



(51) International Patent Classification:

C12N 15/70 (2006.01)

(21) International Application Number:

PCT/US2019/056977

(22) International Filing Date:

18 October 2019 (18.10.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/747,479

18 October 2018 (18.10.2018)

US

(71) Applicant: THE REGENTS OF THE UNIVERSITY OF COLORADO, A BODY CORPORATE [US/US]; 1800

Grant Street, 8th Floor, Denver, Colorado 80203 (US).

(72) Inventors; and

(71) Applicants: GILL, Ryan T. [US/US]; 460 Oneida Street, Boulder, Colorado 80303 (US). BASSALO, Marcelo Col-ika [BR/US]; 3605 Table Mesa Drive, Apt. P-286, Boulder, Colorado 80305 (US).

(74) Agent: HANSON, Roberta J. et al.; 1401 Lawrence Street, Suite 2300, Denver, Colorado 80202 (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,

(54) Title: COMPOSITIONS AND METHODS FOR IDENTIFYING MUTATIONS OF GENES OF MULTI-GENE SYSTEMS HAVING IMPROVED FUNCTION

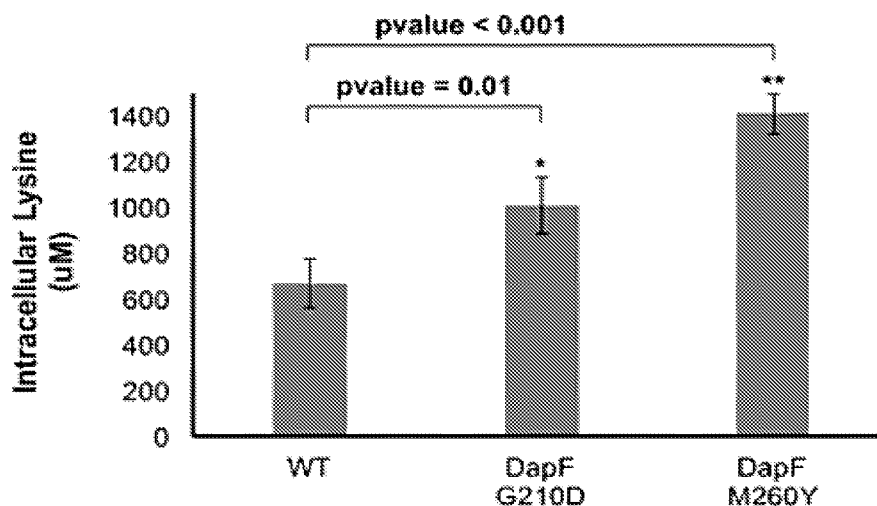


FIG. 10B

(57) Abstract: Embodiments of the present disclosure provide for generating engineered microorganisms for increased tolerance and/or production of a target molecule. In certain embodiments, a microorganism is Escherichia coli (E. coli). In some embodiments, engineered E. coli are created using mutant selection from deep scanning mutagenesis techniques for identifying mutations that result in an increase in production of and/or tolerance for an amino acid. In certain embodiments, an engineered microorganism has increased tolerance and/or production of lysine. In other embodiments, combination compositions and methods of deep scanning mutagenesis techniques and CRISPR/Cas editing systems can be used to efficiently identify mutants having improved function from a saturated pool of mutants in a multi-gene system for synthesizing amino acids.



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).

Published:

- *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*
- *with sequence listing part of description (Rule 5.2(a))*

**COMPOSITIONS AND METHODS FOR IDENTIFYING MUTATIONS OF GENES
OF MULTI-GENE SYSTEMS HAVING IMPROVED FUNCTION
PRIORITY**

[001] This PCT application claims priority to U.S. Provisional Application No. 62/747,479 filed October 18, 2018. This application is incorporated herein by reference in its entirety for all purposes.

STATEMENT REGARDING GOVERNMENT FUNDING

[002] This invention was made with government support under grant number DE-SC0008812 awarded by the U.S. Department of Energy. The government has certain rights in the invention.

STATEMENT REGARDING SEQUENCE LISTING

[003] The instant application contains a Sequence Listing which has been submitted via ASCII copy created on October 18, 2019 referred to as '101877.639065 CU4347B-PCT1_ST25.txt' and is 30 kilobytes having 21 sequences.

FIELD

[004] Embodiments of the present disclosure relate to engineering microorganisms for increasing production of and/or increasing tolerance to a target molecule having an assayable endpoint. In certain embodiments, compositions and methods disclosed herein concern genetically modifying microorganisms through manipulating pathway flux of an amino acid to increase amino acid production and/or tolerance compared to microorganisms not genetically modified. In some embodiments, genetic modifications of these microorganisms can be engineered through methods of deep scanning mutagenesis strategies applied to one or more pathways related to molecular flux of a target molecule. Some embodiments concern genetically modifying a microorganism such as bacteria or yeast. In other embodiments, modified bacteria can be of the *Enterobacteriaceae* family. In yet other embodiments, compositions and methods concern modifying *Escherichia coli* ("*E. coli*"). In certain embodiments, *E. coli* are genetically modified to positively modify amino acid flux relative to wild type. Yet other embodiments disclosed herein relate to use of engineered *E. coli* for increased production and/or tolerance of an amino acid (e.g. lysine, arginine, leucine etc.).

BACKGROUND

[005] Evolution has selected for efficient and robust metabolic and regulatory networks that prevent unnecessary metabolite biosynthesis and optimally distribute resources to maximize overall cellular fitness. Complexity of these networks in combination with limited

approaches to understand their structure and function has limited the ability to re-program cellular networks in effort to modify these systems for a range of applications. Current approaches to re-program cellular networks are directed to modifying single genes of complex pathways but as a consequence of modifying the single genes unwanted modifications to the genes or other genes can be created, limiting the ability to identify changes necessary to achieve a particular endpoint.

[006] Amino acids have many useful applications. Amino acid metabolism is fundamental to all domains of life that includes highly involved pathways with extensive kinetic and regulatory features. Amino acid metabolism is an ideal model for assessing modifications to pathways affecting amino acid flux by having a measureable endpoint, increased amino acid production and/or tolerance. Some uses for amino acids for example, the amino acid lysine, is useful for supplementing animal feedstock as a nutritional supplement, used in pharmaceuticals, and cosmetics, among others. Lysine can be industrially produced by microbial fermentation, but, there are limits to its efficiency, scalability, tolerance and production. Microbial overproducers of lysine have traditionally been identified via “adaptive evolution”, namely, adaptation of the microbes in the presence of antimetabolites (such as the analog S-(2-aminoethyl)-L-cysteine (AEC)) but the underlying genetic basis for the overproduction phenotype is relatively unknown. As an example, sequencing of a lysine-overproducing industrial strain of *Corynebacterium glutamicum* revealed more than 1000 mutations had accumulated in the genome after decades of adaptive evolution but it is unclear which of these mutations contribute to the adapted organism.

SUMMARY

[007] Embodiments of the present disclosure relate to applying, for example, deep scanning technologies in order to introduce and assay for mutations directed to altering one or more pathways related to molecular flux of a target molecule in an organism instead of targeting or selecting for single gene changes. In certain embodiments disclosed herein, microorganisms can be engineered using these deep scanning technologies for increasing production of and/or increasing tolerance to a target molecule having a measurable endpoint such as an amino acid. In some embodiments, methods disclosed herein can be used to screen tens of thousands of mutations introduced to one or more genes affecting one or more biosynthetic pathways of a target molecule to exploit mechanism(s) responsible for producing the target molecule.

[008] In other embodiments, compositions and methods disclosed herein concern genetically modifying microorganisms to increase amino acid production and/or tolerance

compared to microorganisms that are not genetically modified. In some embodiments, genetic modifications to a microorganism are engineered through applications of deep scanning mutagenesis strategies applied to one or more pathways related to molecular flux of a target amino acid. Some embodiments concern genetically modifying bacteria of the *Enterobacteriaceae* family. In yet other embodiments, compositions and methods concern modifying *Escherichia coli*. (“*E. coli*”).

[009] In certain embodiments, *E. coli* are genetically modified to positively affect amino acid flux relative to wild type (e.g. lysine) to increase tolerance and/or increase production of the amino acid by the genetically modified *E. coli*. Yet other embodiments disclosed herein relate to use of engineered *E. coli* for production of lysine.

[0010] Some embodiments of the present disclosure relate to selectively engineering bacteria for producing amino acids (e.g. lysine, arginine etc.). In certain embodiments, compositions and methods disclosed herein concern genetically modifying bacteria to increase amino acid production and/or tolerance compared to bacteria that are not genetically modified. Some embodiments concern genetically modifying bacteria of the *Enterobacteriaceae* family. In yet other embodiments, compositions and methods concern modifying *Escherichia coli*. (“*E. coli*”). In certain embodiments, *E. coli* are genetically modified to increase lysine production, increase lysine tolerance, and/or modify lysine homeostasis relative to their wild type. Yet other embodiments relate to use of these engineered organisms for over production or increased tolerance to produced lysine.

[0011] Certain embodiments relate to introducing genetic mutations in genes of pathways related to amino acid production, amino acid tolerance, amino acid metabolism, and/or amino acid homeostasis in *E. coli*. In accordance with these embodiments, one or more genes of these pathways are modified to increase tolerance of the engineered *E. coli* to lysine and/or to induce over-production of lysine by the engineered *E. coli*. In other embodiments, one or more genes of the engineered *E. coli* are modified in order to enhance lysine homeostasis. In yet other embodiments, one or more genes of the engineered *E. coli* are modified in order to enhance amino acid metabolism (e.g. lysine). In accordance with these embodiments, genetic modifications to certain genes can lead to modifications of genes contributing to all around amino acid metabolism and tolerance. For example, production and tolerance of the amino acid lysine can be altered in a microorganism. In certain embodiments, lysine production, lysine tolerance, lysine metabolism, and/or lysine homeostasis in for example, during 1) lysine biosynthesis, 2) lysine degradation, 3) lysine regulation, and/or 4) lysine transport can be altered in an engineered microorganism contemplated herein.

[0012] In some embodiments, increased lysine production, lysine tolerance, modified lysine metabolism, and/or lysine homeostasis in *E. coli* can be effected through deletions or insertions into the *E. coli* genes. In accordance with these embodiments, these modifications can include genes that encode particular proteins affecting pathways related to lysine production, lysine tolerance, lysine metabolism, and/or lysine homeostasis, for example proteins involved in lysine biosynthesis, lysine degradation, lysine regulation, and lysine transport or export.

[0013] In some embodiments, genetic modifications in the engineered *E. coli* can be mutations to a binding site of one or more polypeptides involved in lysine biosynthesis and/or tolerance. In accordance with these embodiments, binding sites can include a substrate binding site, a co-factor binding site, a DNA binding site, an allosteric factor binding site. In some embodiments, the one or more genetic and/or pathway modifications to the engineered *E. coli* lead to an assayable trait. In accordance with these embodiments, an assayable trait can be with respect to an engineered microorganism having altered lysine metabolism, a decrease in uptake of S-(2-aminoethyl)-L-cysteine (AEC) by the engineered microorganism (e.g. *E. coli*) demonstrating effective lysine flux manipulation for selection purposes.

[0014] In some embodiments, production, metabolism, and/or homeostasis in *E. coli* can be enhanced by introducing mutations such as site-directed mutations or targeted mutations that affect the binding region of targeted genes. In accordance with these embodiments, some mutations can include introducing a single mutation or multiple mutations up to mutating all regions of a gene to alter a binding region for example, introducing a single nucleotide polymorphism (SNP) into the gene in one to all sites or nucleotides that affect binding affinity of the gene for a particular molecule. In certain embodiments, mutations can be introduced or selected for; for example, selecting for a SNP in one or more of genes that encode proteins affecting lysine production, metabolism, and/or homeostasis, including lysine biosynthesis, lysine degradation, lysine regulation, and lysine transport or efflux.

[0015] In some embodiments, genetic modifications for creating an engineered *E. coli* for modulating lysine metabolism can include, but are not limited to, mutating one or more *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*. In other embodiments, genetic modifications for creating an engineered *E. coli* for modulating lysine metabolism can include, but are not limited to, mutating *dapF*, *lysP*, *lysR*, *lysC*, and *lysS*, or combinations thereof. In accordance with these embodiments, the engineered *E. coli* has increased tolerance and/or production of lysine compared to a wild type.

[0016] In certain embodiments, targeted genes for modification can include, but are not limited to, one or more of *dapF*, *lysP*, and *lysR*, genes can be modified. In some embodiments,

introducing one or more SNP(s) introduced to a targeted gene of a microorganism can include, but are not limited to; one or more of *dapF* G210D, *dapF* M260Y, *lysP* T33F, *lysP* Q219I, and *lysR* S36R in order to modulate lysine biosynthesis and/or tolerance in the engineered microorganism; for example, bacteria (*e.g. E. coli*).

[0017] In certain embodiments, promoters are targeted to increase expression of one or more genes in *E. coli* in order to affect lysine production, tolerance, metabolism, and/or homeostasis. Alternatively or in addition, vectors can be designed for transfection of *E. coli* in order to increase lysine production, tolerance, metabolism, and/or homeostasis. In accordance with these embodiments, a vector can include at least a regulated promoter, an editing cassette having a selectable marker, and an associated spacer. In some embodiments, the selectable marker can include tracking a marker that indicates one or more modifications to one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC* in order to allow selection of these modified genes. It is contemplated herein that constructs of use for enhanced lysine production, metabolism, and/or homeostasis in *E. coli* can include swapping promoter regions in order to upregulate or down regulate targeted genes of a bacteria to modify lysine biosynthesis and tolerance in the bacteria.

[0018] In certain embodiments, methods are disclosed for targeting bacterial (*e.g. E. coli*) pathways associated with one or more amino acid (*e.g. lysine*) production and/or tolerance using genetic manipulation in order to obtain engineered bacteria. In certain embodiments, methods for creating engineered bacteria (*e.g. E. coli*) can include, but are not limited to, using a Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based approach. This type of approach provides for reprogramming of gene transcription translation and other effects to elicit particular targeted cellular phenotypes in the bacteria. In some embodiments, these methods can include subsequently producing an engineered bacteria (*e.g. E. coli*) by introducing into the bacteria (*e.g. E. coli*) a vector that encodes one or more mutated genes identified by deep scanning mutagenesis of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*; producing bacteria expressing the vector. In some embodiments, the bacteria can be an engineered *E. coli* having increased lysine tolerance. In other embodiments, the bacteria can be engineered *E. coli* having increased lysine production. In other embodiments, the bacteria can be an engineered *E. coli* having both increased lysine production and increased lysine tolerance.

[0019] In some embodiments, methods of making engineered bacteria or other organisms can concern manipulation of genes involved in the aspartate pathway in a microorganism to make one or more amino acid as a product from the oxaloacetate/aspartate family. In

accordance with these embodiments, amino acids contemplated in this family can include, but are not limited to, lysine, asparagine, methionine, threonine, and/or isoleucine. It is understood by those of skill in the art that aspartate can be converted into lysine, asparagine, methionine and threonine. Threonine can be converted to isoleucine. Typically, enzymes associated with generating these amino acids are subject to feedback inhibition and/or repression. Additional regulation can be found at each branch point of the pathway. This type of regulatory scheme allows control over the total flux of the aspartate pathway in addition to the total flux of individual amino acids. Aspartate pathway uses L-aspartic acid as the precursor for the biosynthesis of one fourth of the building block amino acids.

[0020] In certain embodiments, engineered microorganisms contemplated herein concern microorganisms capable of having increased production and or tolerance to one or more of lysine, arginine, proline, glutamic acid, leucine, methionine, phenylalanine, tryptophan, tyrosine, valine, isoleucine, and/or histidine. In accordance with these embodiments, the following agents can be used for selection and/or detection of a corresponding amino acid contemplated herein S-(2-Aminoethyl)-L-cysteine, canavaninin, Azetidine-2-carboxylic acid, Beta-N-Methylaminoalanine (BMAA), 5-hydroxyleucine, ethionine, selenomethionine,, o-tyrosine, 7-azatryptophan, 3,4-Dihydroxyphenylalanine (DOPA), 4-hydroxyvaline, O-methylthreonine and/or 2-thiazolealanine or other chemical of use to assay for the production of one or more amino acids contemplated herein.

[0021] In some embodiments, methods for making engineered bacteria (*e.g.* *E. coli*) can include introducing into the bacteria (*e.g.* *E. coli*) a first vector having a polynucleotide encoding a nuclease-deactivated CRISPR-associated (Cas) protein; and a second vector of one of at least one short guide RNA (sgRNA) molecule of a CRISPR-associated (Cas) protein binding site and further including a targeting RNA sequence directed to a target polynucleotide. In certain embodiments, the targeting RNA sequence is directed to a target polynucleotide including, but not limited to, one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC* or other gene related to an amino acid synthesis pathway. In other embodiments, methods provide for engineered bacteria expressing the second vector, having an increased lysine tolerance and/or increased lysine production.

[0022] In some embodiments, amino acids (*e.g.* lysine) can be produced from an engineered *E. coli* by culturing the engineered *E. coli* under conditions sufficient to produce the amino acid. In certain embodiments, methods disclosed herein include recovering the amino acid from media of engineered bacteria such as an *E. coli*. In some embodiments, methods can

include harvesting the engineered bacteria such as an *E. coli* and recovering intracellularly produced amino acid.

In some embodiments, engineered bacteria (*e.g.* *E. coli*) disclosed herein can be used for technological or commercial applications. In certain embodiments, engineered bacteria (*e.g.* *E. coli*) disclosed herein can be used for increasing production of and tolerance for an amino acid (*e.g.* lysine) by the engineered bacteria compared to a wild-type bacteria (*e.g.* *E. coli*). In certain embodiments, a 5%, or a 10%, or a 20% or, a 30%, or a 40%, or a 50%, or a 60%, or a 70 %, or an 80% or a 90% or more increase in production and/or tolerance of the amino acid (*e.g.* lysine) can be produced in the engineered bacteria.

Brief Description of the Drawings

[0023] The following drawings form part of the present specification and are included to further demonstrate certain embodiments of the present disclosure. Certain embodiments can be better understood by reference to one or more of these drawings in combination with the detailed description of specific embodiments presented herein.

[0024] **FIG. 1** illustrates an overview of lysine metabolism in exemplary bacteria (*e.g.* *E. coli*) of some embodiments disclosed herein.

[0025] **FIG. 2A** illustrates library coverage assessed through exemplary deep sequencing before exposure to Cas9 of some embodiments disclosed herein.

[0026] **FIG. 2B** illustrates library coverage assessed through exemplary deep sequencing, after exposure to Cas9 of some embodiments disclosed herein.

[0027] **FIG. 3A** illustrates an exemplary enrichment map of variants across *E. coli* targeted genes related to lysine production, metabolism, and/or homeostasis, including 1) lysine biosynthesis, 2) lysine degradation, 3) lysine regulation, and 4) lysine transport of some embodiments disclosed herein.

[0028] **FIG. 3B** illustrates an exemplary map of the number of enriched mutations in genes classified in each of the four exemplary categories mentioned above with increasing concentration of a selection agent of some embodiments disclosed herein.

[0029] **FIG. 3C** illustrates exemplary enrichment scores for each gene represented in **FIG. 4B** of some embodiments disclosed herein.

[0030] **FIG. 4** illustrates the fraction of engineered *E. coli lysP* mutants across increasing selective pressures compared to other mutants of some embodiments disclosed herein.

[0031] **FIG. 5A** illustrates growth of an exemplary engineered *E. coli lysP* T33F mutant compared to wild-type *E. coli* transformed with a non-target gRNA of some embodiments disclosed herein.

[0032] **FIG. 5B** illustrates growth of an exemplary engineered *E. coli lysP* Q219I mutant compared to wild-type *E. coli* cells transformed with a non-target gRNA of some embodiments disclosed herein.

[0033] **FIG. 6** illustrates enrichment of exemplary synonymous mutations observed for *LysP*, *LysR* and *DapF* in engineered *E. coli* of some embodiments disclosed herein.

[0034] **FIG. 7** illustrates an exemplary illustration of mutations conferring selection tolerance in engineered *E. coli* of some embodiments disclosed herein.

[0035] **FIG. 8A** illustrates an exemplary quantification of intracellular lysine levels in wild type *E. coli* and an engineered *E. coli lysR* S36R mutant of some embodiments disclosed herein.

[0036] **FIG. 8B** illustrates differential gene expression for the *lysR* and *lysA* genes in a wild type *E. coli* compared to an engineered *E. coli* having a genetic mutation of some embodiments disclosed herein.

[0037] **FIG. 9** illustrates growth of an exemplary engineered *E. coli* mutant compared to wild type *E. coli* cells transformed with a non-target gRNA of some embodiments disclosed herein.

[0038] **FIG. 10A** illustrates growth of an exemplary engineered *E. coli* mutant compared to wild type *E. coli* transformed with a non-target gRNA of some embodiments disclosed herein.

[0039] **FIG. 10B** illustrates quantification of intracellular lysine concentration in wild type *E. coli* cells and exemplary engineered *E. coli* mutant of some embodiments disclosed herein.

[0040] **FIG. 11** illustrates an exemplary gel demonstrating expression and purification of an engineered *E. coli* variant compared to a positive control microorganism of some embodiments disclosed herein.

[0041] **FIG. 12A** illustrates a negative control for a DapF kinetic experiments, comparing before and after DapF exposure of some embodiments disclosed herein.

[0042] **FIG. 12B** illustrates a positive control for a DapF kinetic experiment, illustrating comparing before and after wild-type DapF exposure of some embodiments disclosed herein.

[0043] **FIG. 13A** illustrates a DapF assay of kinetics of wild type and engineered *E. coli* of some embodiments disclosed herein.

[0044] **FIG. 13B** illustrates differential gene expression for target genes of an engineered *E. coli* compared to wild type *E. coli* of some embodiments disclosed herein.

[0045] **FIG. 14** illustrates an exemplary vector used for a selected mutant using CREATE of some embodiments disclosed herein.

[0046] **FIG. 15** represents an exemplary table illustrating genes related to lysine synthesis in a bacteria and targeted sites in an exemplary library of some embodiments disclosed herein.

[0047] **FIG. 16** represents an exemplary table illustrating amino acids and an exemplary analog thereof for selecting or detecting the presence of its respective amino acid. genes related to lysine synthesis in a bacteria and targeted sites in an exemplary library of some embodiments disclosed herein.

[0048] **FIG. 17** represents a schematic of a workflow strategy to map trajectories of agent resistance in a microorganism using CREATE of some embodiments disclosed herein.

[0049] **FIGS. 18A-18E** represent 18A) model structure of LysR, with the HTH DNA-binding domains; 18B) an enlargement of a mutation illustrating its proximity to the DNA phosphate backbone; 18C) a substitution mutation 18D) an example of absolute quantification of intracellular amino acid levels (*e.g.* lysine) in wild-type and the reconstructed mutant; and 18E) differential gene expression quantified via QPCR for exemplary genes on wild-type and mutant backgrounds of some embodiments disclosed herein.

[0050] **FIGS. 19A-19C** is a table that represents an exemplary targeted system of amino acid synthesis of targeted sites in a library of various sizes of a target protein of some embodiments disclosed herein.

[0051] **FIGS. 20A-20E** represents 5 tables in 20A-20E of a list of parameters of various mutants of some embodiments disclosed herein.

[0052] **FIGS. 21A-21D** represent 21A) a comparison mapping technique of adaptive evolution to deep scanning mutagenesis; 21B) single nucleotide polymorphism (SNP) categories; 21C) a plot of mutants found after adaptation and 21D) mapping of enriched mutations using selective pressure of some embodiments disclosed herein.

[0053] **FIG. 22** represents exemplary growth curves of an amino acid library (*e.g.* lysine) (black) compared to two different negative controls under increasing selective pressures. DSB

(double-strand break) serves as a negative control where a cassette designed to introduce a stop codon at the unrelated gene *galK* and a non-target negative control is also used for some embodiments disclosed herein. $n=3$ for each curve. Positive results were observed related to the amino acid library under increasing selective pressures.

DETAILED DESCRIPTION

[0054] In the following sections, various exemplary compositions and methods are described in order to detail various embodiments of the disclosure. It will be obvious to one of skill in the relevant art that practicing the various embodiments does not require the employment of all or even some of the details outlined herein, but rather that concentrations, times and other details may be modified through routine experimentation. In some cases, well-known methods or components have not been included in the description.

[0055] As disclosed herein “modulation” and “manipulation” of a gene can mean an increase, a decrease, upregulation, downregulation, an induction, a change in encoded activity, a change in binding, a change in stability or the like, of one or more of targeted genes or gene clusters.

[0056] In certain embodiments of the present disclosure, there can be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. *See, e.g.*, Sambrook, Fritsch & Maniatis, *Molecular Cloning: A Laboratory Manual*, Second Edition 1989, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; *Animal Cell Culture*, R. I. Freshney, ed., 1986).

[0057] In certain embodiments of this disclosure, primers used for sequencing and sample preparation per conventional techniques can include sequencing primers and amplification primers. In some embodiments, plasmids and oligomers used per conventional techniques can include synthesized oligomers, oligomer cassettes. Appendix A, Appendix B, Appendix C, Appendix D, Appendix E, Appendix F and Appendix G and the figures and sequence listing of the provisional application are incorporated herein in their entirety for all purposes.

[0058] In microbes and multicellular organisms, amino acid metabolism consists of highly evolved pathways with extensive kinetic and regulatory features. Evolution has selected for efficient and robust metabolic and regulatory networks that prevent unnecessary metabolite biosynthesis and optimally distribute resources to maximize overall cellular fitness. The complexity of such networks, coupled with limited approaches to understand their structure

and function, has broadly limited capabilities for understanding and rewiring cellular networks across a range of applications. Network and pathway engineering strategies have relied primarily upon coarse approaches for modulating function (*e.g.* promoter swaps or complete gene knockouts) at a limited number of loci. Alternatively, adaptive laboratory evolution (ALE) approaches are often employed to produce more refined adjustments (*e.g.* SNPs) for manipulating pathway flux. ALE can lead to a larger number of unintended passenger mutations and limited mechanistic understanding of the improved phenotype. Moreover, both strategies massively under sample the combinatorial space of interest. As such, network and pathway engineering would benefit from improved approaches capable of generating a broad range of targeted mutations that can be mapped with high resolution to the pathway-network level function, mirroring deep scanning mutagenesis strategies that have revolutionized protein engineering. This capability would provide for entirely new paradigms to engineer complex multigenic phenotypes to optimize function through transcription, translation, stability, and kinetics among others that encompass the breadth of what is found in nature. In certain embodiments disclosed herein, sequence to function mapping at a pathway scale has been designed and used.

