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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 December 2024 (12.12.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/252392 A1

(51) International Patent Classification:
C12N 15/85 (2006.01)

(21) International Application Number:
PCT/IL2024/050552

(22) International Filing Date:
05 June 2024 (05.06.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/471,014 05 June 2023 (05.06.2023) US

(71) Applicant: **RAMOT AT TEL-AVIV UNIVERSITY LTD.** [IL/IL]; P.O. Box 39296, 6139201 Tel Aviv (IL).

(72) Inventors: **TULLER, Tamir**; c/o Ramot at Tel-Aviv University Ltd., P.O. Box 39296, 6139201 Tel Aviv (IL). **ARBEL GROISSMAN, Matan**; Shlomo HaMelekh St 69, 6451112 Tel Aviv-Yafo (IL). **TFILIN ARMONI, Ofri**; Isser Harel 8, 4223995 Netanya (IL). **SABAN, Rinat**; Sd. HaGiborim 5, 5241526 Ramat Gan (IL). **HAGGIAG, David**; Moshe Sne 31, 4421670 Kfar Saba (IL).

(74) Agent: **KESTEN,Dov** et al.; Geyra Kesten Frydman, 100 HaHashmonaim St., P.O. Box 52630, 6713317 Tel Aviv (IL).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

(54) Title: COMPUTATIONAL MODELS FOR EFFECTING COPY NUMBER OF PLASMIDS

(57) Abstract: Provided herein are methods of controlling gene expression by modifying plasmid copy number to a pre-determined copy number per a need of a user.

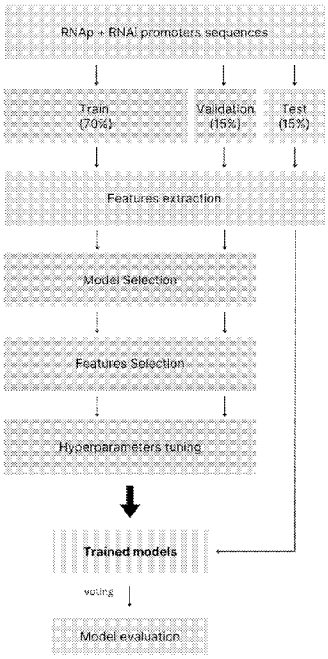


FIGURE 29C

WO 2024/252392 A1

[Continued on next page]

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

COMPUTATIONAL MODELS FOR EFFECTING COPY NUMBER OF PLASMIDS

REFERENCE TO AN ELECTRONIC SEQUENCE LISTING

[001] The contents of the electronic sequence listing (RMT-P-030-PCT.xml; size: 17,315 bytes; and date of creation: May 27, 2024) is herein incorporated by reference in its entirety.

CROSS-REFERENCE TO RELATED APPLICATIONS

[002] This application claims the benefit of priory of U.S. Provisional Patent Application No. 63/471,014, titled "COMPUTATIONAL MODELS FOR EFFECTING COPY NUMBER OF PLASMIDS", filed 5 June 2023, the contents of which are incorporated herein by reference in their entirety.

FIELD OF INVENTION

[003] The present invention is in the field of synthetic biology, and biotechnology, and relates to methods of controlling gene expression by tuning plasmid copy number.

BACKGROUND OF THE INVENTION

[004] Plasmid copy number refers to the number of plasmid copies in a single cell. Plasmid copy number is a crucial factor in synthetic biology, where plasmids are commonly used to transfer and express genes of interest in various organisms. Plasmid copy number can impact the level of protein expression or the level of expression of RNA molecules and help regulate gene expression. Therefore, controlling the copy number can be crucial in designing and optimizing synthetic biology systems.

[005] Current synthetic plasmids are often limited in their copy number selection, which can constrain the potential applications of the plasmids.

[006] There is still a great need to widen the range of copy numbers available for synthetic plasmids, such as by computational modeling of the interactions at the origin of replication (ORI) in different organisms.

SUMMARY OF THE INVENTION

[007] According to the first aspect, there is provided a method for maximizing plasmid copy number in a host cell, the method comprising: (a) modeling origin of replication (*ori*)

associated interactions the affect plasmid copy number in the host cell; (b) determining free energy, affinity, abundance, or any combination thereof, of the *ori* associated interactions; and (c) based on the determined free energy, affinity, abundance, or any combination thereof of step (b) introducing to the plasmid any one of: i. at least one mutation increasing the affinity, reducing the free energy, increasing the abundance, or any combination thereof, of *ori* associated interactions supporting high plasmid copy number of the plasmid; ii. at least one mutation decreasing the affinity, increasing the free energy, reducing the abundance, or any combination thereof, of *ori* associated interactions suppressing high plasmid copy number of the plasmid; or iii. both (i) and (ii).

[008] According to another aspect, there is provided a method for optimizing a copy number of a plasmid to host cell, the method comprising: (a) calculating an activity index for a host cell comprising a pre-determined copy number of the plasmid, wherein an activity index above or below a cut-off range is indicative of the pre-determined copy number of the plasmid being suboptimal for expression of a polynucleotide being integrated into the plasmid in the host cell; and (b) introducing to the plasmid at a pre-determined copy number determined as being suboptimal for expression of the polynucleotide in the host cell any one of: i. at least one mutation increasing the affinity, reducing the free energy, increasing the abundance, or any combination thereof, of *ori* associated interactions supporting high plasmid copy number of the plasmid; ii. at least one mutation decreasing the affinity, increasing the free energy, reducing the abundance, or any combination thereof, of *ori* associated interactions suppressing high plasmid copy number of the plasmid; or iii. both (i) and (ii), wherein the *ori* associated interactions are selected from the group consisting of: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – rop complexation, RNA I – tRNA complexation, RNA II – tRNA complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, and any combination thereof.

[009] According to another aspect, there is provided a method for obtaining a pre-determined copy number of a plasmid in a host cell for a user in need thereof, the method comprising introducing at least one mutation in the *ori* of the plasmid, a constituent of the *ori*, or both, thereby tuning the copy number of the plasmid to a pre-determined copy number in the host cell for the user.

[010] In some embodiments, the *ori* associated interactions are selected from the group consisting of: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – repressor of primer (rop) complexation, RNA I – tRNA complexation, RNA II – tRNA

complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, and any combination thereof.

[011] In some embodiments, the at least one mutation forms an RpoD sigma factor recognition site in RNA I of the *ori*.

[012] In some embodiments, the plasmid comprises a polynucleotide integrated therein.

[013] In some embodiments, the optimizing comprises increasing or decreasing the copy number of the plasmid such that it is optimal for expression of the polynucleotide in the host cell.

[014] In some embodiments, the tRNA is an uncharged tRNA.

[015] In some embodiments, the plasmid is a ColE1-like plasmid.

[016] In some embodiments, the at least one mutation comprises: substitution, deletion, insertion, or any combination thereof.

[017] In some embodiments, the *ori* comprises: RNA I, RNA II, rop, an uncharged tRNA, a promoter thereof, or any combination thereof.

[018] In some embodiments, the at least one mutation is introduced in RNA I or a promoter thereof, and wherein any one of: the mutated RNA I is a deleterious RNA I, the RNA I is characterized by reduced affinity to RNA II compared to a control, the mutated promoter is characterized by increased or decreased transcription rate compared to a control, and any combination thereof.

[019] In some embodiments, the mutation is in RNA II or a promoter thereof, and wherein any one of: the mutated RNA II is characterized by reduced affinity to RNA I compared to a control, the mutated RNA II is characterized by increased affinity to DNA compared to a control, the mutated promoter is characterized by increased or decreased transcription rate compared to a control, the mutated RNA II is mutated at a cleavage site of RNase H, and any combination thereof.

[020] In some embodiments, the mutation is in rop, and wherein the mutated rop is characterized by increased or reduced affinity to a complex comprising RNA I and RNA II (kissing complex) compared to a control.

[021] In some embodiments, the mutated rop stabilizes or destabilizes said kissing complex.

[022] In some embodiments, the method further comprises a step proceeding the introducing comprising contacting the cell with the plasmid comprising the at least one mutation.

[023] In some embodiments, the obtaining comprises increasing or decreasing the copy number of the plasmid such that it is suitable or optimal for expression of a polynucleotide being integrated into the plasmid in the cell.

[024] In some embodiments, the polynucleotide encodes a polypeptide of interest.

[025] Unless otherwise defined, all technical and/or scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the invention pertains. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of embodiments of the invention, exemplary methods and/or materials are described below. In case of conflict, the patent specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and are not intended to be necessarily limiting.

[026] Further embodiments and the full scope of applicability of the present invention will become apparent from the detailed description given hereinafter. However, it should be understood that the detailed description and specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[027] **Fig. 1** includes a plot and a horizontal bar graph showing Lasso For RNAp: R-squared equals 0.31 with the optimal alpha of 0.127.

[028] **Fig. 2** includes a plot and a horizontal bar graph showing Lasso for RNAi: R-squared equals 0.13 with the optimal alpha of 1.162.

[029] **Fig. 3** includes a plot and a horizontal bar graph showing Lasso for the shared model: R-squared equals 0.196 with the optimal alpha of 0.74.

[030] **Fig. 4** includes a horizontal bar graph and a plot showing XGBoost for RNAp: R-squared equals 0.908.

[031] **Fig. 5** includes a plot and a horizontal bar graph showing XGBoost for RNAi: R-squared equals 0.489.

[032] **Fig. 6** includes a plot and a horizontal bar graph showing XGBoost for RNAi: R-squared equals 0.839.

[033] **Fig. 7** includes a plot showing linear regression of feature selection using forward sequential feature selection of for priming RNA.

[034] **Fig. 8** includes a plot showing linear regression of feature selection using forward sequential feature selection of for inhibitor RNA.

[035] **Fig. 9** includes graphs showing XGBoost complexity analysis. The XGBoost model was trained with different numbers of trees with a range of 10-300 trees and default maximum depth of 6, and with different values of maximum depth with a range of 2-20 and a default number of trees of 100. The R2 and Spearman correlation between the true values of the test set and the predicted values of each model were calculated and displayed.

[036] **Fig. 10** includes bar graphs showing features comparison between the models. The models were trained 1,000 times and the 40 features of each model with the highest features importance score on average were intercrossed. The y axis represents the feature importance of the two models: F score for XGBoost and the coefficient for Lasso. The common features are displayed in the graphs.

[037] **Fig. 11** includes dot plots showing correlation of all features correlation by importance (XGboost) and average (Lasso) in inhibitor RNA (RNAi) and priming RNA (RNAp).

[038] **Fig. 12** includes dot plots showing correlation of 40 highest features correlation by importance (XGboost) and average (Lasso) in inhibitor RNA (RNAi) and priming RNA (RNAp).

[039] **Fig. 13** includes dot plots showing correlation of all features correlation by importance (XGboost) and average (Lasso) in inhibitor RNA (RNAi) and priming RNA (RNAp).

[040] **Fig. 14** includes dot plots showing correlation of 100 highest features correlation by importance (XGboost) and average (Lasso) in inhibitor RNA (RNAi) and priming RNA (RNAp).

[041] **Fig. 15** includes correlation between top MIC features and copy number.

[042] **Fig. 16** includes correlation between top Pearson features and copy number.

[043] **Fig. 17** includes correlation between top Spearman features and copy number.

[044] **Fig. 18** includes plots showing promoter correlations. A correlation of MIC being 0.3 and 0.23 with a significant p-value is observed for RNAi, and for RNAp, respectively.

[045] **Fig. 19** includes plots showing sigma factor of RNA polymerase correlations. The rpoD17 and rpoD18 motifs are mutations found in the gene encoding the sigma factor of RNA polymerase in the bacterium *Escherichia coli*. These mutations affect the ability of the sigma factor to bind to DNA and initiate transcription, leading to changes in gene expression patterns in *E. coli*.

[046] **Fig. 20** includes graphs showing AA count and TT count. This feature represents the number of times 'AA' or 'TT' being presented in the promoter sequence.

[047] **Fig. 21** includes PSSM score calculated with the 20% high copy number only, for RNAp. The presented sequence is set forth in SEQ ID NO: 10.

[048] **Fig. 22** includes PSSM score calculated with the 20% high copy number only, for RNAi.

[049] **Fig. 23** includes vertical bar graphs showing the number of times each feature was used in the models out of 5,000 iterations is represented by count (the height of the bin), the feature importance is represented by the features orders on the x axis (from left to right) and the average R2 score of the models in the iterations the feature was used is represented by the color bar.

[050] **Fig. 24** includes a table listing the statistical information of the R2 score of **Fig. 23**.

[051] **Figs. 25A-25F** include graphs, a table, and horizontal bar graphs. **(25A)** RNAp results showing actual copy number vs predicted copy number, with pearson correlation of 0.959 (p.v= $7.92e^{-70}$) and spearman correlation of 0.863 (p.v= $1.3e^{-38}$). **(25B)** RNAi results showing actual copy number vs predicted copy number, with pearson correlation of 0.79 (p.v= 5.13×10^{-13}) and spearman correlation of 0.82 (p.v= 1.62×10^{-14}). **(25C)** A table of the results of **25A** and **25B**. **(25D)** Priming RNA feature importance for CatBoost model, the higher the feature importance score is the higher feature contribution to the predicted copy number. **(25E)** Priming RNA feature importance for XGBoost model, the higher the feature importance score is the higher feature contribution to the predicted copy number. **(25F)** Inhibitory RNA feature importance for CatBoost model, the higher the feature importance score is the higher feature contribution to the predicted copy number.

[052] **Figs. 26A-26D** include plots and horizontal bar graphs. **(26A)** Features distribution corresponding to their copy number for the common features of XGBoost and CatBoot in

the voting model for RNAp. **(26B)** Features distribution corresponding to their copy number for the 2 most important features CatBoot in RNAi model. **(26C)** The absolute mean SHAP value for each feature chosen during the feature selection process in the model based on the RNAp model. **(26D)** The absolute mean SHAP value for each feature chosen during the feature selection process in the model based on the RNAi model.

[053] **Figs. 27A-27B** include graphs. **(27A)** SHAP value for each value of the feature dG_total. There is a specific range of dG_total values with positive SHAP values. **(27B)** copy number as a function of the feature dG_total. It can be seen that high copy number sequences are in the same dG_total range as observed in **27A**, the inventors validated this conclusion statistically by an enrichment p value which was found to be significant.

[054] **Figs. 28A-28D** include illustrations, a graph, and a sequence. **(28A)** A modified origin of replication is synthesized from tool-generated data, followed by PCR amplification of pUC19 to incorporate new flanking regions around the ORI. Subsequently, a restriction enzyme is utilized to target a specific cut site in the original pUC19 sequence. The modified ORI is integrated into the pUC19 backbone via homologous joining, resulting in successful transformation of the modified pUC19 vectors Created with BioRender. **(28B)** Predicted Copy number by the model Vs. ddPCR or qPCR measurements. Pearson correlation of 0.652, pv=0.028. **(28C)** Ctrl-C Webtool- Choose between (1) using the provided plasmids (currently supporting pUC-19 and pBR322), (2) uploading your own plasmid's FASTA file or (3) providing the plasmid's sequence as text. Choose your desired copy number. If you want a grid (range) of copy numbers, first click on the "grid" option, then enter 5 values of your desired grid of copy numbers. Choose whether to visualize the annotated plasmid map or not. Click "Get Modified Plasmid" and wait for the results to appear on the screen. **(28D)** If "visualize the plasmid map" checkbox was marked, clicking on the different site of the plasmid's map illustration will refer the user to the site sequence. Click "Download Modified Sequence" to download the plasmid modified sequence as a FASTA file. The presented sequence is set forth in SEQ ID NO: 11.

[055] **Figs. 29A-29D** include illustrations, a flow chart, and sequences. **(29A)** The interaction strengths between RNA polymerase and the subunit of RNA polymerase σ70 factor and promoter DNA. **(29B)** RNA polymerase binding energy matrix. Colors indicate the contribution of each base-pair to the total binding energy. **(29C)** Model training and evaluation workflow. **(29D)** A few examples of RNAp variants and their corresponding copy number. All the variants are mutated in positions -33 to -30, -8 to -11, and +1. Sequence in 29B is set forth in SEQ ID NO: 12. The sequences of the variants of 29D are set forth in SEQ ID Nos: 13-18.

[056] **Figs. 30A-30B** include vertical bar graphs showing the models performance in the model selection section. **(30A)** RNAp Models exploration showing Pearson correlation and Spearman correlation results on train vs validation. Chosen models are surrounded by a rectangle. The final model is a voting model that utilizes these models (see more in methods). **(30B)** RNAi Models exploration showing Pearson correlation and Spearman correlation results on train vs validation. Chosen models for the final model are surrounded by a rectangle.

DETAILED DESCRIPTION OF THE INVENTION

[057] According to the first aspect, there is provided a method for optimizing a copy number of a plasmid to host cell.

[058] According to another aspect, there is provided a method for maximizing plasmid copy number in a host cell.

[059] According to another aspect, there is provided a method for obtaining a pre-determined copy number of a plasmid in cell for a user in need thereof.

[060] According to another aspect, there is provided a method for tuning a copy number of a plasmid to a pre-determined copy number in cell.

[061] In some embodiments, the method comprises introducing at least one mutation in the origin of replication (*ori*) of the plasmid, a constituent of the *ori*, or both, thereby tuning a copy number of a plasmid to a pre-determined copy number in a cell.

[062] In some embodiments, at least one mutation comprises: substitution, deletion, insertion, or any combination thereof.

[063] In some embodiments, a constituent of an *ori* comprises: RNA I, RNA II, repressor of primer (rop), an uncharged transfer RNA (tRNA), a promoter thereof, or any combination thereof.

[064] In some embodiments, tRNA is or comprises an uncharged tRNA.

[065] In some embodiments, the at least one mutation is introduced in RNA I or a promoter thereof.

[066] In some embodiments, the mutated RNA I is a deleterious RNA I. In some embodiments, the RNA I is characterized by reduced affinity to RNA II compared to a control. In some embodiments, the mutated promoter is characterized by increased or decreased transcription rate compared to a control. In some embodiments, the mutated RNA

I is a deleterious RNA I, the mutated RNA I is characterized by reduced affinity to RNA II compared to a control, the mutated promoter is characterized by increased or decreased transcription rate compared to a control, or any combination thereof.

[067] In some embodiments, the mutation is in RNA II or a promoter thereof.

[068] In some embodiments, the mutated RNA II is characterized by reduced affinity to RNA I compared to a control. In some embodiments, the mutated RNA II is characterized by increased affinity to DNA compared to a control. In some embodiments, the mutated promoter is characterized by increased or decreased transcription rate compared to a control. In some embodiments, the mutated RNA II is mutated at a cleavage site of RNase H.

[069] In some embodiments, the mutated RNA II is characterized by reduced affinity to RNA I compared to a control, the mutated RNA II is characterized by increased affinity to DNA compared to a control, the mutated promoter is characterized by increased or decreased transcription rate compared to a control, the mutated RNA II is mutated at a cleavage site of RNase H, or any combination thereof.

[070] In some embodiments, the mutation is in rop. In some embodiments, a mutated rop is characterized by increased or reduced affinity to a complex comprising RNA I and RNA II (kissing complex) compared to a control.

[071] In some embodiments, a mutated rop stabilizes or destabilizes a kissing complex.

[072] In some embodiments, the method further comprises a step proceeding the introducing, comprising contacting a cell with a plasmid comprising the at least one mutation.

[073] In some embodiments, RNA II is an RNA primer or pre-primer (RNAp).

[074] In some embodiments, RNA I is an inhibitory or interfering RNA (RNAi).

[075] In some embodiments, RNA I and RNA II are at least partially complementary (to one another). In some embodiments, RNA I is at least partially complementary to RNA II. In some embodiments, complementary comprises reverse and complementary. In some embodiments, RNA I is complementary to RNA II in a level that effectively or sufficiently reduces or inhibits hybridization of RNA II to the plasmid. In some embodiments, RNA I is characterized by a complementarity level to RNA II that effectively or sufficiently reduces or inhibits hybridization of RNA II to the plasmid.

[076] In some embodiments, the plasmid is or comprises a ColE1-like plasmid.

[077] Types and sources of ColE1-like plasmids are common and would be apparent to one of ordinary skill in the art, such as reviewed by Chaillou et al., Nucleic Acids Res. 2022; 50(16): 9568–9579.

[078] In some embodiments, the plasmid is or comprises pUC19. In some embodiments, a ColE1-like plasmid is or comprises pUC19.

[079] In some embodiments, at least partially complementary comprises 50-100%, 60-100%, 70-100%, 80-100%, or 90-100% complementary. Each possibility represents a separate embodiment of the invention. In some embodiments, complementarity is over or in an equal-length portion.

[080] In some embodiments, the obtaining, tuning, or both, comprises increasing or decreasing the copy number of the plasmid such that it is suitable or optimal for expression of the polynucleotide in the host cell. In some embodiments, the obtaining, tuning, or both is to a particular or definitive copy number of the plasmid.

[081] In some embodiments, the polynucleotide encodes a polypeptide of interest. In some embodiments, the polypeptide comprises a recombinant polypeptide.

[082] In some embodiments, the cell or the host cell is further cultured under conditions sufficient for expression of the polypeptide of interest. In some embodiments, the cell or the host cell is further cultured under conditions such that the polypeptide of interest sufficiently, effectively, or both, is expressed, secreted, or both. In some embodiments, the cell or the host cell is further cultured under conditions such that the polypeptide of interest is expressed, secreted, or both.

[083] Methods for culturing cells, such as for expression, secretion, or both, of polypeptide(s) of interest are common and would be apparent to one of ordinary skill in the art.

[084] In some embodiments, the methods further comprises culturing the cell comprising the plasmid being modified, optimized, tuned, maximized, or any combination thereof for copy number, according to the method of the invention.

[085] In some embodiments, the method further comprises isolating, purifying, retrieving, or any combination thereof, the polypeptide of interest from the cultured cell or cultured host cell, or a medium in which the cell or host cell being cultured.

[086] Methods for polypeptide isolation and/or purification from cell culture are common and would be apparent to one of ordinary skill in the art of biochemistry.

[087] In some embodiments, the expression “of interest” is to a user or a human user. In some embodiments, the polypeptide of interest comprises any polypeptide, the expression of a recombinant version thereof is of importance to the user. In some embodiments, importance is commercial importance. In some embodiments, the polypeptide is a product or an ingredient thereof. In some embodiments, the polypeptide is a therapeutic agent or an active ingredient thereof.

[088] As used herein, the terms “peptide”, "polypeptide" and "protein" are used interchangeably to refer to a polymer of amino acid residues. In another embodiment, the terms "peptide", "polypeptide" and "protein" as used herein encompass native peptides, peptidomimetics (typically including non-peptide bonds or other synthetic modifications) and the peptide analogues peptoids and semipeptoids or any combination thereof. In another embodiment, the peptides polypeptides and proteins described have modifications rendering them more stable while in the body or more capable of penetrating into cells. In one embodiment, the terms “peptide”, "polypeptide" and "protein" apply to naturally occurring amino acid polymers. In another embodiment, the terms “peptide”, "polypeptide" and "protein" apply to amino acid polymers in which one or more amino acid residue is an artificial chemical analogue of a corresponding naturally occurring amino acid.

[089] In some embodiments, the method comprises: (a) modeling *ori* associated interactions the affect plasmid copy number in the host cell; (b) determining free energy, affinity, abundance, or any combination thereof, of the *ori* associated interactions; and (c) based on the determined free energy, affinity, abundance, or any combination thereof of step (b) introducing to the plasmid at least one mutation.

Ori associated interaction of RNA II or its promoter RNAp

[090] In some embodiments, the modeling comprises determining: (i) free energy of promoter binding; (ii) abundance of complexes comprising the promoter and sigma factor, RNA polymerase, or both; or (iii) any combination of (i) and (ii).

[091] In some embodiments, the promoter comprises a promoter of RNA II. In some embodiments, the complex comprises a promoter of RNA II and sigma factor, a promoter of RNA II and RNA polymerase, or a promoter of RNA II, sigma factor, and RNA polymerase.

[092] In some embodiments, the abundance reflects and the strength of sigma factor recognition sites. In some embodiments, the strength of sigma factor recognition sites is reflected by the abundance, e.g., of the complexes disclosed herein. In some embodiments, the promoter, e.g., RNA II comprises one or more sigma factor recognition sites. In some

embodiments, the promoter, e.g., RNA II comprises at least one sigma factor recognition sites. In some embodiments, the promoter, e.g., RNA II comprises a plurality of sigma factor recognition sites.

[093] As used herein, the term "plurality" refers to any integer being equal to or greater than 2.

[094] In some embodiments, the method comprises determining an *ori* sequence of a plasmid. The method further comprises extracting, from the determined *ori* sequence, a set of features representing transcription related properties. In some embodiments the set of features comprises at least one of: motif score; P-values; PSSM scores (reflecting frequency of nucleotides in high copy number sequences); promoter strength; nucleotide counts; interaction transcription rate measures, or any combination thereof.

[095] In some embodiments, the method comprises determining at least one feature related to the free energy of the promoter. In some embodiments, the set of features comprises at least one feature related to the free energy of the promoter. In some embodiments, the free energy is of an unbound promoter. In some embodiments, the free energy is of a bound promoter. In some embodiments, the free energy is the delta free energy of a bound promoter compared to free energy of an unbound promoter, or vice versa. In some embodiments, a bound promoter is bound to a transcription factor or a plurality thereof, a DNA binding protein, RNA polymerase, or any combination thereof. In some embodiments, a transcription factor or a plurality thereof, a DNA binding protein, or both, comprises a sigma factor.

[096] In some embodiments, free energy is defined or reflected as delta G (dG or ΔG). In some embodiments, free energy is dG is total dG (e.g., dG_total). In some embodiments, total dG comprises or refers to the overall change in free energy between an unbound promoter, such as of RNA II, and a promoter, such as of RNA II bound to a sigma factor, RNA polymerase, or both. In some embodiments, total dG comprises or refers to the overall change in free energy between an unbound RNA II promoter and an RNA II promoter being bound to a sigma factor, RNA polymerase, or both.

[097] In some embodiments, the method further comprises determining the abundance of complexes comprising the promoter and: sigma factor, RNA polymerase, or both, comprises determining the presence of the rpoD16 motif (a sigma factor recognition site) in the promoter. In some embodiments, the method further comprises determining rpoD16_score as a numerical value representing the strength or likelihood of the presence of the rpoD16 motif in the promoter. In some embodiments, the set of features comprises a feature

representing the abundance or presence of rpoD16 motif in the promoter. In some embodiments, a , or, more specifically, a feature representing the abundance or presence of rpoD16 motif in the promoter is or comprises rpoD16_score.

[098] In some embodiments, the method further comprises calculating a vector representation of the determined *ori* sequence, based on the determined set of features, thereby performing an embedding of the *ori* sequence into a respective feature space.

[099] In some embodiments, the method further comprises inferring a respectively pretrained machine learning (ML)-based model on the calculated vector representation to predict the plasmid copy number.

[0100] In some embodiments, the pretrained ML-based model comprises one of gradient boosting algorithms known in the art. In some embodiments, an ML-based model is based on Categorical Boosting (CatBoost) algorithm. In some embodiments, additionally or alternatively, ML-based model is based on Extreme Gradient Boosting (XGBoost) algorithm.

[0101] In some embodiments, ML-based model is pretrained based on samples representing mutations in the RNAp (alternatively used herein for RNA II) promoter region.

[0102] In some embodiments, the method further comprises generating at least one mutation (or analyzing a suggested mutation) to be introduced to the plasmid, based on the predicted plasmid copy number.

[0103] In some embodiments, at least one mutation reducing dG to about -3.7 to -2.5 increases copy number of the plasmid. In some embodiments, the at least one mutation increasing copy number of the plasmid reduces dG to about -3.7 to -2.5. In some embodiments, at least one mutation reducing dG to about -3.8 or less reduces copy number of the plasmid. In some embodiments, the at least one mutation reducing copy number of the plasmid reduces dG to about -3.8 or less. In some embodiments, at least one mutation increasing dG to about -2.4 or more reduces copy number of the plasmid. In some embodiments, the at least one mutation reducing copy number of the plasmid increases dG to about -2.4 or more.

[0104] In some embodiments, dG of about -3.7 to -2.5 is or comprises about -3.4 to -2.7.

[0105] It was determined by the inventors that both total dG and rpoD16_score are strongly correlated with the copy number, in relation with mutations in the RNAp promoter region, which is further discussed in greater detail with reference to **Figs. 25D-25F**. Therefore, the

usage the set of features including either or both total dG and rpoD16_score contributes into the improvement of the technological field, as it provides for higher accuracy and reliability of the plasmid copy number prediction, thereby facilitating the determination of the mutation that would ascertain maximization of plasmid copy number in a host cell.

[0106] As further explained herein, it was determined by the inventors, that using the set of features comprising the rpoD16_score feature in combination with the respectively pretrained ML-based model (e.g., pretrained based on samples representing mutations in the RNAp promoter region), which is constructed based on XGBoost algorithm, as well as using the set of features comprising the total dG feature in combination with the respectively pretrained ML-based model (e.g., pretrained based on samples representing mutations in the RNAp promoter region) which is constructed based on CatBoost algorithm provides for more reliable plasmid copy number prediction.

