, the *lysA* gene, the *lysE* gene, the *zwf* gene, the *pgi* gene, the *ktk* gene, the *fbp* gene, the *ppc* gene, the *pck* gene, the *ddx* gene, the *pyc* gene or the *icd* gene.

107. The heterologous gene of claim 103, wherein the heterologous gene is a gene that is part of a lycopene biosynthetic pathway.

108. The heterologous gene of claim 107, wherein the heterologous gene is selected from the *dxs* gene, the *ispC* gene, the *ispE* gene, the *ispD* gene, the *ispF* gene, the *ispG* gene, the *ispH* gene, the *idi* gene, the *ispA* gene, the *ispB* gene, the *crtE* gene, the *crtB* gene, the *crtI* gene, the *crtY* gene, the *ymgA* gene, the *dxr* gene, the *elbA* gene, the *gdhA* gene, the *appY* gene, the *elbB* gene, or the *ymgB* gene.

109. The heterologous gene of claim 103, wherein the heterologous gene encodes a biopharmaceutical or is a gene in a pathway for generating a biopharmaceutical.

110. The heterologous gene of claim 109, wherein the biopharmaceutical is selected from humulin (rh insulin), intronA (interferon alpha2b), roferon (interferon alpha2a), humatrope (somatropin rh growth hormone), neupogen (filgrastim), detaferon (interferon beta-1b), lispro (fast-acting insulin), rapilysin (reteplase), infergen (interferon alfacon-1), glucagon, beromun (tasonermin), ontak (denileukin diftitox), lantus (long-acting insulin glargine), kineret (anakinra), natreacor (nesiritide), somavert (pegvisomant), calcitonin (recombinant calcitonin salmon), lucentis (ranibizumab), preotact (human parathyroid hormone), kyrstexxal (rh urate oxidase, PEGlyated), nivistim (filgrastim, rhGCSF), voraxaze (glucarpidase), or preos (parathyroid hormone).

FIGURE 1

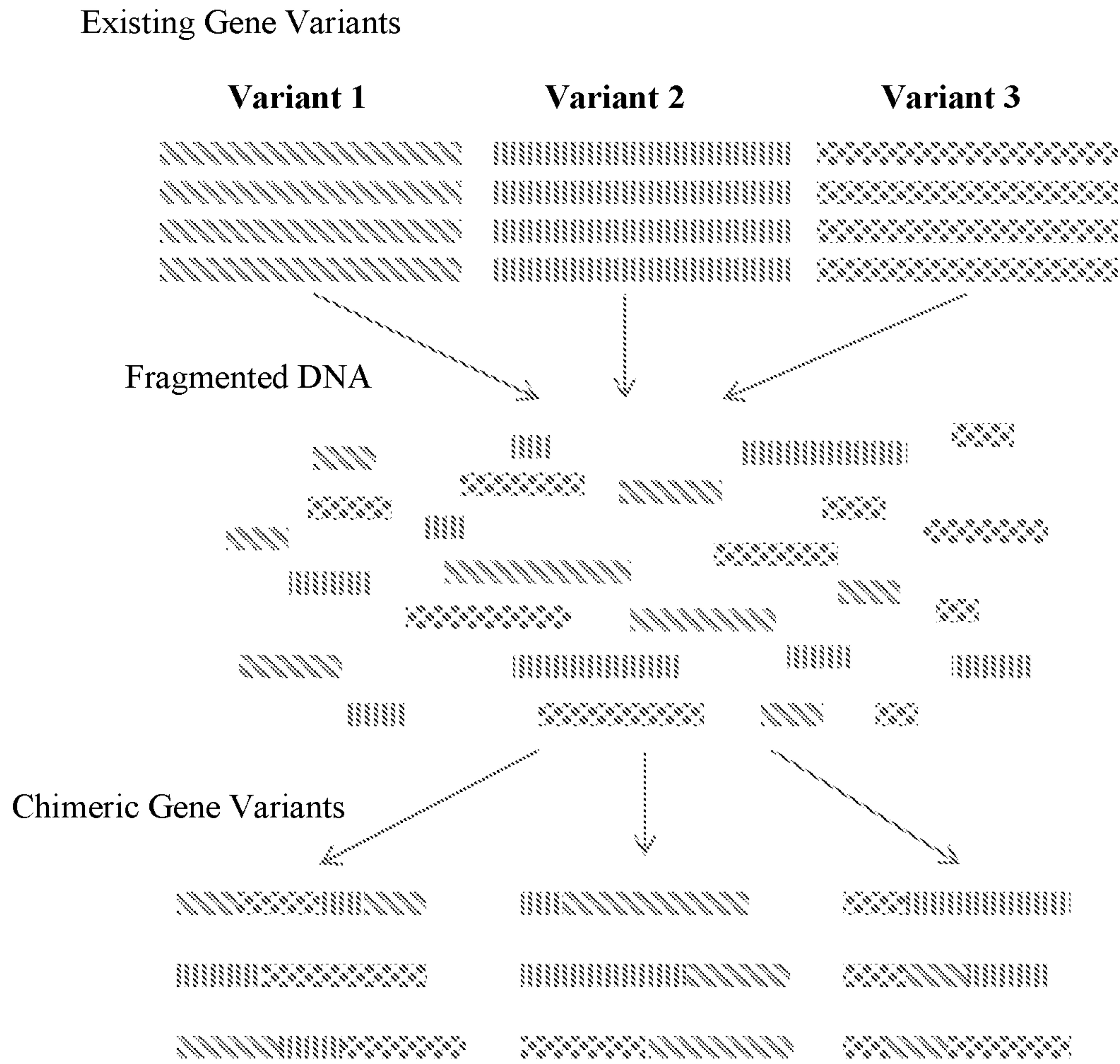


FIGURE 2

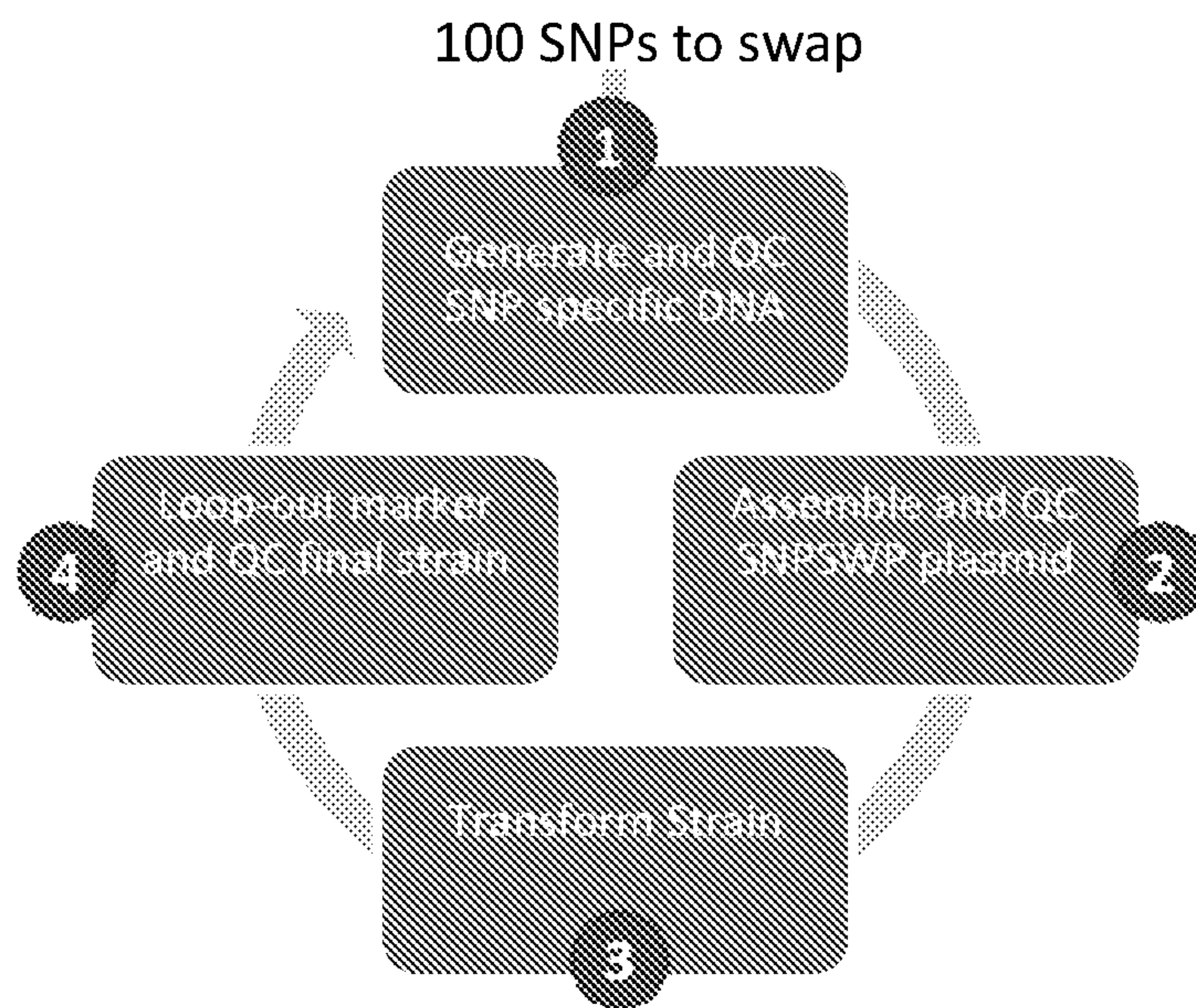


FIGURE 3

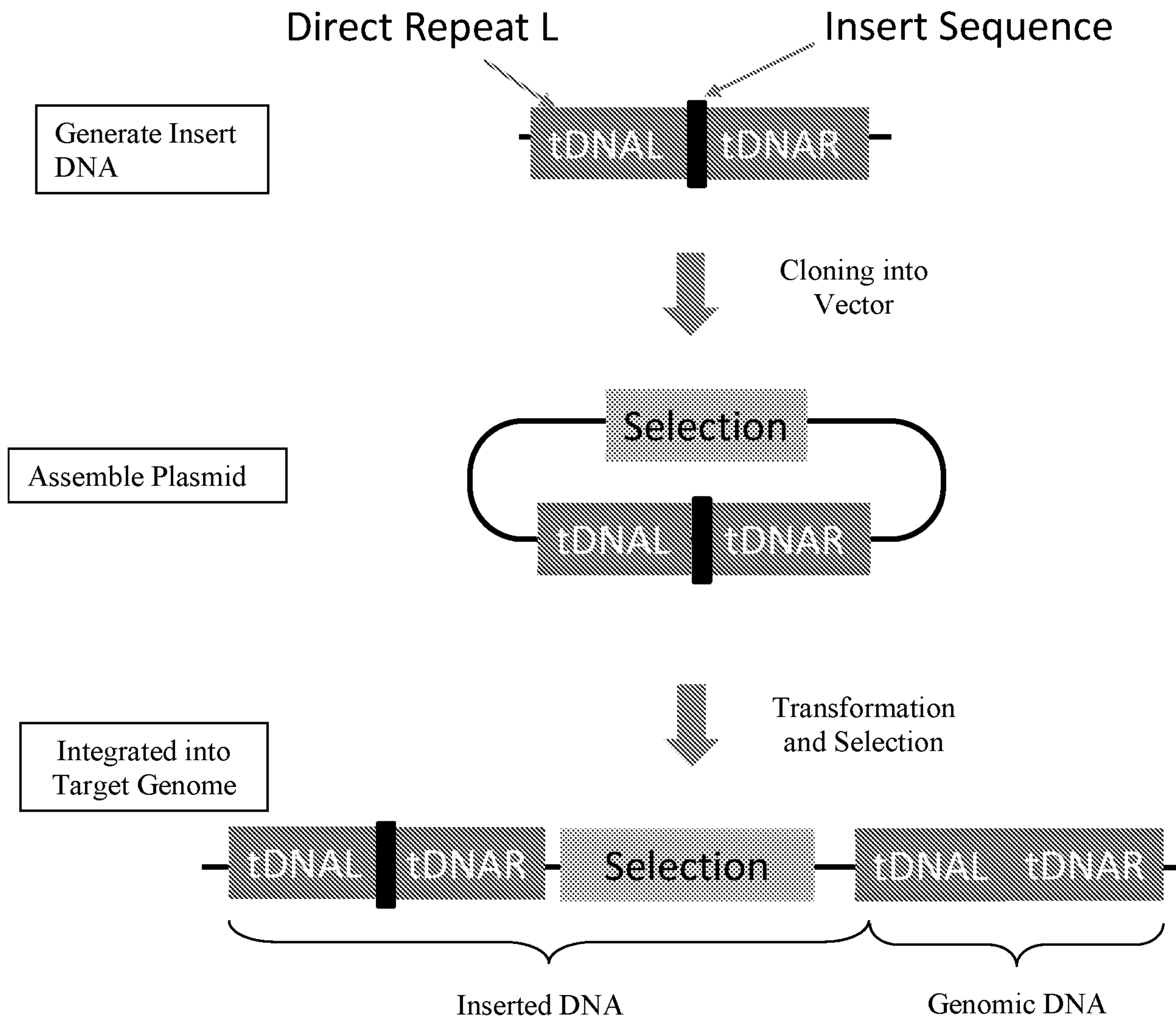


FIGURE 4

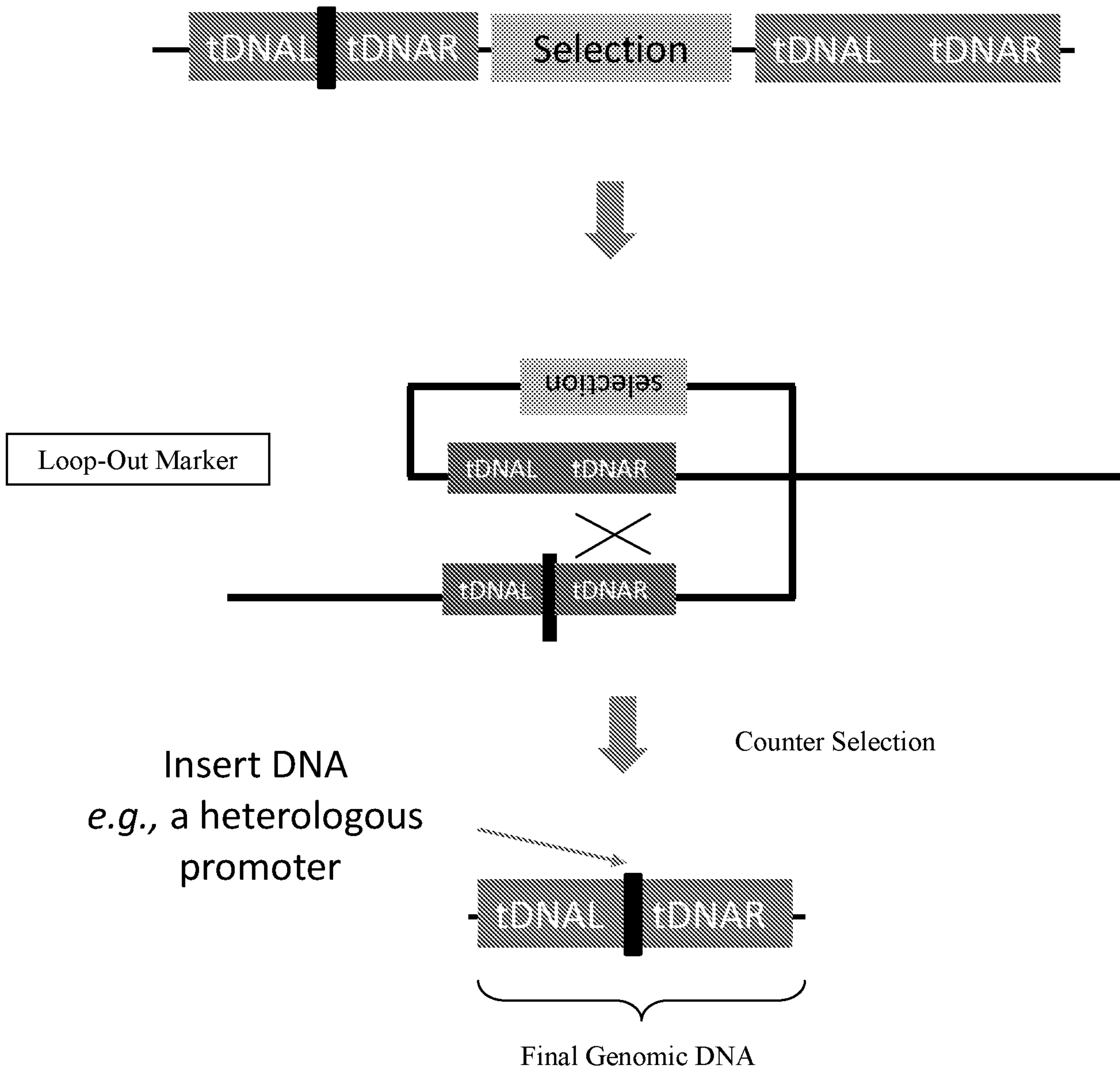


FIGURE 5

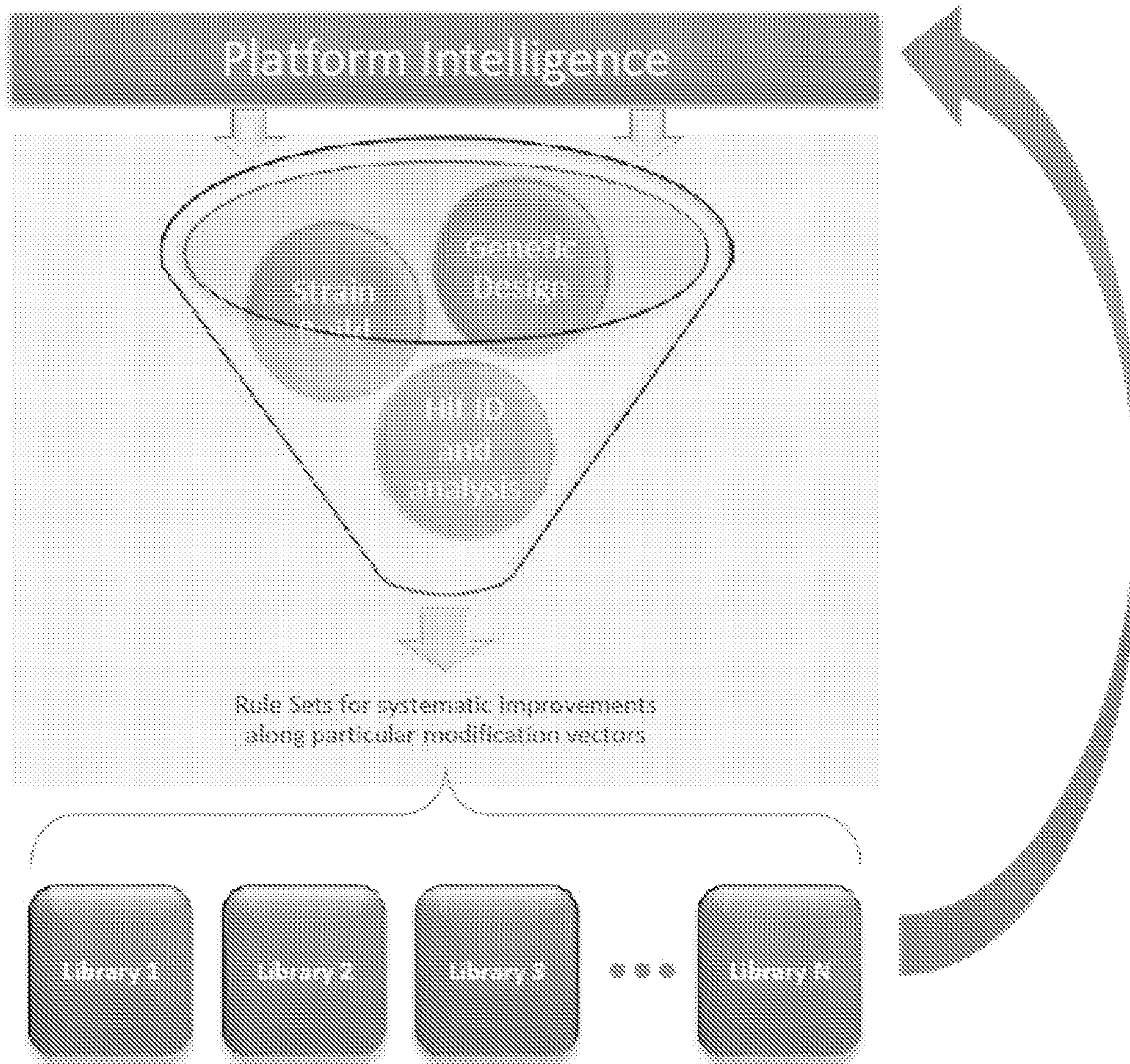


FIGURE 6A

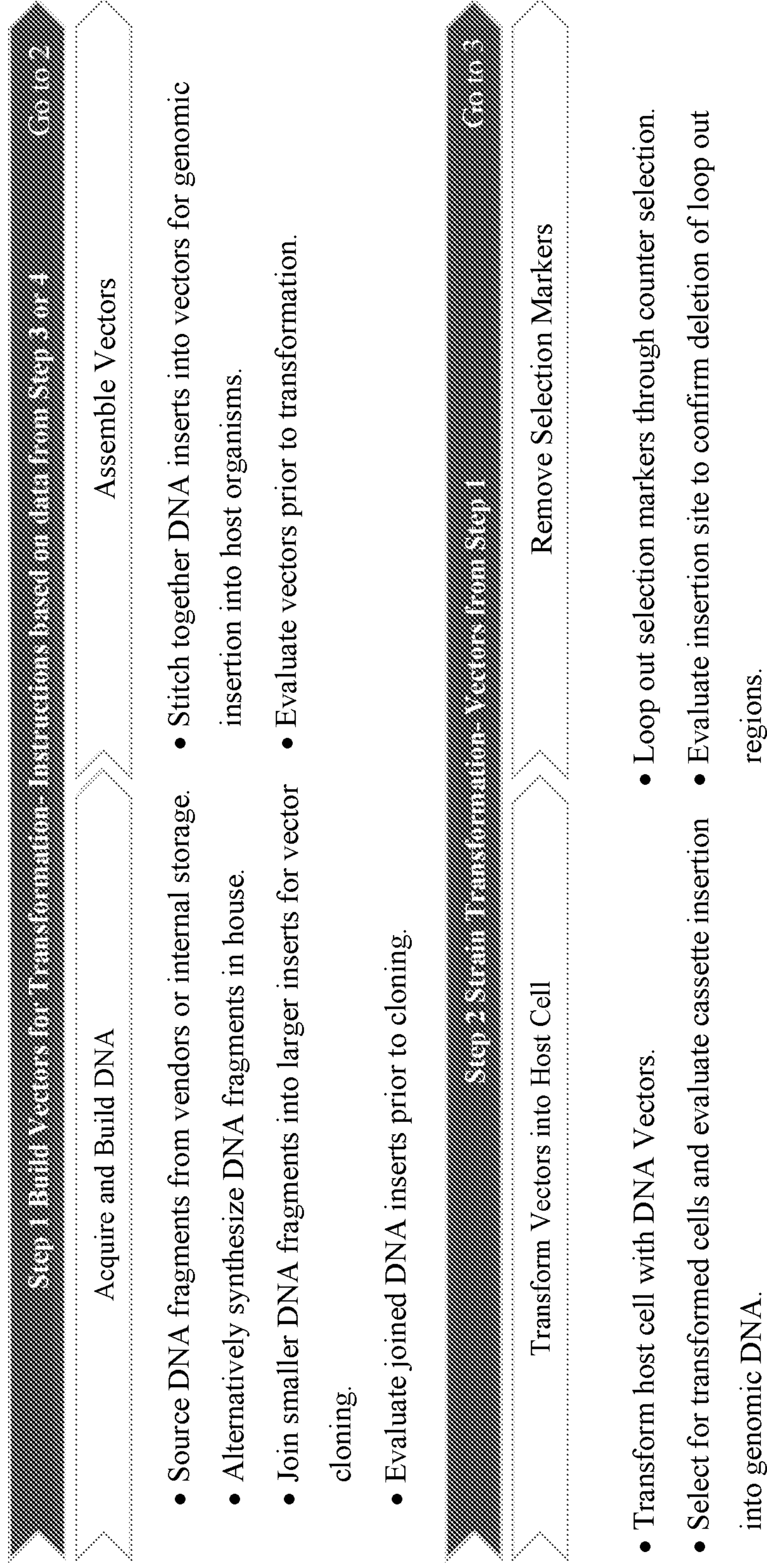
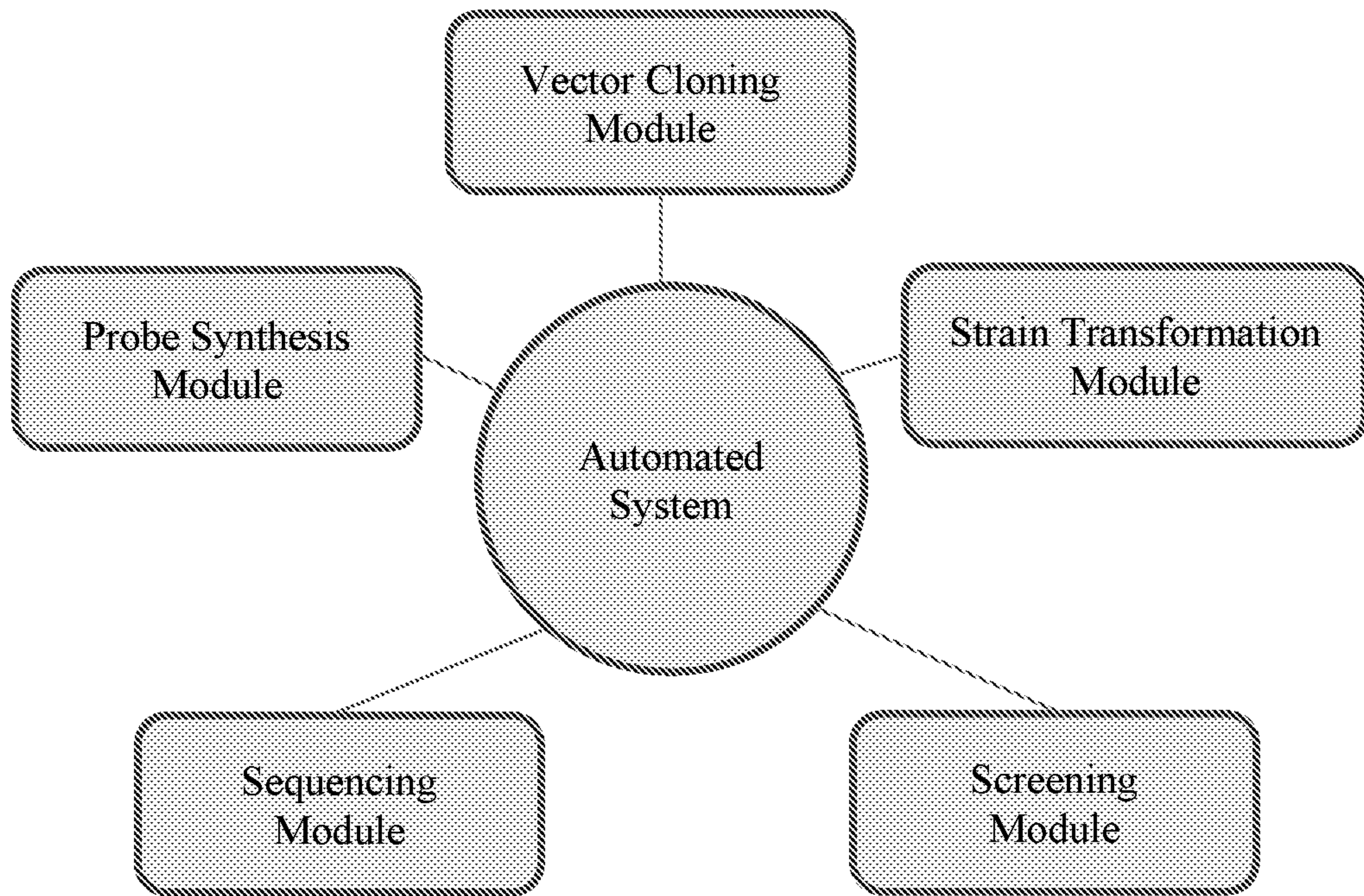