[0059] In certain embodiments, compositions and methods disclosed herein relate to amino acid metabolism and manipulations thereof. Amino acids include large industrial product markets - lysine, for example, is used in the animal feedstock, pharmaceutical and cosmetics industries, having a multi-billion dollar market. Lysine overproducers were traditionally identified via adaptation in the presence of antimetabolites such as the analog S-(2-aminoethyl)-L-cysteine (AEC). Derepression of lysine biosynthesis has been previously implicated as a mechanism of resistance to AEC, however the complexity of this phenotype has also implicated other mechanisms such as improper discrimination by the lysyl-tRNA synthetase machinery. Although recent systems-based approaches are being used to elucidate the biochemical and regulatory mechanisms of lysine overproduction, current strategies rely on individually constructing and testing single sequence-to-activity hypotheses, requiring substantial investment in time and resources.

[0060] In certain embodiments, one tool as used in certain methods disclosed herein to overcome limited abilities to predict the phenotypic consequences of mutations in single proteins is to introduce every possible mutation and couple that to a genotype-phenotype assay platform; for example, deep scanning mutagenesis. As an example, tens of thousands of single and multiple mutations can be investigated in the coding sequence of a target protein to report a local fitness landscape for this protein, using for example, fluorescence as a proxy. Expanding

this concept to a repertoire of proteins connected to one another through a phenotype of interest permits parallel investigation of pathways and networks on a system scale. This requires, however, the ability to individually measure genotype-phenotype relationships for each of the designed mutants across all targeted proteins. It has previously been reported that a method (CRISPR EnAbleD Trackable genome Engineering or CREATE) facilitates parallel mapping of mutations in a massively multiplex scale. CREATE leverages array-based oligo technologies to synthesize and clone hundreds of thousands of cassettes containing a genome-targeting gRNA covalently linked to a dsDNA repair cassette encoding a designed mutation. Until the instant disclosure, these methods have not been applied to amino acid synthesis pathways to optimize production and tolerance to amino acids. In certain methods disclosed herein, after use of CRISPR/Cas (*e.g.* CRISPR/Cas9) genome editing, frequency of each designed mutant can be tracked by high-throughput sequencing using the CREATE plasmid as a barcode, uniquely combining these two technologies. In accordance with these embodiments, using these combination methods, proteins associated with a metabolic pathway can be interrogated in parallel at single nucleotide resolution, validating deep scanning mutagenesis at a pathway-focused scale.

[0061] In certain embodiments, amino acid metabolism pathways were targeted in order to optimize production and tolerance of a target amino acid through analysis of its production pathway. In accordance with these embodiments, the amino acid, lysine, was analyzed through identification of critical modifications to lysine stasis in a microorganism. In certain embodiments, lysine metabolism as an amino acid example in bacteria was used as an exemplary example. In some embodiments, the bacterium used for analysis and validation was *Escherichia coli*. For this example, a saturated mutagenesis library was constructed in binding pockets of key proteins involved in four main categories that affect lysine homeostasis in bacteria: 1) biosynthesis, 2) degradation, 3) regulation and 4) transport (See for example, FIG. 1). In order to assay for amino acid synthesis and other parameters, the library of constructs was challenged by an antimetabolite, S-(2-aminoethyl)-L-cysteine (AEC). Using this approach, in certain embodiments, contribution of the targeted mutations (*e.g.* 16,300 mutations assessed simultaneously) could be evaluated in parallel toward antimetabolite resistance and further for overall pathway flux. Using these compositions and methods, mutations beyond dominant selection winners could be identified as well as identifying mechanisms for altering pathway flux that would have been difficult to predict previously given the number of mutations being evaluated. Assessing genotype-phenotype mapping at a pathway scale, additional factors contributing to amino acid metabolism were

also identified. Combining these compositions and methods for assessing a pathway provides valuable framework for directed engineering of complex multigenic phenotypes for use in commercial purposes. In certain embodiments, directed engineering of targeted amino acids disclosed herein can be used to generate engineered bacteria for the production of target agents (*e.g.* amino acids).

[0062] Some embodiments of the present disclosure relate to selectively engineering bacteria for producing amino acids (*e.g.* lysine, arginine, methionine, etc.). In certain embodiments, compositions and methods disclosed herein concern genetically modifying bacteria to increase amino acid (*e.g.* lysine) production and/or amino acid (*e.g.* lysine) tolerance compared to bacteria that are not genetically modified. Some embodiments concern genetically modifying bacteria of the *Enterobacteriaceae* family. In yet other embodiments, compositions and methods concern modifying *Escherichia coli*. (“*E. coli*”).

[0063] In certain embodiments, *E. coli* are genetically modified to increase lysine production, increase lysine tolerance, and/or modify lysine homeostasis relative to their wild type. Yet other embodiments, relate to use of these engineered organisms for over production or increased tolerance to produced lysine.

[0064] Certain embodiments relate to introducing genetic mutations in one or more genes of pathways related to lysine production, lysine tolerance, lysine metabolism, and/or lysine homeostasis in *E. coli*. In accordance with these embodiments, one or more genes of these pathways are modified to increase tolerance of the engineered *E. coli* to lysine and/or to increase production of lysine by the engineered *E. coli*. In other embodiments, one or more genes of the engineered *E. coli* are modified in order to enhance lysine homeostasis.

ENGINEERED *E. COLI*

[0065] Microorganisms, such as *E. coli*, produce lysine through highly evolved pathways with extensive kinetic and regulatory features. Certain pathways involved in lysine production include: 1) lysine biosynthesis, 2) lysine degradation, 3) lysine regulation, and 4) lysine transport. These pathways include multiple categories of genes and gene regions that can affect lysine production and tolerance in the *E. coli*. Selective mutations or manipulations to one or more genes within these pathways can modulate lysine production, metabolism, tolerance, and/or homeostasis in the bacteria.

[0066] In some embodiments, engineered *E. coli* contain one or more mutations in a single gene or one or more mutations in multiple genes, which form part of one or more pathways for lysine production, metabolism, tolerance, and/or homeostasis. These one or more genes

targeted genes code for proteins with a variety of cellular functions including, but not limited to transcription, repression, and/or regulation of lysine biosynthesis.

[0067] In some embodiments, engineered *E. coli* disclosed herein can contain mutations to a single gene identified using deep scanning methodologies disclosed herein. In certain embodiment, these methodologies can identify a single mutation having significant effect on lysine biosynthesis beyond dominant selection winners as well as identifying mechanisms for altering lysine pathway flux that would have been difficult to predict by known methods given the number of mutations being evaluated. In some embodiments, engineered *E. coli* can contain one or more mutations to multiple genes related to lysine biosynthesis and tolerance. In some embodiments, one or more mutations can include genes unrelated to a lysine regulatory, transport, or biosynthesis pathway identified by deep scanning methodologies. In some embodiments, genetic modifications in the engineered *E. coli* can include mutations to a binding site of one or more polypeptides involved in lysine biosynthesis and/or tolerance. In accordance with these embodiments, binding sites can include a substrate binding site, a co-factor binding site, a DNA binding site, and/or an allosteric factor binding site. In some embodiments, the one or more modifications to the engineered *E. coli* lead to a decrease in uptake of S-(2-aminoethyl)-L-cysteine (AEC) by the engineered *E. coli*.

[0068] In some embodiments, amino acid (*e.g.* lysine) production, metabolism, and/or homeostasis in *E. coli* can be enhanced by introducing mutations such as site-directed mutations or targeted mutations that affect the binding region of the genes identified by methods disclosed herein. In accordance with these embodiments, some mutations can include introducing a single mutation or change in a gene to alter a binding region for example, introducing a single nucleotide polymorphism (SNP) into the gene that affects binding affinity of the gene for a particular molecule. In certain embodiments, mutations can be introduced or selected for; for example, selecting for a one or more SNP in one or more of genes that encode proteins affecting lysine production, metabolism, and/or homeostasis, including lysine biosynthesis, lysine degradation, lysine regulation, and lysine transport or efflux.

[0069] In certain embodiments, targeted genes for lysine production or tolerance in a microorganism can include, but are not limited to, one or more mutations of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC* genes or synonymous gene thereof. In some embodiments, introducing one or more SNP(s) to a target gene can include introducing one or more SNPs of *dapF* G210D, *dapF* M260Y, *lysP* T33F, and *lysP* Q219I in order to modulate lysine biosynthesis and/or tolerance in the engineered *E. coli*.

[0070] In some embodiments, modifications to *E. coli* can be to the gene encoding the protein LysP. In certain embodiments, synonymous mutations to the gene encoding *lysP* can alter expression or stability of LysP. For example in *E. coli*, lysine uptake is mediated by three different transporter systems: ArgT, CadB, and LysP. ArgT codes for a periplasmic binding protein specific to lysine, arginine and ornithine, interacting with the ABC transporter coded by the *hisJQMP* operon. CadB is part of the Cad system, which plays a role in pH homeostasis under acidic conditions. This transporter imports lysine and excretes the decarboxylated product cadaverine in conditions of low external pH and exogenous lysine. LysP is a specific transporter for lysine, but also has a regulatory role in activating the Cad system through transmembrane interactions with CadC.

[0071] In some embodiments, modifications to *E. coli* can be performed by introducing mutations to the gene encoding *dapF*. For example, in *E. coli*, the DapF gene encodes an epimerase catalyzing the penultimate step in the lysine biosynthetic pathway, a conversion of LL-diaminopimelate (LL-DAP) to *meso*-diaminopimelate (*meso*-DAP). In some embodiments, modifications to *E. coli* can be to one or more genes encoding *lysR*, *lysC*, *serC* and *dapD*.

[0072] In some embodiments, modifications to *E. coli* can be to the gene encoding *lysR*. In accordance with these embodiments, a mutation in *lysR* can include modifying amino acid position 36 or synonymous position in *lysR* to be an Arginine instead of a Serine (e.g. *lysR*_S36R) (See for example, **SEQ ID NO: 1**). In some embodiments, a mutation to the gene encoding *lysP*, can include modifying amino acid position 33 in *lysP* to be a Threonine instead of a Phenylalanine (e.g. *lysP*_T33F) (See for example, **SEQ ID NO: 2**). In some embodiments, a mutation to the gene encoding *lysP*, can include modifying amino acid at position 33 in *lysP* to be an Isoleucine instead of a Glutamine (e.g. *lysP*_Q219I) (See for example, **SEQ ID NO: 3**). In some embodiments, a mutation to the gene encoding *dapF*, can include modifying amino acid at position 210 in *dapF* to an Aspartic Acid instead of a Glycine (e.g. *dapF*_G210D) (See for example, **SEQ ID NO: 4**). In some embodiments, a mutation to the gene encoding *dapF*, can include modifying amino acid at position 260 in *dapF* is a Methionine instead of a Tyrosine (e.g. *dapF*_M260Y) (See for example, **SEQ ID NO: 5**).

[0073] In certain embodiments, the engineered *E. coli* disclosed herein can have a decrease in the uptake of the lysine analog S-(2-aminoethyl)-L-cysteine (“AEC”) when AEC is added to culture media demonstrating efficacy of selective mutations in the *E. coli*. AEC is an antimetabolite, an analog to lysine that competes with canonical lysine for binding to the lysil-tRNA synthetase (LysRS), leading to protein misfolding and reduced cell growth.

Additionally, AEC blocks lysine biosynthesis by interacting with riboswitches, inhibiting bacterial growth in the absence of an external lysine source. As a result, mutations that lead to increase in lysine production and/or tolerance can out-compete the effects of AEC and restore cell growth. In certain embodiments, the engineered *E. coli* disclosed herein can have increased tolerance to AEC as compared to control *E. coli* when AEC is added to the bacterial culture media. In other embodiments, AEC can be used to select bacterial mutants of use for increased production or tolerance to lysine for commercial purposes.

[0074] In certain embodiments, mutations in regulatory genes of *E. coli* for example, in the lysine-regulated riboswitch controlling expression of the aspartokinase *lysC* affect uptake of AEC and confirm modification of lysine flux. Using this regulatory gene target as an example, diaminopimelic acid is bound to the regulator and the regulator *lysR* activates the last enzymatic step in lysine biosynthesis (*lysA*). The LysR family of transcription regulators is ubiquitous in bacteria and includes a conserved N-terminal helix-turn-helix (HTH) DNA-binding domain and a less conserved C-terminal co-inducer binding domain. For example, in one exemplary embodiment, a *lysR* mutant (*lysR_S36R*) substitution, as disclosed herein, lies on the DNA-binding (HTH) domain. In certain methods the HTH DNA binding domain can be manipulated to increase lysine production using compositions and methods disclosed herein. In some embodiments, a *lysR* mutant can lead to an increase in intracellular lysine production by the engineered *E. coli*.

[0075] In certain embodiments, modifications to genes or gene regions, as disclosed herein, can lead to increased tolerance of the engineered *E. coli* to the presence of higher concentrations of intracellular lysine. In certain embodiments, modifications to genes or gene pathways, as disclosed herein, can lead to increased lysine production of the engineered *E. coli*. In certain embodiments, modifications to genes or gene regions, as disclosed herein, can lead to increased lysine homeostasis of the engineered *E. coli*. In certain embodiments, modifications to genes or gene regions, as disclosed herein, can lead to increased lysine metabolism of the engineered *E. coli*. In yet other embodiments, modulation of genes or genetic regions contemplated herein in *E. coli* can allow the engineered *E. coli* to increase production of lysine in a shorter period of time compared to control *E. coli*. In accordance with these embodiments, engineered *E. coli* having a modified tolerance to the presence of high intracellular lysine levels can be used to produce larger quantities of lysine for example, for industrial applicability, reducing production costs and scaling up production, for example.

[0076] In accordance with these embodiments, modifications to increase lysine production, tolerance, metabolism, and/or homeostasis in bacteria can be the result of increased copy

number of one or more gene or genetic regions. In other embodiments, modifications to increase lysine production and/or tolerance can be the result of upregulation or down regulation; of expression of one or more genes in bacteria, or can result from a decreased copy number of one or more such gene. In some embodiments, modifications of the bacterial genes can include point mutations either selected for or introduced, resulting in one or more SNPs within a gene affecting the production of lysine and/or tolerance of lysine in the engineered bacteria.

[0077] In some embodiments, manipulated *E. coli* phenotypes contemplated herein can be used for technological or commercial applications. In accordance with these embodiments, one or more genes can be manipulated in *E. coli* in order for an engineered *E. coli* to tolerate high lysine concentrations for example, lysine levels typically toxic to a wild-type *E. coli*. Certain embodiments disclosed herein concern introducing one or more mutations to one or more genes or gene regions disclosed herein in *E. coli* in order to modulate lysine production, metabolism, and/or homeostasis in the *E. coli*.

[0078] In some embodiments, engineered bacteria (e.g. *E. coli*) disclosed herein can have increased productivity of a target amino acid such as lysine, arginine, methionine or other target amino acid compared to an unmodified control *E. coli*. In certain embodiments, a 5%, or a 10%, or a 20%, or a 30%, or a 40%, or a 50%, or a 60%, or a 70 %, or an 80% or a 90% or more increase in production and/or tolerance of a target amino acid such as lysine, arginine, methionine or other target amino acid can be obtained. In other embodiments, a commensurate level of improved amino acid homeostasis is observed in the engineered *E. coli* compared to unmodified control *E. coli*.

METHODS FOR GENERATING ENGINEERED ORGANISMS

[0079] Adaptive mutations can be a central driver of evolution; however, the abundance and relative contribution of these mutations to resulting cellular phenotypes are poorly understood, even in well-studied organisms. Network and pathway engineering strategies have relied primarily upon coarse approaches for modulating function (e.g. promoter swaps or complete gene knockouts) at a limited number of loci. Alternatively, adaptive laboratory evolution or “directed evolution” approaches have been employed, which while producing more refined adjustments for manipulating pathway flux can lead to a larger number of unintended passenger mutations and limited mechanistic understanding of the improved phenotype. A fundamental limitation to directed evolution or targeted selection of a particular phenotype, is the inability to effectively manipulate complex phenotypes in a laboratory setting, where the relevant

combinatorial mutational space is often much larger than can be searched on laboratory time scales or budgets. Further, off-target mutations can decrease overall fitness of an organism and lead to “dead-end” phenotypes, preventing further improvement of an evolved strain.

[0080] Traditionally in directed evolution studies, a vast majority of introduced mutations within a protein negatively affect protein function and stability (e.g. 30-50% are strongly deleterious, 50-70% are slightly deleterious or neutral), with only a handful (0.01-1%) improving function. This highlights the importance of the use of tagging (aka barcoding) or other method of tracking to identify a plurality of mechanisms for altering a phenotype of interest (e.g. increased pathway flux vs decreased inhibitor flux), allowing exploration beyond a local optima in the fitness landscape.

[0081] In certain embodiments, methods disclosed herein are able to overcome limited abilities and errors of directed evolution in predicting the phenotypic consequences of mutations in single proteins. In some embodiments disclosed herein, methods include introduction of every possible mutation to a target gene or genetic pathway and combining these mutations to a genotype-phenotype assay platform. In certain embodiments, this permits expansion of deep scanning concepts to a repertoire of proteins connected to one another through a phenotype of interest, allowing parallel investigation of pathways and/or networks on a system scale when partnered with individual measurements of genotype-phenotype relationships for each mutant across all targeted proteins. In some embodiments, probable protein targets with no known functional sites can be scanned and mutations of these regions can divulge roles previously unidentified due to lack of ability to pinpoint activity/function relationships.

[0082] One aspect of the present disclosure provides methods for mapping multiple loss-of-function or gain-of-function mutations in one or more transporter protein(s) related to lysine biosynthesis, identifying these one or more transporter protein(s) as an important resistance route to modifying amino acid flux.

[0083] In certain embodiment, exemplary lysine transporter proteins of the one or more transporters include but are not limited to *lysP*, *argT*, and *cadB*. In some embodiments, certain synonymous mutations in the one or more transporter protein(s) can disrupt transporter function, for example by, impacting transcription or translation rate or affecting proper folding or function of the one or more transporter(s) in the membrane. In certain embodiments, synonymous mutations in the one or more transporter protein(s) can enhance transporter function, such that the transporter protein(s) functions are enhanced.

[0084] Another aspect of the present disclosure provides methods for mapping multiple loss-of-function or gain-of-function mutations in one or more regulatory protein(s), identifying these one or more regulatory proteins as contributory molecules to lysine regulation and/or tolerance. In certain embodiments, exemplary regulatory proteins can include, but are not limited to, *argP* and *lysR*. In certain embodiments, one or more mutations in the one or more regulatory protein(s) can disrupt regulatory function. In certain embodiments, one or more mutations in one or more regulatory protein(s) can enhance regulatory function, for example, when regulatory protein(s) improve efficiency in lysine production, tolerance metabolism, and/or homeostasis.

[0085] In another aspect of the present disclosure, methods are provided for mapping multiple loss-of-function and/or gain-of-function mutations in one or more genes encoding one or more amino acid (e.g. lysine) degradation-related protein(s), identifying important contributory proteins. In certain embodiments, proteins contributing to lysine degradation can include, but are not limited to, *ldcC*, *cadA*, *cadB*, *cadC* and combinations thereof. In some embodiments, several synonymous mutations in one or more degradation-related protein(s) can also disrupt amino acid degradation function. In other embodiments, several synonymous mutations in one or more degradation-related protein(s) can enhance transport function.

[0086] In yet another aspect of the present disclosure, methods are provided for mapping multiple loss-of-function and/or gain-of-function mutations in one or more genes encoding amino acid (e.g. lysine) biosynthesis protein(s), identifying these genes for encoding these proteins as important target for this biosynthesis.

[0087] In certain embodiments, exemplary proteins contributing to the amino acid lysine biosynthesis can include, but are not limited to, *dapA*, *dapB*, *dapE*, *lysU*, *lysS*, *asd*, *dapF*, *argD*, *lysA*, *lysR*, *lysC*, *serC*, and *dapD* or synonymous genes thereof. In certain embodiments, mutations in one or more biosynthesis protein(s) can disrupt regulatory function, impacting transcription or translation rate or binding of related signaling proteins responsible for regulation.

[0088] In some embodiments, methods using mapping techniques can permit assess to identifying phenotypic consequences of one or more mutations under conditions of stringent selective pressures. In other embodiments, a library redesign, including, for example different sets of mutations targeting each site, can allow high resolution mapping of effects of silent substitutions on wild-type protein function.

[0089] One aspect of the present disclosure provides for methods for deeply mapping, through the use of barcodes, to quantify beyond a typical selection winner. In some

embodiments, this mapping can be used to identify mutations of one or more genes for directing pathway optimization. In certain embodiments, parallel integration of genes can be leveraged to uncover regulatory interactions on a systems scale. In particular exemplary embodiments, mutations identified by these deep scanning methods resulting in amino acid (*e.g.* lysine) overproduction in a microorganism can be used to create industrial strains having these traits in order to increase production to manufacturing scale.

[0090] In some embodiments of this disclosure, deep scanning mutagenesis strategies can be used to profile genetic mutations in a microorganism from a single gene to entire metabolic pathway(s). In some embodiments of this disclosure, deep scanning mutagenesis can be used to elucidate multiple routes having one or more mutated genes that affect amino acid (*e.g.* lysine) production, metabolism, and/or homeostasis leading to identification of contribution of individual genes within particular pathways to modify amino acid production. This process permits development of engineered microorganisms (*e.g. E. coli*) capable of increased amino acid (*e.g.* lysine) production and/or homeostasis.

[0091] In some embodiments, one approach for engineering complex phenotypes can be to use genome engineering tools. In accordance with these embodiments, methods can include the use of trackable, precision genome editing referred to as Clustered regularly interspersed short palindromic repeats (CRISPR). CRISPR systems exist in many bacterial genomes and have been found to play an important role in adaptive bacteria immunity. Genome engineering as detailed herein can use CREATE, (Crispr enabled trackable genome engineering), a CRISPR-based technology that involves synthesizing constructs which contain an editing cassette and CRISPR-RNA sequentially. CREATE methods achieve highly efficient editing/mutating using a single vector that encodes both an editing cassette and a guide RNA (gRNA). CREATE allows parallel mapping of mutations in a multiplex scale. CREATE leverages array-based oligo technologies to synthesize and clone hundreds of thousands of cassettes containing a genome-targeting gRNA covalently linked to a dsDNA repair cassette encoding a designed mutation. In certain methods, a CREATE editing cassette can introduce a silent protospacer adjacent motif (PAM). In accordance with these embodiments, PAM mutation(s) can be an insertion, deletion or substitution of one or more nucleotides that mutates the sequence of the PAM such that the mutated PAM (PAM mutation) is not recognized by the CRISPR system. A cell including a PAM mutation can be said to be "immune" to CRISPR-mediated killing. After CRISPR/Cas9 or CRISPR/Cas12a or other similar genome editing, frequency of each designed mutant can be tracked by high-throughput sequencing using the CREATE plasmid as a "barcode."

[0092] In certain embodiments, methods for mapping genotype-phenotype relationships on a multiplex protein scale can include identifying a target multigenic system; applying full codon mutagenesis to genes of the target multigenic system to create a mutant library; transfecting the mutant library into a host; applying CrispR/Cas genomic engineering to the genes of the multigene system to create a mutant library; and using deep-scanning mutagenesis to analyze the mutant library. In accordance with these embodiments, these technologies can be used in combination to predict genotype-phenotype relationships of mutants from the mutant pool (*e.g.* tens of thousands of mutants of two or more genes of a system). In certain methods, the deep-scanning mutagenesis method is CRISPR-EnAbleD Trackable genome Engineering (CREATE). In other embodiments, the target multigenic system comprises a pathway. In certain embodiments, pathway can include a synthesis pathway or a regulatory pathway or a transport pathway or a storage pathway or similar. In accordance with these embodiments, a synthesis pathway can include an amino acid synthesis pathway. In other embodiments, a pathway contemplated herein can be a pathway that produces mapable and/or assayable product. In other methods, the pathway can produce an assayable end-product. In some embodiments, combination methods disclosed herein can identify one or more mutant having improved characteristics compared to the wild-type gene(s). wherein the target multigenic system is optimized based on the identified one or more mutants.

[0093] In other embodiments, compositions contemplated herein include a host cell, a saturated mutant library of a targeted multi-gene system and CRISPR/Cas plasmid construct system.

[0094] In certain embodiments disclosed herein, methods for demonstrating construction and mapping of libraries can include about tens of thousands of mutations in four *E. coli* regulatory pathways that increase lysine production, tolerance, metabolism, and/or homeostasis. These methods can enable identification of specific mutants conferring increased amino acid production, tolerance, metabolism, and/or homeostasis in microorganisms (*e.g.* *E. coli*). In some embodiments, engineered *E. coli* can be created that produce amino acids (*e.g.* lysine) at concentrations of about 40%, to about 50% to about 60% or greater than wild-type microorganisms.

[0095] In certain embodiments, to map mutations to a functional outcome, for example, engineered bacteria (*e.g.* *E. coli*) effects on amino acid (*e.g.* lysine) flux as identified by the methods disclosed above can be used to create a mutation library exposed to an amino acid analog. In accordance with these embodiments, a lysine analog, S-(2-aminoethyl)-L-cysteine ("AEC") can be used for analysis of lysine flux in bacteria. This antimetabolite is an analog to

lysine and competes with canonical lysine for binding to the lysil-tRNA synthetase (LysRS), leading to protein misfolding and reduced cell growth. Additionally, AEC blocks lysine biosynthesis by interacting with riboswitches, inhibiting bacterial growth in the absence of an external lysine source. In this example, mutations that lead to lysine overproduction to out-compete AEC; thereby restore cell growth. In some embodiments, whole genome sequencing can be performed under a selective AEC concentration to identify contribution of individual genes within particular pathways related to lysine flux. Combination methods disclosed herein allow for development of engineered microorganisms that produce target molecules by simple to complex pathways to manipulate production of the target molecules (e.g. amino acids).