Ori associated interaction of RNA I (RNAi)

[0107] In some embodiments, the method comprises determining the probability of adjacent RNA nucleotides of RNA I to hybridize. In some embodiments, the determining comprises determining the probability of secondary structure formation(s) of RNA I. In some embodiments, the determining is by using RNAFold or an equivalent thereof. In some embodiments, the determining comprises determining the probability of the nucleotide in index 375 to pair with the nucleotide in index 385 of the first 400 bases of RNA I (e.g., seq_end_400_375_385).

[0108] In some embodiments, the set of features further comprises the feature representing the determined probability of secondary structure(s) formation (RNAFold).

[0109] In some embodiments, the determined probability of secondary structure(s) formation is further analyzed using Catboost. In some embodiments, the method further comprises analyzing the determined probability of secondary structure(s) formation using Catboost.

[0110] In some embodiments, the method comprises determining the presence of the *de-novo* motif in a promoter of RNA I. In some embodiments, the *de-novo* motif comprises the nucleic acid sequence: TGCGCGTTTAAT (SEQ ID NO: 7) or having 80-100% sequence identity thereto. In some embodiments, the *de-novo* motif is present with higher rates in promoters of high copy number plasmids, compared to control promoters of low copy number plasmids. In some embodiments, the method comprises determining the presence of the nucleic acid sequence: TGCGCGTTTAAT (SEQ ID NO: 7) in a promoter of RNA I. In

some embodiments, the method further comprises analyzing the determined presence of the *de-novo* motif in a promoter of RNA I.

[0111] In some embodiments, the set of features further comprises a feature representing a presence of the nucleic acid sequence: TGCGCGTTTAAT (SEQ ID NO: 7).

[0112] In some embodiments, the ML-based model is pretrained based on samples representing mutations in the RNAi (as used herein interchangeably with RNA I) promoter region.

[0113] As further explained herein, it was determined by the inventors, that using the set of features comprising the feature representing the probability of secondary structure(s) formation (RNAFold) in combination with the respectively pretrained ML-based model (e.g., pretrained based on samples representing mutations in the RNAi promoter region), which is constructed based on CatBoost algorithm, as well as using the set of features comprising the feature representing presence of the nucleic acid sequence: TGCGCGTTTAAT (SEQ ID NO: 7) in combination with the respectively pretrained ML-based model (e.g., pretrained based on samples representing mutations in the RNAi promoter region) which is constructed based on CatBoost algorithm, provides for more reliable plasmid copy number prediction.

[0114] In some embodiments, the abovementioned variants of ML-based model implementations may be used in combination, e.g., be arranged as an ensemble of ML-based models, for further improvement of the plasmid copy number prediction.

[0115] In some embodiments, the at least one mutation comprises any one of: (i) at least one mutation increasing the affinity, reducing the free energy, increasing the abundance, or any combination thereof, of *ori* associated interactions supporting high plasmid copy number of the plasmid; (ii) at least one mutation decreasing the affinity, increasing the free energy, reducing the abundance, or any combination thereof, of *ori* associated interactions suppressing high plasmid copy number of the plasmid; or (iii) both (i) and (ii).

[0116] In some embodiments, the *ori* associated interactions are selected from: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – rop complexation, RNA I – tRNA complexation, RNA II – tRNA complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, or any combination thereof.

[0117] In some embodiments, the at least one mutation forms or introduces an RpoD sigma factor recognition site in the RNA II promoter. In some embodiments, the at least one mutation increases the abundance or presence of an RpoD sigma factor recognition site(s) in the RNA II promoter. In some embodiments, the at least one mutation forms or introduces the nucleic acid sequence TGCGCGTTTAAT (SEQ ID NO: 7) in a promoter of RNA I. In some embodiments, the at least one mutation increases the abundance or presence of the nucleic acid sequence TGCGCGTTTAAT (SEQ ID NO: 7) in a promoter of RNA I. In some embodiments, the at least one mutation forms or introduces at least one secondary structure in RNA I or a promoter thereof. In some embodiments, the at least one mutation increases the abundance or presence of at least one secondary structure in RNA I or a promoter thereof. In some embodiments, at the least one mutation reduces the total dG of a complex comprising the RNA II promoter and: a sigma factor, RNA polymerase, or both.

[0118] In some embodiments, increase the abundance or presence of a feature as disclosed herein, e.g., dG, rpo16D, secondary structure, SEQ ID NO: 7, etc., comprises increase the probability to determine the presence of the feature. In some embodiments, increase the abundance or presence of a feature as disclosed herein, e.g., dG, rpo16D, secondary structure, SEQ ID NO: 7, etc., means that the probability to determine the presence of the feature is increased.

[0119] In some embodiments, reduces or reduce or increase or increases is compared to a control promoter being: non-modified, non-mutated, wildtype, naïve, or any combination thereof.

[0120] In some embodiments, the method comprises: (a) calculating an activity index for a host cell comprising a pre-determined copy number of the plasmid, wherein an activity index above or below a cut-off range is indicative of the pre-determined copy number of the plasmid being suboptimal for expression of a polynucleotide being integrated into the plasmid in the host cell; and (b) introducing to the plasmid at a pre-determined copy number determined as being suboptimal for expression of the polynucleotide in the host cell any one of: (i) at least one mutation increasing the affinity of *ori* associated interactions supporting high plasmid copy number of the plasmid; (ii) at least one mutation decreasing the affinity of *ori* associated interactions suppressing high plasmid copy number of the plasmid; or (iii) both (i) and (ii).

[0121] In some embodiments, the optimizing comprises increasing or decreasing the copy number of the plasmid such that it is optimal for expression of the polynucleotide in the host

cell. In some embodiments, the optimizing is to a particular or definitive copy number of the plasmid.

[0122] In some embodiments, an activity index equivalent to a cut-off range is indicative of the pre-determined copy number of the plasmid being optimal for expression of a polynucleotide being integrated into the plasmid in the host cell.

[0123] In some embodiments, the method is an in vitro method. In some embodiments, the method is a computerized method.

[0124] The term "expression" as used herein refers to the biosynthesis of a gene product, including the transcription and/or translation of said gene product. Thus, expression of a nucleic acid molecule may refer to transcription of the nucleic acid fragment (e.g., transcription resulting in mRNA or other functional RNA) and/or translation of RNA into a precursor or mature protein (polypeptide).

[0125] Expressing a gene within a cell is well known to one skilled in the art. It can be carried out by, among many methods, transfection, viral infection, or direct alteration of the cell's genome. In some embodiments, the gene is in an expression vector such as plasmid or viral vector. One such example of an expression vector containing p16-Ink4a is the mammalian expression vector pCMV p16 INK4A available from Addgene.

[0126] A vector nucleic acid sequence generally contains at least an origin of replication for propagation in a cell and optionally additional elements, such as a heterologous polynucleotide sequence, expression control element (e.g., a promoter, enhancer), selectable marker (e.g., antibiotic resistance), poly-Adenine sequence.

[0127] The vector may be a DNA plasmid delivered via non-viral methods or via viral methods. The viral vector may be a retroviral vector, a herpesviral vector, an adenoviral vector, an adeno-associated viral vector or a poxviral vector. The promoters may be active in mammalian cells. The promoters may be a viral promoter.

[0128] In some embodiments, the gene is operably linked to a promoter. The term "operably linked" is intended to mean that the nucleotide sequence of interest is linked to the regulatory element or elements in a manner that allows for expression of the nucleotide sequence (e.g. in an in vitro transcription/translation system or in a host cell when the vector is introduced into the host cell).

[0129] In some embodiments, the vector is introduced into the cell by standard methods including electroporation (e.g., as described in From et al., Proc. Natl. Acad. Sci. USA 82,

5824 (1985)),Heat shock, infection by viral vectors, high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface (Klein et al., Nature 327. 70-73 (1987)), and/or the like.

[0130] The term "promoter" as used herein refers to a group of transcriptional control modules that are clustered around the initiation site for an RNA polymerase i.e., RNA polymerase II. Promoters are composed of discrete functional modules, each consisting of approximately 7-20 bp of DNA, and containing one or more recognition sites for transcriptional activator or repressor proteins.

[0131] In some embodiments, nucleic acid sequences are transcribed by RNA polymerase II (RNAP II and Pol II). RNAP II is an enzyme found in eukaryotic cells. It catalyzes the transcription of DNA to synthesize precursors of mRNA and most snRNA and microRNA.

[0132] In some embodiments, mammalian expression vectors include, but are not limited to, pcDNA3, pcDNA3.1 (±), pGL3, pZeoSV2(±), pSecTag2, pDisplay, pEF/myc/cyto, pCMV/myc/cyto, pCR3.1, pSinRep5, DH26S, DHBB, pNMT1, pNMT41, pNMT81, which are available from Invitrogen, pCI which is available from Promega, pMbac, pPbac, pBK-RSV and pBK-CMV which are available from Strategene, pTRES which is available from Clontech, and their derivatives.

[0133] In some embodiments, expression vectors containing regulatory elements from eukaryotic viruses such as retroviruses are used by the present invention. SV40 vectors include pSVT7 and pMT2. In some embodiments, vectors derived from bovine papilloma virus include pBV-1MTHA, and vectors derived from Epstein Bar virus include pHEBO, and p2O5. Other exemplary vectors include pMSG, pAV009/A+, pMTO10/A+, pMAMneo-5, baculovirus pDSVE, and any other vector allowing expression of proteins under the direction of the SV-40 early promoter, SV-40 later promoter, metallothionein promoter, murine mammary tumor virus promoter, Rous sarcoma virus promoter, polyhedrin promoter, or other promoters shown effective for expression in eukaryotic cells.

[0134] In some embodiments, recombinant viral vectors, which offer advantages such as lateral infection and targeting specificity, are used for in vivo expression. In one embodiment, lateral infection is inherent in the life cycle of, for example, retrovirus and is the process by which a single infected cell produces many progeny virions that bud off and infect neighboring cells. In one embodiment, the result is that a large area becomes rapidly infected, most of which was not initially infected by the original viral particles. In one embodiment, viral vectors are produced that are unable to spread laterally. In one

embodiment, this characteristic can be useful if the desired purpose is to introduce a specified gene into only a localized number of targeted cells.

[0135] Various methods can be used to introduce the expression vector of the present invention into cells. Such methods are generally described in Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (1989, 1992), in Ausubel et al., Current Protocols in Molecular Biology, John Wiley and Sons, Baltimore, Md. (1989), Chang et al., Somatic Gene Therapy, CRC Press, Ann Arbor, Mich. (1995), Vega et al., Gene Targeting, CRC Press, Ann Arbor Mich. (1995), Vectors: A Survey of Molecular Cloning Vectors and Their Uses, Butterworths, Boston Mass. (1988) and Gilboa et at. [Biotechniques 4 (6): 504-512, 1986] and include, for example, stable or transient transfection, lipofection, electroporation and infection with recombinant viral vectors. In addition, see U.S. Pat. Nos. 5,464,764 and 5,487,992 for positive-negative selection methods.

[0136] In one embodiment, plant expression vectors are used. In one embodiment, the expression of a polypeptide coding sequence is driven by a number of promoters. In some embodiments, viral promoters such as the 35S RNA and 19S RNA promoters of CaMV [Brisson et al., Nature 310:511-514 (1984)], or the coat protein promoter to TMV [Takamatsu et al., EMBO J. 6:307-311 (1987)] are used. In another embodiment, plant promoters are used such as, for example, the small subunit of RUBISCO [Coruzzi et al., EMBO J. 3:1671-1680 (1984); and Brogli et al., Science 224:838-843 (1984)] or heat shock promoters, e.g., soybean hsp17.5-E or hsp17.3-B [Gurley et al., Mol. Cell. Biol. 6:559-565 (1986)]. In one embodiment, constructs are introduced into plant cells using Ti plasmid, Ri plasmid, plant viral vectors, direct DNA transformation, microinjection, electroporation and other techniques well known to the skilled artisan. See, for example, Weissbach & Weissbach [Methods for Plant Molecular Biology, Academic Press, NY, Section VIII, pp 421-463 (1988)]. Other expression systems such as insects and mammalian host cell systems, which are well known in the art, can also be used by the present invention.

[0137] It will be appreciated that other than containing the necessary elements for the transcription and translation of the inserted coding sequence (encoding the polypeptide), the expression construct of the present invention can also include sequences engineered to optimize stability, production, purification, yield or activity of the expressed polypeptide.

[0138] The term "nucleic acid" is well known in the art. A "nucleic acid" as used herein will generally refer to a molecule (i.e., a strand) of DNA, RNA or a derivative or analog thereof,

comprising a nucleobase. A nucleobase includes, for example, a naturally occurring purine or pyrimidine base found in DNA (e.g., an adenine "A," a guanine "G," a thymine "T" or a cytosine "C") or RNA (e.g., an A, a G, an uracil "U" or a C).

[0139] The terms "polynucleotide," "polynucleotide sequence," "nucleic acid sequence," and "nucleic acid molecule" are used interchangeably herein. These terms encompass nucleotide sequences and the like. A polynucleotide may be a polymer of RNA or DNA that is single- or double-stranded, that optionally contains synthetic, non-natural or altered nucleotide bases.

[0140] As used herein, the term "primer" includes an oligonucleotide, either natural or synthetic, that is capable, upon forming a duplex with a polynucleotide template, of acting as a point of initiation of nucleic acid synthesis and being extended from its 3′ end along the template so that an extended duplex is formed. Primers within the scope of the present invention bind adjacent to a target sequence. A "primer" may be considered a short polynucleotide, generally with a free 3′-OH group that binds to a target or template potentially present in a sample of interest by hybridizing with the target, and thereafter promoting polymerization of a polynucleotide complementary to the target.

[0141] The term "interaction strength" as used herein refers to hybridization free energy between a nucleic acid molecule and a ribosomal RNA. Lower and more negative free energy is related to stronger hybridization and stronger interaction strength. Hybridization free energy can be computed based on the Vienna package RNAcoFold, which computes a common secondary structure of two RNA molecules. According to some embodiments, the interaction strength can be defined by a scale of strong, intermediate and weak.

[0142] The term "hybridization" or "hybridizes" as used herein refers to the formation of a duplex between nucleotide sequences which are sufficiently complementary to form duplexes via Watson-Crick base pairing. Two nucleotide sequences are "complementary" to one another when those molecules share base pair organization homology. "Complementary" nucleotide sequences will combine with specificity to form a stable duplex under appropriate hybridization conditions. For instance, two sequences are complementary when a section of a first sequence can bind to a section of a second sequence in an anti-parallel sense wherein the 3′-end of each sequence binds to the 5′-end of the other sequence and each A, T (U), G and C of one sequence is then aligned with a T (U), A, C and G, respectively, of the other sequence. RNA sequences can also include complementary

G=U or U=G base pairs. Thus, two sequences need not have perfect homology to be "complementary" under the invention.

[0143] As used herein, the term "free energy" refers to the Gibbs free energy (ΔG), reflecting the thermodynamic potential that measures the hybridization reaction between a given oligonucleotide and its DNA or RNA complement, a given DNA sequence, such as of a promoter, e.g., of RNA II, and one or more proteins, such as sigma factor, RNA polymerase, or both.

[0144] According to another aspect, there is provided a computer program product comprising a non-transitory computer-readable storage medium having program code embodied thereon.

[0145] In some embodiments, the program code being executable by at least one hardware processor to: (a) receive a sequence of a plasmid; (b) receive a list of interactions associated with the origin of replication (*ori*) of the plasmid; (c) calculate mutations within at least one sequence associated with an *ori* associated interaction that: (i) increases the affinity, reduces the free energy, increases the abundance, or any combination thereof, of ori associated interactions supporting high plasmid copy number of the plasmid; (ii) decreases the affinity, increases the free energy, reduces the abundance, or any combination thereof, of ori associated interactions suppressing high plasmid copy number of the plasmid; or both (i) and (ii); and (d) provide an output (sequence of a) plasmid characterized by a maximized copy number, an optimized copy number, or a predetermined copy number, comprising at least one sequence associated with an *ori* associated interaction comprising at least one calculated mutation.

[0146] In some embodiments, the ori associated interaction is selected from: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – repressor of primer (rop) complexation, RNA I – tRNA complexation, RNA II – tRNA complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, or any combination thereof.

[0147] In some embodiments, the computer program product is for maximizing plasmid copy number. In some embodiments, the computer program product is for optimizing a copy number of a plasmid to host cell. In some embodiments, the computer program product is for obtaining a pre-determined copy number of a plasmid in a host cell for a user in need thereof.

[0148] In some embodiments, the computer program product calculates mutation within at least one sequence associated with ori associated interactions in the plasmid. In some embodiments, the computer program product provides an output (sequence of a) plasmid characterized by a maximized copy number, an optimized copy number, or a predetermined copy number, comprising at least one sequence associated with an *ori* associated interaction comprising at least one calculated mutation.

[0149] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0150] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0151] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention.

[0152] These computer readable program instructions may be provided to a processor of a general-purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[0153] Embodiments may comprise a computer program that embodies the functions described and illustrated herein, wherein the computer program is implemented in a computer system that comprises instructions stored in a machine-readable medium and a processor that executes the instructions. However, it should be apparent that there could be many different ways of implementing embodiments in computer programming, and the

embodiments should not be construed as limited to any one set of computer program instructions. Further, a skilled programmer would be able to write such a computer program to implement one or more of the disclosed embodiments described herein. Therefore, disclosure of a particular set of program code instructions is not considered necessary for an adequate understanding of how to make and use embodiments. Further, those skilled in the art will appreciate that one or more aspects of embodiments described herein may be performed by hardware, software, or a combination thereof, as may be embodied in one or more computing systems. Moreover, any reference to an act being performed by a computer should not be construed as being performed by a single computer as more than one computer may perform the act.

General

[0154] As used herein, the term "about" when combined with a value refers to plus and minus 10% of the reference value. For example, a length of about 1,000 nanometers (nm) refers to a length of 1,000 nm±100 nm.

[0155] It is noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a polynucleotide" includes a plurality of such polynucleotides and reference to "the polypeptide" includes reference to one or more polypeptides and equivalents thereof known to those skilled in the art, and so forth. It is further noted that the claims may be drafted to exclude any optional element. As such, this statement is intended to serve as antecedent basis for use of such exclusive terminology as "solely," "only" and the like in connection with the recitation of claim elements, or use of a "negative" limitation.

[0156] In those instances where a convention analogous to "at least one of A, B, and C, etc." is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (e.g., "a system having at least one of A, B, and C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms, or both terms. For example, the phrase "A or B" will be understood to include the possibilities of "A" or "B" or "A and B."

[0157] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. All combinations of the embodiments pertaining to the invention are specifically embraced by the present invention and are disclosed herein just as if each and every combination was individually and explicitly disclosed. In addition, all sub-combinations of the various embodiments and elements thereof are also specifically embraced by the present invention and are disclosed herein just as if each and every such sub-combination was individually and explicitly disclosed herein.

[0158] Additional objects, advantages, and novel features of the present invention will become apparent to one ordinarily skilled in the art upon examination of the following examples, which are not intended to be limiting. Additionally, each of the various embodiments and aspects of the present invention as delineated hereinabove and as claimed in the claims section below finds experimental support in the following examples.

[0159] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination or as suitable in any other described embodiment of the invention. Certain features described in the context of various embodiments are not to be considered essential features of those embodiments, unless the embodiment is inoperative without those elements.

EXAMPLES

[0160] Generally, the nomenclature used herein and the laboratory procedures utilized in the present invention include molecular, biochemical, microbiological and recombinant DNA techniques. Such techniques are thoroughly explained in the literature. See, for example, "Molecular Cloning: A laboratory Manual" Sambrook et al., (1989); "Current Protocols in Molecular Biology" Volumes I-III Ausubel, R. M., ed. (1994); Ausubel et al., "Current Protocols in Molecular Biology", John Wiley and Sons, Baltimore, Maryland (1989); Perbal, "A Practical Guide to Molecular Cloning", John Wiley & Sons, New York (1988); Watson et al., "Recombinant DNA", Scientific American Books, New York; Birren et al. (eds) "Genome Analysis: A Laboratory Manual Series", Vols. 1-4, Cold Spring Harbor Laboratory

Press, New York (1998); methodologies as set forth in U.S. Pat. Nos. 4,666,828; 4,683,202; 4,801,531; 5,192,659 and 5,272,057; "Cell Biology: A Laboratory Handbook", Volumes I-III Cellis, J. E., ed. (1994); "Culture of Animal Cells - A Manual of Basic Technique" by Freshney, Wiley-Liss, N. Y. (1994), Third Edition; "Current Protocols in Immunology" Volumes I-III Coligan J. E., ed. (1994); Stites et al. (eds), "Basic and Clinical Immunology" (8th Edition), Appleton & Lange, Norwalk, CT (1994); Mishell and Shiigi (eds), "Strategies for Protein Purification and Characterization - A Laboratory Course Manual" CSHL Press (1996); all of which are incorporated by reference. Other general references are provided throughout this document.

EXAMPLE 1

RNA II and RNA I Promoter sequence affects plasmid copy number

[0161] In 2022, Rouches et al., generated a library of pUC19 plasmids, each containing a different promoter sequence for either RNA II or RNA I. Notably, each plasmid only contains variations in the promoter regions of either RNA II or RNA I. The library was subsequently transformed into *E. coli* to measure the copy number of each plasmid. To further analyze the data, the inventors created a position-specific scoring matrix (PSSM) for the top and bottom 20% of high and low copy number plasmids.

Models

Copy Number Predictor

[0162] The inventors are building a copy number predictor based on a plasmid sequence. The predictor is based on the XGBoost model. The model gets sequence-related features for RNAI, RNAII promoters, including gaps, motifs, promotes strength, MIC, pssm, one hot encoding, and more. The inventors aim at extended the model also to RNAI and RNAII sequences and model their interactions.

Plasmid Sequence Generator

[0163] By understanding the importance of each of the features to the predictor, the inventors aim at reverse engineering the system to create different sequences that will correspond to a given plasmid copy number.

Modeling Metabolic Burden

[0164] The inventors aim to model the metabolic burden of the plasmids on the host cell and the effect on gene expression and bioproduction. This can be achieved by building a whole-

cell simulation with a pool of ribosomes under different conditions that translates plasmid genes.

Modeling Plasmid Stability and Variability

[0165] The inventors aim to model plasmid loss over generations and the variability of plasmid copy number in different cell populations.

Generalizing the Models

[0166] By understanding ColE1-like plasmids, the inventors aim to generalize the current model to broader plasmid families.

Abiotic Factors

[0167] The inventors aim to gain new insights to optimize the system, including the effect of different abiotic factors on plasmid copy number. For example, different abiotic factors can affect plasmid supercoiling, which is known to affect plasmid copy number.

Modeling Multi-plasmid Synthetic Biology (SynBio) Systems

[0168] The current goal is to develop computational models that can simulate multi-plasmid SynBio systems, where each plasmid in use has a different copy number. By incorporating the different copy numbers of each plasmid, models will enable further investigation on how the expression of genes on each plasmid is affected and how this impacts the overall behavior of the system.

Methods

Current Model Description

Data

[0169] The data from [Rouches, M.V., Xu, Y., Cortes, L.B.G., et al. (2022)] contains 366 inhibitory RNA promoter variants and 833 priming RNA promoter variants with their corresponding copy number.

[0170] The variants were edited at positions (-33,-30), (-11,-8), and +1 for priming RNA and (-33,-30), (-10,-7), and +1 for inhibitory RNA.

[0171] The original sequence for priming RNA is: TTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTT (SEQ ID NO: 1), and for inhibitory RNA is: TTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGA (SEQ ID NO: 2).

[0172] The data contains additional features such as growth rate and predicted promoter strength (KbT).

[0173] RNA polymerase recognizes DNA sequences through the promoter. The promoter identification is mediated by the subunit of RNA polymerase σ factor in bacteria. The housekeeping σ70 factor of *Escherichia coli* recognizes two DNA sequence elements '−10' hexamer (5′-TATAAT-3′) located at positions -7 to-12 and '−35' hexamer (5′-TTGACA-3′) located at positions -30 to -35. (Singh et al., 2011) According to the data, the RNAp and RNAi promoters were mutated at 9 positions. It is assumed that these positions were changed because they correspond to -10 and -35 hexamers.

Features

PSSM

[0174] The promoter sequences of priming RNA and inhibitory RNA corresponding to the top and bottom 20% copy number values were used to compute a position-specific scoring matrix (PSSM) for differentiating between high and low copy number plasmids.

[0175] The PSSM score, calculated with the 20% high copy number only, was added as a feature called 'pssm_score'.

Motifs

[0176] Motifs files were downloaded from (1) DPINTERACT database which is a comprehensive library of binding sites for E. coli DNA-binding proteins containing a set of 68 motifs, between 10 and 50 in width (average width 24.5), and (2) SwissRegulon *E.coli* database which is a database containing genome-wide annotations of regulatory motifs, promoters and TF binding sites (TFBSs) in promoter regions across model organisms, contains a set of 97 motifs, between 6 and 30 in width (average width 20.1).

[0177] The PV of each motif was calculated with FIMO module, and sequences with PV greater than $1 \cdot e^{-3}$ were replaced with 1.

Promoter strength

[0178] The promoter strength was calculated using an energy matrix [K_BT] taken from journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002811.

[0179] The matrix covers base pairs [-41:-1] where 0 denotes the transcription start site. Each row corresponds to a given position; each column corresponds to a value for that base pair.

[0180] The promoter strength was calculated for the entire sequence and for the two edited zones of the priming and inhibitory RNAs.

[0181] One hot encoding features: each promoter sequence was split into 144 (=36×4) columns so that each column represents the index and nucleotide combination, and the value is set to 1 if it matches the current sequence and 0 otherwise.

[0182] Other features: Adapted from "PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences", including:

Table 1. Features of DNA, RNA, and protein sequences

Feature name	Number of features	Feature Description
zCurve	3 features	$x\_axis = (\Sigma A + \Sigma G) - (\Sigma C + \Sigma T)$ $y\_axis = (\Sigma A + \Sigma C) - (\Sigma G + \Sigma T)$ $z\_axis = (\Sigma A + \Sigma T) - (\Sigma G + \Sigma C)$
gcContent	1 feature	$\frac{\Sigma G + \Sigma C}{\Sigma A + \Sigma C + \Sigma G + \Sigma T} * 100\%$
ATGC ratio	1 feature	$\frac{\Sigma A + \Sigma T}{\Sigma G + \Sigma C}$
Cumulative Skew	2 features	$GC\ Skew = \frac{\Sigma G - \Sigma C}{\Sigma G + \Sigma C}$ $AT\ Skew = \frac{\Sigma A - \Sigma T}{\Sigma A + \Sigma T}$
Pseudo KNC	when –kGap = 2, 20 features	Features will be numbers of A, C, G, T, AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT of the whole sequence of DNA respectively.
monoMonoKGap	when –kGap = 2, 32 features	Features will be numbers of A_A, A_C, A_G, A_T, C_A, C_C, C_G, C_T, G_A, G_C, G_G, G_T, T_A, T_C, T_G, T_T,

		A__A, A__C, A__G, A__T, C__A, C__C, C__G, C__T, G__A, G__C, G__G, G__T, T__A, T__C, T__G, T__T of the whole sequence of DNA respectively.
monoDiKGap	when –kGap = 2, 128 features	Feature structure will be X_XX, and X__XX of the whole sequence of DNA respectively.
monoTriKGap	when –kGap = 2, 512 features	Feature structure will be X_XXX, and X__XXX of the whole sequence of DNA respectively.
diMonoKGap	when –kGap = 2, 128 features	Feature structure will be XX_X, and XX__X of the whole sequence of DNA respectively.
diDiKGap	when –kGap = 2, 512 features	Feature structure will be XX_XX, and XX__XX of the whole sequence of DNA respectively.
diTriKGap	when –kGap = 2, 2048 features	Feature structure will be XX_XXX, and XX__XXX of the whole sequence of DNA respectively.
triMonoKGap	when –kGap = 2, 512 features	Feature structure will be XXX_X, and XXX__X of the whole sequence of DNA respectively.
triDiKGap	when –kGap = 2, 2048 features	Feature structure will be XXX_XX, and XXX__XX of the whole sequence of DNA respectively.

Features correlation with copy number

[0183] The inventors calculated correlations for every feature with the copy number and saved it as a df/excel file.

[0184] Correlations that were calculated - Pearson + pv, Spearman + pv, MIC + pv (MIC pv was calculated by doing 1000 permutations for each feature and the copy number, looking for the percentage of permutations that get equal or higher MIC than the original feature-copy number correlation). The permutations' MIC score was also saved into a df/excel file.

[0185] MIC (maximal information coefficient) is a measure of the strength of the functional and non-functional relationships association between two variables.

[0186] Spearman assesses how well the relationship between two variables can be described using a monotonic function, and Pearson is a measure of linear correlation between two sets of data.

Data preparation

[0187] Numeric features were scaled using a standard scaler, and zero variance features were removed.

All the features were calculated for the original promoter sequences, and the shared features matrix was concatenated as follows:

Features names	
Inhibitory RNA variants (366)	Priming RNA original sequence (×366)
Inhibitory RNA original sequence (×833)	Priming RNA variants (833)

Where the features of the original sequences were duplicated as the number of the other RNA variants.