FIGURE 6B



FIGURE 7



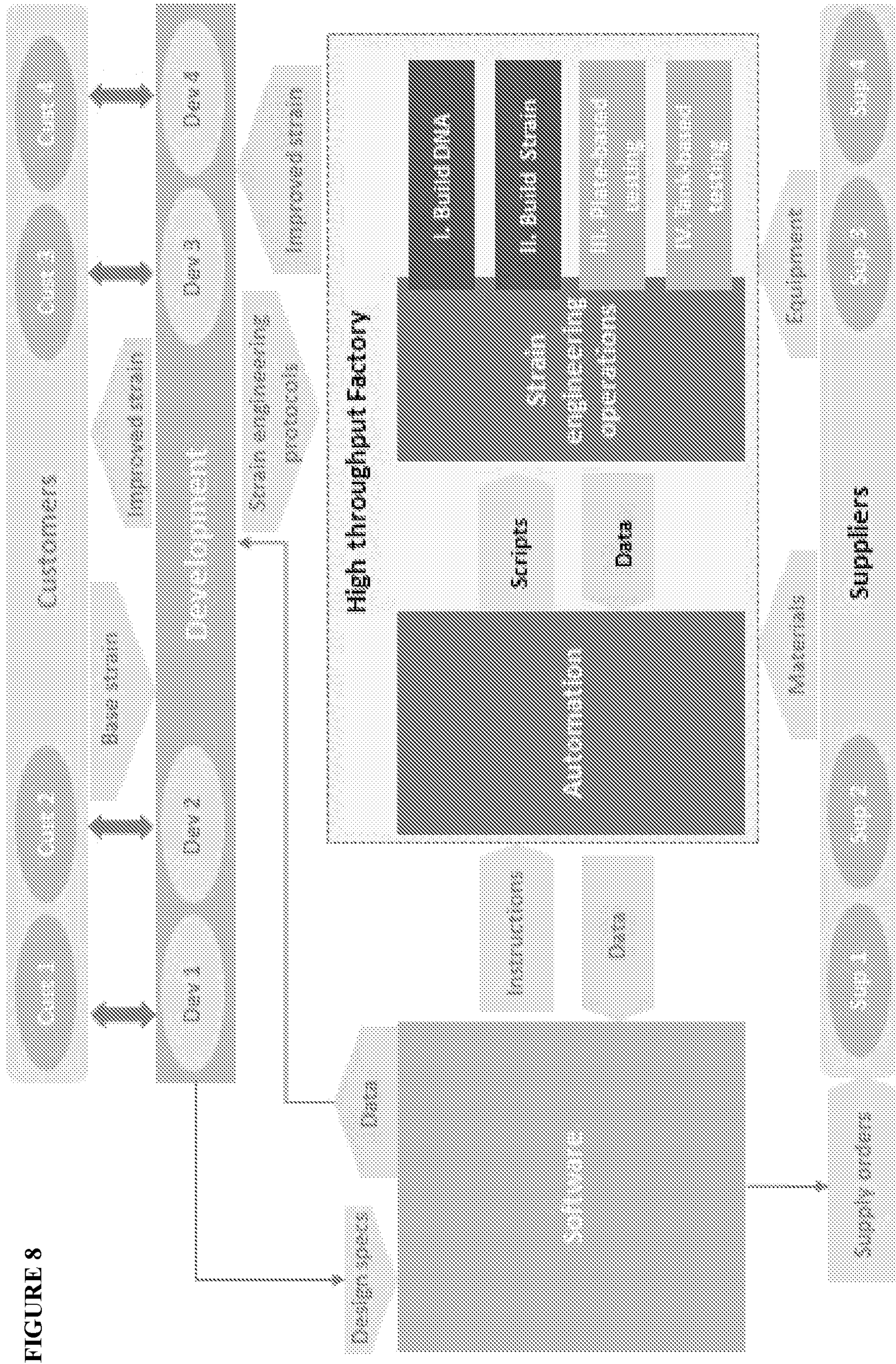


FIGURE 8

FIGURE 9

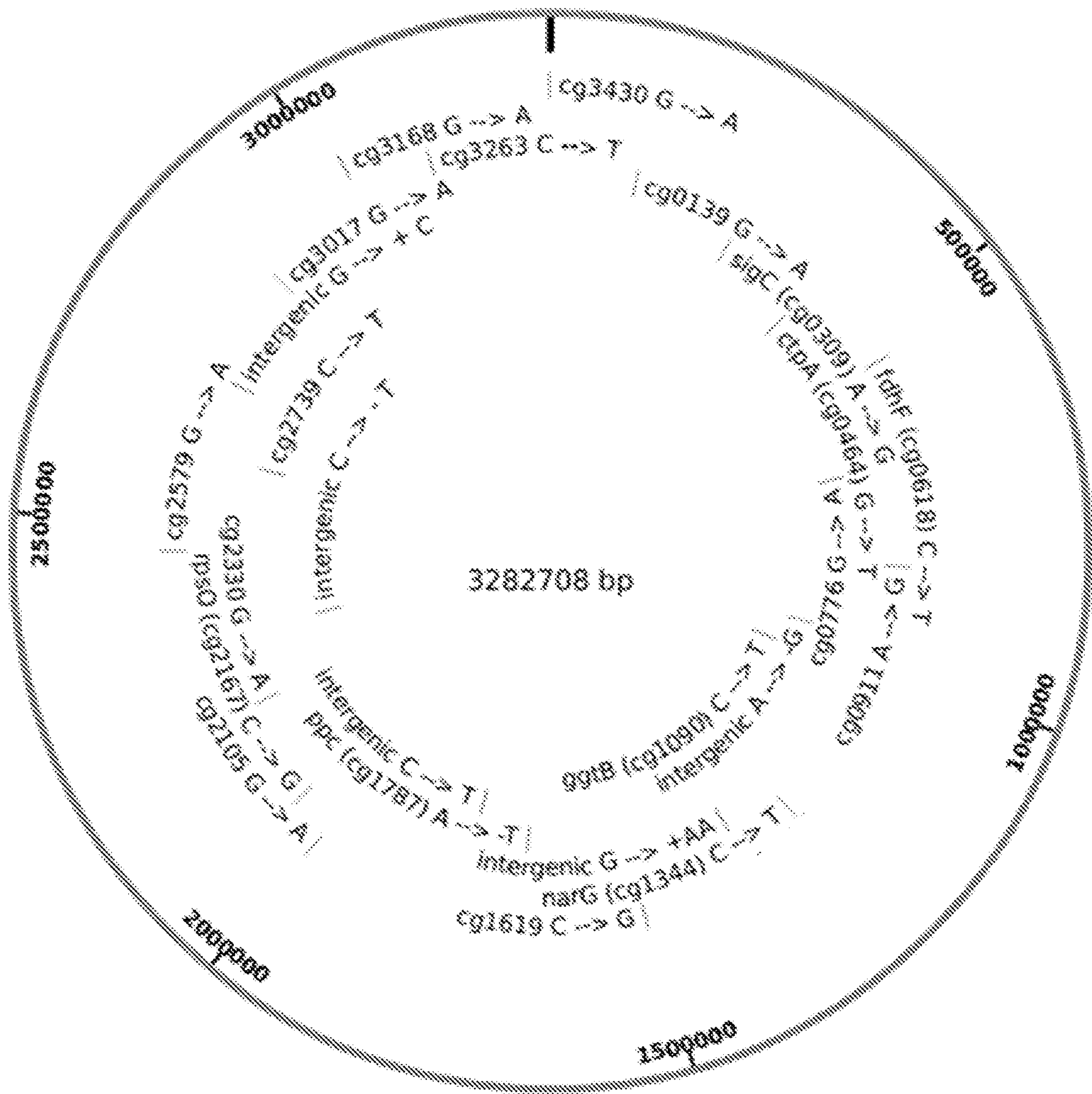


FIGURE 10

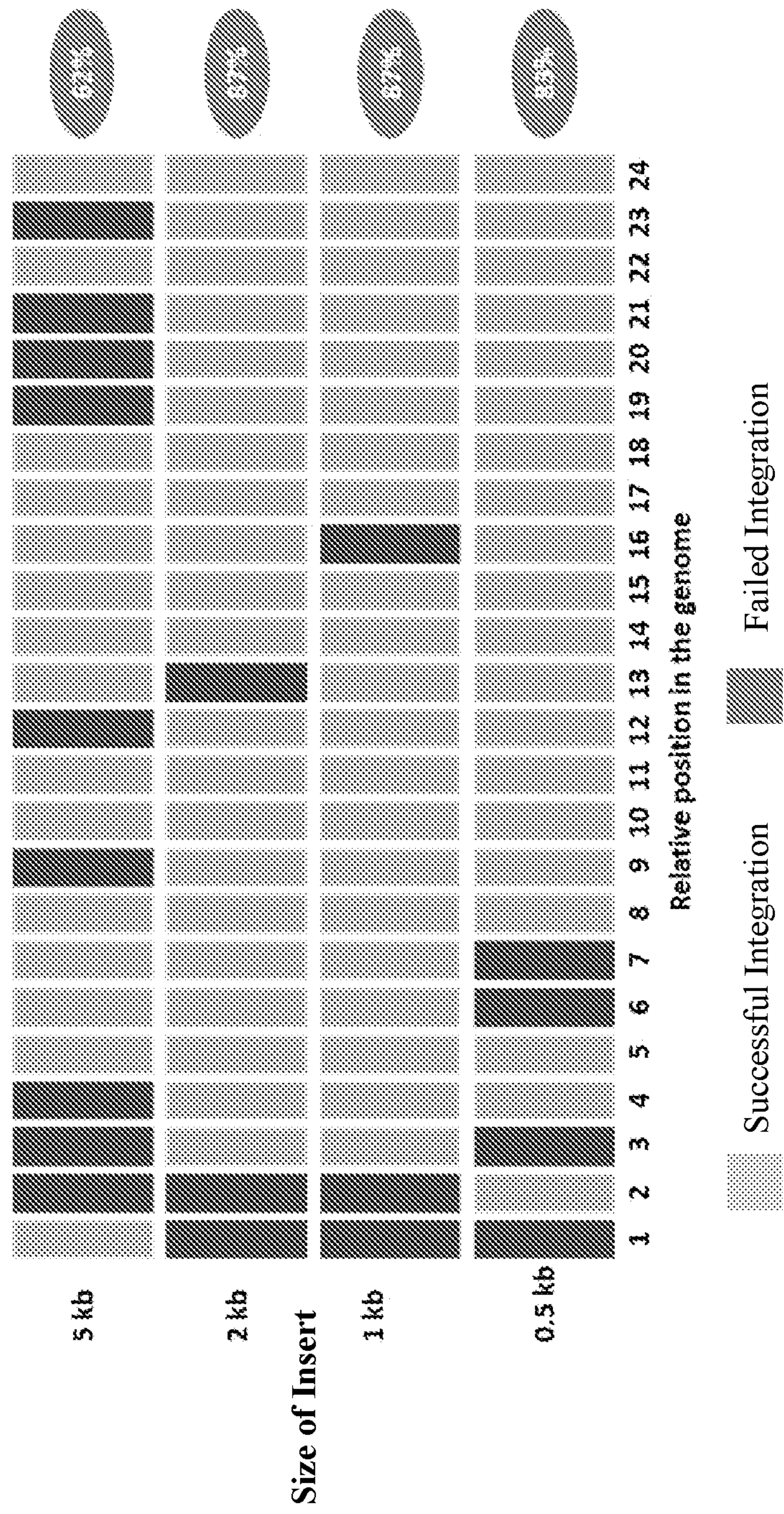


FIGURE 11

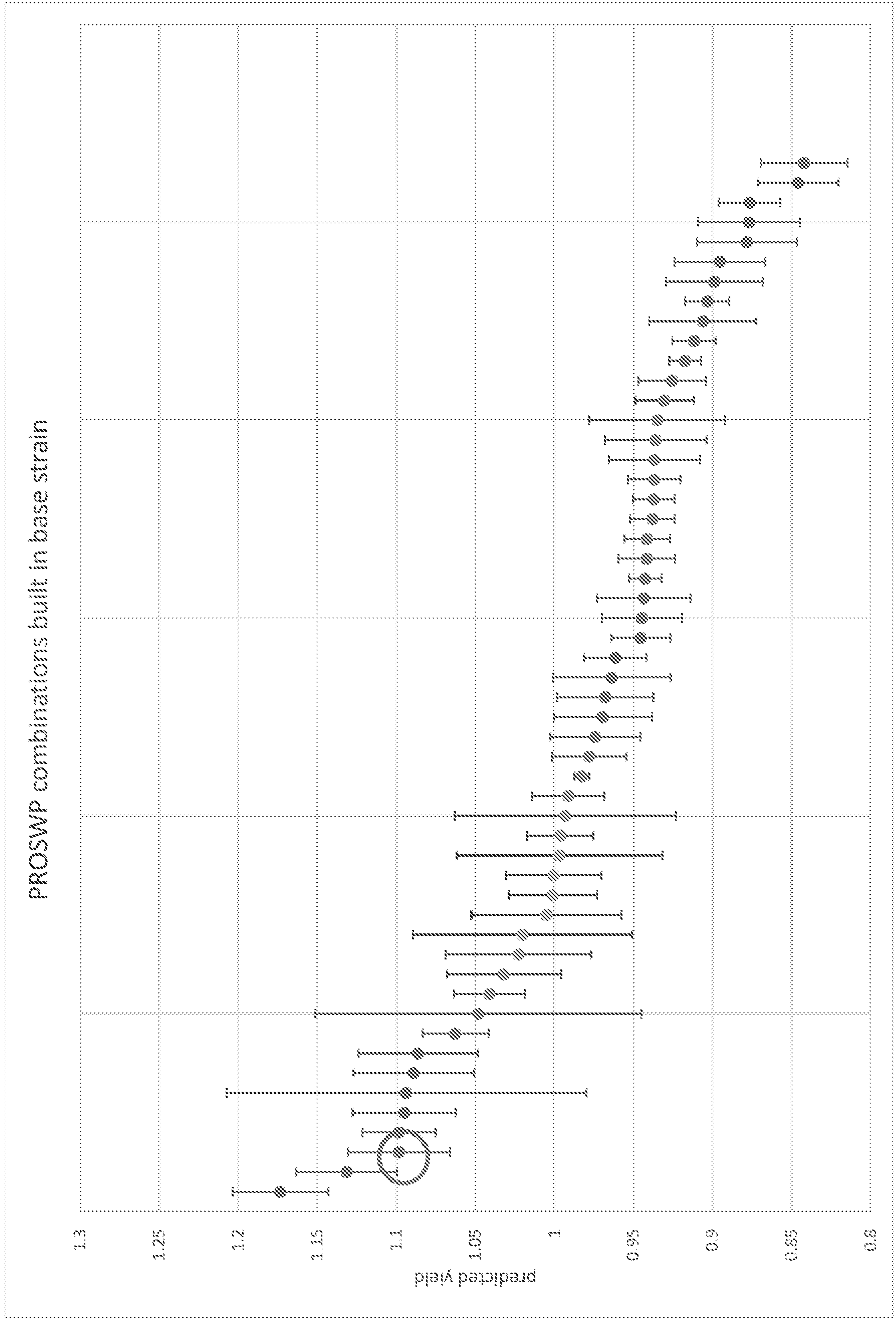
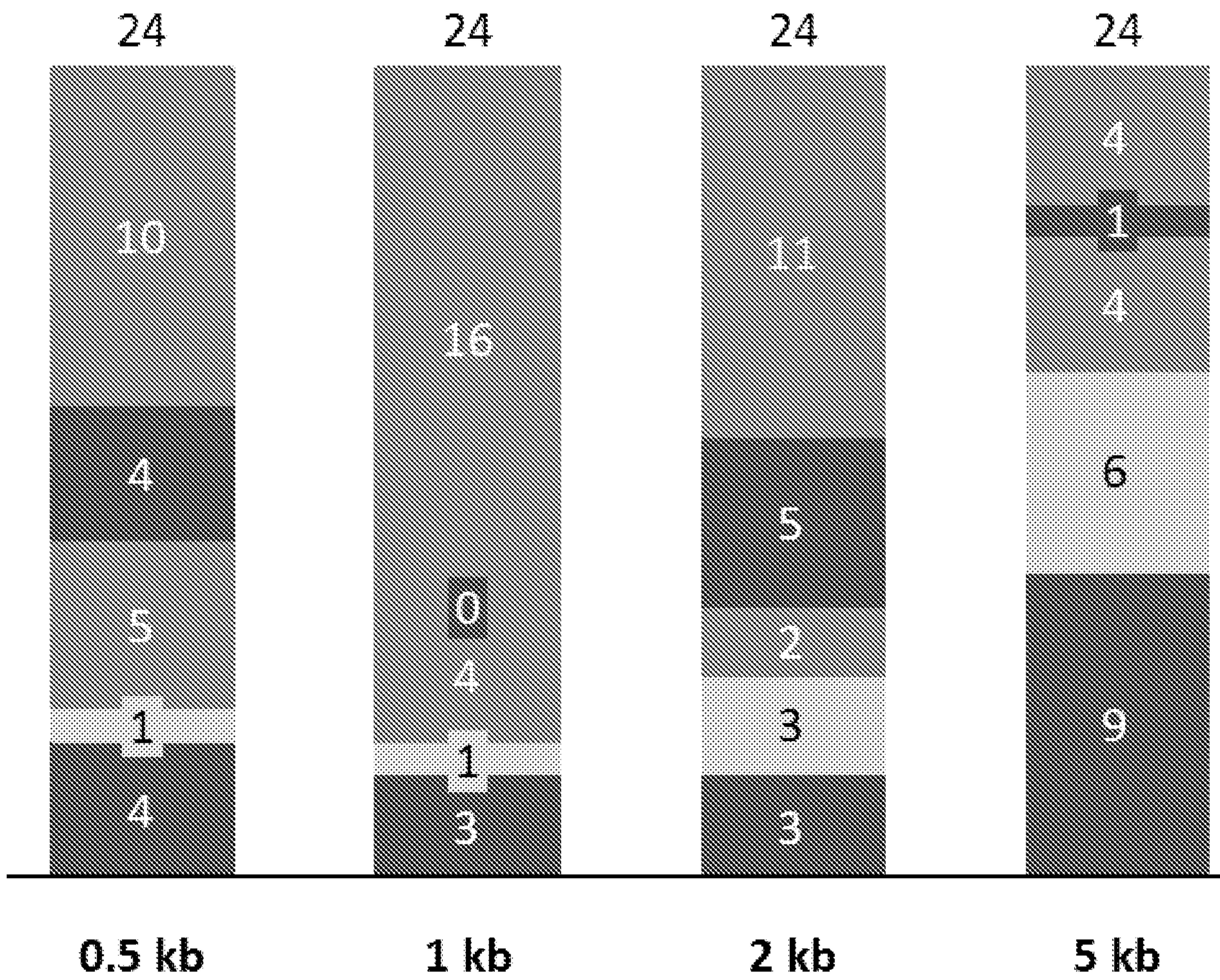


FIGURE 12






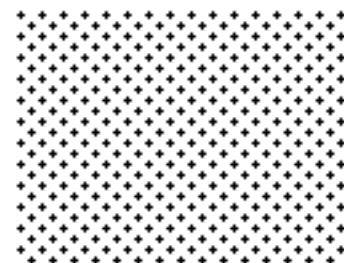


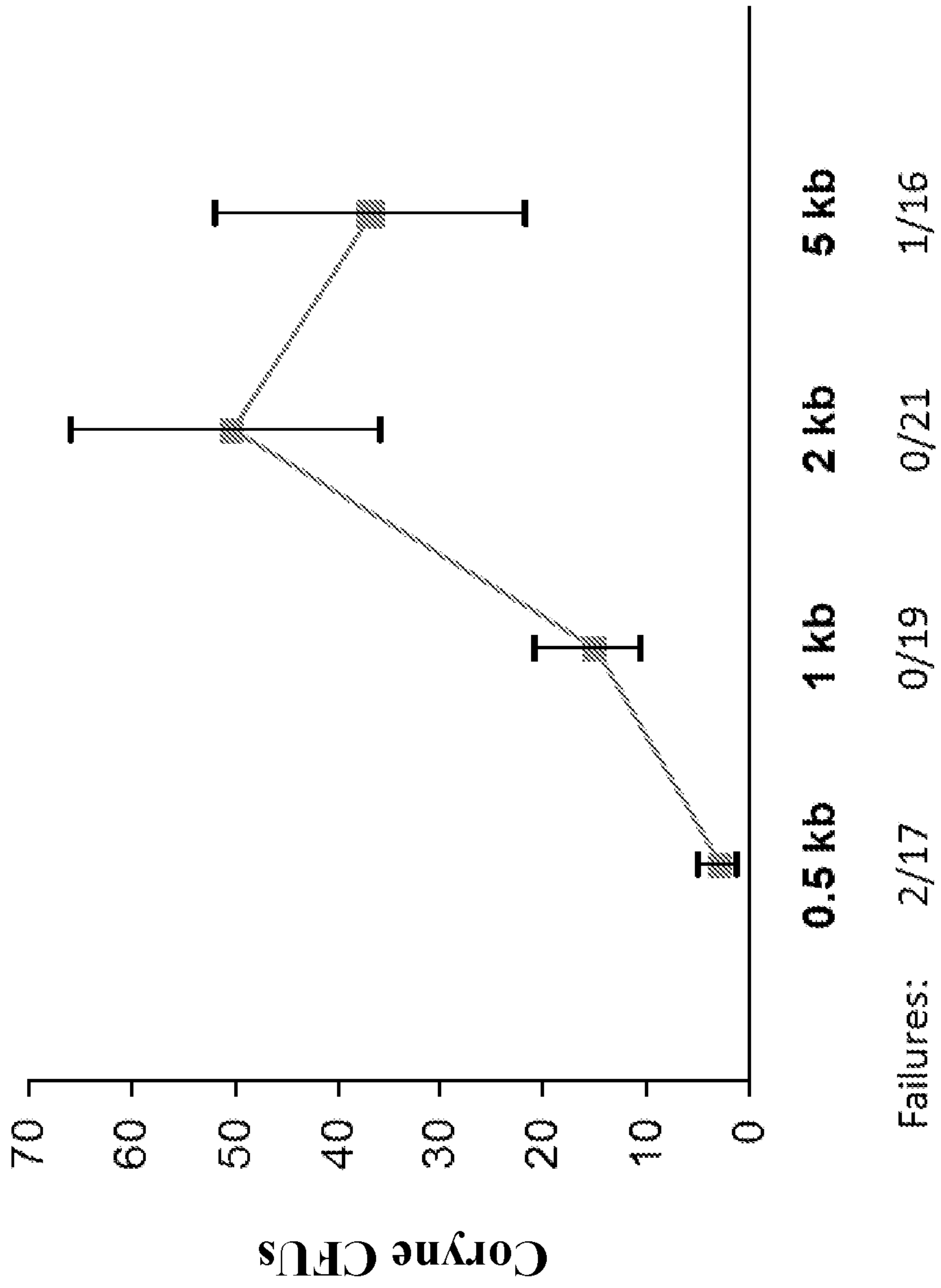
-  4 correct colonies
-  3 correct colonies
-  2 correct colonies
-  1 correct colony
-  fail
-  (0 correct colonies)