[0096] In some embodiments, gene summaries can be then mapped to one or more categories that can affect molecular production; for example of amino acids or other bacterial molecules. In one example, gene summaries can be then mapped to one or more categories that can affect molecular production; for example, lysine production, metabolism, and/or homeostasis including lysine biosynthesis, lysine degradation, lysine regulation, or lysine transport. In some embodiments, deep mapping analysis leads to generation of a comprehensive map of evolutionary trajectories resulting from selective (e.g. AEC) resistance that can identify combinations of one or more mutations that can affect amino acid (e.g. lysine) production, metabolism, and/or homeostasis. In some embodiments, the use of trackable barcodes for each mutant can enable unique characterization of each mutant, beyond those having dominant characteristics, enabling the identification and contribution of all mutations, even minor contributory effects of silent mutations. In some embodiments, methods disclosed herein enable creation of comprehensive map(s) which can highlight which pathway features can be optimized and which specific mutations can lead to phenotypic improvement in amino acid production, metabolism, and/or homeostasis (e.g. lysine, arginine, methionine etc.).

[0097] In some embodiments, engineered *E. coli* of certain embodiments disclosed herein can be used for technological applications. In other embodiments, one or more genes can be mutated in *E. coli* in order for the engineered *E. coli* to tolerate increased lysine concentrations compared to a wild-type *E. coli*. In some embodiments, engineered *E. coli* as disclosed herein can have increased lysine production compared to a control *E. coli*. In certain embodiments, a 10%, or a 20%, or a 30%, or a 40%, or a 50%, or a 60%, or a 70 %, or an 80% or a 90% or more increase in production of lysine can be engineered. In other embodiments, a commensurate level of improved lysine tolerance or homeostasis can be observed in the modified *E. coli* compared to unmodified control *E. coli*.

USE OF ENGINEERED *E. COLI*

[0098] In certain embodiments, engineered microorganisms as described herein can be used in the industrial or commercial production of amino acids or similar target molecules. In some embodiments, the production of intracellular lysine in the engineered *E. coli* can include production of about 800 μM and to about 1400 μM or more per engineered *E. coli* cell. Other embodiments disclosed herein concern increasing intracellular production of lysine to concentrations of about 800 μM , of about 1000 μM , of about 1200 μM , or of about 1400 μM or more. Other embodiments disclosed herein concern increasing lysine expression in engineered *E. coli* as calculated as expression fold change compared to wild type *E. coli* as between about 1.5-10 fold as compare to wild type *E. coli*, in some embodiments about 1.25 fold compared to wild type *E. coli*, in some embodiments about 1.5 fold compared to wild type *E. coli*, in some embodiments about 2 fold compared to wild type *E. coli*, in some embodiments about 5 fold compared to wild type *E. coli*, in some embodiments about 10 fold compared to wild type *E. coli*. In some embodiments, engineered *E. coli* can increase export of lysine to surrounding media for harvesting and reuse of the engineered organisms.

[0099] In other embodiments, kits are contemplated of use to transport or house engineered microorganisms (*e.g.* *E. coli*) having modified amino acid flux (*e.g.* lysine) regarding production, tolerance metabolism, and/or homeostasis. In some embodiments, kits can include components for culturing and growing engineered microorganisms to produce the amino acids or similar molecule (*e.g.* lysine).

[00100] Additional objects, advantages, and novel features of this disclosure will become apparent to those skilled in the art upon review of the following examples in light of this disclosure. The following examples are not intended to be limiting.

EXAMPLES

[00101] The following examples are included to illustrate various embodiments. It should be appreciated by those of skill in the art that the techniques disclosed in the examples that follow represent techniques discovered to function well in the practice of the claimed methods, compositions and apparatus. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the disclosure.

Example 1

Lysine library design and selection strategy

[00102] In one exemplary method, to develop an amino acid (*e.g.* lysine) deep scanning mutagenesis library, genes encoding target regulatory proteins were selected from four categories that affect lysine flux: lysine transport (3 genes), lysine regulation (2 genes), lysine biosynthesis (12 genes), and lysine degradation (2 decarboxylation genes) (**FIG. 1**). Based on the observations above, 16,300 mutations targeting four primary routes that affect lysine flux were designed: lysine biosynthesis (12 genes), lysine degradation (2 decarboxylation genes), lysine transport (3 genes), and regulation of genes in such pathways (2 genes) (**FIG. 1**). For each targeted gene, full saturation mutagenesis libraries were designed and constructed of all residues within a 6 Å shell from known or model-predicted binding sites, encompassing substrate, co-factor, DNA binding or allosteric factors. A comprehensive description of all targeted sites and the respective cassette sequences were identified (data not shown). This strategy permits the scanning of probable targets for proteins with no known functional sites, and an average higher than 50% of known functional sites in the remaining proteins (data not shown). Overall, 3.5-32% of all residues for each of the 19 genes involved in lysine metabolism were fully saturated and illustrated in a histogram plot (data not shown). The constructed plasmid libraries were deep sequenced to confirm coverage. 99% of the designs were cloned successfully into the plasmid backbone, with 91-93% surviving after exposure to Cas9 across two biological replicates (data not shown).

[00103] In order to assess coverage at the genomic level and confirm that edits are indeed introduced in the genome, one targeted genomic window was deep sequenced from each of 5 genes across biological replicates. Overall, 22.6% – 61.6% of the designed edits were measured in these regions (data not shown). Further calculation suggests that the overall genomic editing efficiency can be estimated at 1.6-3.7%, taking into consideration the ratio of edited reads to wild-type, as well as the probability of cells being edited at that specific locus versus the other targeted loci (dataset not shown). These results demonstrate that we are effectively introducing edits at the targeted genomic loci.

[00104] For each targeted protein, 3D structures were collected from the RCSB Protein Data Bank, if available, or modeled using SWISS-MODEL50 or I-TASSER51. A 6 angstrom shell from the binding sites was built using PyMOL (v.1.8.6.2) scripts to select sites for mutagenesis (see **Fig. 19A-19C** for selected sites and structure details). (additional data not shown)

[00105] To map mutations to lysine pathway function, the exemplary library was exposed to the lysine analog S-(2-aminoethyl)-L-cysteine or AEC. This analog competes with canonical

lysine for binding to the lysyl-tRNA synthetase (LysRS), leading to protein misfolding and reduced growth. Additionally, AEC blocks lysine biosynthesis by interacting with riboswitches, inhibiting bacterial growth in the absence of an external lysine source (**FIG. 7**). In certain methods, designer mutations were investigated that influence lysine regulation and overproduction allowing lysine to outcompete AEC and thereby restore cell growth. Sequencing of the plasmid cassettes (herein referenced as barcodes) before and after growth in the presence of AEC allows parallel tracking of each designed mutant in the library, permitting highly parallel mapping to be performed and to assess their contribution to tolerance and their inference to lysine flux (FIG. 17, workflow illustration).

Example 2

Mapping the impact of each pathway category on tolerance and function

[00106] In accordance with these exemplary methods, the lysine deep scanning mutagenesis library exhibited enhanced growth when compared to wild-type cells transformed with either a non-targeting gRNA or a gRNA targeting the unrelated loci *galK* (double-stranded break control or DSB) across a range of AEC concentrations (data not shown). There were no significant growth differences between the non-target and DSB controls under AEC selection, suggesting that the improved growth phenotype observed in the library is not a consequence of DSB-induced adaptation. After 30 hrs, both negative controls began to grow in up to 1,000 μ M AEC, suggesting that spontaneous mutations can also confer AEC tolerance.

[00107] After sequencing the lysine library barcodes before and after selection, fitness contribution of each designer mutation to AEC resistance can be inferred in parallel (data not shown), and then summarized at the gene level. Mutations in several genes demonstrate consistent enrichment across several selective conditions (e.g. *lysP* and *dapF*). The majority of genes, however, demonstrate concentration dependent enrichment, consistent with the expectation that different genes will affect network function to differing levels. Mutations in *dapB*, *lysA* and *lysU* were not significantly enriched in any of the selections performed. Note that when grown in the absence of AEC, the library has an enrichment score centered around 0 (data not shown), indicating that growth in minimal media is not strongly biasing the library.

[00108] In another exemplary method, gene summaries were then mapped to the four design categories described earlier, resulting in a comprehensive map of trajectories leading to AEC resistance (data not shown). Mutations in transporters are the most effective resistance route, which is not surprising as any loss-of-function mutation could prevent cellular uptake of AEC

from the media. The use of barcodes for each mutant enabled us to characterize beyond the dominant selection winner, uncovering the contribution of the remaining categories, as will be discussed below. This analysis provides a comprehensive map of the various strategies typically pursued in pathway-network engineering, highlighting what pathway features need to be optimized and which specific mutations could lead to phenotypic improvement. These exemplary methods improve identification of specific targets, while using genomic reconstruction and validation in order to confirm phenotypic improvement. In additional methods outlined below, different aspects and mutations of this map are analyzed, highlighting important features for genotype-phenotype mapping at such scale.

Example 3

Transporter loss-of-function dominates the selected population

[00109] In other exemplary methods, lysine uptake was analyzed. Lysine uptake is mediated by three different transporter systems in *E. coli* (**FIG. 1**). ArgT codes for a periplasmic binding protein specific to lysine, arginine and ornithine, interacting with the ABC transporter coded by the hisJQMP operon. CadB is part of the Cad system, which plays a role in pH homeostasis under acidic conditions. This transporter imports lysine and excretes the decarboxylated product cadaverine in conditions of low external pH and presence of exogenous lysine. LysP is a specific transporter for lysine, but also has a regulatory role in activating the Cad system through transmembrane interactions with CadC. Mutations in lysP were identified as the most highly enriched, including the dominant selection winner (data not shown). No enrichment for lysP mutations were observed when cells were grown in the absence of AEC (data not shown). This correlates with previous findings that identified lysP mutations in AEC resistant strains

[00110] When mapped at single amino acid resolution, significantly enriched mutations were observed across all targeted regions in lysP (data not shown). The relatively even distribution of enriched mutations across all targeted positions in the gene suggests loss-of-function, and thereby abrogated AEC transport. These mutations map across a substantial spatial fraction of the modeled structure (image not shown), further supporting disruption of LysP function. Genome modified strains were individually constructed for two highly enriched mutations, T33F and Q219I. These two mutants grow similar to wild-type cells (transformed with a non-targeting gRNA) in the absence of AEC, but exhibit superior growth under increasing AEC concentrations (**FIG 5A** and **5B**).

[00111] Further, enrichment of synonymous mutations in lysP under AEC selection were observed. It's well established that synonymous mutations can have an effect on the levels, stability and folding of both mRNA and proteins, making these observed mutations candidates for further observation regarding altering expression or stability of LysP and thereby conferring AEC tolerance. Several synonymous mutations were enriched under weak selective pressure (10 μ M AEC), suggesting that small fluctuations in LysP levels may be sufficient to confer low levels of resistance (**FIG. 6**). As selective pressure is increased up to 10,000 μ M AEC, fewer synonymous mutations were still enriched, suggesting that these mutations are introducing more drastic effects on LysP levels. Overall, the frequency of synonymous mutations affecting LysP activity is substantially higher than that observed for other targeted proteins (**FIG. 6**), highlighting an unusually strong effect of synonymous substitutions on this transporter. Since this effect is not restricted to the beginning of the gene (commonly associated with regulation of translation initiation), this result could indicate that co-translational folding is essential for LysP function. It is noted that changes in codon usage or disruption of important transcript secondary structures would alter ribosome attenuation sites required for proper folding, for example.

[00112] These observations demonstrate that deep scanning strategy maps tolerance mutations of use in creating selective construct for improved tolerance and/or production. The high fraction of lysP mutants in the selected population (>95%) suggest that this is a critical trajectory to evolve AEC resistance. Directed evolution studies have demonstrated that the vast majority of mutations within a protein are known to negatively affect protein function and stability (c.a. 30-50% are strongly deleterious, 50-70% are slightly deleterious or neutral), with only a handful (0.01-1%) typically improving-altering function. As such, it is not surprising that the dominant clones in these selections involved loss-of-function mutations, which further validates the importance of rapid screening of thousands of mutations simultaneously in order to identify mutants with improving effects. More importantly, this outcome highlights the importance of the use of barcoding or another method for deeply scanning selected libraries to identify a plurality of mechanisms for altering the phenotype of interest (*e.g.* increased pathway flux vs decreased inhibitor flux), allowing exploration beyond a local optimum in the fitness landscape.

Example 4

Beyond the dominant selection winner: a non-obvious mechanism in DapF

[00113] With the strong dominance of *lysP* mutations (>95%) in the selected population, identifying hits beyond the main selection winner would be challenging with traditional approaches. In other exemplary methods, to demonstrate parallel tracking in this technology, hits in the remainder (<5%) of the population were validated. Among the biosynthetic genes, mutations in *dapF* were highly enriched across multiple selective pressures. This gene encodes an epimerase catalyzing the penultimate step in the biosynthetic pathway, a conversion of LL-diaminopimelate (LL-DAP) to meso-diaminopimelate (meso-DAP). *DapF* mutations ranked as the most enriched non-*lysP* mutant under 100 μ M AEC and the second most under 1,000 μ M AEC, although no strong enrichment was observed under 10,000 μ M AEC. Two highly enriched mutants, G210D and M260Y, for further analysis (data not shown).

[00114] In certain exemplary methods, both G210D and M260Y substitutions which lie close to the protein catalytic site (data not shown), suggesting an effect on catalytic activity. After genomic reconstruction, both mutants grew similarly to wild-type cells in the absence of AEC, but displayed distinct phenotypes when put under selective pressure. *DapF* G210D mutants had high growth rates up to 10,000 μ M AEC (data not shown), confirming the barcode enrichment previously observed. However, *DapF* M260Y grew similarly to wild-type cells in the presence of AEC (data not shown). The *DapF* G210D mutant was regrown and independently tested and the same phenotype having superior growth in the presence of AEC was demonstrated. In order to rule out adaptive mutations in *lysP*, this locus was sequenced after the selective growth and observed no mutations in this region. Mass spectrometry quantification revealed a significantly higher intracellular level of lysine in both mutants compared to wild-type cells (**FIG. 9** and **10A**), with G210D accumulating 51% more lysine and M260Y accumulating 111% more (**FIG 10B**).

[00115] To further investigate the mechanism behind these *dapF* mutations, wild-type and the mutant *DapF* variants were purified and their kinetics measured *in vitro* (data not shown). Surprisingly, both *DapF* mutants are kinetically impaired relative to the wild-type variant (data not shown). qPCR profiling of the entire biosynthetic pathway revealed one gene with statistically significant increase in gene expression, the diaminopimelate decarboxylase *lysA* (data not shown). *LysA* is responsible for the last enzymatic step in lysine biosynthesis, and it's known to be repressed by lysine and induced by diaminopimelic acid through the regulator *LysR*. As such, the increased expression of *lysA* in a *dapF* impaired background suggests that a larger pool of LL-DAP works as a stronger co-effector to activate *lysA* than the wild-type mixture of both LL-DAP and meso-DAP.

[00116] It has been observed that these surprising results uncovered a counter-intuitive interplay between lower kinetics and lysine overproduction. This finding highlights the limited ability to predict genotype-phenotype relationships in the context of an entire pathway, similar to what has been observed in the protein engineering field. Therefore, deep scanning mutagenesis proves to be a valuable strategy to identify novel regulatory mechanisms on pathway scale.

Example 5

Validating other hits: decoupling noise from real enrichment

[00117] Since plasmid barcodes are used as a proxy for identifying genomic edits, lack of correlation introduces noise that can lead to false positives in the enrichment scores. In theory, plasmid-genome correlation should be strong for real hits with strong enrichment, and weaker for non-enriched variants. To investigate this further, the regulator category was analyzed and investigated regarding a weakly enriched mutation in LysR, as well as a strongly enriched mutation in ArgP.

[00118] Regulatory mutations are well known to confer AEC resistance, mainly in the lysine-regulated riboswitch controlling expression of the aspartokinase lysC (data not shown). The regulator LysR, which upon binding to diaminopimelic acid activates the last enzymatic step in lysine biosynthesis (lysA, **FIG. 1**), exhibited few weakly enriched mutations in this exemplary library (data not shown). In these exemplary methods, a LysR S36R substitution was examined, a mutant that had significant enrichment scores at 1,000 μ M AEC (p-value: 0.007), while at lower concentrations enrichment was not significant (p-value of 0.14 at 10 μ M AEC and 0.12 at 100 μ M AEC).

[00119] The LysR family of transcription regulators is ubiquitous in bacteria and comprises a conserved N-terminal helix-turn-helix (HTH) DNA-binding domain and a less conserved C-terminal co-inducer binding domain. The LysR S36R mutation lies on the DNA-binding (HTH) domain. However, after reconstruction and genomic verification of this edit, it was observed that mutants do not display any alteration in intracellular lysine levels (**FIG. 18B**). Further, it was observed that strains harboring the S36R mutation grew slower than wild-type cells transformed with a non-targeting gRNA (data not shown). These results suggest that the enrichment observed at the plasmid-barcode level for LysR is possibly a false positive. This section discusses some of the caveats of CREATE. This example illustrates how a mutant that appeared as a positive hit in the CREATE screening (e.g. LysR_S36R), turned out to be a false

positive because upon further investigation, the mutation did not improve lysine production or increase tolerance to AEC.

[00120] However, the engineered *E. coli* *lysR*_S36R mutants grew more slowly than wild-type cells transformed with a non-targeting gRNA (Supplementary Figure 6), which was not in accordance to the barcode enrichment described in Example 4. This finding demonstrates some of the complexity in the relationship between the selection environment and the fitness effect. In unicellular asexual organisms such as bacteria, fitness in a competitive environment can be mainly attributed to three parameters: 1) lag phase duration; 2) exponential growth rate; and 3) maximum yield and at saturation. Mutations can affect fitness through differing degrees on each of these parameters. Moreover, the effect of each parameter is further confounded in more complex populations, in which clonal interference has a strong effect on shaping the adaptation dynamics and evolutionary outcomes. Therefore, it is reasonable to speculate that the LysR S36R substitution cannot directly compete against adaptive mutations arising in wild-type cells, but does enrich over time in a competition among all other variants in the population. This further underlies the utility of the method described herein in which the profound depth of barcode sequencing allows for the identification of real mutations that display weak enrichment.

[00121] On the other hand, the ArgP regulator displayed much stronger enrichment scores for a E246Q substitution (**FIG. 18C**), with a p-value of 1.6×10^{-6} at 100 μM AEC, 8.1×10^{-8} at 1,000 μM AEC, and 1.59×10^{-5} at 10,000 μM AEC. ArgP, which also belongs to the LysR family of transcriptional regulators, can bind to lysine in order to inhibit transcription of several genes in the biosynthetic lysine pathway (**FIG. 1**), acting as one of the main negative feedback mechanisms. The E246Q substitution lies on the C-terminal co-inducer binding domain, although the apparent role for this residue is unclear. After genomic reconstruction, it was observed that strains harboring the ArgP E246Q mutation accumulated 124% more intracellular lysine (data not shown), likely responsible for the barcode enrichment previously observed, although the reconstructed mutant could also not outcompete the wild-type strain (data not shown), similarly to the results observed for the DapF M260Y mutation. This example suggests that the system was able to identify mutations that would not be uncovered by standard techniques but mechanisms behind this mutation remain unclear. This mutation had improved effects on amino acid metabolism; increase lysine production compared to controls. This observation reinforces the power of the strategy, as non-obvious solutions can be identified through these methods.

[00122] Overall, these observations support the concept that strongly enriched mutations are more likely to yield a real signal than mutations displaying weak enrichment scores. However, as discussed below, adaptive mutations could also introduce noise in the form of strong enrichment scores. Therefore, genomic reconstruction and validation are essential in order to confirm targets identified by this approach. Further, a more stringent p-value threshold with improved statistical methods could filter a larger fraction of false-positives in the sample.

[00123] Complex phenotypes are often engineered through directed evolution or other random mutagenic strategies. While successful for phenotype optimization in industrial strains, off-target mutations can decrease overall cell fitness and lead to “dead-end” phenotypes, preventing further improvement of the evolved strain. New tools that combine targeted deep-scanning mutagenesis with genotype-phenotype mapping provide a powerful framework to explore distinct hypotheses in parallel, uncovering mechanisms that would be difficult to rationalize in complex systems. This concept was evident for the DapF mutations investigated here, in which lower kinetics counter-intuitively improved lysine accumulation in the strains.

[00124] Further, the ability to map deeply, through the use of barcodes, enabled quantification beyond the main selection winner. Transporter loss-of-function was a clear solution to the AEC challenge, dominating most of the selected population. Therefore, looking beyond lysP mutations would be challenging with traditional strategies. However, using methods disclosed herein other hits that were being masked by the enrichment of lysP mutations were identified, highlighting the value of parallel genotype-phenotype mapping. It is noted that some of the mutants described herein could not outcompete adaptive mutations that inactivated lysP, growing similarly to wild-type cells even though a clear improvement in lysine accumulation was observed. This finding underlines the complex relationship between the selection environment and the fitness effect. In unicellular asexual organisms such as bacteria, fitness in a competitive environment can be mainly attributed to three parameters: (1) lag phase duration, (2) exponential growth rate, (3) maximum yield at saturation. Mutations can affect fitness through differing degrees on each of these parameters. Moreover, the effect of each parameter is further confounded in more complex populations, in which clonal interference has a strong effect on shaping the adaptation dynamics and evolutionary outcomes. These different adaptive niches could explain the results observed here, and recently developed tools could aid in the elucidation of these evolutionary niches at the population scale.

[00125] These methods identified successful mapping of different AEC resistance routes in parallel. A few parameters must be taken in consideration when attempting genotype-

phenotype mapping on a pathway scale. First, applications that include strong selective pressures are more likely to succeed. With the relatively low (2-4%) editing efficiencies reported herein, the screening burden would be too high for most screening throughputs. Second, while these technologies efficiently narrow the search space to a few hypotheses (genes and specific mutants) of interest, reconstruction in wild-type backgrounds and subsequent validation is essential. Third, sequencing depth remains an important consideration. In this study, the selective dominance of lysP mutations likely prohibited the investigation of every single designed edit. A rarefaction curve should be included in future studies in order to assess the required sequencing depth. Finally, strategies to improve map accuracy would be valuable additions. As an example, the use of single cell-specific barcodes could improve the confidence of mapping, so that each single mutation is mapped as a population of cells. Transferring barcodes from plasmids to genomes could also decrease cell-to-cell variation and hence decrease noise in barcode enrichments. In the specific case of lysine metabolism, comparing mutations identified in the presence of different antimetabolites or with screening-based approaches using lysine biosensors could be a valuable contribution.

[00126] Exemplary methods herein demonstrated expansion of deep scanning mutagenesis strategies from a single gene to an entire metabolic pathway. In parallel, multiple routes of AEC resistance were identified, encompassing mutations in transporters, regulators and biosynthetic genes. This technology, as well as future implementations that addresses some of the limitations described above, should accelerate the ability to investigate, understand and control, complex multigenic phenotypes, providing knowledge that will contribute to the forward engineering of these traits.

Materials and methods

Genome edited strains, plasmids and general cloning procedures

[00127] Genome editing and individual mutant validation was performed in a wild-type *Escherichia coli* str. K-12 substr. MG1655 strain. A custom pSIM5-Cas9 dual-vector was built by cloning the araC-pBAD-Cas9 fragment from pX2-Cas9 vector (Addgene #85811) into the temperature sensitive pSIM5 plasmid containing the lambda red genes. This pSIM5-Cas9 dual-vector was transformed into *E. coli* MG1655 prior to the library introduction. The editing cassettes containing the homology arm and genome-targeting gRNA were cloned in the same backbone previously used for CREATE .

[00128] Cloning procedures that did not involve libraries were performed using CPEC. Briefly, fragments containing 40bp homology arms were PCR amplified using Phusion High-Fidelity PCR Master Mix (New England BioLabs), treated with DpnI to remove methylated plasmid templates when necessary, and purified from 1% agarose gels using the QIAquick Gel Extraction Kit (QIAGEN). CPEC assembly was performed using 300 ng of backbone and equimolar insert amounts. After 10 cycles of reaction, the product was dialyzed and transformed via electroporation into E. cloni 10GF' ELITE Electrocompetent Cells (Lucigen). Cloning procedures for the library preparation will be detailed below.

Library design

[00129] For each targeted protein in this study, 3D structures were collected from the RCSB Protein Data Bank if available or modeled using SWISS-MODEL or I-TASSER. A 6 Å shell from binding sites was built using PyMOL (v.1.8.6.2) scripts to select sites for mutagenesis. A comprehensive list of all selected sites and structure details were identified (data not shown). In total, 19 genes and 815 sites were selected. For each selected site, a full codon saturation mutagenesis was introduced using the most frequent codons, resulting in a total of 16,300 variants. For each variant, the gRNA and homology arm designs were automated using previously described Python scripts. Briefly, the cassette design included the following features: a library-specific 18 nt priming site for subpooling, a 12 nt variant-specific priming site (not used in this study), a 118 nt homology arm encoding the specific genomic edit and a synonymous PAM mutation in close proximity, the constitutive promoter J23119 (35 nt), a 3 bp spacing sequence (ATC), the 20 nt spacer region required for Cas9 targeting, followed by 24 nt of the 5' end of the canonical *S. pyogenes* gRNA. The full list of cassette sequences can be provided but is not shown.

Library construction

[00130] The designed library was synthesized as 230-mers by Agilent Technologies in a custom array and delivered pooled as lyophilized single-stranded DNA. As described in more details previously, the oligo pool was subjected to an Alexa Fluor 488-labeled strand extension reaction and purified in a 6% SDS-PAGE gel to remove indels introduced in the synthesis process. From the resulting purified oligo pool, the lysine library was amplified as a single subpool using predefined library-specific priming sites included in the cassette design. The amplification was optimized to minimize overamplification in an effort to reduce product crossover. The PCR reaction was performed using Phusion High-Fidelity PCR Master Mix

(New England BioLabs) and the following reaction conditions: 98°C for 60 seconds, followed by 8 cycles of 98°C_{30s}/68°C_{30s}/72°C_{90s}, followed by 10 cycles of 98°C_{30s}/72°C_{90s} and then a final extension at 72°C for 3 minutes. The library product was purified from 1% agarose gels using the QIAquick Gel Extraction Kit (QIAGEN).