Models

[0188] The data was split into a 15% test set and 85% train set.

Least Absolute Shrinkage and Selection Operator (Lasso)

[0189] Lasso is a linear regression model that performs feature selection and regularization, making it a powerful tool for predicting outcomes in high-dimensional datasets.

[0190] The lasso model uses L1 regularization, which adds a penalty term equal to the absolute value of the coefficients of the regression parameters in order to shrink the coefficients of less important features to zero. This results in a more parsimonious model that only includes the most relevant features in the final prediction while also improving the model's performance by reducing overfitting.

[0191] The optimization objective for Lasso is:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where the regression coefficients are listed as beta_i. The default lambda regulation term is 1.

[0192] Lasso regression can be used for both classification and regression tasks, and is particularly useful in cases where the number of features is much larger than the number of observations. It has been widely used in applications such as gene expression analysis, image processing, and financial modeling, among others.

XGBoost

[0193] XGBoost stands for Extreme Gradient Boosting. This algorithm is based on the gradient boosting trees algorithm but improves its speed and has been proven to work well even on sparse data.

[0194] The loss function used is: mean squared error. The algorithm starts by guessing 0.5 as an answer, then it calculates the difference between the real training data results and the guess (from now on called residuals). The following step will be to build a tree that tries to minimize the residuals. Each tree starts from a single leaf, the leaves containing all of the residuals. In addition, each leaf will get a Similarity Score: Similarity Score=(sum of residuals in the leaf)^2/(num of residuals + regularization value). Then the inventors start splitting the tree based on one feature at a time, and for each feature, the training data is split into quantiles, and the leaf is split for each option, the inventors choose the best-split feature and value that results in the highest Gain. The gain of each split possible is calculated by the following equation: Gain =(left children similarity score + right children similarity score) - parent similarity score.

[0195] A tree has a depth limit, and when one reaches this limit, a new tree that again tries to minimize the residuals calculated by the last tree is stared, in which the output of the matching leaf is: (sum of the residuals)/(num of residuals + regularization value). Then, the inventors add the results of each tree multiplied by a learning rate. The inventors keep building trees until the residuals are small enough or the maximum number of trees is reached.

Hyperparameters tuning

XGBoost

[0196] A converging random search approach was used in order to find the best hyperparameters for both RNAi and RNAp. The number of trees, maximum depth of each tree, learning rate, gamma, and colsample bytree were optimized.

Lasso model with Cross Validation (LassoCV)

[0197] LassoVC was used in order to determine the optimal alpha value (listed as lambda in the optimization objective):

[0198] Specify a range of alpha values to be evaluated.

[0199] Divide the data into k-folds for cross-validation.

[0200] For each alpha value, fit a Lasso regression model to the training data and evaluate its performance on the validation data using some metric (e.g., mean squared error).

[0201] Average the performance metrics over the k-folds for each alpha value.

[0202] Select the alpha value that gives the best performance (i.e., lowest mean squared error or highest R-squared).

[0203] Optionally, refit the Lasso regression model using the selected alpha value and all of the available data.

Evaluation

[0204] In order to evaluate the model's performance, an R-squared score was used. It represents the proportion of variance (of Y) that has been explained by the independent variables in the model. It provides an indication of goodness of fit and, therefore a measure of how well unseen samples are likely to be predicted by the model through the proportion of explained variance.

[0205] The estimated R-squared is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$.

Best Features Analysis

[0206] The features analysis is made for the most correlated features and the most important features that were used in the models.

Most correlated features

[0207] Analyzing the features-copy number correlations, some features showed noticeable results that are summarized in the following tables:

Table 2. Priming RNA high correlated features

Feature	Pearson|pv	Spearman|pv	MIC|pv
pssm_score	0.312|10^{-20}	0.396|10^{-24}	0.216|~0
Promoter Strength	0.04|0.25	0.06|0.1	0.234|~0
rpoD17	-0.22|10^{-10}	-0.05|0.13	0.4|~0
AA_count	0.24|10^{-12}	0.304|10^{-19}	0.18|~0

Inhibitory RNA high correlated features

Feature	Pearson|pv	Spearman|pv	MIC|pv
pssm_score	0.303|10^{-9}	10^{-9}|0.32	0.22|0.003
Promoter Strength	-0.1|0.04	-0.16|0.002	0.301|~0
rpoD18	-0.07|0.15	-0.12|0.02	0.337|~0
TT_count	-0.375|10^{-10}	-0.384|10^{-14}	0.331|~0

[0208] To validate the model results, the inventors perform various molecular biology techniques and conduct several experiments, as elaborated herein below.

[0209] Modification of the col-E1 Origin of Replication: To generate a collection of Col-E1 plasmids, the existing origin of replication is enzymatically digested using various high-fidelity restriction enzymes (described in the methods section), followed by ligation using high-fidelity DNA ligase, of an alternate origin of replication sequence. For the synthesis of custom-designed origins of replication, the inventors procure the sequences from a manufacturer affiliated with the International Gene Synthesis Consortium (IGSC).

[0210] Adding a genetic marker to the plasmid. To help with various assays and experiments, chromoprotein is added to the plasmid. The inventors use 3 different chromoproteins throughout the current experiments, all of which are official parts from the

iGEM registry (BBa_K1692032, BBa_K1033917, BBa_K1033926). Using restriction enzymes and DNA ligase, one chromoprotein is added to the plasmid.

[0211] Quantitative PCR (qPCR) is used to measure the copy number of a given plasmid. This is achieved by measuring the ratio between two single-copy genes (one found in the chromosome and the other on the plasmid itself). This way plasmid copy number can be calculated using the following equation: [plasmid copy number] = [copy of plasmid gene] / [copy of chromosomal gene].

Wet-lab Methods

Strain, media, and growth conditions

[0212] To prepare Luria-Bertani (LB) medium, 20 g of LB powder was added to 1 liter of water, and the mixture was autoclaved to ensure sterility. Once the solution cooled down to 60 °C, antibiotics were added in a 1:1,000 ratio to create a selective medium.

[0213] For LB agar plate preparation, 20 g of LB powder and 20 g of agar were added to 1 liter of water. The mixture was thoroughly mixed using a stirrer and then autoclaved. After the liquid reached a temperature of 60 °C, antibiotics were added in a 1:1,000 ratio. A flame was lit, and the LB agar solution was carefully poured into sterile Petri dishes to solidify.

[0214] For patch plating, a sterile toothpick was used to carefully lift off a bacterial colony from a culture plate. The colony was then transferred onto the surface of a fresh agar plate. The plate was subsequently incubated overnight at 37 °C to allow for colony growth and expansion.

[0215] To prepare a starter culture, a sterile toothpick was used to lift off a bacterial colony from a culture plate of *E. coli* DH5-alpha strain. The colony was inserted into a starter culture containing 3 to 5 ml of growth media. The starter culture was then incubated at 37 °C with shaking at 250-300 rpm overnight to promote bacterial growth and achieve a sufficient cell density for subsequent experiments.

pUC19 transformation to E. coli DH5-alpha

[0216] To transform the pUC19 plasmid into *E. coli* DH5-alpha, a 10 µl aliquot of double-distilled water (DDW) was combined with the plasmid powder. Next, 3 µl of the resulting plasmid solution was added to competent DH5-alpha bacteria and thoroughly mixed. The mixture was then subjected to a 10-minute incubation in a -20 °C freezer to enhance the transformation efficiency. Subsequently, the bacteria were transferred to a preheated 42 °C water bath for 2 minutes. To facilitate bacterial growth and recovery, 1 ml of LB medium

was added to the suspension, which was then incubated at 37 °C for 1 hour. Following incubation, the bacterial culture was centrifuged at maximum speed for 2 minutes using a pull-down mode, resulting in cell pellet formation. The supernatant was carefully removed, and the pellet was resuspended by adding 1 ml of DDW and pipetting thoroughly. Finally, the transformed bacteria were plated onto an LB agar plate supplemented with an appropriate antibiotic. The plate was incubated overnight at 37 °C to enable the formation of colonies, indicative of successful transformation.

EXAMPLE 2

Materials and Methods

RNAp and RNAi Promoters variants library

[0217] The inventors analyzed the pUC19 plasmid library by Rouches et al., consisting of distinct promoter variants for RNAp or RNAi. The mutations are specifically confined to the promoter regions of RNAp or RNAi. The data contains 366 RNAi promoter variants and 833 RNAp promoter variants with their corresponding copy number. The variants were edited at positions (-33,-30), (-11,-8), and +1 for RNAp and (-33,-30), (-10,-7) , and +1 for RNAi (**Fig. 25F**). The inventors divided the data into 3 sets, 15% test set,15% validation set, and 70% train set. During this process, the inventors ensured that the copy number distribution in each set remained consistent with the original dataset.

Sequence features for machine learning

[0218] The promoter's sequences can't be used directly as input for the models, so, about 6,000 features were engineered to represent the sequence properties that correspond to transcriptions in prokaryotes.

[0219] The analyzed features were:

- **Motifs:** Motifs files were downloaded from the DPINTERACT database and the SwissRegulon *E.coli* database. The score and P-value of each motif were calculated with the FIMO module, and sequences with no matching motif were replaced with a 0 score, and p-value of 1. (330 features)
- **PSSM score:** The position-specific scoring matrix (PSSM) score of each sequence is calculated based on the 20% high copy number matrix, as follows:

$$PSSM\ score = \log_2(\frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k))$$

where $i \in (1, \ldots, N)$ $with\ N\ as\ the\ sequence\ length, j \in (A, C, T, G)$, and $I(a = k)$ is an indicator function where $I(a = k)$ is 1 if $a = k$ and 0 otherwise. (1 feature)

- **Promoter strength:** The promoter strength was calculated as described in Brewster et al. (3 features)
- **One hot encoding features:** each promoter sequence was split into 144 (=36×4) columns so that each column represents the index and nucleotide combination, and the value is set to 1 if it matches the current sequence and 0 otherwise. (144 features)
- **Nucleotide counts features:** nucleotide counts features from an article by Muhammod et al. (5945 features)
- **dG total and TX rate:** these features represent the interactions between the promoter sequence and RNA polymerase with or without sigma factor and the transcription rate, as described in LaFleur, T.L. et al. (3 features)
- **Entropy:** The promoter sequence entropy was calculated as follows:

$$\boldsymbol{S} = -\frac{\sum_{i=0}^{n} p_i log_2(p_i)}{2}$$

Where p_i is the probability of each nucleotide to appear in each index i. (1 feature)

- **De Novo motif detection:** motifs that appear with higher rates in groups of promoters that have high copy number against a reference group of promoters with low copy number, and vice versa, were detected using HOMER (Hypergeometric Optimization of Motif EnRichment). (29 features).

Secondary Structure Features for RNAi model

[0220] The model based on RNAi promoters, using the above features, produced less successful results than the model based on RNAp promoters. To improve this, the inventors added features describing the RNAp secondary structure, considering the complementarity between RNAi sequences and RNAp promoters. Mutations in RNAi's promoter can impact RNAp's structure, leading to alterations in its secondary structure. The inventors investigated the replication mechanism of the ColE1 plasmid family, specifically pUC19, where RNAp's persistence on the DNA template is vital for initiating replication. A key interaction involves a G-rich loop in the RNAp transcript binding to a C-rich stretch on the DNA, affecting RNAp conformations and the formation of RNA-DNA hybrids necessary for DNA synthesis.

[0221] RNAp contains three binding sites: alpha, beta, and gamma. The RNAp transcript undergoes structural changes during transcription, forming three stem-loop domains and later an extended stem-loop. This folding is influenced by the pairing of alpha-beta or beta-

gamma sequences, which determines the transcript's ability to form the hybrid for DNA synthesis. RNAi's structure, with three stem loops and an unpaired region, binds to RNAp, influencing its folding and favoring beta-gamma pairing. This interaction affects primer formation, essential for replication. The current study used RNAi and RNAp sequences from the pUC19 plasmid, analyzing RNAi promoter mutations. The inventors focused on nine specific mutation positions in the RNAi promoter, corresponding to key points in the RNAp sequence. The inventors employed the ViennaRNA package, particularly RNAfold and RNAeval, to calculate RNAp's secondary structure, considering the impact of RNAi promoter mutations. This approach allowed the inventors to generate features describing RNAp's secondary structure at different transcription stages, enhancing the current RNAi-based model's accuracy. The features (additional ~25,000 features to the features detailed above) are calculated for a range of lengths of RNAp in order to describe its structure during transcription.

Table 3. RNAp secondary structure features

Feature Group Name	**Number of Features**	**Feature Description**
Base pairs match ratio	22 features	We calculate the following ratio: $\frac{number\ of\ paired\ bases\ in\ specific\ area}{number\ of\ paired\ bases\ in\ pUC19\ sequence\ in\ this\ area}$ We get the paired bases using RNAFold's output. **Specific features examples:** • alpha_beta_match_seq_end_130 – it means the inventors used the first 130 bases of the RNA sequences and checked alpha-beta area. • alpha_beta_extended_match_seq_end_400 – it means the inventors used the first 400 bases of the RNA sequences and checked extended alpha-beta area. Note: the inventors used varied lengths of RNA because during its formation it changes shapes, and it can be interesting to see the difference in structures in different stages.
Base pairing probabilities	22,877 features	RNAFold can output in an additional file the probability of each nucleotide to pair with a nearby nucleotide. The inventors parsed the file and took each such probability as a feature. **Specific feature example:**

		seq_end_400_375_385 – it means the inventors used the first 400 bases of the RNA sequences and checked the probability of the nucleotide in index 375 to pair with nucleotide in index 385.
MFE and Centroid comparison	1,730 features	RNAFold can predict the secondary structure of a given RNA sequence using 'MFE' and 'centroid'. The inventors compared how similar the two predictions are. RNAFold's documentation states that the more similar they are the prediction is more reliable, so the inventors thought it could be a good feature as well. The features are whether base pairs in different indexes exist in both predicted structures. **Specific feature example:** seq_end_180 mfe == centroid 118_141 – it means that for each RNA sequencethe inventors used the first 180 bases as an input for RNAFold and got two predicted secondary structures: MFE and Centorid, for each base pair the inventors check if it appears in both forms, in this case the inventors check if the pair combined of nucleotides at indexes 118 and 141 appears in both forms.
Stem loops MFE	20 features	Another output the inventors can extract is the minimal free energy (MFE) of specific areas in the secondary structure (using RNAFold and then RNAeval – another tool in ViennaRNA package). **Specific feature example:** mfe_seq_end_300_sl_IV – it means the inventors used the first 300 bases of the RNA sequences and calculated MFE for stem loop IV.
Base pairing of specific areas	6 features	We checked how many bases are paired in specific areas in the predicted secondary structure of the RNA sequences. **Specific feature example:** base_pairing_mfe_alpha_range – it means the inventors checked in the MFE predicted structure how many bases are paired in the alpha area.
RNA form representation	36 features	We tried to represent the form using some basic characteristics of the secondary structure output. The inventors calculated: • Number of unpaired bases • Number of 'arcs' – a term the inventors use for continuous sequence of unpaired bases that form an open ended circular shape. • Number of 'interior loops' – a term the inventors use for continuous sequences of unpaired bases that form an interior loop shape (similar to the tip of RNA hairpin)

		Specific features examples: • unpaired_cnt_seq_end_82 - counts number of bases that don't pair with other bases in the first 82 bases of the RNA sequences • arc_cnt_seq_end_102 – counts how many open ended circles the inventors find in the first 102 bases of the RNA sequences. interior_loop_cnt_seq_end_172 – counts how many interior loops the inventors find in the first 172 bases of the RNA sequences.
RNA form in Window	28 features	These features include: mfe, unpaired bases count, 'arcs' count and 'interior loops' count but in local windows of size 70. Since the bonds of RNA's secondary structure are usually local it makes sense to check these features in small windows. **Specific features examples:** • mfe_window_130_200 – calculates MFE of RNA between index 130 to 200. • unpaired_cnt_window_120_190 – counts number of bases that don't pair with other bases in the secondary structure of RNA between index 120 to 190. • arc_cnt_window_110_180 – counts how many open ended circles the inventors find in the window (here between index 110 to 180). interior_loop_cnt_window_100_170 – counts how many interior loops the inventors find in the window (here between 100 to 170).

Model training and evaluation

[0222] The inventors explored various models, including Neural Networks, Linear Regression models, Random Forest, and Ensemble Learning models, while tuning their hyperparameters with Optuna (more details below). The inventors assessed the Spearman and Pearson correlations on both the training and validation sets. Notably, XGBoost, CatBoost, Light GBM, and Random Forest demonstrated superior performance. Consequently, the inventors advanced to the subsequent stages, feature selection and hyperparameter optimization, with these selected models.

[0223] For hyperparameter optimization, the best set of hyperparameters that fit the data, and the model were selected using Optuna. Optuna automates the search for optimal hyperparameter configurations using Bayesian optimization, where each trial employs a

different set of hyperparameter values and evaluates an evaluation score on the validation set.

[0224] The feature selection was performed in two steps: (1) Mutual information-based feature selection: the inventors calculated the mutual information (MI) between each feature and the copy number, selecting those with MI values greater than the mean MI as potential candidates. Subsequently, the inventors assessed the MI between the remaining features, identifying pairs of features exhibiting strong correlations within the top percentile. For each pair of correlated features, the inventors chose the one displaying a higher MI with the copy number, ensuring the retention of the most informative attributes. (2) Boruta-Shap based feature selection: Boruta-SHAP is a feature selection algorithm that combines the Boruta algorithm with SHAP values to select important features: (a) Boruta is a tree-based feature selection algorithm that works by creating random shadow features (noise) and comparing the importance of the original features to the importance of the shadow features. Features that are more important than the shadow features are kept, while features that are not more important than the shadow features are removed. (b) SHAP values are a way to measure the individual impact of each feature on the prediction of a model. SHAP values are calculated by considering all possible combinations of features and their impact on the prediction.

[0225] Using Boruta-SHAP works by using SHAP values to rank the importance of the features and the shallow features.

[0226] Ultimately, the CatBoost model was employed for predicting the final outcomes of the test set in the case of RNAi. In the context of RNAp, a voting model was adopted, wherein the final prediction is determined by averaging the predictions from the ensemble learning model, CatBoost, and XGBoost. This voting model approach not only enhances performance but also mitigates noise and overfitting. The model performance was evaluated using Pearson and Spearman correlations, R-squared score, and mean absolute error (MAE). The model development workflow is demonstrated in **Fig. 29C**.

Model explainability

[0227] The most important features of each model were analyzed and visualized to summarize their main characteristics. XGBoost employs a distinct method for assessing feature importance, known as "gain". This method calculates the contribution of each feature to the model by considering the improvement in accuracy brought about by splitting on that feature, averaged over all trees. The gain is a measure of the average contribution of a feature to the model's prediction performance: the higher the gain, the more important the feature.

On the other hand, the CatBoost model uses a combination of the SHAP values and internal algorithms to rank features, providing a measure of the impact each feature has on the variation in the predicted outcome. This model-agnostic approach is beneficial as it offers a consistent way to interpret feature importance across different types of models. The inventors visualized the features' distribution by their corresponding copy number using violin plots for discrete features and scatter histograms for continuous features, in order to identify patterns, trends, and outliers. SHAP (Shapley Additive exPlanations), an interpretable AI approach leveraging Shapley values, is a model-agnostic technique designed to elucidate the impact of individual features on model predictions. This method assigns a Shapley value to each feature, providing a quantitative measure of its contribution to the model's output for a specific instance. Features with high absolute Shapley values are indicative of having a substantial influence on predictions.

[0228] In the current analysis, the inventors employed SHAP to examine the selected features and assess their influence on the decision-making process of the models. This approach allows the inventors to identify the critical features and understand how specific values of each feature contribute to the prediction of a high copy number. This interpretability framework was particularly valuable as the inventors sought to enhance the comprehension of the colE1 replication system's mechanisms. By delving into the model structure and leveraging SHAP, the inventors aimed to unravel the intricate dynamics underlying the replication system and glean insights into its functional intricacies.

Origin of replication replacement

[0229] The modified origin of replications were synthesized based on the tool's output. The original pUC19 plasmid was amplified using PCR to introduce new flanking regions around the ORI, then a cut site was targeted with a restriction enzyme, to cut the original ORI. The modified ORI was inserted using homologues joining with NEBuilder® HiFi DNA Assembly kit to the pUC19 backbone. The altered plasmids were introduced into *E. coli* DH5 alpha through a chemical transformation and verified using sanger sequencing. The process is demonstrated in **Fig. 29**.

Real-time qPCR using SYBR Green

[0230] All qPCR reactions were performed in technical replicates on an applied biosystems Real-Time PCR Detection instrument using the following cycling conditions for all reactions: 95 °C for 10 min, and 40 cycles of 3 sec at 95 °C and 30 sec at 60 °C. Primers

were designed for the plasmid and the bacteria and were tested with melt-curve (detailed in chart below).

Table 4. Primers which were used in the study.

Target	Primer Sequence	Position	Length	Melting temperature
PlasmidF'	cgtgtcgcccttattcc ctt (SEQ ID NO: 3)	ampicillin resistance	20 bp	56.5 C°
PlasmidR'	cccaactgatcttcagc a (SEQ ID NO: 4)	ampicillin resistance	18 bp	50.9 C°
Chromosome F'	cgagaaactggcgatc ctta (SEQ ID NO: 5)	E. coli dxs gene	20 bp	60 C°
Chromosome R'	cttcatcaagcggttta cac (SEQ ID NO: 6)	E. coli dxs gene	20 bp	60 C°

[0231] To perform the qPCR, a colony was picked and mixed in DDW into a dilution of 10^{-6}. A reaction mix was prepared with: 10 µl fast SYBR Green Master Mix, 2 µl of each primer and 5 µl DDW. Nineteen (19) µl of the reaction mix was transferred into qPCR plate, and 1 µL of the colony sample was added. The PCR plate went through spin down for 1 minute at 800 rpm prior to its loading in the instrument.

Plasmid copy number determination

[0232] A standard curve was created for the quantification of the targeted DNA of the plasmid and the bacteria. By finding the amplifications efficiency (E), the plasmids copy number were calculated using the equation: $PCN = (Ec)^{Ctc}/(Ep)^{Ctp}$.

Plasmid copy number transformation

[0233] The biological measurements in Rouches et al. were of a relative copy number, meaning based on growth parameters. These measurements were highly correlated (in terms of ranking) with copy number; hence they used a polynomial fit based on ddPCR measurements to convert their measurements into final copy number. With this

understanding, and with the additional qPCR biological copy number measurements, the inventors performed a mathematical transformation to the data used to train our model for improving the model performance.

[0234] Starting from the relative copy number, the inventors improved the mathematical fitting of the data to the measured copy number. For this purpose, the inventors added 30% more data points to the sequences used for the fitting while saving aside an equal number of data points for validation. Analyzing the data, the inventors observed a power log behavior connecting the measured copy number to the relative copy number; therefore, the inventors decided to perform a linear regression of the logarithmic values to transform the relative copy number to copy number estimations.

Results

Models Results

[0235] A model able to predict the copy number from the ORI sequence will have practical implementations, allowing researchers to design new plasmids with a specific copy number tailored to their individual needs. Creating it as an explainable model will also allow the inventors to gain new insights into the biology of Col-E1-like plasmids replication, insights that can be capitalized upon in future research. The inventors used a database of tagged ORI sequences (meaning different ORI variants each with a corresponding copy number) that was recently published, wherein two libraries of ORI variants, one with mutations in the RNAp promoter region of the ORI and the second with changes in the RNAi promoter region were created. The inventors therefore chose to create two separate models, for each of the datasets. Setting out to build the framework for our model the inventors first set to find the best fitting modal for the task. To test different models, the inventors split the data into training, test, and validation sets (70%, 15%, and 15% respectively) and verified that all sets feature a copy number distribution similar to that of the entire dataset.

[0236] In order to use the promoter sequences for machine learning models, which cannot process raw sequences, around 6,000 features were engineered to represent transcription-related properties in prokaryotes. These features include motif scores and P-values derived from known databases calculated using the FIMO module; PSSM scores that reflect the frequency of nucleotides in high copy number sequences; promoter strength; nucleotide counts; interaction and transcription rate measures; and *de novo* motif detection using HOMER and more. The features were selected to capture the complex biological factors that

influence transcription in prokaryotes, facilitating the development of predictive models (more details in the methods section).

[0237] The inventors trained 8 different models for RNAp and RNAi each, and tested their performance on the test set results are shown in **Fig. 30**. On the RNAp dataset, CatBoost and XGBoost both performed best, with similar results (see more details about these models in the supplementary). In order to enhance performance and mitigate noise and overfitting the inventors utilized a voting model using both of them, averaging their predictions (model performance shown in **Fig. 25**). On the RNAi set, CatBoost outperformed the rest of the models (results shown in **Figs. 25A-25B**). As can be seen, both models exhibited a very high correlation between the model predictions and the test set, demonstrating the ability of the model to predict the copy number of plasmids that were not used for its training, based on the RNAp or RNAi promoter sequences alone.

Model Explainability

[0238] By identifying which features of the ORI sequence are most influential in determining the copy number, the inventors could uncover new aspects of the sequence or structural elements that are critical for the replication process and gain insights into the biological mechanisms that regulate ColE1 replication. To do so, the inventors evaluated the features distribution and their relation to copy number, examined the impact of individual features on model predictions, and finally, use these insights to interpret the ColE1 replication system's mechanisms.

Features Distribution and Their Relation to Copy Number

[0239] The inventors ranked all the features for each separate model according to their importance to the model prediction, results shown in **Figs. 25D-25F**. The most important feature for RNAp was dG_total for CatBoost and rpoD16_score for XGBoost. dG_total is the difference in free energy between an unbound promoter and a promoter bound to a sigma factor + RNA polymerase. rpoD16_score is a numerical value representing the strength or likelihood of the presence of the rpoD16 motif, a specific bacterial sigma factor recognition site. Higher scores suggest a higher probability of successful biosynthesis. For RNAi CatBoost model, the most important features were seq_end_400_375_385 and TGCGCGTTTAAT (SEQ ID NO: 7)_denovo_high. RNAFold can provide the probability of each nucleotide pairing with a nearby nucleotide. seq_end_400_375_385 means the inventors used the first 400 bases of the RNA sequences and checked the probability of the nucleotide in index 375 to pair with the nucleotide in index 385. The sequence

TGCGCGTTTAAT (SEQ ID NO: 7) is a *de-novo* motif that appears with higher rates in groups of promoters that have high copy number, against a reference group of promoters with a low copy number.

[0240] The current feature distribution analysis aimed to understand the underlying patterns, trends, and outliers in the data. While analyzing the common and most important features that were used in each model, utilizing violin plots for discrete features and scatter histograms for continuous features, allowed the inventors to observe the distribution and prevalence of copy number variations across different categories. These plots highlighted diverse feature behaviors concerning copy number variations. **Figs. 26A-26B** display the selected features along with their respective copy numbers and the distribution patterns for these features. It displays the common features between the XGBoost and CatBoost models used in the RNAp model, as well as the two most significant features of the CatBoost model used in the RNAi model.

[0241] In the RNAp related features distribution graph (**Fig. 26A**), rpoD16_score demonstrates a pronounced peak, signifying a strong effect at a specific score on the copy number. Additionally, the plot for the feature dG_total indicates a spread of values with most data points concentrated around -3.4 to -2.7, implying a region of stability that could be critical for high copy numbers. For RNAi related features (**Fig. 26B**), the scatter plot for seq_end_400_375_385 shows a dense accumulation of points at the lower end of the scale, which may indicate a significant correlation with increased copy numbers. Conversely, the feature TGGCGGTTTAAT (SEQ ID NO: 8)_denovo_HIGH displays a spread of data points, with notable bars extending toward middle values, hinting at the possibility of distinct conditions or states influencing the copy number within this feature.

Impact of Individual Features on Model Predictions Using SHAP

[0242] SHAP is a tool used to interpret machine learning models by assigning each feature an importance value for a particular prediction. It is based on Shapley values from game theory, which allocates fair payoffs to players depending on their contribution to the total payout. SHAP helps to understand which features most strongly influence the model's predictions, revealing the impact of each feature on the predicted outcome. This insight allows the inventors to identify the features that are most significant in determining the copy number, thereby providing a deeper understanding of the underlying plasmid replication system. The SHAP analysis yielded a quantitative measure of each feature's influence on the model's predictions, spotlighting the features with high absolute Shapley values across both

RNAp and RNAi models. **Figs. 27A-27B** show the SHAP value for each feature chosen to train the models during the feature selection process. In the RNAp model, dG_total surfaced as the most influential feature, leading to significant positive impacts on predicted copy numbers. The strong positive influence of dG_total was particularly striking against the backdrop of more moderate effects from other features, positioning dG_total as a potentially critical determinant in the RNAp model's decision-making process. Within the RNAi model, *TGCGGTTTAAT* (SEQ ID NO: 9) *denovo_HIGH* was identified as having the highest mean absolute SHAP values, signifying its substantial role in the model's output.