FIGURE 13



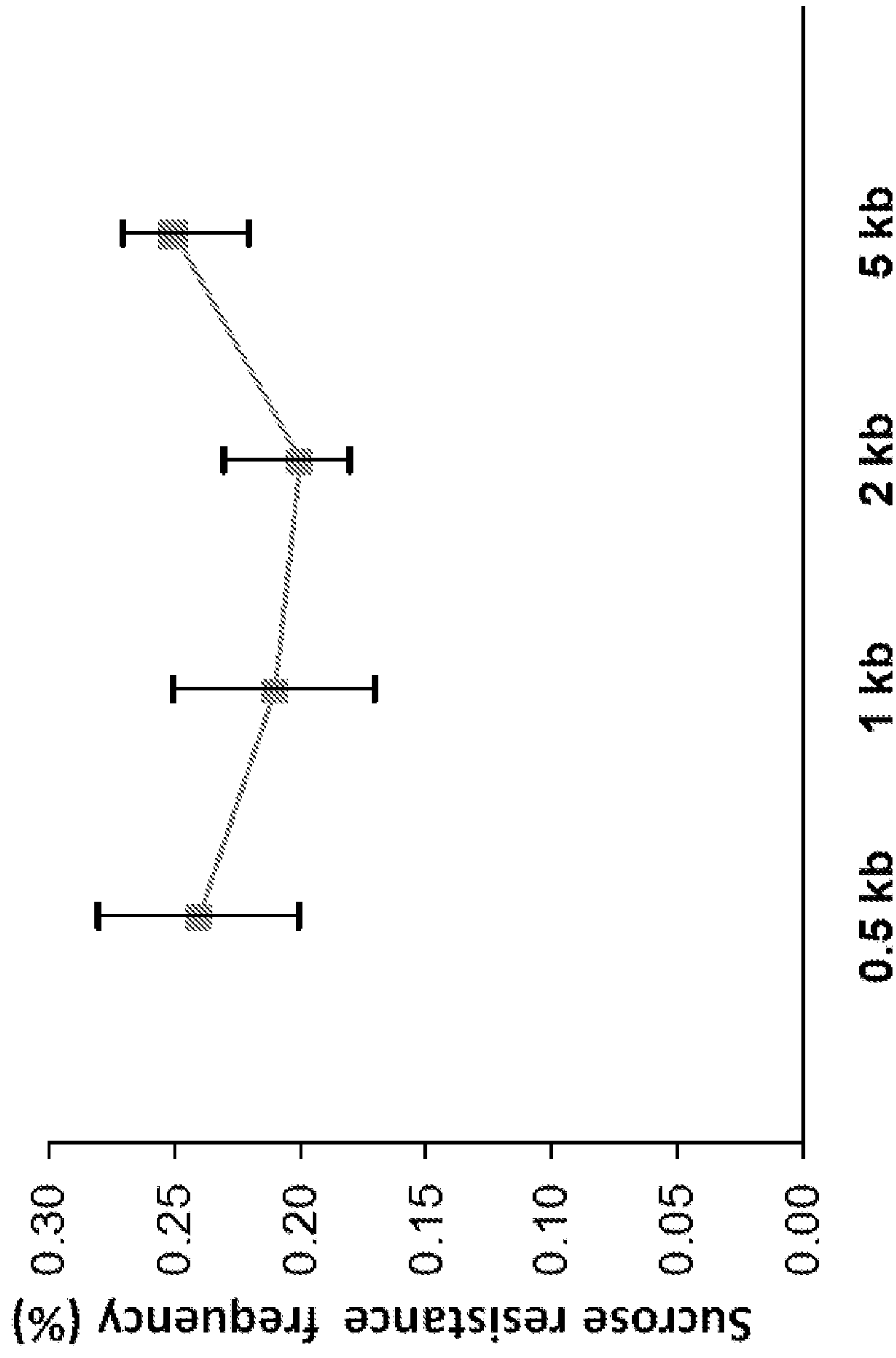


FIGURE 14

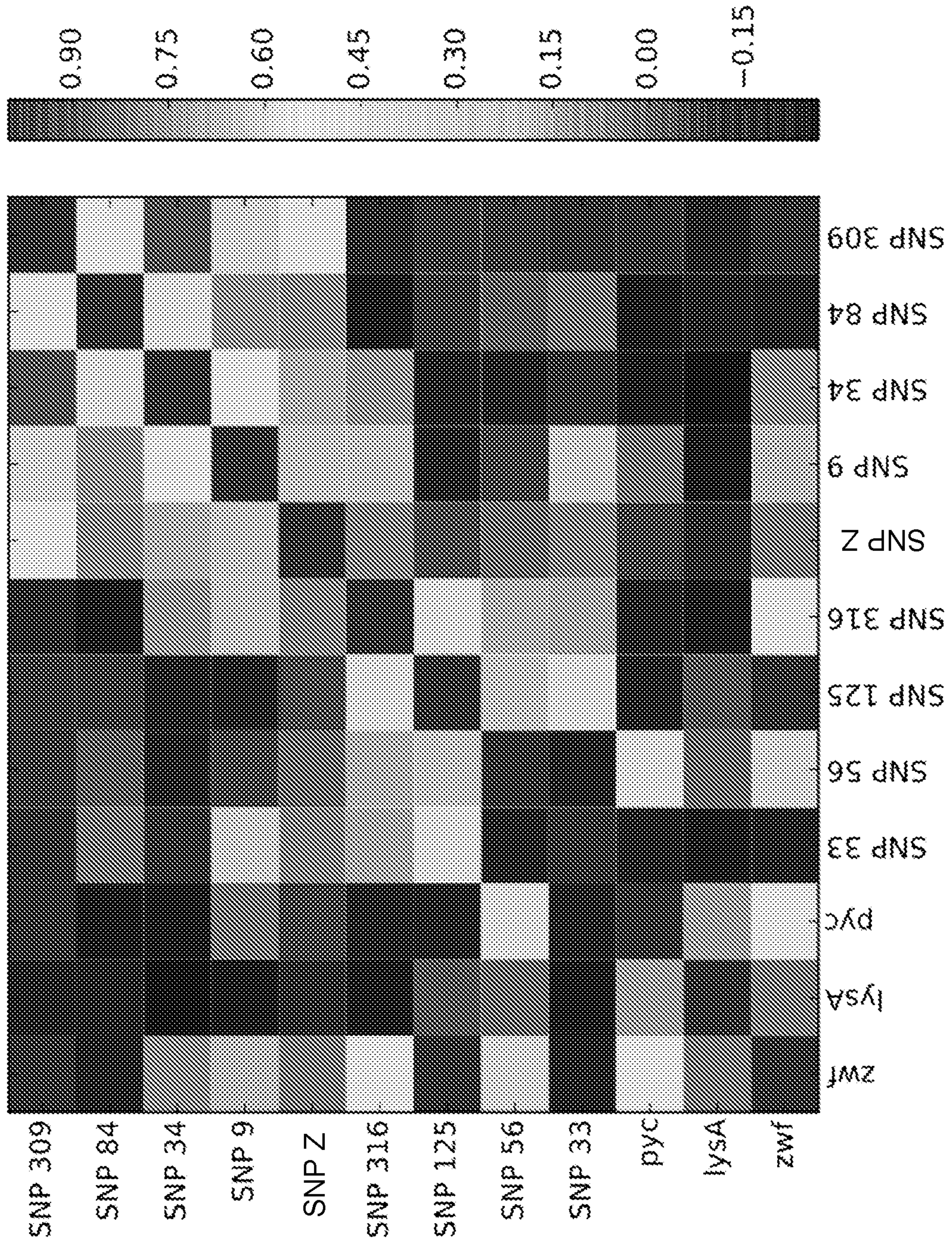


FIGURE 15

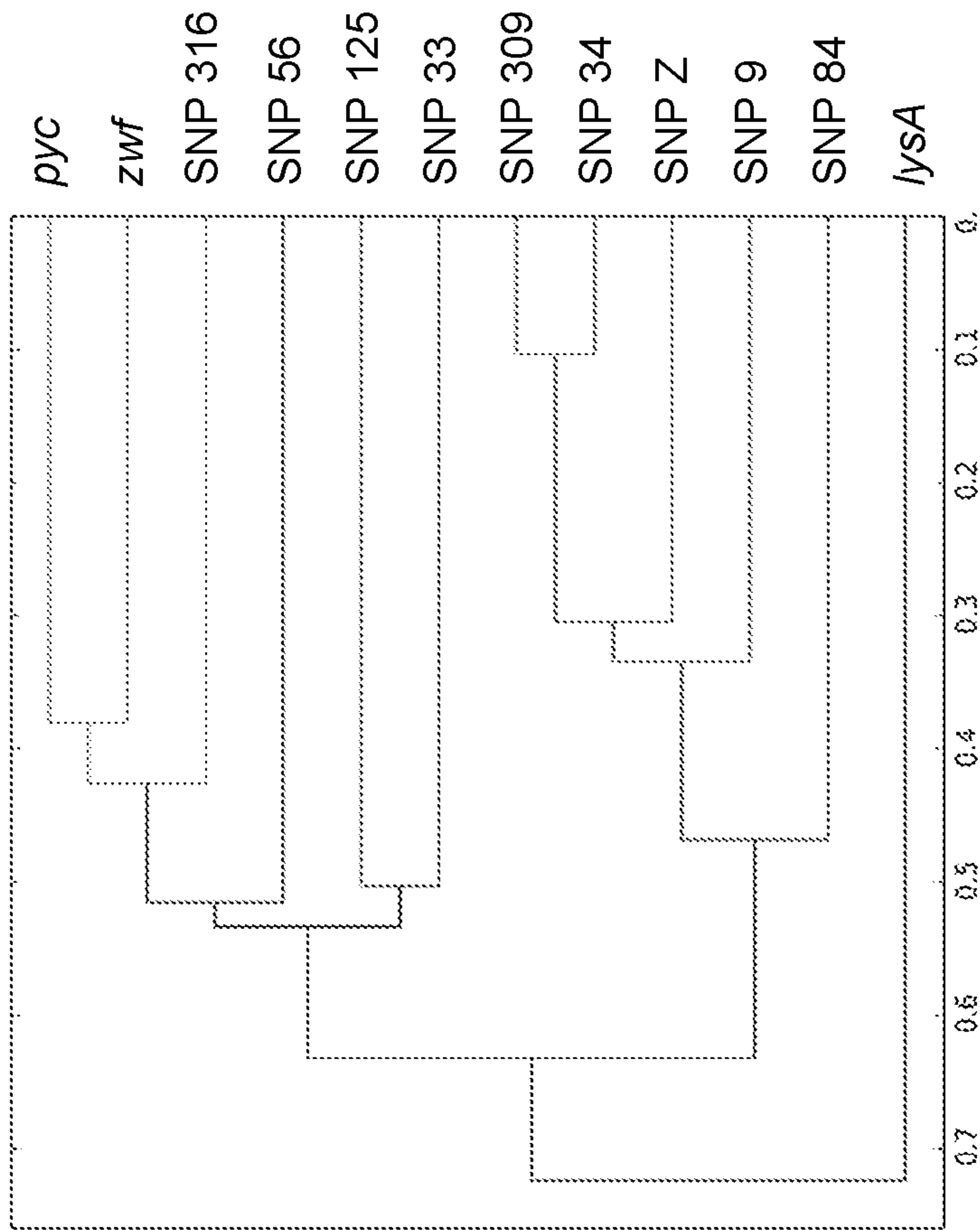


FIGURE 16 A

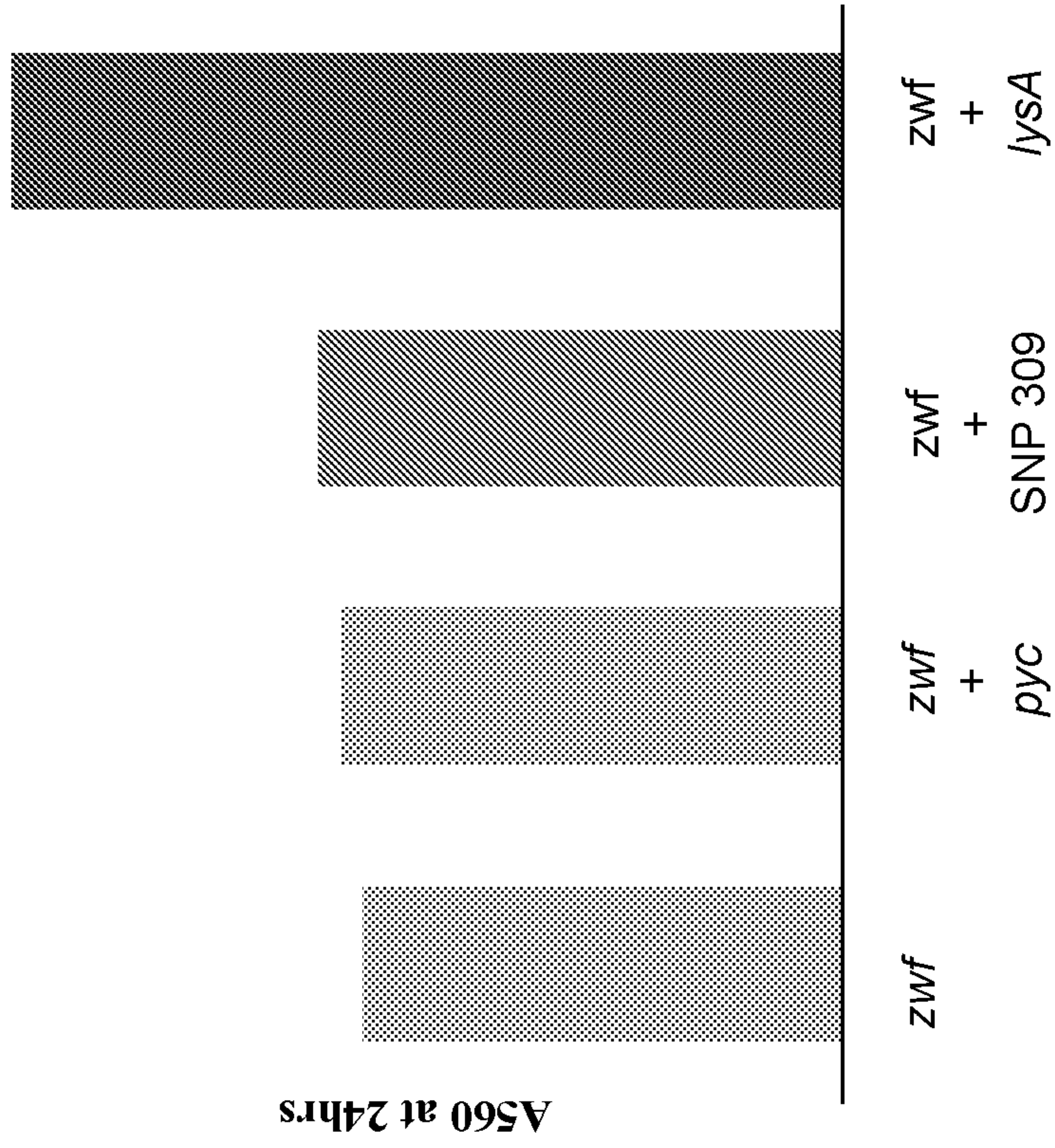


FIGURE 16 B

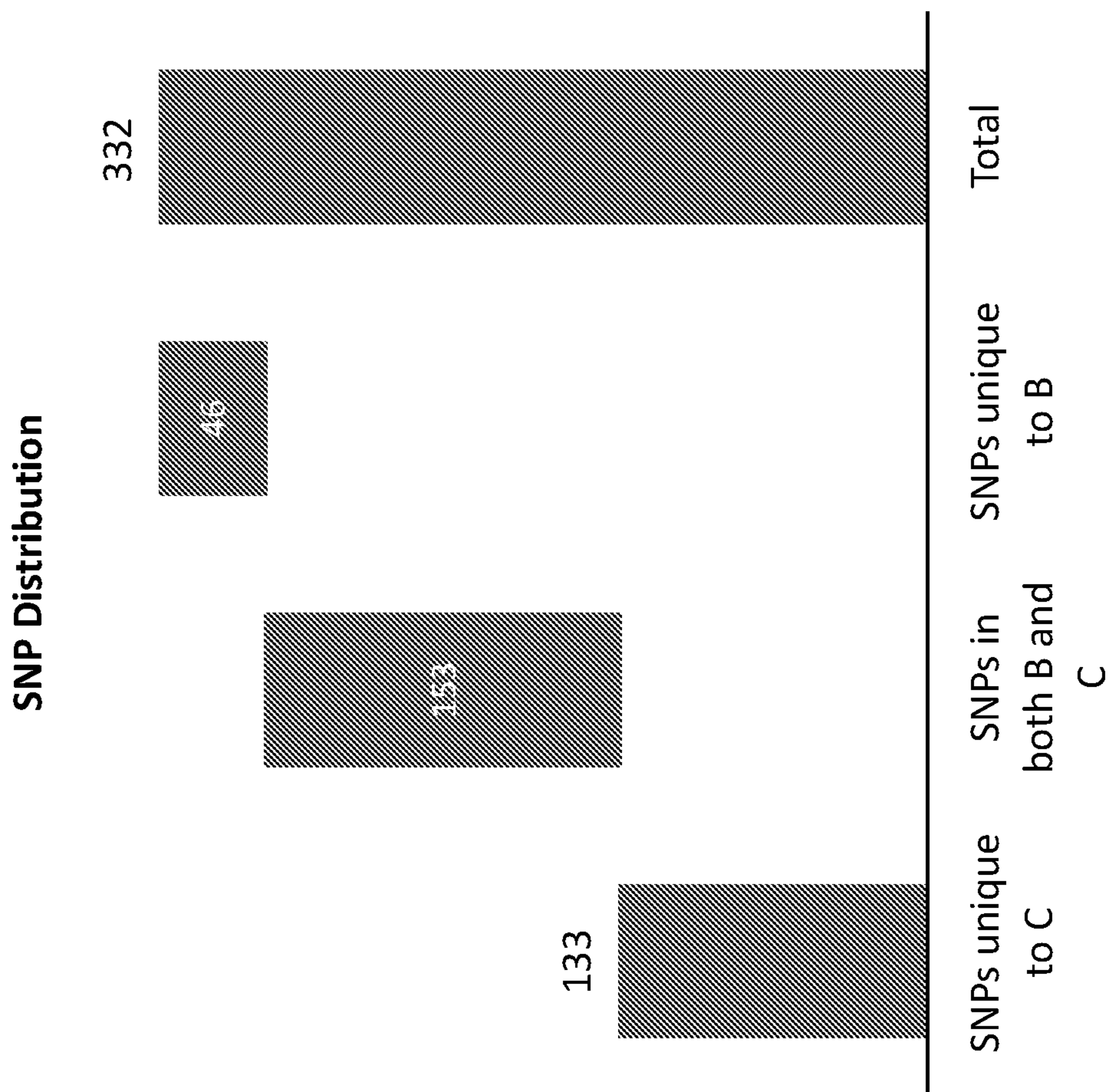


FIGURE 17 A

FIGURE 17 B

FIGURE 18

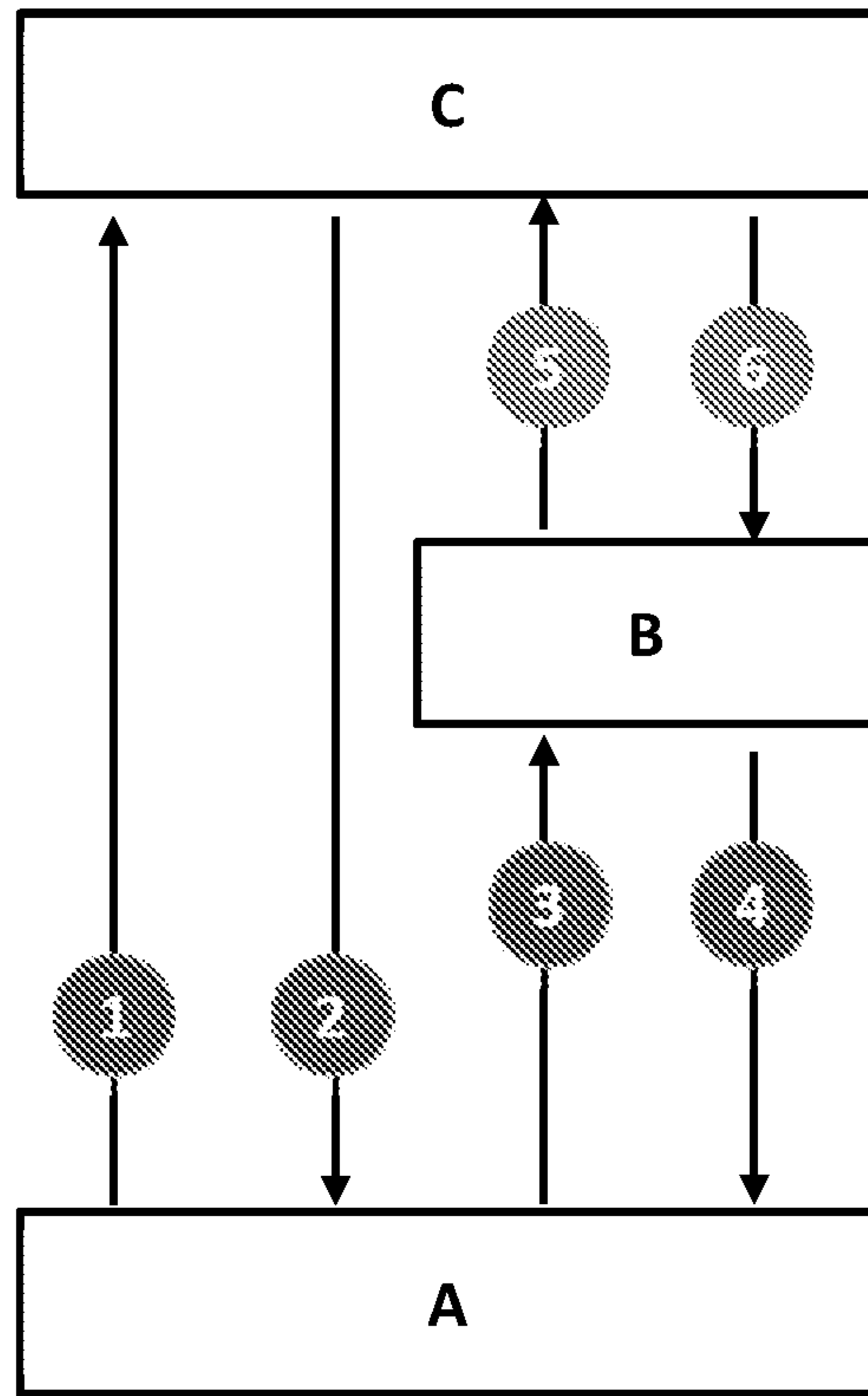


FIGURE 19

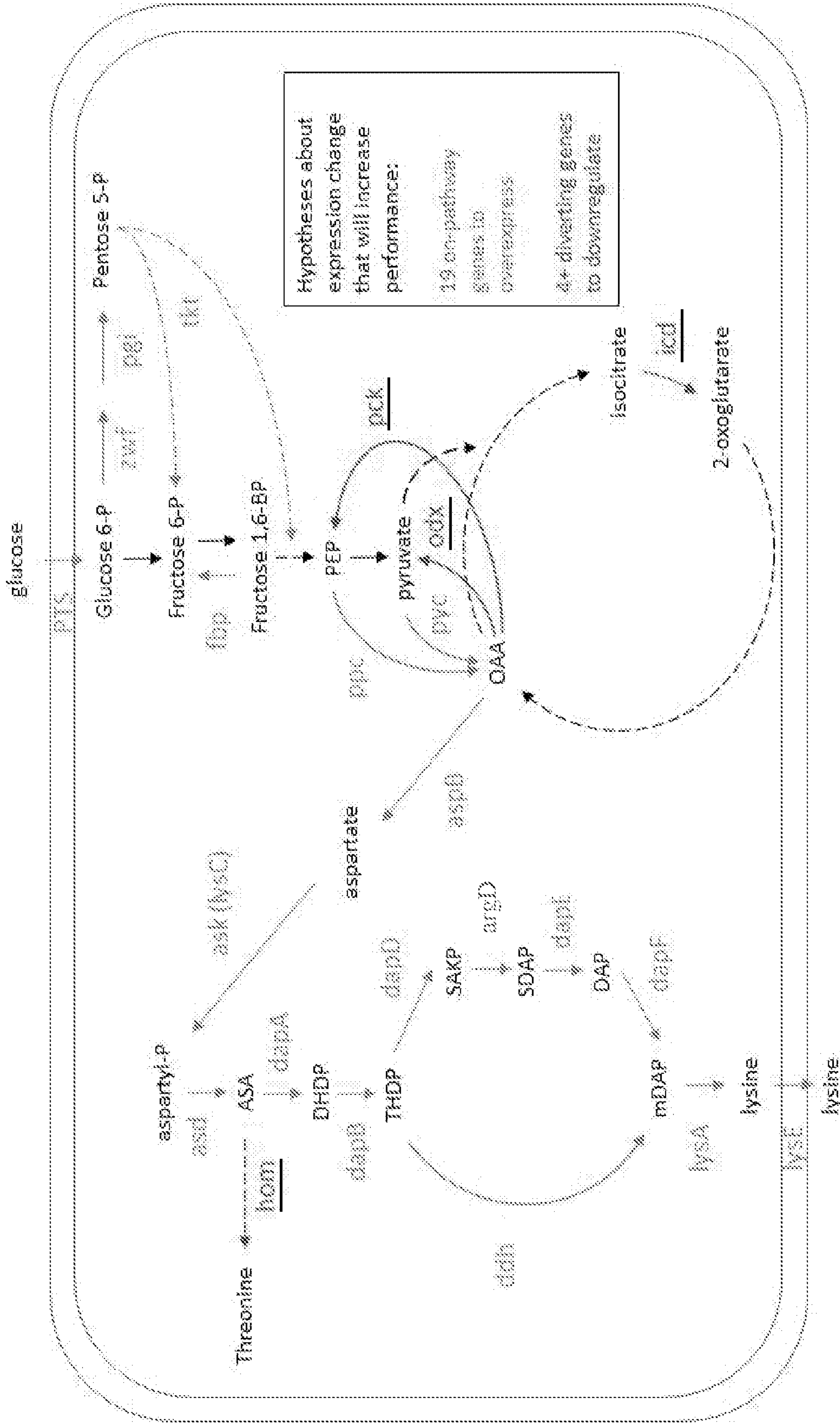


FIGURE 20

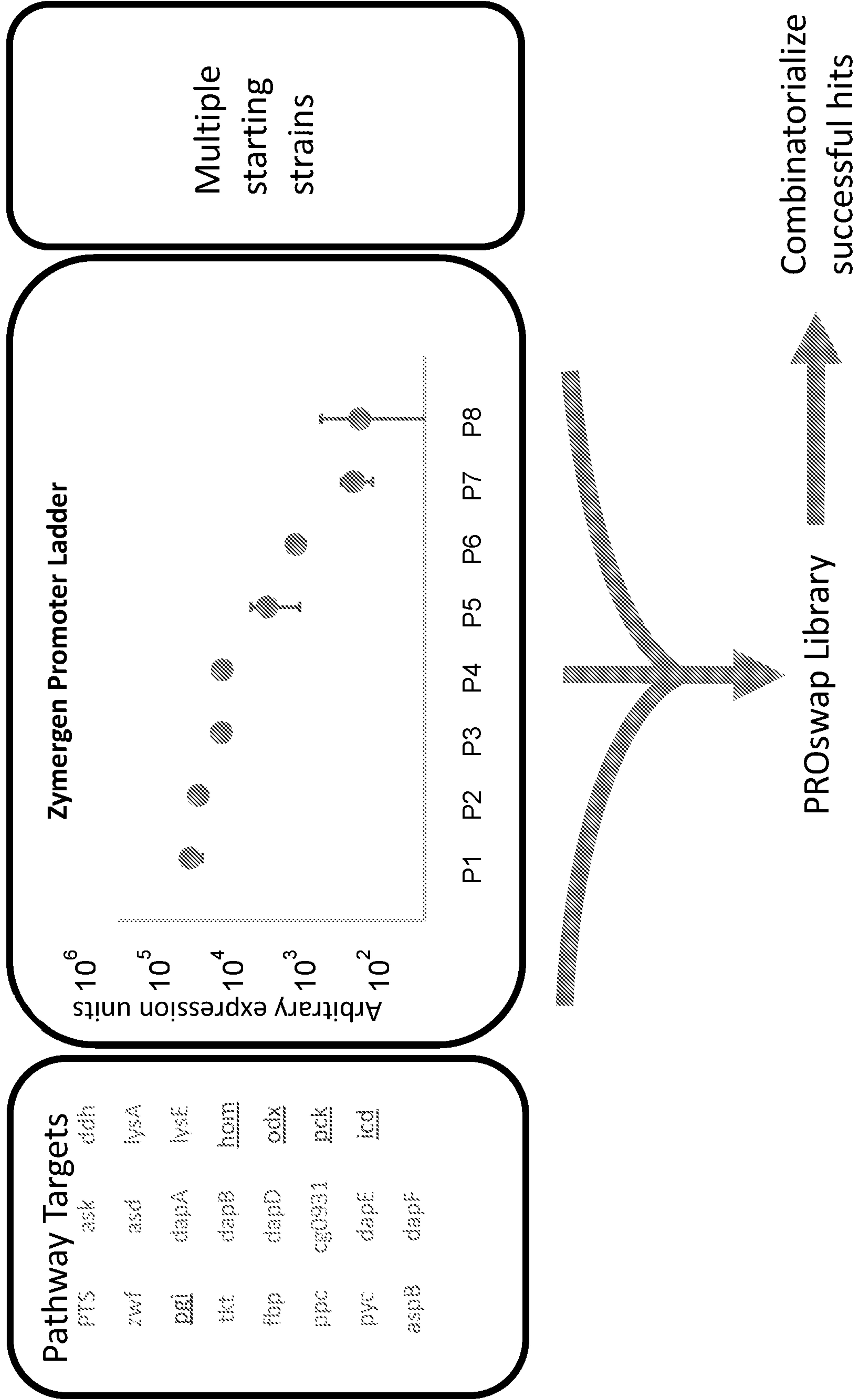