[00131] The amplified library was cloned using Gibson Assembly Hi-Fi 1-Step Kit (SGI-DNA), with 300 ng of the linearized backbone and 30 ng of the library insert. The cloning reaction was dialyzed and then transformed via electroporation into E. cloni 10GF' ELITE Electrocompetent Cells (Lucigen), in a single electroporation using a 0.2 cm gap cuvette (GenePulser, Bio Rad). Cloning efficiency was estimated by counting colonies in LB agar plates. Overall, >60X coverage (total CFUs/number of library variants) were achieved at the cloning stage. Subsequently, the library was grown in LB media to saturation and plasmid was extracted using the QIAprep Spin Miniprep Kit (QIAGEN). The plasmid library was then transformed in *E. coli* MG1655 following a modified recombineering protocol. Briefly, the strain previously transformed with the dual Cas9/pSIM5 vector was grown at 30°C in LB media in 250 mL flasks under 200 rpm until mid-log phase ($OD_{600} = 0.4-0.5$). Cells were then induced with 0.2% arabinose (for Cas9 induction) and placed in a 42°C shaking water bath for 15 minutes (for lambda red induction). Next, cells were kept on ice for 5 minutes and made electrocompetent. To ensure coverage, 2 µg of the plasmid library was transformed in a single electroporation using a 0.2 cm gap cuvette (GenePulser, Bio Rad). Two independent transformations were performed for the library (biological duplicates), followed by recovery in 5 mL of LB media supplemented with 0.2% arabinose for 3 hours at 30°C. Afterwards, cells were plated in LB media with the proper antibiotics to calculate transformation efficiency and transferred to 30 mL of liquid LB media with antibiotics for 8 hours before proceeding to selective conditions. Overall, >300X coverage was achieved at this stage (total CFUs/number of library variants). Both the cloning and recombineered libraries were sequenced using an Illumina MiSeq run to assess the real plasmid library coverage (threshold set at 100% full matching cassettes, Figure EV1). Deep sequencing procedures for plasmid libraries are detailed below.

AEC selections and high-throughput sequencing of the library barcodes

[00132] Selection was performed in 30 mL of M9 minimal media containing 5X M9 Minimal Salts (BD Biosciences), 2 mM magnesium sulfate, 0.1 mM calcium chloride, 1% glucose, 100 µg/mL carbenicillin (to select for the library plasmid) and varying S-(2-aminoethyl)-L-cysteine (AEC) concentrations (0-10,000 µM). The library culture growing for 8 hours in LB media

(described above) was washed with PBS and 10 μ L was used to inoculate the selective media. Cultures were kept at 37°C under 200 rpm. Two different selection controls were included, all subjected to the same construction procedure described above: (1) a non-targeting control, containing a plasmid with a gRNA that does not target the *E. coli* genome, (2) a double-stranded break control, containing a plasmid with a CREATE cassette designed to introduce a stop codon at the unrelated gene *galK*.

[00133] Selections up to 1,000 μ M AEC were harvested to sequence the library barcodes at 30 hours post-inoculation, and the 10,000 μ M AEC selections were harvested at 40 hours post-inoculation. To do so, 3 mL of the selection cultures were pelleted and plasmid DNA was extracted using the QIAprep Spin Miniprep Kit (QIAGEN). Next, custom Illumina compatible primers were used to barcode each selection using Phusion High-Fidelity PCR Master Mix (New England BioLabs), 300 ng of the plasmid prep, 3% DMSO, and the following cycling conditions: 98°C for 30 seconds, 20 cycles of 98°C_{10s}/68°C_{15s}/72°C_{20s}, followed by a final extension of 72°C for 5 minutes. PCR products were purified from 1% agarose gels using the QIAquick Gel Extraction Kit (QIAGEN), pooled together in equimolar amounts, and sequenced using an Illumina MiSeq 2x150 paired end reads run.

Processing of the library barcode reads and statistical analysis

[00134] Reads were demultiplexed and then merged using the PANDAseq assembler (v2.10). Merged reads were matched to the database of all designed cassettes using the usearch_global algorithm (v9.2.64), with an identity threshold of 95% and minimal alignment length of 150 bp. These parameters were chosen so that chimeras in the designs could be evaluated. 40 possible hits were allowed for each query, which were subsequently sorted by percent identity and the best-matching cassette was chosen. To generate read counts for each designed cassette, only reads that had a full alignment and an identity higher than 99% were used. The number of reads obtained at each processing step was outlined, data not shown.

[00135] The next processing steps of the read counts were done using the Pandas data analysis python package (v0.20.2). First, since low-count variants are subject to counting error, variants with initial counts (pre-selection) of less than 10 were not included in the individual biological replicate analysis. Then, variants with 0 counts post-selection were replaced to 0.5 in order to allow the subsequent calculation steps. For each individual biological replicate, enrichment scores were calculated as the logarithm (base 2) of the ratio of frequencies between post-selection to pre-selection. Frequencies were determined by dividing the read counts for

each variant by the total experimental counts. Finally, a weighted average was used to combine the enrichment scores obtained in the two biological replicates, according the formula:

$$Wavg = \frac{\sum_{i=1}^N (Ci \times Wi)}{\sum_{i=1}^N Ci}$$

where, Wavg is the weighted average score, i is the biological replicate, C is the read count obtained for the variant in the biological replicate and W is the enrichment score calculated for the variant in the biological replicate.

[00136] To assess significance, the average of enrichment scores for all synonymous mutations included in the library was calculated (average μ of wild-type enrichment). Bootstrap analysis (resampled with replacement 20,000 times) was performed to obtain a 95% confidence interval for the wild-type enrichment average μ . Variants were considered as significantly enriched if their weighted enrichment scores were at least $\mu \pm 2*\sigma$ (i.e. p-value ≤ 0.05 assuming a normal distribution of synonymous mutations enrichment scores), with σ being the standard deviation. For individual mutants chosen to be investigated further in this study, the p-value of their respective enrichment scores was calculated using the probability density function of all mutants under the specific selective pressure.

Deep sequencing of selected genomic regions

[00137] Selected genomic pockets (data not shown) were PCR amplified with primers that included the Nextera adapter sequences as overhangs (Forward primer: SEQ ID NO: 20, 5' - TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG - [locus- specific sequence] - 3'; Reverse primer: SEQ ID NO: 21, 5' - GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG - [locus- specific sequence] - 3'). Samples were then prepared with the Nextera XT DNA Library Prep Kit (Illumina) and sequenced on an Illumina NextSeq 2x150 paired end reads run. Sequencing reads were merged using the PANDAseq assembler (v2.10) and trimmed to the selected positions (these positions exclude the primer binding site, illustration not shown). A database was generated containing all expected sequence variants for the full length between the sequenced positions, which is the wild-type sequence and all designed edits incorporated into the respective positions. Reads were then matched at 100% identity to this database using custom Python scripts. The number of reads obtained at each processing step was identified.

Individual mutant reconstruction

[00138] To individually reconstruct the mutants investigated in these methods, the same cassette sequence included in the library for that specific variant was obtained. The cassette was then cloned, sequence verified and introduced in *E. coli* MG1655 using the same procedure described above. Then, the specific genomic edit was confirmed through Sanger sequencing of the target site.

Absolute quantification of intracellular lysine levels

[00139] Saturated overnight cultures of the reconstructed mutants were used to inoculate 100 mL of the minimal media used for selections (without any AEC present). Inoculums were made to an initial OD₆₀₀ of 0.01, and cultures were grown in shake flasks at 37°C under 200 rpm until OD₆₀₀ reached 0.5. At this stage, cells were plated to calculate CFUs/mL, washed with PBS, pelleted by centrifugation and stored at -80°C for metabolite extraction. The frozen cell pellets were extracted in ice cold lysis buffer, a 5:3:2 ratio of MeOH:ACN:H₂O, containing amino acid standard mix at a final concentration of 1 µM (MSK-A2-1.2 standard amino acid mix, purchased from Cambridge Isotope Laboratories, Inc. - Tewksbury, MA). Samples were vortexed for 30 minutes at 4°C with 1mm glass beads. Insoluble proteins and lipids were pelleted by centrifugation at 4°C for 10 minutes at 12,000g. Supernatants were collected and analyzed using a Thermo Vanquish UHPLC coupled online to a Thermo Q Exactive mass spectrometer. UHPLC-MS methods and data analysis approaches were performed as described previously. The intracellular concentration of wild-type control samples was normalized to 1, and the experimental samples are reported as fold-change relative to these wild-type levels.

Expression and purification of the DapF mutants

[00140] The *dapF* variants were PCR amplified from boiled cells that contained the desired mutation (wild type *E. coli* MG1655 for the wild type *dapF* sequence; reconstructed *dapF* mutants for the G210D and M260Y variants). The PCR products were then cloned and sequence verified into a custom made pET-3 backbone, containing the histidine tag (6x) on either the 5' or 3' end of the genes to test for optimal expression. *Corynebacterium glutamicum* DAP Dehydrogenase was synthesized from Eurofins Genomics and also cloned in the pET-based vector. Expression was done in a *E. coli* BL21 strain using LB media, which was induced with 1 mM IPTG when OD₆₀₀ reached 0.6. Induced cultures were grown at 30°C overnight under 200 rpm, harvested by centrifugation, and the pellet stored at -80°C for protein purification.

[00141] Proteins were purified using the Ni-NTA Spin Kit (QIAGEN), following the protocol for purification of tagged proteins under native conditions. Purified samples were run on a denaturing PAGE gel (Mini-PROTEAN TGX Stain-Free Precast Gels, Bio-Rad) to confirm purity and quantified using the Thermo Fisher Scientific Pierce 660 nm Protein Assay Reagent. Purified proteins were used fresh for the kinetic assay (never frozen).

In vitro assay to measure DapF kinetics

[00142] Enzymatic activity of the DapF variants was determined *in vitro* using a modified DAP epimerase-DAP dehydrogenase coupled spectrophotometric assay (Cox *et al.*, 2002). Briefly, 100 mM Tris (pH 7.8), 0.1 mM diaminopimelic acid (racemic mixture), 0.44 mM NADP⁺ and 1 mM DTT was added to a cuvette and incubated at 37°C for 10 minutes to equilibrate the temperature. Then, 1.8 mM DAP Dehydrogenase was added and the absorbance was recorded at 340 nm until it reached a plateau (i.e. all meso-DAP was depleted) (Appendix Figure S4). Next, varying amounts of the purified DapF variants were added, and the absorbance at 340 nm followed through time. The assay was performed with 400 μ L final volume in a NanoDrop One^C UV-Vis Spectrophotometer (Thermo Fisher Scientific Inc.).

Quantitative analysis of gene expression

[00143] Wild type *E. coli* MG1655 and the analyzed reconstructed mutants were grown under the same conditions as described for absolute intracellular lysine quantification. At the harvest stage ($OD_{600} = 0.5$), 1 mL of the culture was treated with RNAprotect Bacteria Reagent (QIAGEN) to stabilize the RNA and the resulting pellet frozen at -80°C. Total RNA was then extracted using the RNeasy Mini Kit (QIAGEN) with an on-column DNase digestion. cDNA was synthesized using the SuperScript IV First-Strand Synthesis System (Invitrogen). Power SYBR Green Master Mix (Thermo Fisher Scientific Inc.) was then used for the QPCR reactions, which was run on a QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific Inc.) with the following conditions: 95°C for 30 seconds, 40 cycles of 95°C_{30s}/65°C_{30s}/72°C_{30s}, followed by the standard melting curve protocol. Three different housekeeping genes were tested as QPCR endogenous controls: the 5S ribosomal RNA (*rrfA*), syroheme synthase (*cysG*) and the integration host factor B (*ihfB*). After testing each endogenous control, *ihfB* exhibited variability among samples, and so *rrfA* and *cysG* were chosen as endogenous controls for the analysis. Relative expression was calculated using the $\Delta\Delta C_t$ method on the Thermo Fisher Cloud Data Analysis Apps (qPCR Module).

Adaptive evolution and whole genome sequencing

[00144] The adaptive evolution experiments were performed with wild-type *E. coli* MG1655 (without any plasmids) in 30 mL of the same minimal media used for selections, containing 1000 μ M AEC. Cells were grown at 37°C under 200 rpm in two different regimes: (1) growth for 48 hours (single-batch) since the inoculation; (2) growth for 5 days, with passages to new media every 24 hours (100 μ L was transferred in each passage). Additionally, wild-type *E. coli* MG1655 cells were also grown for 48 hours in minimal media without any AEC present (parent strain genome). Next, the final cultures were streaked to agar plates of the same selective media and single colonies were processed for whole genome sequencing. To do so, genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega), libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina) and sequenced on an Illumina MiSeq 2x150 paired end reads run.

[00145] Reads were then mapped to the reference *Escherichia coli* str. K-12 substr. MG1655 genome (RefSeq NC_000913.3), using Bowtie2 (v2.3.2) in the *sensitive* preset and *end-to-end* mode. After mapping, SNPs calling was done through samtools (v1.5) with the following filtering parameters: (1) Phred quality score higher than 20, (2) SNP read depth higher than 10, (3) SNP frequency higher than 50%. Finally, the SNPs called in the sequenced parent genome were subtracted from the SNPs called in the adapted strains, yielding the final list of SNPs. The number of reads obtained at each processing step was identified, data not shown.

The foregoing discussion of the disclosure has been presented for purposes of illustration and description. The foregoing is not intended to limit the disclosure to the form or forms disclosed herein. Although the description of the disclosure has included description of one or more embodiments and certain variations and modifications, other variations and modifications are within the scope of the disclosure, *e.g.*, as can be within the skill and knowledge of those in the art, after understanding the present disclosure. It is intended to obtain rights which include alternative embodiments to the extent permitted, including alternate, interchangeable and/or equivalent structures, functions, ranges or steps to those claimed, whether or not such alternate, interchangeable and/or equivalent structures, functions, ranges or steps are disclosed herein, and without intending to publicly dedicate any patentable subject matter.

WHAT IS CLAIMED IS:

1. An engineered *E. coli* comprising, one or more modifications to one or more genes comprising one or more genes of lysine flux comprising lysine biosynthesis, lysine regulation, lysine degradation, and lysine transport, wherein the modifications increase at least one of lysine tolerance and lysine production in the engineered *E. coli*.
2. The engineered *E. coli* according to claim 1, wherein the one or more genes comprise one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC* and *dapD*.
3. The engineered *E. coli* according to claim 1 or claim 2, wherein the one or more modifications comprise or further comprise one or more mutations to one or more gene comprising one or more of *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*.
4. The engineered *E. coli* according to claim 1, wherein the one or more modifications comprise, one or more mutations of a binding site of one or more polypeptide, comprising a substrate binding site, a co-factor binding site, a DNA binding site, an allosteric factor binding site or combination thereof.
5. The engineered *E. coli* according to claim 1, wherein the one or more modifications reduce uptake of S-(2-aminoethyl)-L-cysteine (AEC) by the engineered *E. coli* compared to a control.
6. The engineered *E. coli* according to any one of claims 1 to 5, wherein the one or more modifications comprise one or more single nucleotide polymorphisms (SNP) to the one or more genes.
7. The engineered *E. coli* according to claim 6, wherein the one or more single nucleotide polymorphisms (SNP) comprise one or more of *dapF* G210D, *dapF* M260Y, *lysP* T33F, and *lysP* Q219I.

8. The engineered *E. coli* according to claim 1, wherein the one or more modifications lead to an amino acid sequence represented by one or more polypeptide represented by SEQ ID NOs: 1-5.
9. The engineered *E. coli* according to claim 1, wherein the one or more modifications comprise one or more mutations to *lysP*.
10. The engineered *E. coli* according to claim 1, wherein the one or more modifications comprise one or more mutations to *dapF*.
11. The engineered *E. coli* according to claim 1, wherein the one or more modifications comprise one or more mutations to *lysR*.
12. The engineered *E. coli* according to claim 1, wherein the one or more modifications lead to an increase in *lysA* expression.
13. The engineered *E. coli* according to claim 1 wherein the engineered *E. coli* comprises increased production of lysine compared to a control *E. coli* not having the one or more modifications.
14. A method for producing lysine from the engineered *E. coli* according to any one of claims 1 to 13 comprising:
 - culturing the engineered *E. coli* under conditions sufficient to produce increased concentrations of lysine, and
 - recovering lysine from the engineered *E. coli*.
15. The method according to claim 14, wherein the one or more engineered *E. coli* comprise: one or more mutations of a binding site of one or more polypeptide, comprising a substrate binding site, a co-factor binding site, a DNA binding site, an allosteric factor binding site or combination thereof.
16. The method according to claim 14, wherein the increased production of lysine in the engineered *E. coli* comprises at least a 5.0% increase in lysine compared to a wild type *E. coli*.

17. The method according to claim 14, wherein the one or more modifications comprise one or more single nucleotide polymorphisms (SNP) to the one or more genes.
18. The method according to claim 17, wherein the one or more single nucleotide polymorphisms (SNP) comprise one or more of *dapF* G210D, *dapF* M260Y, *lysP* T33F, and *lysP* Q219I.
19. A vector comprising:
a promoter;
an editing cassette comprising a selectable marker, wherein the selectable marker comprises one or more modifications to one or more genes comprising one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*; and
an associated spacer.
20. The vector according to claim 19, wherein the vector is a plasmid.
21. The vector according to claim 19, wherein the one or more genes encode one or more polypeptides represented by SEQ ID NOs: 1-5.
22. A method for producing an engineered *E. coli* comprising:
introducing into *E. coli* a vector that encodes one or more modifications to one or more genes comprising one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*; and
obtaining viable engineered *E. coli* expressing the vector, wherein the engineered *E. coli* have at least one of increased lysine tolerance and increased lysine production.
23. The method according to claim 22, wherein the vector is a plasmid.
24. The method according to claim 22, wherein the one or more modifications introduce one or more single nucleotide polymorphisms (SNP) to the one or more genes.
25. A method for producing an engineered *E. coli* comprising:

introducing into *E. coli*, a first vector that comprises a polynucleotide having a sequence that encodes a nuclease-deactivated CRISPR-associate (Cas) protein; and

a second vector comprising at least one short guide RNA (sgRNA) molecule comprising a CRISPR-associated (Cas) protein binding site and a targeting RNA sequence directed to a target a nucleic acid sequence of one or more genes comprising one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC*, *dapD*, *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*; and

obtaining viable engineered *E. coli* expressing the second vector, wherein the engineered *E. coli* have at least one of increased lysine tolerance and increased lysine production.

26. The method according to claim 25, wherein the Cas protein is Cas9 or Cas 12a.

27. The method according to claim 25, further comprising a traceable barcode positioned outside of an open reading frame in the targeting RNA sequence, wherein the traceable barcode is linked to a modification of the one or more genes.

28. The method according to claim 25, wherein the one or more modifications introduce one or more single nucleotide polymorphisms (SNP) to the one or more genes.

29. A kit comprising:

one or more containers; and

an engineered *E. coli* comprising, one or more modifications to one or more genes comprising one or more genes of lysine biosynthesis, lysine regulation, lysine degradation, and lysine transport, wherein the modifications increase at least one of lysine tolerance and lysine production in the engineered *E. coli*.

30. The kit according to claim 29, wherein the one or more modifications comprise one or more mutations to one or more gene comprising one or more of *dapF*, *lysP*, *lysR*, *lysC*, *serC* and *dapD*.

31. The kit according to claim 29 or 30, wherein the one or more modifications comprise one or more mutations to one or more gene comprising one or more of *cadA*, *argT*, *dapE*, *dapA*, *lysA*, *lysS*, *argP*, *argD*, *asd*, *lysU*, *cadB*, *dapB*, and *ldcC*.

32. The kit according to claim 29, 30 or 31, wherein the one or more modifications comprise:

one or more mutations of a binding site of one or more polypeptide, comprising a substrate binding site, a co-factor binding site, a DNA binding site, an allosteric factor binding site or combination thereof.

33. The kit according to any one of claims 29 to 32, wherein the one or more modifications comprise one or more single nucleotide polymorphisms (SNPs) to the one or more genes.

34. The kit according to any one of claims 29 to 33, wherein the one or more single nucleotide polymorphisms (SNPs) comprise one or more of *dapF* G210D, *dapF* M260Y, *lysP* T33F, *lysP* Q219I, and *lysR* S36R.

35. An engineered microorganism comprising, one or more modifications to one or more genes comprising one or more genes of amino acid flux comprising amino acid biosynthesis, amino acid regulation, amino acid degradation, and amino acid transport, wherein the modifications increase at least one of amino acid tolerance and amino acid production in the engineered microorganism expressing one or more amino acid.

36. The engineered microorganism according to claim 35, wherein the one or more amino acid comprises one or more of lysine, arginine, proline, glutamic acid, leucine, methionine, phenylalanine, tryptophan, tyrosine, valine, isoleucine, and histidine.

37. The engineered microorganism according to claim 35 or 36, wherein the one or more amino acid has a corresponding selection analog comprising one or more of S-(2-Aminoethyl)-L-cysteine, canavanine, azetidine-2-carboxylic acid, beta-N-Methylaminoalanine (BMAA), 5-hydroxyleucine, ethionine, selenomethionine, o-tyrosine, 7-Azatryptophan, 3,4-Dihydroxyphenylalanine (DOPA), 4-hydroxyvaline, O-Methylthreonine, and 2-Thiazolealanine.