Interpretation of the ColE1 Replication System's Mechanisms

[0243] Since dG_total appeared to be significant in the current analysis, the inventors continued to analyze it using a SHAP force plot. The dG_total feature represents the overall change in free energy (Gibbs free energy) between an unbound promoter and a promoter bound to a sigma factor + RNA polymerase which is associated with the plasmid replication process. A lower dG_total value indicates a more thermodynamically favorable biosynthesis reaction. The analysis of the ColE1 replication system mechanisms, particularly through the lens of the effect of the dG_total feature on the copy number, offers valuable insights into the plasmid replication process. The dG_total, as depicted in **Fig. 27**, plays a critical role in the replication efficiency, with a specific range appearing to be optimal. This optimal range, located between -3.4 to -2.7, corresponds to a more thermodynamically favorable reaction for plasmid replication, thereby indicating a threshold effect where certain energy levels are conducive to higher copy numbers. The presence of clustering within this range suggests a stable and efficient replication process, with dG_total values outside this range potentially leading to less favorable replication outcomes. This range represents the reaction's thermodynamic 'sweet spot' – a state of maximum favorability characterized by the stable interaction of molecular complexes, efficient regulatory mechanisms, environmental adaptation, and optimal resource utilization. In the context of the ColE1 replication system, understanding this optimal dG_total range is crucial for enhancing the efficiency and predictability of plasmid replication processes, which is of significant interest in genetic engineering and biotechnology applications.

PCN Measurements and Correlation to the Model Predictions

[0244] The inventors were successfully able to replace the origins of replications of a pUC19 plasmid with new modified sequences taken from the literature or generated by the current model using a high throughput method. This innovative approach allowed the inventors to

easily construct a library of plasmid sequences which played a crucial role in the verification and improvement of the current model. This procedure involves synthesizing modified origins of replication based on tool outputs, introducing them into the pUC19 plasmid via PCR amplification and homologous recombination, and then verifying the modified plasmids after transformation into *E. coli* using Sanger sequencing, confirming the presence of the intended mutations within the RNAp and RNAi promoters (see more details in **Fig. 28A** and in the methods section). Following the successful verification of mutations, the inventors proceeded to calculate the copy numbers of the newly constructed plasmids. To minimize potential errors stemming from DNA isolation procedures, the inventors adopted a colony qPCR approach to measure the plasmid copy numbers for the various constructs within our library. The results of this copy number analysis provided valuable insights into the replication dynamics of these engineered plasmids. Based on these findings, as anticipated, it is evident that substituting the promoter sequences at the origin of replication has an impact on the plasmid copy number.

[0245] The inventors used Rouches et al. ddPCR measurements, in addition to in-house qPCR measurements to improve the current model. As done in Rouches et al., the inventors also performed a mathematical transformation to the copy number data used to train the current model for improving the model performance (see more details in the methods section). The inventors achieved a high correlation on both the test set and the current biological results, as can be seen in **Fig. 28B**. This further verifies the reliability of the current model and the success of introducing a machine-learning model for copy number engineering and prediction.

Webtool for predicting Plasmid Copy Number

[0246] The inventors used the machine learning model results and developed a software tool (available at: pcn-gradient.vercel.app) that enables researchers to design and predict plasmid copy numbers simply and efficiently. The model's high performance allows the current tool to accurately predict the copy number of any plasmid, based on its sequence and the host strain in which it will be used. The tool offers several key features to enhance plasmid sequence manipulation and customization. As demonstrated in **Fig. 28**, users can input a plasmid sequence (FASTA or GenBank files) along with the desired copy number, and the tool generates a modified sequence with ORI adjustments to achieve the specified copy number. Its versatility extends to both modifying existing plasmids for a desired copy number and creating plasmids from scratch with the designated copy number. For tailored adjustments, the tool efficiently generates FASTA files, ensuring precision in achieving the

desired copy number. Additionally, users can request multiple copy numbers simultaneously, receiving dedicated sequences for each. The tool's output includes clear and concise plasmid maps, seamlessly integrated with annotations within the sequence, providing comprehensive information. Accessible on both computers and mobile devices, the tool caters to diverse user needs. Furthermore, developers can seamlessly integrate the tool into their projects using the provided API, enhancing its adaptability and accessibility.

[0247] This accomplishment contributes to the field of synthetic biology in a notable way. For the first time, a model has been successfully developed to predict plasmid copy number, solely based on the ORI sequence. This is particularly important because it addresses a critical challenge in genetic engineering, and researchers can now more accurately engineer plasmids for desired traits, since plasmids are essential tools in gene cloning, recombinant protein synthesis, genetic editing, complex genetic circuits, and many more applications.

Discussion

[0248] In this study, the inventors suggest a novel approach for dealing with the challenge of controlling the expression levels of genes: accurate model-based fine-tuning of plasmids copy number. The current approach can be used to modulate the exact protein levels in the cell. A high copy number can enhance protein yield but may also lead to metabolic burden and instability, whereas a low copy number may result in insufficient product formation.

[0249] Here the inventors suggest for the first time a model that can predict plasmid copy number based on the ORI sequence. Using the computational model, a library of plasmid sequences was constructed, and a colony qPCR method was used to measure copy numbers, validating the model's predictions and allowing for improvements. The current model can be used both for understanding the regulation of plasmid copy number and for engineering it.

[0250] The inventors developed two separate machine learning models to predict copy number from RNAp and RNAi promoter sequences using a database of tagged ORI sequences. The models demonstrated high correlation in predicting plasmid copy numbers not used for training the model. The inventors found that the sequence features related to transcription regulation of the RNA genes encoded in the ORI (e.g., dG_total and rpoD16_score) are the most important features in the current models. These features represent aspects related to transcription efficiency such as the free energy of promoter binding and the strength of sigma factor recognition sites. Based on the feature analyses, insights into the ColE1 replication system were gained: there is a specific range of "dG_total" values that will optimize the copy number where a too high or too low values

will result in lower copy numbers. Lower efficiency of interaction may be related to lower levels of RNAp and thus less replication and less plasmids; on the other hand, a very high "dG_total" value can result in a polymerase that is bound too strongly to the promotor and might have problems disengaging and starting replication, thus limiting copy number. Alternatively, it is possible that very high levels of RNAp due to very high interaction with the RNA polymerase is too burdensome to the cell, resulting in the end in low copy number as the cell negatively regulates the plasmid copy number using some unknown mechanism.

[0251] Interestingly, the current analysis suggests that the interactions between RNAp and RNAi within the ColE1 plasmid family, particularly in the pUC19 model, through their respective secondary structures play a pivotal role in the regulation of plasmid copy number. Specifically, the inclusion of secondary structure features into the current computational model has significantly enhanced its predictive accuracy. In the current study the inventors were able to quantify how mutations in the RNAi promoter impact the RNAp secondary structure, which is critical for designing strategies to optimize plasmid stability and efficiency in synthetic biology applications. Finally, the inventors developed a web-based software tool which offers a user-friendly platform for researchers to efficiently design, predict, and modify plasmid sequences, thereby streamlining research workflows and enhancing the precision of genetic manipulations.

[0252] While certain features of the invention have been illustrated and described herein, many modifications, substitutions, changes, and equivalents will now occur to those of ordinary skill in the art. It is, therefore, to be understood that the appended claims are intended to cover all such modifications and changes as fall within the true spirit of the invention.

CLAIMS

What is claimed:

1. A method for maximizing plasmid copy number in a host cell, the method comprising:

    a. modeling origin of replication (*ori*) associated interactions the affect plasmid copy number in said host cell;

    b. determining free energy, affinity, abundance, or any combination thereof, of said *ori* associated interactions; and

    c. based on said determined free energy, affinity, abundance, or any combination thereof of step (b) introducing to said plasmid any one of:

        i. at least one mutation increasing the affinity, reducing the free energy, increasing the abundance, or any combination thereof, of *ori* associated interactions supporting high plasmid copy number of said plasmid;

        ii. at least one mutation decreasing the affinity, increasing the free energy, reducing the abundance, or any combination thereof, of *ori* associated interactions suppressing high plasmid copy number of said plasmid; or

        iii. both (i) and (ii).

2. The method of claim 1, wherein said *ori* associated interactions are selected from the group consisting of: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – repressor of primer (rop) complexation, RNA I – tRNA complexation, RNA II – tRNA complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, and any combination thereof.

3. The method of claim 2, wherein said at least one mutation forms an RpoD sigma factor recognition site in RNA I of said *ori*.

4. The method of any one of claims 1 to 3, wherein said plasmid comprises a polynucleotide integrated therein.

5. A method for optimizing a copy number of a plasmid to host cell, the method comprising:

a. calculating an activity index for a host cell comprising a pre-determined copy number of said plasmid, wherein an activity index above or below a cut-off range is indicative of said pre-determined copy number of said plasmid being suboptimal for expression of a polynucleotide being integrated into said plasmid in said host cell; and

b. introducing to said plasmid at a pre-determined copy number determined as being suboptimal for expression of said polynucleotide in said host cell any one of:

i. at least one mutation increasing the affinity, reducing the free energy, increasing the abundance, or any combination thereof, of *ori* associated interactions supporting high plasmid copy number of said plasmid;

ii. at least one mutation decreasing the affinity, increasing the free energy, reducing the abundance, or any combination thereof, of *ori* associated interactions suppressing high plasmid copy number of said plasmid; or

iii. both (i) and (ii),

wherein said *ori* associated interactions are selected from the group consisting of: RNA I – RNA II complexation, RNA II – DNA complexation, RNA I – RNA II – rop complexation, RNA I – tRNA complexation, RNA II – tRNA complexation, RNA II – RNase H complexation, RNA II – sigma factor complexation, RNA II – sigma factor – RNA polymerase, and any combination thereof.

6. The method of claim 5, wherein said optimizing comprises increasing or decreasing said copy number of said plasmid such that it is optimal for expression of said polynucleotide in said host cell.

7. The method of any one of claims 2 to 6, wherein said tRNA is an uncharged tRNA.

8. The method of any one of claims 1 to 7, wherein said plasmid is a ColE1-like plasmid.

9. A method for obtaining a pre-determined copy number of a plasmid in a host cell for a user in need thereof, the method comprising introducing at least one mutation in the *ori* of said plasmid, a constituent of said *ori*, or both, thereby tuning the copy number of the plasmid to a pre-determined copy number in the host cell for the user.

10. The method of claim 9, wherein said at least one mutation comprises: substitution, deletion, insertion, or any combination thereof.

11. The method of claim 9 or 10, wherein said constituent of said *ori* comprises: RNA I, RNA II, rop, an uncharged tRNA, a promoter thereof, or any combination thereof.

12. The method of claim 11, wherein said at least one mutation is introduced in RNA I or a promoter thereof, and wherein any one of: said mutated RNA I is a deleterious RNA I, said RNA I is characterized by reduced affinity to RNA II compared to a control, said mutated promoter is characterized by increased or decreased transcription rate compared to a control, and any combination thereof.

13. The method claim 12, wherein said mutation is in RNA II or a promoter thereof, and wherein any one of: said mutated RNA II is characterized by reduced affinity to RNA I compared to a control, said mutated RNA II is characterized by increased affinity to DNA compared to a control, said mutated promoter is characterized by increased or decreased transcription rate compared to a control, said mutated RNA II is mutated at a cleavage site of RNase H, and any combination thereof.

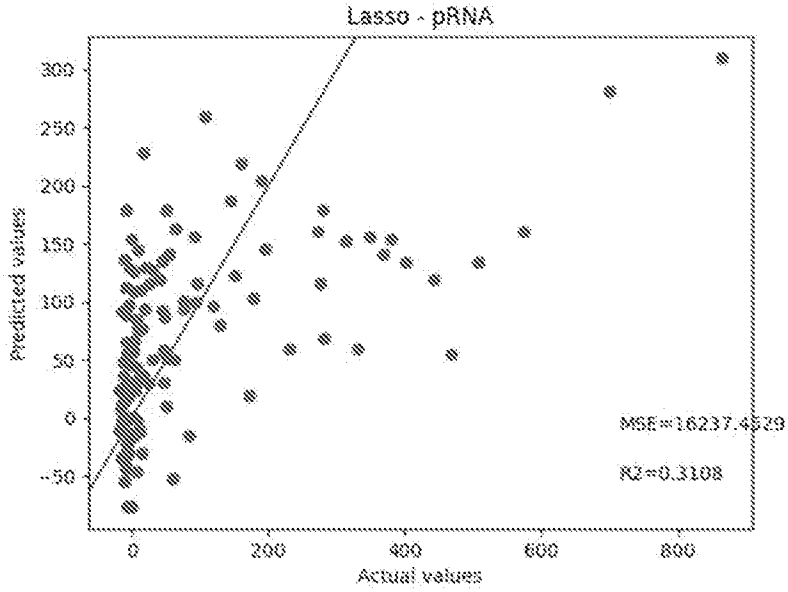
14. The method of claim 13, wherein said mutation is in rop, and wherein said mutated rop is characterized by increased or reduced affinity to a complex comprising RNA I and RNA II (kissing complex) compared to a control.

15. The method of claim 14, wherein said mutated rop stabilizes or destabilizes said kissing complex.

16. The method of any one of claims 9 to 15, further comprising a step proceeding said introducing comprising contacting said cell with said plasmid comprising said at least one mutation.

17. The method of any one of claims 9 to 16, wherein said obtaining comprises increasing or decreasing said copy number of said plasmid such that it is suitable or optimal for expression of a polynucleotide being integrated into said plasmid in said cell.

18. The method of any one of claims 4 to 17, wherein said polynucleotide encodes a polypeptide of interest.


Lasso - pRNA
Predicted values
Actual values
MSE=16237.4529
R2=0.3108
300
250
200
150
100
50
0
-50
0
200
400
600
800

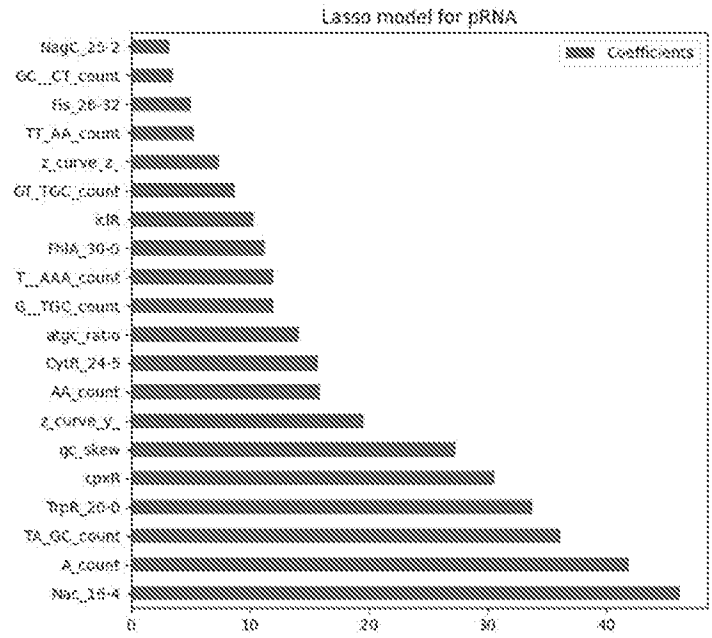

Lasso model for pRNA
Coefficients
NagC_25-2
GC__CT_count
fis_26-32
TT_AA_count
Z_curve_z_
GT_TGC_count
KdR
PNA_30-0
T__AAA_count
G__TGC_count
atgc_ratio
CytR_24-5
AA_count
Z_curve_y_
GC_skew
CpxR
TrpR_20-0
TA_GC_count
A_count
Nac_15-4
0
10
20
30
40

FIGURE 1

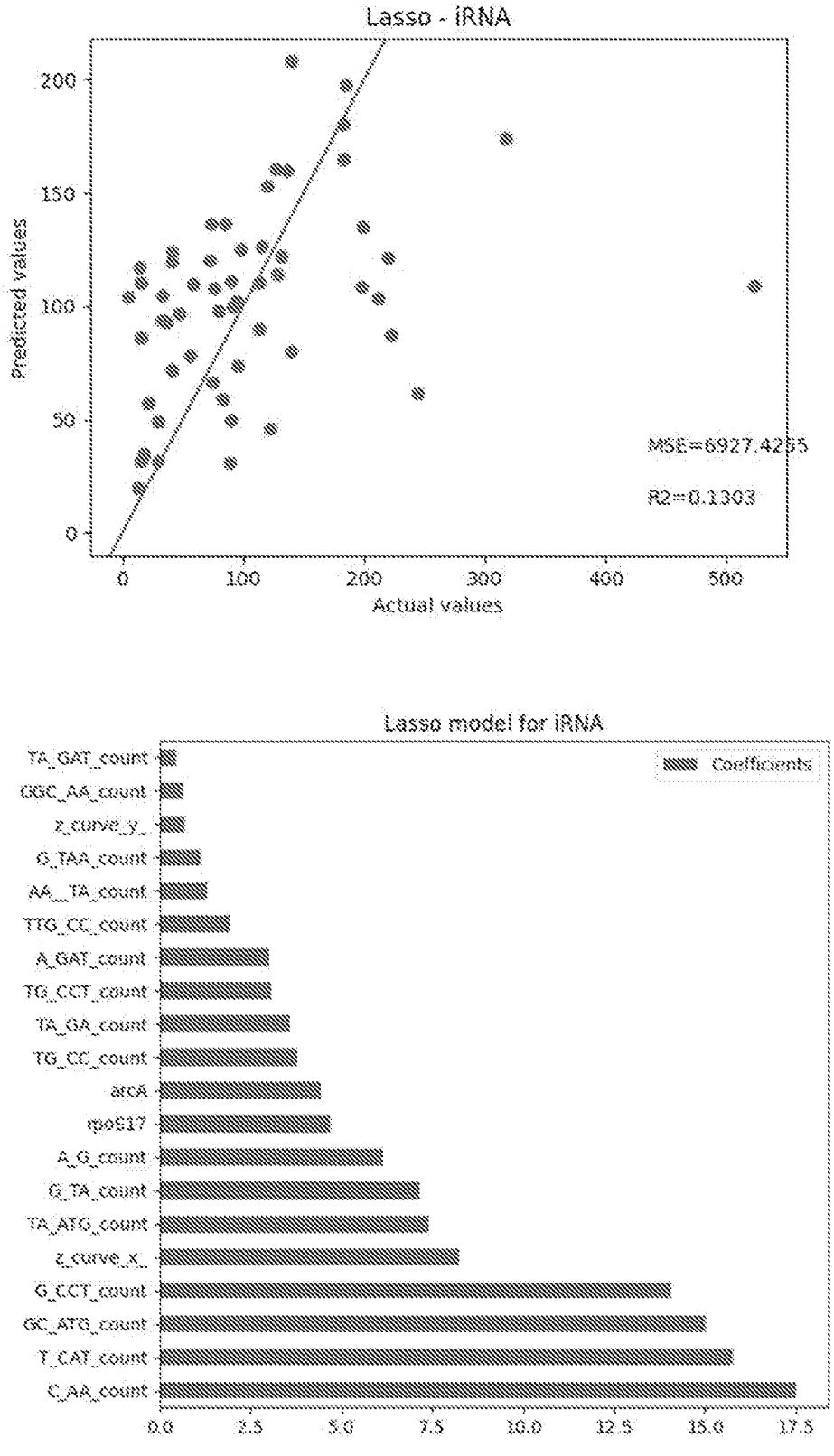
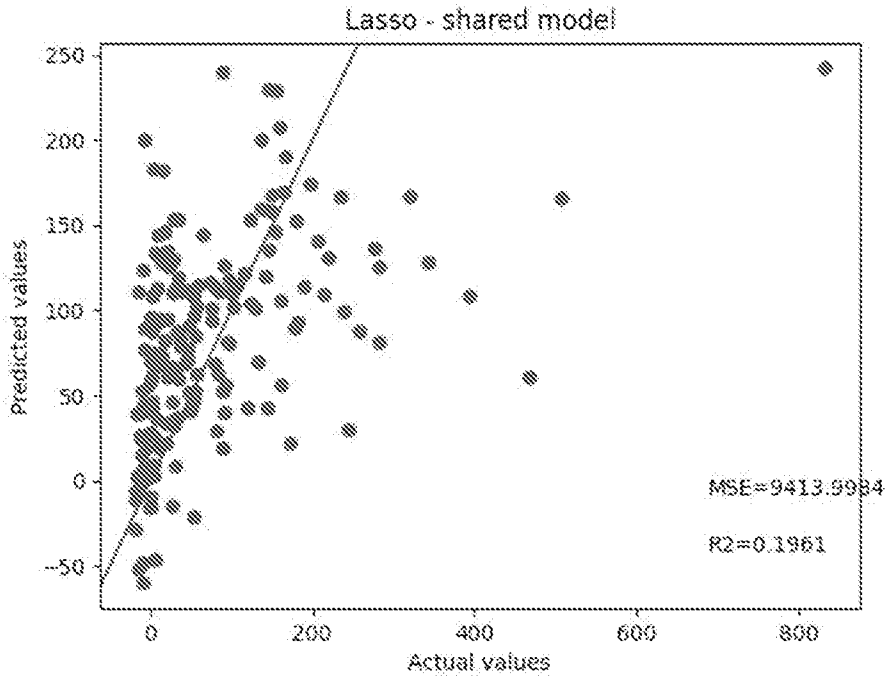

Lasso - iRNA
Predicted values
Actual values
MSE=6927.4255
R2=0.1303
Lasso model for iRNA
Coefficients
TA_GAT_count
GGC_AA_count
z_curve_y_
G_TAA_count
AA__TA_count
TTG_CC_count
A_GAT_count
TG_CCT_count
TA_GA_count
TG_CC_count
arcA
rpoS17
A_G_count
G_TA_count
TA_ATG_count
z_curve_x_
G_CCT_count
GC_ATG_count
T_CAT_count
C_AA_count

FIGURE 2


Lasso - shared model
Predicted values
Actual values
MSE=9413.9984
R2=0.1961
250
200
150
100
50
0
-50
0
200
400
600
800

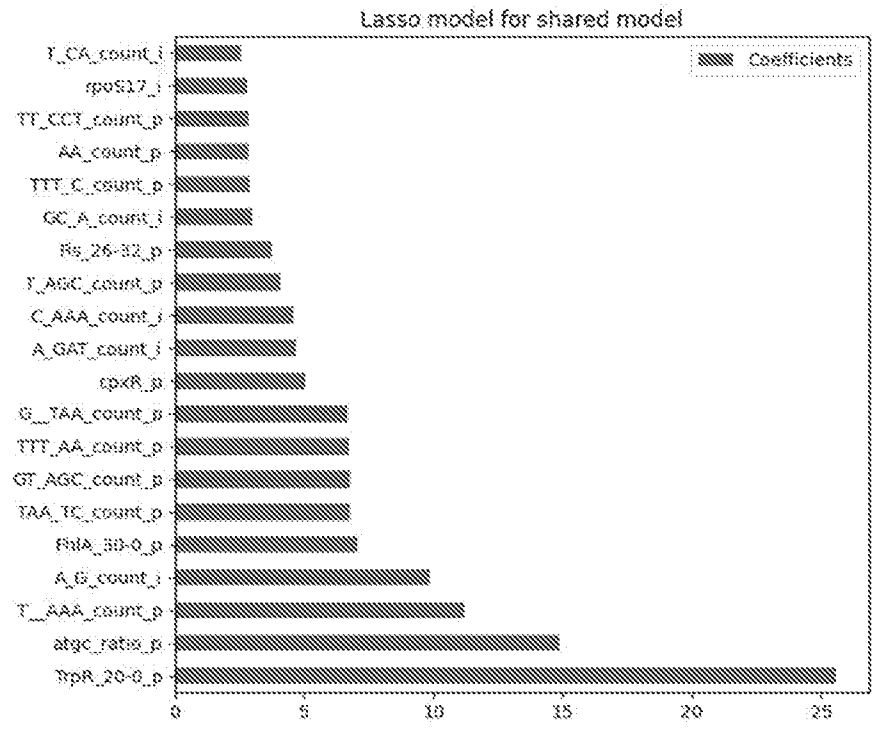
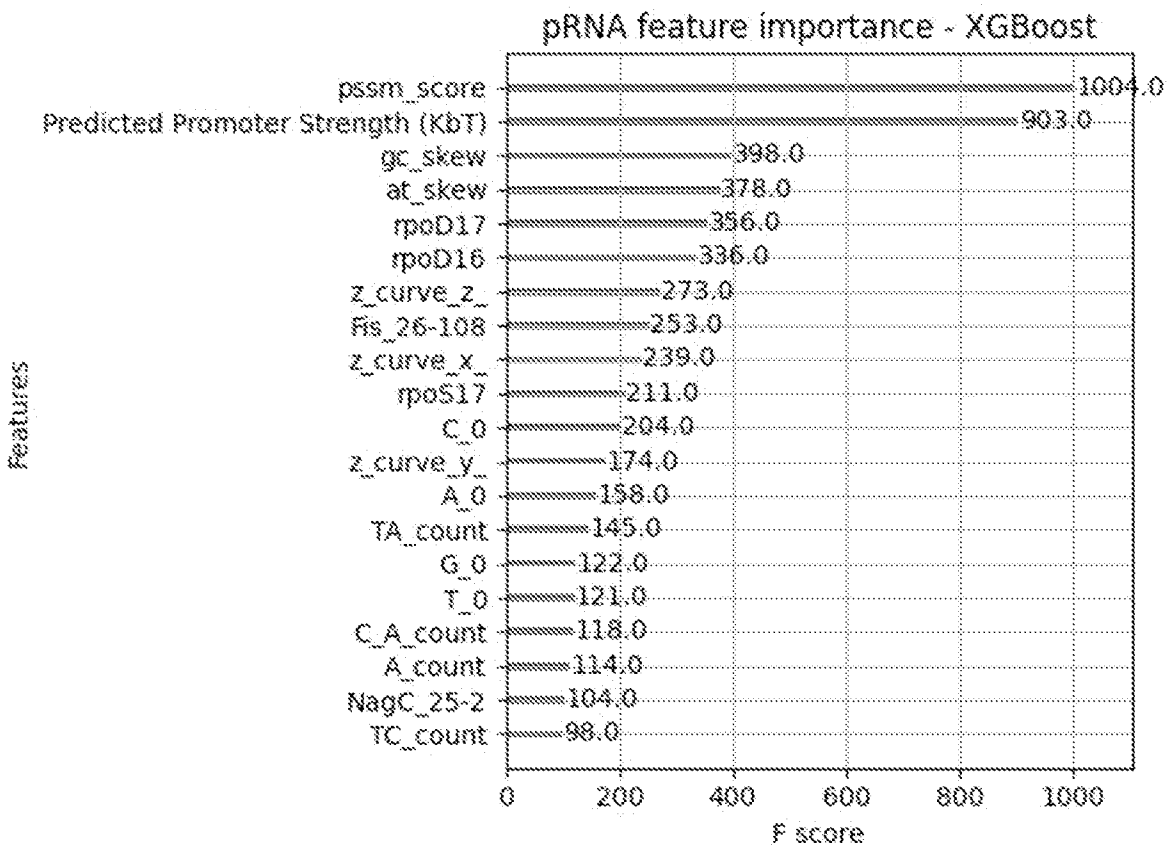

Lasso model for shared model
Coefficients
T_CA_count_i
rpoS17_i
TT_CCT_count_p
AA_count_p
TTT_C_count_p
GC_A_count_i
Rs_26-32_p
T_AGC_count_p
C_AAA_count_i
A_GAT_count_i
cpxR_p
G__TAA_count_p
TTT_AA_count_p
GT_AGC_count_p
TAA_TC_count_p
RhlA_30-0_p
A_G_count_i
T__AAA_count_p
atgc_ratio_p
TrpR_20-0_p
0
5
10
15
20
25

FIGURE 3


pRNA feature importance - XGBoost
pssm_score
Predicted Promoter Strength (KbT)
gc_skew
at_skew
rpoD17
rpoD16
z_curve_z_
Rs_26-108
z_curve_x_
rpoS17
C_0
z_curve_y_
A_0
TA_count
G_0
T_0
C_A_count
A_count
NagC_25-2
TC_count
1004.0
903.0
398.0
378.0
356.0
336.0
273.0
253.0
239.0
211.0
204.0
174.0
158.0
145.0
122.0
121.0
118.0
114.0
104.0
98.0
Features
0
200
400
600
800
1000
F score