FIGURE 21

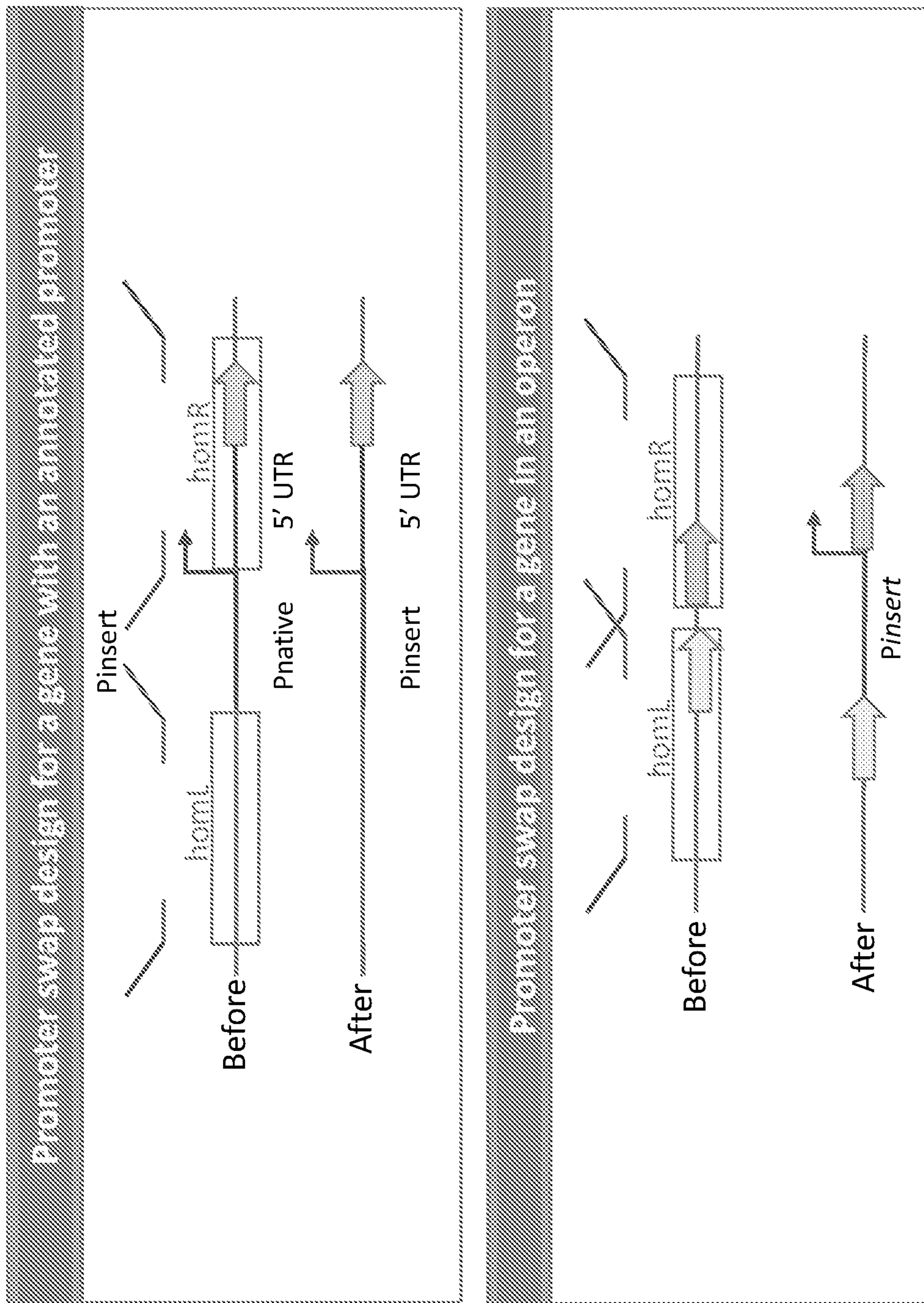


FIGURE 22

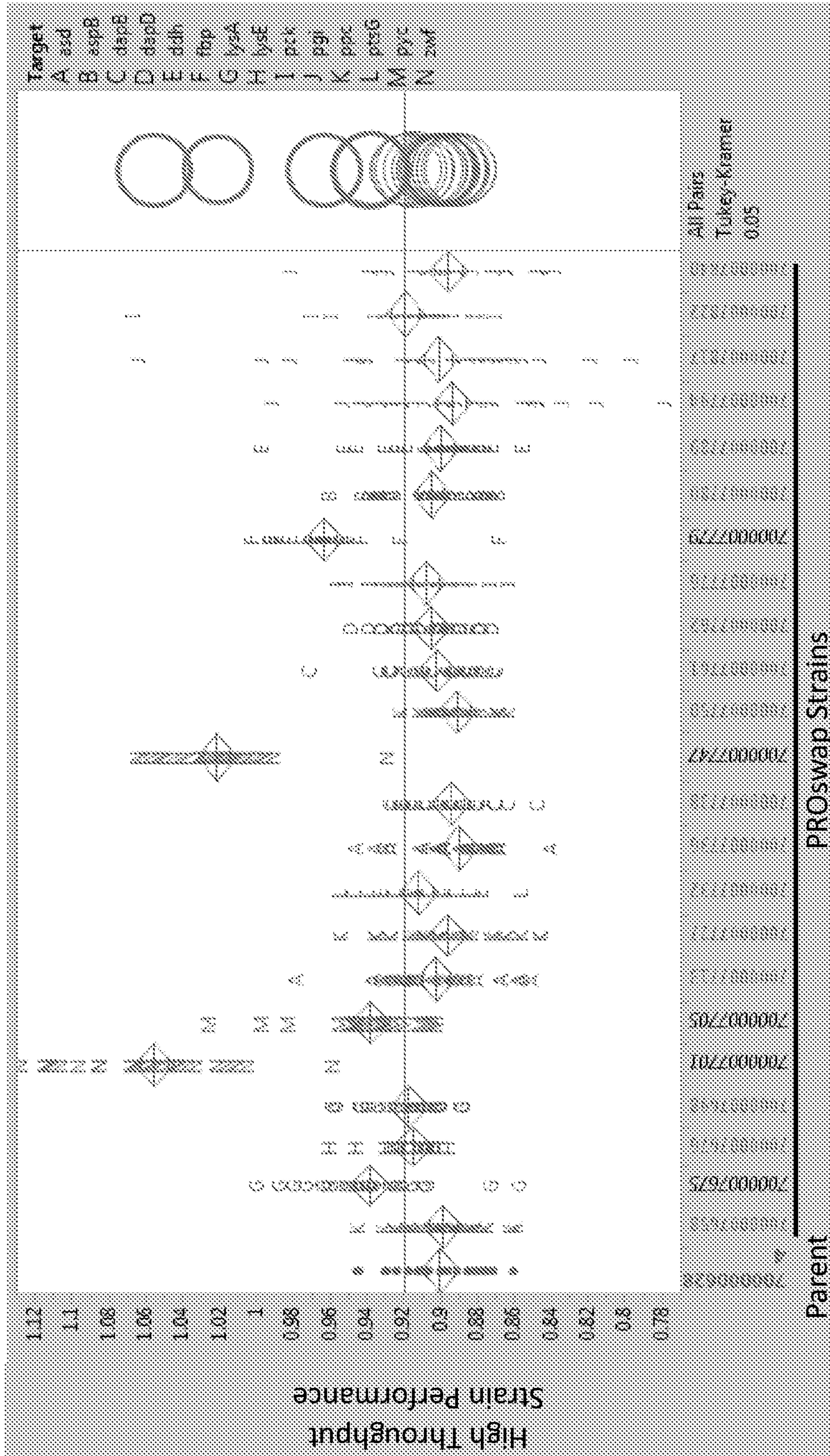


FIGURE 23

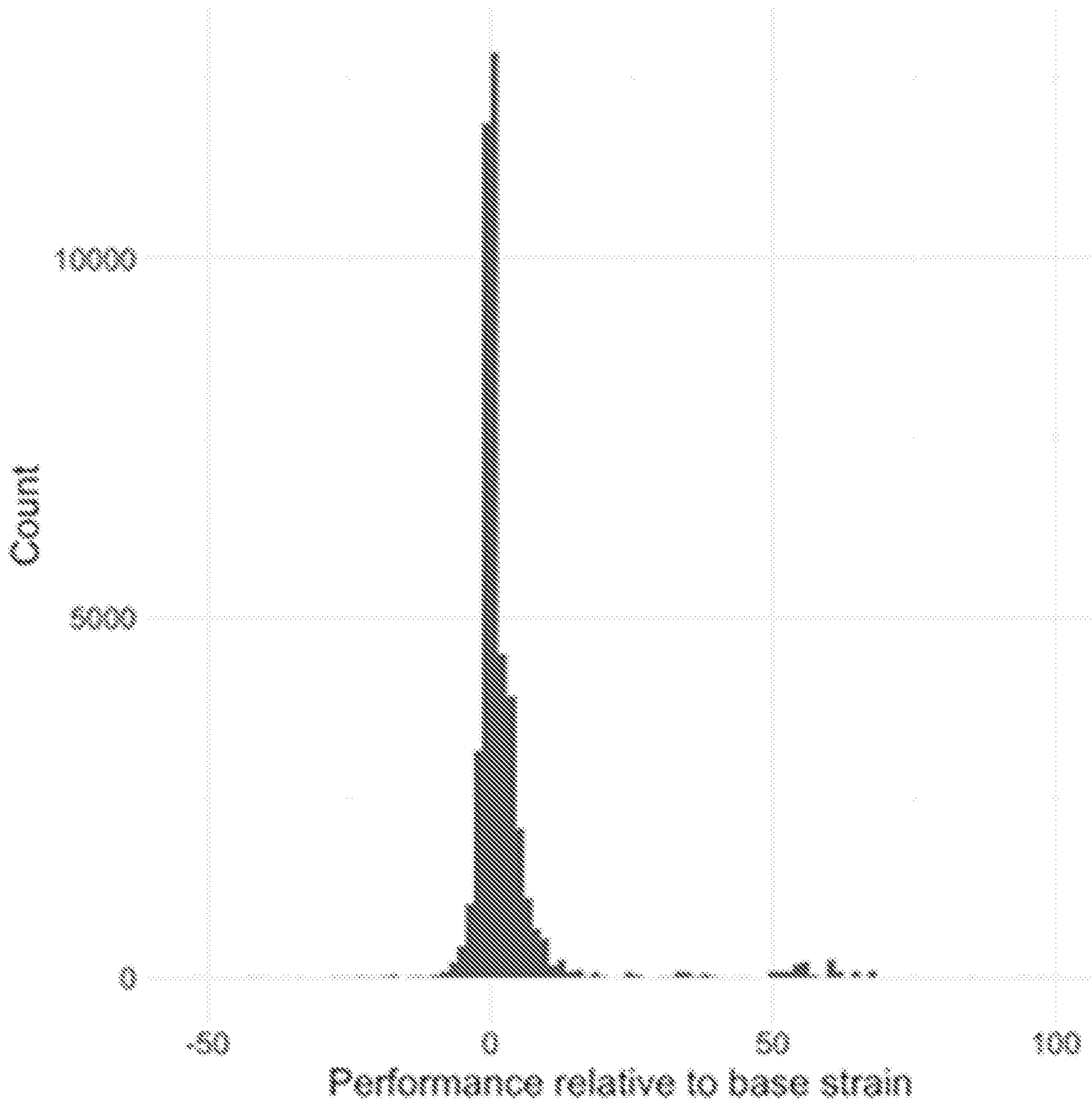


FIGURE 24

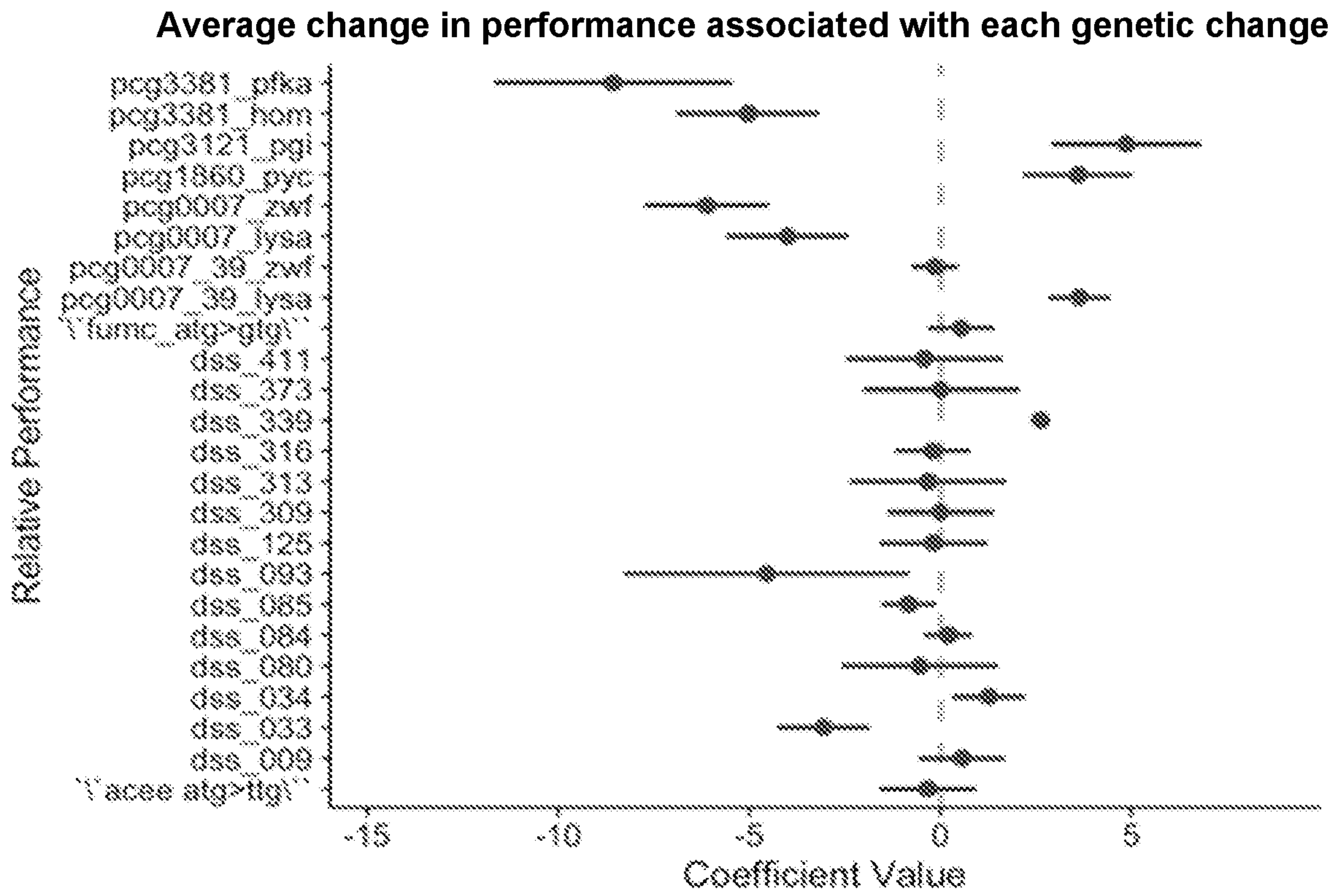


FIGURE 25

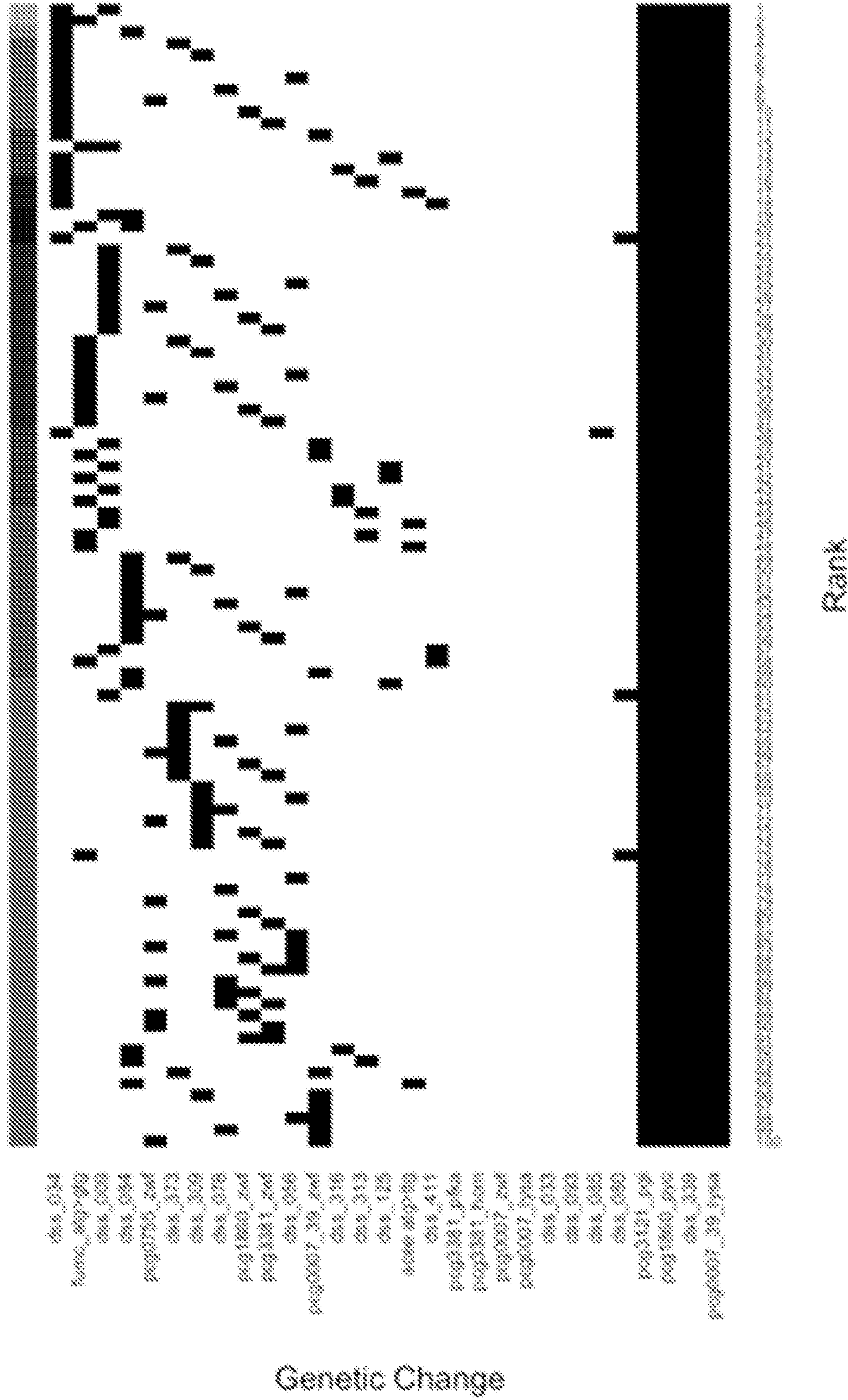


FIGURE 26

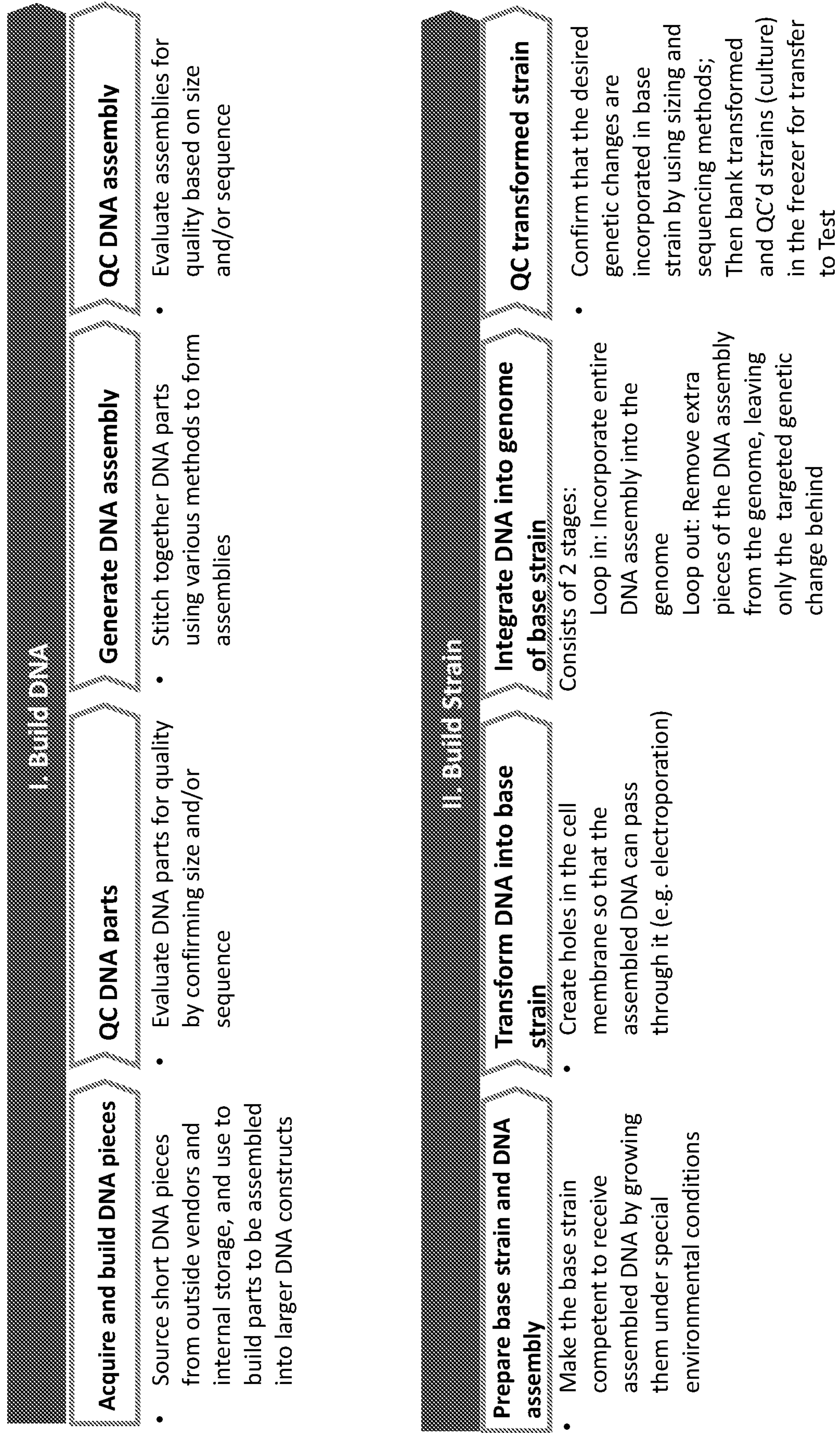


FIGURE 27

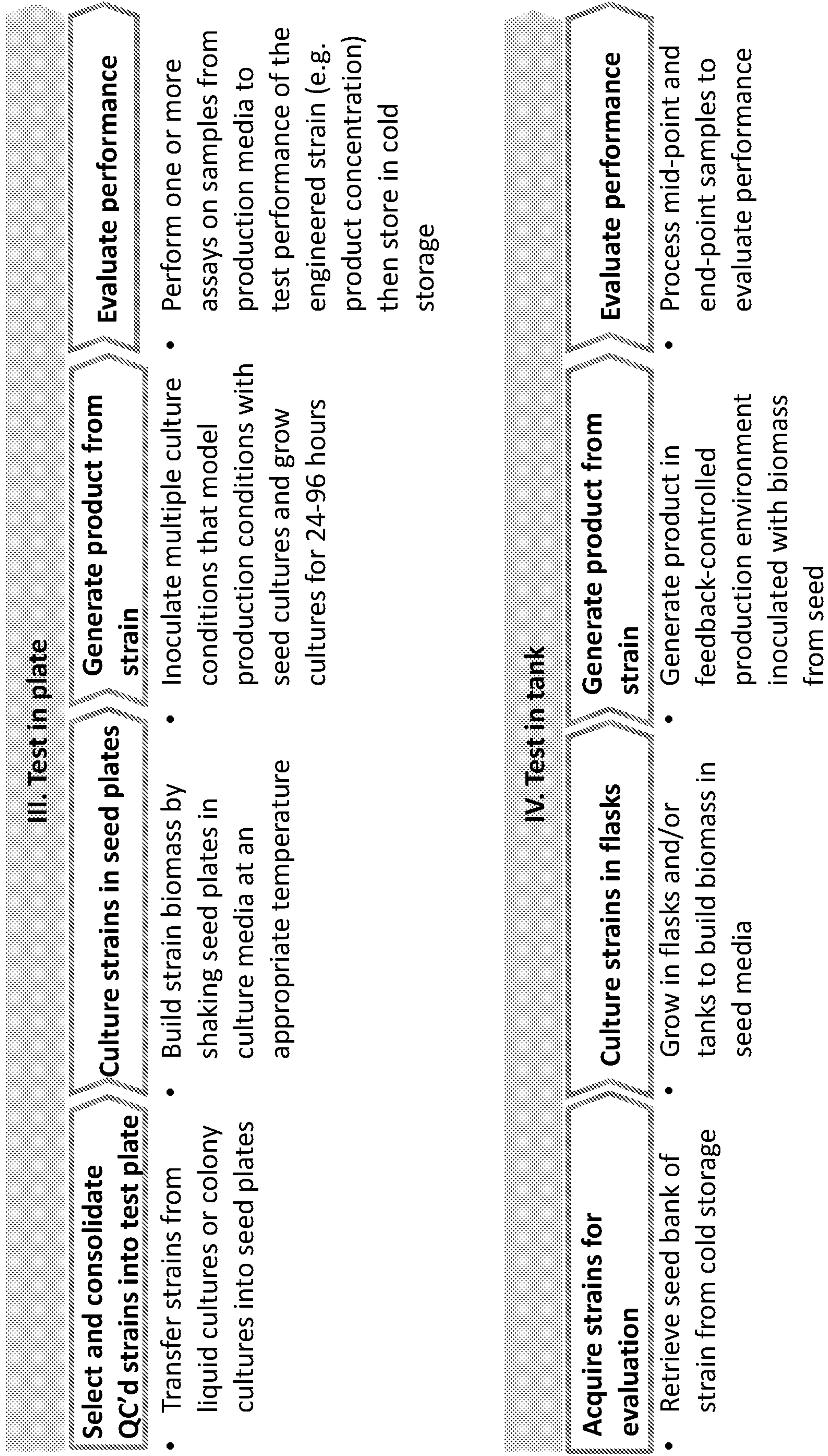


FIGURE 28

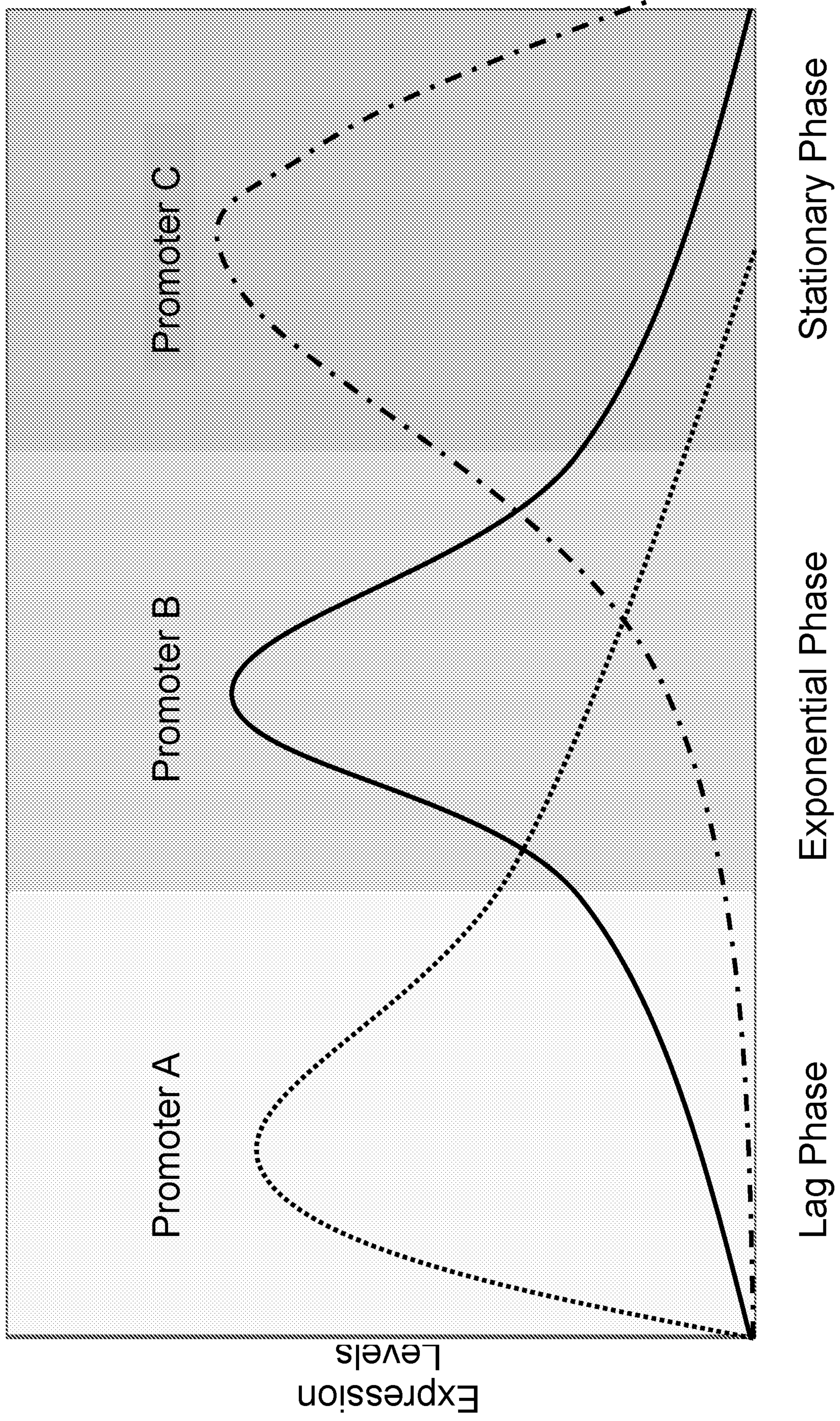


FIGURE 29