38. A composition comprising an engineered *E. coli* according to any one of claims 1 to 13 and media.

39. The composition according to claim 38, further comprising supplements.

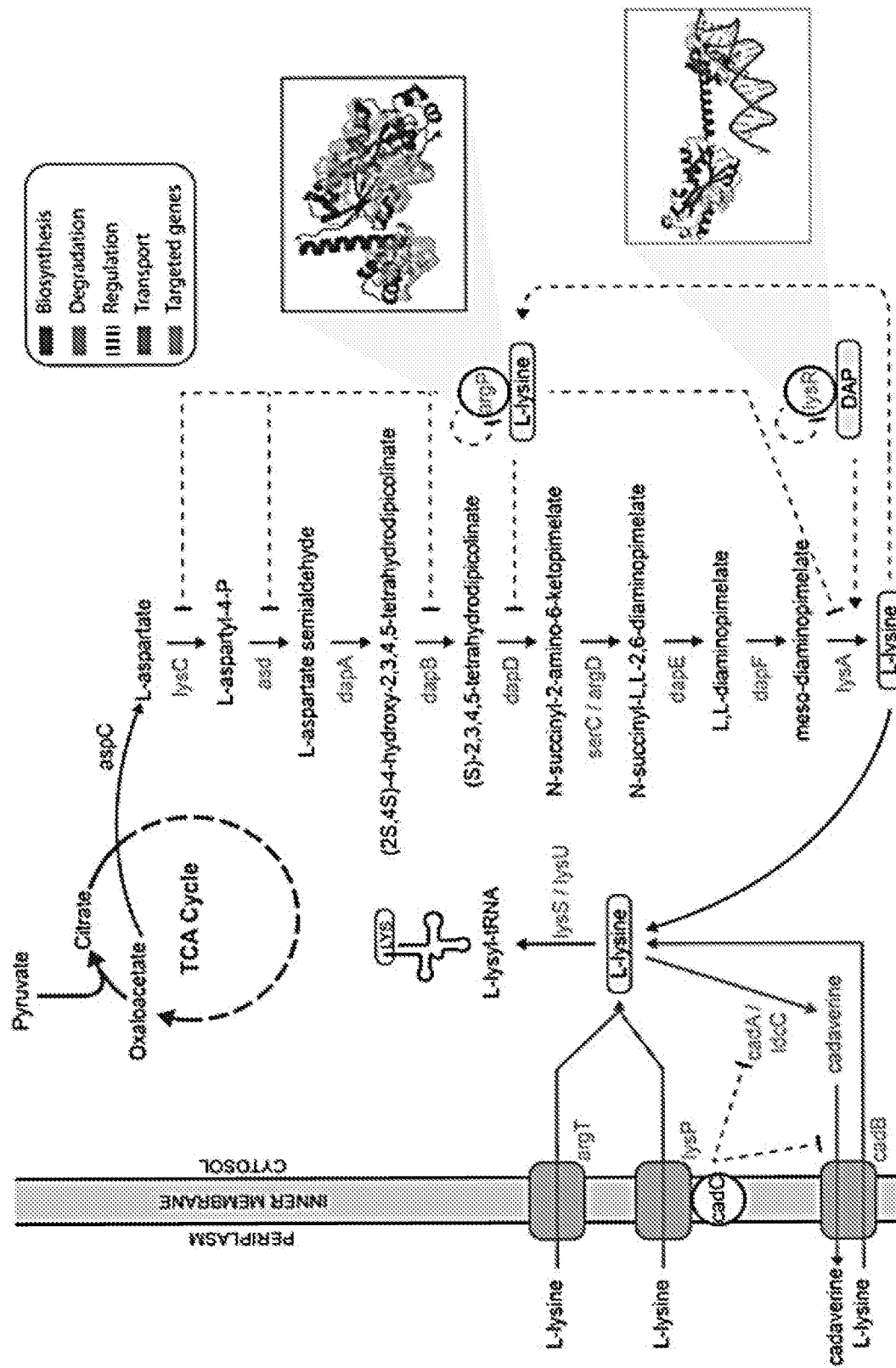


FIG. 1

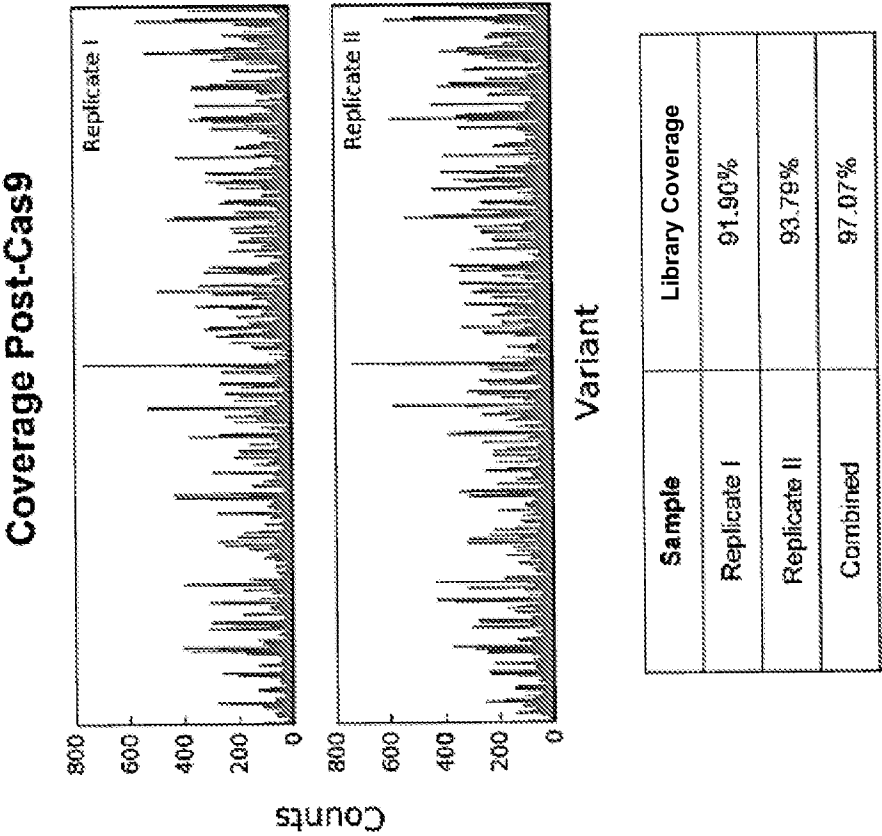


FIG. 2B

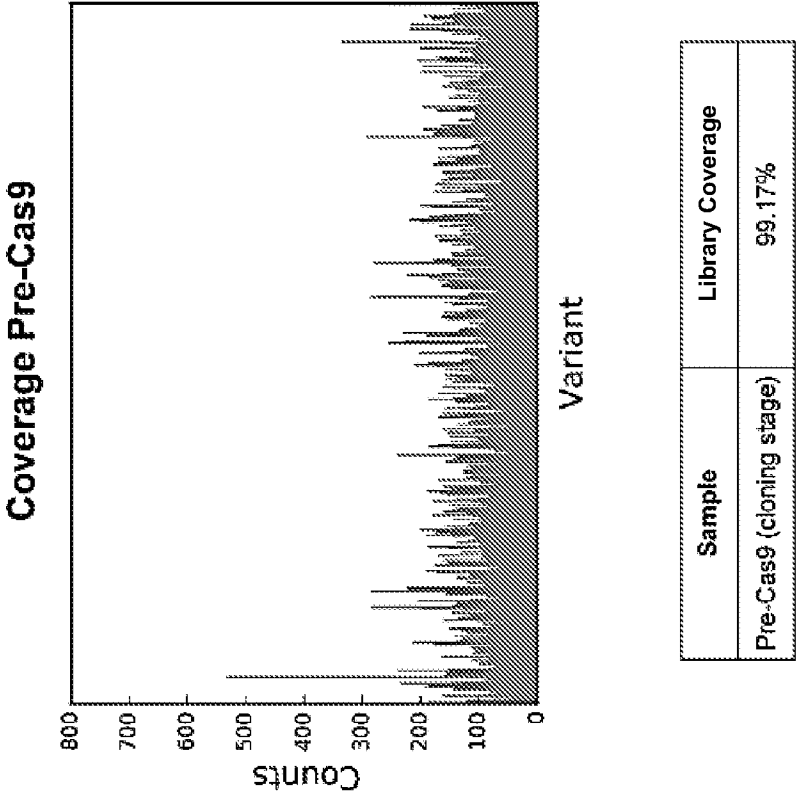


FIG. 2A

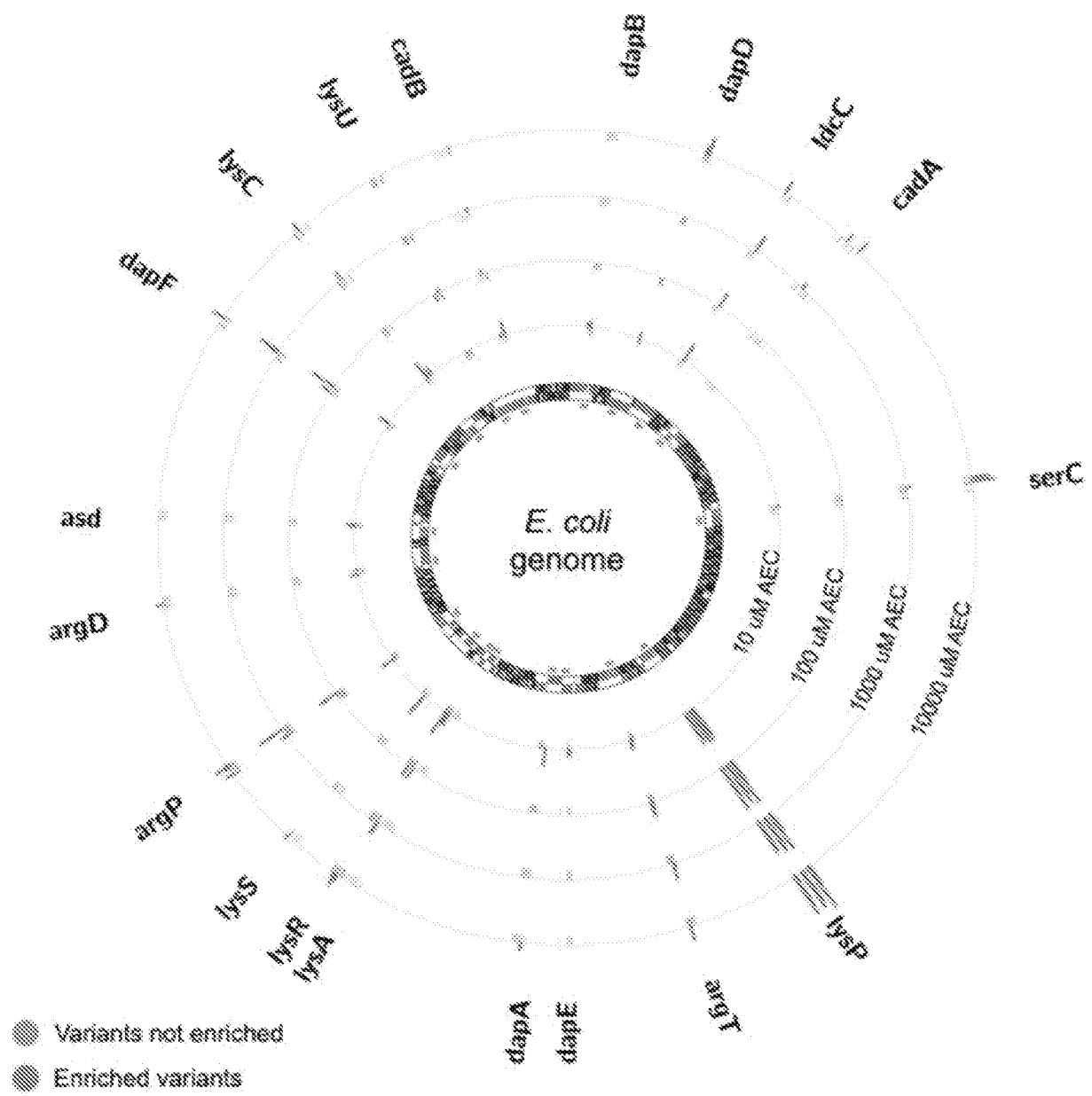


FIG. 3A

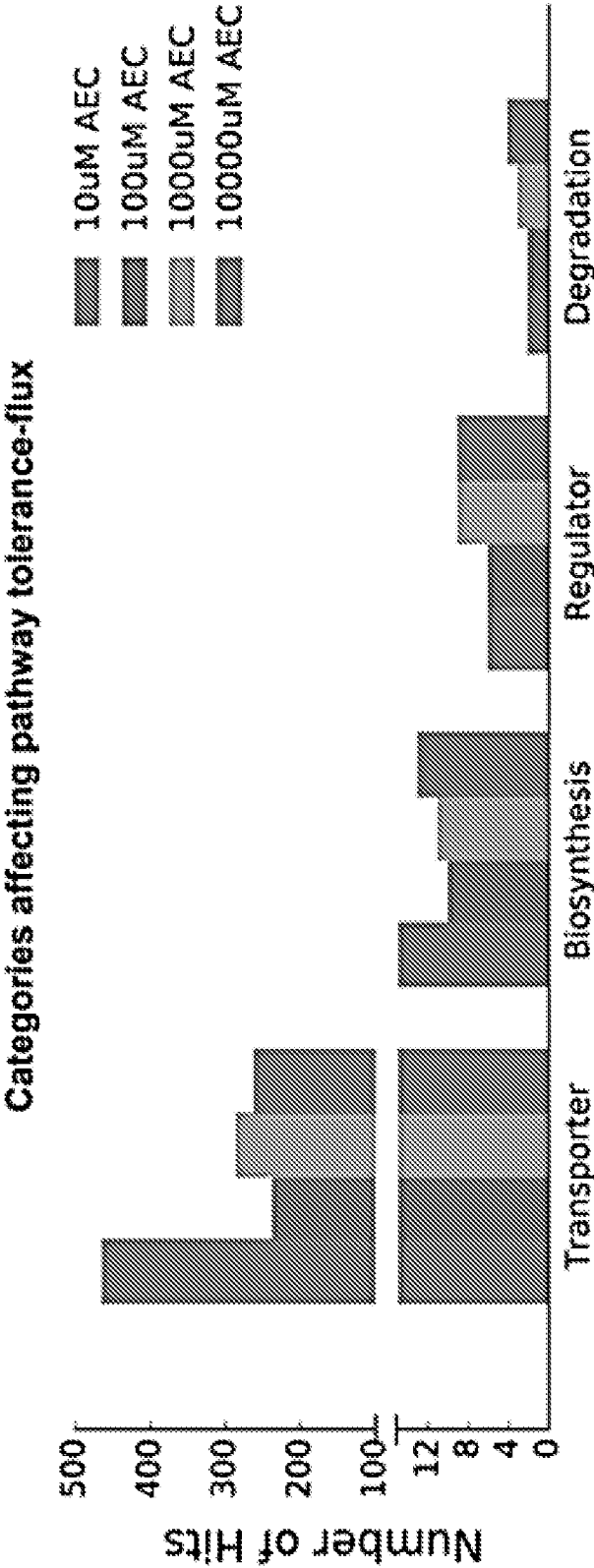


FIG. 3B

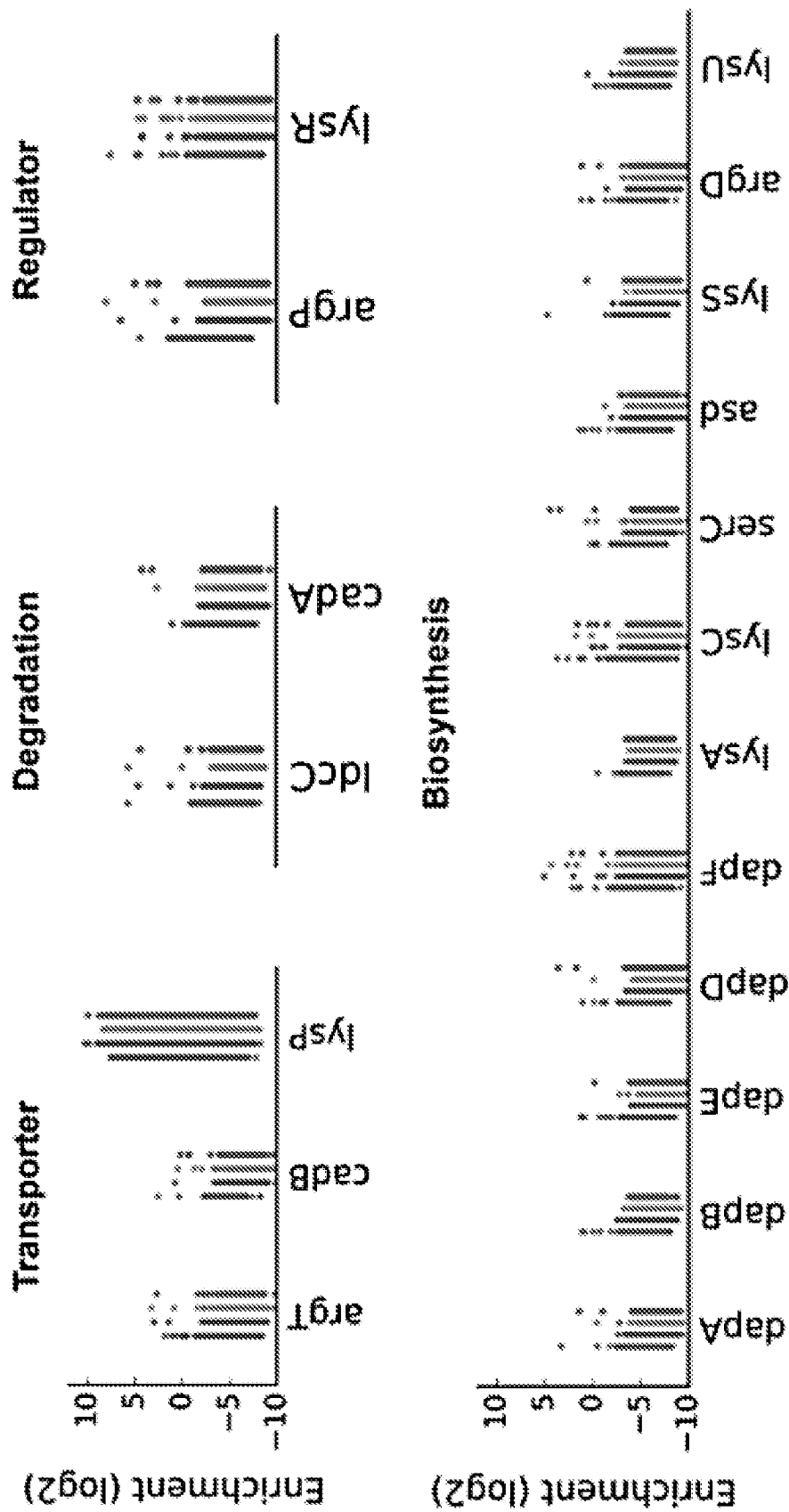


FIG. 3C

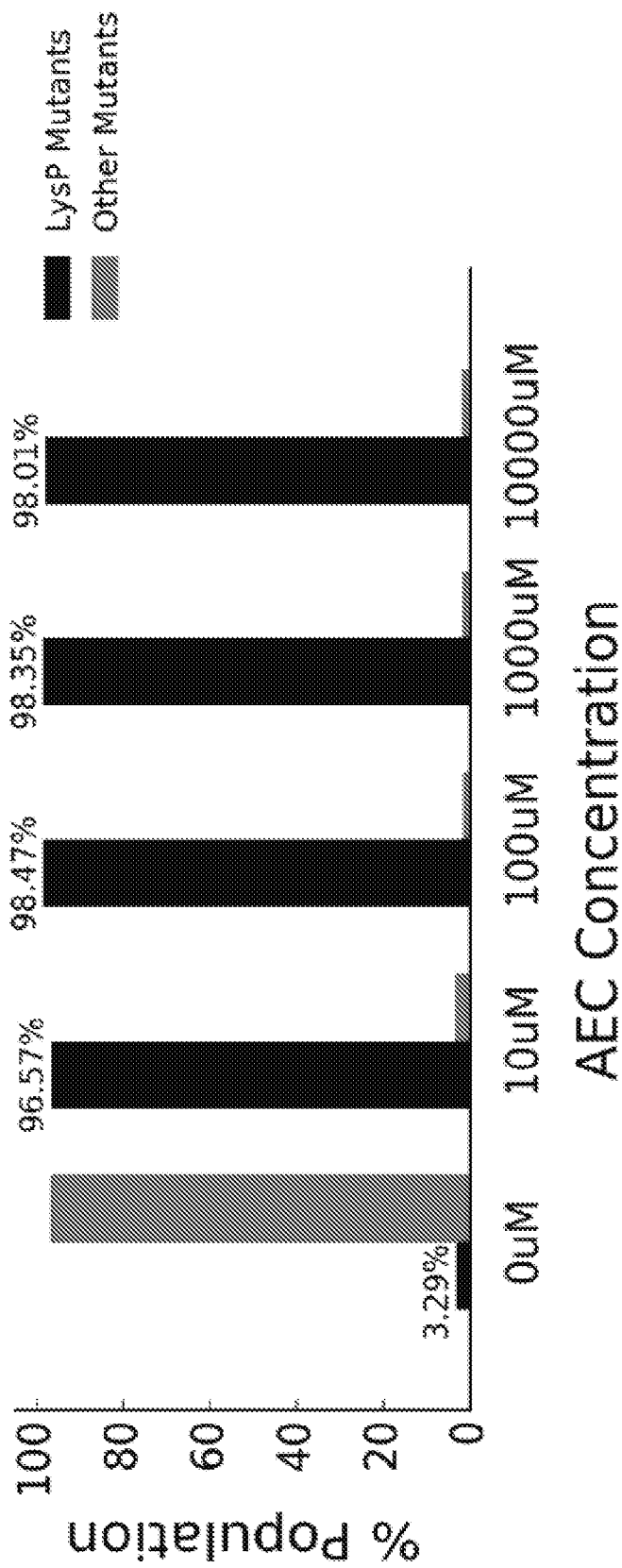


FIG. 4

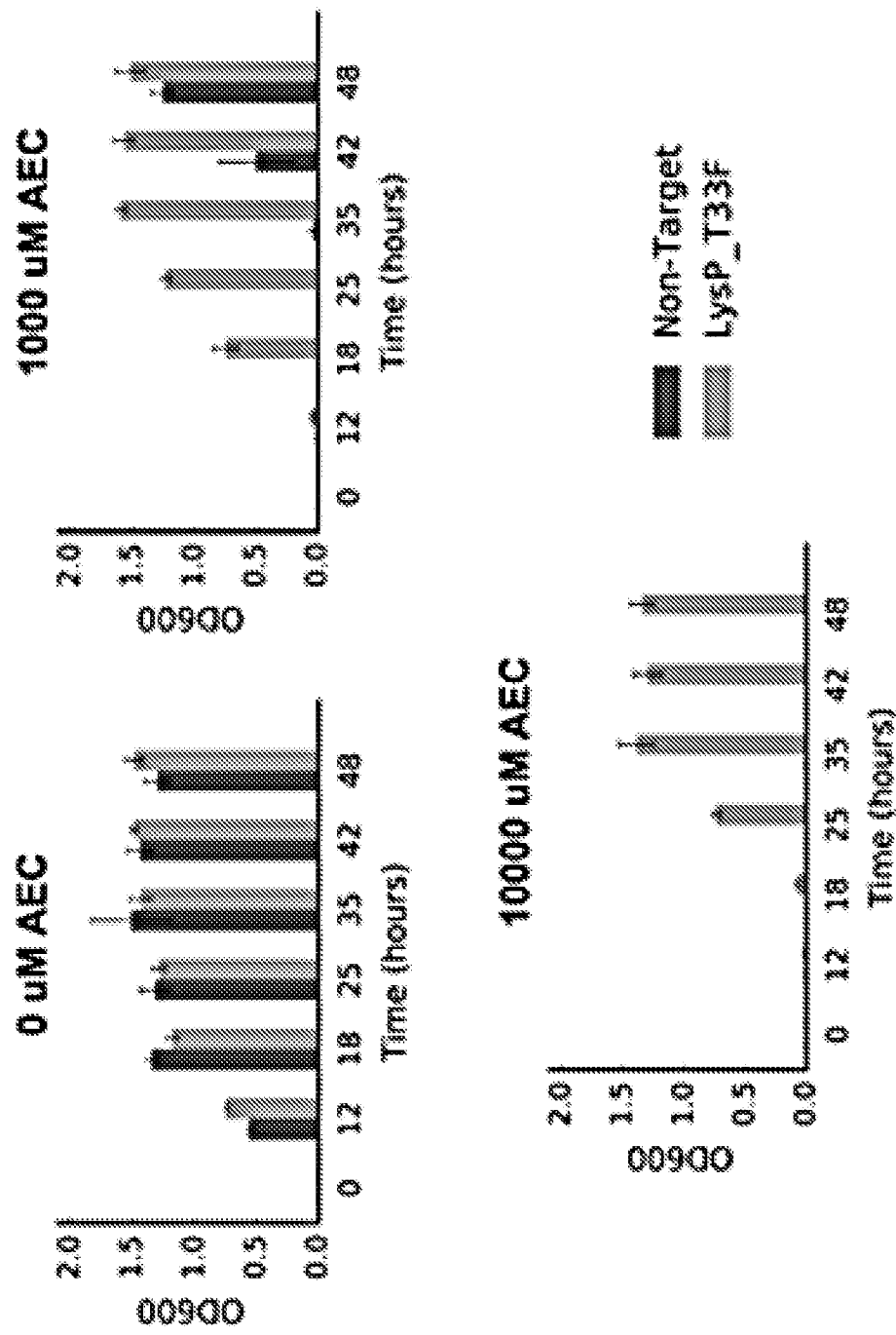


FIG. 5A

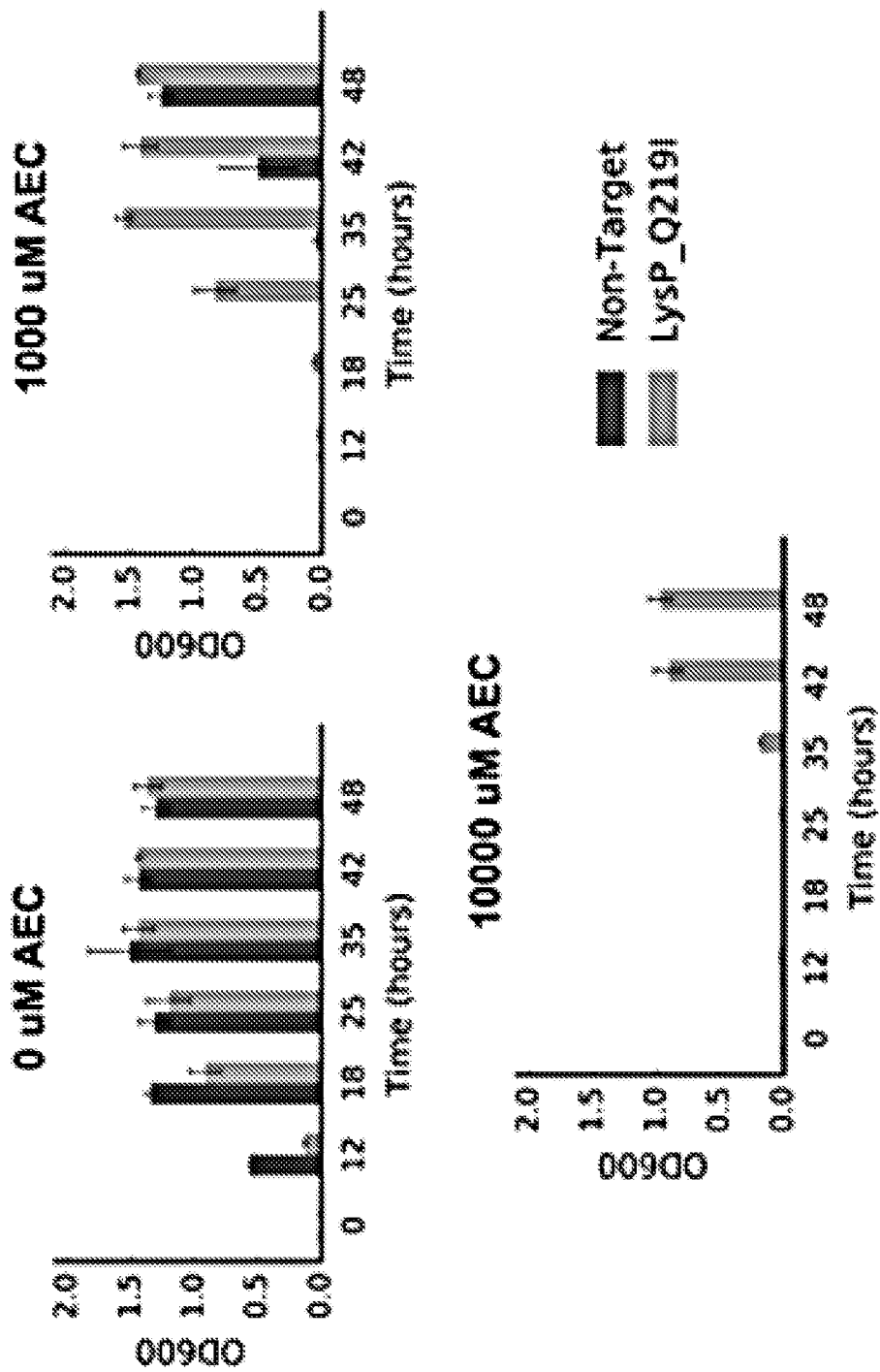


FIG. 5B

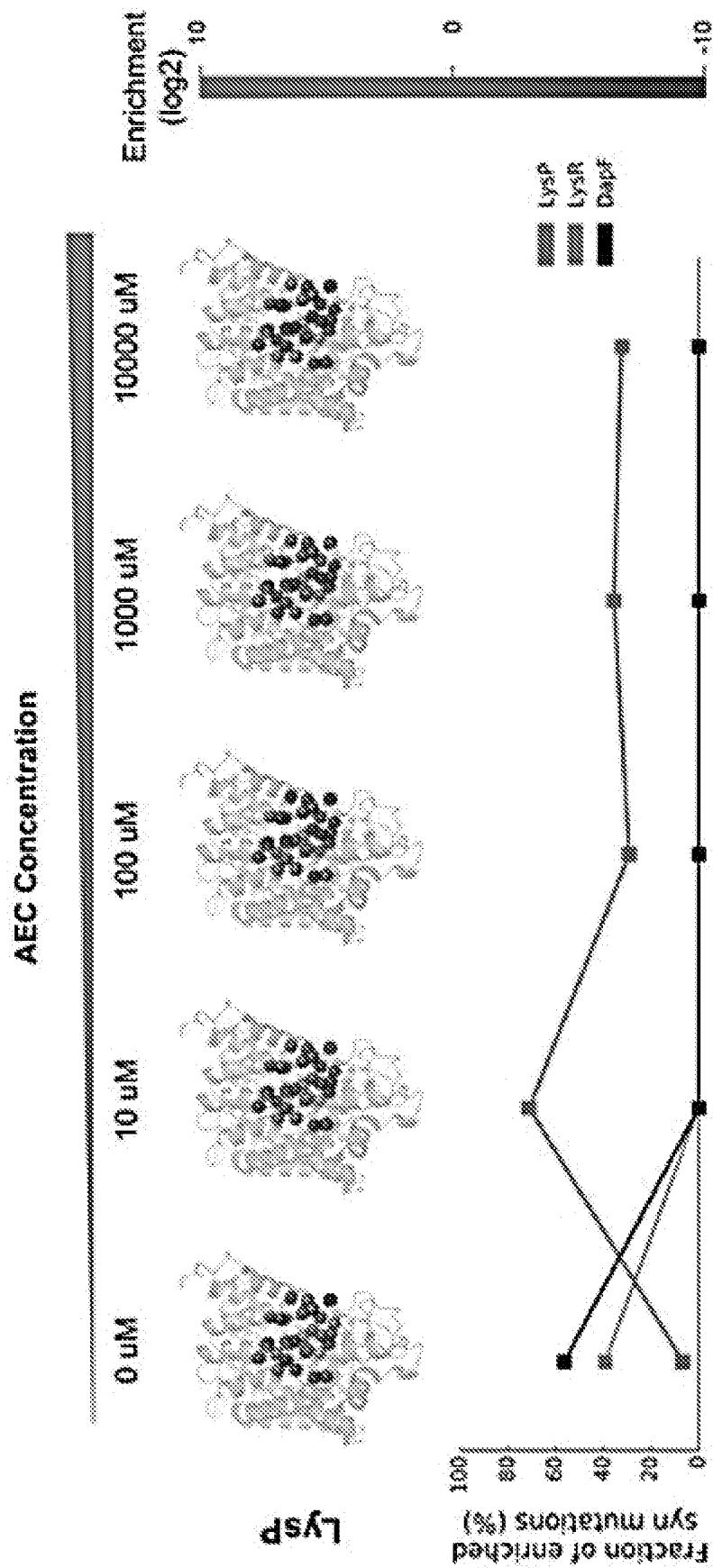


FIG. 6

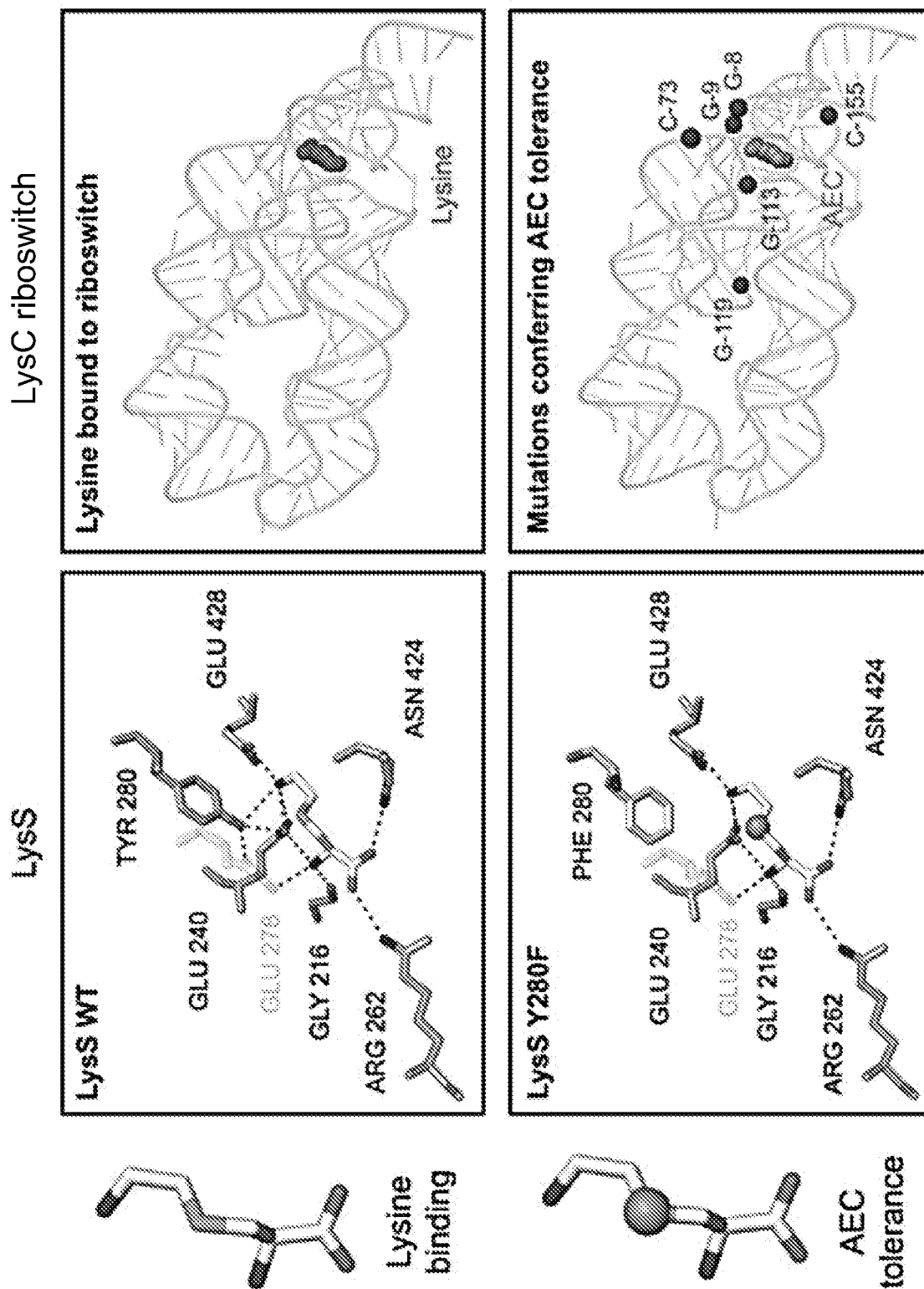


FIG. 7

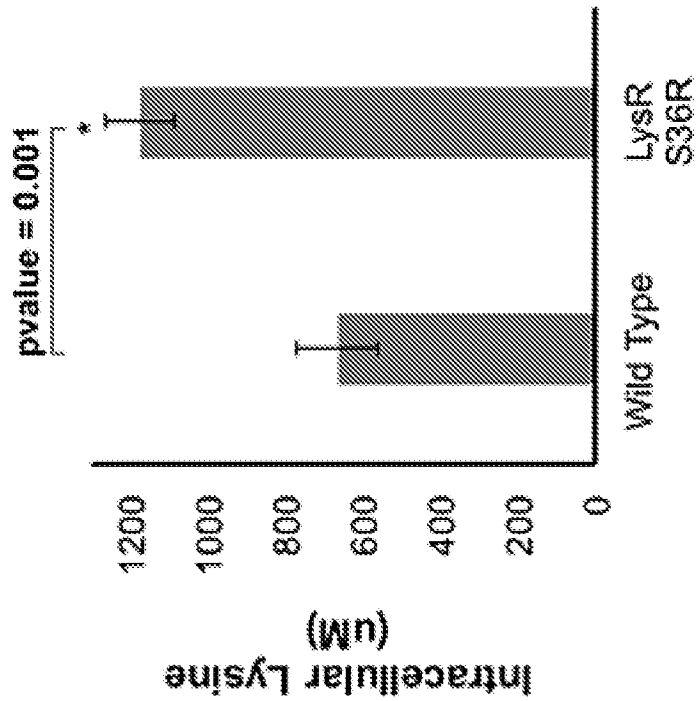


FIG. 8B

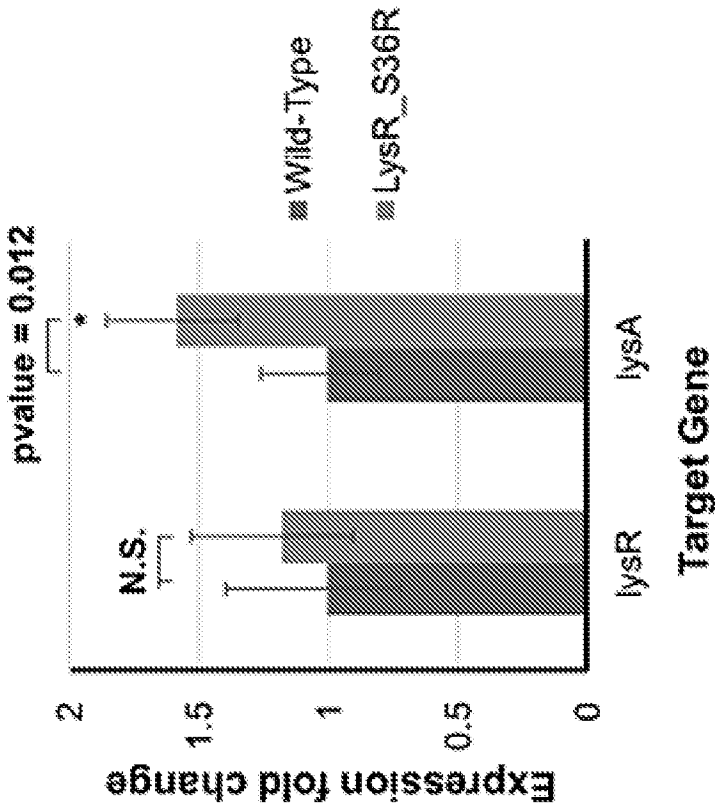


FIG. 8A

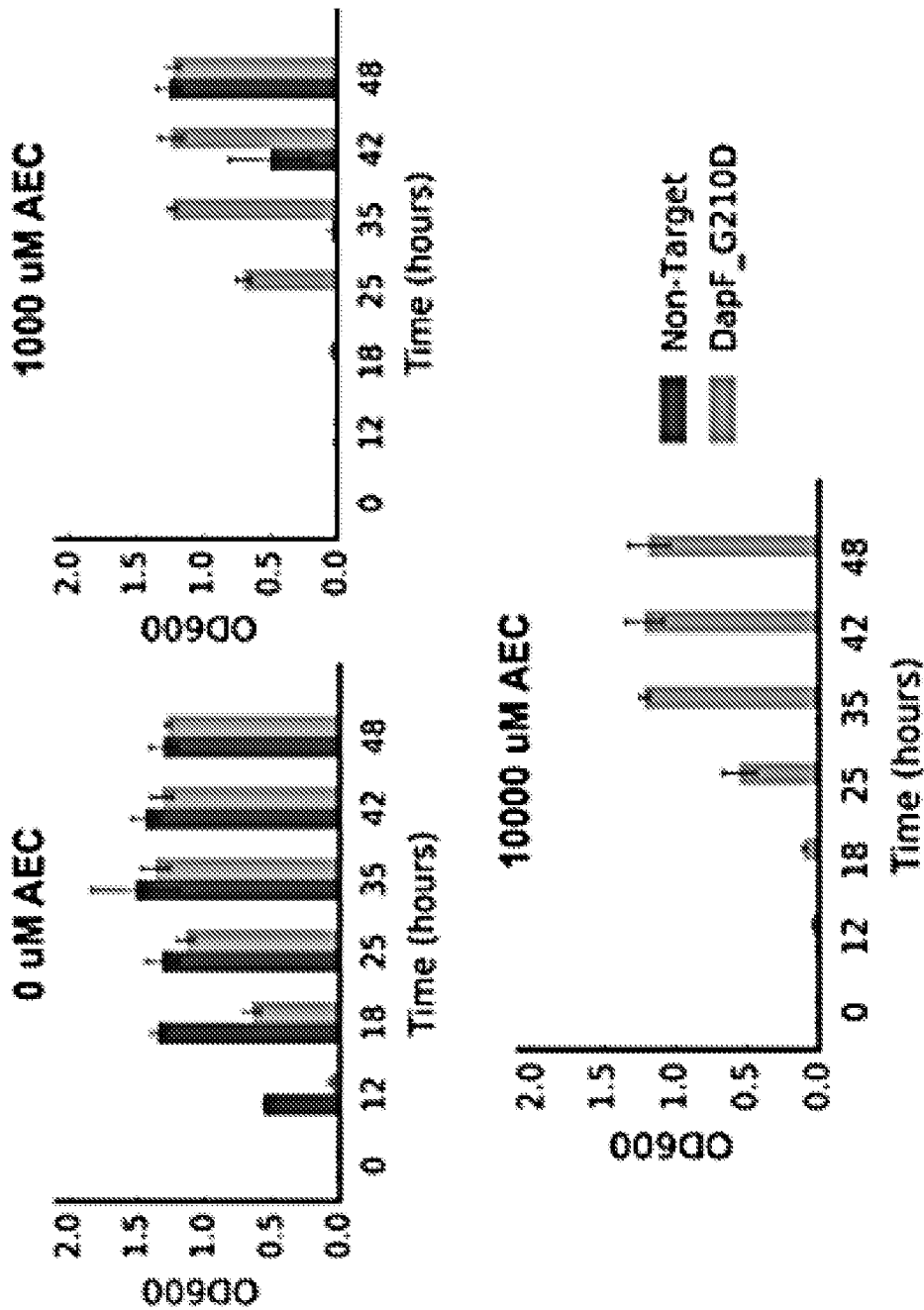


FIG. 9

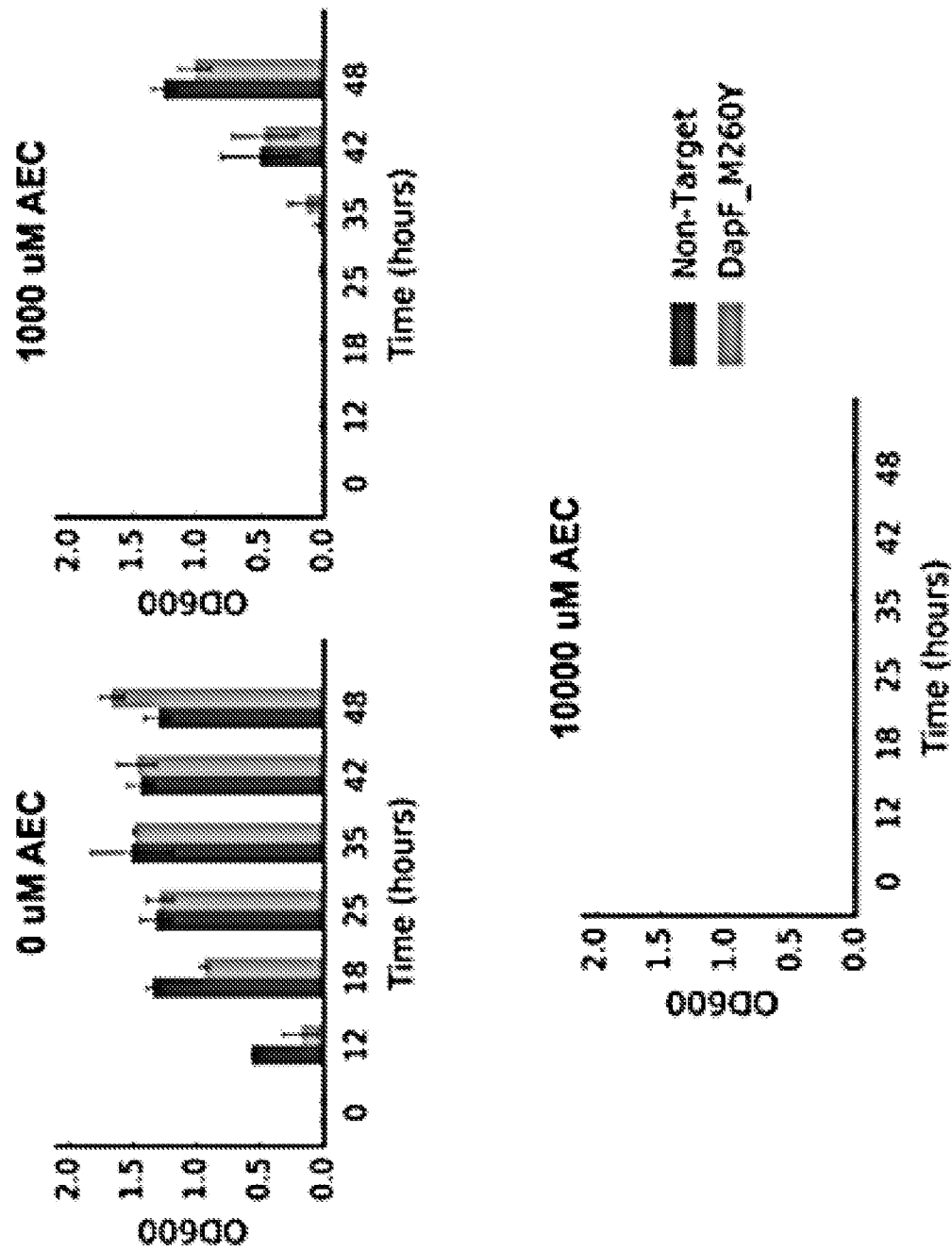


FIG. 10A

14/34

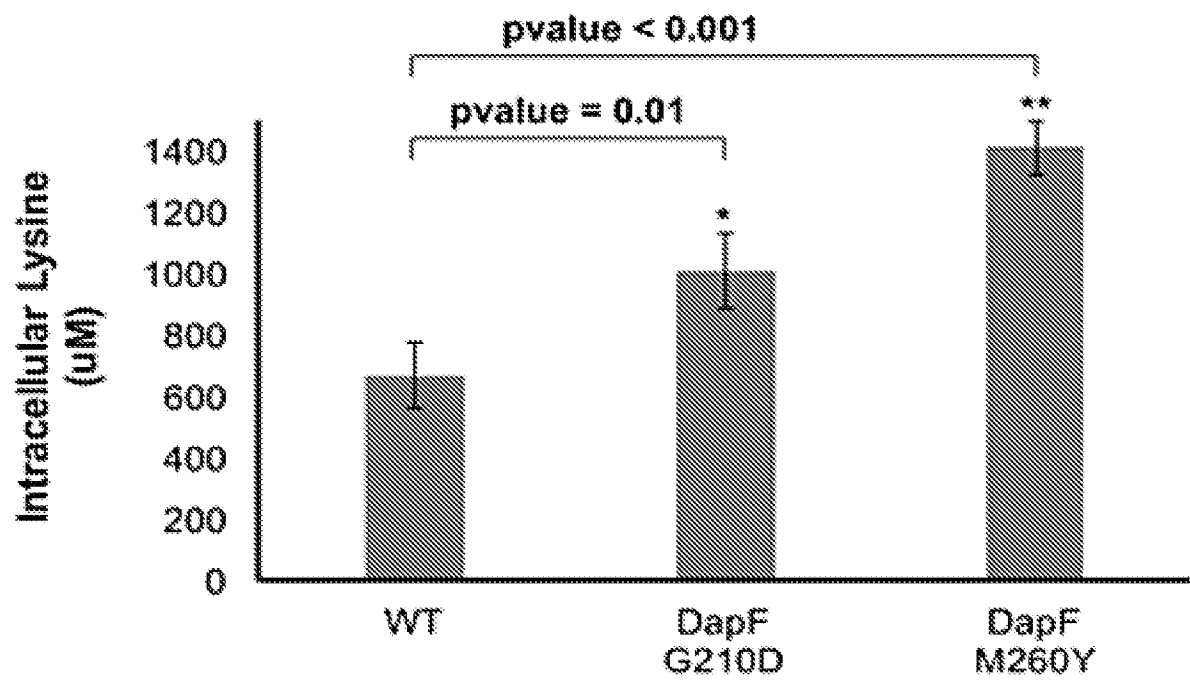


FIG. 10B

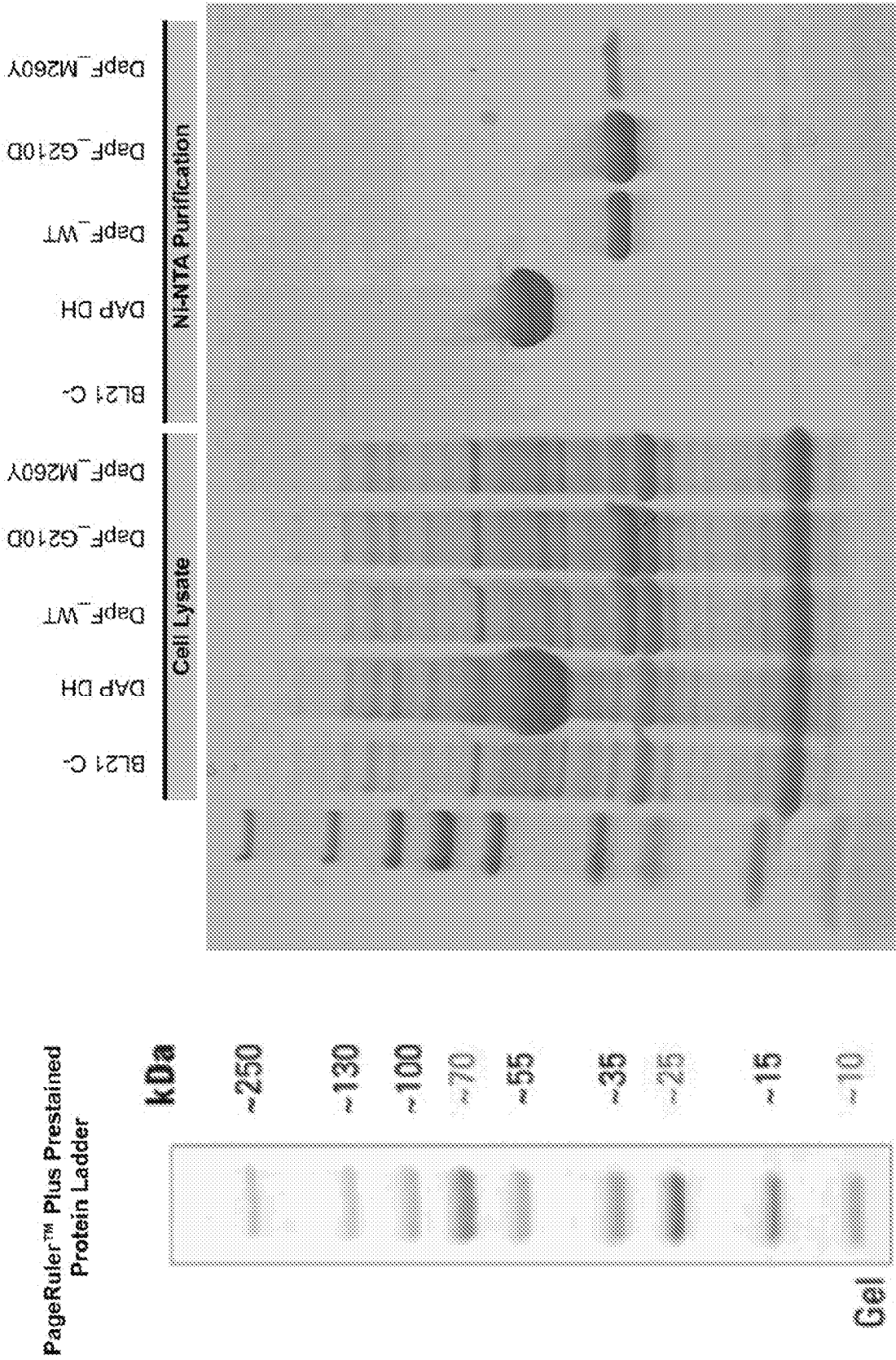


FIG. 11

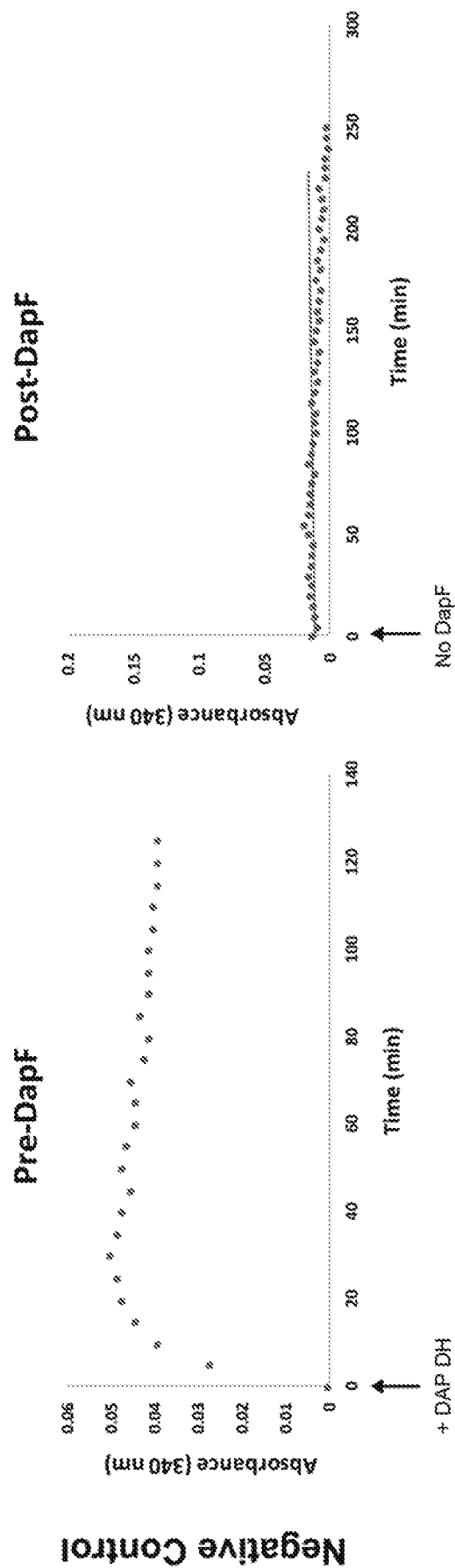


FIG. 12A

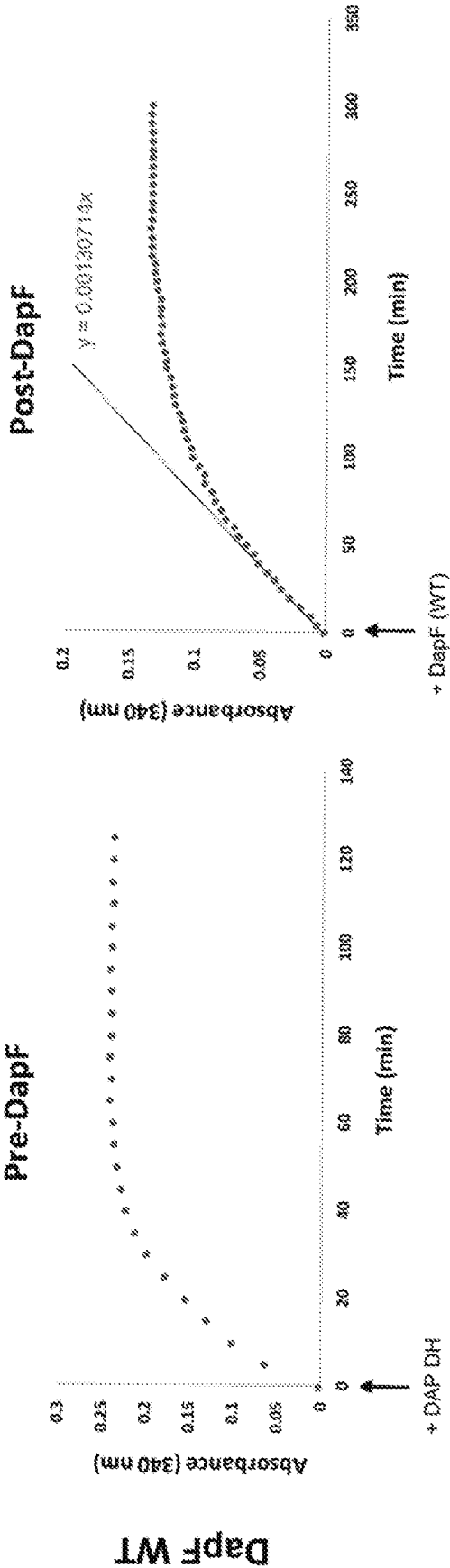


FIG. 12B

18/34

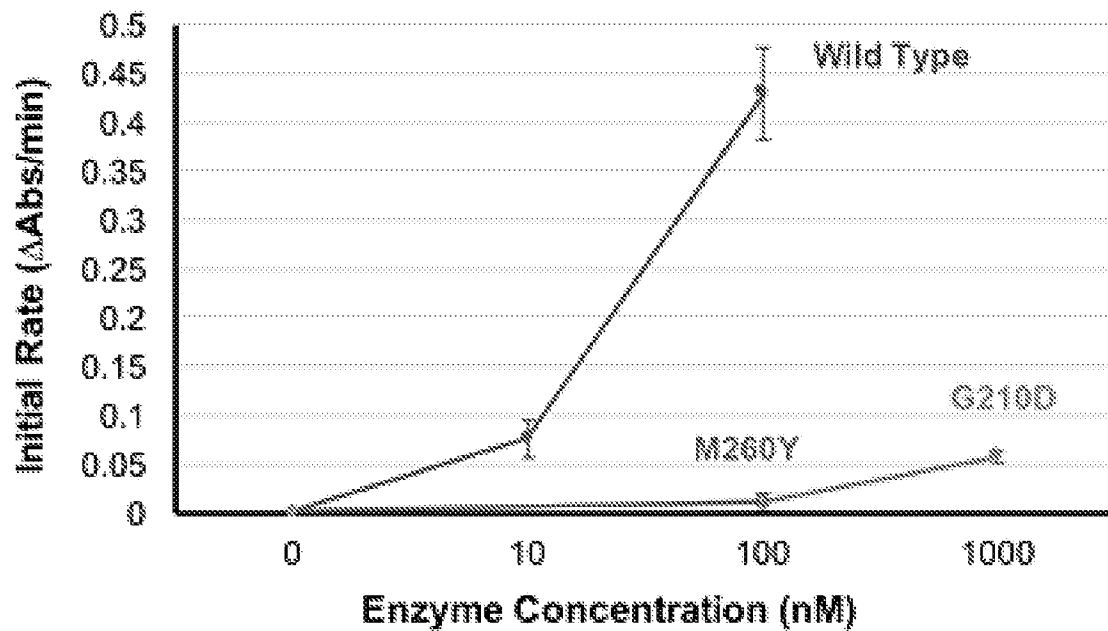


FIG. 13A

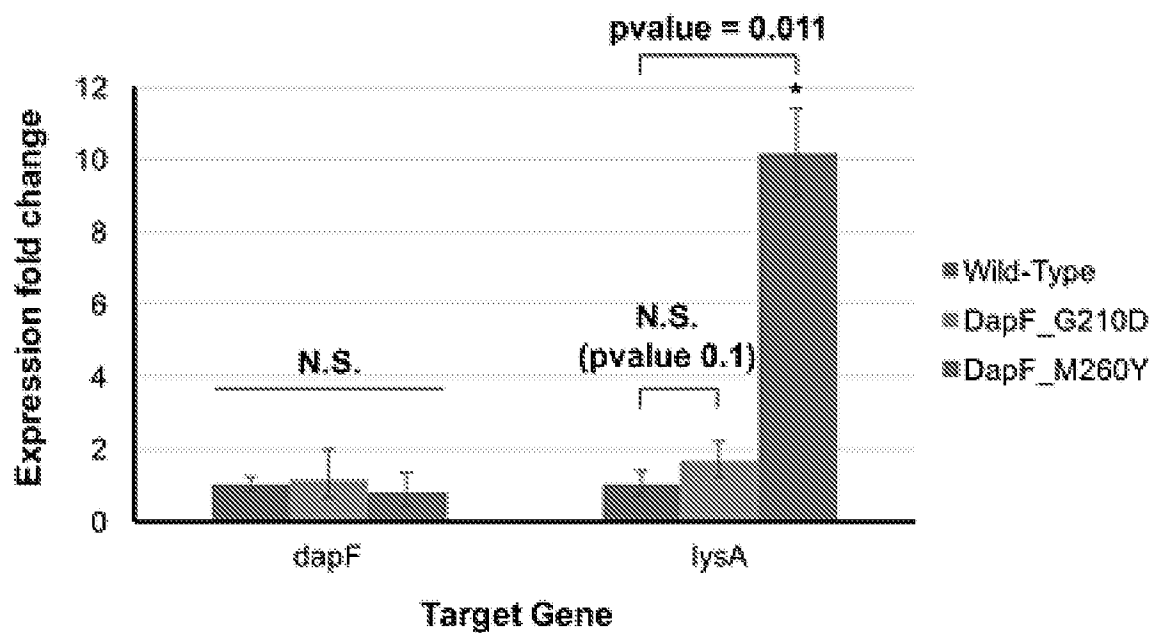


FIG. 13B

19/34

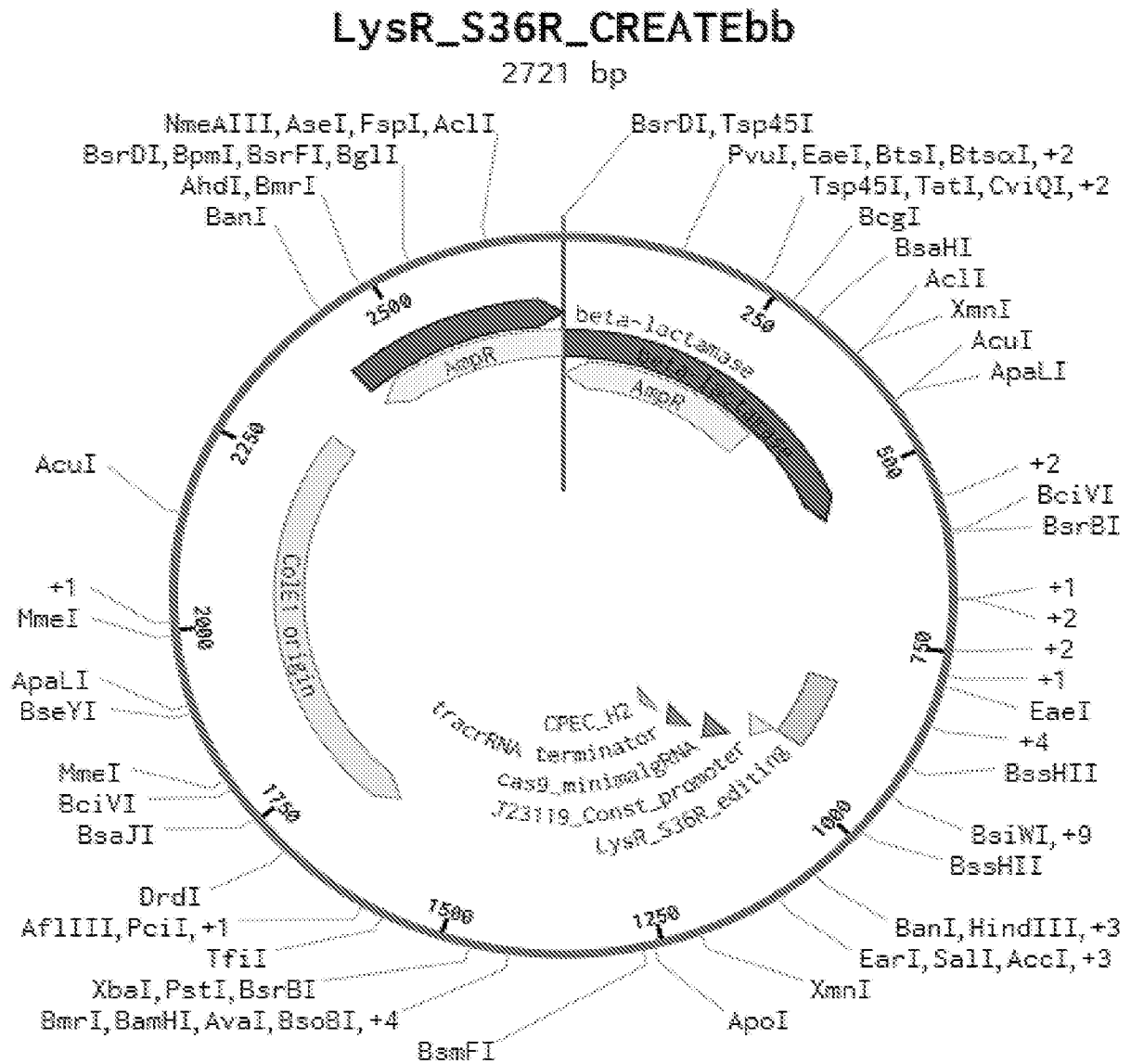


FIG. 14

20/34

Target Protein	Targeted Sites in the Library	Library Size
dapA	7, 8, 9, 40, 42, 43, 44, 45, 46, 47, 48, 76, 80, 81, 101, 106, 107, 108, 133, 135, 138, 161, 163, 185, 186, 187, 188, 203, 204, 205, 244, 248, 297, 298	680
dapB	12, 13, 14, 15, 16, 17, 18, 38, 39, 79, 80, 81, 83, 84, 88, 102, 103, 104, 126, 127, 128, 129, 130, 133, 163, 170, 217, 237, 239, 240, 242, 243	640
dapE	62, 63, 64, 65, 66, 67, 68, 97, 98, 99, 100, 101, 132, 133, 134, 135, 136, 161, 162, 163, 348	420
dapD	90, 92, 108, 110, 111, 126, 128, 129, 144, 145, 146, 147, 148, 170, 171, 172, 173, 174, 188, 189, 190, 191, 192, 204, 205, 206, 207, 213, 214, 216	600
dapF	7, 8, 9, 10, 11, 12, 13, 43, 44, 62, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 118, 155, 156, 157, 158, 159, 190, 208, 209, 210, 211, 214, 215, 216, 217, 218, 219, 220, 221, 222, 242, 247, 260, 263	880
argP	12, 14, 15, 16, 17, 18, 19, 20, 23, 32, 43, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 113, 121, 124, 125, 126, 127, 128, 129, 130, 134, 144, 146, 147, 159, 160, 161, 162, 163, 164, 193, 194, 195, 196, 197, 198, 200, 221, 222, 223, 224, 225, 226, 227, 232, 239, 241, 243, 244, 246, 247, 248, 250, 265, 266, 267, 285, 286, 287, 288, 289, 290, 291, 292, 293	1880
lysA	52, 54, 55, 75, 142, 157, 164, 191, 192, 226, 227, 232, 268, 270, 271, 305, 306, 307, 308, 311, 313, 315, 377, 378, 379, 381, 382	540