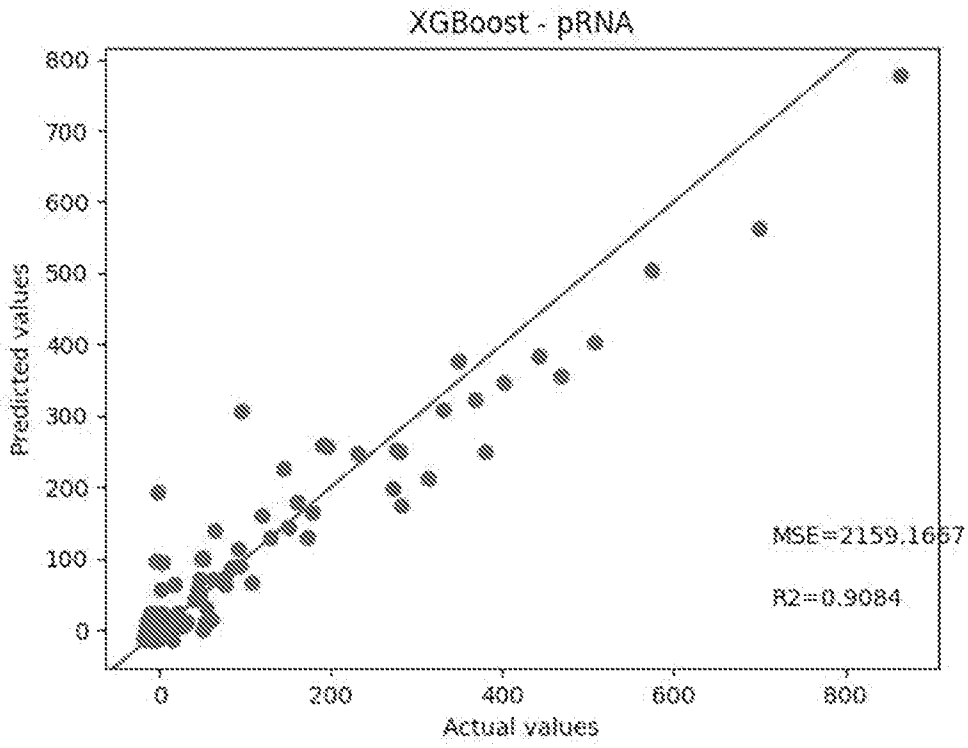
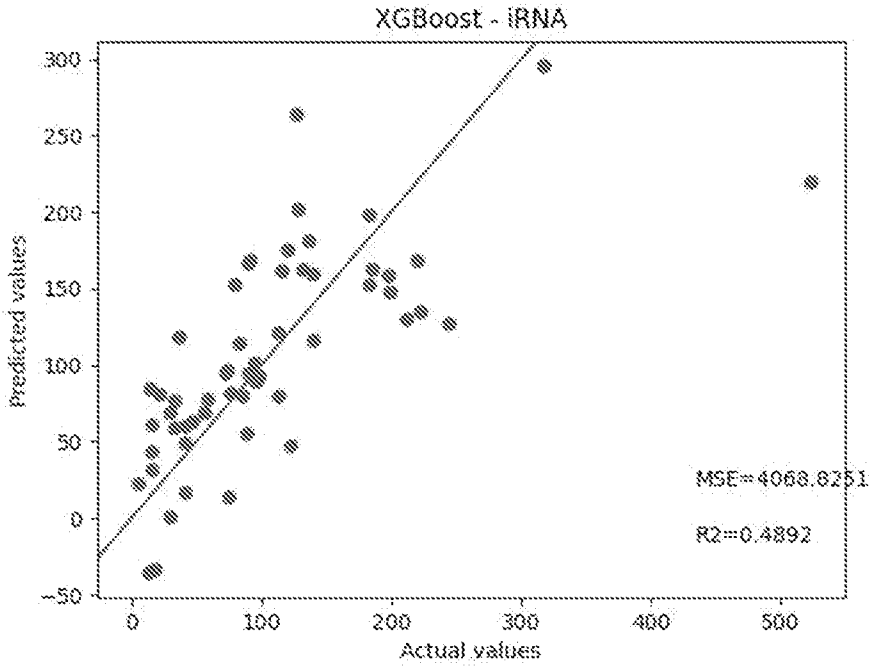

XGBoost - pRNA
Predicted values
Actual values
0
100
200
300
400
500
600
700
800
MSE=2159.1687
R2=0.9084

FIGURE 4


XGBoost - iRNA
Predicted values
Actual values
MSE=4068.8251
R2=0.4892
300
250
200
150
100
50
0
−50
0
100
200
300
400
500

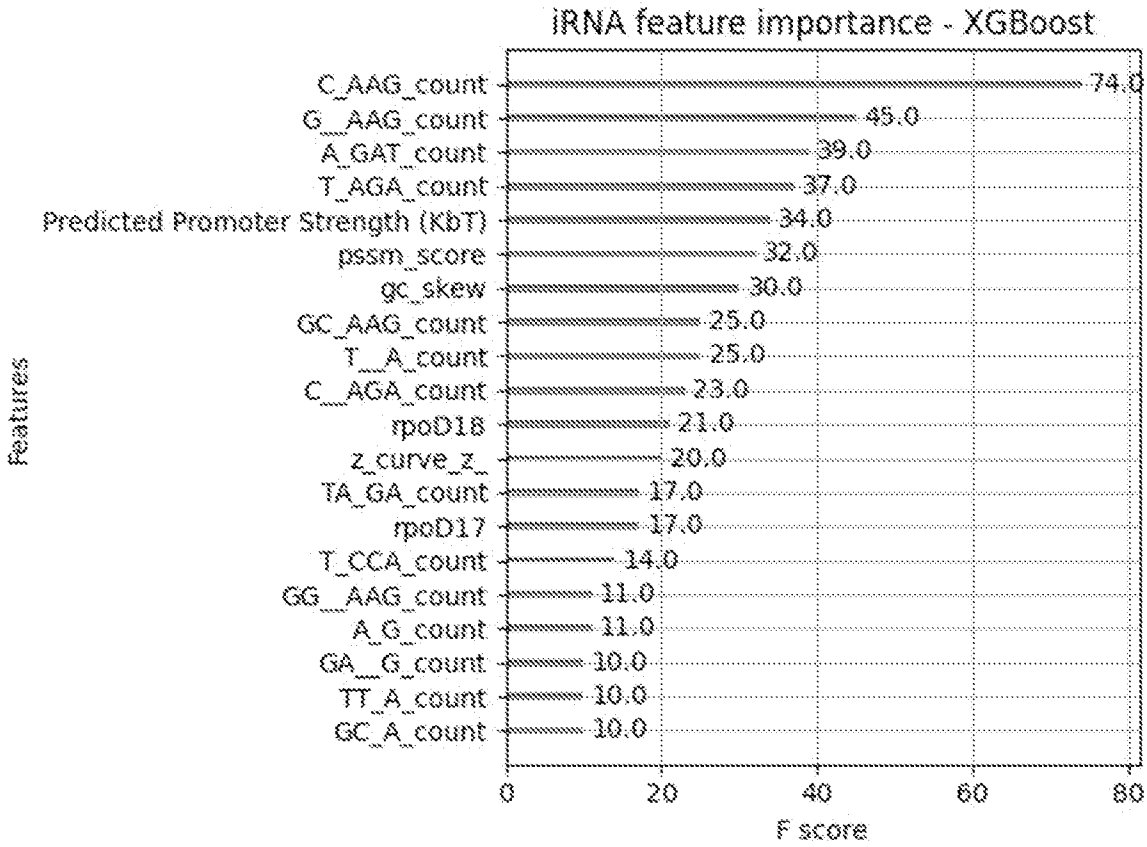
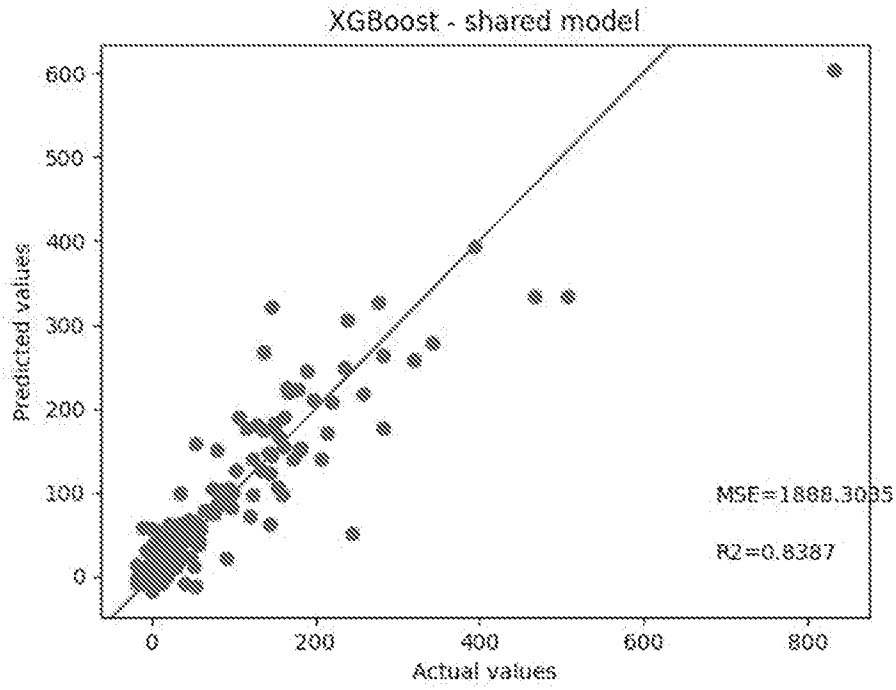

iRNA feature importance - XGBoost
Features
F score
C_AAG_count 74.0
G__AAG_count 45.0
A_GAT_count 39.0
T_AGA_count 37.0
Predicted Promoter Strength (KbT) 34.0
pssm_score 32.0
gc_skew 30.0
GC_AAG_count 25.0
T__A_count 25.0
C__AGA_count 23.0
rpoD18 21.0
z_curve_z_ 20.0
TA_GA_count 17.0
rpoD17 17.0
T_CCA_count 14.0
GG__AAG_count 11.0
A_G_count 11.0
GA__G_count 10.0
TT_A_count 10.0
GC_A_count 10.0
0
20
40
60
80

FIGURE 5


XGBoost - shared model
Predicted values
Actual values
MSE=1888.3095
R2=0.8387
0
100
200
300
400
500
600
0
200
400
600
800

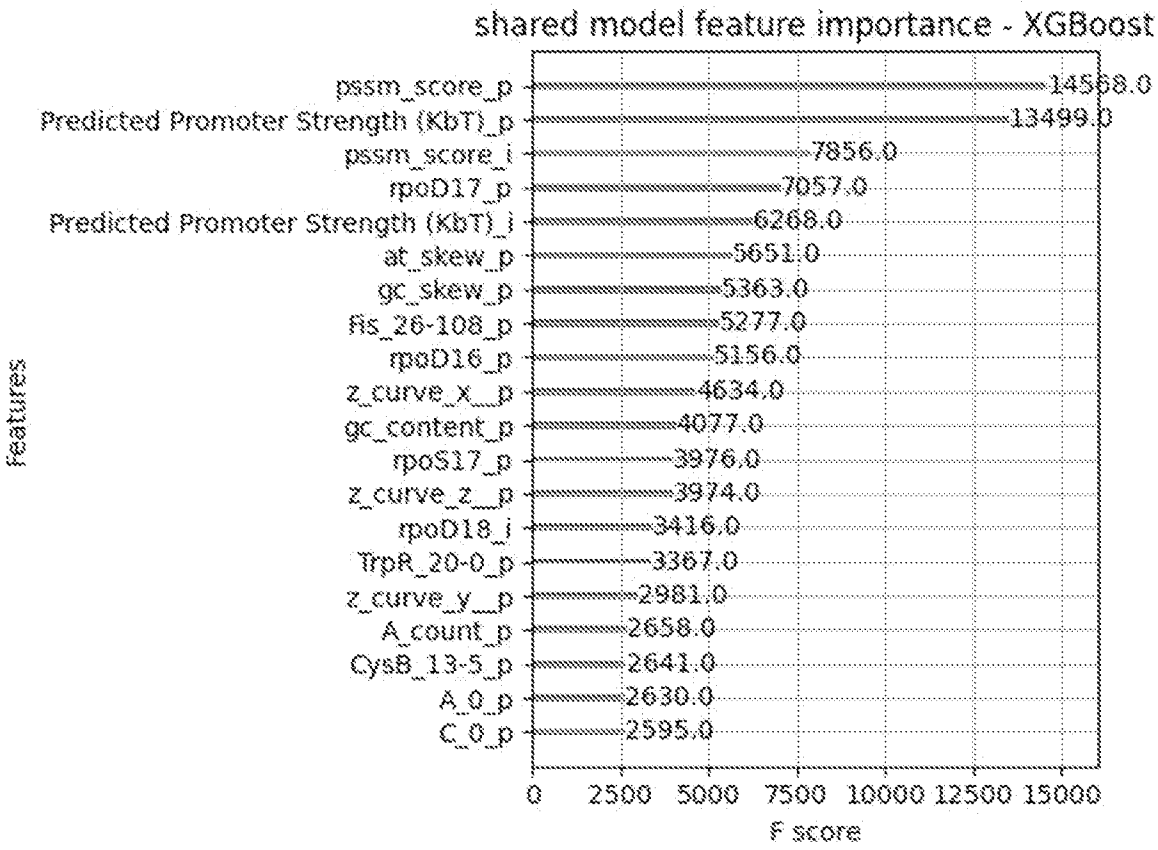

shared model feature importance - XGBoost
Features
F score
pssm_score_p 14568.0
Predicted Promoter Strength (KbT)_p 13499.0
pssm_score_i 7856.0
rpoD17_p 7057.0
Predicted Promoter Strength (KbT)_i 6268.0
at_skew_p 5651.0
gc_skew_p 5363.0
fis_26-108_p 5277.0
rpoD16_p 5156.0
z_curve_x__p 4634.0
gc_content_p 4077.0
rpoS17_p 3976.0
z_curve_z__p 3974.0
rpoD18_i 3416.0
TrpR_20-0_p 3367.0
z_curve_y__p 2981.0
A_count_p 2658.0
CysB_13-5_p 2641.0
A_0_p 2630.0
C_0_p 2595.0
0
2500
5000
7500
10000
12500
15000

FIGURE 6

```
Mean squared error (RNAi):  7138.504671330525
R-squared value (RNAi):  0.10735855160621633
Index(['G_CC_count', 'A_AA_count', 'C__AAA_count', 'C_AA_count',
       'T__ACC_count', 'AT_A_count', 'rpoD17', 'TT_CT_count', 'GC_TAA_count',
       'GA_C_count', 'T_AAT_count', 'GA_AT_count', 'A_G_count', 'GC_TTG_count',
       'TC_count', 'G_TA_count', 'T__C_count', 'A__T_count'],
      dtype='object')
```

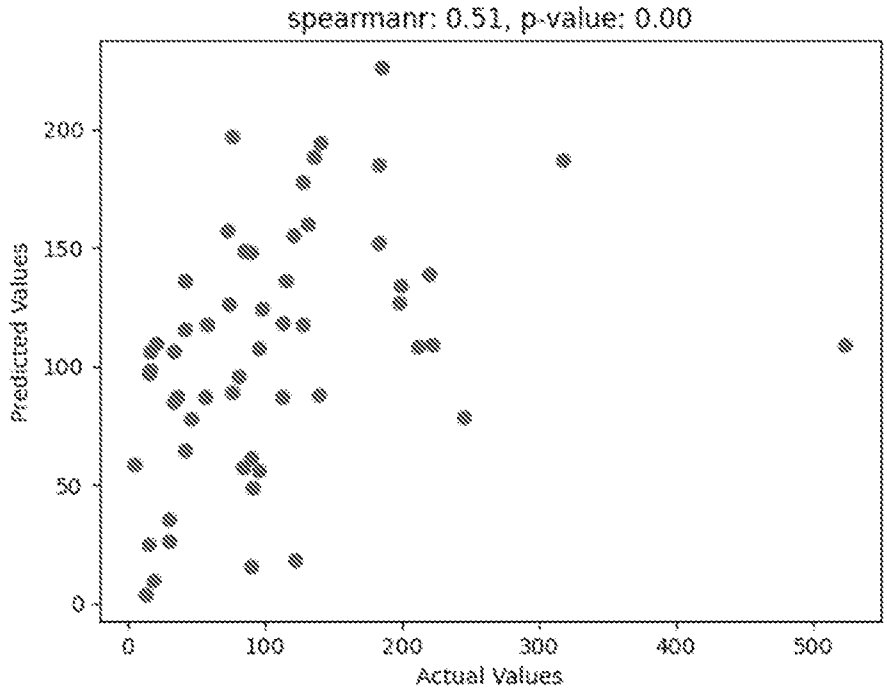


FIGURE 7

```
Mean squared error (RNAp):  17377.212539579934
R-squared value (RNAp):  0.26241453061788547
Index(['modE', 'NarP_8-3', 'G_ATA_count', 'TA_ATC_count', 'T_AGC_count',
       'T_G_count', 'rpoD17', 'cpxR', 'T__AAC_count', 'CA__CTT_count',
       'rpoD18', 'A__C_count', 'Fis_26-25', 'T_CCT_count', 'TT_BAT_count',
       'Fis_26-32', 'G_0', 'T__C_count'],
      dtype='object')
```

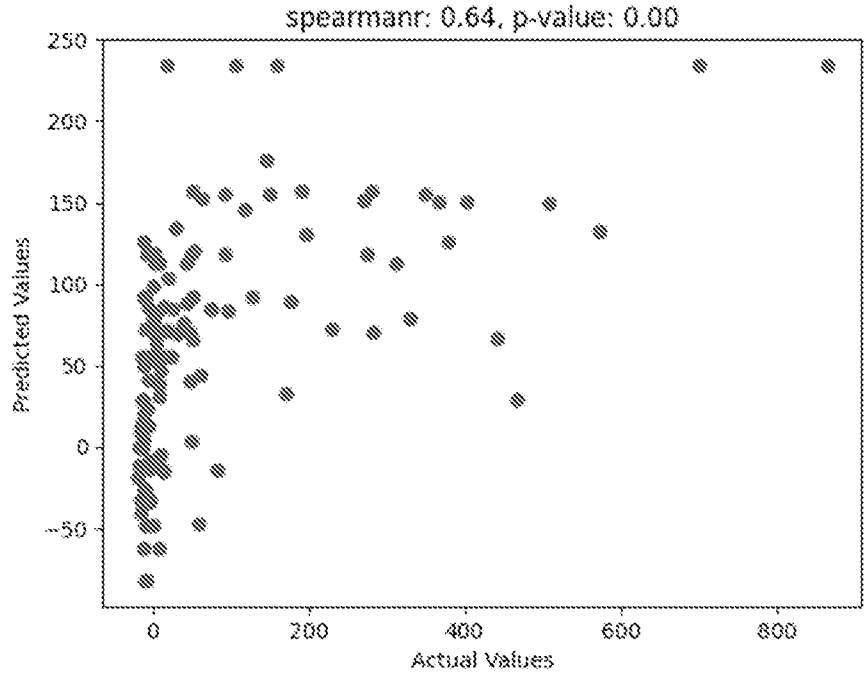


FIGURE 8

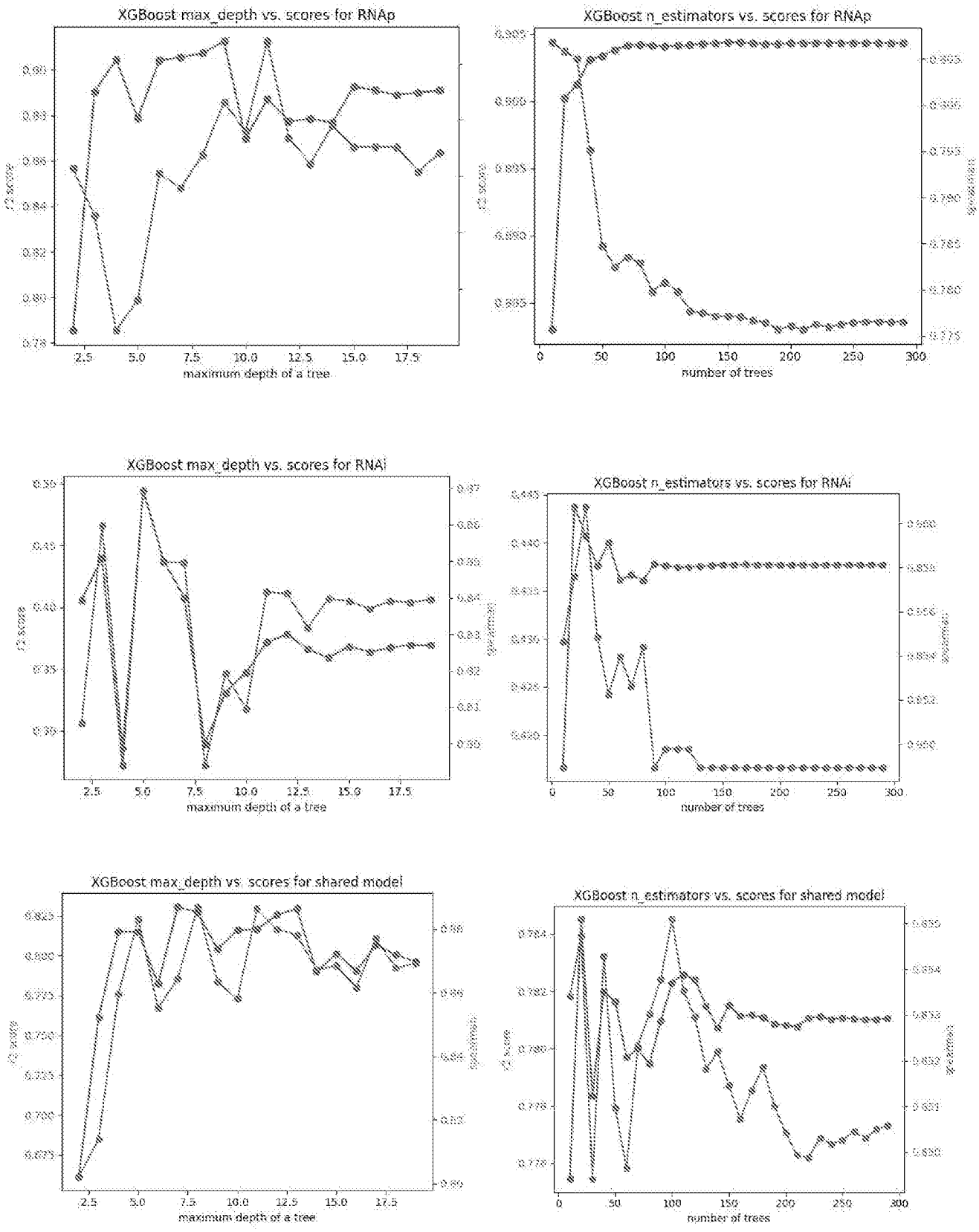

XGBoost max_depth vs. scores for RNAp
maximum depth of a tree
XGBoost n_estimators vs. scores for RNAp
number of trees
XGBoost max_depth vs. scores for RNAi
maximum depth of a tree
XGBoost n_estimators vs. scores for RNAi
number of trees
XGBoost max_depth vs. scores for shared model
maximum depth of a tree
XGBoost n_estimators vs. scores for shared model
number of trees