Substrate Concentration Over Time

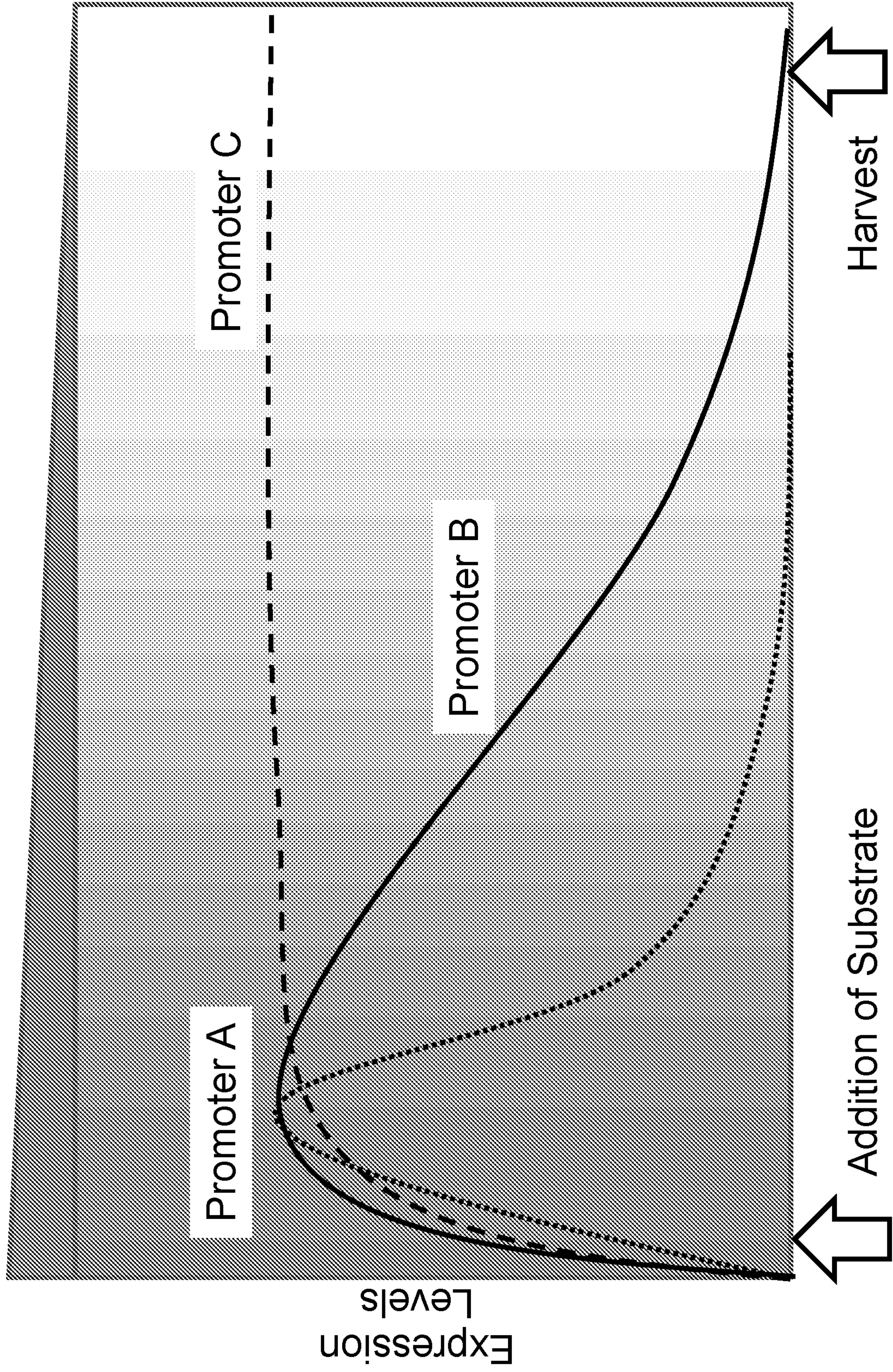


FIGURE 30

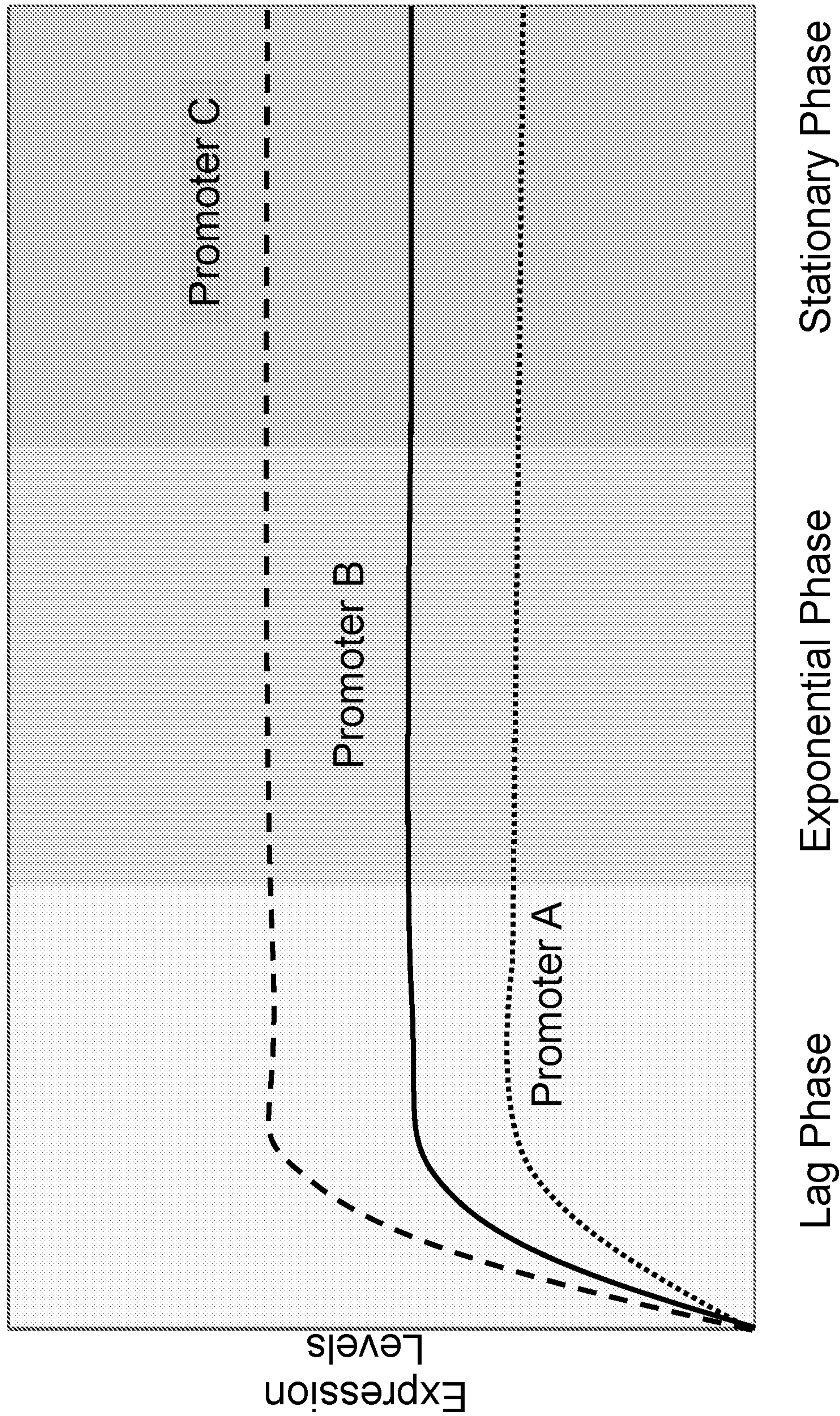


FIGURE 31

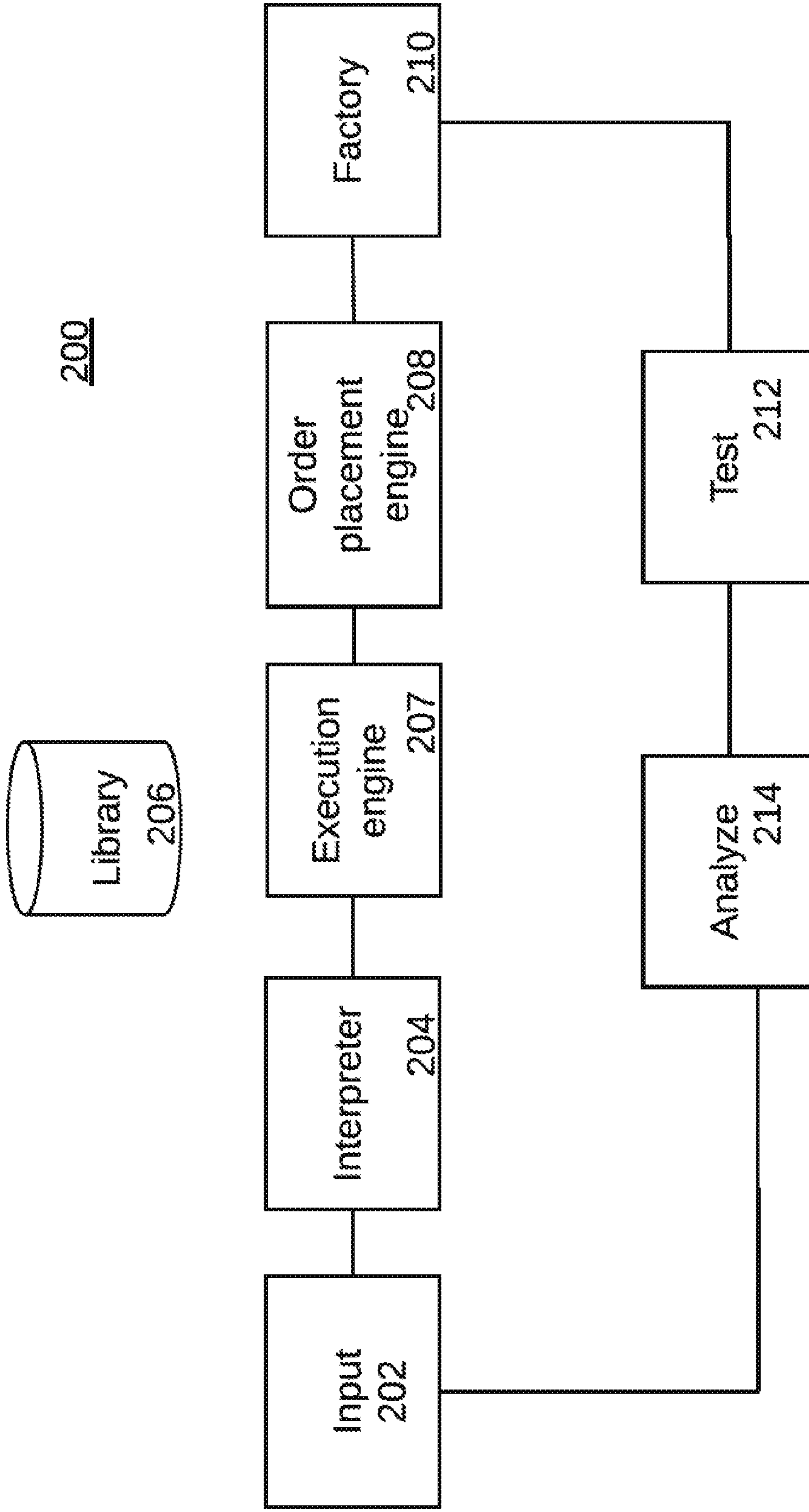


FIGURE 32

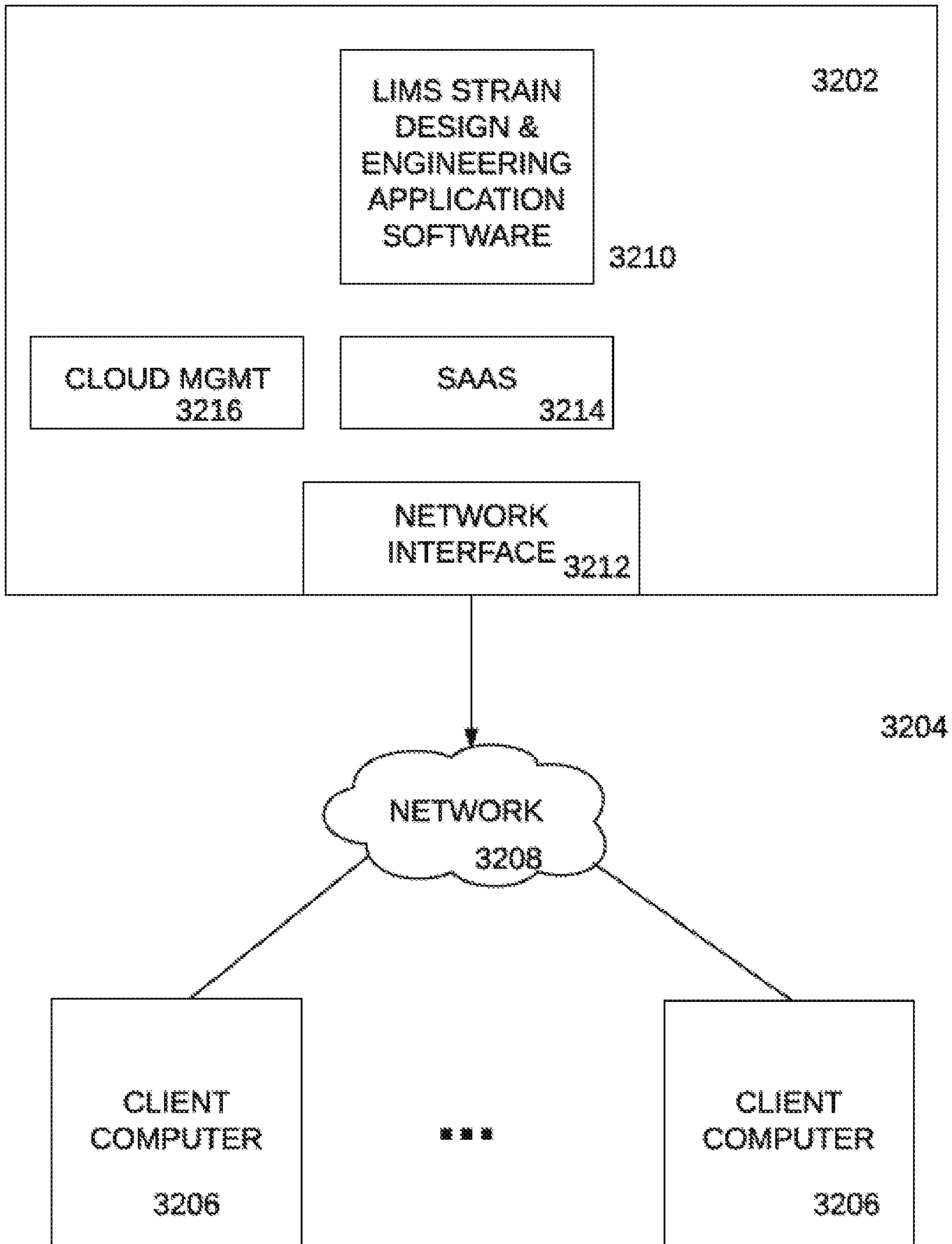


FIGURE 33

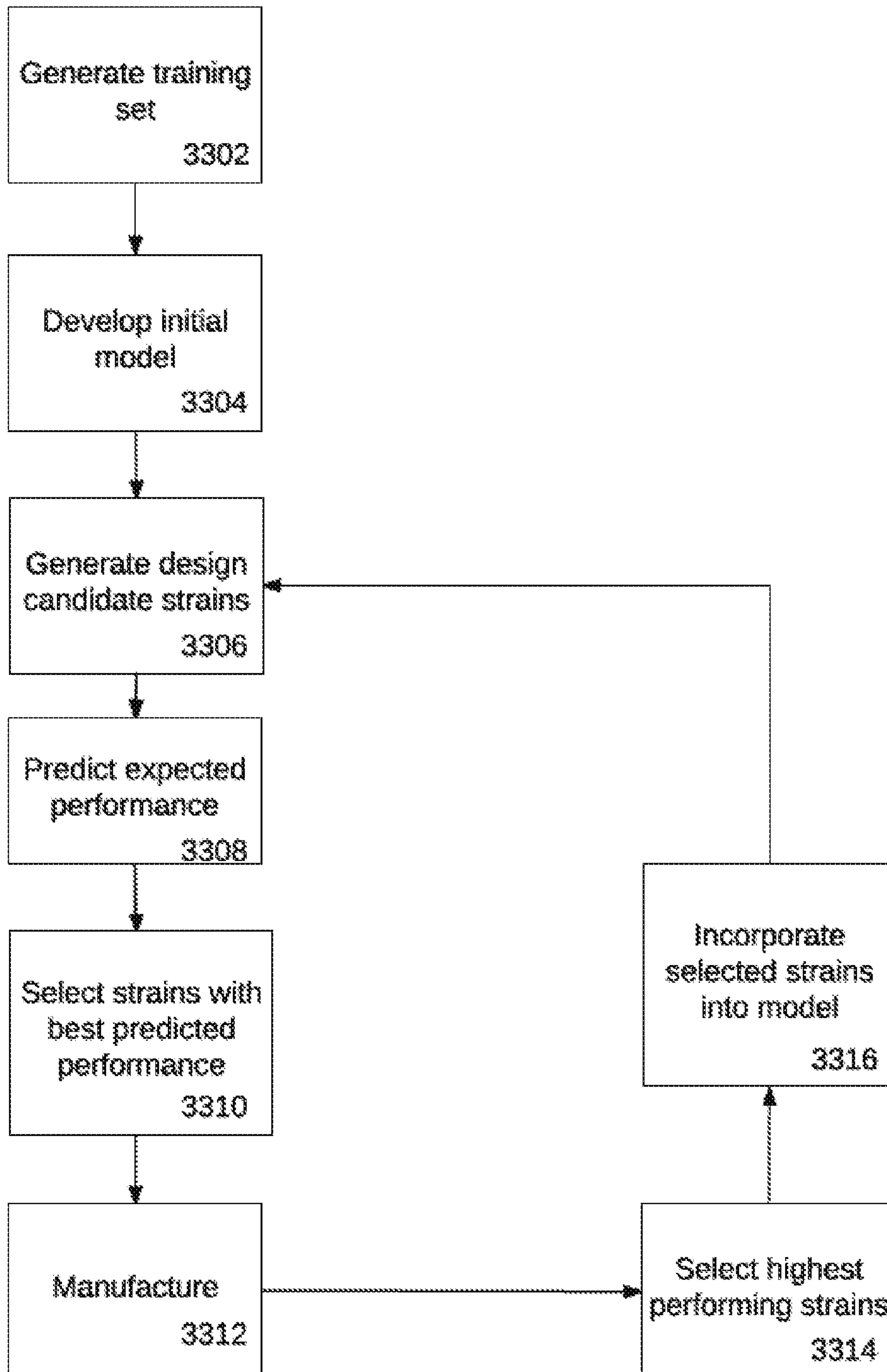


FIGURE 34

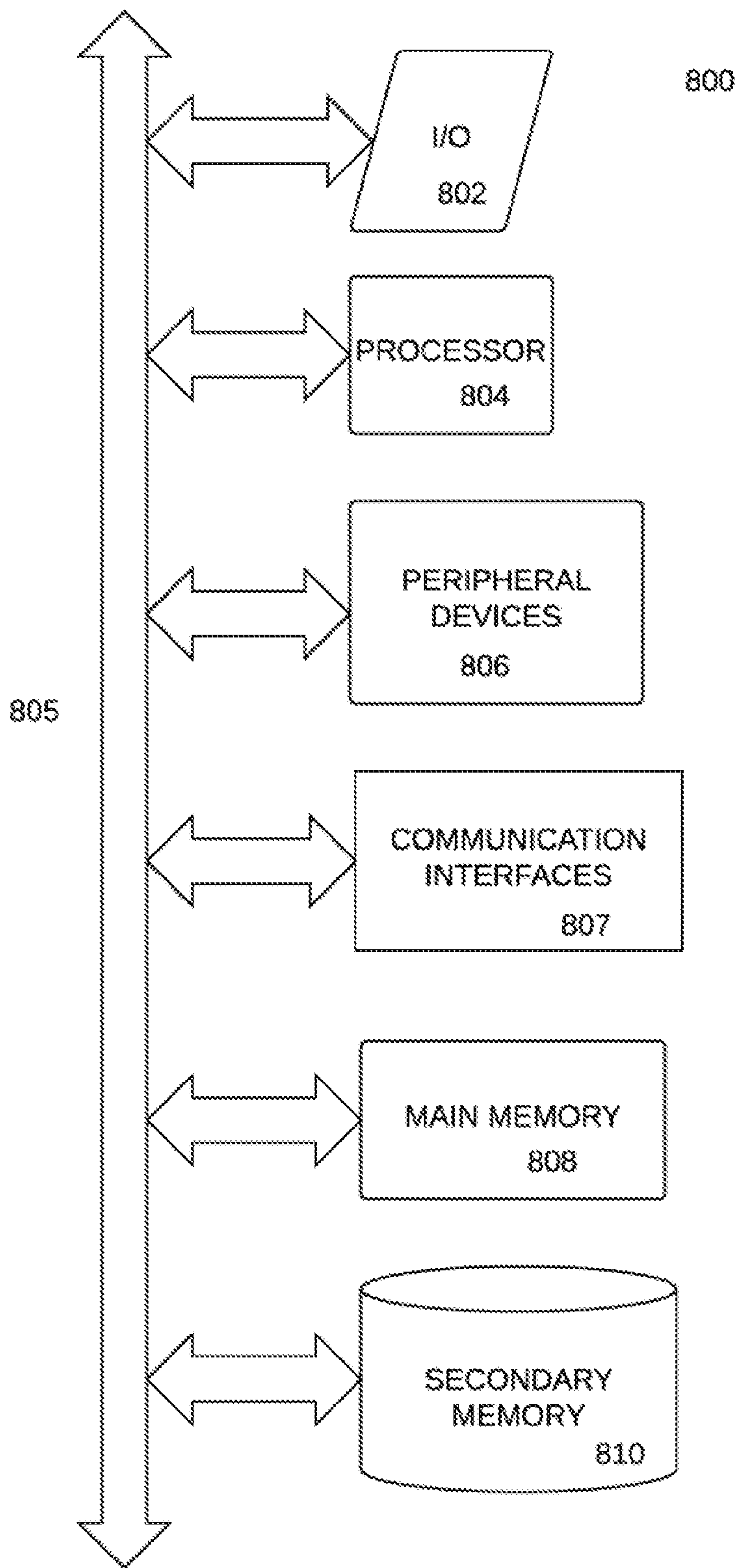


FIGURE 35

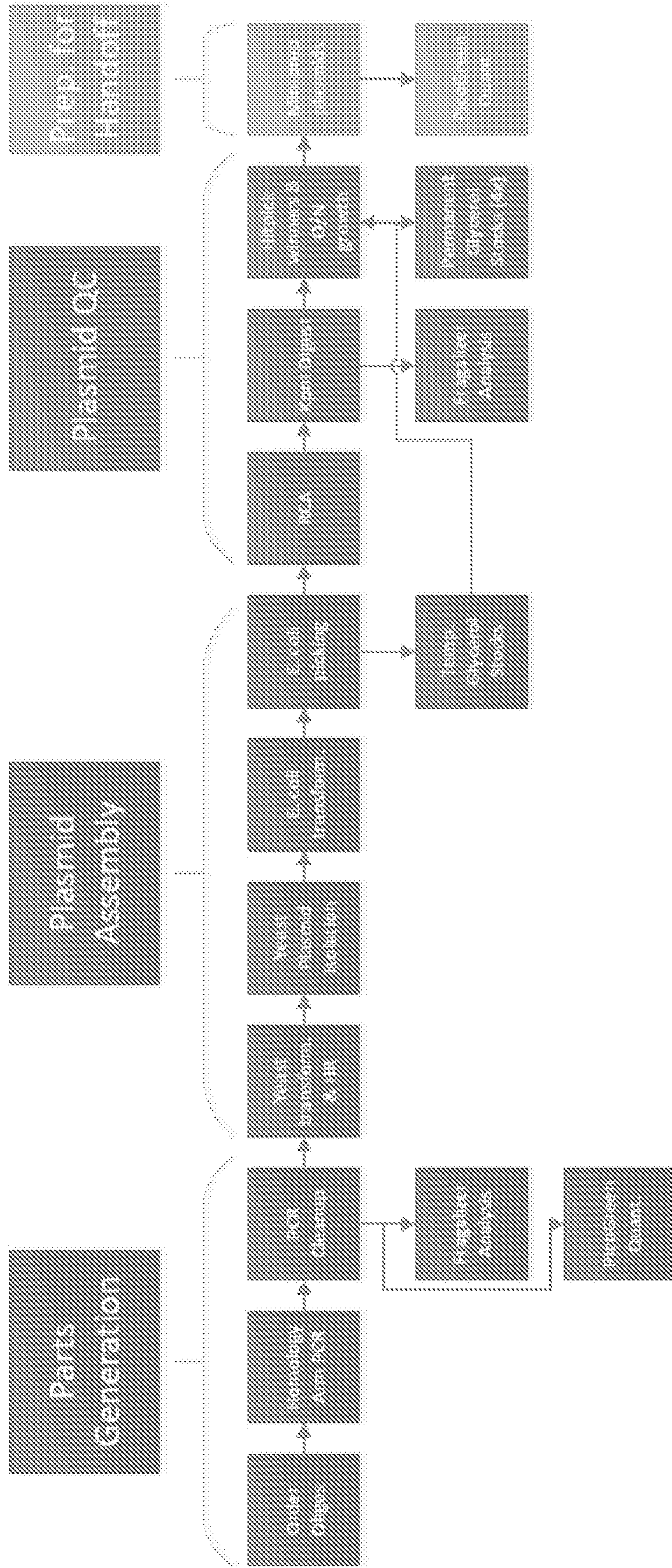


FIGURE 36

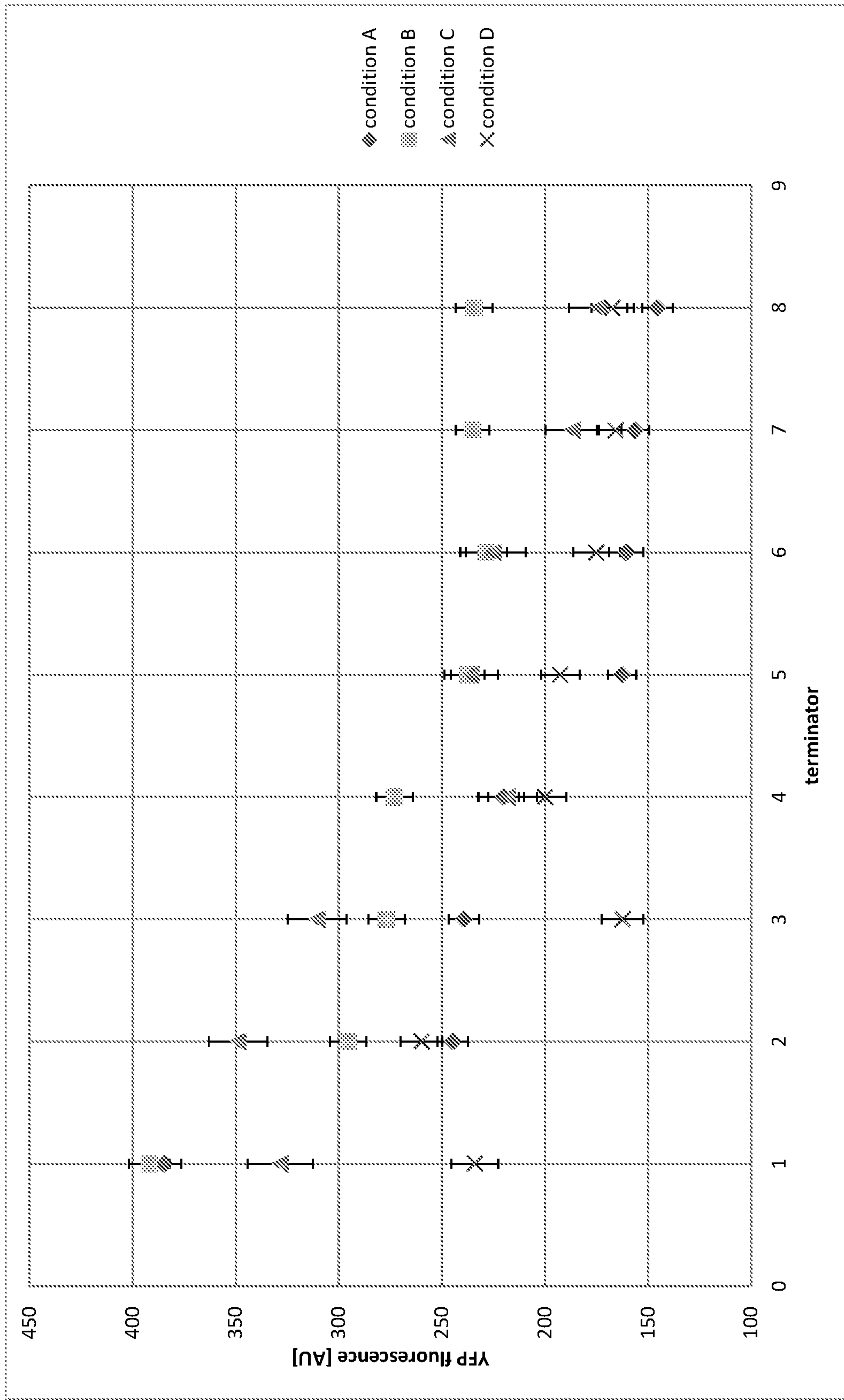


FIGURE 37