lysC	8, 10, 11, 12, 39, 40, 45, 119, 184, 197, 198, 199, 200, 201, 202, 220, 221, 222, 223, 225, 226, 227, 228, 229, 230, 231, 232, 251, 255, 256, 257, 258, 318, 320, 321, 322, 323, 324, 325, 326, 337, 338, 339, 340, 344, 345, 346, 347, 348, 354	1000
argT	32, 33, 34, 35, 36, 37, 39, 40, 52, 72, 73, 74, 75, 77, 78, 90, 91, 92, 93, 94, 95, 96, 99, 105, 110, 112, 113, 114, 139, 140, 141, 142, 143, 144, 145, 146, 163, 164, 165, 181, 182, 183, 184, 186, 212, 213, 214, 216, 217, 218	1000
ldcC	218, 219, 220, 221, 222, 224, 244, 245, 246, 247, 248, 304, 305, 306, 330, 331, 332, 333, 362, 363, 364, 366, 367, 375, 376	500
cadA	75, 97, 206, 218, 219, 220, 221, 222, 224, 245, 247, 304, 330, 332, 333, 364, 365, 366, 367, 368, 369, 417, 418, 558, 559, 562, 563, 564, 565, 568, 585, 716	640
cadB	16, 18, 19, 20, 21, 22, 23, 25, 32, 43, 46, 47, 51, 96, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 231, 235, 289, 349	560
serC	9, 10, 12, 41, 42, 74, 75, 76, 77, 78, 80, 102, 149, 151, 153, 154, 174, 176, 177, 195, 196, 197, 198, 199, 200, 201, 238, 239, 240, 328, 329, 335, 339, 363, 364	700
asd	8, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 167, 168, 169, 170, 171, 193, 194, 240, 241, 243, 244	660
lysP	29, 30, 31, 32, 33, 35, 36, 43, 53, 54, 55, 56, 57, 58, 61, 62, 65, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 245, 246, 249, 250, 253, 375, 379, 383, 386, 427, 431	780
lysR	4, 5, 6, 7, 8, 11, 15, 19, 20, 21, 22, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 89, 101, 102, 103, 104, 105, 106, 130, 131, 132, 133, 136, 148, 149, 150, 151, 152, 153, 154, 155, 160, 162, 166, 168, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 223, 224, 225, 226, 227, 228, 230, 242, 243, 244, 245, 246, 247, 266, 267, 268, 269, 270, 272, 278	1940
lysS	215, 216, 217, 218, 237, 238, 239, 240, 241, 243, 244, 260, 262, 274, 276, 277, 278, 279, 280, 289, 399, 400, 412, 413, 414, 421, 423, 424, 425, 426, 427, 428, 429, 471, 472, 473, 474, 475, 476, 477, 505	820
argD	21, 51, 78, 79, 80, 106, 107, 108, 109, 110, 112, 141, 142, 143, 144, 193, 195, 197, 198, 226, 227, 228, 229, 252, 254, 255, 281, 282, 283, 284, 285, 377	640
lysU	215, 216, 217, 218, 219, 237, 238, 239, 240, 241, 243, 244, 260, 261, 262, 263, 264, 265, 266, 267, 268, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 289, 376, 380, 392, 399, 400, 412, 413, 414, 415, 416, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 452, 454, 460, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 483, 491, 495	1420
		Total Size: 16300

FIG. 15

Amino Acid Target for Production in Microorganism(s)	Analog (for selection/detection)
Lysine	S-(2-Aminoethyl)-L-cysteine
Arginine	Canavanine
Proline	Azetidine-2-carboxylic acid
Glutamic Acid	Beta-N-Methylaminoalanine (BMAA)
Leucine	5-hydroxyleucine
Methionine	Ethionine, Selenomethionine