FIGURE 9

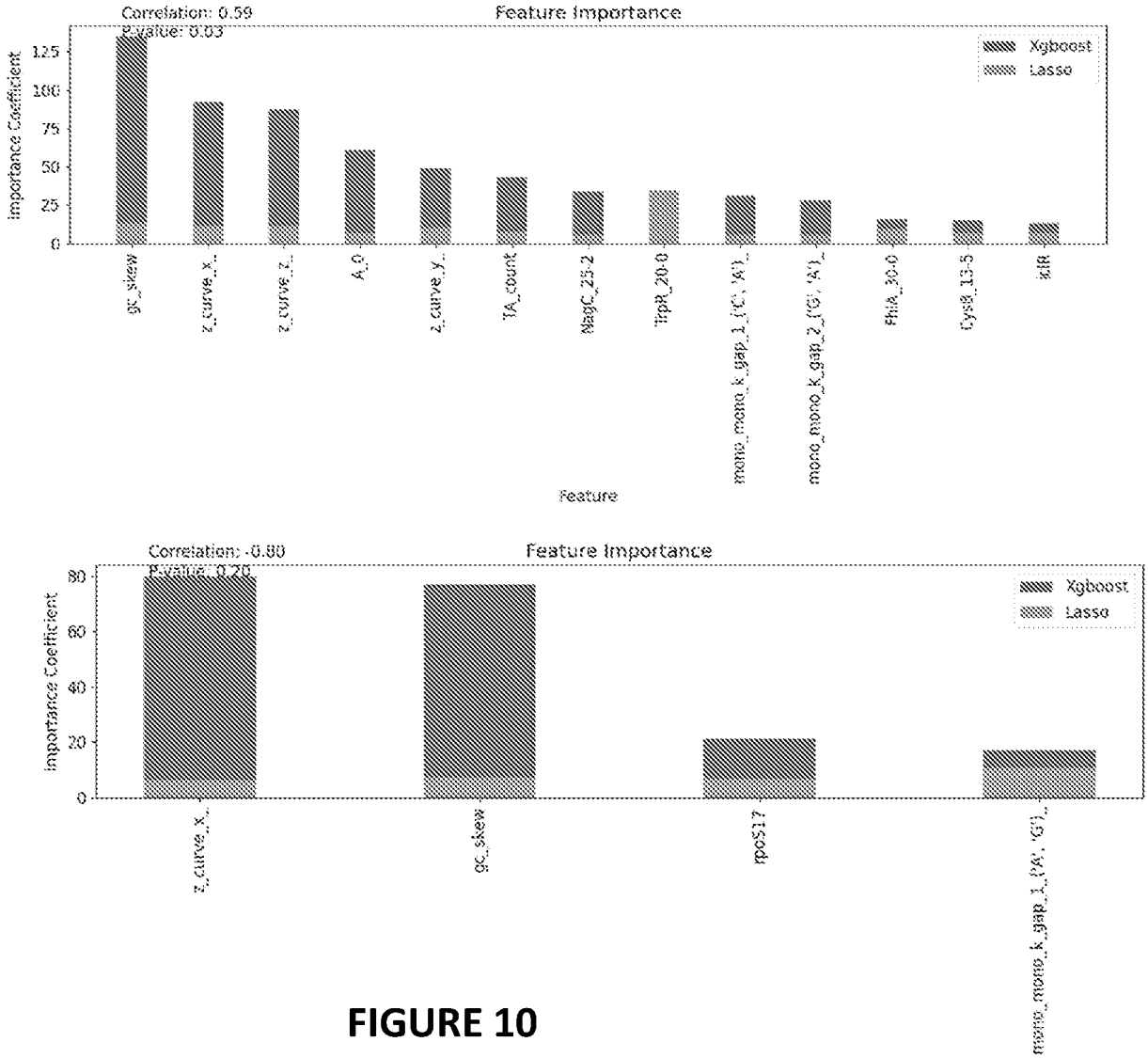


FIGURE 10

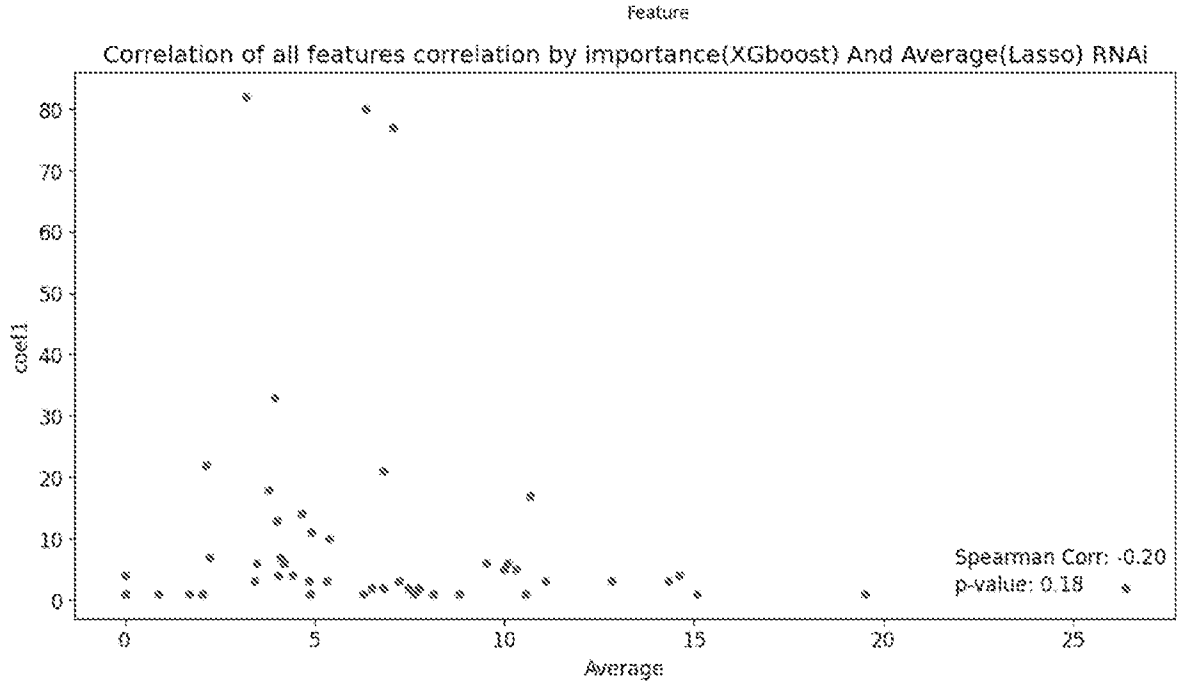


FIGURE 11

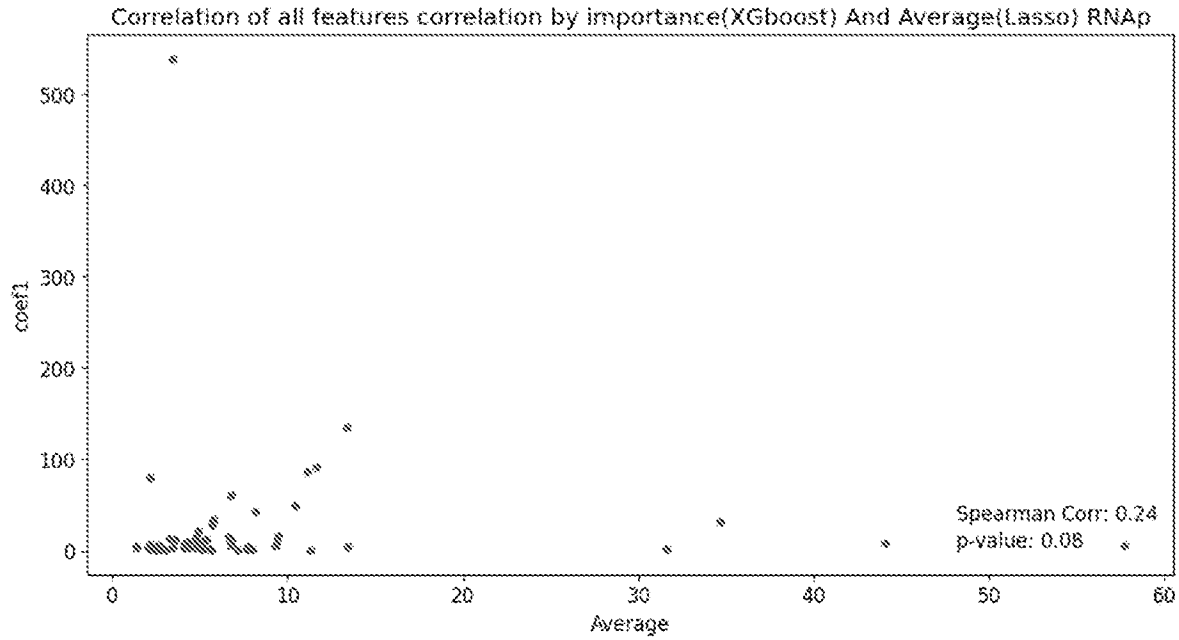


FIGURE 11 – Continue

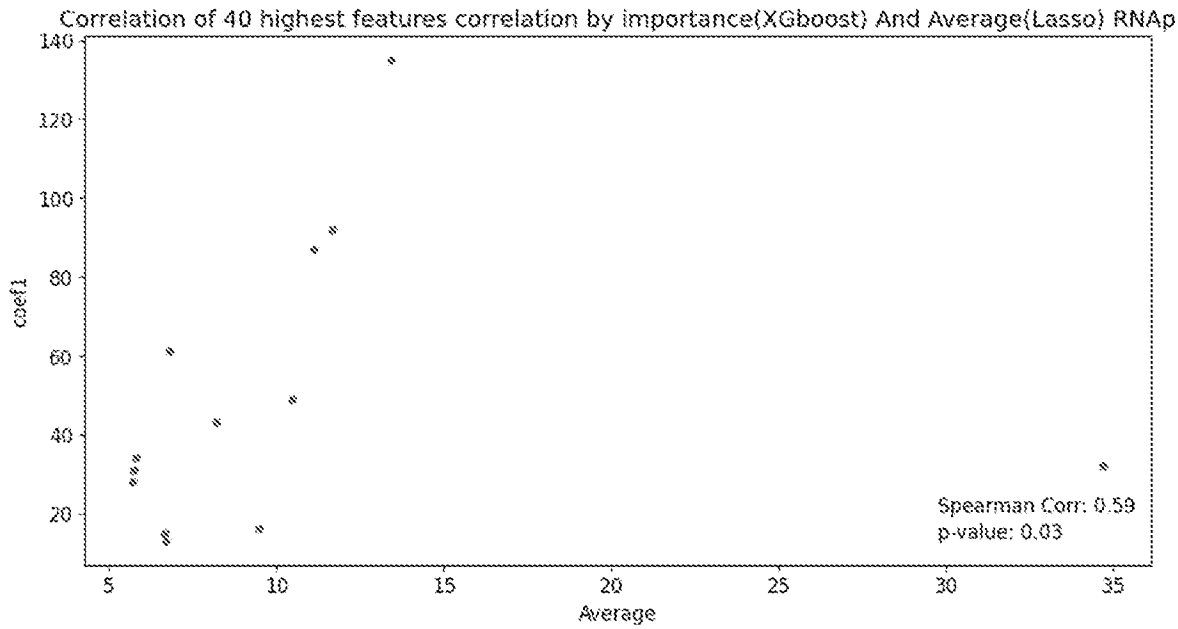


FIGURE 12

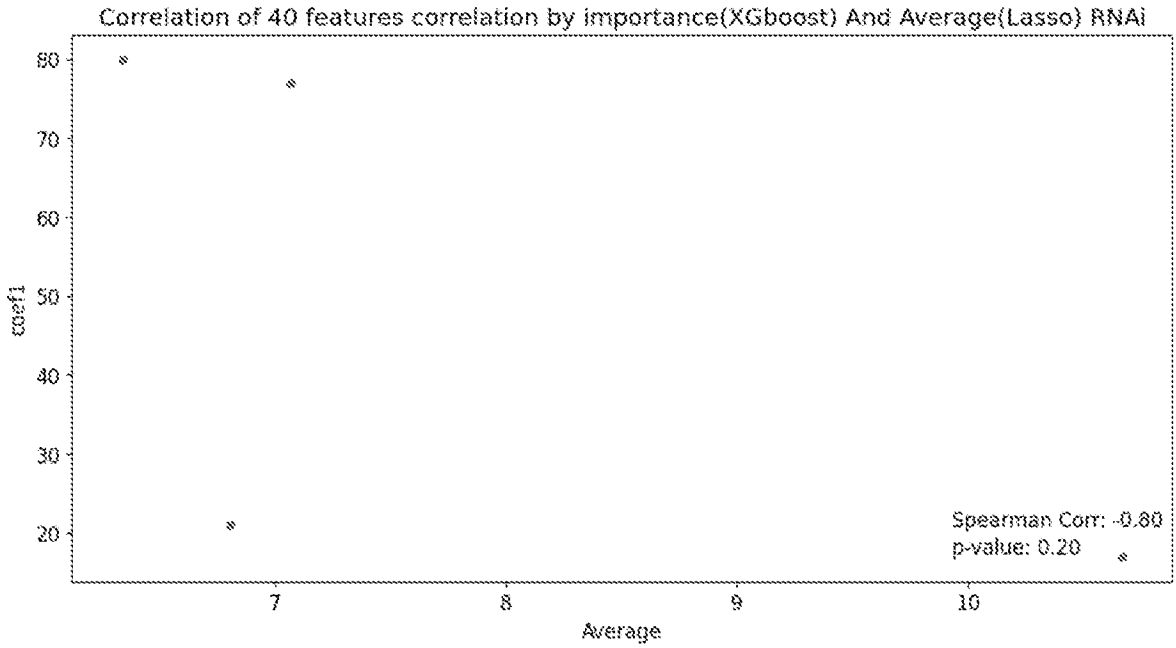


FIGURE 12 – Continue

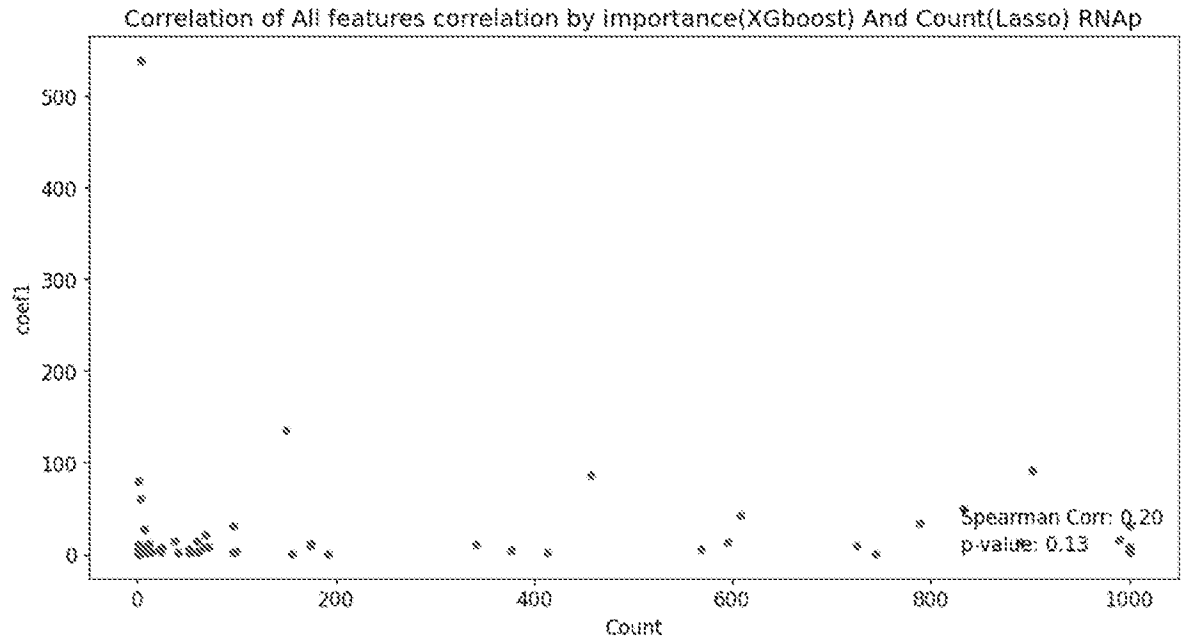


FIGURE 13

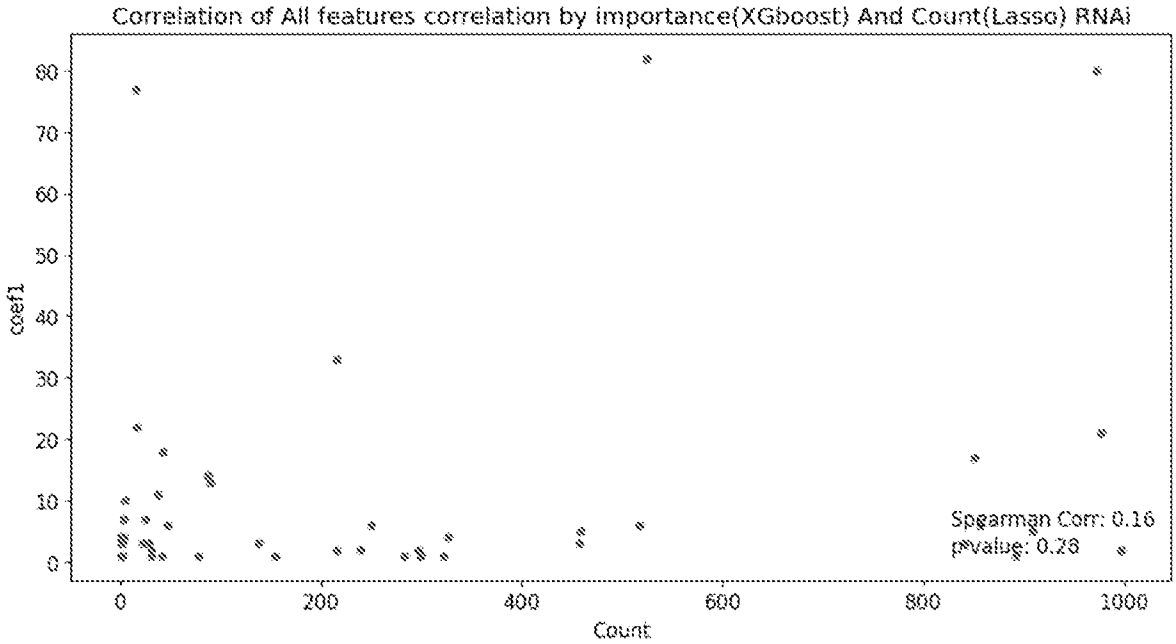


FIGURE 13 – Continue

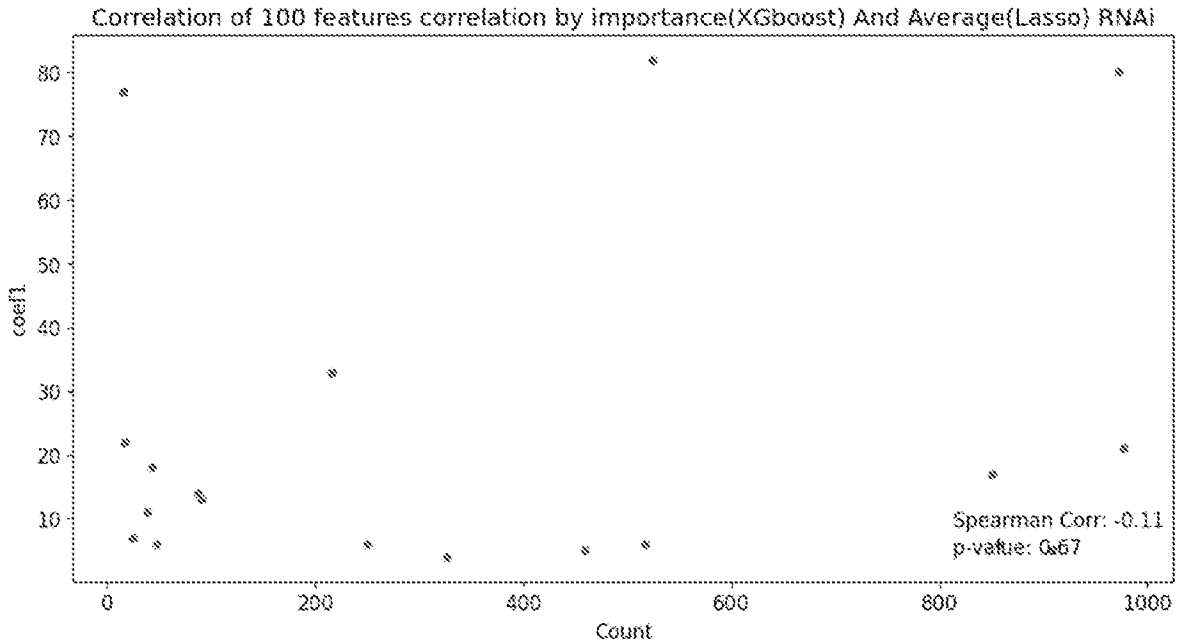


FIGURE 14

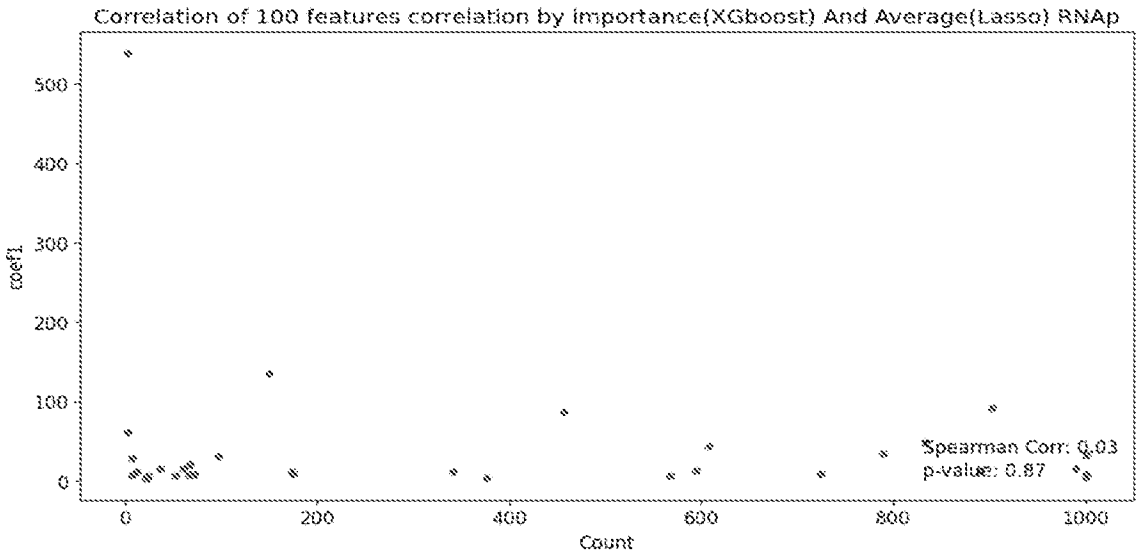


FIGURE 14 – Continue

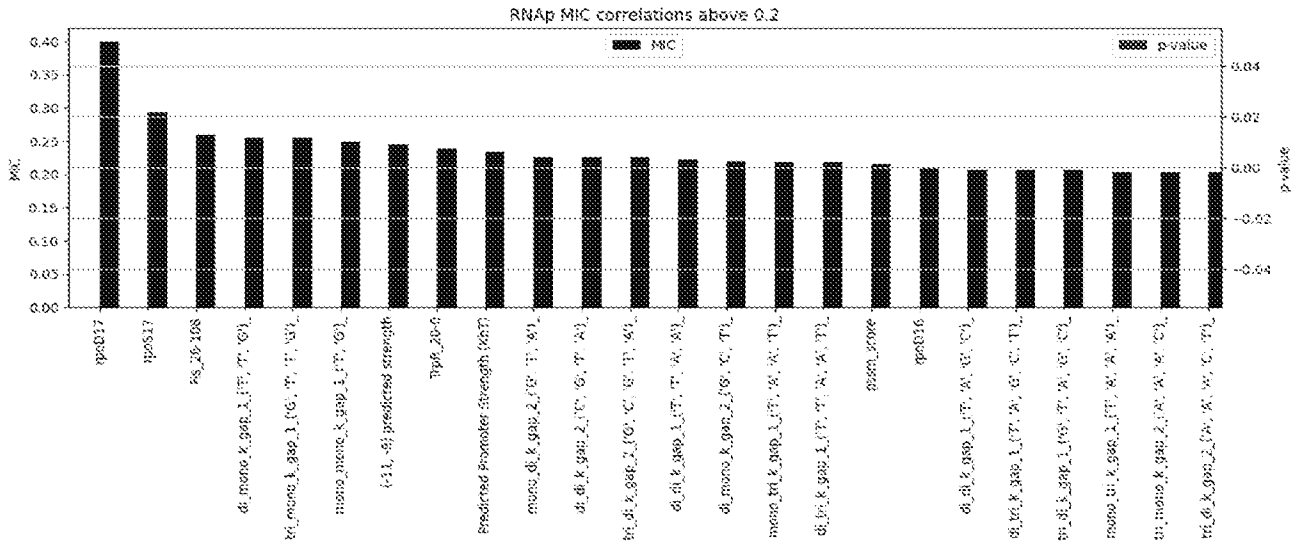


FIGURE 15

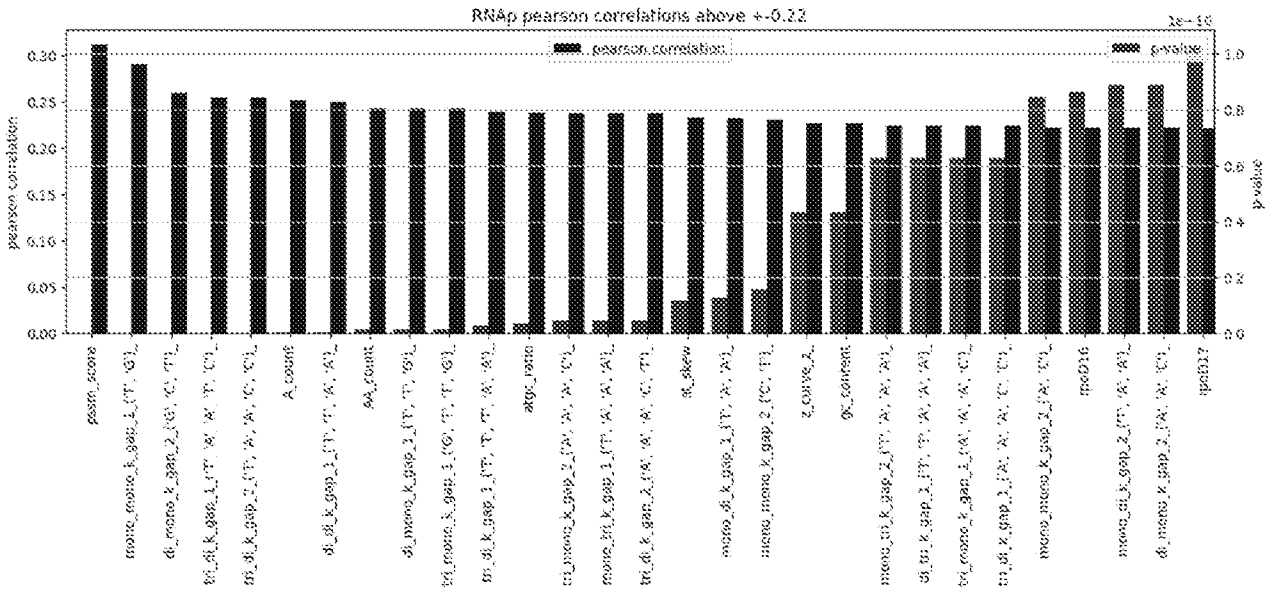


FIGURE 16

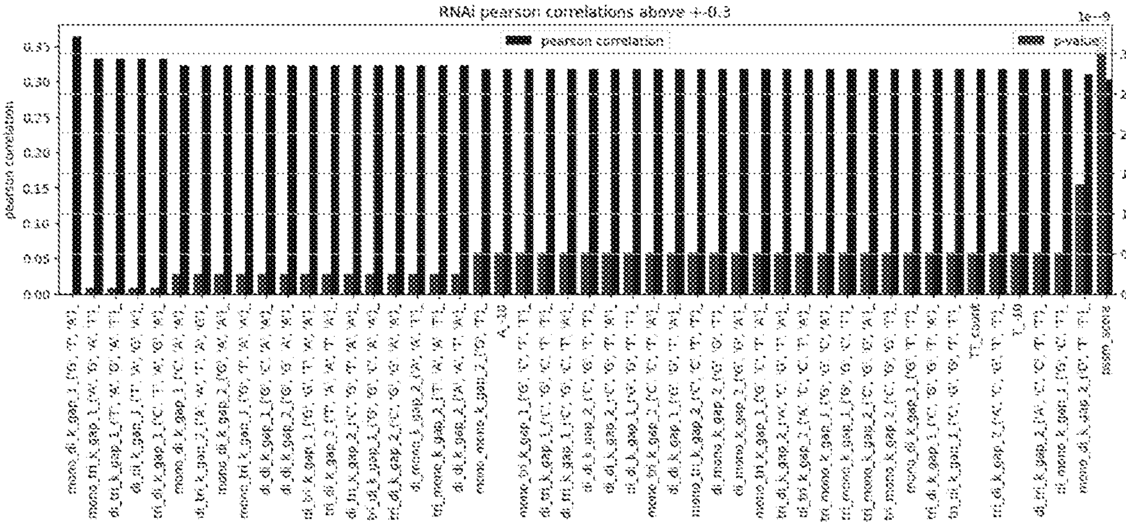


FIGURE 16 – Continue

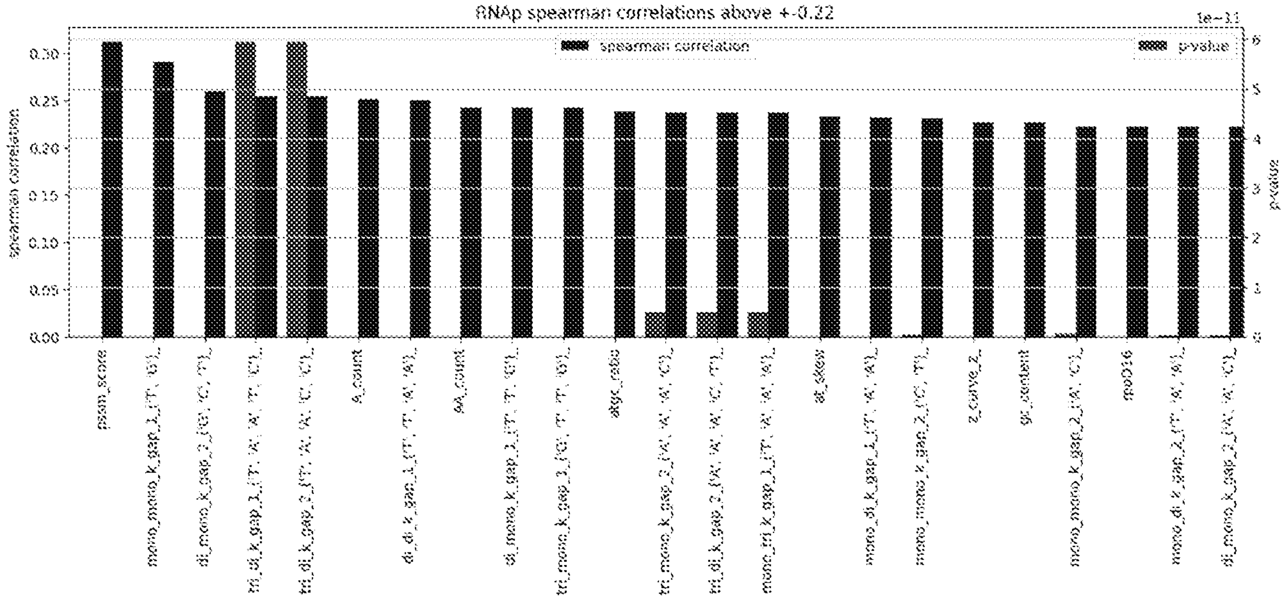


FIGURE 17

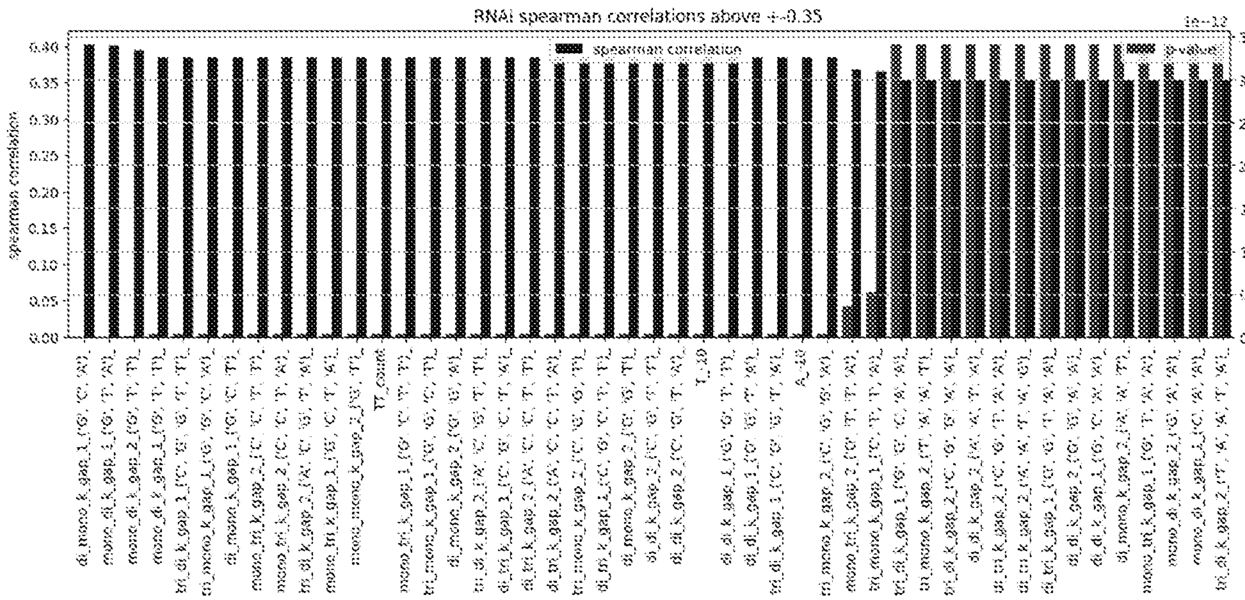


FIGURE 17 – Continue

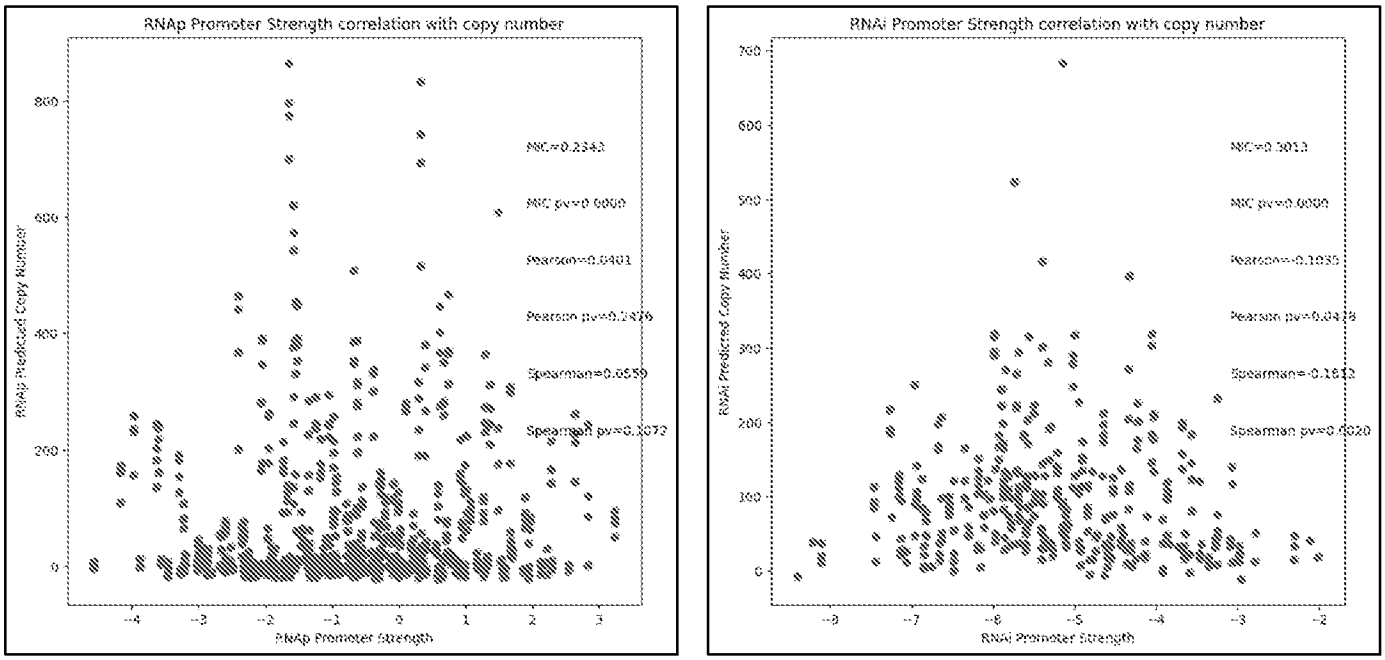


FIGURE 18

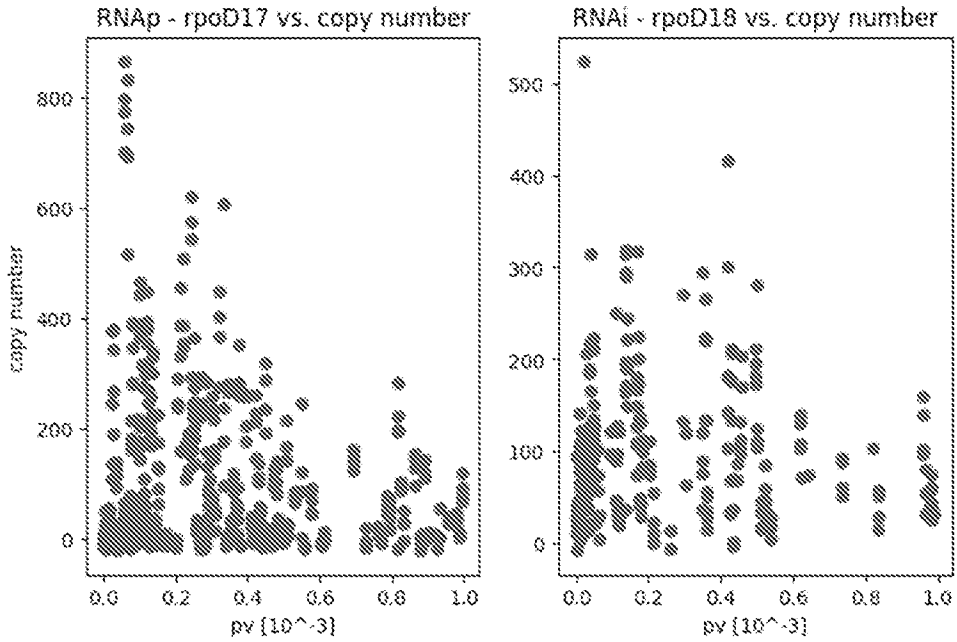
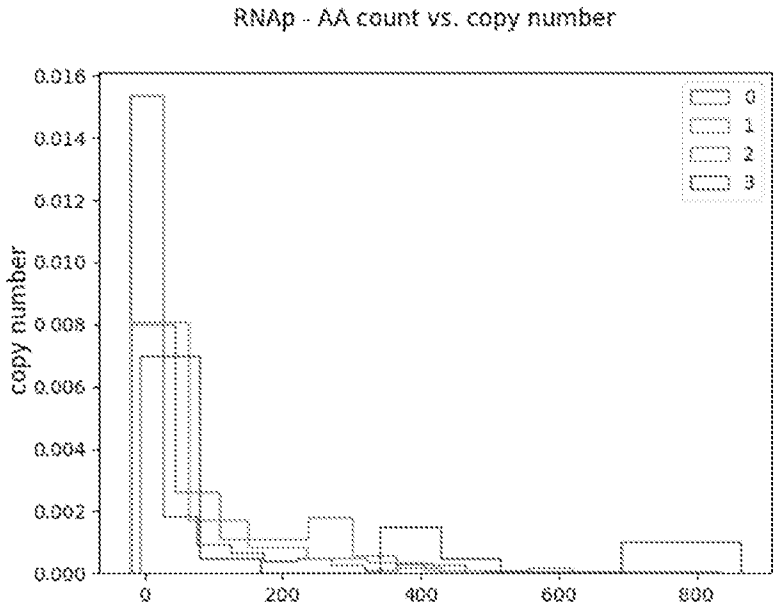


FIGURE 19


RNAp - AA count vs. copy number
copy number
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
0
1
2
3

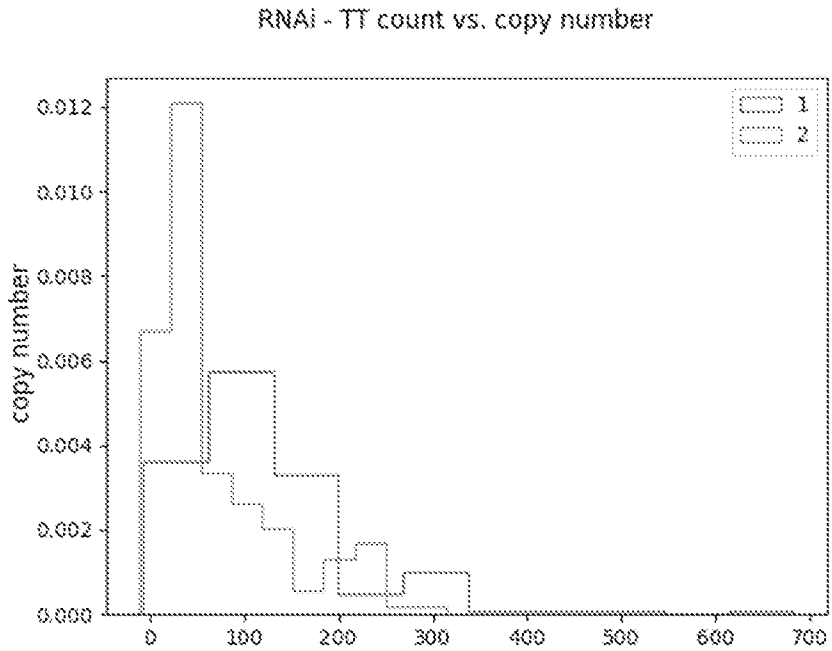
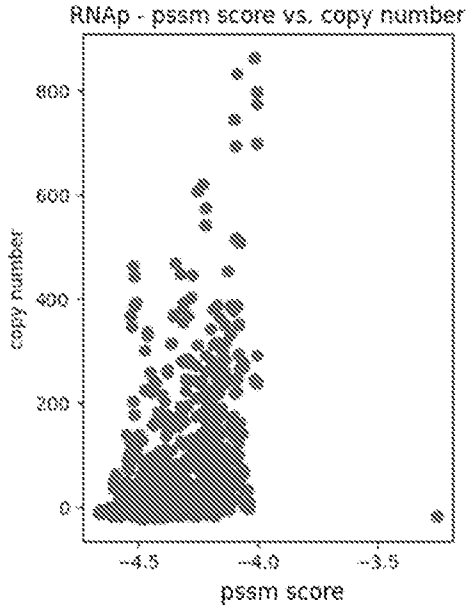

RNAi - TT count vs. copy number
copy number
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
100
200
300
400
500
600
700
1
2

FIGURE 20


RNAp - pssm score vs. copy number
copy number
0
200
400
600
800
−4.5
−4.0
−3.5
pssm score

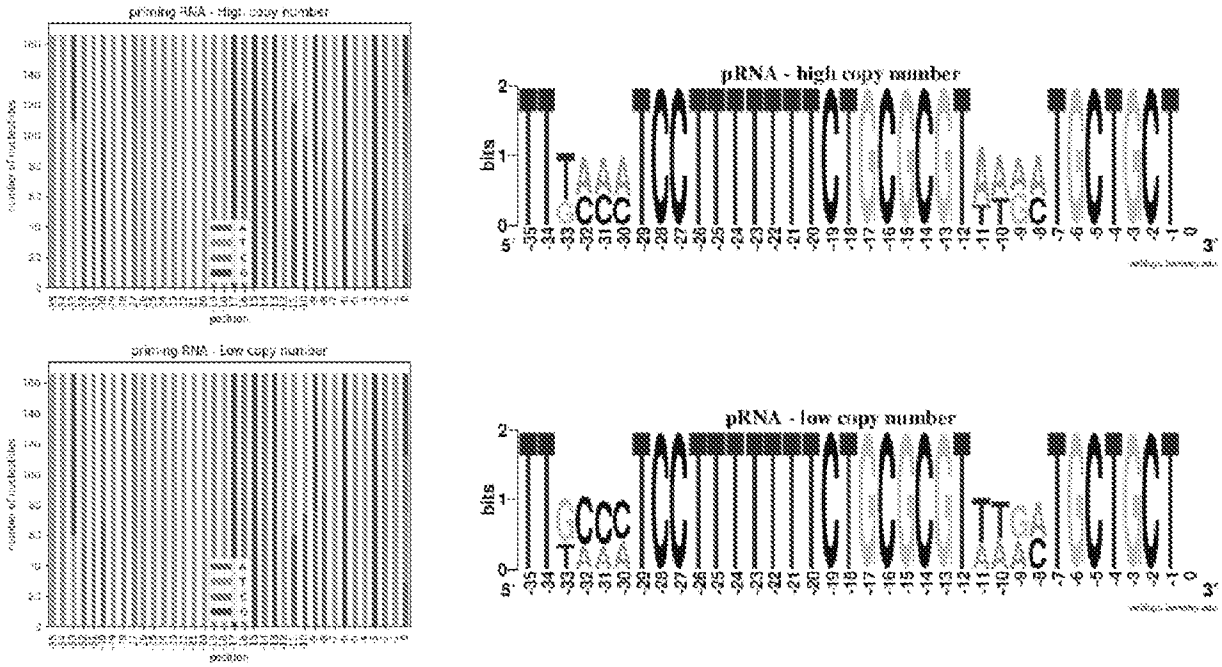
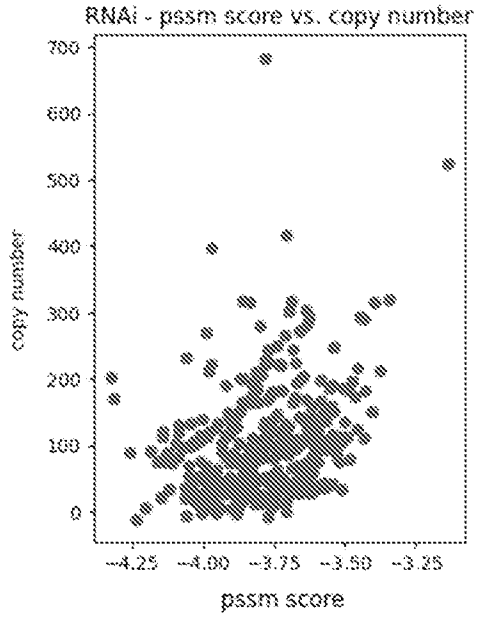
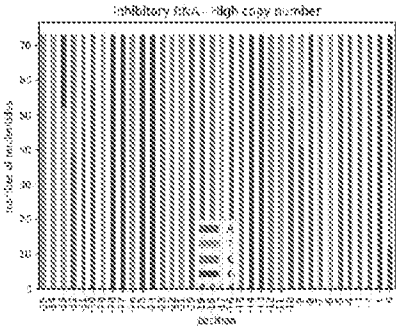
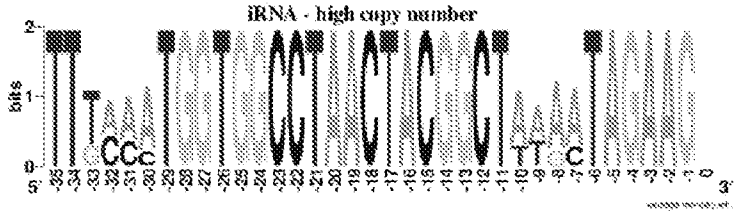