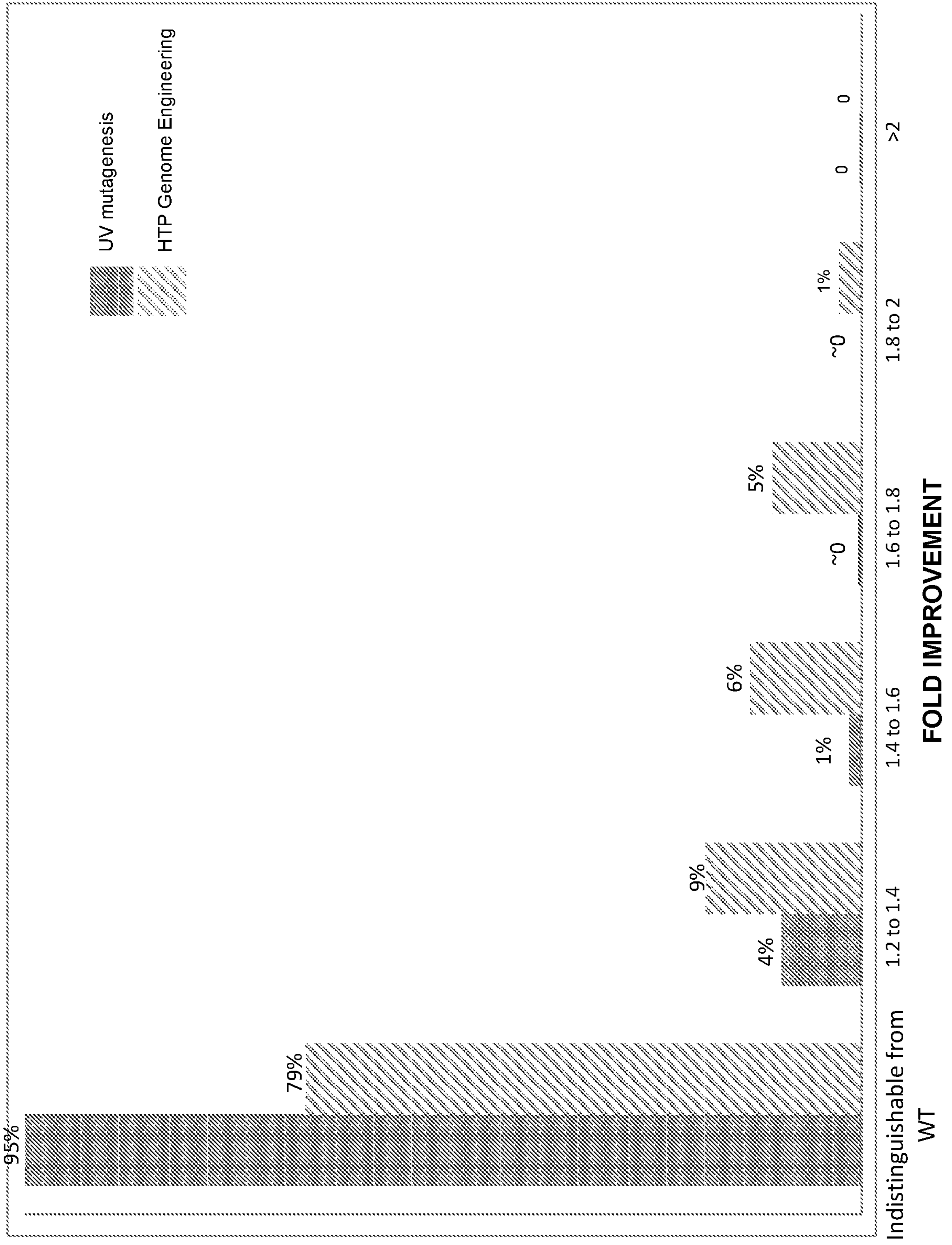


FIGURE 38

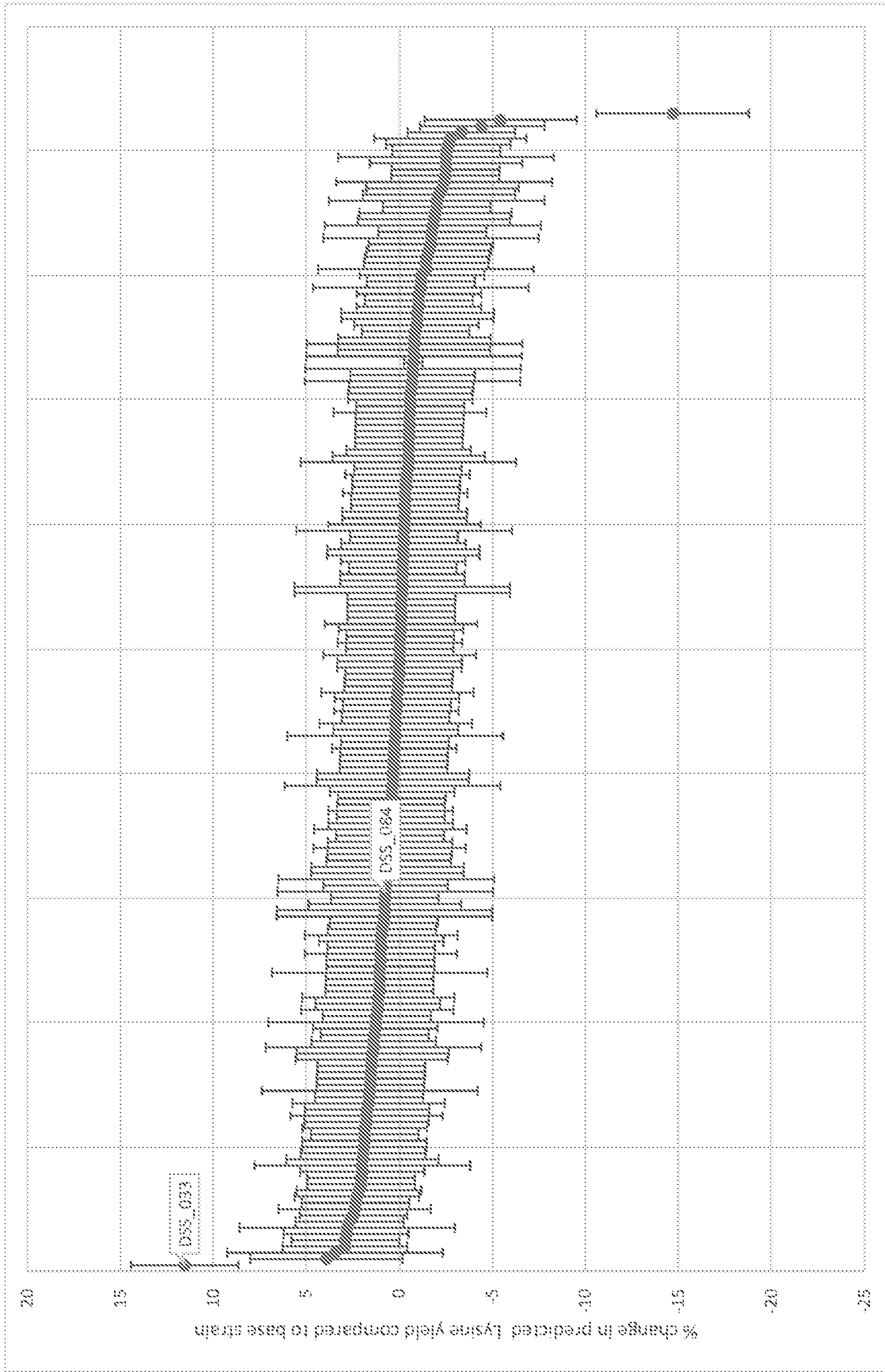


FIGURE 39

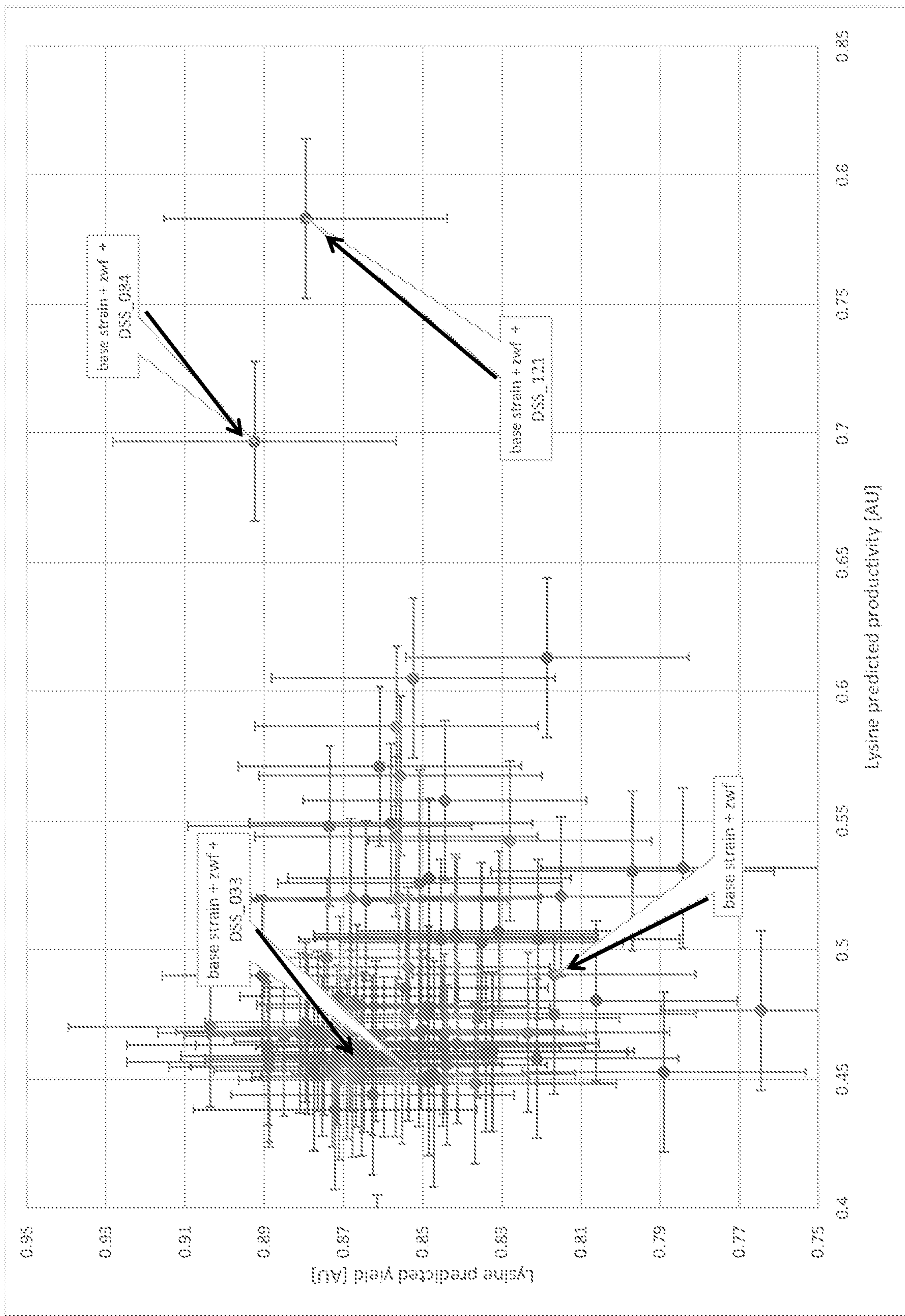


FIGURE 40

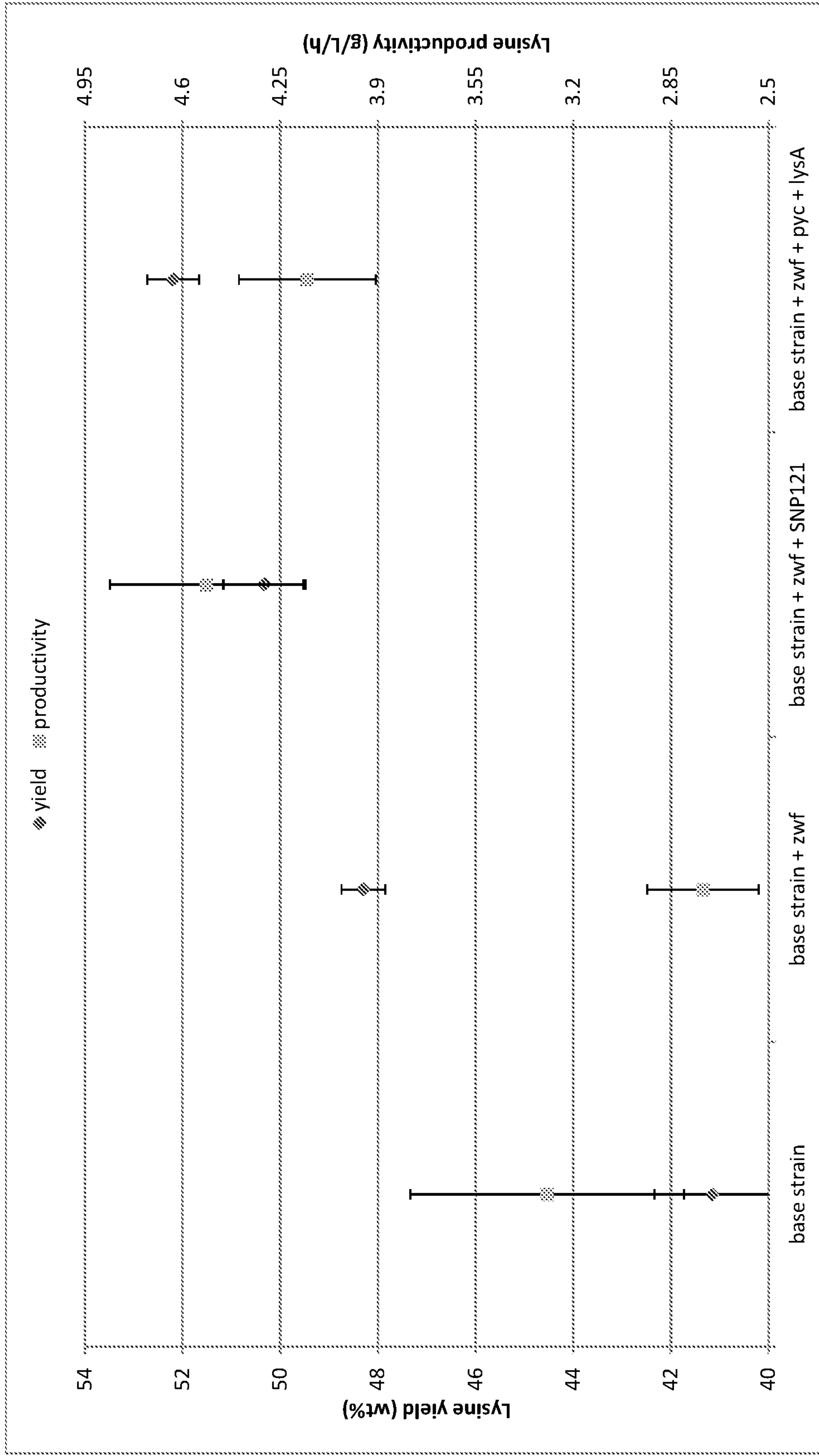


FIGURE 41

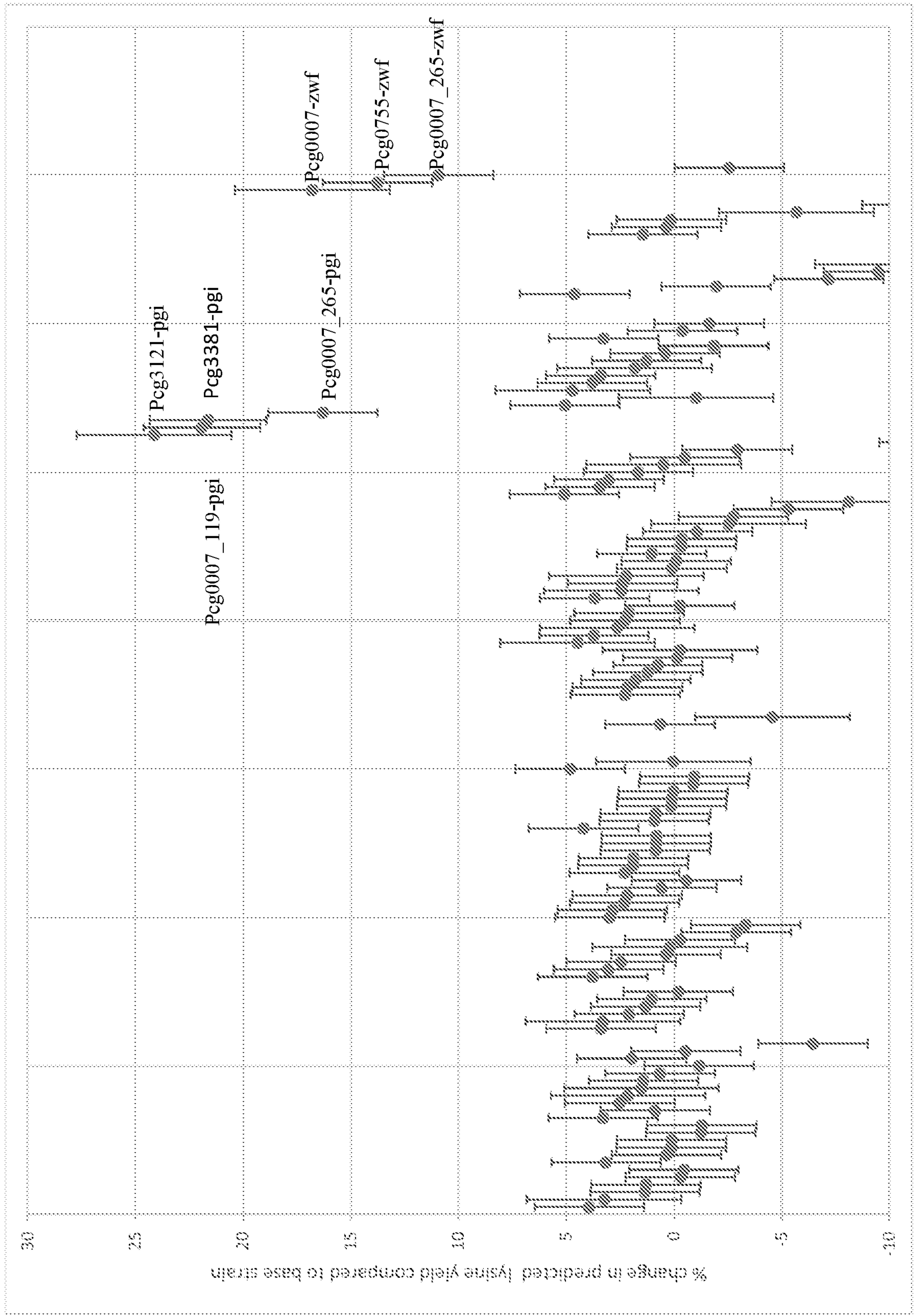


FIGURE 42

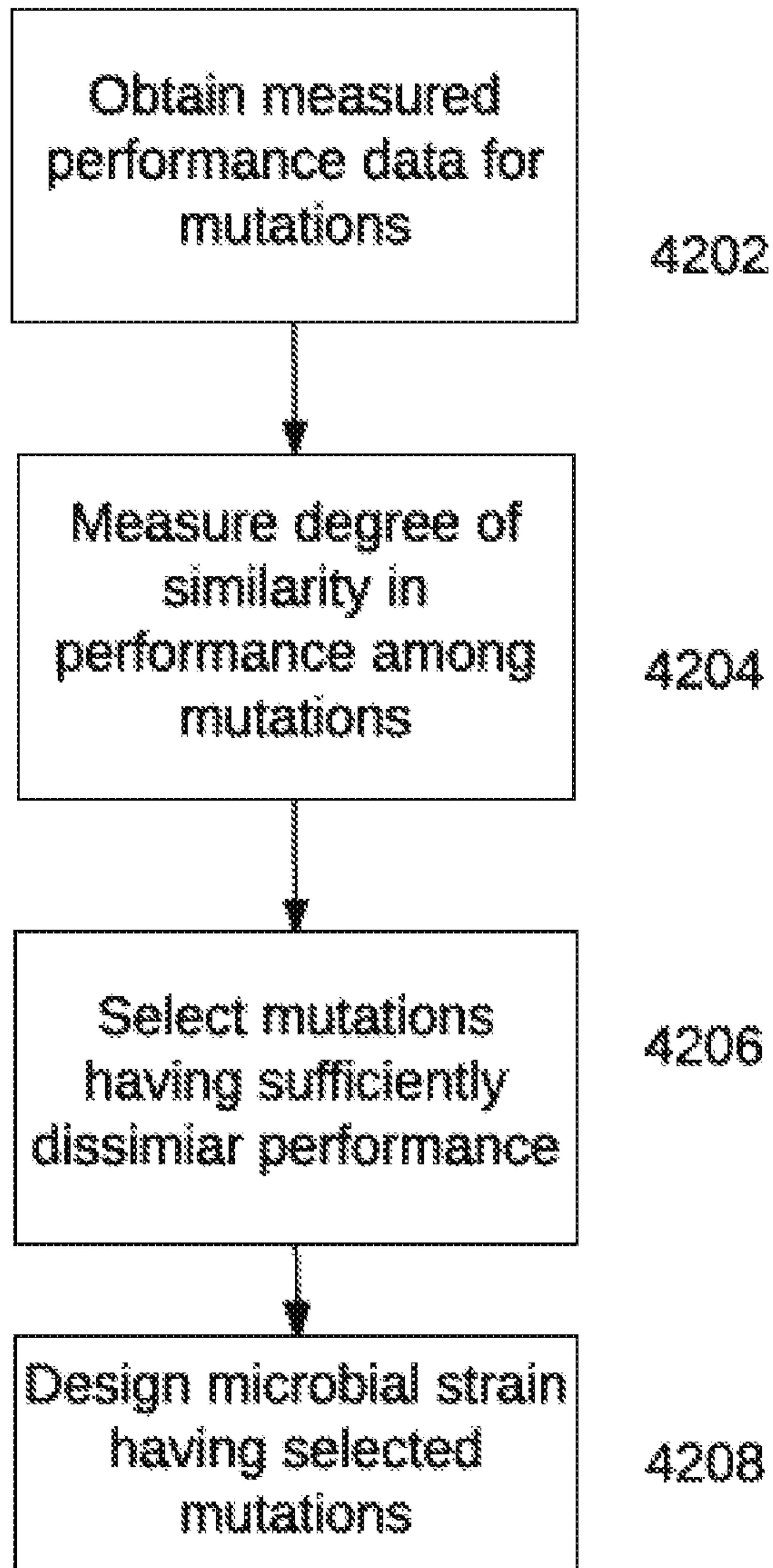


FIGURE 43

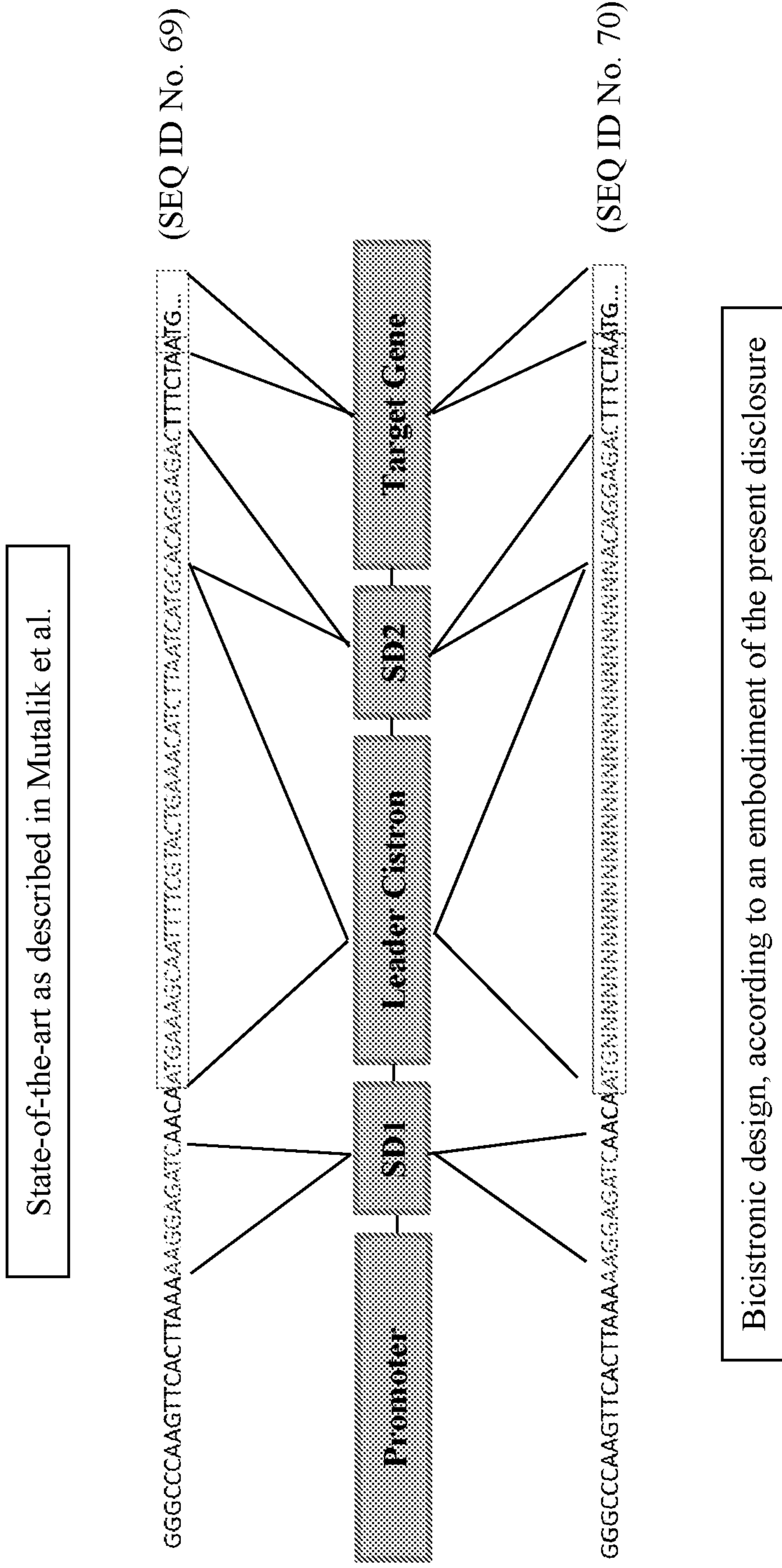


FIGURE 44

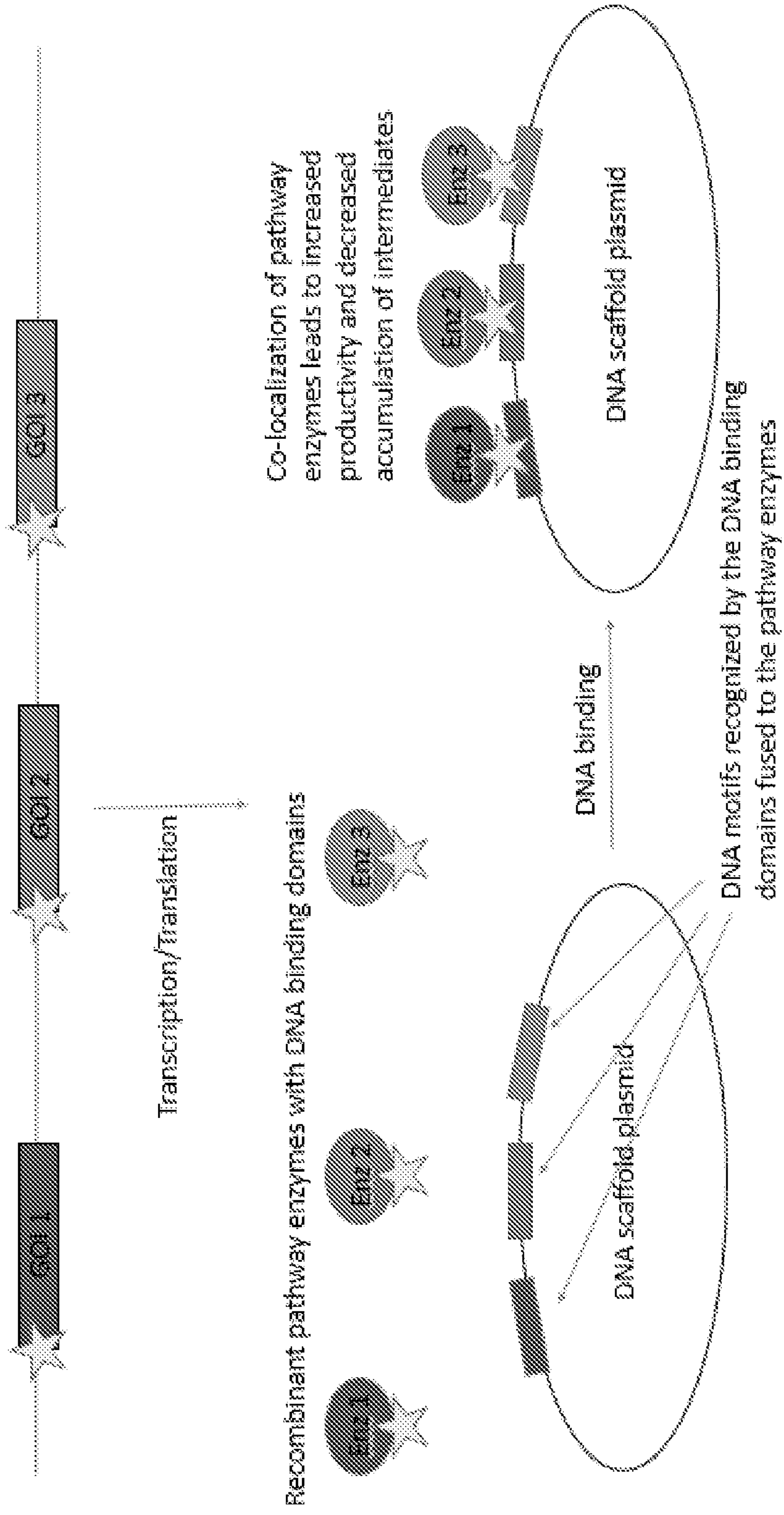
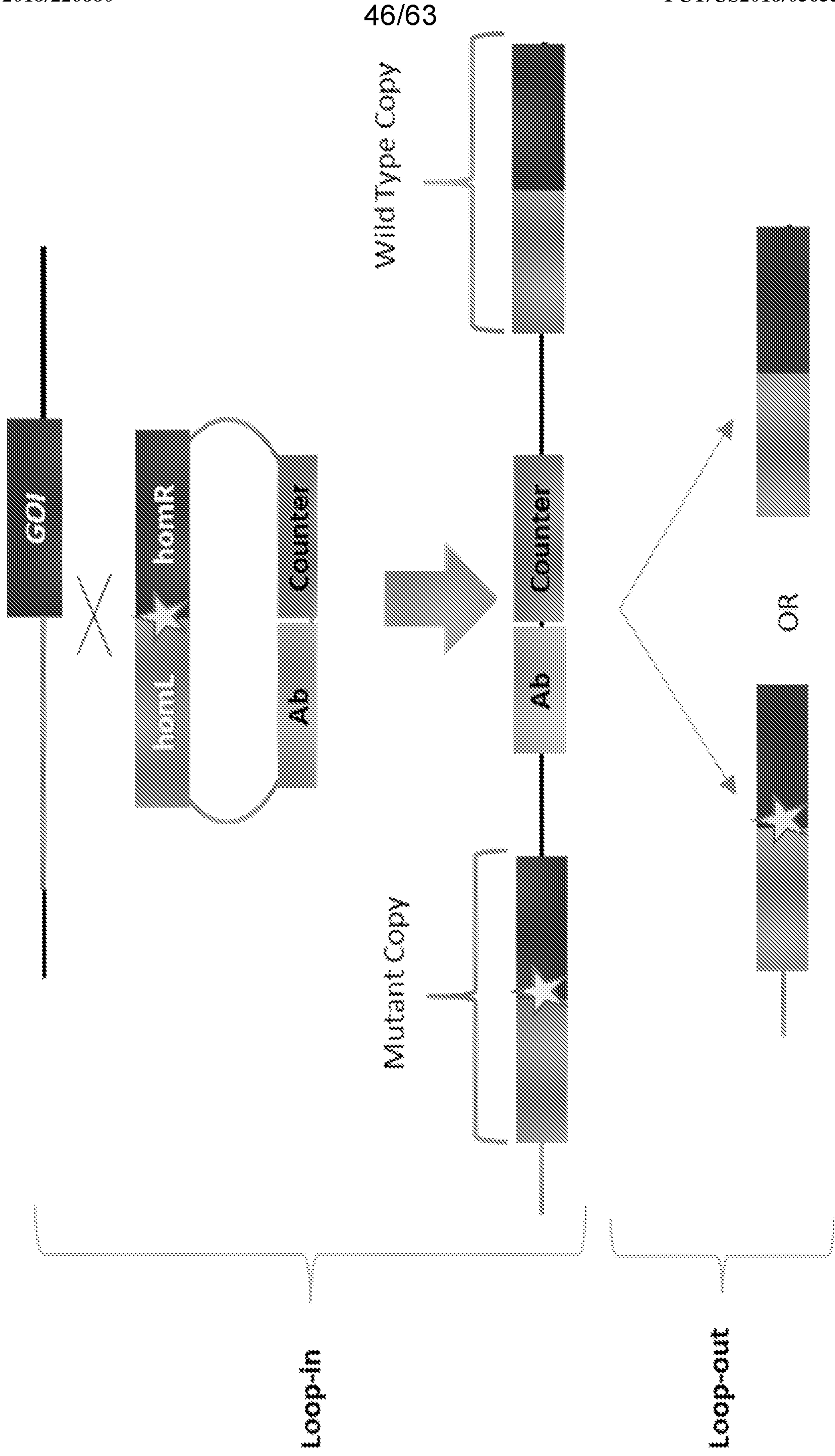