Phenylalanine	o-tyrosine
Tryptophan	7-Azatryptophan
Tyrosine	3,4-Dihydroxyphenylalanine (DOPA)
Valine	4-hydroxyvaline
Isoleucine	O-Methylthreonine
Histidine	2-Thiazolealanine

FIG. 16

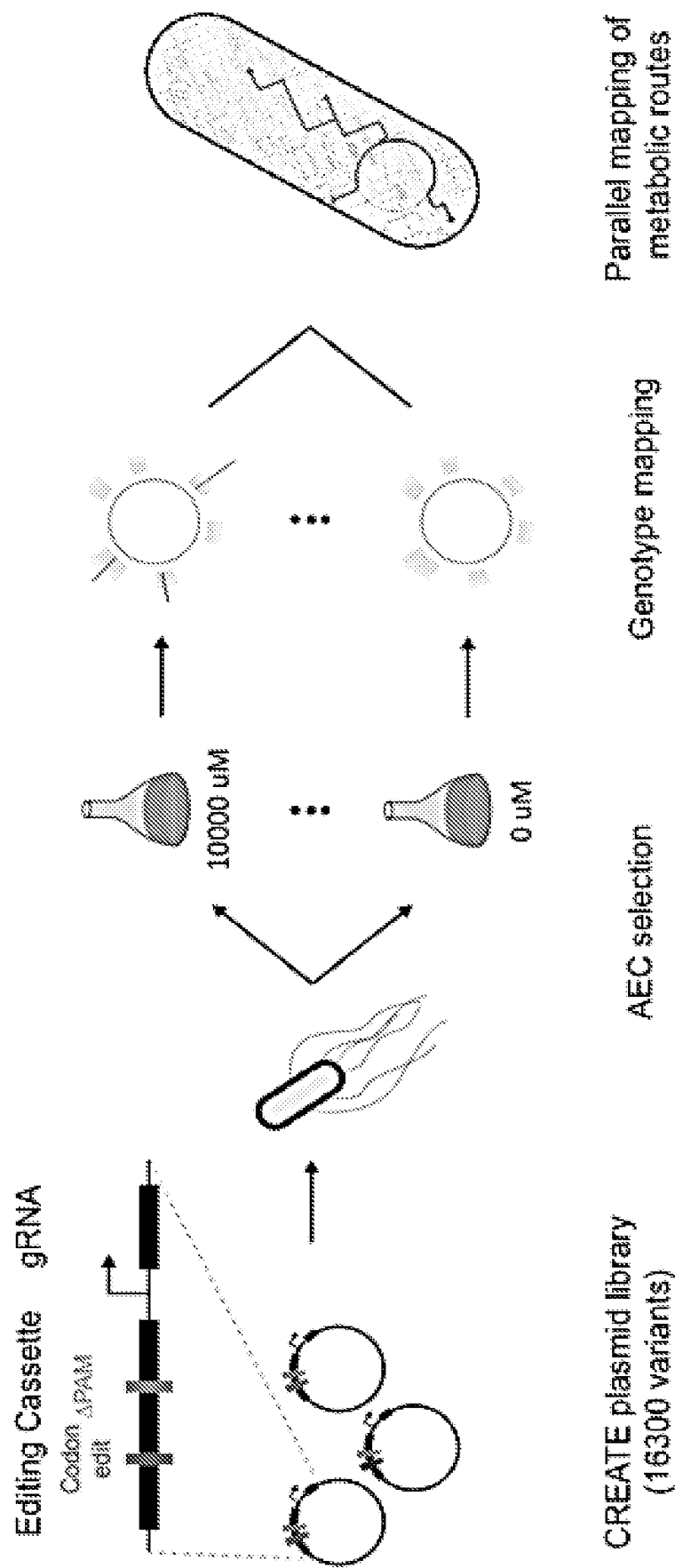


FIG. 17

FIG. 18A

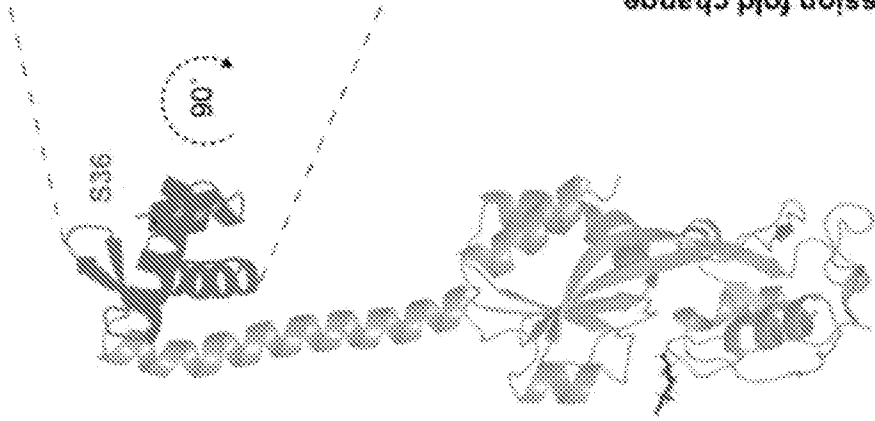


FIG. 18B

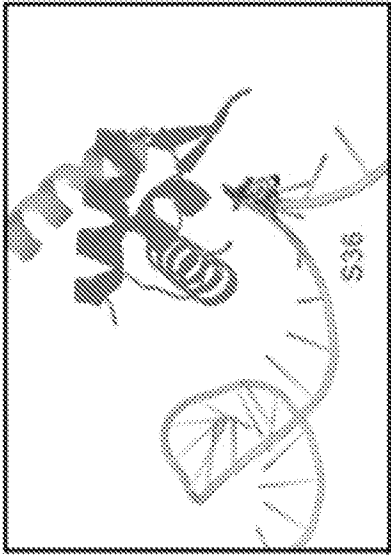


FIG. 18C

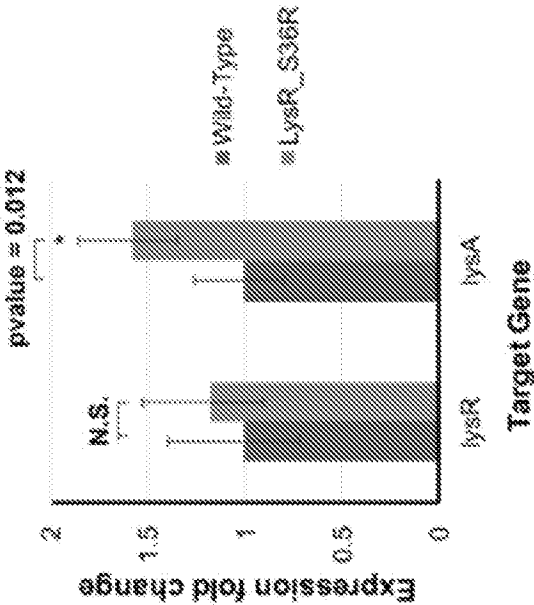
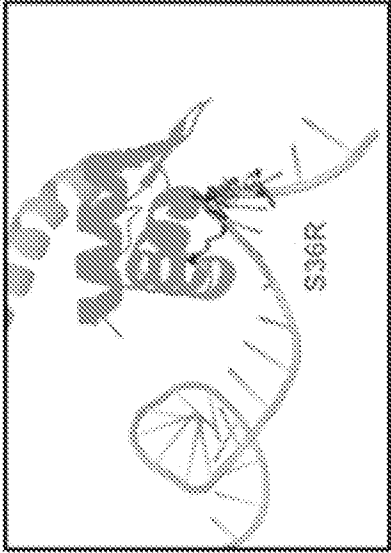


FIG. 18D

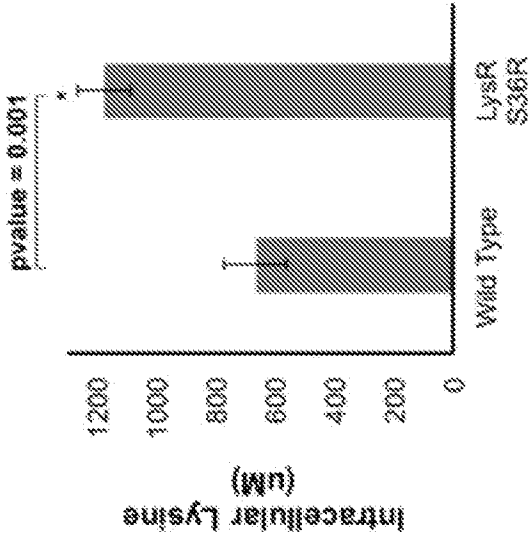


FIG. 18E

Target Protein	Total Size (Amino Acids)	Protein Structure	Targeted Sites in the Library	Library Size	Functional Sites	Functional Sites Covered in Library (%)
dapA	292	PDB (4EOU; 3I7S)	7, 8, 9, 40, 42, 43, 44, 45, 46, 47, 48, 76, 80, 81, 101, 106, 107, 108, 133, 135, 138, 161, 163, 185, 186, 187, 188, 203, 204, 205, 244, 248, 297, 298	680	44, 45, 49, 80, 84, 106, 107, 133, 161, 203	80%
dapB	273	PDB (1DRU; 1DRW; 1ARZ)	12, 13, 14, 15, 16, 17, 18, 38, 39, 79, 80, 81, 83, 84, 88, 102, 103, 104, 126, 127, 128, 129, 130, 133, 163, 170, 217, 237, 239, 240, 242, 243	640	12-17, 38, 39, 102-104, 126-129, 159, 160, 163, 169, 170	85%
dapE	375	SWISS-MODEL	62, 63, 64, 65, 66, 67, 68, 97, 98, 99, 100, 101, 132, 133, 134, 135, 136, 161, 162, 163, 348	420	140, 141	0%
dapD	274	PDB (38XY; 3TDT)	90, 92, 108, 110, 111, 126, 128, 129, 144, 145, 146, 147, 148, 170, 171, 172, 173, 174, 188, 189, 190, 191, 192, 204, 205, 206, 207, 213, 214, 216	600	66, 68, 99, 133, 134, 162, 348	100%
dapF	274	PDB (4UJZ; 4IKO)	7, 8, 9, 10, 11, 12, 13, 43, 44, 62, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 118, 155, 156, 157, 158, 159, 190, 208, 209, 210, 211, 214, 215, 216, 217, 218, 219, 220, 221, 222, 242, 247, 260, 263	880	11, 44, 64, 73, 74, 75, 157, 159, 190, 208, 209, 217, 218, 219, 268	86.66%
argP	297	SWISS-MODEL	12, 14, 15, 16, 17, 18, 19, 20, 23, 32, 43, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 113, 121, 124, 125, 126, 127, 128, 129, 130, 134, 144, 146, 147, 159, 160, 161, 162, 163, 164, 193, 194, 195, 196, 197, 198, 200, 221, 222, 223, 224, 225, 226, 227, 232, 239, 241, 243, 244, 246, 247, 248, 250, 265, 266, 267, 285, 286, 287, 288, 289, 290, 291, 292, 293	1880	-	-
lysA	420	PDB (1KNW; 1KOO)	52, 54, 55, 75, 142, 157, 164, 191, 192, 226, 227, 232, 268, 270, 271, 305, 306, 307, 308, 311, 313, 315, 377, 378, 379, 381, 382	540	191, 227, 268-271, 307, 311, 342, 343, 378	72.72%
lysC	449	PDB (2JOW; 2JOX)	8, 10, 11, 12, 39, 40, 45, 119, 184, 197, 198, 199, 200, 201, 202, 220, 221, 222, 223, 225, 226, 227, 228, 229, 230, 231, 232, 251, 255, 256, 257, 258, 318, 320, 321, 322, 323, 324, 325, 326, 337, 338, 339, 340, 344, 345, 346, 347, 348, 354	1000	8-11, 45, 119, 198-201, 221, 222, 227, 232, 257, 258, 318-321, 324, 325, 338-340, 345, 346	92.59%

FIG. 19A

Target Protein	Total Size (Amino Acids)	Protein Structure	Targeted Sites in the Library	Library Size	Functional Sites	Functional Sites Covered in Library (%)
argT	260	SWISS-MODEL	32, 33, 34, 35, 36, 37, 39, 40, 52, 72, 73, 74, 75, 77, 78, 90, 91, 92, 93, 94, 95, 96, 99, 105, 110, 112, 113, 114, 139, 140, 141, 142, 143, 144, 145, 146, 163, 164, 165, 181, 182, 183, 184, 186, 212, 213, 214, 216, 217, 218	1000	-	-
ldcC	713	PDB (5FKZ); SWISS-MODEL	218, 219, 220, 221, 222, 224, 244, 245, 246, 247, 248, 304, 305, 306, 330, 331, 332, 333, 362, 363, 364, 366, 367, 375, 376	500	-	-
cadA	715	PDB (3N75)	75, 97, 206, 218, 219, 220, 221, 222, 224, 245, 247, 304, 330, 332, 333, 364, 365, 366, 367, 368, 369, 417, 418, 558, 559, 562, 563, 564, 565, 568, 585, 716	640	-	-
cadB	444	SWISS-MODEL	16, 18, 19, 20, 21, 22, 23, 25, 32, 43, 46, 47, 51, 96, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 231, 235, 289, 349	560	43, 57, 73, 89, 90, 107, 204, 235, 303, 366, 368, 423	16.66%
serC	362	PDB (1BJO)	9, 10, 12, 41, 42, 74, 75, 76, 77, 78, 80, 102, 149, 151, 153, 154, 174, 176, 177, 195, 196, 197, 198, 199, 200, 201, 238, 239, 240, 328, 329, 335, 339, 363, 364	700	9, 42, 76, 77, 102, 153, 174, 197, 239, 240	100%
asd	367	PDB (1T4D)	8, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 167, 168, 169, 170, 171, 193, 194, 240, 241, 243, 244	660	10-13, 37, 38, 73, 102, 135, 162, 165-169, 173, 193, 241, 244, 267, 274, 350	36.36%
lysP	489	SWISS-MODEL; I-TASSER	29, 30, 31, 32, 33, 35, 36, 43, 53, 54, 55, 56, 57, 58, 61, 62, 65, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 245, 246, 249, 250, 253, 375, 379, 383, 386, 427, 431	780	-	-
lysR	311	SWISS-MODEL; I-TASSER	4, 5, 6, 7, 8, 11, 15, 19, 20, 21, 22, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 89, 101, 102, 103, 104, 105, 106, 130, 131, 132, 133, 136, 148, 149, 150, 151, 152, 153, 154, 155, 160, 162, 166, 168, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 223, 224, 225, 226, 227, 228, 230, 242, 243, 244, 245, 246, 247, 266, 267, 268, 269, 270, 272, 278	1940	-	-

FIG. 19B

Total Size			Library Size	Functional Sites		
Target Protein	(Amino Acids)	Protein Structure		Targeted Sites in the Library	Covered in Library (%)	
lysS	505	PDB (1BBU, 1BBW, 1KRS, 1KRT)	215, 216, 217, 218, 237, 238, 239, 240, 241, 243, 244,	820	415, 422	0%
			260, 262, 274, 276, 277, 278, 279, 280, 289, 399, 400,			
			412, 413, 414, 421, 423, 424, 425, 426, 427, 428, 429,			
			471, 472, 473, 474, 475, 476, 477, 505			
argD	406	SWISS-MODEL	21, 51, 78, 79, 80, 106, 107, 108, 109, 110, 112, 141,	640	108, 109, 141, 144, 226-229, 283, 284	100%
			142, 143, 144, 193, 195, 197, 198, 226, 227, 228, 229,			
			252, 254, 255, 281, 282, 283, 284, 285, 377			
lysU	505	PDB (1E1O, 1E1T, 1E22, 1E24, 1LYL)	215, 216, 217, 218, 219, 237, 238, 239, 240, 241, 243,	1420	415, 422	100%
			244, 260, 261, 262, 263, 264, 265, 266, 267, 268, 270,			
			271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 289,			
			376, 380, 392, 399, 400, 412, 413, 414, 415, 416, 419,			
			420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 452,			
			454, 460, 471, 472, 473, 474, 475, 476, 477, 478, 479,			
			480, 481, 483, 491, 495			
Total Size:				16300		

FIG. 19C

27/34

Strain ID	Mean Genome Coverage	Position	Reference	Mutation	QUAL	Depth	Frequency	Genomic Context
SEQ ID NO. 11								
	Parent_Strain	258673	A	[AGGCCCGTATATT]	217	13	100%	variation in mobile element inserted in RNA polymerase holoenzyme assembly factor (crl) gene
	Parent_Strain	258675	C	[G]	33.42	13	100%	variation in mobile element inserted in RNA polymerase holoenzyme assembly factor (crl) gene
SEQ ID NO. 12								
	Parent_Strain	258675	C	[CAGGCCCGTATATTCG]	225	13	100%	variation in mobile element inserted in RNA polymerase holoenzyme assembly factor (crl) gene
	Parent_Strain	1299502	A	[ATAATCAATTGTTAAA TT]	78	20	85%	variation in mobile element.
SEQ ID NO. 13								
	Parent_Strain	1299503	G	[GAATCAATTGTTAAAT TATT]	99	19	85%	variation in mobile element.
	Parent_Strain	1300691	T	[TTGGTAGAGGAAAG]	108	25	86%	variation in mobile element.
SEQ ID NO. 14								
	Parent_Strain	1300692	C	[CGGTAGAGGAAAGT]	117	24	100%	variation in mobile element.
	Parent_Strain	1300693	C	[CACTGCCCTTGGTAGA GGAAAGTG]	108	25	100%	variation in mobile element.
SEQ ID NO. 15								
	Parent_Strain	1979269	C	[CACAACCACTCAATT ACTTA, CTCAATTACTTA]	34.6591	17	100%	variation in mobile element.
	Parent_Strain	1979270	CCA	[CAAACTCAATTACT TAACA, CCAAACACTCAATTA CTTAACA]	97	20	100%	variation in mobile element.
SEQ ID NO. 16								
	Parent_Strain	2173358	CCAC	[CC]	84	54	95%	galactitol-specific PTS enzyme IIC component (gatC)
	Parent_Strain	2173360	ACCCC	[ACC]	225	53	100%	galactitol-specific PTS enzyme IIC component (gatC)
SEQ ID NO. 17								
	Parent_Strain	3560455	C	[CG]	225	64	100%	transcriptional repressor (glpR)
	Parent_Strain	4296380	AC	[ACGC]	225	74	100%	transcriptional repressor (glpR)

FIG. 20A

28/34

Strain ID	Mean Genome Coverage	Position	Reference	Mutation	QUAL	Depth	Frequency	Genomic Context
POST RE-ITERING OF PARENT STRAIN SNPs								
Strain IA	66.2X	354288	T	[G]	56	39	53%	propionyl-CoA synthetase (prpE)
Strain IA	66.2X	932321	C	[T]	225	63	100%	intergenic region between trxB and lrp genes.
Strain IA	66.2X	2386862	G	[A]	33.1368	84	55%	intergenic between yfbl and yfbM
Strain IA	66.2X	3460777	A	[C]	64	46	53%	Type II secretion system protein (GspF)
Strain IB	66.1X	932321	C	[T]	225	68	100%	intergenic region between trxB and lrp genes.
Strain IB	66.1X	2386862	G	[A]	56	55	67%	intergenic between yfbl and yfbM
Strain IB	66.1X	3665702	C	[A]	52	65	53%	DNA-binding transcriptional dual regulator (gadX)
Strain IC	77X	932321	C	[T]	225	65	100%	intergenic region between trxB and lrp genes.
Strain IC	77X	2386862	G	[A]	50	91	59%	intergenic between yfbl and yfbM
Strain IIA	58X	2247947	CG	[C]	225	61	100%	lysine-H ⁺ symporter (lysp)
Strain IIA	58X	2386862	G	[A]	61	61	52%	intergenic between yfbl and yfbM
Strain IIA	58X	3665702	C	[A]	56	70	54%	DNA-binding transcriptional dual regulator (gadX)
Strain IIA	58X	3813272	C	[T]	40.1609	39	54%	nfp
Strain IIB	59.2X	2247947	CG	[C]	225	65	100%	lysine-H ⁺ symporter (lysp)
Strain IIB	59.2X	4604344	C	[T]	50	69	57%	DNA-binding transcriptional regulator (bglI)
Strain IIB	59.2X	4604350	C	[T]	60	69	51%	DNA-binding transcriptional regulator (bglI)
Strain IIC	58.7X	2247947	CG	[C]	225	58	100%	lysine-H ⁺ symporter (lysp)

FIG. 20B

29/34

Strain IIC	58.7X	2386862	G	[A]	58	51	69%	intergenic between yfbl and yfcm DNA-binding transcriptional dual regulator (gadX)
Strain IIC	58.7X	3665702	C	[A]	95	83	53%	
Strain IIIA	51.5X	2247947	CG	[C]	225	63	100%	lysine:H ⁺ symporter (lysP)
Strain IIIA	51.5X	2386862	G	[A]	56	63	65%	intergenic between yfbl and yfcm
Strain IIIA	51.5X	2386883	C	[T]	77	65	54%	intergenic between yfbl and yfcm
Strain IIIA	51.5X	3665698	T	[C]	69	70	53%	DNA-binding transcriptional dual regulator (gadX)
Strain IIIB	67.9X	932321	C	[T]	225	61	100%	Intergenic region between trxB and lrp genes. aliphatic sulfonate ABC transporter ATP binding subunit (ssuB)
Strain IIIB	67.9X	993917	A	[C]	57	46	55%	
Strain IIIB	67.9X	2386862	G	[A]	54	69	53%	intergenic between yfbl and yfcm DNA-binding transcriptional dual regulator (gadX)
Strain IIIB	67.9X	3665702	C	[A]	114	100	54%	
Strain IIIC	59X	327222	C	[T]	222	62	64%	choline dehydrogenase (betA)
Strain IIIC	59X	932321	C	[T]	225	56	100%	Intergenic region between trxB and lrp genes.
Strain IIIC	59X	1051242	T	[A]	222	101	60%	upstream of stress protein (cspH)
Strain IIIC	59X	2386862	G	[A]	60	60	56%	intergenic between yfbl and yfcm

FIG. 20C

30/34

Strain ID	Mean Genome Coverage	Position	Reference	Mutation	QUAL	Depth	Frequency	Genomic Context
PARENTING OF PARENT STRAIN								
Strain 1A	95.5X	2117851	G	[T]	225	102	100%	putative colanic acid biosynthesis pyruvyl transferase (weak)
Strain 1A	95.5X	2247776	A	[G]	225	106	100%	lysine-H ⁺ symporter (ysp)
Strain 1A	95.5X	2388862	G	[A]	30.3177	55	59%	intergenic between yfbl and yfblM
Strain 1B	82.1X	1331661	T	[TAGAGCG]	221	81	81%	DNA topoisomerase 1 (topA)
Strain 1B	82.1X	1331662	CGAGCGTGA	[CGAGCGTGAGCG]	225	78	100%	DNA topoisomerase 1 (topA)
Strain 1B	82.1X	1331664	GCGTGA	[TGAGCGTGA]	228	78	97%	DNA topoisomerase 1 (topA)
Strain 1B	82.1X	2388862	G	[A]	33.3284	73	60%	intergenic between yfbl and yfblM
Strain 1B	82.1X	3204103	GC	[GCG]	225	79	100%	undecaprenyl pyrophosphate phosphatase (bacA)
Strain 1B	82.1X	3665698	T	[C]	104	91	51%	DNA-binding transcriptional dual regulator (gadX)
Strain 1B	82.1X	3665702	C	[A]	87	92	62%	DNA-binding transcriptional dual regulator (gadX)
Strain 1C	87.7X	1331265	T	[G]	225	69	100%	DNA topoisomerase 1 (topA)
Strain 1C	87.7X	2388862	G	[A]	55	64	67%	intergenic between yfbl and yfblM
Strain 1C	87.7X	3665702	C	[A]	66	128	50%	DNA-binding transcriptional dual regulator (gadX)
Strain 2A	94X	2247878	G	[A]	228	123	99%	lysine-H ⁺ symporter (ysp)
Strain 2A	94X	2251365	G	[T]	228	85	99%	endonuclease IV (nfo)
Strain 2A	94X	3665698	T	[C]	104	81	55%	DNA-binding transcriptional dual regulator (gadX)
Strain 2A	94X	3665702	C	[A]	113	81	58%	DNA-binding transcriptional dual regulator (gadX)
Strain 2B	70.1X	993917	A	[C]	54	65	51%	aliphatic sulfonate ABC transporter ATP binding subunit (ssuB)
Strain 2B	70.1X	1331412	A	[T]	225	70	100%	DNA topoisomerase 1 (topA)
Strain 2B	70.1X	2388862	G	[A]	39.3011	58	60%	intergenic between yfbl and yfblM

FIG. 20D

31/34

Strain 2B	70.1X	3665702	C	[A]	64	80	54%	DNA-binding transcriptional dual regulator igadX
Strain 2C	95.9X	2117851	G	[T]	222	103	51%	putative colanic acid biosynthesis pyruvyl transferase (weaK)
Strain 2C	95.9X	2347509	CA	[C]	222	112	63%	lysine H ⁺ symporter (lysP)
Strain 2C	95.9X	2348396	G	[C]	210	83	54%	lysine H ⁺ symporter (lysP)
Strain 2C	95.9X	2386852	G	[A]	49,1981	79	54%	intergenic between yfbl and yfblM

FIG. 20E

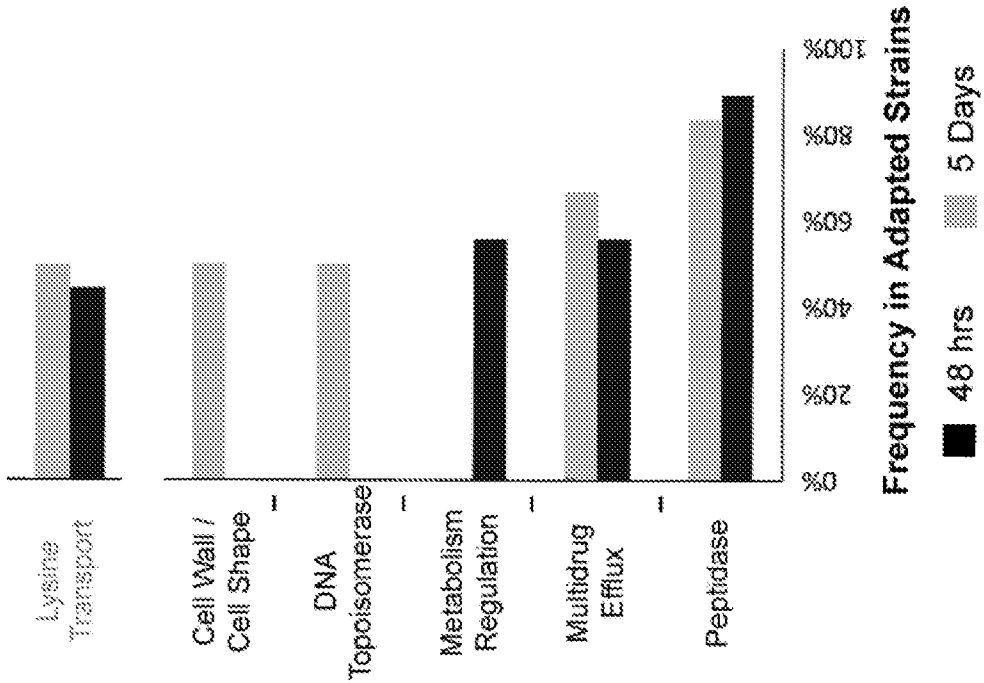


FIG. 21B

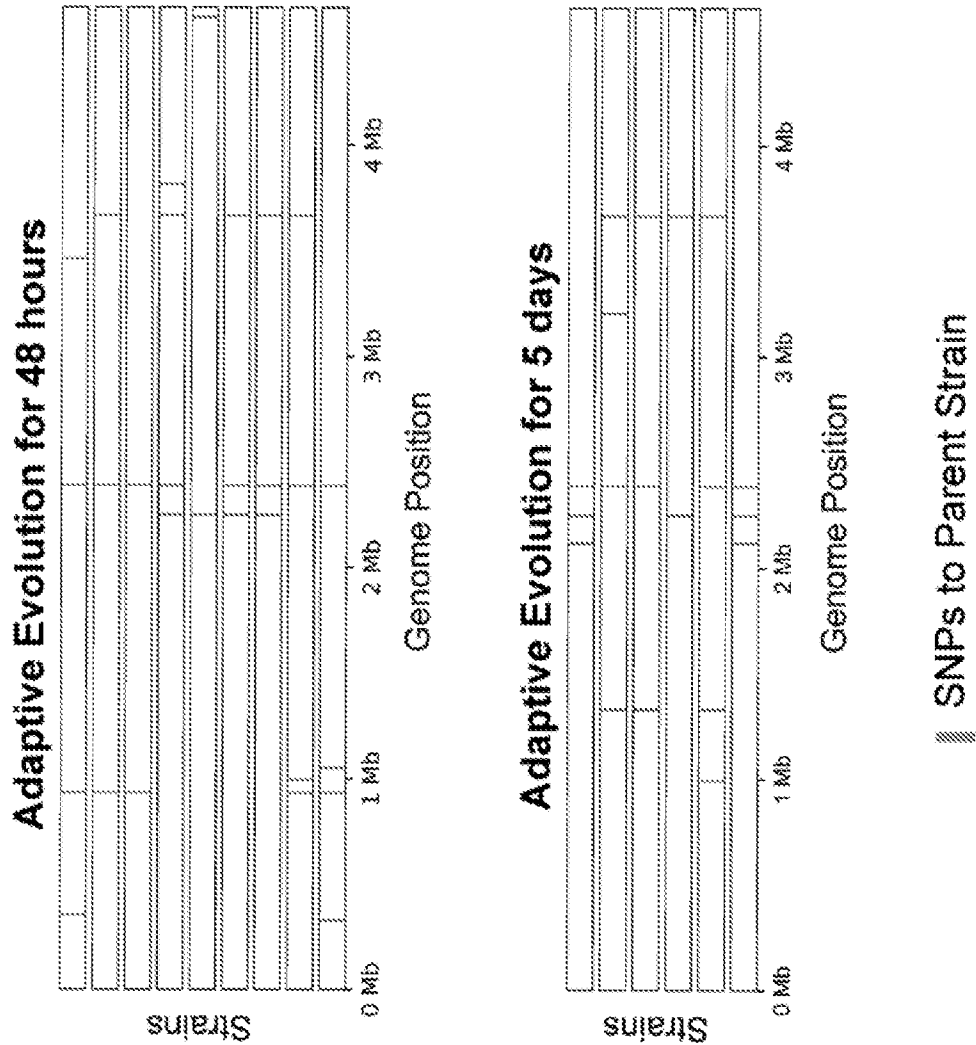


FIG. 21A

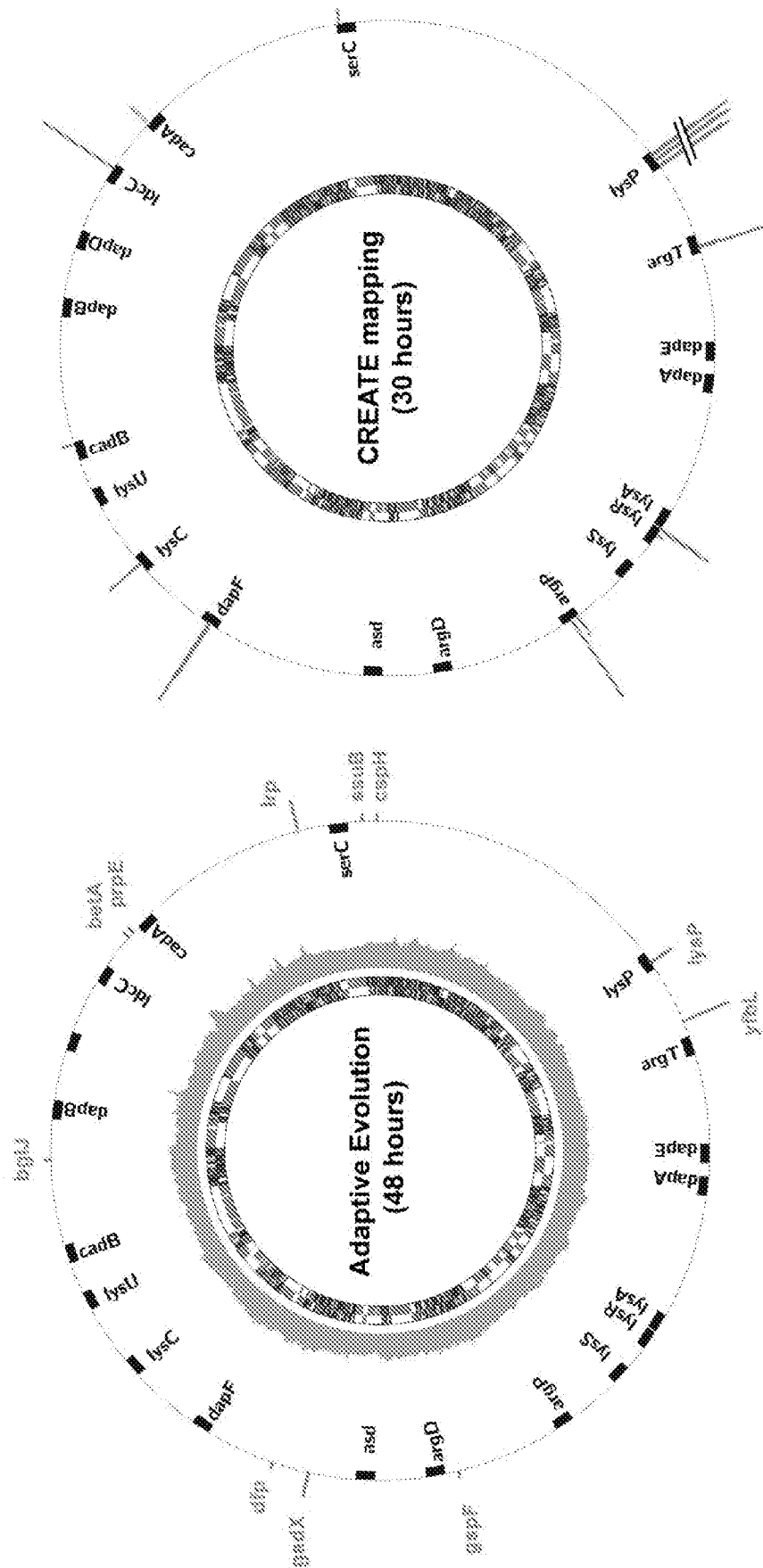


FIG. 21D

FIG. 21C

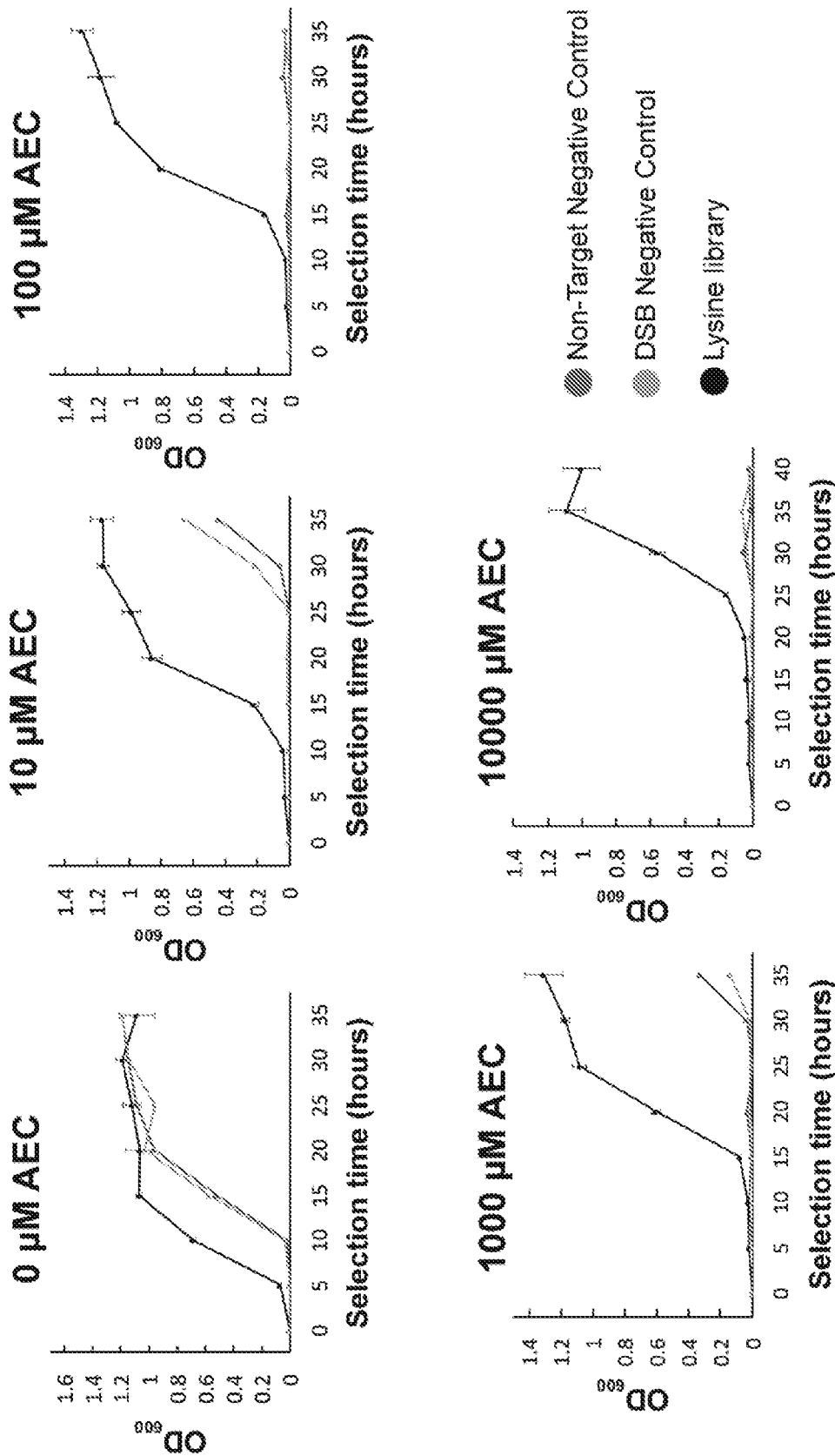


FIG. 22