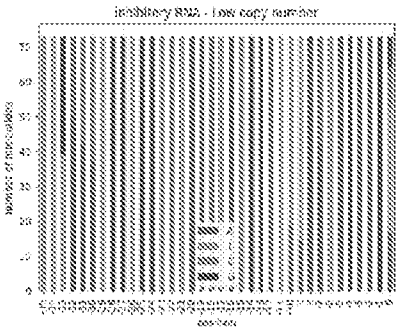

priming RNA - High copy number
priming RNA - Low copy number
position
pRNA - high copy number
pRNA - low copy number
bits
5'
3'

FIGURE 21


RNAi - pssm score vs. copy number
copy number
pssm score
700
600
500
400
300
200
100
0
-4.25
-4.00
-3.75
-3.50
-3.25




iRNA - high copy number
bits



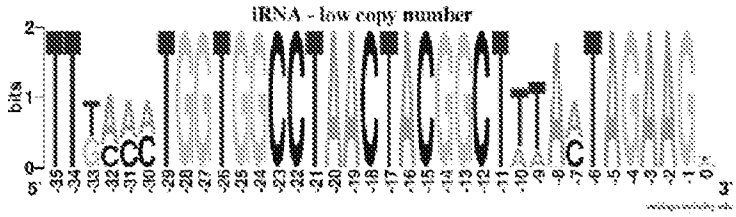

iRNA - low copy number
bits

FIGURE 22

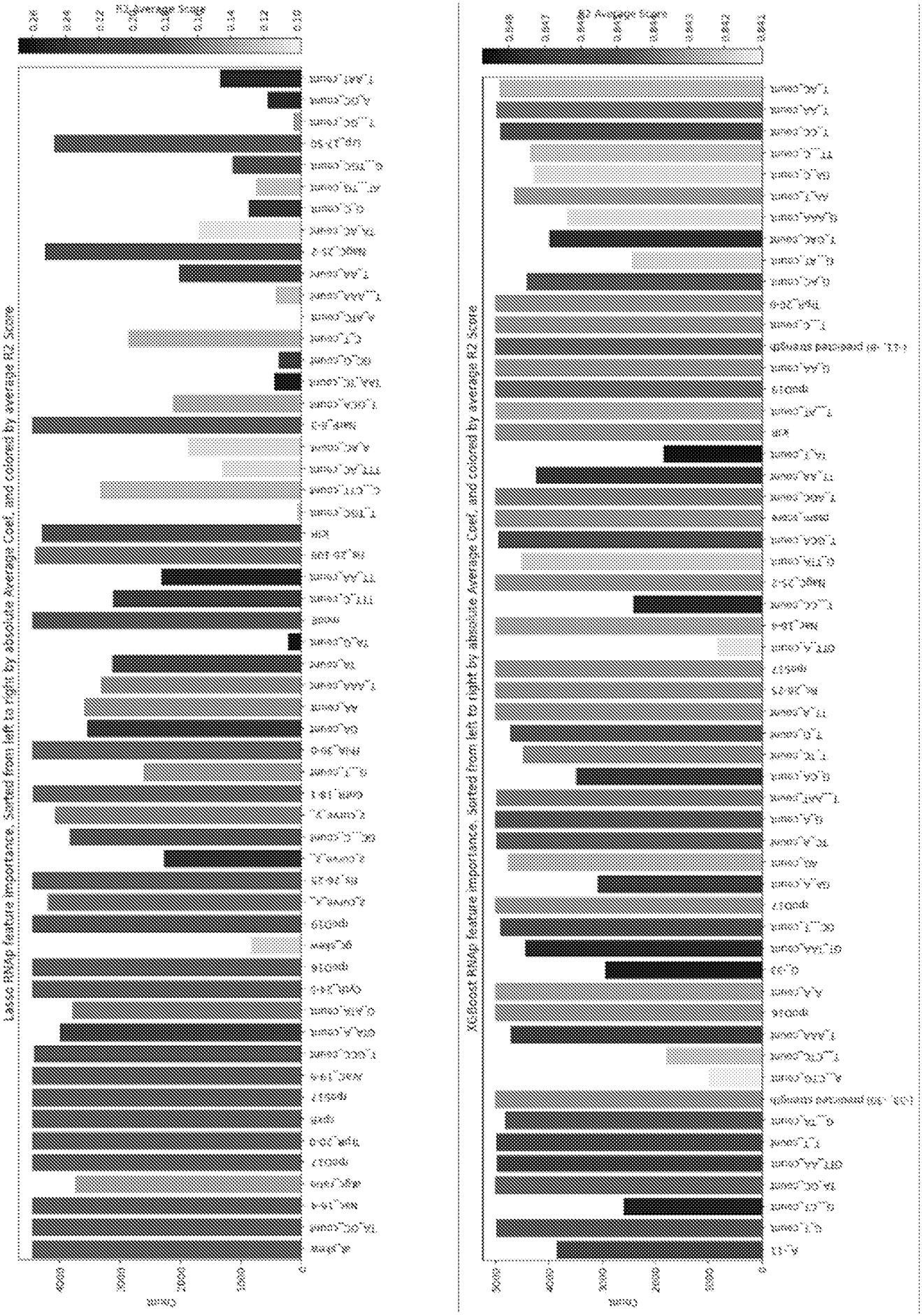

Lasso RNAp feature importance. Sorted from left to right by absolute Average Coef, and colored by average R2 Score
R2 Average Score
Count
XGBoost RNAp feature importance. Sorted from left to right by absolute Average Coef, and colored by average R2 Score
R2 Average Score
Count

FIGURE 23

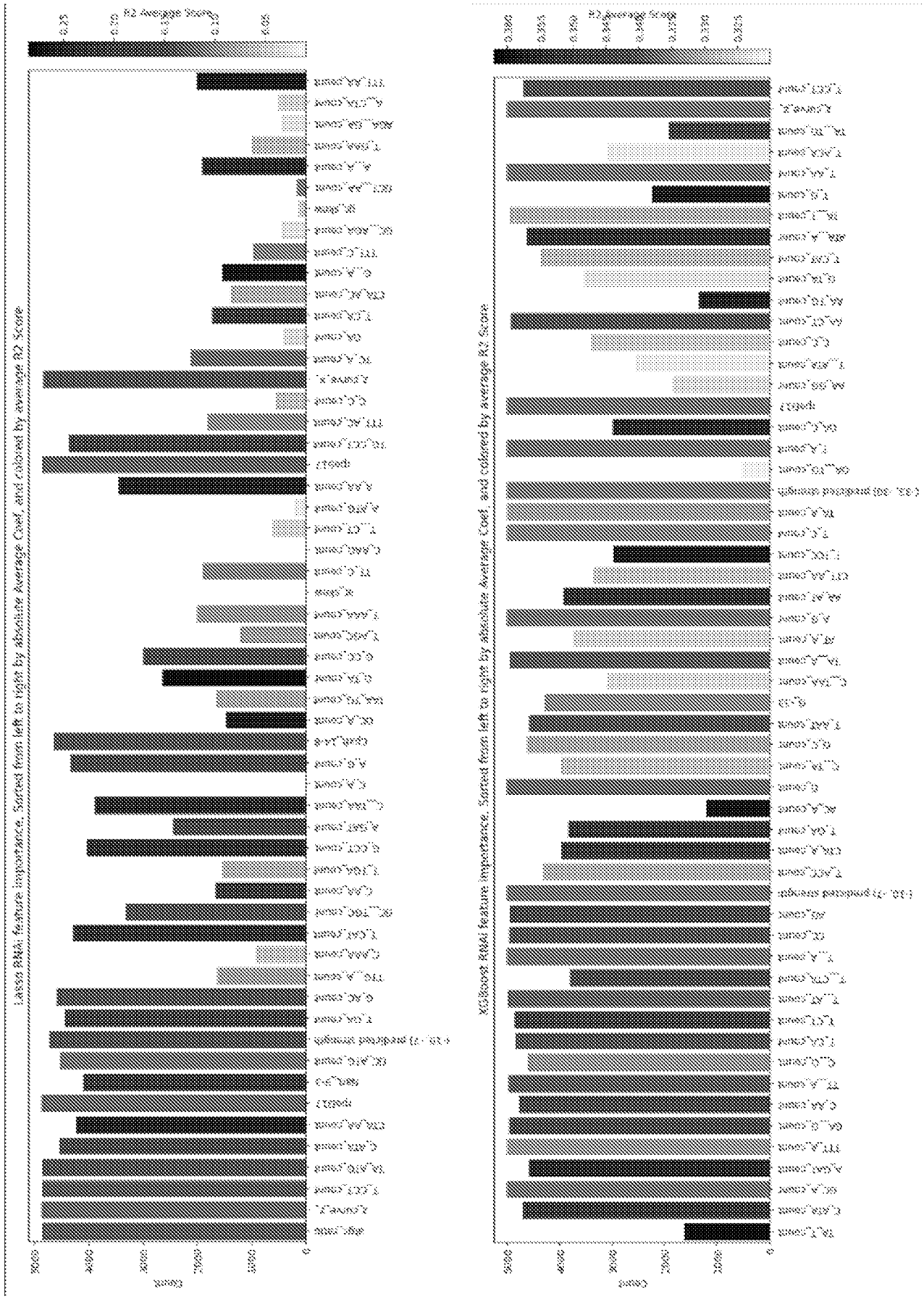


FIGURE 23 – Continue

	R2 Score RNAp lasso	R2 Score RNAi lasso	R2 Score RNAp xgb	R2 Score RNAi xgb
count	4450.000000	4863.000000	5000.000000	5000.000000
mean	0.226170	0.254521	0.846486	0.347957
std	0.081628	0.112658	0.072620	0.193220
min	0.000087	0.000548	-0.109136	-1.143647
25%	0.171176	0.169604	0.818282	0.229121
50%	0.238637	0.252497	0.860893	0.356016
75%	0.289420	0.336580	0.892105	0.488304
max	0.418862	0.619473	0.957422	0.844549

FIGURE 24

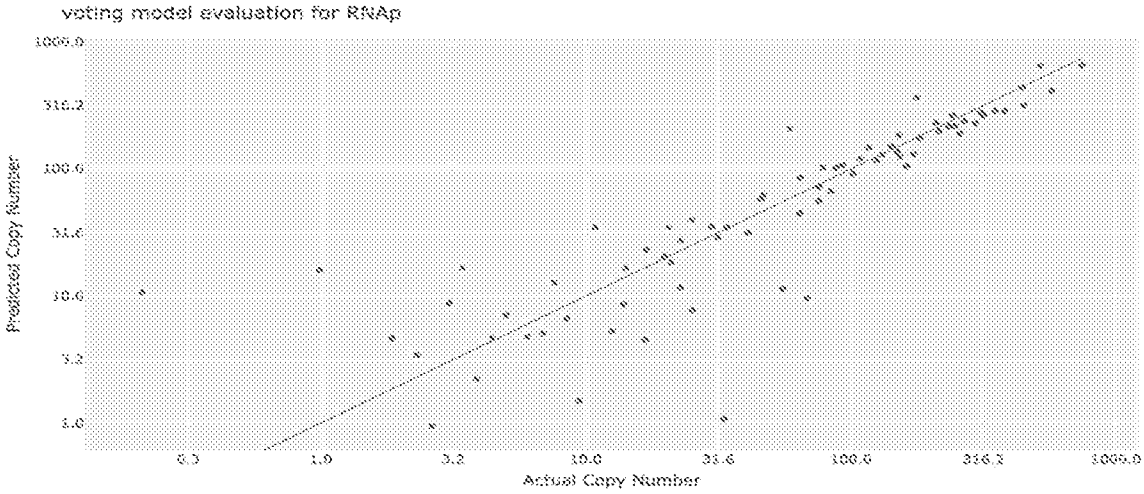


FIGURE 25A

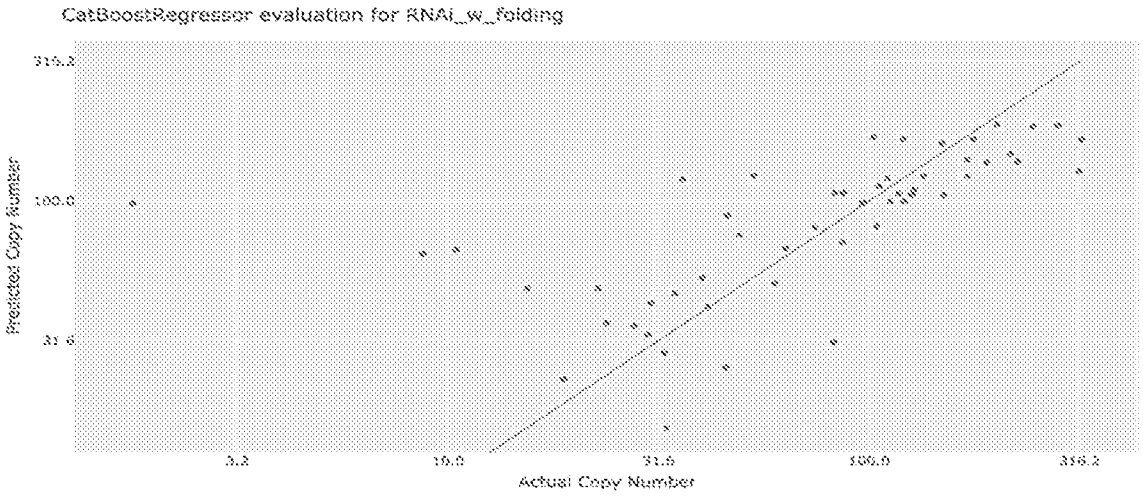


FIGURE 25B

	Pearson correlation	Spearman correlation	R-squared score	MAE
RNAp (voting model)	0.959 (p.v=7.92e-70)	0.863 (p.v=1.30e-38)	0.981	21.062
RNAi (CatBoost)	0.793 (p.v=5.13e-13)	0.821 (p.v=1.62e-14)	0.588	34.404