FIGURE 45



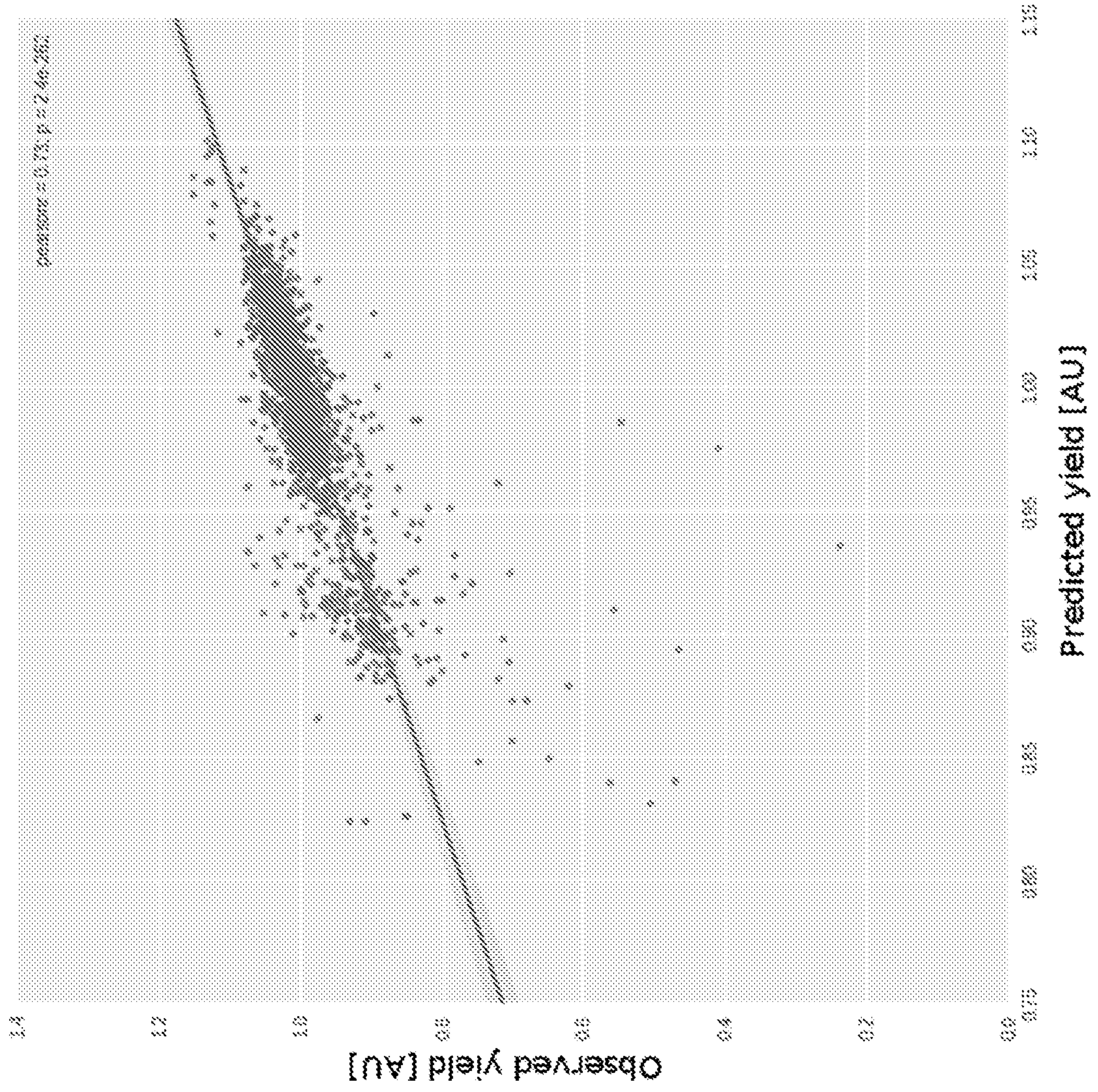


FIGURE 46

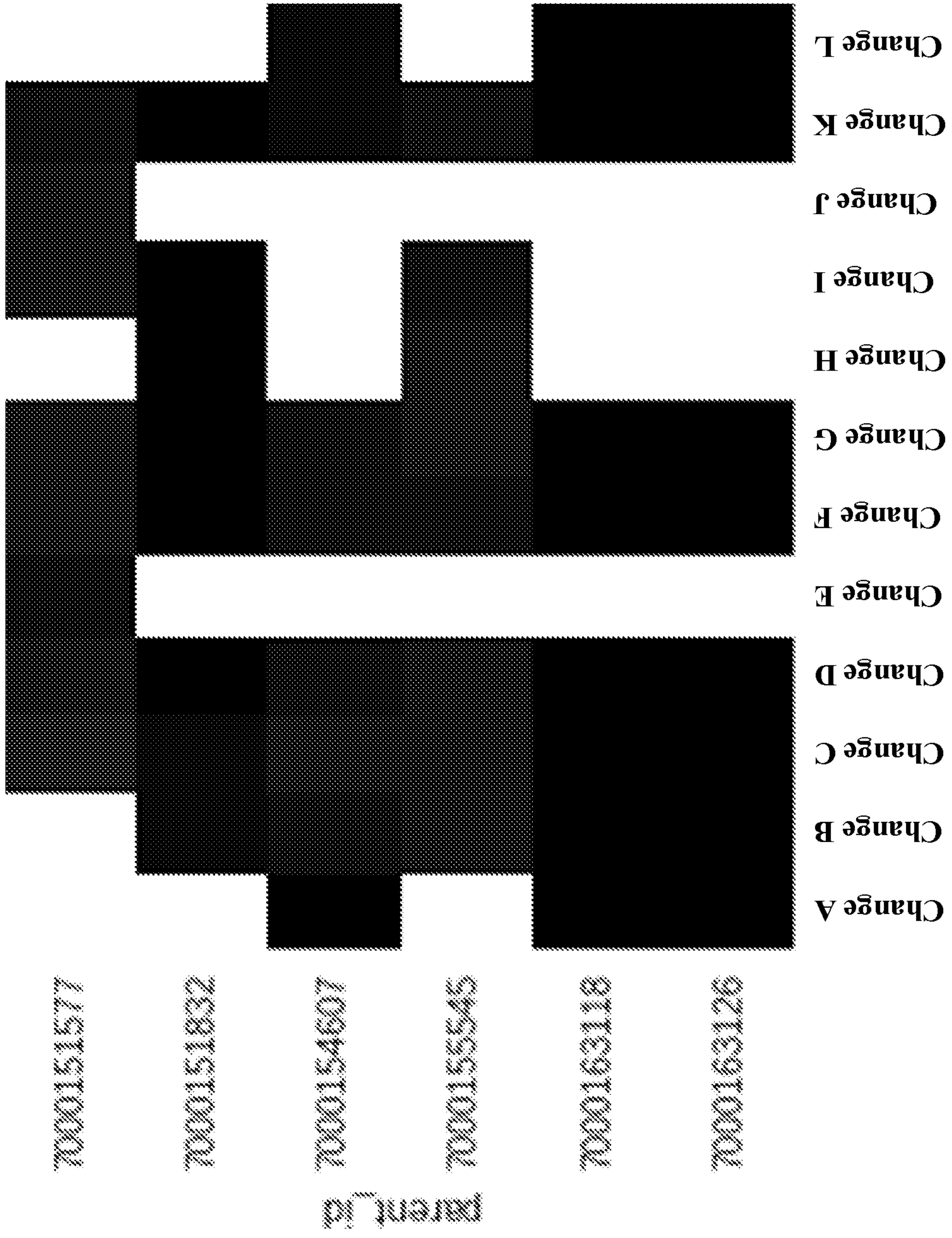


FIGURE 47

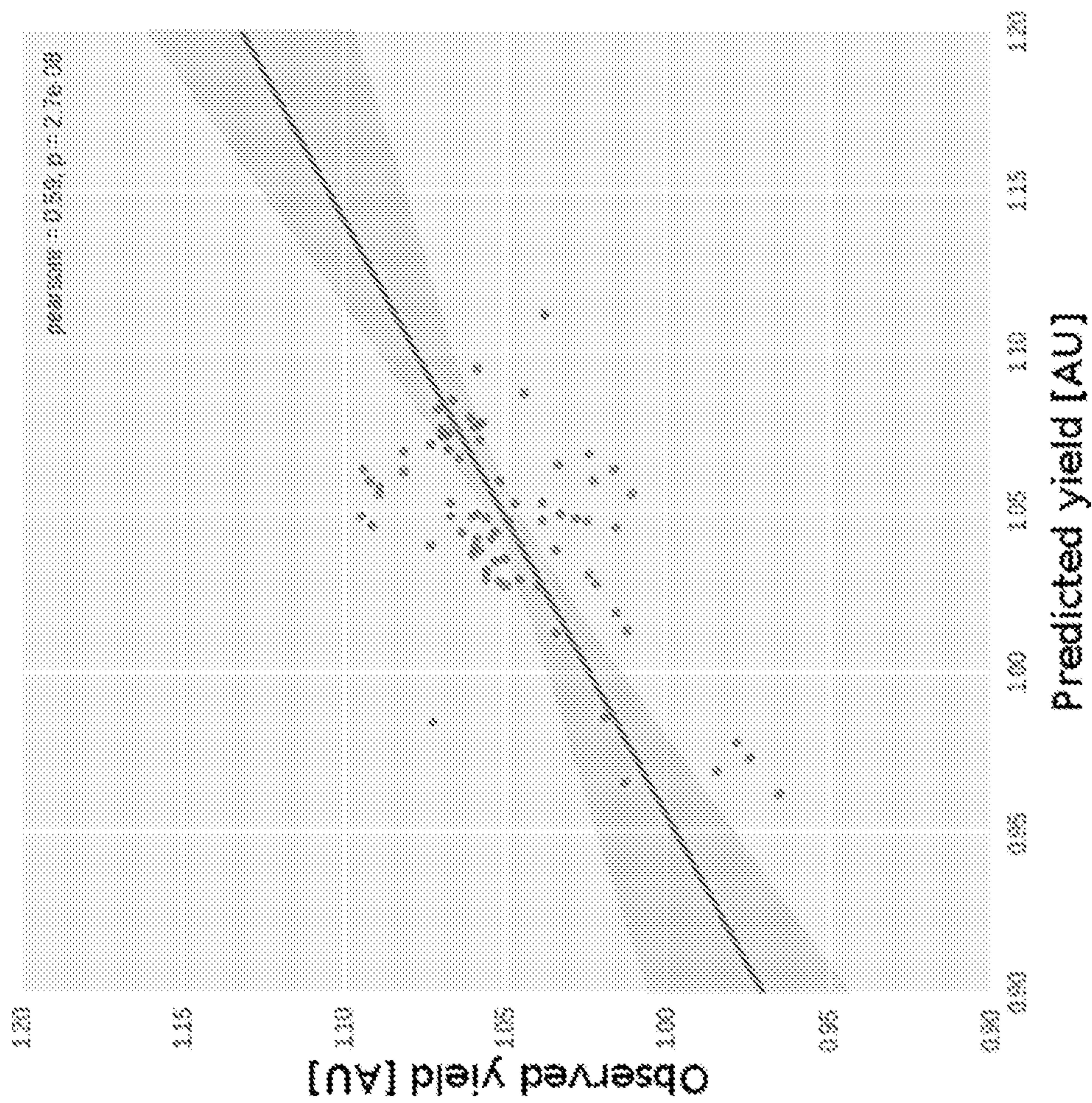


FIGURE 48

FIGURE 49

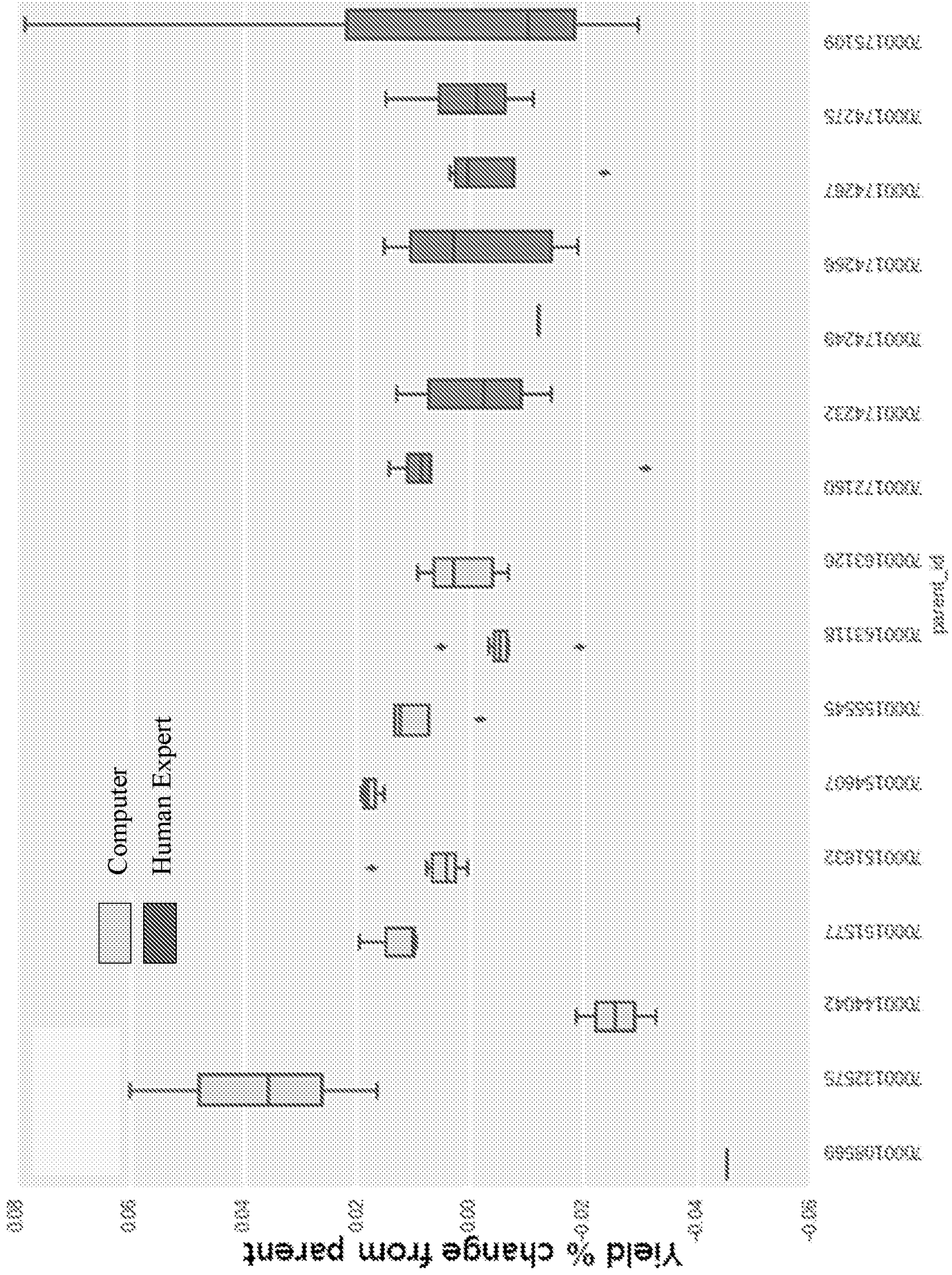


FIGURE 50

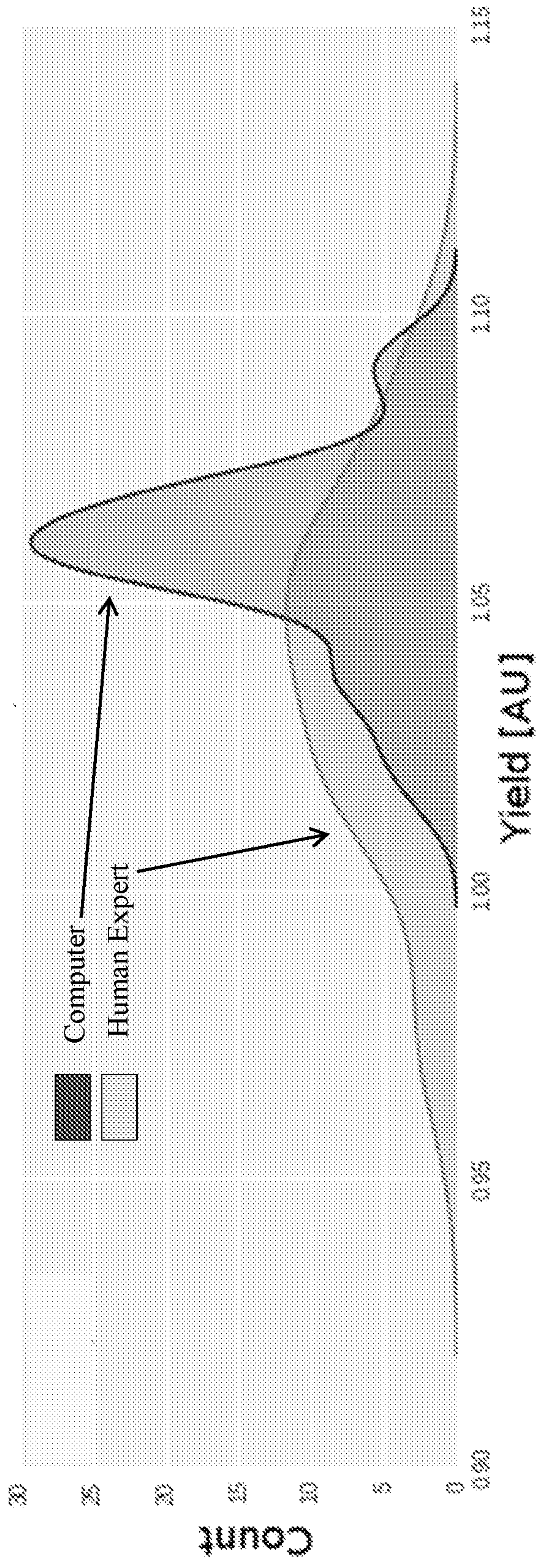


FIGURE 51

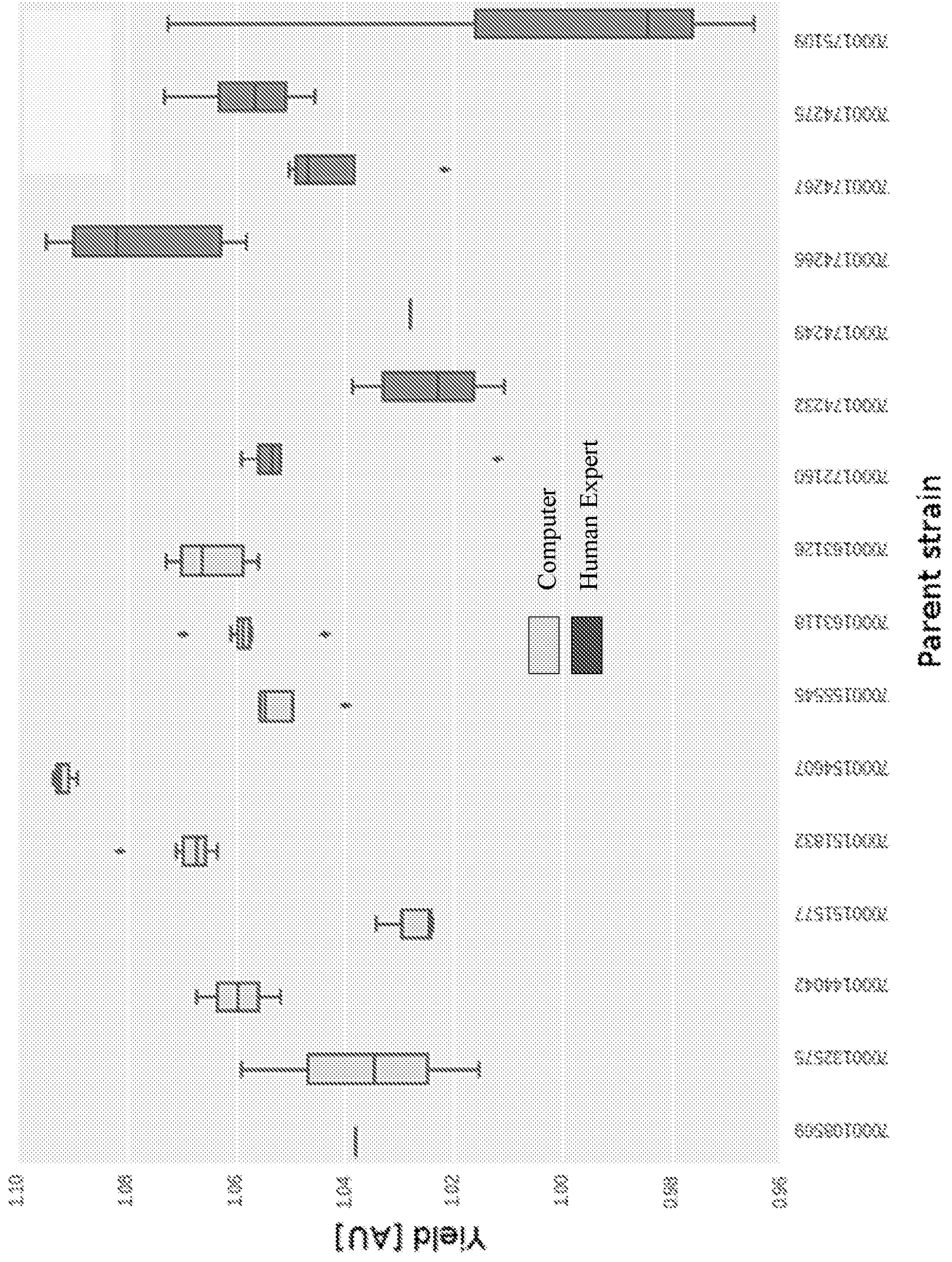


FIGURE 52

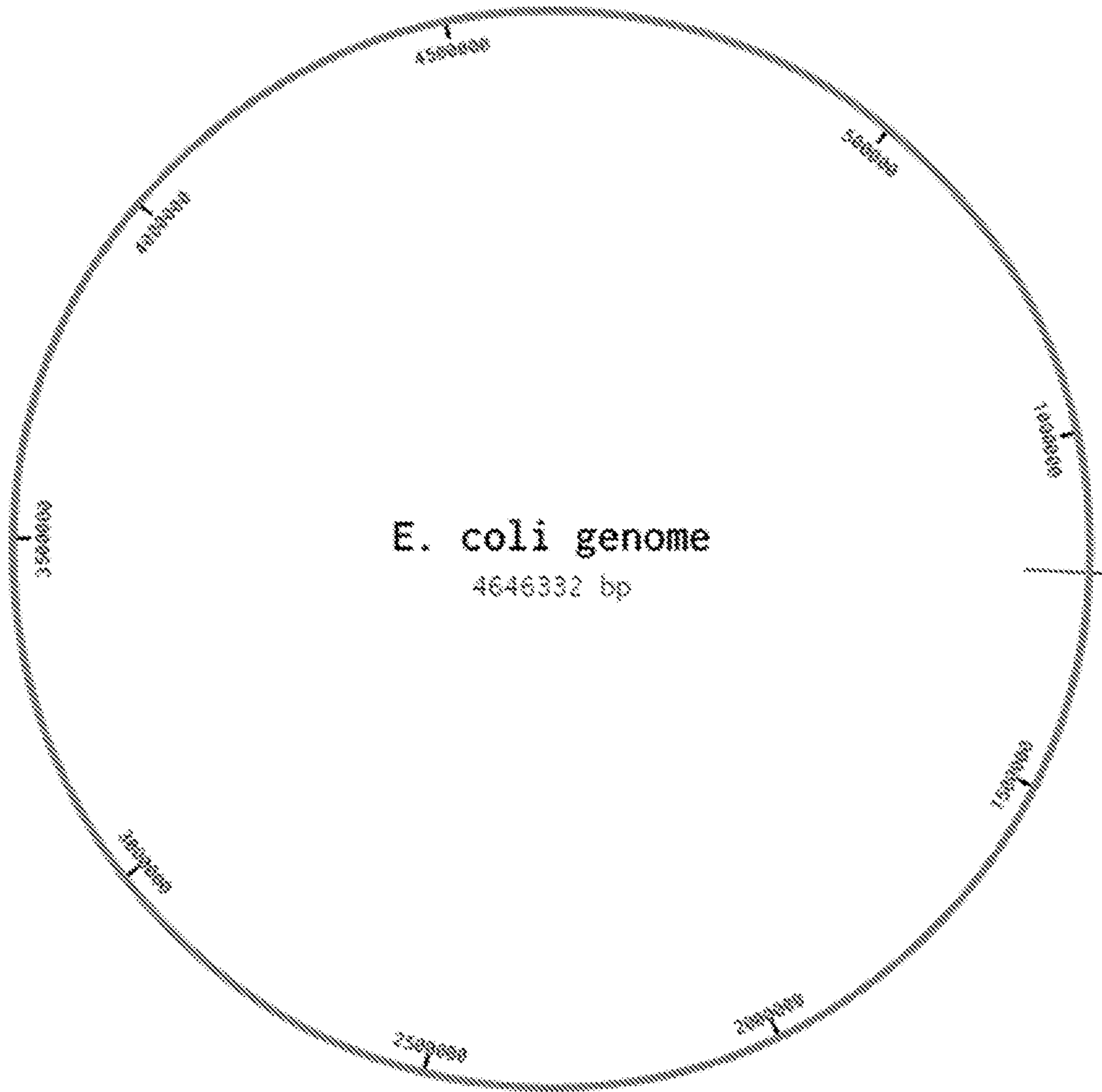


FIGURE 53

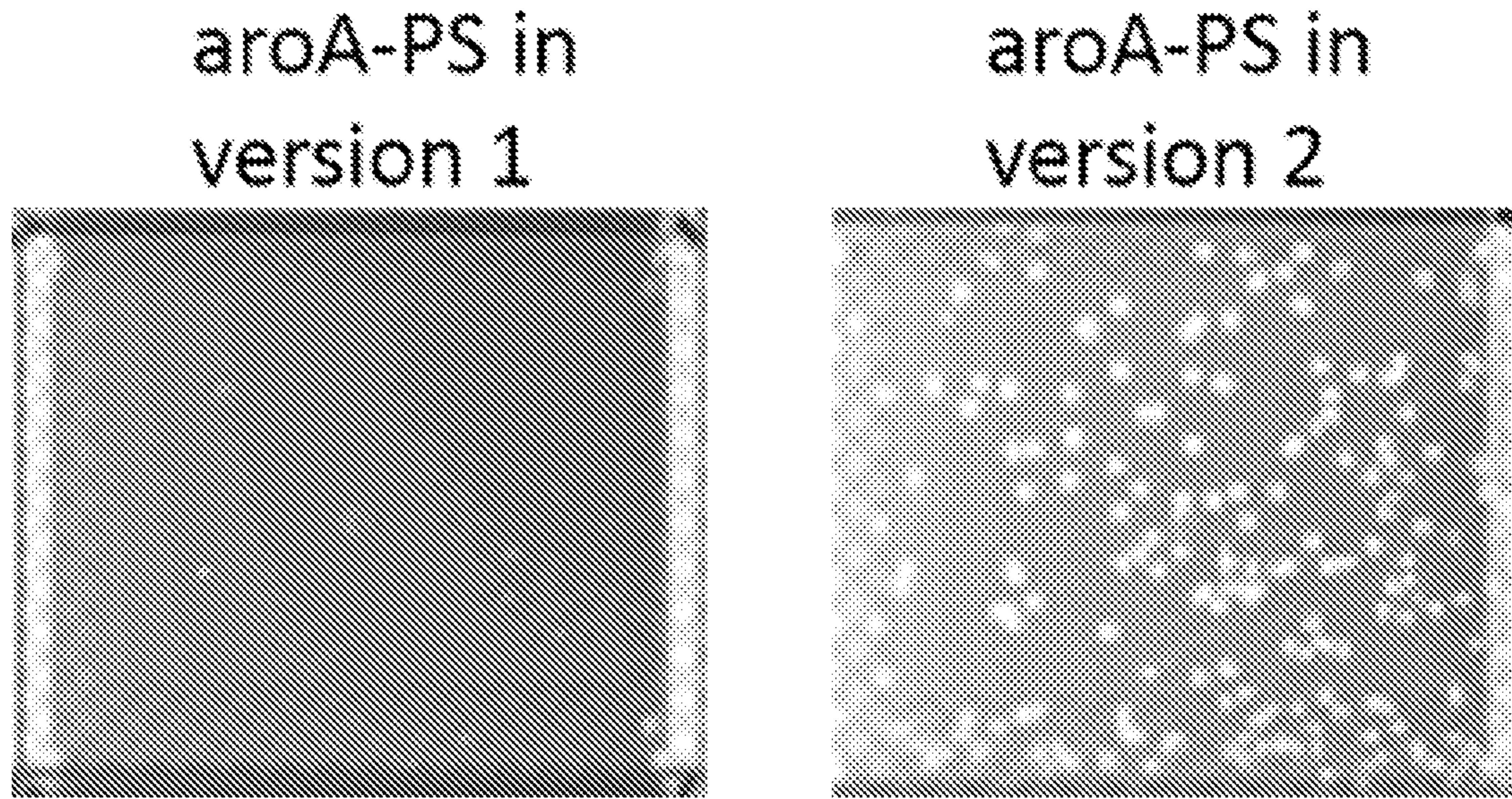


FIGURE 54



FIGURE 55

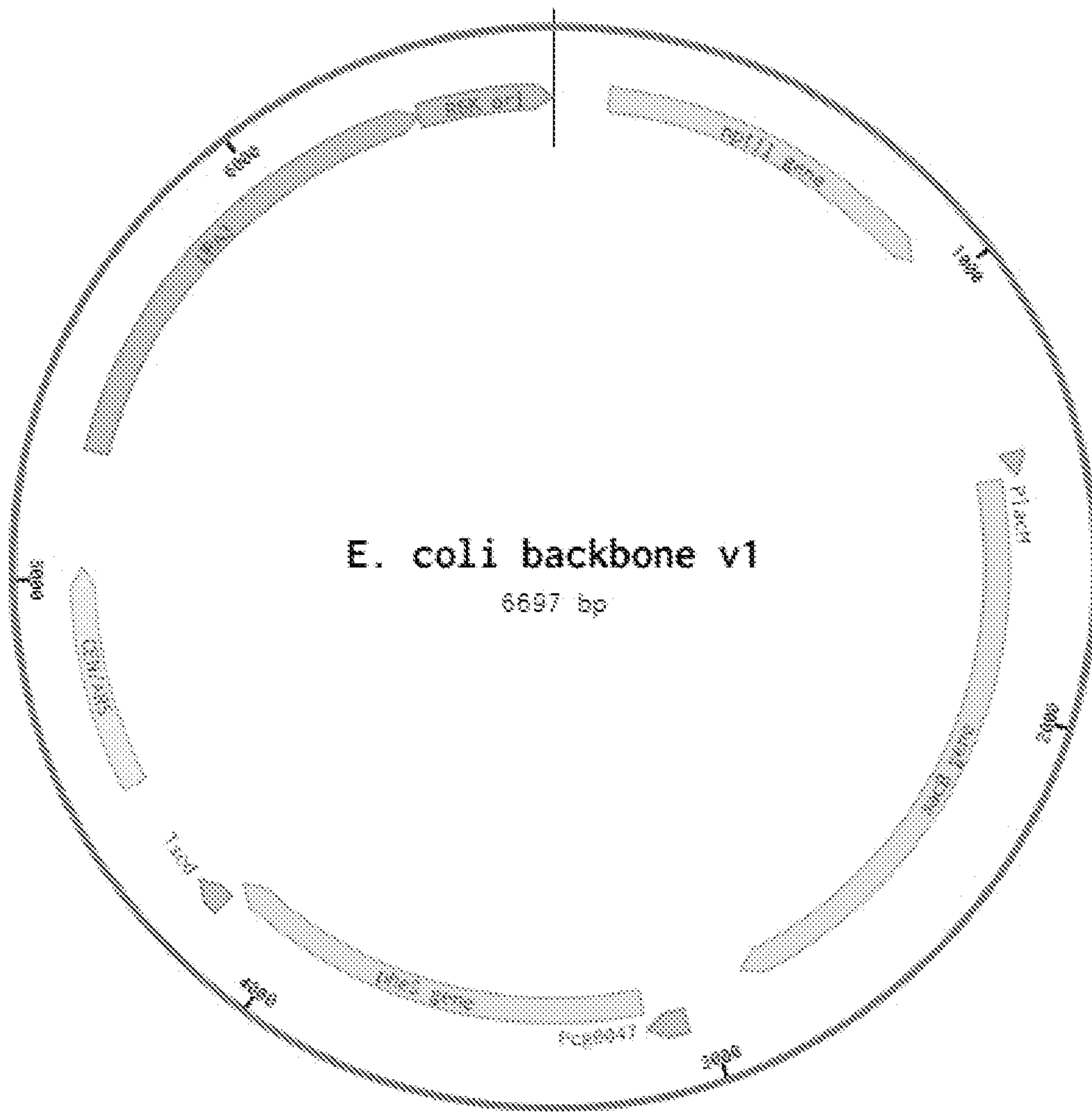


FIGURE 56

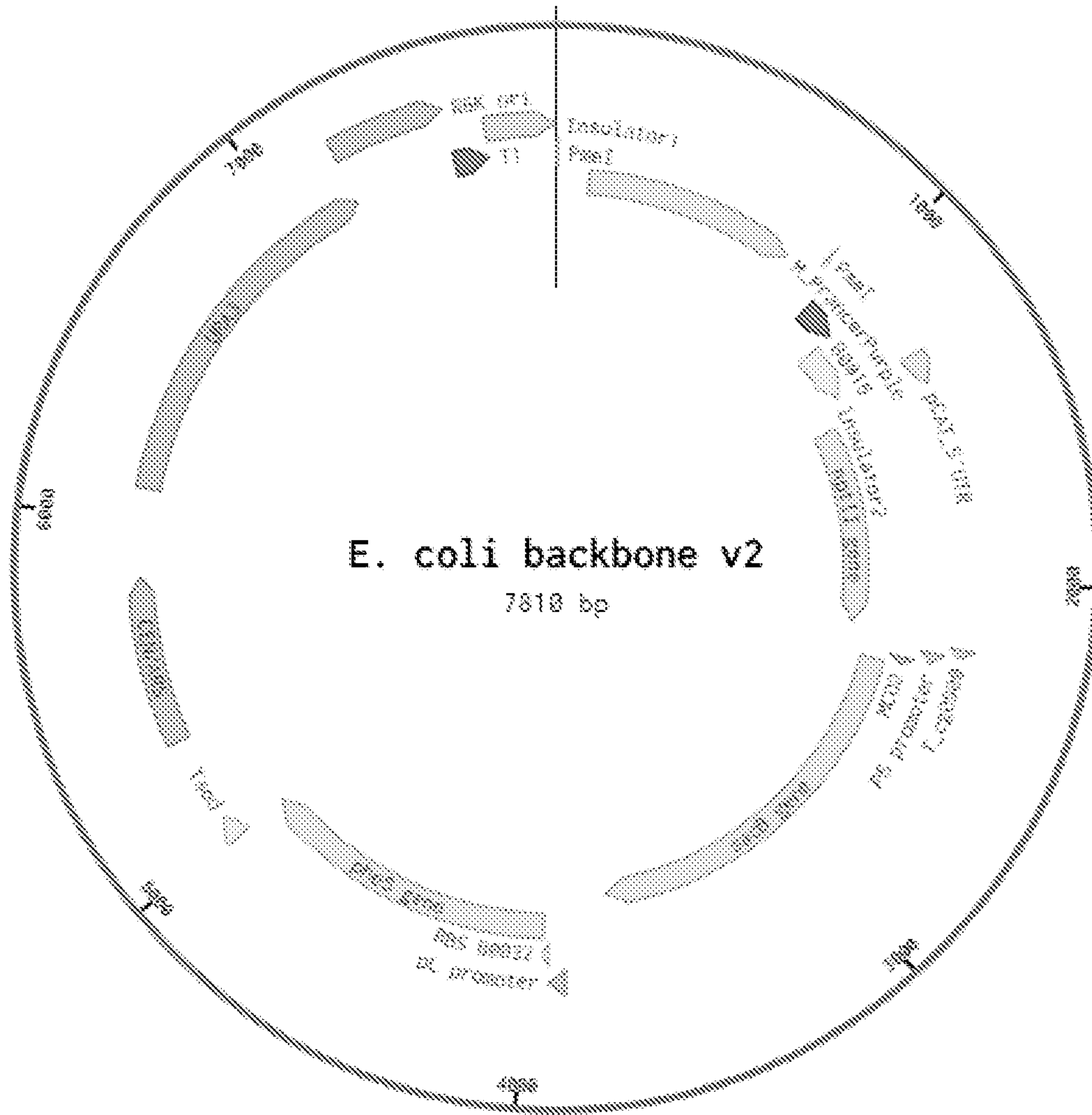


FIGURE 57

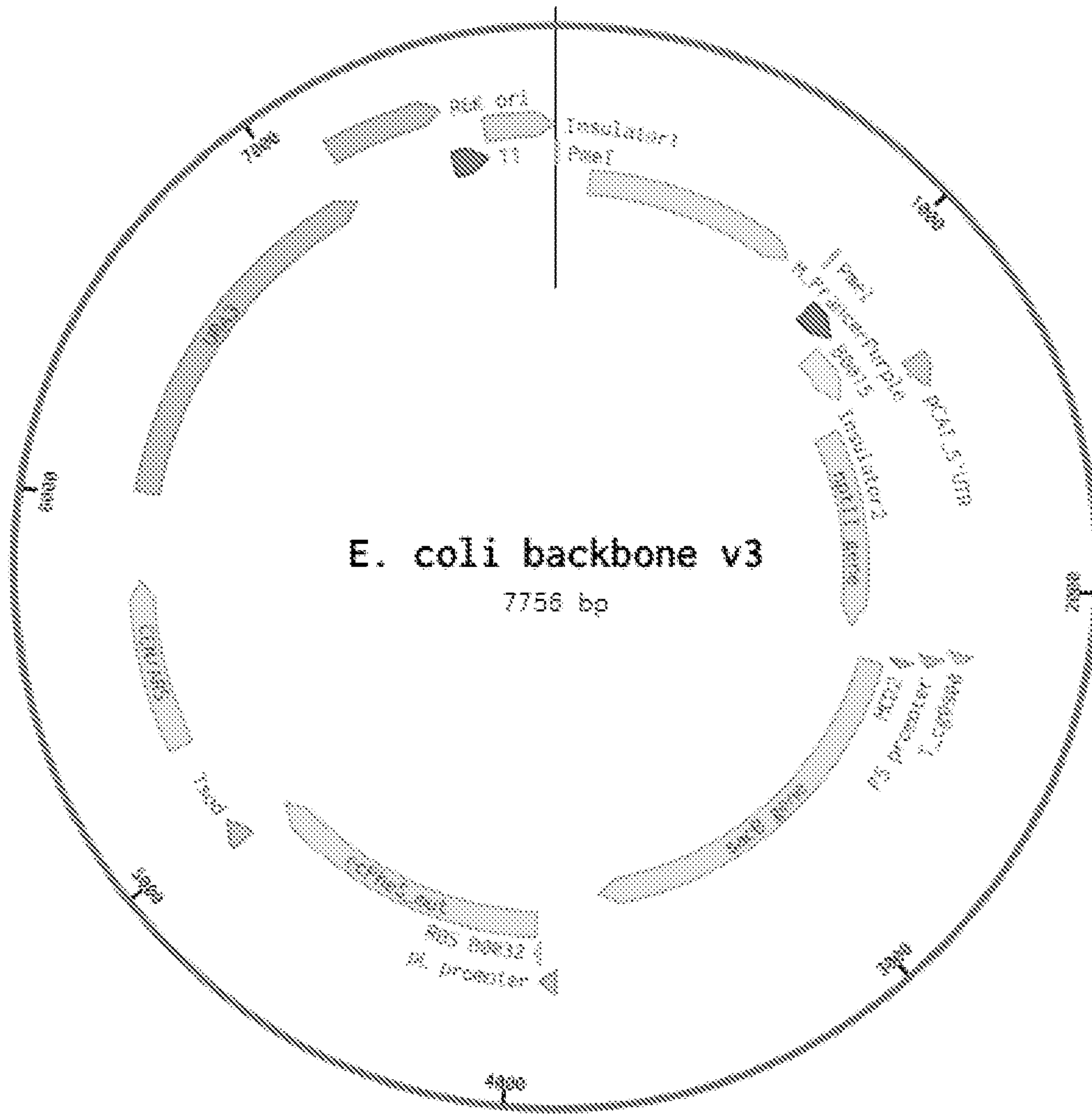


FIGURE 58

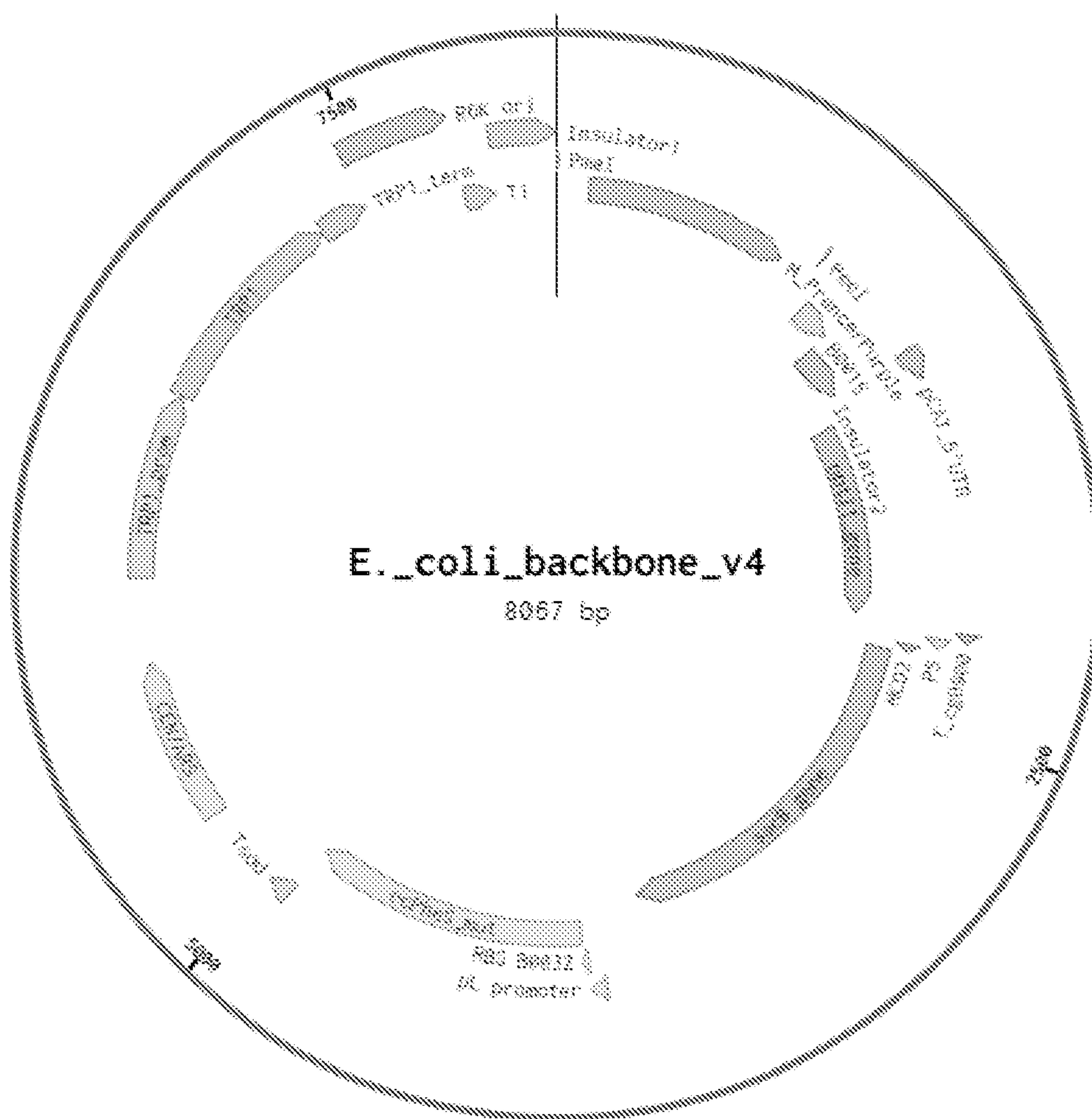


FIGURE 59

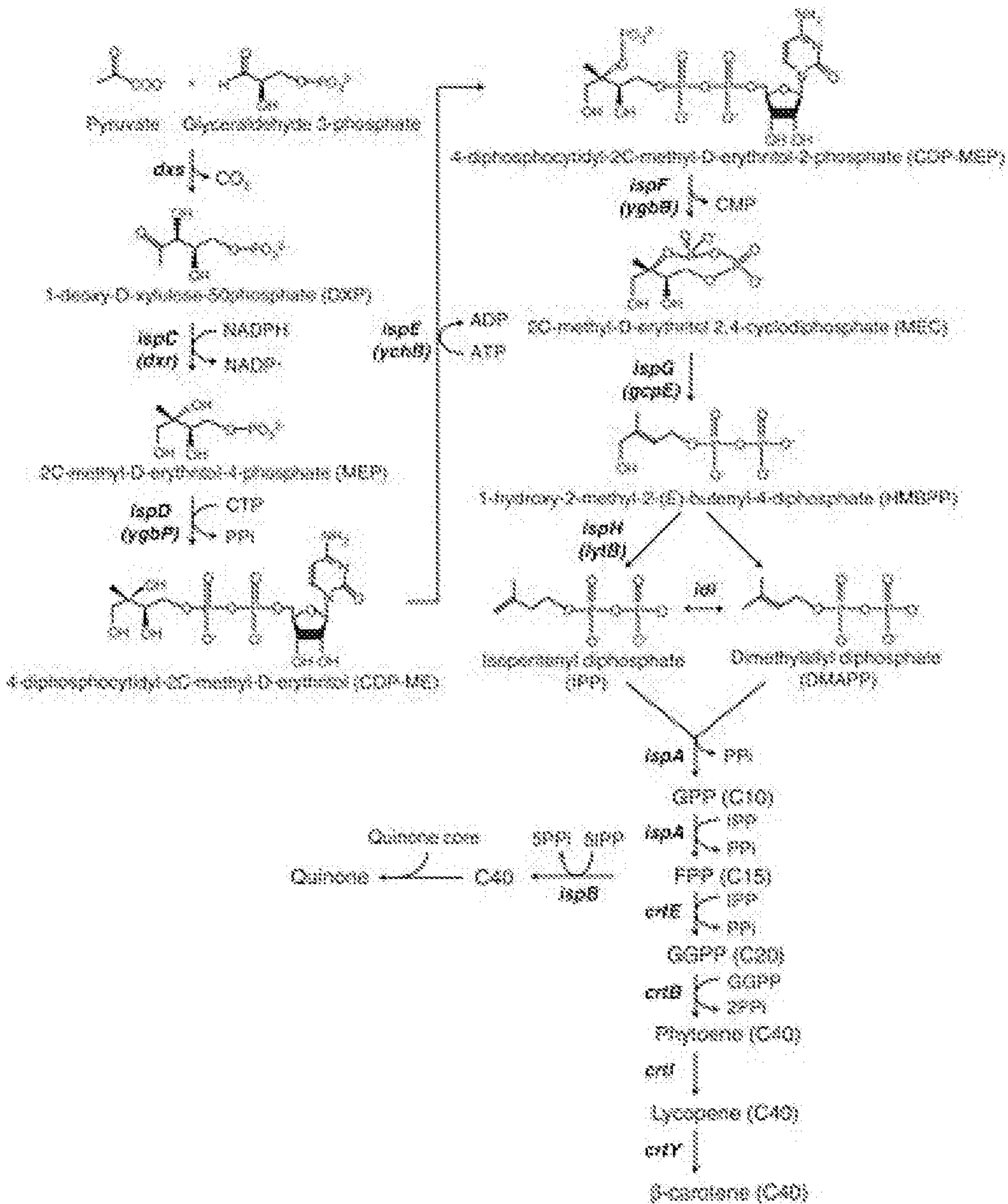


FIGURE 60

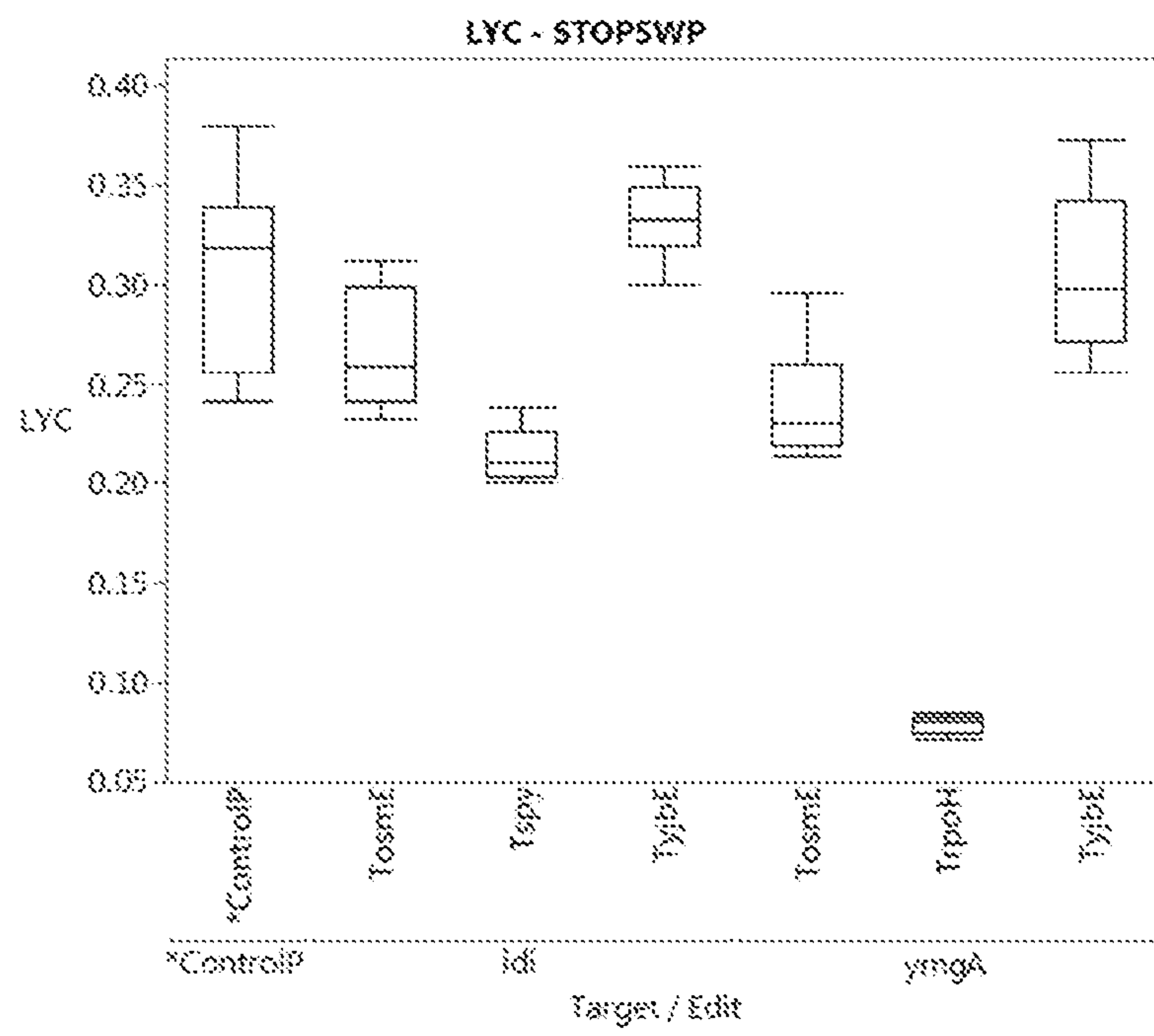


FIGURE 61

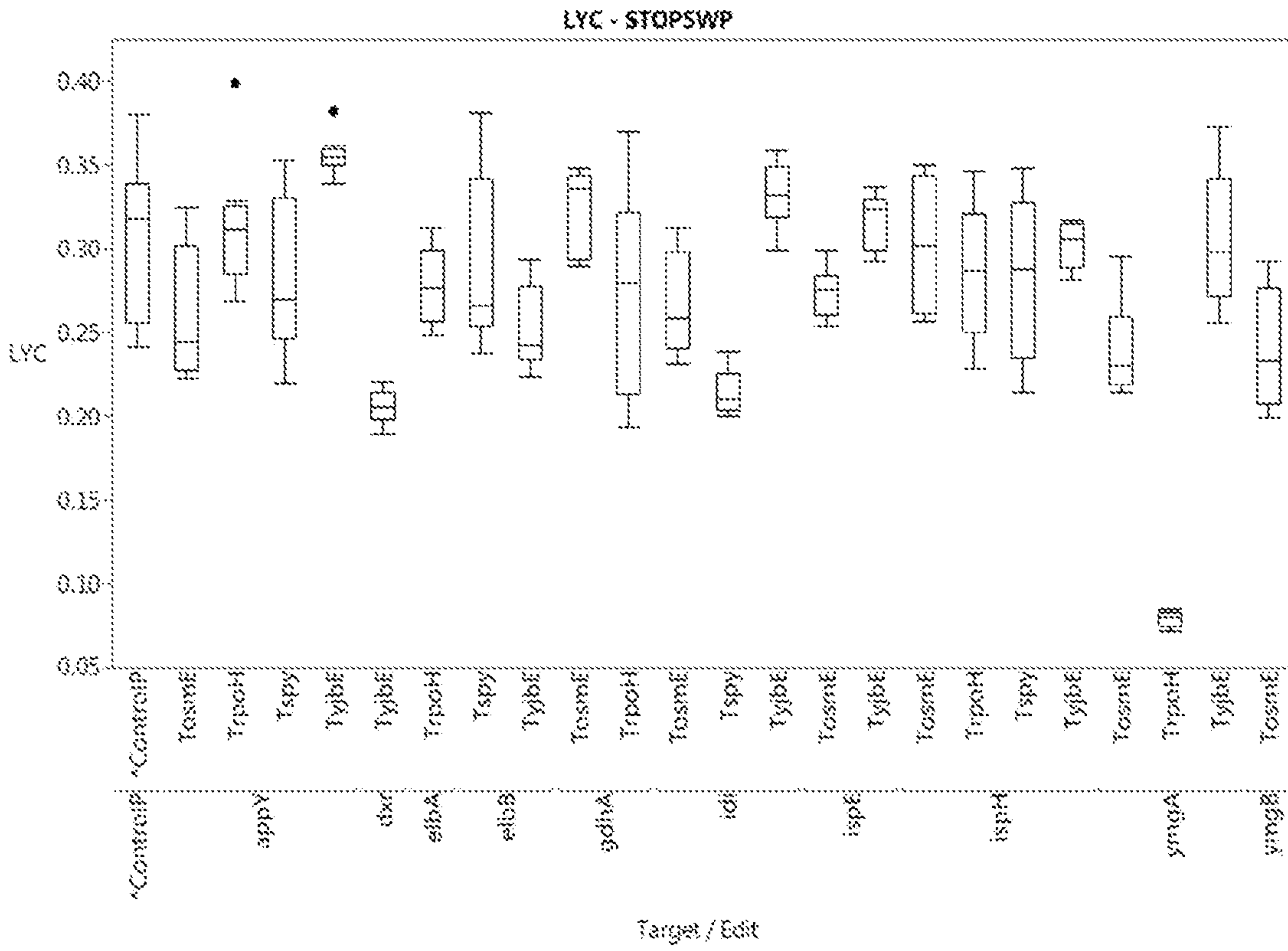


FIGURE 62

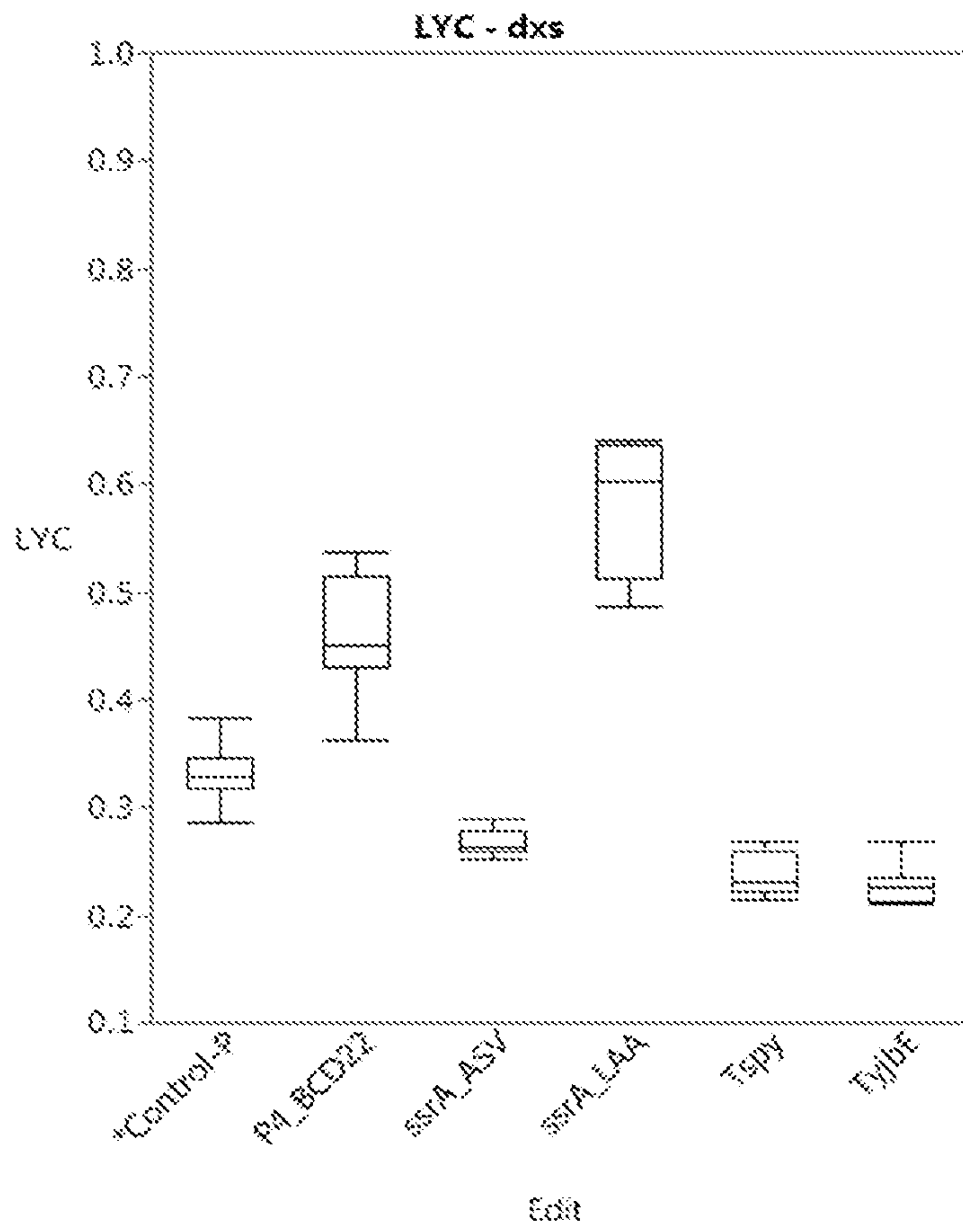


FIGURE 63

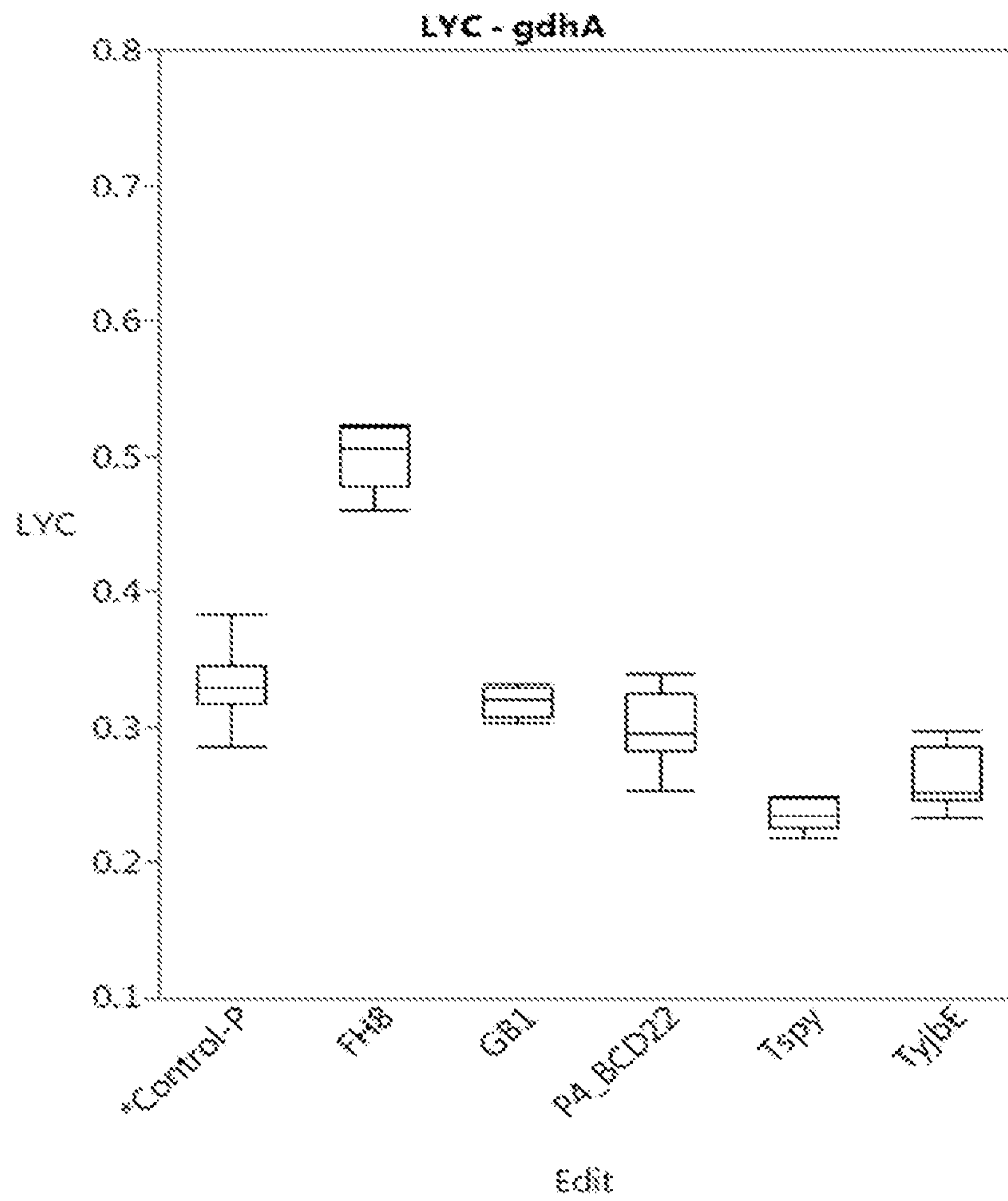


FIGURE 5