FIGURE 25C

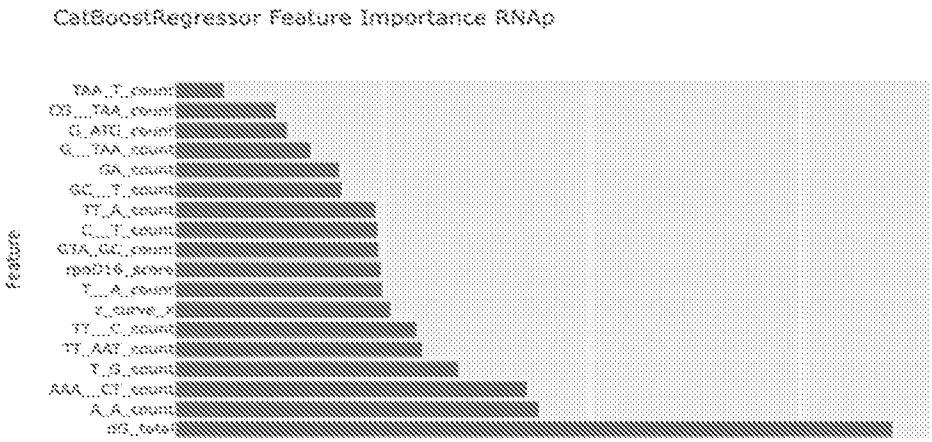


FIGURE 25D

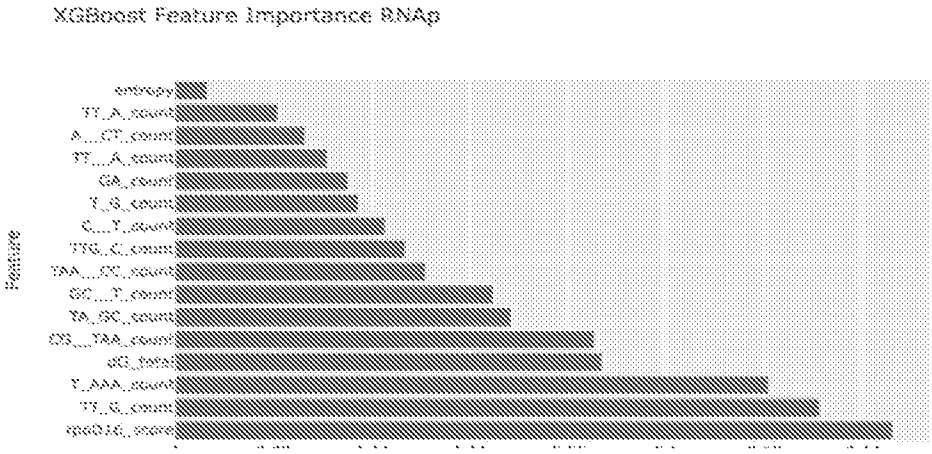


FIGURE 25E

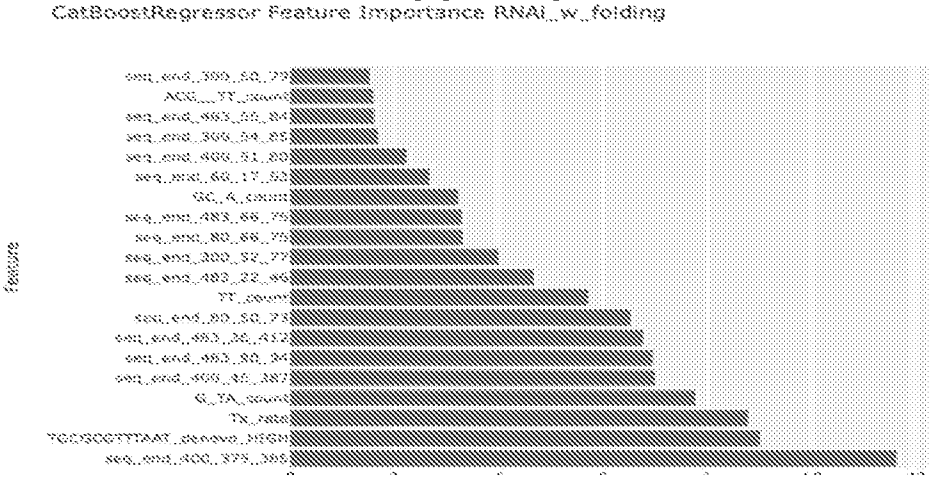


FIGURE 25F

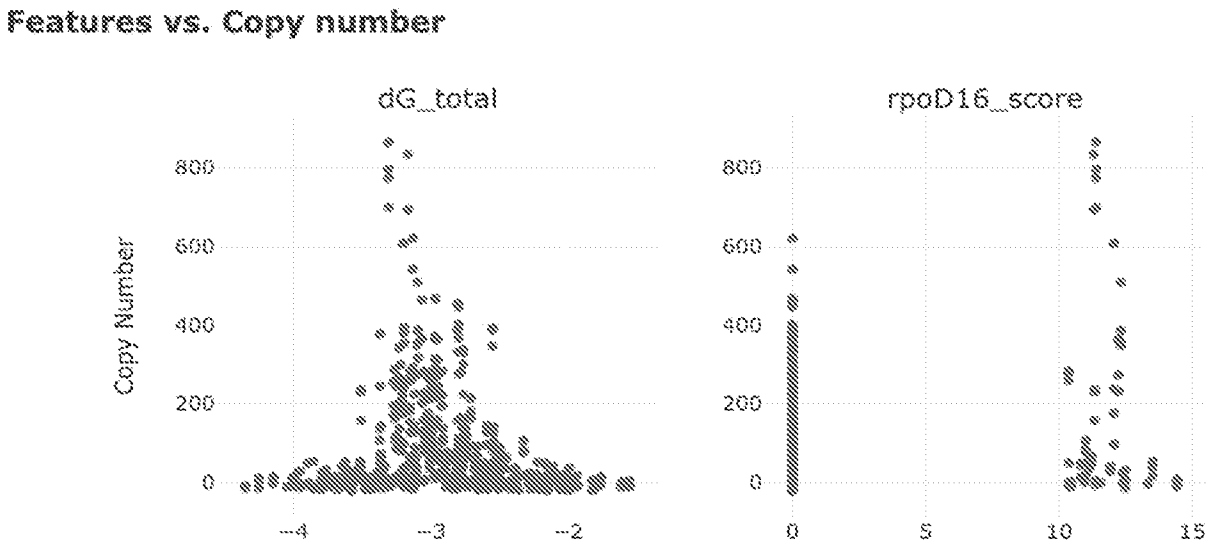


FIGURE 26A

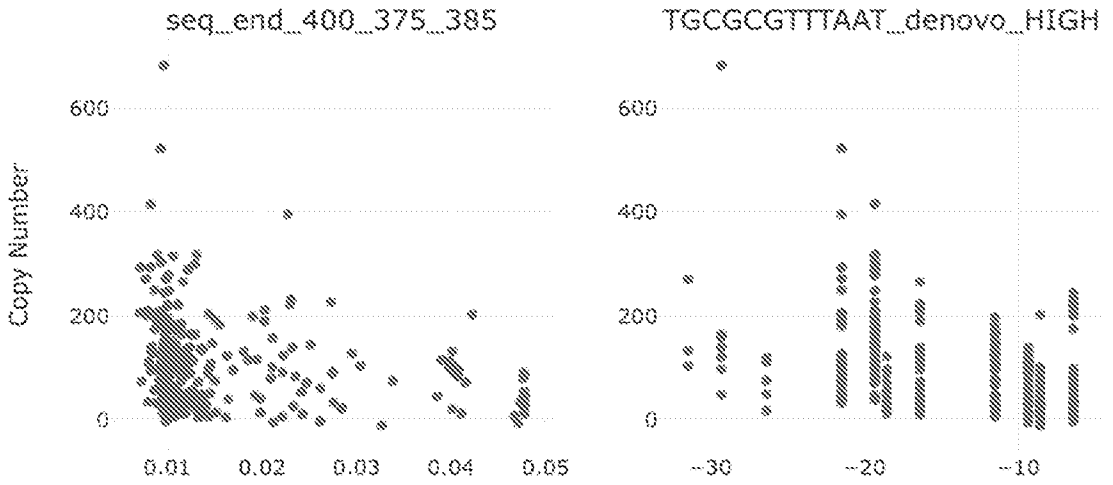


FIGURE 26B

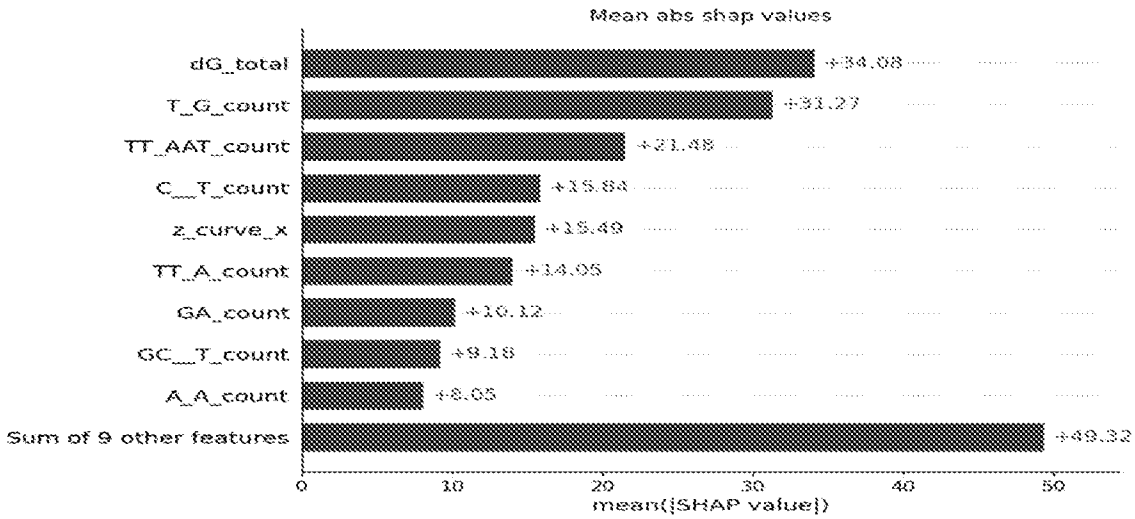


FIGURE 26C

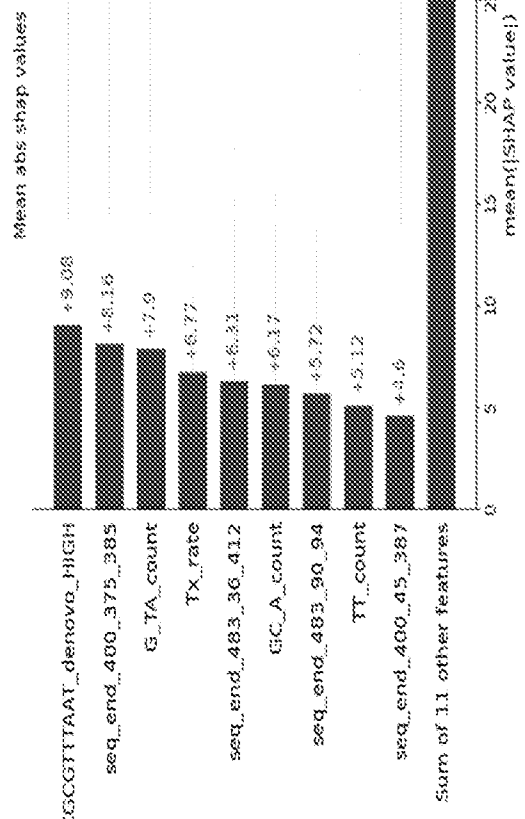


FIGURE 26D

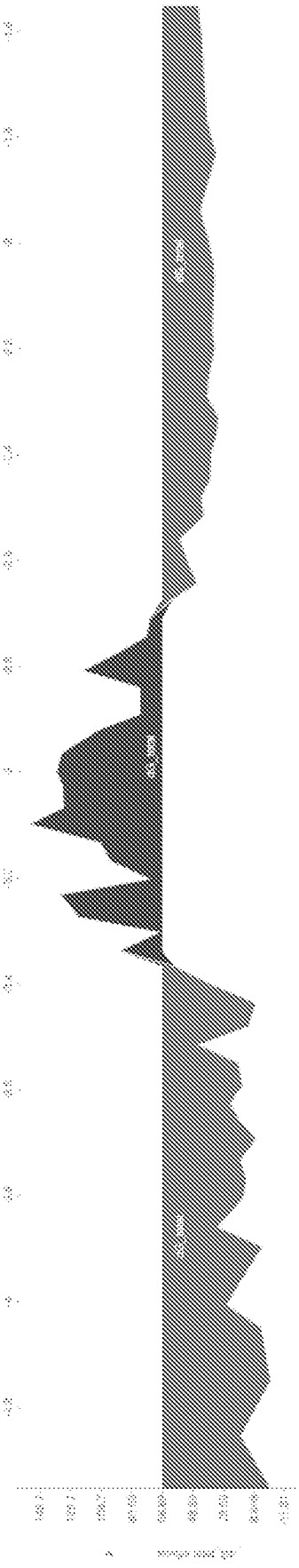


FIGURE 27A

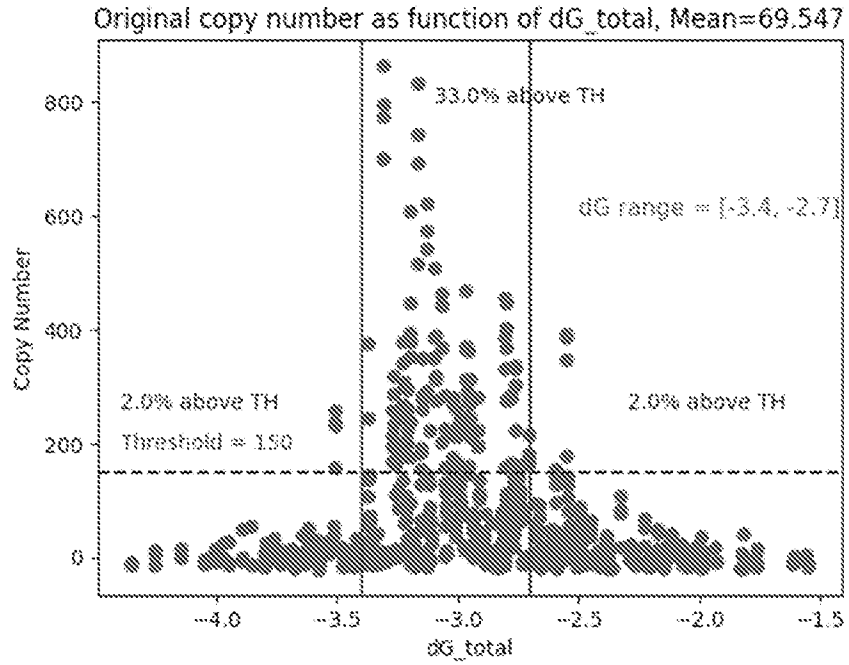


FIGURE 27B

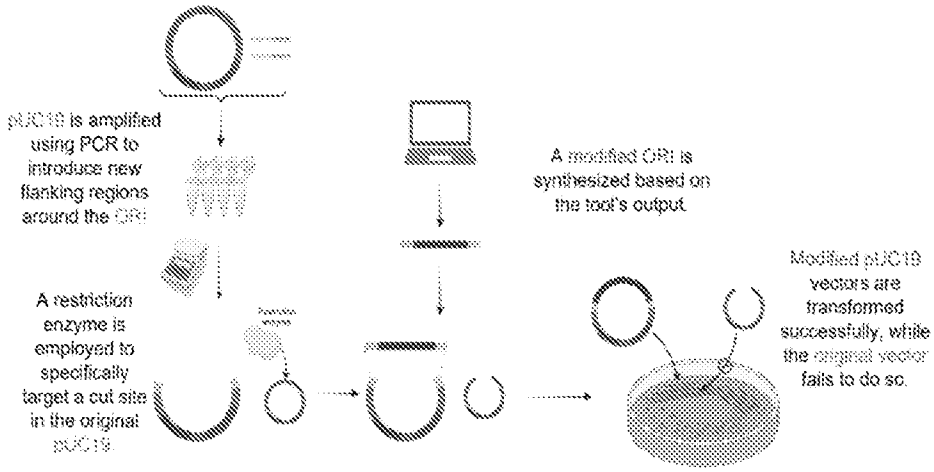


FIGURE 28A

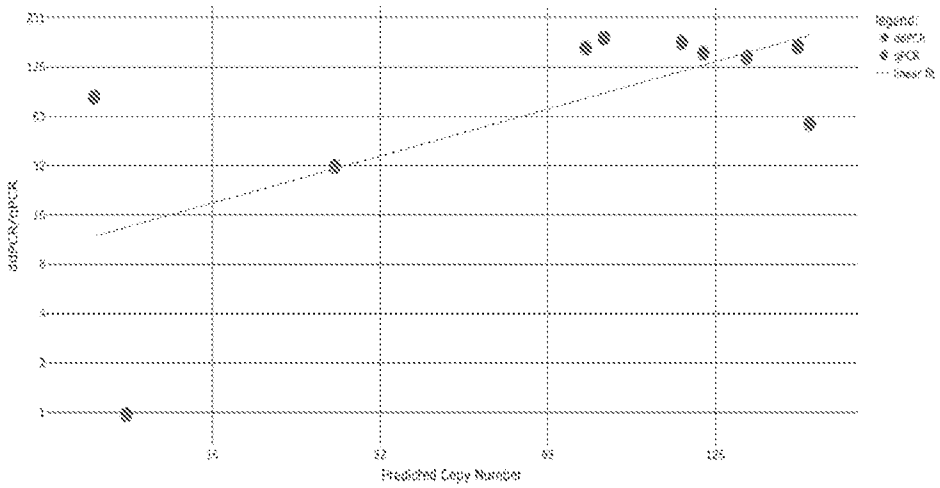


FIGURE 28B

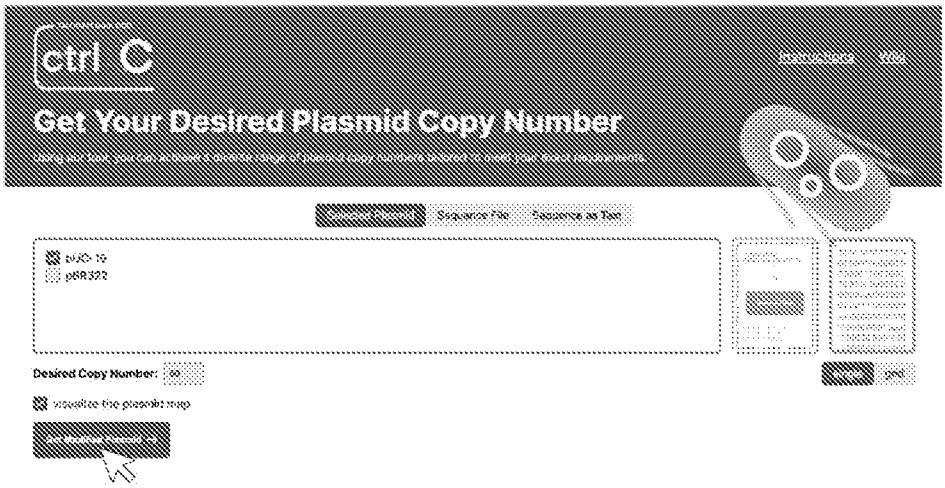


FIGURE 28C

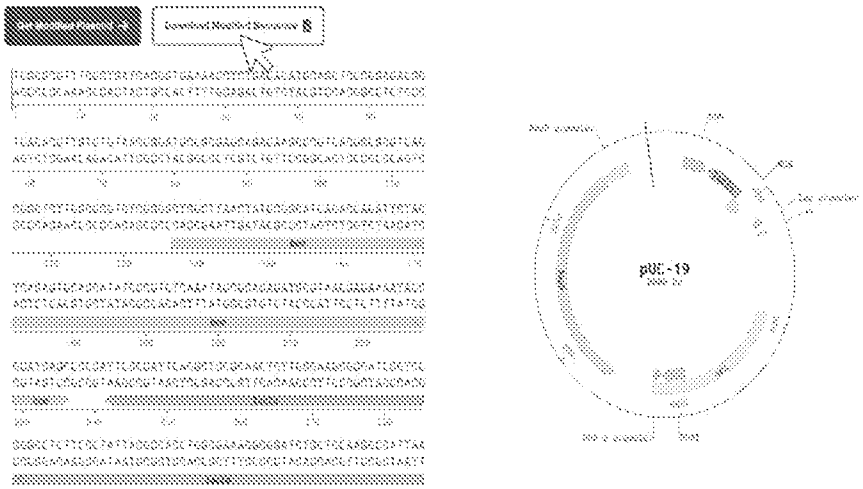


FIGURE 28D

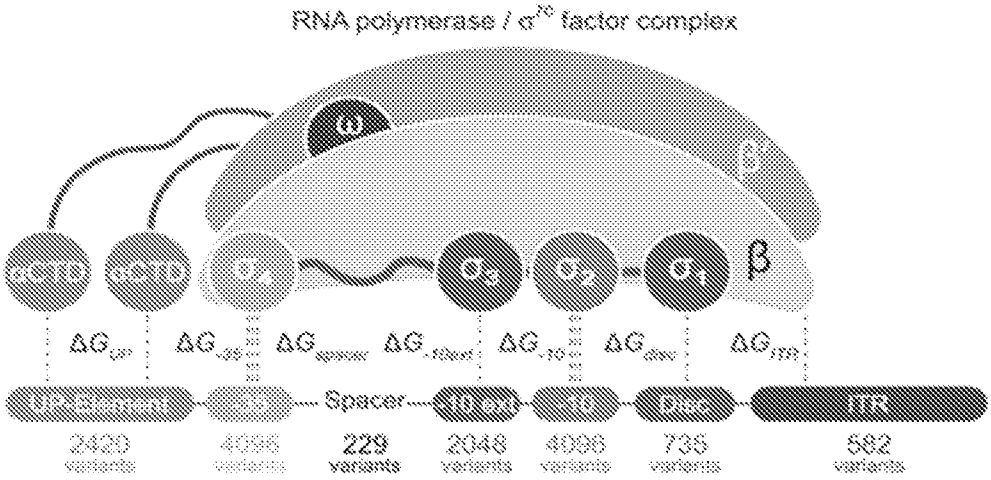


FIGURE 29A

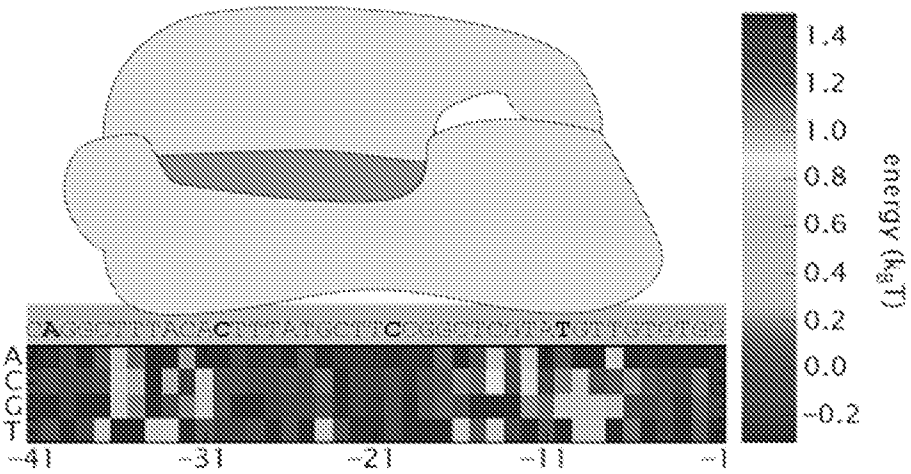


FIGURE 29B

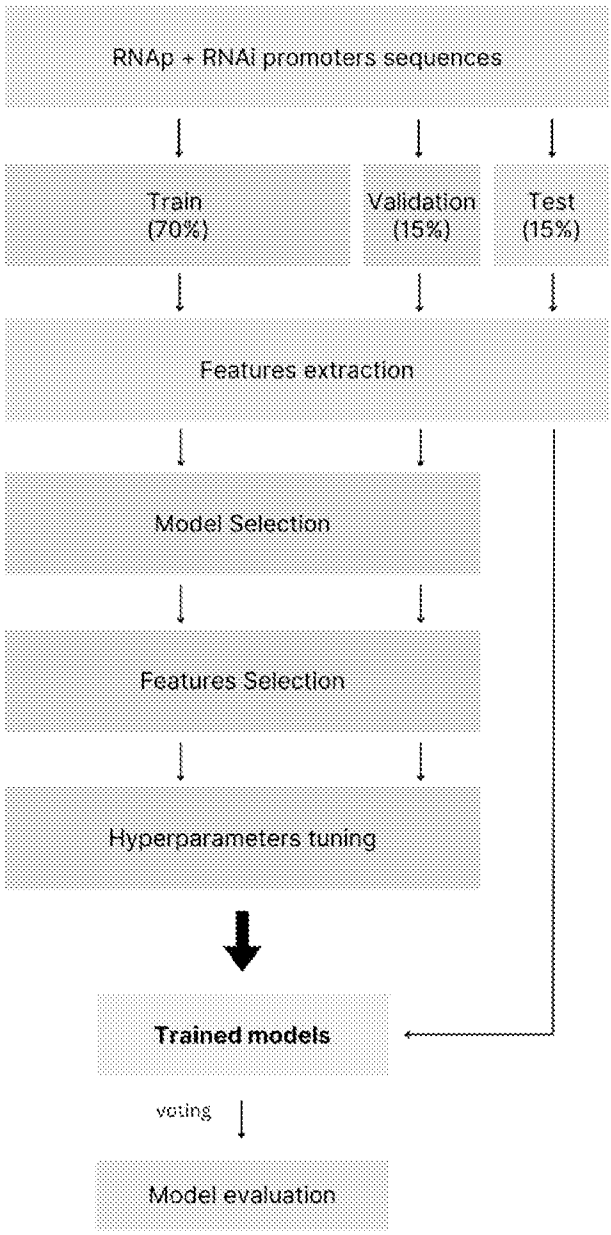
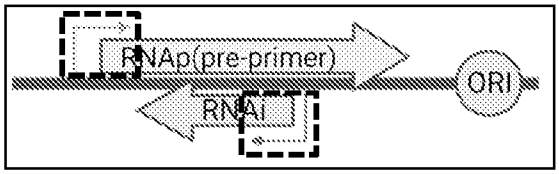


FIGURE 29C



Promoter Sequence	Copy Number
TTTCCCTCCTTTTTTTCTGCGCGTTTACTGCTGCTT	3.028
TTTAAATCCTTTTTTTCTGCGCGTAAGCTGCTGCTG	52.859
TTTACATCCTTTTTTTCTGCGCGTTAAATGCTGCTT	101.804
TTTCCCTCCTTTTTTTCTGCGCGTAAAATGCTGCTG	243.766
TTGAAATCCTTTTTTTCTGCGCGTAAGCTGCTGCTT	388.156
TTTAAATCCTTTTTTTCTGCGCGTAAAATGCTGCTT	620.393

FIGURE 29D

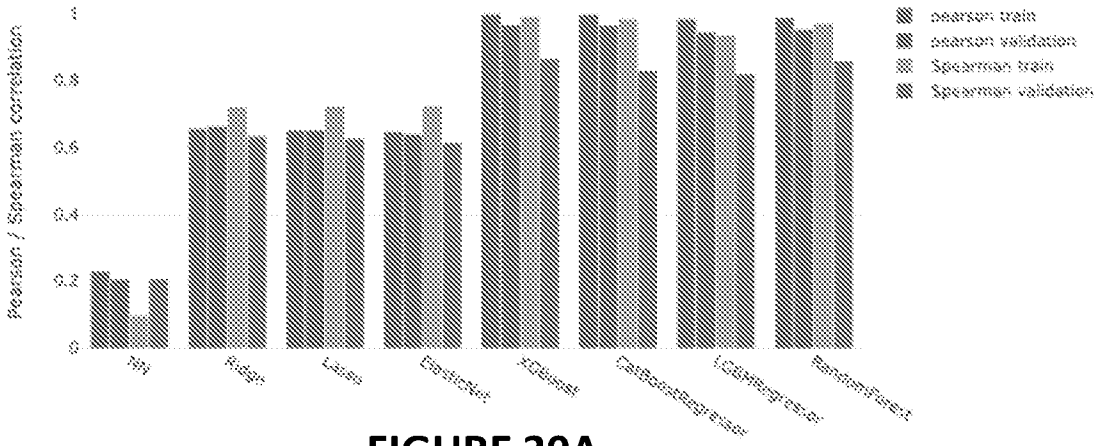


FIGURE 30A

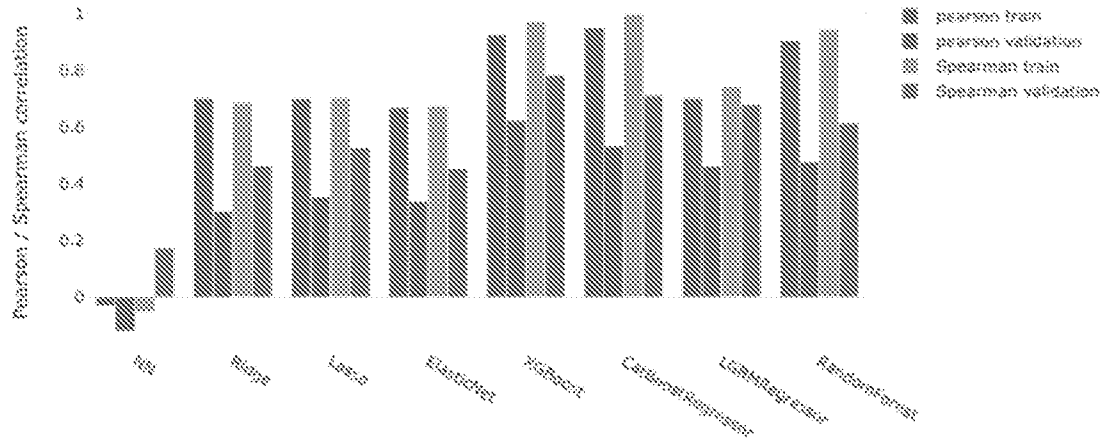


FIGURE 30B

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IL2024/050552

A. CLASSIFICATION OF SUBJECT MATTER

C12N 15/85(2024.01)i
CPC:C12N 15/85; C12N 2800/10; C12N 2830/55

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12N 15/85
CPC:C12N 15/85; C12N 2800/10; C12N 2830/55

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Databases consulted: Esp@cenet, Google Patents, Google Scholar Search terms used: Plasmid copy number ori rop mutations

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Rouches, M. V., Xu, Y., Cortes, L. B. G., and Lambert, G. (2022). A plasmid system with tunable copy number. Nature communications, 13(1), 3908.2022/07/07))	
Y	whole document	14,15
X	whole document, especially abstract; page 2, 8th para-page 3, 6th para; page 5, 3rd -5th para; page 7, 3rd para - page 8, 2nd para; figures 1 and 5	1,2,4,9-13,16-18
	Fang, D. I. A. N. N. E. (2004). Attempts to use PCR site-directed mutagenesis to create a non-functional rop gene in the plasmid pBR322. J. Exp. Microbiol. Immunol, 6, 45-51.2004/12/) 31)	
Y	abstract; page 45, 2nd para	14,15

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 August 2024	**12 August 2024**

Name and mailing address of the ISA/IL

Israel Patent Office
Technology Park, Bldg.5, Malcha, Jerusalem, 9695101,
Israel
Israel

Telephone No. **972-73-3927174**
Email: **pctoffice@justice.gov.il**

Authorized officer

MAZEL Alexander

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IL2024/050552

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
   a. ☑ forming part of the international application as filed.
   b. ☐ furnished subsequent to the international filing date for the purposes of international search (Rule 13*ter*.1(a)),
      ☐ accompanied by a statement to the effect that the sequence listing does not go beyond the disclosure in the international application as filed.

2. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, this report has been established to the extent that a meaningful search could be carried out without a WIPO Standard ST.26 compliant sequence listing.

3. Additional comments: