(51) International Patent Classification:
C12N 15/10 (2006.01)

(21) International Application Number:
PCT/US2018/036230

(22) International Filing Date:
06 June 2018 (06.06.2018)

(25) Filing Language:
English

(26) Publication Language:
English

(30) Priority Data:
62/515,965 06 June 2017 (06.06.2017) US


(72) Inventors: KELLY, Peter; c/o Zymegen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).
           ENYEART, Peter; c/o Zymegen Inc., 5980 Horton Street, Suite 105, Emeryville, California 94608 (US).


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

(54) Title: HIGH THROUGHPUT TRANSPOSON MUTAGENESIS

(57) Abstract: The present disclosure is directed to a method of high-throughput (HTP) microbial genomic engineering, which utilizes in vivo transposon mutagenesis to develop strain libraries for the perturbation of microbial phenotypes.
HIGH THROUGHPUT TRANSPOSON MUTAGENESIS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This Application claims the benefit of priority to U.S. Provisional Application No. 62/515,965, filed on June 6, 2017, the contents of which are hereby incorporated by reference in their entirety.

FIELD

[0002] The present disclosure is directed to a method of high-throughput (HTP) microbial genomic engineering, which utilizes in vivo transposon mutagenesis to develop strain libraries for the perturbation of microbial phenotypes.

STATEMENT REGARDING SEQUENCE LISTING

[0003] The Sequence Listing associated with this application is provided in text format in lieu of a paper copy, and is hereby incorporated by reference into the specification. The name of the text file containing the Sequence Listing is ZYMR_014_01WO_SeqList_ST25.txt. The text file is 14 KB, was created on June 6, 2018, and is being submitted electronically via EFS-Web.

BACKGROUND

[0004] Humans have been harnessing the power of microbial cellular biosynthetic pathways for millennia to produce products of interest, the oldest examples of which include alcohol, vinegar, cheese, and yogurt. These products are still in large demand today and have also been accompanied by an ever increasing repertoire of products producible by microbes. The advent of genetic engineering technology has enabled scientists to design and program novel biosynthetic pathways into a variety of organisms to produce a broad range of industrial, medical, and consumer products. Indeed, microbial cellular cultures are now used to produce products ranging from small molecules, antibiotics, vaccines, insecticides, enzymes, fuels, and industrial chemicals. Given the large number of products produced by modern industrial microbes, it comes as no surprise that
engineers are under tremendous pressure to improve the speed and efficiency by which a given microorganism is able to produce a target product.

[0005] A variety of approaches have been used to improve the economy of biologically-based industrial processes by “improving” the microorganism involved. For example, many industries rely on microbial strain improvement programs in which the parent strains of a microbial culture are continuously mutated through exposure to chemicals or UV radiation and are subsequently screened for performance increases, such as in productivity, yield and titer. This mutagenesis process is extensively repeated until a strain demonstrates a suitable increase in product performance. The subsequent “improved” strain is then utilized in commercial production.

[0006] However, identification of improved industrial microbial strains through a traditional mutagenesis process is time consuming and inefficient. The process, by its very nature, is haphazard, inefficient, and slow.

[0007] Thus, there is a need in the art for new methods of engineering microbes, which accelerate the process of discovering and consolidating beneficial mutations.

SUMMARY OF THE DISCLOSURE

[0008] The present disclosure addresses this need in the art, by providing a high-throughput (HTP) method of microbial genomic engineering, which offers dramatic improvements over the slow and inefficient methods currently practiced in the art.

[0009] The HTP microbial genomic engineering platform utilizes a suite of HTP toolsets to derive microbial strain libraries that allow for the fast and efficient identification of genetic perturbations leading to improved host phenotype. For instance, the HTP microbial genomic engineering platform described herein utilizes in vivo transposon mutagenesis to perturb the genome of host microbes, which enables the creation of diverse microbial strain libraries that can be utilized to improve host phenotype.

[0010] The disclosed HTP genomic engineering platform is computationally driven and integrates molecular biology, automation, and advanced machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets to create HTP genetic design libraries, which are derived from, inter alia, scientific insight and iterative pattern recognition.
[0011] As aforementioned, the taught HTP genetic design libraries function as drivers of the genomic engineering process, by providing libraries of particular genomic alterations for testing in a microbe. The microbes engineered utilizing a particular library, or combination of libraries, are efficiently screened in a HTP manner for a resultant outcome, e.g. production of a product of interest. This process of utilizing the HTP genetic design libraries to define particular genomic alterations for testing in a microbe and then subsequently screening host microbial genomes harboring the alterations is implemented in an efficient and iterative manner. In some aspects, the iterative cycle or “rounds” of genomic engineering campaigns can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more iterations/cycles/rounds.

[0012] Thus, in some aspects, the present disclosure teaches methods of conducting at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000 or more “rounds” of HTP genetic engineering (e.g., rounds of SNP swap, PRO swap, STOP swap, transposon mutagenesis, or combinations thereof).

[0013] In some embodiments the present disclosure teaches a linear approach, in which each subsequent HTP genetic engineering round is based on genetic variation identified in the previous round of genetic engineering. In other embodiments the present disclosure teaches a non-linear approach, in which each subsequent HTP genetic engineering round is based on genetic variation identified in any previous round of genetic engineering, including previously conducted analysis, and separate HTP genetic engineering branches.

[0014] The data from these iterative cycles enables large scale data analytics and pattern recognition, which is utilized by the integrative platform to inform subsequent rounds of HTP genetic design library implementation. Consequently, the HTP genetic design libraries utilized in the taught platform are highly dynamic tools that benefit from large scale data pattern recognition algorithms and become more informative through each iterative round of microbial engineering.
[0015] In some embodiments, the genetic design libraries of the present disclosure comprise at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000 or more individual genetic changes (e.g., at least X number of promoter: gene combinations in the PRO swap library or transposon gain-of-function libraries).

[0016] In some embodiments, the present disclosure teaches a high-throughput (HTP) method of genomic engineering to evolve a microbe to acquire a desired phenotype, comprising: a) perturbing the genomes of an initial plurality of microbes having the same microbial strain background using transposon mutagenesis, to thereby create an initial HTP genetic design transposon mutagenesis microbial strain library comprising individual microbial strains with unique genetic variations; b) screening and selecting individual microbial strains of the initial HTP genetic design transposon mutagenesis microbial strain library for the desired phenotype; c) providing a subsequent plurality of microbes that each comprise a unique combination of genetic variation, the genetic variation selected from the genetic variation present in at least two individual microbial strains screened in the preceding step, to thereby create a subsequent HTP genetic design transposon mutagenesis microbial strain library; d) screening and selecting individual microbial strains of the subsequent HTP genetic design transposon mutagenesis microbial strain library for the desired phenotype; e) repeating steps c)-d) one or more times, in a linear or non-linear fashion, until a microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new HTP genetic design transposon mutagenesis microbial strain library comprising individual microbial strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual microbial strains of a preceding HTP genetic design transposon mutagenesis microbial strain library.

[0017] In some embodiments, the present disclosure teaches methods of making a subsequent plurality of microbes that each comprise a unique combination of genetic variations, wherein each of the combined genetic variations is derived from the initial HTP genetic design transposon
mutagenesis microbial strain library or the HTP genetic design transposon mutagenesis microbial strain library of the preceding step.

[0018] In some embodiments, the combination of genetic variations in the subsequent plurality of microbes will comprise a subset of all the possible combinations of the genetic variations in the initial HTP genetic design transposon mutagenesis microbial strain library or the HTP genetic design transposon mutagenesis microbial strain library of the preceding step.

[0019] In some embodiments, the present disclosure teaches that the subsequent HTP genetic design microbial strain library is a partial combinatorial microbial strain library derived from the genetic variations in the initial HTP genetic design microbial strain library or the HTP genetic design microbial strain library of the preceding step.

[0020] For example, if the prior HTP genetic design microbial strain library only had genetic variations A, B, C, and D, then a partial combinatorial of the variations could include a subsequent HTP genetic design microbial strain library comprising three microbes each comprising either the AB, AC, or AD unique combinations of genetic variations (order in which the mutations are represented is unimportant). A full combinatorial microbial strain library derived from the genetic variations of the HTP genetic design library of the preceding step would include six microbes, each comprising either AB, AC, AD, BC, BD, or CD unique combinations of genetic variations.

[0021] In some embodiments, the methods of the present disclosure teach perturbing the genome utilizing at least one method selected from the group consisting of: random mutagenesis, targeted sequence insertions, targeted sequence deletions, targeted sequence replacements, transposon mutagenesis, or any combination thereof.

[0022] In some embodiments of the presently disclosed methods, the initial plurality of microbes comprise unique genetic variations derived from an industrial production strain microbe.

[0023] In some embodiments of the presently disclosed methods, the initial plurality of microbes comprise industrial production strain microbes denoted SiGen1 and any number of subsequent microbial generations derived therefrom denoted SiGenn.

[0024] In some embodiments, the present disclosure teaches a transposon mutagenesis method of genomic engineering to evolve a microbe to acquire a desired phenotype, the method comprising the steps of: a) providing a transposase enzyme and a DNA payload sequence. In some
embodiments, the transposase enzyme and DNA payload sequence form a transposase-DNA payload complex. In some embodiments, the transposon mutagenesis results in random insertion of a transposon into the genome of the plurality of microbes. In some embodiments, the transposase is derived from EZ-Tn5 transposon system. In some embodiments, the DNA payload sequence is flanked by mosaic elements (ME) that can be recognized by the transposase. The specific sequence of the DNA payload can be varied to bias toward a loss of function or gain of function effect of transposon insertion into the target genome.

[0025] In some embodiments, the transposon mutagenesis causes a loss-of-function (LoF) or a gain-of-function (GoF) phenotype. In some embodiments, the DNA payload can be a loss-of-function (LoF) transposon, or a gain-of-function (GoF) transposon. In some embodiments, the DNA payload comprises a selection marker. In some embodiments, the selection marker is antibiotic resistance. In some embodiments, the DNA payload comprises a counter-selection marker. In some embodiments, the counter-selection marker is used to facilitate loop-out of a DNA payload containing the selectable marker, which enables marker recycling and thus further rounds of engineering. In some embodiments, the GoF transposon comprises a GoF element. In some embodiments, the GoF transposon comprises a promoter sequence and/or a solubility tag sequence. In some embodiments, the GoF transposon comprises an antibiotic marker and a strong promoter. In some embodiments, the methods further comprise b) combining the transposase and the DNA payload sequence to form a complex, and c) transforming the transposase-DNA payload complex to a microbial strain, thus resulting in random integration of the DNA payload sequence in the genome of the microbial strain. In some embodiments, strains comprising the random integration of DNA payload form an initial transposon mutagenesis library.

[0026] In some embodiments, the methods further comprise d) screening and selecting individual microbial strains of the initial transposon mutagenesis microbial strain library for the desired phenotype. In some embodiments, the methods further comprise e) providing a subsequent plurality of microbes that each comprise a unique combination of genetic variation, the genetic variation selected from the genetic variation present in at least two individual microbial strains screened in the preceding step, to thereby create a subsequent transposon mutagenesis microbial strain library. In some embodiments, the methods further comprise f) screening and selecting individual microbial strains of the subsequent transposon mutagenesis microbial strain library for the desired phenotype. In some embodiments, the methods further comprise g) repeating steps e)-
f) one or more times, in a linear or non-linear fashion, until a microbe has acquired the desired phenotype, wherein each subsequent iteration creates a new transposon mutagenesis microbial strain library comprising individual microbial strains harboring unique genetic variations that are a combination of genetic variation selected from amongst at least two individual microbial strains of a preceding transposon mutagenesis microbial strain library.

[0027] In some embodiments, the present disclosure teaches iteratively improving the design of candidate microbial strains by (a) accessing a predictive model populated with a training set comprising (1) inputs representing genetic changes to one or more background microbial strains and (2) corresponding performance measures; (b) applying test inputs to the predictive model that represent genetic changes, the test inputs corresponding to candidate microbial strains incorporating those genetic changes; (c) predicting phenotypic performance of the candidate microbial strains based at least in part upon the predictive model; (d) selecting a first subset of the candidate microbial strains based at least in part upon their predicted performance; (e) obtaining measured phenotypic performance of the first subset of the candidate microbial strains; (f) obtaining a selection of a second subset of the candidate microbial strains based at least in part upon their measured phenotypic performance; (g) adding to the training set of the predictive model (1) inputs corresponding to the selected second subset of candidate microbial strains, along with (2) corresponding measured performance of the selected second subset of candidate microbial strains; and (h) repeating (b)-(g) until measured phenotypic performance of at least one candidate microbial strain satisfies a performance metric. In some cases, during a first application of test inputs to the predictive model, the genetic changes represented by the test inputs comprise genetic changes to the one or more background microbial strains; and during subsequent applications of test inputs, the genetic changes represented by the test inputs comprise genetic changes to candidate microbial strains within a previously selected second subset of candidate microbial strains.

[0028] In some embodiments, selection of the first subset may be based on epistatic effects. This may be achieved by: during a first selection of the first subset: determining degrees of dissimilarity between performance measures of the one or more background microbial strains in response to application of a plurality of respective inputs representing genetic changes to the one or more background microbial strains; and selecting for inclusion in the first subset at least two candidate microbial strains based at least in part upon the degrees of dissimilarity in the performance
measures of the one or more background microbial strains in response to application of genetic changes incorporated into the at least two candidate microbial strains.

[0029] In some embodiments, the present disclosure teaches applying epistatic effects in the iterative improvement of candidate microbial strains, the method comprising: obtaining data representing measured performance in response to corresponding genetic changes made to at least one microbial background strain; obtaining a selection of at least two genetic changes based at least in part upon a degree of dissimilarity between the corresponding responsive performance measures of the at least two genetic changes, wherein the degree of dissimilarity relates to the degree to which the at least two genetic changes affect their corresponding responsive performance measures through different biological pathways; and designing genetic changes to a microbial background strain that include the selected genetic changes. In some cases, the microbial background strain for which the at least two selected genetic changes are designed is the same as the at least one microbial background strain for which data representing measured responsive performance was obtained.

[0030] In some embodiments, the present disclosure teaches HTP strain improvement methods utilizing only a single type of genetic microbial library. For example, in some embodiments, the present disclosure teaches HTP strain improvement methods utilizing only transposon mutagenesis libraries.

[0031] In other embodiments, the present disclosure teaches HTP strain improvement methods utilizing two or more types of genetic microbial libraries. For example, in some embodiments, the present disclosure teaches HTP strain improvement methods combining SNP swap and transposon mutagenesis libraries. In some embodiments, the present disclosure teaches HTP strain improvement methods combining PRO swap and transposon mutagenesis libraries. In some embodiments, the present disclosure teaches HTP strain improvement methods combining STOP swap and transposon mutagenesis libraries. In yet other embodiments, the HTP strain improvement methods of the present disclosure can be combined with one or more traditional strain improvement methods.

[0032] In some embodiments, the HTP strain improvement methods of the present disclosure result in an improved host cell. That is, the present disclosure teaches methods of improving one or more host cell properties. In some embodiments the improved host cell property is selected from
the group consisting of: volumetric productivity, specific productivity, yield or titer, of a product of interest produced by the host cell. In some embodiments, the improved host cell property is volumetric productivity. In some embodiments, the improved host cell property is specific productivity. In some embodiments, the improved host cell property is yield.

[0033] In some embodiments, the HTP strain improvement methods of the present disclosure result in a host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one host cell property over a control host cell that is not subjected to the HTP strain improvements methods (e.g., an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therein). In some embodiments, the HTP strain improvement methods of the present disclosure are selected from the group consisting of: SNP swap, PRO swap, STOP swap, transposon mutagenesis, and combinations thereof.

[0034] In some embodiments, the transposon mutagenesis methods of the present disclosure result in a host cell that exhibits a 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%, 150%, 200%, 250%, 300% or more of an improvement in at least one host cell property over a control host cell that is not subjected to the transposon mutagenesis methods (e.g., an X% improvement in yield or productivity of a biomolecule of interest, incorporating any ranges and subranges therein).

BRIEF DESCRIPTION OF THE FIGURES
[0035] FIGURE 1 depicts a DNA recombination method of the present disclosure for increasing variation in diversity pools. DNA sections, such as genome regions from related species, can be cut via physical or enzymatic/chemical means. The cut DNA regions are melted and allowed to reanneal, such that overlapping genetic regions prime polymerase extension reactions. Subsequent melting/extension reactions are carried out until products are reassembled into chimeric DNA, comprising elements from one or more starting sequences.

[0036] FIGURE 2 outlines methods of the present disclosure for generating new host organisms with selected sequence modifications (e.g., 100 SNPs to swap). Briefly, the method comprises (1) desired DNA inserts are designed and generated by combining one or more synthesized oligos in an assembly reaction, (2) DNA inserts are cloned into transformation plasmids, (3) completed plasmids are transferred into desired production strains, where they are integrated into the host strain genome, and (4) selection markers and other unwanted DNA elements are looped out of the host strain. Each DNA assembly step may involve additional quality control (QC) steps, such as cloning plasmids into E.coli bacteria for amplification and sequencing.

[0037] FIGURE 3 depicts assembly of transformation plasmids of the present disclosure, and their integration into host organisms. The insert DNA is generated by combining one or more synthesized oligos in an assembly reaction. DNA inserts containing the desired sequence are flanked by regions of DNA homologous to the targeted region of the genome. These homologous regions facilitate genomic integration, and, once integrated, form direct repeat regions designed for looping out vector backbone DNA in subsequent steps. Assembled plasmids contain the insert DNA, and optionally, one or more selection markers.

[0038] FIGURES 4A-B depict the DNA assembly, transformation, and strain screening steps of one of the embodiments of the present disclosure. FIGURE 4A depicts the steps for building DNA fragments, cloning the DNA fragments into vectors, transforming the vectors into host strains, and looping out selection sequences through counter selection. FIGURE 4B depicts the steps for high-throughput culturing, screening, and evaluation of selected host strains. This figure also depicts the optional steps of culturing, screening, and evaluating selected strains in culture tanks.

[0039] FIGURE 5 depicts one embodiment of the automated system of the present disclosure. The present disclosure teaches use of automated robotic systems with various modules capable of cloning, transforming, culturing, screening and/or sequencing host organisms.
[0040] FIGURE 6 depicts the results of a second round HTP engineering PRO swap program. Top promoter::gene combinations identified during the first PRO swap round were analyzed according to the methods of the present disclosure to identify combinations of the mutations that would be likely to exhibit additive or combinatorial beneficial effects on host performance. Second round PRO swap mutants thus comprised pair combinations of various promoter::gene mutations. The resulting second round mutants were screened for differences in host cell yield of a selected biomolecule. A combination pair of mutations that had been predicted to exhibit beneficial effects is emphasized with a circle.

[0041] FIGURE 7 is a similarity matrix computed using the correlation measure. The matrix is a representation of the functional similarity between SNP variants. The consolidation of SNPs with low functional similarity is expected to have a higher likelihood of improving strain performance, as opposed to the consolidation of SNPs with higher functional similarity.

[0042] FIGURES 8A-B depict the results of an epistasis mapping experiment. Combination of SNPs and PRO swaps with low functional similarities yields improved strain performance. FIGURE 8A depicts a dendrogram clustered by functional similarity of all the SNPs/PRO swaps. FIGURE 8B depicts host strain performance of consolidated SNPs as measured by product yield. Greater cluster distance correlates with improved consolidation performance of the host strain.

[0043] FIGURES 9A-B depict SNP differences among strain variants in the diversity pool. FIGURE 9A depicts the relationship among the strains of this experiment. Strain A is the wild-type host strain. Strain B is an intermediate engineered strain. Strain C is the industrial production strain. FIGURE 9B is a graph identifying the number of unique and shared SNPs in each strain.

[0044] FIGURE 10 illustrates the distribution of relative strain performances for the input data under consideration. A relative performance of zero indicates that the engineered strain performed equally well to the in-plate base strain. The processes described herein are designed to identify the strains that are likely to perform significantly above zero.

[0045] FIGURE 11 illustrates example gene targets to be utilized in a promoter swap process.

[0046] FIGURE 12 illustrates an exemplary promoter library that is being utilized to conduct a promoter swap process for the identified gene targets. Promoters utilized in the PRO swap (i.e. promoter swap) process are P_{1}-P_{s}, the sequences and identity of which can be found in Table 1.
[0047] FIGURE 13 illustrates that promoter swapping genetic outcomes depend on the particular gene being targeted.

[0048] FIGURE 14 illustrates the composition of changes for the top 100 predicted strain designs. The x-axis lists the pool of potential genetic changes (dss mutations are SNP swaps, and Pgc mutations are PRO swaps), and the y-axis shows the rank order. Black cells indicate the presence of a particular change in the candidate design, while white cells indicate the absence of that change. In this particular example, all of the top 100 designs contain the changes pga3121_pgi, pga1860_pyc, dss_339, and pga0007_39_lysa. Additionally, the top candidate design contains the changes dss_034, dss_009.

[0049] FIGURE 15 depicts the DNA assembly and transformation steps of one of the embodiments of the present disclosure. The flow chart depicts the steps for building DNA fragments, cloning the DNA fragments into vectors, transforming the vectors into host strains, and looping out selection sequences through counter selection.

[0050] FIGURE 16 depicts the steps for high-throughput culturing, screening, and evaluation of selected host strains. This figure also depicts the optional steps of culturing, screening, and evaluating selected strains in culture tanks.

[0051] FIGURE 17 depicts expression profiles of illustrative promoters exhibiting a range of regulatory expression, according to the promoter ladders of the present disclosure. Promoter A expression peaks at the lag phase of bacterial cultures, while promoter B and C peak at the exponential and stationary phase, respectively.

[0052] FIGURE 18 depicts expression profiles of illustrative promoters exhibiting a range of regulatory expression, according to the promoter ladders of the present disclosure. Promoter A expression peaks immediately upon addition of a selected substrate, but quickly returns to undetectable levels as the concentration of the substrate is reduced. Promoter B expression peaks immediately upon addition of the selected substrate and lowers slowly back to undetectable levels together with the corresponding reduction in substrate. Promoter C expression peaks upon addition of the selected substrate, and remains highly expressed throughout the culture, even after the substrate has dissipated.
[0053] **FIGURE 19** depicts expression profiles of illustrative promoters exhibiting a range of constitutive expression levels, according to the promoter ladders of the present disclosure. Promoter A exhibits the lowest expression, followed by increasing expression levels promoter B and C, respectively.

[0054] **FIGURE 20** diagrams an embodiment of LIMS system of the present disclosure for strain improvement.

[0055] **FIGURE 21** diagrams a cloud computing implementation of embodiments of the LIMS system of the present disclosure.

[0056] **FIGURE 22** depicts an embodiment of the iterative predictive strain design workflow of the present disclosure.

[0057] **FIGURE 23** diagrams an embodiment of a computer system, according to embodiments of the present disclosure.

[0058] **FIGURE 24** is a flowchart illustrating the consideration of epistatic effects in the selection of mutations for the design of a microbial strain, according to embodiments of the disclosure.

[0059] **FIGURE 25** depicts linear maps of plasmids for transposon mutagenesis in *S. spinosa*. Loss-of-Function (LoF) transposon, Gain-of-Function (GoF) transposon, and Gain-of-Function (GoF) Recyclable Transposon are shown.

**DETAILED DESCRIPTION**

**Definitions**

[0060] While the following terms are believed to be well understood by one of ordinary skill in the art, the following definitions are set forth to facilitate explanation of the presently disclosed subject matter.

[0061] The term “a” or “an” refers to one or more of that entity, *i.e.* can refer to a plural referents. As such, the terms “a” or “an”, “one or more” and “at least one” are used interchangeably herein. In addition, reference to “an element” by the indefinite article “a” or “an” does not exclude the possibility that more than one of the elements is present, unless the context clearly requires that there is one and only one of the elements.
[0062] As used herein the terms “cellular organism” “microorganism” or “microbe” should be taken broadly. These terms are used interchangeably and include, but are not limited to, the two prokaryotic domains, Bacteria and Archaea, as well as certain eukaryotic fungi and protists. In some embodiments, the disclosure refers to the “microorganisms” or “cellular organisms” or “microbes” of lists/tables and figures present in the disclosure. This characterization can refer to not only the identified taxonomic genera of the tables and figures, but also the identified taxonomic species, as well as the various novel and newly identified or designed strains of any organism in the tables or figures. The same characterization holds true for the recitation of these terms in other parts of the Specification, such as in the Examples.

[0063] The term “prokaryotes” is art recognized and refers to cells which contain no nucleus or other cell organelles. The prokaryotes are generally classified in one of two domains, the Bacteria and the Archaea. The definitive difference between organisms of the Archaea and Bacteria domains is based on fundamental differences in the nucleotide base sequence in the 16S ribosomal RNA.

[0064] The term “Archaea” refers to a categorization of organisms of the division Mendosicutes, typically found in unusual environments and distinguished from the rest of the prokaryotes by several criteria, including the number of ribosomal proteins and the lack of muramic acid in cell walls. On the basis of ssrRNA analysis, the Archaea consist of two phylogenetically-distinct groups: Crenarchaeota and Euryarchaeota. On the basis of their physiology, the Archaea can be organized into three types: methanogens (prokaryotes that produce methane); extreme halophiles (prokaryotes that live at very high concentrations of salt (NaCl); and extreme (hyper) thermophilus (prokaryotes that live at very high temperatures). Besides the unifying archaeal features that distinguish them from Bacteria (i.e., no murein in cell wall, ester-linked membrane lipids, etc.), these prokaryotes exhibit unique structural or biochemical attributes which adapt them to their particular habitats. The Crenarchaeota consists mainly of hyperthermophilic sulfur-dependent prokaryotes and the Euryarchaeota contains the methanogens and extreme halophiles.

[0065] “Bacteria” or “eubacteria” refers to a domain of prokaryotic organisms. Bacteria include at least 11 distinct groups as follows: (1) Gram-positive (gram+) bacteria, of which there are two major subdivisions: (1) high G+C group (Actinomycetes, Mycobacteria, Micrococcus, others) (2) low G+C group (Bacillus, Clostridia, Lactobacillus, Staphylococci, Streptococci, Mycoplasmas);
(2) Proteobacteria, *e.g.*, Purple photosynthetic+non-photosynthetic Gram-negative bacteria (includes most “common” Gram-negative bacteria); (3) Cyanobacteria, *e.g.*, oxygenic phototrophs; (4) Spirochetes and related species; (5) Planctomyces; (6) *Bacteroides, Flavobacteria*; (7) *Chlamydia*; (8) Green sulfur bacteria; (9) Green non-sulfur bacteria (also anaerobic phototrophs); (10) Radioreistant micrococci and relatives; (11) *Thermotoga* and *Thermosiphon thermophiles*.

[0066] A “eukaryote” is any organism whose cells contain a nucleus and other organelles enclosed within membranes. Eukaryotes belong to the taxon Eukarya or Eukaryota. The defining feature that sets eukaryotic cells apart from prokaryotic cells (the aforementioned Bacteria and Archaea) is that they have membrane-bound organelles, especially the nucleus, which contains the genetic material, and is enclosed by the nuclear envelope.

[0067] The terms “genetically modified host cell,” “recombinant host cell,” and “recombinant strain” are used interchangeably herein and refer to host cells that have been genetically modified by the cloning and transformation methods of the present disclosure. Thus, the terms include a host cell (*e.g.*, bacteria, yeast cell, fungal cell, CHO, human cell, etc.) that has been genetically altered, modified, or engineered, such that it exhibits an altered, modified, or different genotype and/or phenotype (*e.g.*, when the genetic modification affects coding nucleic acid sequences of the microorganism), as compared to the naturally-occurring organism from which it was derived. It is understood that in some embodiments, the terms refer not only to the particular recombinant host cell in question, but also to the progeny or potential progeny of such a host cell.

[0068] The term “wild-type microorganism” or “wild-type host cell” describes a cell that occurs in nature, *i.e.* a cell that has not been genetically modified.

[0069] The term “genetically engineered” may refer to any manipulation of a host cell’s genome (*e.g.* by insertion, deletion, mutation, or replacement of nucleic acids).

[0070] The term “control” or “control host cell” refers to an appropriate comparator host cell for determining the effect of a genetic modification or experimental treatment. In some embodiments, the control host cell is a wild type cell. In other embodiments, a control host cell is genetically identical to the genetically modified host cell, save for the genetic modification(s) differentiating the treatment host cell. In some embodiments, the present disclosure teaches the use of parent strains as control host cells (*e.g.*, the S1 strain that was used as the basis for the strain improvement
program). In other embodiments, a host cell may be a genetically identical cell that lacks a specific promoter or SNP being tested in the treatment host cell.

[0071] As used herein, the term “allele(s)” means any of one or more alternative forms of a gene, all of which alleles relate to at least one trait or characteristic. In a diploid cell, the two alleles of a given gene occupy corresponding loci on a pair of homologous chromosomes.

[0072] As used herein, the term “locus” (loci plural) means a specific place or places or a site on a chromosome where for example a gene or genetic marker is found.

[0073] As used herein, the term “genetically linked” refers to two or more traits that are co-inherited at a high rate during breeding such that they are difficult to separate through crossing.

[0074] A “recombination” or “recombination event” as used herein refers to a chromosomal crossing over or independent assortment.

[0075] As used herein, the term “phenotype” refers to the observable characteristics of an individual cell, cell culture, organism, or group of organisms which results from the interaction between that individual’s genetic makeup (i.e., genotype) and the environment.

[0076] As used herein, the term “chimeric” or “recombinant” when describing a nucleic acid sequence or a protein sequence refers to a nucleic acid, or a protein sequence, that links at least two heterologous polynucleotides, or two heterologous polypeptides, into a single macromolecule, or that re-arranges one or more elements of at least one natural nucleic acid or protein sequence. For example, the term “recombinant” can refer to an artificial combination of two otherwise separated segments of sequence, e.g., by chemical synthesis or by the manipulation of isolated segments of nucleic acids by genetic engineering techniques.

[0077] As used herein, a “synthetic nucleotide sequence” or “synthetic polynucleotide sequence” is a nucleotide sequence that is not known to occur in nature or that is not naturally occurring. Generally, such a synthetic nucleotide sequence will comprise at least one nucleotide difference when compared to any other naturally occurring nucleotide sequence.

[0078] As used herein, the term “nucleic acid” refers to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides, or analogs thereof. This term refers to the primary structure of the molecule, and thus includes double- and single-stranded DNA, as well as double- and single-stranded RNA. It also includes modified nucleic acids such as methylated
and/or capped nucleic acids, nucleic acids containing modified bases, backbone modifications, and the like. The terms “nucleic acid” and “nucleotide sequence” are used interchangeably.

[0079] As used herein, the term “gene” refers to any segment of DNA associated with a biological function. Thus, genes include, but are not limited to, coding sequences and/or the regulatory sequences required for their expression. Genes can also include non-expressed DNA segments that, for example, form recognition sequences for other proteins. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters.

[0080] As used herein, the term “homologous” or “homologue” or “ortholog” is known in the art and refers to related sequences that share a common ancestor or family member and are determined based on the degree of sequence identity. The terms “homology,” “homologous,” “substantially similar” and “corresponding substantially” are used interchangeably herein. They refer to nucleic acid fragments wherein changes in one or more nucleotide bases do not affect the ability of the nucleic acid fragment to mediate gene expression or produce a certain phenotype. These terms also refer to modifications of the nucleic acid fragments of the instant disclosure such as deletion or insertion of one or more nucleotides that do not substantially alter the functional properties of the resulting nucleic acid fragment relative to the initial, unmodified fragment. It is therefore understood, as those skilled in the art will appreciate, that the disclosure encompasses more than the specific exemplary sequences. These terms describe the relationship between a gene found in one species, subspecies, variety, cultivar or strain and the corresponding or equivalent gene in another species, subspecies, variety, cultivar or strain. For purposes of this disclosure homologous sequences are compared. “Homologous sequences” or “homologues” or “orthologs” are thought, believed, or known to be functionally related. A functional relationship may be indicated in any one of a number of ways, including, but not limited to: (a) degree of sequence identity and/or (b) the same or similar biological function. Preferably, both (a) and (b) are indicated. Homology can be determined using software programs readily available in the art, such as those discussed in Current Protocols in Molecular Biology (F.M. Ausubel et al., eds., 1987) Supplement 30, section 7.718, Table 7.71. Some alignment programs are MacVector (Oxford Molecular Ltd, Oxford, U.K.), ALIGN Plus (Scientific and Educational Software, Pennsylvania) and AlignX (Vector NTI, Invitrogen, Carlsbad, CA). Another alignment program is Sequencher (Gene Codes, Ann Arbor, Michigan), using default parameters.
[0081] As used herein, the term “endogenous” or “endogenous gene,” refers to the naturally occurring gene, in the location in which it is naturally found within the host cell genome. In the context of the present disclosure, operably linking a heterologous promoter to an endogenous gene means genetically inserting a heterologous promoter sequence in front of an existing gene, in the location where that gene is naturally present. An endogenous gene as described herein can include alleles of naturally occurring genes that have been mutated according to any of the methods of the present disclosure.

[0082] As used herein, the term “exogenous” is used interchangeably with the term “heterologous,” and refers to a substance coming from some source other than its native source. For example, the terms “exogenous protein,” or “exogenous gene” refer to a protein or gene from a non-native source or location, and that have been artificially supplied to a biological system.

[0083] As used herein, the term “nucleotide change” refers to, e.g., nucleotide substitution, deletion, and/or insertion, as is well understood in the art. For example, mutations contain alterations that produce silent substitutions, additions, or deletions, but do not alter the properties or activities of the encoded protein or how the proteins are made.

[0084] As used herein, the term “protein modification” refers to, e.g., amino acid substitution, amino acid modification, deletion, and/or insertion, as is well understood in the art.

[0085] As used herein, the term “at least a portion” or “fragment” of a nucleic acid or polypeptide means a portion having the minimal size characteristics of such sequences, or any larger fragment of the full length molecule, up to and including the full length molecule. A fragment of a polynucleotide of the disclosure may encode a biologically active portion of a genetic regulatory element. A biologically active portion of a genetic regulatory element can be prepared by isolating a portion of one of the polynucleotides of the disclosure that comprises the genetic regulatory element and assessing activity as described herein. Similarly, a portion of a polypeptide may be 4 amino acids, 5 amino acids, 6 amino acids, 7 amino acids, and so on, going up to the full length polypeptide. The length of the portion to be used will depend on the particular application. A portion of a nucleic acid useful as a hybridization probe may be as short as 12 nucleotides; in some embodiments, it is 20 nucleotides. A portion of a polypeptide useful as an epitope may be as short as 4 amino acids. A portion of a polypeptide that performs the function of the full-length polypeptide would generally be longer than 4 amino acids.


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

As used herein, “promoter” refers to a DNA sequence capable of controlling the expression of a coding sequence or functional RNA. In some embodiments, the promoter sequence consists
of proximal and more distal upstream elements, the latter elements often referred to as enhancers. Accordingly, an “enhancer” is a DNA sequence that can stimulate promoter activity, and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, or even comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of some variation may have identical promoter activity.

[0090] As used herein, the phrases “recombinant construct”, “expression construct”, “chimeric construct”, “construct”, and “recombinant DNA construct” are used interchangeably herein. A recombinant construct comprises an artificial combination of nucleic acid fragments, e.g., regulatory and coding sequences that are not found together in nature. For example, a chimeric construct may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. Such construct may be used by itself or may be used in conjunction with a vector. If a vector is used then the choice of vector is dependent upon the method that will be used to transform host cells as is well known to those skilled in the art. For example, a plasmid vector can be used. The skilled artisan is well aware of the genetic elements that must be present on the vector in order to successfully transform, select and propagate host cells comprising any of the isolated nucleic acid fragments of the disclosure. The skilled artisan will also recognize that different independent transformation events will result in different levels and patterns of expression (Jones et al., (1985) EMBO J. 4:2411-2418; De Almeida et al., (1989) Mol. Gen. Genetics 218:78-86), and thus that multiple events must be screened in order to obtain lines displaying the desired expression level and pattern. Such screening may be accomplished by Southern analysis of DNA, Northern analysis of mRNA expression, immunoblotting analysis of protein expression, or phenotypic analysis, among others. Vectors can be plasmids, viruses, bacteriophages, pro-viruses, phagemids, transposons, artificial chromosomes, and the like, that replicate autonomously or can integrate into a chromosome of a host cell. A vector can also be a naked RNA polynucleotide, a naked DNA polynucleotide, a polynucleotide composed of both
DNA and RNA within the same strand, a poly-lysine-conjugated DNA or RNA, a peptide-conjugated DNA or RNA, a liposome-conjugated DNA, or the like, that is not autonomously replicating. As used herein, the term "expression" refers to the production of a functional end-product e.g., an mRNA or a protein (precursor or mature).

[0091] “Operably linked” means in this context the sequential arrangement of the promoter polynucleotide according to the disclosure with a further oligo- or polynucleotide, resulting in transcription of the further polynucleotide.

[0092] The term "product of interest" or "biomolecule" as used herein refers to any product produced by microbes from feedstock. In some cases, the product of interest may be a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, etc. For example, the product of interest or biomolecule may be any primary or secondary extracellular metabolite. The primary metabolite may be, inter alia, ethanol, citric acid, lactic acid, glutamic acid, glutamate, lysine, threonine, tryptophan and other amino acids, vitamins, polysaccharides, etc. The secondary metabolite may be, inter alia, an antibiotic compound like penicillin, or an immunosuppressant like cyclosporin A, a plant hormone like gibberellin, a statin drug like lovastatin, a fungicide like griseofulvin, etc. The product of interest or biomolecule may also be any intracellular component produced by a microbe, such as: a microbial enzyme, including: catalase, amylase, protease, pectinase, glucose isomerase, cellulase, hemicellulase, lipase, lactase, streptokinase, and many others. The intracellular component may also include recombinant proteins, such as: insulin, hepatitis B vaccine, interferon, granulocyte colony-stimulating factor, streptokinase and others.

[0093] The term “carbon source” generally refers to a substance suitable to be used as a source of carbon for cell growth. Carbon sources include, but are not limited to, biomass hydrolysates, starch, sucrose, cellulose, hemicellulose, xylose, and lignin, as well as monomeric components of these substrates. Carbon sources can comprise various organic compounds in various forms, including, but not limited to polymers, carbohydrates, acids, alcohols, aldehydes, ketones, amino acids, peptides, etc. These include, for example, various monosaccharides such as glucose, dextrose (D-glucose), maltose, oligosaccharides, polysaccharides, saturated or unsaturated fatty acids, succinate, lactate, acetate, ethanol, etc., or mixtures thereof. Photosynthetic organisms can
additionally produce a carbon source as a product of photosynthesis. In some embodiments, carbon sources may be selected from biomass hydrolysates and glucose.

[0094] The term “feedstock” is defined as a raw material or mixture of raw materials supplied to a microorganism or fermentation process from which other products can be made. For example, a carbon source, such as biomass or the carbon compounds derived from biomass are a feedstock for a microorganism that produces a product of interest (e.g. small molecule, peptide, synthetic compound, fuel, alcohol, etc.) in a fermentation process. However, a feedstock may contain nutrients other than a carbon source.

[0095] The term “volumetric productivity” or “production rate” is defined as the amount of product formed per volume of medium per unit of time. Volumetric productivity can be reported in gram per liter per hour (g/L/h).

[0096] The term “specific productivity” is defined as the rate of formation of the product. Specific productivity is herein further defined as the specific productivity in gram product per gram of cell dry weight (CDW) per hour (g/g CDW/h). Using the relation of CDW to OD₆₀₀ for the given microorganism specific productivity can also be expressed as gram product per liter culture medium per optical density of the culture broth at 600 nm (OD) per hour (g/L/h/OD).

[0097] The term “yield” is defined as the amount of product obtained per unit weight of raw material and may be expressed as g product per g substrate (g/g). Yield may be expressed as a percentage of the theoretical yield. “Theoretical yield” is defined as the maximum amount of product that can be generated per a given amount of substrate as dictated by the stoichiometry of the metabolic pathway used to make the product.

[0098] The term “titre” or “titer” is defined as the strength of a solution or the concentration of a substance in solution. For example, the titre of a product of interest (e.g. small molecule, peptide, synthetic compound, fuel, alcohol, etc.) in a fermentation broth is described as g of product of interest in solution per liter of fermentation broth (g/L).

[0099] The term “total titer” is defined as the sum of all product of interest produced in a process, including but not limited to the product of interest in solution, the product of interest in gas phase if applicable, and any product of interest removed from the process and recovered relative to the initial volume in the process or the operating volume in the process.
[0100] As used herein, the term “HTP genetic design library” or “library” refers to collections of genetic perturbations according to the present disclosure. In some embodiments, the libraries of the present disclosure may manifest as i) a collection of sequence information in a database or other computer file, ii) a collection of genetic constructs encoding for the aforementioned series of genetic elements, or iii) host cell strains comprising the genetic elements. In some embodiments, the libraries of the present disclosure may refer to collections of individual elements (e.g., collections of promoters for PRO swap libraries, collections of terminators for STOP swap libraries, or transposon mutagenesis libraries). In other embodiments, the libraries of the present disclosure may also refer to combinations of genetic elements, such as combinations of promoter::genes, gene::terminator, gene deletions or pertubations, or even promoter::gene:terminators. In some embodiments, the libraries of the present disclosure further comprise meta data associated with the effects of applying each member of the library in host organisms. For example, a library as used herein can include a collection of promoter::gene sequence combinations, together with the resulting effect of those combinations on one or more phenotypes in a particular species, thus improving the future predictive value of using the combination in future promoter swaps.

[0101] As used herein, the term “SNP” refers to Small Nuclear Polymorphism(s). In some embodiments, SNPs of the present disclosure should be construed broadly, and include single nucleotide polymorphisms, sequence insertions, deletions, inversions, and other sequence replacements. As used herein, the term “non-synonymous” or non-synonymous SNPs” refers to mutations that lead to coding changes in host cell proteins.

[0102] A “high-throughput (HTP)” method of genomic engineering may involve the utilization of at least one piece of automated equipment (e.g. a liquid handler or plate handler machine) to carry out at least one step of the method.

[0103] The term “transposon” refers to a polynucleotide that is able to excise from a donor polynucleotide, for instance, a vector, and integrate into a target site, for instance, a cell’s genomic DNA. A transposon may include a polynucleotide that includes a nucleic acid sequence flanked by cis-acting nucleotide sequences located at the termini of the transposon. A nucleic acid sequence is “flanked by” cis-acting nucleotide sequences if at least one cis-acting nucleotide sequence is positioned 5′ to the nucleic acid sequence, and at least one cis-acting nucleotide is
positioned 3’ to the nucleic acid sequence. A nucleic acid sequence flanked by cis-acting nucleotide sequences may be referred to herein as a “flanked sequence.” cis-acting nucleotide sequences include at least one inverted repeat at each end of the transposon, to which a transposase binds. The “flanked sequence” or “transposon payload” may include one or more nucleic acid sequences that act as insertional mutagens. An insertional mutagen is a nucleic acid sequence whose insertion will affect the level of expression or the nature of the product expressed by a coding region near or in which the flanked sequence is inserted by transposition. When the nature of the product expressed is altered, the nucleic acid is referred to as a “disruptive sequence.” When the level of expression is altered, the nucleic acid is referred to as an “affective sequence”. Transposons of the present disclosure may include one or more insertional mutagens, which may be disruptive and/or affective sequences.

[0104] The term “Pro Swap” as used herein refers to methods of selecting promoters with optimal expression properties to produce beneficial effects on an overall-host strain phenotype. In some embodiments, these methods include methods of identifying one or more promoters and/or generating variants of one or more promoters within a host cell, which exhibit a range of expression strengths, or superior regulatory properties. A particular combination of these identified and/or generated promoters can be grouped together as a promoter ladder.

[0105] The term “SNP Swap” as used herein refers to the systematic introduction or removal of individual Small Nuclear Polymorphism nucleotide mutations (i.e. SNPs) across strains. In some embodiments, the resultant microbes that are engineered via this process form HTP genetic design libraries. In some embodiments, SNP swapping involves the reconstruction of host organisms with optimal combination of target SNP “building blocks” with identified beneficial performance effects. Thus, in some embodiments, SNP swapping involves consolidating multiple beneficial mutations into a single strain background, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations. In other embodiments, SNP swapping also involves removing multiple mutations identified as detrimental from a strain, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations. In some
embodiments, the SNP swapping methods of the present disclosure include both the addition of beneficial SNPs, and removing detrimental and/or neutral mutations.

[0106] The term “STOP Swap” as used herein refers to method of improving host cell productivity (e.g. through the modulation of transcription via the modulation of gene terminator sequences) through the optimization of cellular gene transcription. In some embodiments, the present disclosure teaches methods of selecting termination sequences (“terminators”) with optimal expression properties to produce beneficial effects on overall-host strain productivity. In some embodiments, this method includes identifying one or more terminators and/or generating variants of one or more terminators within a host cell which exhibit a range of expression strengths (e.g. terminator ladders). A particular combination of these identified and/or generated terminators can be grouped together as a terminator ladder.

Traditional Methods of Strain Improvement

[0107] Traditional approaches to strain improvement can be broadly categorized into two types of approaches: directed strain engineering, and random mutagenesis.

[0108] Directed engineering methods of strain improvement involve the planned perturbation of a handful of genetic elements of a specific organism. These approaches are typically focused on modulating specific biosynthetic or developmental programs, and rely on prior knowledge of the genetic and metabolic factors affecting the pathways. In its simplest embodiments, directed engineering involves the transfer of a characterized trait (e.g., gene, promoter, or other genetic element capable of producing a measurable phenotype) from one organism to another organism of the same, or different species.

[0109] Random approaches to strain engineering involve the random mutagenesis of parent strains, coupled with extensive screening designed to identify performance improvements. Approaches to generating these random mutations include exposure to ultraviolet radiation, or mutagenic chemicals such as Ethyl methanesulfonate. Though random and largely unpredictable, this traditional approach to strain improvement had several advantages compared to more directed genetic manipulations. *First*, many industrial organisms were (and remain) poorly characterized in terms of their genetic and metabolic repertoires, rendering alternative directed improvement approaches difficult, if not impossible.
[0110] Second, even in relatively well characterized systems, genotypic changes that result in industrial performance improvements are difficult to predict, and sometimes only manifest themselves as epistatic phenotypes requiring cumulative mutations in many genes of known and unknown function.

[0111] Additionally, for many years, the genetic tools required for making directed genomic mutations in a given industrial organism were unavailable, or very slow and/or difficult to use.

[0112] The extended application of the traditional strain improvement programs, however, yield progressively reduced gains in a given strain lineage, and ultimately lead to exhausted possibilities for further strain efficiencies. Beneficial random mutations are relatively rare events, and require large screening pools and high mutation rates. This inevitably results in the inadvertent accumulation of many neutral and/or detrimental (or partly detrimental) mutations in “improved” strains, which ultimately create a drag on future efficiency gains.

[0113] Another limitation of traditional cumulative improvement approaches is that little to no information is known about any particular mutation’s effect on any strain metric. This fundamentally limits a researcher’s ability to combine and consolidate beneficial mutations, or to remove neutral or detrimental mutagenic “baggage.”

[0115] These include techniques such as protoplast fusion and whole genome shuffling that facilitate genomic recombination across mutated strains. For some industrial microorganisms such as yeast and filamentous fungi, natural mating cycles can also be exploited for pairwise genomic recombination. In this way, detrimental mutations can be removed by ‘back-crossing’ mutants with parental strains and beneficial mutations consolidated. Moreover, beneficial mutations from two different strain lineages can potentially be combined, which creates additional improvement possibilities over what might be available from mutating a single strain lineage on its own. However, these approaches are subject to many limitations that are circumvented using the methods of the present disclosure.

[0116] For example, traditional recombinant approaches as described above are slow and rely on a relatively small number of random recombination crossover events to swap mutations, and are therefore limited in the number of combinations that can be attempted in any given cycle, or time period. In addition, although the natural recombination events in the prior art are essentially random, they are also subject to genome positional bias.

[0117] Most importantly, the traditional approaches also provide little information about the influence of individual mutations and due to the random distribution of recombined mutations many specific combinations cannot be generated and evaluated.

[0118] To overcome many of the aforementioned problems associated with traditional strain improvement programs, the present disclosure sets forth a unique HTP genomic engineering platform that is computationally driven and integrates molecular biology, automation, data analytics, and machine learning protocols. This integrative platform utilizes a suite of HTP molecular tool sets that are used to construct HTP genetic design libraries. These genetic design libraries will be elaborated upon below.

[0119] The taught HTP platform and its unique microbial genetic design libraries fundamentally shift the paradigm of microbial strain development and evolution. For example, traditional mutagenesis-based methods of developing an industrial microbial strain will eventually lead to microbes burdened with a heavy mutagenic load that has been accumulated over years of random mutagenesis.

[0120] The ability to solve this issue (i.e. remove the genetic baggage accumulated by these microbes) has eluded microbial researchers for decades. However, utilizing the HTP platform
disclosed herein, these industrial strains can be “rehabilitated,” and the genetic mutations that are deleterious can be identified and removed. Congruently, the genetic mutations that are identified as beneficial can be kept, and in some cases improved upon. The resulting microbial strains demonstrate superior phenotypic traits (e.g., improved production of a compound of interest), as compared to their parental strains.

[0121] Furthermore, the HTP platform taught herein is able to identify, characterize, and quantify the effect that individual mutations have on microbial strain performance. This information, i.e. what effect does a given genetic change x have on host cell phenotype y (e.g., production of a compound or product of interest), is able to be generated and then stored in the microbial HTP genetic design libraries discussed below. That is, sequence information for each genetic permutation, and its effect on the host cell phenotype are stored in one or more databases, and are available for subsequent analysis (e.g., epistasis mapping, as discussed below). The present disclosure also teaches methods of physically saving/storing valuable genetic permutations in the form of genetic insertion constructs, or in the form of one or more host cell organisms containing the genetic permutation (e.g., see libraries discussed below.)

[0122] When one couples these HTP genetic design libraries into an iterative process that is integrated with a sophisticated data analytics and machine learning process a dramatically different methodology for improving host cells emerges. The taught platform is therefore fundamentally different from the previously discussed traditional methods of developing host cell strains. The taught HTP platform does not suffer from many of the drawbacks associated with the previous methods. These and other advantages will become apparent with reference to the HTP molecular tool sets and the derived genetic design libraries discussed below.

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

[0123] As aforementioned, the present disclosure provides a novel HTP platform and genetic design strategy for engineering microbial organisms through iterative systematic introduction and removal of genetic changes across strains. The platform is supported by a suite of molecular tools, which enable the creation of HTP genetic design libraries and allow for the efficient implementation of genetic alterations into a given host strain.
[0124] The HTP genetic design libraries of the disclosure serve as sources of possible genetic alterations that may be introduced into a particular microbial strain background. In this way, the HTP genetic design libraries are repositories of genetic diversity, or collections of genetic perturbations, which can be applied to the initial or further engineering of a given microbial strain. Techniques for programming genetic designs for implementation to host strains are described in pending US Patent Application, Serial No. 15/140,296, entitled “Microbial Strain Design System and Methods for Improved Large Scale Production of Engineered Nucleotide Sequences,” incorporated by reference in its entirety herein.

[0125] The HTP molecular tool sets utilized in this platform may include, inter alia: (1) Promoter swaps (PRO Swap), (2) SNP swaps, (3) Start/Stop codon exchanges, (4) STOP swaps, (5) Sequence optimization, and (6) Transposon Mutagenesis or a combination thereof. The HTP methods of the present disclosure also teach methods for directing the consolidation/combinatorial use of HTP tool sets, including (7) Epistasis mapping protocols. As aforementioned, this suite of molecular tools, either in isolation or combination, enables the creation of HTP genetic design host cell libraries.

[0126] As will be demonstrated, utilization of the aforementioned HTP genetic design libraries in the context of the taught HTP microbial engineering platform enables the identification and consolidation of beneficial “causative” mutations or gene sections and also the identification and removal of passive or detrimental mutations or gene sections. This new approach allows rapid improvements in strain performance that could not be achieved by traditional random mutagenesis or directed genetic engineering. The removal of genetic burden or consolidation of beneficial changes into a strain with no genetic burden also provides a new, robust starting point for additional random mutagenesis that may enable further improvements.

[0127] In some embodiments, the present disclosure teaches that as orthogonal beneficial changes are identified across various, discrete branches of a mutagenic strain lineage, they can also be rapidly consolidated into better performing strains. These mutations can also be consolidated into strains that are not part of mutagenic lineages, such as strains with improvements gained by directed genetic engineering.

[0128] In some embodiments, the present disclosure differs from known strain improvement approaches in that it analyzes the genome-wide combinatorial effect of mutations across multiple
disparate genomic regions, including expressed and non-expressed genetic elements, and uses gathered information (e.g., experimental results) to predict mutation combinations expected to produce strain enhancements.

[0129] In some embodiments, the present disclosure teaches: i) industrial microorganisms, and other host cells amenable to improvement via the disclosed inventions, ii) generating diversity pools for downstream analysis, iii) methods and hardware for high-throughput screening and sequencing of large variant pools, iv) methods and hardware for machine learning computational analysis and prediction of synergistic effects of genome-wide mutations, and v) methods for high-throughput strain engineering.

[0130] The following molecular tools and libraries are discussed in terms of illustrative microbial examples. Persons having skill in the art will recognize that the HTP molecular tools of the present disclosure are compatible with any host cell, including eukaryotic cellular, and higher life forms.

[0131] Each of the identified HTP molecular tool sets—which enable the creation of the various HTP genetic design libraries utilized in the microbial engineering platform—will now be discussed.

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

[0132] In some embodiments, the present disclosure teaches methods of selecting promoters with optimal expression properties to produce beneficial effects on overall-host strain phenotype (e.g., yield or productivity).

[0133] For example, in some embodiments, the present disclosure teaches methods of identifying one or more promoters and/or generating variants of one or more promoters within a host cell, which exhibit a range of expression strengths (e.g. promoter ladders discussed *infra*), or superior regulatory properties (e.g., tighter regulatory control for selected genes). A particular combination of these identified and/or generated promoters can be grouped together as a promoter ladder, which is explained in more detail below.

[0134] The promoter ladder in question is then associated with a given gene of interest. Thus, if one has promoters P₁-P₉ (representing eight promoters that have been identified and/or generated to exhibit a range of expression strengths) and associates the promoter ladder with a single gene
of interest in a microbe (i.e. genetically engineer a microbe with a given promoter operably linked to a given target gene), then the effect of each combination of the eight promoters can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered microbes have an otherwise identical genetic background except the particular promoter(s) associated with the target gene.

[0135] The resultant microbes that are engineered via this process form HTP genetic design libraries.

[0136] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed via this process, with each member strain being representative of a given promoter operably linked to a particular target gene, in an otherwise identical genetic background, the library being termed a “promoter swap microbial strain library.”

[0137] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given promoter \( x \) operably linked to a given gene \( y \)—the collection being termed a “promoter swap library.”

[0138] Further, one can utilize the same promoter ladder comprising promoters \( P_1-P_8 \) to engineer microbes, wherein each of the 8 promoters is operably linked to 10 different gene targets. The result of this procedure would be 80 microbes that are otherwise assumed genetically identical, except for the particular promoters operably linked to a target gene of interest. These 80 microbes could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any data storage construct, including a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could be, for example, a given promoter’s (e.g. \( P_1-P_8 \)) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more of promoters \( P_1-P_8 \) to a given gene target.

[0139] The aforementioned examples of eight promoters and 10 target genes is merely illustrative, as the concept can be applied with any given number of promoters that have been grouped together based upon exhibition of a range of expression strengths and any given number of target genes. Persons having skill in the art will also recognize the ability to operably link two or more promoters in front of any gene target. Thus, in some embodiments, the present disclosure teaches promoter
swap libraries in which 1, 2, 3 or more promoters from a promoter ladder are operably linked to one or more genes.

[0140] In summary, utilizing various promoters to drive expression of various genes in an organism is a powerful tool to optimize a trait of interest. The molecular tool of promoter swapping, developed by the inventors, uses a ladder of promoter sequences that have been demonstrated to vary expression of at least one locus under at least one condition. This ladder is then systematically applied to a group of genes in the organism using high-throughput genome engineering. This group of genes is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity. In some embodiments, the selection of genes can include all the genes in a given host. In other embodiments, the selection of genes can be a subset of all genes in a given host, chosen randomly.

[0141] The resultant HTP genetic design microbial strain library of organisms containing a promoter sequence linked to a gene is then assessed for performance in a high-throughput screening model, and promoter-gene linkages which lead to increased performance are determined and the information stored in a database. The collection of genetic perturbations (i.e. given promoter x operably linked to a given gene y) form a “promoter swap library,” which can be utilized as a source of potential genetic alterations to be utilized in microbial engineering processing. Over time, as a greater set of genetic perturbations is implemented against a greater diversity of host cell backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest.

[0142] Transcription levels of genes in an organism are a key point of control for affecting organism behavior. Transcription is tightly coupled to translation (protein expression), and which proteins are expressed in what quantities determines organism behavior. Cells express thousands of different types of proteins, and these proteins interact in numerous complex ways to create function. By varying the expression levels of a set of proteins systematically, function can be altered in ways that, because of complexity, are difficult to predict. Some alterations may increase
performance, and so, coupled to a mechanism for assessing performance, this technique allows for the generation of organisms with improved function.

[0143] In the context of a small molecule synthesis pathway, enzymes interact through their small molecule substrates and products in a linear or branched chain, starting with a substrate and ending with a small molecule of interest. Because these interactions are sequentially linked, this system exhibits distributed control, and increasing the expression of one enzyme can only increase pathway flux until another enzyme becomes rate limiting.

[0144] Metabolic Control Analysis (MCA) is a method for determining, from experimental data and first principles, which enzyme or enzymes are rate limiting. MCA is limited however, because it requires extensive experimentation after each expression level change to determine the new rate limiting enzyme. Promoter swapping is advantageous in this context, because through the application of a promoter ladder to each enzyme in a pathway, the limiting enzyme is found, and the same thing can be done in subsequent rounds to find new enzymes that become rate limiting. Further, because the read-out on function is better production of the small molecule of interest, the experiment to determine which enzyme is limiting is the same as the engineering to increase production, thus shortening development time. In some embodiments the present disclosure teaches the application of PRO swap to genes encoding individual subunits of multi-unit enzymes. In yet other embodiments, the present disclosure teaches methods of applying PRO swap techniques to genes responsible for regulating individual enzymes, or whole biosynthetic pathways.

[0145] In some embodiments, the promoter swap tool of the present disclosure can is used to identify optimum expression of a selected gene target. In some embodiments, the goal of the promoter swap may be to increase expression of a target gene to reduce bottlenecks in a metabolic or genetic pathway. In other embodiments, the goal of the promoter swap may be to reduce the expression of the target gene to avoid unnecessary energy expenditures in the host cell, when expression of the target gene is not required.

[0146] In the context of other cellular systems like transcription, transport, or signaling, various rational methods can be used to try and find out, a priori, which proteins are targets for expression change and what that change should be. These rational methods reduce the number of perturbations that must be tested to find one that improves performance, but they do so at significant cost. Gene
deletion studies identify proteins whose presence is critical for a particular function, and important genes can then be over-expressed. Due to the complexity of protein interactions, this is often ineffective at increasing performance. Different types of models have been developed that attempt to describe, from first principles, transcription or signaling behavior as a function of protein levels in the cell. These models often suggest targets where expression changes might lead to different or improved function. The assumptions that underlie these models are simplistic and the parameters difficult to measure, so the predictions they make are often incorrect, especially for non-model organisms. With both gene deletion and modeling, the experiments required to determine how to affect a certain gene are different than the subsequent work to make the change that improves performance. Promoter swapping sidesteps these challenges, because the constructed strain that highlights the importance of a particular perturbation is also, already, the improved strain.

[0147] Thus, in particular embodiments, promoter swapping is a multi-step process comprising:

[0148] 1. Selecting a set of “x” promoters to act as a “ladder.” Ideally these promoters have been shown to lead to highly variable expression across multiple genomic loci, but the only requirement is that they perturb gene expression in some way.

[0149] 2. Selecting a set of “n” genes to target. This set can be every open reading frame (ORF) in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps or previous SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection. In other embodiments, the “n” targeted genes can comprise non-protein coding genes, including non-coding RNAs.

[0150] 3. High-throughput strain engineering to rapidly-and in some embodiments, in parallel-carry out the following genetic modifications: When a native promoter exists in front of target gene n and its sequence is known, replace the native promoter with each of the x promoters in the ladder. When the native promoter does not exist, or its sequence is unknown, insert each of the x promoters in the ladder in front of gene n (see e.g., Figure 13). In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the library is an instance of x promoter operably linked to n target, in an otherwise identical genetic
context. As previously described combinations of promoters can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0151] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0152] This foundational process can be extended to provide further improvements in strain performance by, inter alia: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then sequential regeneration of the library of perturbations into an improved member or members of the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches (see Figure 12).

[0153] The molecular tool, or technique, discussed above is characterized as promoter swapping, but is not limited to promoters and can include other sequence changes that systematically vary the expression level of a set of targets. Other methods for varying the expression level of a set of genes could include: a) a ladder of ribosome binding sites (or Kozak sequences in eukaryotes); b) replacing the start codon of each target with each of the other start codons (i.e start/stop codon exchanges discussed infra); c) attachment of various mRNA stabilizing or destabilizing sequences to the 5’ or 3’ end, or at any other location, of a transcript, d) attachment of various protein stabilizing or destabilizing sequences at any location in the protein.

[0154] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

2. **SNP Swaps: A Molecular Tool for the Derivation of SNP Swap Microbial Strain Libraries**
[0155] In certain embodiments, SNP swapping is not a random mutagenic approach to improving a microbial strain, but rather involves the systematic introduction or removal of individual Small Nuclear Polymorphism nucleotide mutations (i.e. SNPs) (hence the name “SNP swapping”) across strains.

[0156] The resultant microbes that are engineered via this process form HTP genetic design libraries.

[0157] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed via this process, with each member strain being representative of the presence or absence of a given SNP, in an otherwise identical genetic background, the library being termed a “SNP swap microbial strain library.”

[0158] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given SNP being present or a given SNP being absent—the collection being termed a “SNP swap library.”

[0159] In some embodiments, SNP swapping involves the reconstruction of host organisms with optimal combinations of target SNP “building blocks” with identified beneficial performance effects. Thus, in some embodiments, SNP swapping involves consolidating multiple beneficial mutations into a single strain background, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations.

[0160] In other embodiments, SNP swapping also involves removing multiple mutations identified as detrimental from a strain, either one at a time in an iterative process, or as multiple changes in a single step. Multiple changes can be either a specific set of defined changes or a partly randomized, combinatorial library of mutations. In some embodiments, the SNP swapping methods of the present disclosure include both the addition of beneficial SNPs, and removing detrimental and/or neutral mutations.

[0161] SNP swapping is a powerful tool to identify and exploit both beneficial and detrimental mutations in a lineage of strains subjected to mutagenesis and selection for an improved trait of interest. SNP swapping utilizes high-throughput genome engineering techniques to systematically determine the influence of individual mutations in a mutagenic lineage. Genome sequences are
determined for strains across one or more generations of a mutagenic lineage with known performance improvements. High-throughput genome engineering is then used systematically to recapitulate mutations from improved strains in earlier lineage strains, and/or revert mutations in later strains to earlier strain sequences. The performance of these strains is then evaluated and the contribution of each individual mutation on the improved phenotype of interest can be determined. As aforementioned, the microbial strains that result from this process are analyzed/characterized and form the basis for the SNP swap genetic design libraries that can inform microbial strain improvement across host strains.

[0162] Removal of detrimental mutations can provide immediate performance improvements, and consolidation of beneficial mutations in a strain background not subject to mutagenic burden can rapidly and greatly improve strain performance. The various microbial strains produced via the SNP swapping process form the HTP genetic design SNP swapping libraries, which are microbial strains comprising the various added/deleted/or consolidated SNPs, but with otherwise identical genetic backgrounds.

[0163] As discussed previously, random mutagenesis and subsequent screening for performance improvements is a commonly used technique for industrial strain improvement, and many strains currently used for large scale manufacturing have been developed using this process iteratively over a period of many years, sometimes decades. Random approaches to generating genomic mutations such as exposure to UV radiation or chemical mutagens such as ethyl methanesulfonate were a preferred method for industrial strain improvements because: 1) industrial organisms may be poorly characterized genetically or metabolically, rendering target selection for directed improvement approaches difficult or impossible; 2) even in relatively well characterized systems, changes that result in industrial performance improvements are difficult to predict and may require perturbation of genes that have no known function, and 3) genetic tools for making directed genomic mutations in a given industrial organism may not be available or very slow and/or difficult to use.

[0164] However, despite the aforementioned benefits of this process, there are also a number of known disadvantages. Beneficial mutations are relatively rare events, and in order to find these mutations with a fixed screening capacity, mutations rates must be sufficiently high. This often results in unwanted neutral and partly detrimental mutations being incorporated into strains along
with beneficial changes. Over time this ‘mutagenic burden’ builds up, resulting in strains with deficiencies in overall robustness and key traits such as growth rates. Eventually ‘mutagenic burden’ renders further improvements in performance through random mutagenesis increasingly difficult or impossible to obtain. Without suitable tools, it is impossible to consolidate beneficial mutations found in discrete and parallel branches of strain lineages.

[0165] SNP swapping is an approach to overcome these limitations by systematically recapitulating or reverting some or all mutations observed when comparing strains within a mutagenic lineage. In this way, both beneficial (‘causative’) mutations can be identified and consolidated, and/or detrimental mutations can be identified and removed. This allows rapid improvements in strain performance that could not be achieved by further random mutagenesis or targeted genetic engineering.

[0166] Removal of genetic burden or consolidation of beneficial changes into a strain with no genetic burden also provides a new, robust starting point for additional random mutagenesis that may enable further improvements.

[0167] In addition, as orthogonal beneficial changes are identified across various, discrete branches of a mutagenic strain lineage, they can be rapidly consolidated into better performing strains. These mutations can also be consolidated into strains that are not part of mutagenic lineages, such as strains with improvements gained by directed genetic engineering.

[0168] Other approaches and technologies exist to randomly recombine mutations between strains within a mutagenic lineage. These include techniques such as protoplast fusion and whole genome shuffling that facilitate genomic recombination across mutated strains. For some industrial microorganisms such as yeast and filamentous fungi, natural mating cycles can also be exploited for pairwise genomic recombination. In this way, detrimental mutations can be removed by ‘back-crossing’ mutants with parental strains and beneficial mutations consolidated. When directed mutational changes are desired, SNP swapping methods of the present disclosure may be used.

[0169] For example, as these approaches rely on a relatively small number of random recombination crossover events to swap mutations, it may take many cycles of recombination and screening to optimize strain performance. In addition, although natural recombination events are essentially random, they are also subject to genome positional bias and some mutations may be difficult to address. These approaches also provide little information about the influence of
individual mutations without additional genome sequencing and analysis. SNP swapping overcomes these fundamental limitations as it is not a random approach, but rather the systematic introduction or removal of individual mutations across strains.

[0170] In some embodiments, the present disclosure teaches methods for identifying the SNP sequence diversity present among the organisms of a diversity pool. A diversity pool can be a given number \( n \) of microbes utilized for analysis, with the microbes’ genomes representing the “diversity pool.”

[0171] In particular aspects, a diversity pool may be an original parent strain (\( S_1 \)) with a “baseline” or “reference” genetic sequence at a particular time point (\( S_{1\text{Gen}_1} \)) and then any number of subsequent offspring strains (\( S_{2:a} \)) that were derived/developed from the \( S_1 \) strain and that have a different genome (\( S_{2:a\text{Gen}_{2:a}} \)), in relation to the baseline genome of \( S_1 \).

[0172] For example, in some embodiments, the present disclosure teaches sequencing the microbial genomes in a diversity pool to identify the SNPs present in each strain. In one embodiment, the strains of the diversity pool are historical microbial production strains. Thus, a diversity pool of the present disclosure can include for example, an industrial reference strain, and one or more mutated industrial strains produced \textit{via} traditional strain improvement programs.

[0173] In some embodiments, the SNPs within a diversity pool are determined with reference to a “reference strain.” In some embodiments, the reference strain is a wild-type strain. In other embodiments, the reference strain is an original industrial strain prior to being subjected to any mutagenesis. The reference strain can be defined by the practitioner and does not have to be an original wild-type strain or original industrial strain. The base strain is merely representative of what will be considered the “base,” “reference” or original genetic background, by which subsequent strains that were derived, or were developed from the reference strain, are to be compared.

[0174] Once all SNPs in the diversity pool are identified, the present disclosure teaches methods of SNP swapping and screening methods to delineate (\textit{i.e.} quantify and characterize) the effects (\textit{e.g.} creation of a phenotype of interest) of SNPs individually and/or in groups.
In some embodiments, the SNP swapping methods of the present disclosure comprise the step of introducing one or more SNPs identified in a mutated strain (e.g., a strain from amongst S$_2$ Gen$_{2-n}$) to a reference strain (S$_1$ Gen$_1$) or wild-type strain (“wave up”).

In other embodiments, the SNP swapping methods of the present disclosure comprise the step of removing one or more SNPs identified in a mutated strain (e.g., a strain from amongst S$_2$ Gen$_{2-n}$) (“wave down”).

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


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

The present disclosure similarly teaches swapping start codons. In some embodiments, the present disclosure teaches use of the ATG (AUG) start codon utilized by most organisms (especially eukaryotes). In some embodiments, the present disclosure teaches that prokaryotes use ATG (AUG) the most, followed by GTG (GUG) and TTG (UUG).
[0180] In other embodiments, the present disclosure teaches replacing ATG start codons with TTG. In some embodiments, the present disclosure teaches replacing ATG start codons with GTG. In some embodiments, the present disclosure teaches replacing GTG start codons with ATG. In some embodiments, the present disclosure teaches replacing GTG start codons with TTG. In some embodiments, the present disclosure teaches replacing TTG start codons with ATG. In some embodiments, the present disclosure teaches replacing TTG start codons with GTG.

[0181] In other embodiments, the present disclosure teaches replacing TAA stop codons with TAG. In some embodiments, the present disclosure teaches replacing TAA stop codons with TGA. In some embodiments, the present disclosure teaches replacing TGA stop codons with TAA. In some embodiments, the present disclosure teaches replacing TGA stop codons with TAG. In some embodiments, the present disclosure teaches replacing TAG stop codons with TAA. In some embodiments, the present disclosure teaches replacing TAG stop codons with TGA.

4. Stop swap: A Molecular Tool for the Derivation of Optimized Sequence Microbial Strain Libraries

[0182] In some embodiments, the present disclosure teaches methods of improving host cell productivity through the optimization of cellular gene transcription. Gene transcription is the result of several distinct biological phenomena, including transcriptional initiation (RNAP recruitment and transcriptional complex formation), elongation (strand synthesis/extension), and transcriptional termination (RNAP detachment and termination). Although much attention has been devoted to the control of gene expression through the transcriptional modulation of genes (e.g., by changing promoters, or inducing regulatory transcription factors), comparatively few efforts have been made towards the modulation of transcription via the modulation of gene terminator sequences.

[0183] The most obvious way that transcription impacts on gene expression levels is through the rate of Pol II initiation, which can be modulated by combinations of promoter or enhancer strength and trans-activating factors (Kadonaga, JT. 2004 “Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors” Cell. 2004 Jan 23; 116(2):247-57). In eukaryotes, elongation rate may also determine gene expression patterns by influencing alternative splicing (Cramer P. et al., 1997 “Functional association between promoter structure and transcript alternative splicing.” Proc Natl Acad Sci U S A. 1997 Oct 14; 94(21):11456-60). Failed
termination on a gene can impair the expression of downstream genes by reducing the accessibility of the promoter to Pol II (Greger IH. et al., 2000 “Balancing transcriptional interference and initiation on the GAL7 promoter of Saccharomyces cerevisiae.” Proc Natl Acad Sci U S A. 2000 Jul 18; 97(15):8415-20). This process, known as transcriptional interference, is particularly relevant in lower eukaryotes, as they often have closely spaced genes.

**[0184]** Termination sequences can also affect the expression of the genes to which the sequences belong. For example, studies show that inefficient transcriptional termination in eukaryotes results in an accumulation of unspliced pre-mRNA (see West, S., and Proudfoot, N.J., 2009 “Transcriptional Termination Enhances Protein Expression in Human Cells” Mol Cell. 2009 Feb 13; 33(3-9); 354-364). Other studies have also shown that 3’ end processing, can be delayed by inefficient termination (West, S et al., 2008 “Molecular dissection of mammalian RNA polymerase II transcriptional termination.” Mol Cell. 2008 Mar 14; 29(5):600-10.). Transcriptional termination can also affect mRNA stability by releasing transcripts from sites of synthesis.

*Termination of transcription mechanism in eukaryotes*

**[0185]** Transcriptional termination in eukaryotes operates through terminator signals that are recognized by protein factors associated with the RNA polymerase II. In some embodiments, the cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) transfer from the carboxyl terminal domain of RNA polymerase II to the poly-A signal. In some embodiments, the CPSF and CstF factors also recruit other proteins to the termination site, which then cleave the transcript and free the mRNA from the transcription complex. Termination also triggers polyadenylation of mRNA transcripts. Illustrative examples of validated eukaryotic termination factors, and their conserved structures are discussed in later portions of this document.

*Termination of transcription in prokaryotes*

**[0186]** In prokaryotes, two principal mechanisms, termed Rho-independent and Rho-dependent termination, mediate transcriptional termination. Rho-independent termination signals do not require an extrinsic transcription-termination factor, as formation of a stem-loop structure in the RNA transcribed from these sequences along with a series of Uridine (U) residues promotes release of the RNA chain from the transcription complex. Rho-dependent termination, on the other hand, requires a transcription-termination factor called Rho and cis-acting elements on the mRNA.
The initial binding site for Rho, the Rho utilization (rut) site, is an extended (~70 nucleotides, sometimes 80-100 nucleotides) single-stranded region characterized by a high cytidine/low guanosine content and relatively little secondary structure in the RNA being synthesized, upstream of the actual terminator sequence. When a polymerase pause site is encountered, termination occurs, and the transcript is released by Rho's helicase activity.

**Terminator Swapping (STOP swap)**

[0187] In some embodiments, the present disclosure teaches methods of selecting termination sequences (“terminators”) with optimal expression properties to produce beneficial effects on overall-host strain productivity.

[0188] For example, in some embodiments, the present disclosure teaches methods of identifying one or more terminators and/or generating variants of one or more terminators within a host cell, which exhibit a range of expression strengths (e.g. terminator ladders discussed infra). A particular combination of these identified and/or generated terminators can be grouped together as a terminator ladder, which is explained in more detail below.

[0189] The terminator ladder in question is then associated with a given gene of interest. Thus, if one has terminators T₁-Tₙ (representing eight terminators that have been identified and/or generated to exhibit a range of expression strengths when combined with one or more promoters) and associates the terminator ladder with a single gene of interest in a host cell (i.e. genetically engineer a host cell with a given terminator operably linked to the 3’ end of to a given target gene), then the effect of each combination of the terminators can be ascertained by characterizing each of the engineered strains resulting from each combinatorial effort, given that the engineered host cells have an otherwise identical genetic background except the particular promoter(s) associated with the target gene. The resultant host cells that are engineered via this process form HTP genetic design libraries.

[0190] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed via this process, with each member strain being representative of a given terminator operably linked to a particular target gene, in an otherwise identical genetic background, the library being termed a “terminator swap microbial strain library” or “STOP swap microbial strain library.”
Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given terminator $x$ operably linked to a given gene $y$—the collection being termed a “terminator swap library” or “STOP swap library.”

Further, one can utilize the same terminator ladder comprising promoters $T_1$-$T_8$ to engineer microbes, wherein each of the eight terminators is operably linked to 10 different gene targets. The result of this procedure would be 80 host cell strains that are otherwise assumed genetically identical, except for the particular terminators operably linked to a target gene of interest. These 80 host cell strains could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any database, including without limitation, a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could include, for example, a given terminators’ (e.g., $T_1$-$T_8$) effect when operably linked to a given gene target. This data/information can also be the broader set of combinatorial effects that result from operably linking two or more of promoters $T_1$-$T_8$ to a given gene target.

The aforementioned examples of eight terminators and 10 target genes is merely illustrative, as the concept can be applied with any given number of promoters that have been grouped together based upon exhibition of a range of expression strengths and any given number of target genes.

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

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

[0196] Thus, in particular embodiments, terminator swapping is a multi-step process comprising:

[0197] 1. Selecting a set of “x” terminators to act as a “ladder.” Ideally these terminators have been shown to lead to highly variable expression across multiple genomic loci, but the only requirement is that they perturb gene expression in some way.

[0198] 2. Selecting a set of “n” genes to target. This set can be every ORF in a genome, or a subset of ORFs. The subset can be chosen using annotations on ORFs related to function, by relation to previously demonstrated beneficial perturbations (previous promoter swaps, STOP swaps, or SNP swaps), by algorithmic selection based on epistatic interactions between previously generated perturbations, other selection criteria based on hypotheses regarding beneficial ORF to target, or through random selection. In other embodiments, the “n” targeted genes can comprise non-protein coding genes, including non-coding RNAs.

[0199] 3. High-throughput strain engineering to rapidly and in parallel carry out the following genetic modifications: When a native terminator exists at the 3’ end of target gene n and its sequence is known, replace the native terminator with each of the x terminators in the ladder. When the native terminator does not exist, or its sequence is unknown, insert each of the x terminators in the ladder after the gene stop codon.

[0200] In this way a “library” (also referred to as a HTP genetic design library) of strains is constructed, wherein each member of the library is an instance of x terminator linked to n target, in an otherwise identical genetic context. As previously described, combinations of terminators
can be inserted, extending the range of combinatorial possibilities upon which the library is constructed.

[0201] 4. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0202] This foundational process can be extended to provide further improvements in strain performance by, *inter alia*: (1) Consolidating multiple beneficial perturbations into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations can be either a specific set of defined changes or a partly randomized, combinatorial library of changes. For example, if the set of targets is every gene in a pathway, then sequential regeneration of the library of perturbations into an improved member or members of the previous library of strains can optimize the expression level of each gene in a pathway regardless of which genes are rate limiting at any given iteration; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches.

[0203] The approach is exemplified in the present disclosure with industrial microorganisms, but is applicable to any organism where desired traits can be identified in a population of genetic mutants. For example, this could be used for improving the performance of CHO cells, yeast, insect cells, algae, as well as multi-cellular organisms, such as plants.

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

[0204] In one embodiment, the methods of the provided disclosure comprise codon optimizing one or more genes expressed by the host organism. Methods for optimizing codons to improve expression in various hosts are known in the art and are described in the literature (see U.S. Pat. App. Pub. No. 2007/0292918, incorporated herein by reference in its entirety). Optimized coding sequences containing codons preferred by a particular prokaryotic or eukaryotic host (see also, Murray *et al.* (1989) Nucl. Acids Res. 17:477-508) can be prepared, for example, to increase the rate of translation or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, as compared with transcripts produced from a non-optimized sequence.
[0205] Protein expression is governed by a host of factors including those that affect transcription, mRNA processing, and stability and initiation of translation. Optimization can thus address any of a number of sequence features of any particular gene. As a specific example, a rare codon induced translational pause can result in reduced protein expression. A rare codon induced translational pause includes the presence of codons in the polynucleotide of interest that are rarely used in the host organism may have a negative effect on protein translation due to their scarcity in the available tRNA pool.

[0206] Alternate translational initiation also can result in reduced heterologous protein expression. Alternate translational initiation can include a synthetic polynucleotide sequence inadvertently containing motifs capable of functioning as a ribosome binding site (RBS). These sites can result in initiating translation of a truncated protein from a gene-internal site. One method of reducing the possibility of producing a truncated protein, which can be difficult to remove during purification, includes eliminating putative internal RBS sequences from an optimized polynucleotide sequence.

[0207] Repeat-induced polymerase slippage can result in reduced heterologous protein expression. Repeat-induced polymerase slippage involves nucleotide sequence repeats that have been shown to cause slippage or stuttering of DNA polymerase which can result in frameshift mutations. Such repeats can also cause slippage of RNA polymerase. In an organism with a high G+C content bias, there can be a higher degree of repeats composed of G or C nucleotide repeats. Therefore, one method of reducing the possibility of inducing RNA polymerase slippage, includes altering extended repeats of G or C nucleotides.

[0208] Interfering secondary structures also can result in reduced heterologous protein expression. Secondary structures can sequester the RBS sequence or initiation codon and have been correlated to a reduction in protein expression. Stemloop structures can also be involved in transcriptional pausing and attenuation. An optimized polynucleotide sequence can contain minimal secondary structures in the RBS and gene coding regions of the nucleotide sequence to allow for improved transcription and translation.

[0209] For example, the optimization process can begin by identifying the desired amino acid sequence to be expressed by the host. From the amino acid sequence a candidate polynucleotide or DNA sequence can be designed. During the design of the synthetic DNA sequence, the
frequency of codon usage can be compared to the codon usage of the host expression organism and rare host codons can be removed from the synthetic sequence. Additionally, the synthetic candidate DNA sequence can be modified in order to remove undesirable enzyme restriction sites and add or remove any desired signal sequences, linkers or untranslated regions. The synthetic DNA sequence can be analyzed for the presence of secondary structure that may interfere with the translation process, such as G/C repeats and stem-loop structures.


[0210] The present transposon mutagenesis HTP molecular tool solves two problems: First, there is a lack of understanding of genotype-phenotype relationships. Even in well-studied organisms, large portions of the genomic landscape remain poorly understood. Further, well-understood genetic elements may interact in unexpected ways. Second, with slow-growing or genetically recalcitrant organisms, especially those with large genomes, it is time and/or cost prohibitive to perform targeted genetic perturbations on all possible genetic targets.

[0211] To solve these problems, the present disclosure provides methods for readily and randomly modulating/perturbing/engineering genetic elements of host organisms using in vivo transposon mutagenesis.

[0212] Transposon mutagenesis can be used to create libraries that harbor different genetic perturbations/changes (e.g. gain-of-function or loss-of-function) and implicate new genetic targets to further improve a host’s phenotype.

[0213] Without being bound by theory, in general, transposons are characterized by having short (typically less than 50 bp), transposon-specific terminal DNA sequences. In many cases, these terminal sequences are inverted versions of the same, or closely related, sequences. The transposase binds specifically to the terminal inverted repeat sequences to form a transposase-DNA synaptic complex, which catalyzes the transposition events. The transposons may further include any desired DNA sequence (e.g. any payload gene, selectable marker, promoters, primer binding sites, site-specific recombination sites, T7 RNA polymerase promoters, reporter genes, terminators, etc.).
[0214] Certain tools described in the present disclosure concerns existing polymorphs of genes in microbial strains, but do not create novel mutations that may be useful for improving performance of the microbial strains. The present disclosure teaches a transposon mutagenesis system that randomly integrates the payload DNA into the genome to create mutations that can be further screened for those leading to improved features of the host strains, which in turn cause beneficial effects on overall-host strain phenotype (e.g., yield or productivity).

[0215] For example, in some embodiments, the present disclosure teaches methods of generating mutations/alterations/insertions/deletions (i.e. genetic perturbations) within a host cell genome, which are created by a transposon mutagenesis process. Any particular genomic alteration generated in this process can be grouped together as a transposon mutagenesis library (also termed a transposon mutagenesis diversity library), which is explained in more detail below.

[0216] The resultant microbes that are engineered via this process form HTP genetic design libraries.

[0217] The HTP genetic design library can refer to the actual physical microbial strain collection that is formed via this process, with each member strain being representative of a given mutation/alteration/insertion/deletion (i.e. genetic perturbations) created by transposon mutagenesis, in an otherwise identical genetic background, the strain library being termed a “transposon mutagenesis microbial strain library.”

[0218] Furthermore, the HTP genetic design library can refer to the collection of genetic perturbations—in this case a given perturbation created by transposon mutagenesis—the collection being termed a “transposon mutagenesis library.”

[0219] The microbes from the transposon mutagenesis microbial strain library can be subjected to additional rounds of HTP. The microbes from the transposon mutagenesis microbial strain library could be appropriately screened and characterized and give rise to another HTP genetic design library. The characterization of the microbial strains in the HTP genetic design library produces information and data that can be stored in any data storage construct, including a relational database, an object-oriented database or a highly distributed NoSQL database. This data/information could be, for example, a genetic perturbation's effect on host cell growth or production of a molecule in the host cell. This data/information can also be the broader set of combinatorial effects that result from two or more genetic perturbations.
[0220] The transposon mutagenesis microbial strain library can be subjected to additional rounds of cyclical engineering to further improve the desired phenotype (e.g. tryptophan yield). The additional rounds of engineering may consist of transposon mutagenesis or other library types described herein such as SNP Swap, PRO Swap, or random mutagenesis. The improved strains may be screened against a desired phenotype to identify variants with improved performance, and may also be consolidated with other strain variants exhibiting an improved phenotype to produce a further improved strain through the additive effect of distinct beneficial mutations.

[0221] Persons having skill in the art will recognize the ability to consolidate a genetic perturbation created by transposon mutagenesis with any other genetic perturbation. Thus, in some embodiments, the present disclosure teaches transposon mutagenesis microbial strain libraries with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more genetic perturbations created by transposon mutagenesis.

[0222] In summary, utilizing various mutations/alterations/insertions/deletions (also referred to as genetic perturbations) created by transposon mutagenesis in an organism is a powerful tool to optimize a trait of interest. The molecular tool of utilizing transposon mutagenesis to create HTP libraries, developed by the inventors, uses a collection of mutations/alterations/insertions/deletions having vary effect on a trait of interest. This collection is then systematically applied in the organism using high-throughput genome engineering. This group of mutations/alterations/insertions/deletions is determined to have a high likelihood of impacting the trait of interest based on any one of a number of methods. These could include selection based on known function, or impact on the trait of interest, or algorithmic selection based on previously determined beneficial genetic diversity. In some embodiments, the selection of mutations/alterations/insertions/deletions can include all the genes in a given host. In other embodiments, the selection of mutations/alterations/insertions/deletions can be a subset of all genes in a given host, chosen randomly. In other embodiments, the selection of mutations/alterations/insertions/deletions can be a subset of all genes involved in the synthesis of a given molecule.

[0223] The resultant HTP genetic design microbial strain library of organisms containing genetic perturbations created by transposon mutagenesis is then assessed for performance in a high-throughput screening model, and genetic perturbations which lead to increased performance are
determined and the information stored in a database. The collection of genetic perturbations (e.g. mutations/alterations/insertions/deletions) form a “transposon mutagenesis library,” which can be utilized as a source of potential genetic alterations in future microbial engineering processing. Over time, as a greater set of genetic perturbations is implemented against a greater diversity of host cell backgrounds, each library becomes more powerful as a corpus of experimentally confirmed data that can be used to more precisely and predictably design targeted changes against any background of interest.

[0224] In some embodiments, the transposon mutagenesis library of the present disclosure can be used to identify optimum expression of a gene target. In some embodiments, the goal may be to increase activity of a target gene to reduce bottlenecks in a metabolic or genetic pathway. In other embodiments, the goal may be to reduce the activity of the target gene to avoid unnecessary energy expenditures in the host cell, when expression of the target gene is not required.

[0225] Thus, in particular embodiments, transposon mutagenesis is a multi-step process comprising:

[0226] 1. Selecting a transposon system for mutagenesis and applying the system in a given microbial strain to generate mutations (or any other genetic perturbation, but mutation will be used for simplicity in this synopsis) caused by the transposon. Ideally the system is shown to lead to random integration of transposon into the genome of a selected microbial strain. Such integration perturbs gene expression in some way.

[0227] 2. High-throughput strain engineering to rapidly select strains having integrated transposon in its genome. In this way a “library” (also referred to as a HTP genetic design library, i.e. a transposon mutagenesis microbial strain library ) of strains is constructed, wherein each member of the library is a strain comprising a transposon mutation, in an otherwise identical genetic context. As previously described, combinations of mutations can be consolidated, extending the range of combinatorial possibilities upon which the library is constructed.

[0228] 3. High-throughput screening of the library of strains in a context where their performance against one or more metrics is indicative of the performance that is being optimized.

[0229] This foundational process can be extended to provide further improvements in strain performance by, inter alia: (1) Consolidating multiple beneficial perturbations (e.g. mutations)
into a single strain background, either one at a time in an interactive process, or as multiple changes in a single step. Multiple perturbations (e.g. mutations) can be either a specific set of defined changes or a partly randomized, combinatorial library of changes, regardless of the gene function that has been modified by the mutations; (2) Feeding the performance data resulting from the individual and combinatorial generation of the library into an algorithm that uses that data to predict an optimum set of perturbations based on the interaction of each perturbation; and (3) Implementing a combination of the above two approaches.

[0230] In some embodiments, the transposon has preference for insertion at GC-rich regions. In some embodiments, the transposon requires GC-bases at the insertion site. In some embodiments, the transposon has preference for AT-rich regions at the insertion site. In some embodiments, the transposon requires AT-bases at the insertion site.

[0231] In some embodiments, the transposon payload includes a non-coding DNA sequence that can alter the nature of the product expressed by a coding region when the transposon inserts the nucleic acid sequence in or near that coding region in a cell. Any nucleotide sequence that will alter the nature of the product expressed by a coding region present in the cell can be used.

[0232] In some embodiments, the transposon payload includes a non-coding DNA sequence that can alter the level of expression of a coding region when the transposon inserts near that coding region in a cell. This affective sequence may either increase or decrease the level of expression of a coding region. Any nucleotide sequence that will alter the level of expression of a coding region present in a cell can be used.

[0233] In some embodiments, the one or more non-coding or coding DNA sequences include, but are not limited to, promoters, terminator sequences, stop codons, optimized codons, splice acceptor sites, splice donor sites, silencer elements, SNPs, solubility tags, bar codes, enhancers, matrix attachment sequences, transcription binding sites, frame-shift mutations, selectable markers, and counter-selectable markers.

[0234] In some embodiments, the transposon payload includes a selectable marker. Selectable markers that may be used in the present disclosure include but are not limited to, drug resistance markers (e.g. hygromycin, kanamycin, beta-lactamase resistance, puromycin, or the neomycin analog G418), detectable markers (e.g. fluorescent proteins, luciferase, chloramphenicol acetyl
transferase, and beta-galactosidase), mFabI, chloramphenicol resistance, and auxotrophic markers (e.g. URA, LYS, cscA).

[0235] In some embodiments, the transposon payload includes a counter-selectable marker including, but not limited to, URA3/5-FOA counter-selection system, sacB, tetAR, rpsL, ccdB, pheS, and thymidine kinase.

[0236] The transposon payload may be varied to elicit diverse phenotypic responses. For example, in a loss-of-function (LoF) library, the payload may include a marker that allows for the selection of successful transposon integration events. In another example, in a gain-of-function library, the payload may include promoters or solubility tags. In other embodiments, the payload may include counter-selectable markers that facilitate loop-out of a portion of the payload containing the selectable marker, thus allowing serial transposon mutagenesis.

[0237] In some embodiments, the transposon has a high frequency of transposition. In some embodiments, the transposon has a high frequency of transposition so that it is possible to achieve saturated mutagenesis (e.g. insert into every gene in the genome at least once).

[0238] Any appropriate transposon system may be used in the present disclosure. In some embodiments, the transposon is a cut-and-paste transposon. In some embodiments, the transposon is a replicative transposon. In some embodiments, the transposon is a retro element, where transposition is accomplished through a process involving reverse transcription. In some embodiments, the transposon and transposase systems are selected from the group including, but not limited to, Tn1, Tn2, Tn3, Tn4, Tn5, Tn6, Tn7, Tn10, mariner, Himar 1, Tol2, Frog Prince, P-elements, Passport, Tn4001, Ty1, Ty2, Ty3, Ty4, Ty5, synthetic transposons, Sleeping Beauty, piggyback, or derivatives thereof. In some embodiments, the transposon system is the Tn5 transpososome system.

[0239] In some embodiments, the transposon is a composite transposon made up of two or more transposon payloads. In some embodiments, the one or more transposon payloads is complexed with the transposase. In some embodiments, the complexed transposon payload and transposase allows for in vivo transposition. In some embodiments, the complexed transposase is polypeptide. In some embodiments, the complexed transposase is a polynucleotide encoding a transposase polypeptide. In some embodiments, the complexed transposase is Tn5 transposase.
[0240] In some embodiments, the transposon includes polynucleotides that mediate site-specific integration. Site-specific integration sequences that may be used in the present disclosure include, but are not limited to LoxP (for use with Cre recombinase) and FRT (for use with FLP recombinase).

[0241] In some embodiments, the transposon inserts randomly into the genome. In some embodiments, the transposon inserts randomly into the genome and causes loss of function mutations. In some embodiments the transposon inserts into the promoter of a gene. In some embodiments, the transposon randomly inserts into an open read frame and prevents transcription or translation of the disrupted gene (e.g. a loss-of-function mutation). In some embodiments, the transposon inserts into an upstream regulatory element of a gene. In some embodiments, the transposon randomly inserts in a site proximal to the gene and increases gene expression (e.g. a gain-of-function mutation). In some embodiments, the transposon inserts into the promoter or upstream regulatory element of a gene and causes a gain-of-function mutation. In some embodiments, the transposon inserts into the promoter or upstream regulatory element of a gene and causes a loss-of-function mutation. In some embodiments, the transposon inserts into a gene and causes an early termination mutation. In some embodiments the early termination mutation causes a loss-of-function mutation.

[0242] In some embodiments, the transposon integrates into the genomic DNA at the insertion site. In some embodiments, the transposon is stably inherited by the microbial organism.

[0243] In some embodiments, the transposon inserts one or more DNA sequences (e.g. transposon payload) at the insertion site in the genome. In some embodiments, the transposon includes one or more disruptive sequences and/or one or more affective sequences, or a combination thereof.

[0244] In some embodiments, the transposon results in deletion of a portion of genomic DNA. In some embodiments, the deletion of a portion of genomic DNA is accomplished through Cre-catalyzed excision of DNA.

[0245] The transposon may be delivered to a cell using any appropriate vector. In some embodiments, a vector may include at least one transposon, at least two transposons, at least 3 transposons, at least 4 transposons, at least 5 transposons, at least 6 transposons, at least 7 transposons, at least 8 transposons, at least 9 transposons, at least 10 transposons, or more.
[0246] In some embodiments, the vector includes a coding region encoding a transposase. As used herein, the term “transposase” refers to a polypeptide that binds an inverted repeat or a direct repeat of a transposon and catalyzes the excision of a transposon from a donor polynucleotide (e.g. a vector) and subsequent integration of the transposon into the genomic DNA of a cell. The transposase may be present as a polypeptide. Alternatively, the transposase may be present as a polynucleotide that includes a coding sequence encoding a transposase. The polynucleotide may be RNA (e.g. mRNA) or DNA. The polynucleotide encoding a transposase may be on a vector, or present in a chromosome. When the transposase is present as a coding sequence encoding the transposase, in some aspects of the disclosure, the coding sequence may be present on the same polynucleotide (e.g. a vector) that includes the transposon (i.e. in cis). In some embodiments, the transposase coding sequence may be present on a second polynucleotide (e.g. a vector), i.e. in trans.

[0247] The present disclosure provides methods for using the transposons and vectors disclosed herein. The vectors may be transformed into the target cell, evaluated, and cloned using any appropriate means known in the art. The method may include observing the cells to determine if a phenotype has changed.

[0248] The methods disclosed may include mapping the location of the transposons present in a cell. In some embodiments, the area of insertion may be identified by sequence analysis. Sequence analysis may be performed by any appropriate means in the art, including but not limited to, PCR-based techniques (e.g. inverse PCR or linker-mediated PCR techniques). In some embodiments, sequence analysis comprises use of a transposon-specific primer (Tn primer) coupled with an arbitrary primer to PCR-amplify one of the transposon boundaries, which is subsequently sequenced in order to identify the target DNA immediately adjacent to the transposon end sequence. In some embodiments, sequence analysis comprises use of a transposon-specific primer and primers designed to known sequences in the microbial genome (e.g. “footprinting”). In some embodiments, sequence analysis may be performed by assaying unique sequences built into the transposon (e.g. a specific 20-mer or a bar-code) which may be identified by hybridization. In some embodiments, sequence analysis includes microarray analysis. In some embodiments, sequence analysis includes in situ hybridization. In some embodiments, sequence analysis using a restriction endonuclease capable of cleaving a restriction site within the transposon.
7. Epistasis Mapping – A Predictive Analytical Tool Enabling Beneficial Genetic Consolidations

[0249] In some embodiments, the present disclosure teaches epistasis mapping methods for predicting and combining beneficial genetic alterations into a host cell. The genetic alterations may be created by any of the aforementioned HTP molecular tool sets (e.g., promoter swaps, SNP swaps, start/stop codon exchanges, sequence optimization, transposon mutagenesis) and the effect of those genetic alterations would be known from the characterization of the derived HTP genetic design microbial strain libraries. Thus, as used herein, the term epistasis mapping includes methods of identifying combinations of genetic alterations (e.g., beneficial SNPs or beneficial promoter/target gene associations, or beneficial mutations from a transposon mutagenesis campaign) that are likely to yield increases in host performance.

[0250] In embodiments, the epistasis mapping methods of the present disclosure are based on the idea that the combination of beneficial mutations from two different functional groups is more likely to improve host performance, as compared to a combination of mutations from the same functional group. See, e.g., Costanzo, The Genetic Landscape of a Cell, Science, Vol. 327, Issue 5964, Jan. 22, 2010, pp. 425-431 (incorporated by reference herein in its entirety).

[0251] Mutations from the same functional group are more likely to operate by the same mechanism, and are thus more likely to exhibit negative or neutral epistasis on overall host performance. In contrast, mutations from different functional groups are more likely to operate by independent mechanisms, which can lead to improved host performance and in some instances synergistic effects.

[0252] Thus, in some embodiments, the present disclosure teaches methods of analyzing SNP mutations to identify SNPs predicted to belong to different functional groups. In some embodiments, SNP functional group similarity is determined by computing the cosine similarity of mutation interaction profiles (similar to a correlation coefficient, see Figure 8A). The present disclosure also illustrates comparing SNPs via a mutation similarity matrix (see Figure 7) or dendrogram (see Figure 8A). The same concept could be applied to a genetic perturbation brought about by transposon mutagenesis.
[0253] Thus, the epistasis mapping procedure provides a method for grouping and/or ranking a diversity of genetic mutations applied in one or more genetic backgrounds for the purposes of efficient and effective consolidations of the mutations into one or more genetic backgrounds.

[0254] In aspects, consolidation is performed with the objective of creating novel strains which are optimized for the production of target biomolecules. Through the taught epistasis mapping procedure, it is possible to identify functional groupings of mutations, and such functional groupings enable a consolidation strategy that minimizes undesirable epistatic effects.

[0255] As previously explained, the optimization of microbes for use in industrial fermentation is an important and difficult problem, with broad implications for the economy, society, and the natural world. Traditionally, microbial engineering has been performed through a slow and uncertain process of random mutagenesis. Such approaches leverage the natural evolutionary capacity of cells to adapt to artificially imposed selection pressure. Such approaches are also limited by the rarity of beneficial mutations, the ruggedness of the underlying fitness landscape, and more generally underutilize the state of the art in cellular and molecular biology.

[0256] Modern approaches leverage new understanding of cellular function at the mechanistic level and new molecular biology tools to perform targeted genetic manipulations to specific phenotypic ends. In practice, such rational approaches are confounded by the underlying complexity of biology. Causal mechanisms are poorly understood, particularly when attempting to combine two or more changes that each has an observed beneficial effect. Sometimes such consolidations of genetic changes yield positive outcomes (measured by increases in desired phenotypic activity), although the net positive outcome may be lower than expected and in some cases higher than expected. In other instances, such combinations produce either net neutral effect or a net negative effect. This phenomenon is referred to as epistasis, and is one of the fundamental challenges to microbial engineering (and genetic engineering generally).

[0257] As aforementioned, the present HTP genomic engineering platform solves many of the problems associated with traditional microbial engineering approaches. The present HTP platform uses automation technologies to perform hundreds or thousands of genetic mutations at once. In particular aspects, unlike the rational approaches described above, the disclosed HTP platform enables the parallel construction of thousands of mutants to more effectively explore large subsets of the relevant genomic space, as disclosed in U.S. Application No. 15/140,296, entitled Microbial
Strain Design System And Methods For Improved Large-Scale Production Of Engineered Nucleotide Sequences, incorporated by reference herein in its entirety. By trying “everything,” the present HTP platform sidesteps the difficulties induced by our limited biological understanding.

[0258] However, at the same time, the present HTP platform faces the problem of being fundamentally limited by the combinatorial explosive size of genomic space, and the effectiveness of computational techniques to interpret the generated data sets given the complexity of genetic interactions. Techniques are needed to explore subsets of vast combinatorial spaces in ways that maximize non-random selection of combinations that yield desired outcomes.

[0259] Somewhat similar HTP approaches have proved effective in the case of enzyme optimization. In this niche problem, a genomic sequence of interest (on the order of 1000 bases), encodes a protein chain with some complicated physical configuration. The precise configuration is determined by the collective electromagnetic interactions between its constituent atomic components. This combination of short genomic sequence and physically constrained folding problem lends itself specifically to greedy optimization strategies. That is, it is possible to individually mutate the sequence at every residue and shuffle the resulting mutants to effectively sample local sequence space at a resolution compatible with the Sequence Activity Response modeling.

[0260] However, for full genomic optimizations for biomolecules, such residue-centric approaches are insufficient for some important reasons. First, because of the exponential increase in relevant sequence space associated with genomic optimizations for biomolecules. Second, because of the added complexity of regulation, expression, and metabolic interactions in biomolecule synthesis. The present inventors have solved these problems via the taught epistasis mapping procedure.

[0261] The taught method for modeling epistatic interactions, between a collection of mutations for the purposes of more efficient and effective consolidation of the mutations into one or more genetic backgrounds, is groundbreaking and highly needed in the art.

[0262] When describing the epistasis mapping procedure, the terms “more efficient” and “more effective” refers to the avoidance of undesirable epistatic interactions among consolidation strains with respect to particular phenotypic objectives.
[0263] As the process has been generally elaborated upon above, a more specific workflow example will now be described.

[0264] First, one begins with a library of M mutations and one or more genetic backgrounds (e.g., parent bacterial strains). Neither the choice of library nor the choice of genetic backgrounds is specific to the method described here. But in a particular implementation, a library of mutations may include exclusively, or in combination: SNP swap libraries, Promoter swap libraries, Transposon mutagenesis libraries, or any other mutation library described herein, or any combination thereof.

[0265] In one implementation, only a single genetic background is provided. In this case, a collection of distinct genetic backgrounds (microbial mutants) will first be generated from this single background. This may be achieved by applying the primary library of mutations (or some subset thereof) to the given background for example, application of a HTP genetic design library of particular SNPs or a HTP genetic design library of particular promoters to the given genetic background, to create a population (perhaps 100’s or 1,000’s) of microbial mutants with an identical genetic background except for the particular genetic alteration from the given HTP genetic design library incorporated therein. As detailed below, this embodiment can lead to a combinatorial library or pairwise library.

[0266] In another implementation, a collection of distinct known genetic backgrounds may simply be given. As detailed below, this embodiment can lead to a subset of a combinatorial library.

[0267] In a particular implementation, the number of genetic backgrounds and genetic diversity between these backgrounds (measured in number of mutations or sequence edit distance or the like) is determined to maximize the effectiveness of this method.

[0268] A genetic background may be a natural, native or wild-type strain or a mutated, engineered strain. N distinct background strains may be represented by a vector \( \mathbf{b} \). In one example, the background \( \mathbf{b} \) may represent engineered backgrounds formed by applying \( N \) primary mutations \( \mathbf{m}_0 = (m_1, m_2, \ldots, m_N) \) to a wild-type background strain \( b_0 \) to form the \( N \) mutated background strains \( \mathbf{b} = \mathbf{m}_0 \mathbf{b}_0 = (m_1 \mathbf{b}_0, m_2 \mathbf{b}_0, \ldots, m_N \mathbf{b}_0) \), where \( m_i \mathbf{b}_0 \) represents the application of mutation \( m_i \) to background strain \( b_0 \).
[0269] In either case (i.e. a single provided genetic background or a collection of genetic backgrounds), the result is a collection of N genetically distinct backgrounds. Relevant phenotypes are measured for each background.

[0270] Second, each mutation in a collection of M mutations $m_1$ is applied to each background within the collection of N background strains $b$ to form a collection of M x N mutants. In the implementation where the N backgrounds were themselves obtained by applying the primary set of mutations $m_0$ (as described above), the resulting set of mutants will sometimes be referred to as a combinatorial library or a pairwise library. In another implementation, in which a collection of known backgrounds has been provided explicitly, the resulting set of mutants may be referred to as a subset of a combinatorial library. Similar to generation of engineered background vectors, in embodiments, the input interface 202 receives the mutation vector $m_1$ and the background vector $b$, and a specified operation such as cross product.

[0271] Continuing with the engineered background example above, forming the MxN combinatorial library may be represented by the matrix formed by $m_1 \times m_0 b_0$, the cross product of $m_1$ applied to the N backgrounds of $b = m_0 b_0$, where each mutation in $m_1$ is applied to each background strain within $b$. Each ith row of the resulting MxN matrix represents the application of the ith mutation within $m_1$ to all the strains within background collection $b$. In one embodiment, $m_1 = m_0$ and the matrix represents the pairwise application of the same mutations to starting strain $b_0$. In that case, the matrix is symmetric about its diagonal (M=N), and the diagonal may be ignored in any analysis since it represents the application of the same mutation twice.

[0272] In embodiments, forming the MxN matrix may be achieved by inputting into the input interface 202 the compound expression $m_1 \times m_0 b_0$. The component vectors of the expression may be input directly with their elements explicitly specified, via one or more DNA specifications, or as calls to the library 206 to enable retrieval of the vectors during interpretation by interpreter 204. As described in U.S. Patent Application, Serial No. 15/140,296, entitled “Microbial Strain Design System and Methods for Improved Large Scale Production of Engineered Nucleotide Sequences,” via the interpreter 204, execution engine 207, order placement engine 208, and factory 210, the LIMS system 200 generates the microbial strains specified by the input expression.

[0273] Third, with reference to the flowchart of Figure 24, the analysis equipment 214 (Figure 20) measures phenotypic responses for each mutant within the MxN combinatorial library matrix
(4202). As such, the collection of responses can be construed as an $M \times N$ Response Matrix $R$. Each element of $R$ may be represented as $r_{ij} = y(m_i, b_j)$, where $y$ represents the response (performance) of background strain $b_j$ within engineered collection $b$ as mutated by mutation $m_i$. For simplicity, and practicality, we assume pairwise mutations where $m_1 = m_0$. Where, as here, the set of mutations represents a pairwise mutation library, the resulting matrix may also be referred to as a gene interaction matrix or, more particularly, as a mutation interaction matrix.

[0274] Those skilled in the art will recognize that, in some embodiments, operations related to epistatic effects and predictive strain design may be performed entirely through automated means of the LIMS system 200, e.g., by the analysis equipment 214, or by human implementation, or through a combination of automated and manual means. When an operation is not fully automated, the elements of the LIMS system 200, e.g., analysis equipment 214, may, for example, receive the results of the human performance of the operations rather than generate results through its own operational capabilities. As described elsewhere herein, components of the LIMS system 200, such as the analysis equipment 214, may be implemented wholly or partially by one or more computer systems. In some embodiments, in particular where operations related to predictive strain design are performed by a combination of automated and manual means, the analysis equipment 214 may include not only computer hardware, software or firmware (or a combination thereof), but also equipment operated by a human operator such as that listed in Table 3 below, e.g., the equipment listed under the category of “Evaluate performance.”

[0275] Fourth, the analysis equipment 212 normalizes the response matrix. Normalization consists of a manual and/or, in this embodiment, automated processes of adjusting measured response values for the purpose of removing bias and/or isolating the relevant portions of the effect specific to this method. With respect to Figure 24, the first step 4202 may include obtaining normalized measured data. In general, in the claims directed to predictive strain design and epistasis mapping, the terms “performance measure” or “measured performance” or the like may be used to describe a metric that reflects measured data, whether raw or processed in some manner, e.g., normalized data. In a particular implementation, normalization may be performed by subtracting a previously measured background response from the measured response value. In that implementation, the resulting response elements may be formed as $r_{ij} = y(m_i, b_j) - y(m_i)$, where $y(m_i)$ is the response of the engineered background strain $b_j$ within engineered collection $b$ caused by application of primary mutation $m_i$ to parent strain $b_0$. Note that each row of the normalized response matrix is
treated as a response profile for its corresponding mutation. That is, the $i$th row describes the relative effect of the corresponding mutation $m_i$ applied to all the background strains $b_j$ for $j=1$ to $N$.

**[0276]** With respect to the example of pairwise mutations, the combined performance/response of strains resulting from two mutations may be greater than, less than, or equal to the performance/response of the strain to each of the mutations individually. This effect is known as “epistasis,” and may, in some embodiments, be represented as $e_{ij} = y(m_i, m_j) - (y(m_i) + y(m_j))$. Variations of this mathematical representation are possible, and may depend upon, for example, how the individual changes biologically interact. As noted above, mutations from the same functional group are more likely to operate by the same mechanism, and are thus more likely to exhibit negative or neutral epistasis on overall host performance. In contrast, mutations from different functional groups are more likely to operate by independent mechanisms, which can lead to improved host performance by reducing redundant mutative effects, for example. Thus, mutations that yield dissimilar responses are more likely to combine in an additive manner than mutations that yield similar responses. This leads to the computation of similarity in the next step.

**[0277]** *Fifth*, the analysis equipment 214 measures the similarity among the responses—in the pairwise mutation example, the similarity between the effects of the $i$th mutation and $j$th (e.g., primary) mutation within the response matrix (4204). Recall that the $i$th row of $R$ represents the performance effects of the $i$th mutation $m_i$ on the $N$ background strains, each of which may be itself the result of engineered mutations as described above. Thus, the similarity between the effects of the $i$th and $j$th mutations may be represented by the similarity $s_{ij}$ between the $i$th and $j$th rows, $\rho_i$ and $\rho_j$, respectively, to form a similarity matrix $S$, an example of which is illustrated in Figure 7. Similarity may be measured using many known techniques, such as cross-correlation or absolute cosine similarity, e.g., $s_{ij} = \text{abs}(\cos(\rho_i, \rho_j))$.

**[0278]** As an alternative or supplement to a metric like cosine similarity, response profiles may be clustered to determine degree of similarity. Clustering may be performed by use of a distance-based clustering algorithms (e.g. k-mean, hierarchical agglomerative, etc.) in conjunction with suitable distance measure (e.g. Euclidean, Hamming, etc). Alternatively, clustering may be performed using similarity based clustering algorithms (e.g. spectral, min-cut, etc.) with a suitable similarity measure (e.g. cosine, correlation, etc). Of course, distance measures may be mapped to
similarity measures and vice-versa via any number of standard functional operations (e.g., the exponential function). In one implementation, hierarchical agglomerative clustering may be used in conjunction absolute cosine similarity. (See Figure 8A).

[0279] As an example of clustering, let C be a clustering of mutations m into k distinct clusters. Let C be the cluster membership matrix, where c is the degree to which mutation i belongs to cluster j, a value between 0 and 1. The cluster-based similarity between mutations i and j is then given by Cj×Ci (the dot product of the ith and jth rows of C). In general, the cluster-based similarity matrix is given by CC\(^T\) (that is, C times C-transpose). In the case of hard-clustering (a mutation belongs to exactly one cluster), the similarity between two mutations is 1 if they belong to the same cluster and 0 if not.

[0280] As is described in Costanzo, The Genetic Landscape of a Cell, Science, Vol. 327, Issue 5964, Jan. 22, 2010, pp. 425-431 (incorporated by reference herein in its entirety), such a clustering of mutation response profiles relates to an approximate mapping of a cell's underlying functional organization. That is, mutations that cluster together tend to be related by an underlying biological process or metabolic pathway. Such mutations are referred to herein as a “functional group.” The key observation of this method is that if two mutations operate by the same biological process or pathway, then observed effects (and notably observed benefits) may be redundant. Conversely, if two mutations operate by distant mechanism, then it is less likely that beneficial effects will be redundant.

[0281] Sixth, based on the epistatic effect, the analysis equipment 214 selects pairs of mutations that lead to dissimilar responses, e.g., their cosine similarity metric falls below a similarity threshold, or their responses fall within sufficiently separated clusters, (e.g., in Figure 7 and Figure 8A) as shown in Figure 24 (4206). Based on their dissimilarity, the selected pairs of mutations should consolidate into background strains better than similar pairs.

[0282] Based upon the selected pairs of mutations that lead to sufficiently dissimilar responses, the LIMS system (e.g., all of or some combination of interpreter 204, execution engine 207, order placer 208, and factory 210) may be used to design microbial strains having those selected mutations (4208). In embodiments, as described below and elsewhere herein, epistatic effects may be built into, or used in conjunction with the predictive model to weight or filter strain selection.
[0283] It is assumed that it is possible to estimate the performance (a.k.a. score) of a hypothetical strain obtained by consolidating a collection of mutations from the library into a particular background via some preferred predictive model. A representative predictive model utilized in the taught methods is provided in the below section entitled “Predictive Strain Design” that is found in the larger section of: “Computational Analysis and Prediction of Effects of Genome-Wide Genetic Design Criteria.”

[0284] When employing a predictive strain design technique such as linear regression, the analysis equipment 214 may restrict the model to mutations having low similarity measures by, e.g., filtering the regression results to keep only sufficiently dissimilar mutations. Alternatively, the predictive model may be weighted with the similarity matrix. For example, some embodiments may employ a weighted least squares regression using the similarity matrix to characterize the interdependencies of the proposed mutations. As an example, weighting may be performed by applying the "kernel" trick to the regression model. (To the extent that the "kernel trick" is general to many machine learning modeling approaches, this re-weighting strategy is not restricted to linear regression.)

[0285] Such methods are known to one skilled in the art. In embodiments, the kernel is a matrix having elements $1 - w \cdot s_{ij}$ where 1 is an element of the identity matrix, and $w$ is a real value between 0 and 1. When $w = 0$, this reduces to a standard regression model. In practice, the value of $w$ will be tied to the accuracy ($r^2$ value or root mean square error (RMSE)) of the predictive model when evaluated against the pairwise combinatorial constructs and their associate effects $y(m_i, m_j)$. In one simple implementation, $w$ is defined as $w = 1 - r^2$. In this case, when the model is fully predictive, $w = 1 - r^2 = 0$ and consolidation is based solely on the predictive model and epistatic mapping procedure plays no role. On the other hand, when the predictive model is not predictive at all, $w = 1 - r^2 = 1$ and consolidation is based solely on the epistatic mapping procedure. During each iteration, the accuracy can be assessed to determine whether model performance is improving.

[0286] It should be clear that the epistatic mapping procedure described herein does not depend on which model is used by the analysis equipment 214. Given such a predictive model, it is possible to score and rank all hypothetical strains accessible to the mutation library via combinatorial consolidation.
[0287] In some embodiments, to account for epistatic effects, the dissimilar mutation response profiles may be used by the analysis equipment 214 to augment the score and rank associated with each hypothetical strain from the predictive model. This procedure may be thought of broadly as a re-weighting of scores, so as to favor candidate strains with dissimilar response profiles (e.g., strains drawn from a diversity of clusters). In one simple implementation, a strain may have its score reduced by the number of constituent mutations that do not satisfy the dissimilarity threshold or that are drawn from the same cluster (with suitable weighting). In a particular implementation, a hypothetical strain’s performance estimate may be reduced by the sum of terms in the similarity matrix associated with all pairs of constituent mutations associated with the hypothetical strain (again with suitable weighting). Hypothetical strains may be re-ranked using these augmented scores. In practice, such re-weighting calculations may be performed in conjunction with the initial scoring estimation.

[0288] The result is a collection of hypothetical strains with score and rank augmented to more effectively avoid confounding epistatic interactions. Hypothetical strains may be constructed at this time, or they may be passed to another computational method for subsequent analysis or use.

[0289] Those skilled in the art will recognize that epistasis mapping and iterative predictive strain design as described herein are not limited to employing only pairwise mutations, but may be expanded to the simultaneous application of many more mutations to a background strain. In another embodiment, additional mutations may be applied sequentially to strains that have already been mutated using mutations selected according to the predictive methods described herein. In another embodiment, epistatic effects are imputed by applying the same genetic mutation to a number of strain backgrounds that differ slightly from each other, and noting any significant differences in positive response profiles among the modified strain backgrounds.

Organisms Amenable to Genetic Design

[0290] The disclosed HTP genomic engineering platform is exemplified with industrial microbial cell cultures (e.g., Corynebacterium, E. coli, A. niger, and Saccharopolyspora spp), but is applicable to any host cell organism where desired traits can be identified in a population of genetic mutants.

[0291] Thus, as used herein, the term “microorganism” should be taken broadly. It includes, but is not limited to, the two prokaryotic domains, Bacteria and Archaea, as well as certain eukaryotic
fungi and protists. However, in certain aspects, “higher” eukaryotic organisms such as insects, plants, and animals can be utilized in the methods taught herein.

[0292] Suitable host cells include, but are not limited to: bacterial cells, algal cells, plant cells, fungal cells, insect cells, and mammalian cells. In one illustrative embodiment, suitable host cells include *E. coli* (e.g., SHuffle™ competent *E. coli* available from New England BioLabs in Ipswich, Mass.).

[0293] Suitable host strains of the *E. coli* species comprise: Enterotoxigenic *E. coli* (ETEC), Enteropathogenic *E. coli* (EPEC), Enteroinvasive *E. coli* (EIEC), Enterohemorrhagic *E. coli* (EHEC), Uropathogenic *E. coli* (UPEC), Verotoxin-producing *E. coli*, *E. coli* O157:H7, *E. coli* O104:H4, *Escherichia coli* O121, *Escherichia coli* O104:H21, *Escherichia coli* K1, and *Escherichia coli* NC101.


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

In some embodiments, the present disclosure teaches enterotoxigenic *E. coli* (ETEC), such as strains BC 5581 (O78:H11), BC 5583 (O2:K1), BC 8221 (O118), BC 8222 (O148:H-), BC 8223 (O111), BC 8224 (O110:H-), BC 8225 (O148), BC 8226 (O118), BC 8227 (O25:H42), BC 8229 (O6), BC 8231 (O153:H45), BC 8232 (O9), BC 8233 (O148), BC 8234 (O128), BC 8235 (O118), BC 8237 (O111), BC 8238 (O110:H17), BC 8240 (O148), BC 8241 (O6:H16), BC 8243 (O153), BC 8244 (O15:H-), BC 8245 (O20), BC 8269 (O125a.c:H-), BC 8313 (O6:H6), BC 8315 (O153:H-), BC 8329, BC 8334 (O118:H12), and BC 8339.

In some embodiments, the present disclosure teaches enteropathogenic *E. coli* (EPEC), such as strains BC 7567 (O86), BC 7568 (O128), BC 7571 (O114), BC 7572 (O119), BC 7573 (O125), BC 7574 (O124), BC 7576 (O127a), BC 7577 (O126), BC 7578 (O142), BC 7579 (O26), BC 7580 (OK26), BC 7581 (O142), BC 7582 (O55), BC 7583 (O158), BC 7584 (O-), BC 7585 (O-), BC 7586 (O-), BC 8330, BC 8550 (O26), BC 8551 (O55), BC 8552 (O158), BC 8553 (O26), BC 8554 (O158), BC 8555 (O86), BC 8556 (O128), BC 8557 (OK26), BC 8558 (O55), BC 8560 (O158), BC 8561 (O158), BC 8562 (O114), BC 8563 (O86), BC 8564 (O128), BC 8565 (O158), BC 8566 (O158), BC 8567 (O158), BC 8568 (O111), BC 8569 (O128), BC 8570 (O114), BC 8571 (O128), BC 8572 (O128), BC 8573 (O158), BC 8574 (O158), BC 8575 (O158), BC 8576 (O158), BC 8577 (O158), BC 8578 (O158), BC 8581 (O158), BC 8583 (O128), BC 8584 (O158), BC 8585 (O128), BC 8586 (O158), BC 8588 (O26), BC 8589 (O86), BC 8590 (O127), BC 8591 (O128),
BC 8592 (O114), BC 8593 (O114), BC 8594 (O114), BC 8595 (O125), BC 8596 (O158), BC 8597 (O26), BC 8598 (O26), BC 8599 (O158), BC 8605 (O158), BC 8606 (O158), BC 8607 (O158), BC 8608 (O128), BC 8609 (O55), BC 8610 (O114), BC 8615 (O158), BC 8616 (O128), BC 8617 (O26), BC 8618 (O86), BC 8619, BC 8620, BC 8621, BC 8622, BC 8623, BC 8624 (O158), and BC 8625 (O158).

[0299] In some embodiments, the present disclosure also teaches methods for the engineering of *Shigella* organisms, including *Shigella flexneri*, *Shigella dysenteriae*, *Shigella boydii*, and *Shigella sonnei*.

[0300] Other suitable host organisms of the present disclosure include microorganisms of the genus *Corynebacterium*. In some embodiments, preferred *Corynebacterium* strains/species include: *C. efficiens*, with the deposited type strain being DSM44549, *C. glutamicum*, with the deposited type strain being ATCC13032, and *C. ammoniagenes*, with the deposited type strain being ATCC6871. In some embodiments the preferred host of the present disclosure is *C. glutamicum*.

[0301] Suitable host strains of the genus *Corynebacterium*, in particular of the species *Corynebacterium glutamicum*, are in particular the known wild-type strains: *Corynebacterium glutamicum* ATCC13032, *Corynebacterium acetoglutamicum* ATCC15806, *Corynebacterium acetoacidophilum* ATCC13870, *Corynebacterium melassecola* ATCC17965, *Corynebacterium thermoaminogenes* FERM BP-1539, *Brevibacterium flavum* ATCC14067, *Brevibacterium lactofermentum* ATCC13869, and *Brevibacterium divaricatum* ATCC14020; and L-amino acid-producing mutants, or strains, prepared therefrom, such as, for example, the L-lysine-producing strains: *Corynebacterium glutamicum* FERM-P 1709, *Brevibacterium flavum* FERM-P 1708, *Brevibacterium lactofermentum* FERM-P 1712, *Corynebacterium glutamicum* FERM-P 6463, *Corynebacterium glutamicum* FERM-P 6464, *Corynebacterium glutamicum* DM58-1, *Corynebacterium glutamicum* DG52-5, *Corynebacterium glutamicum* DSM5714, and *Corynebacterium glutamicum* DSM12866.

[0302] The term “*Micrococcus glutamicus*” has also been in use for *C. glutamicum*. Some representatives of the species *C. efficiens* have also been referred to as *C. thermoaminogenes* in the prior art, such as the strain FERM BP-1539, for example.
[0303] In some embodiments, the host cell of the present disclosure is a eukaryotic cell. Suitable eukaryotic host cells include, but are not limited to: fungal cells, algal cells, insect cells, animal cells, and plant cells. Suitable fungal host cells include, but are not limited to: Ascomycota, Basidiomycota, Deuteromycota, Zygomycota, Fungi imperfecti. Certain preferred fungal host cells include yeast cells and filamentous fungal cells. Suitable filamentous fungi host cells include, for example, any filamentous forms of the subdivision Eumycotina and Oomycota. (see, e.g., Hawksworth et al., In Ainsworth and Bisby's Dictionary of The Fungi, 8th edition, 1995, CAB International, University Press, Cambridge, UK, which is incorporated herein by reference). Filamentous fungi are characterized by a vegetative mycelium with a cell wall composed of chitin, cellulose and other complex polysaccharides. The filamentous fungi host cells are morphologically distinct from yeast.

[0304] In certain illustrative, but non-limiting embodiments, the filamentous fungal host cell may be a cell of a species of: Achlya, Acremonium, Aspergillus, Aureobasidium, Bjerkandera, Ceriporiopsis, Cephalosporium, Chrysosporium, Cochliobolus, Corynascus, Cryphonectria, Cryptococcus, Coprinus, Coriolus, Diplodia, Endothis, Fusarium, Gibberella, Gliocladium, Humicola, Hypocrea, Myceliophthora (e.g., Myceliophthora thermophila), Mucor, Neurospora, Penicillium, Podospora, Phlebia, Pironyces, Pyricularia, Rhizomucor, Rhizopus, Schizophyllum, Scytalidium, Sporotrichum, Talaromyces, Thermoascus, Thielavia, Trametes, Toiyocladium, Trichoderma, Verticillium, Volvariella, or teleomorphs, or anamorphs, and synonyms or taxonomic equivalents thereof. In one embodiment, the filamentous fungus is selected from the group consisting of A. nidulans, A. oryzae, A. sojae, and Aspergillus of the A. niger Group. In an embodiment, the filamentous fungus is Aspergillus niger.

[0305] In another embodiment, specific mutants of the fungal species are used for the methods and systems provided herein. In one embodiment, specific mutants of the fungal species are used which are suitable for the high-throughput and/or automated methods and systems provided herein. Examples of such mutants can be strains that protoplast very well; strains that produce mainly or, more preferably, only protoplasts with a single nucleus; strains that regenerate efficiently in microtiter plates, strains that regenerate faster and/or strains that take up polynucleotide (e.g., DNA) molecules efficiently, strains that produce cultures of low viscosity such as, for example, cells that produce hyphae in culture that are not so entangled as to prevent isolation of single clones
and/or raise the viscosity of the culture, strains that have reduced random integration (e.g., disabled non-homologous end joining pathway) or combinations thereof.

[0306] In yet another embodiment, a specific mutant strain for use in the methods and systems provided herein can be strains lacking a selectable marker gene such as, for example, uridine-requiring mutant strains. These mutant strains can be either deficient in orotidine 5-phosphate decarboxylase (OMP) or orotate p-ribosyl transferase (OPRT) encoded by the pyrG or pyrE gene, respectively (T. Goosen et al., Curr Genet. 1987, 11:499 503; J. Begueret et al., Gene. 1984 32:487 92.

[0307] In one embodiment, specific mutant strains for use in the methods and systems provided herein are strains that possess a compact cellular morphology characterized by shorter hyphae and a more yeast-like appearance.

[0308] Suitable yeast host cells include, but are not limited to: Candida, Hansenula, Saccharomyces, Schizosaccharomyces, Pichia, Kluyveromyces, and Yarrowia. In some embodiments, the yeast cell is Hansenula polymorpha, Saccharomyces cerevisiae, Saccharomyces carlsbergensis, Saccharomyces diastaticus, Saccharomyces norbensis, Saccharomyces kluyveri, Schizosaccharomyces pombe, Pichia pastoris, Pichia finlandica, Pichia trehalophila, Pichia kodamae, Pichia membranaefaciens, Pichia opuntiae, Pichia thermotolerans, Pichia salictaria, Pichia quercuum, Pichia pijperi, Pichia stipitis, Pichia methanolica, Pichia angusta, Kluyveromyces lactis, Candida albicans, or Yarrowia lipolytica.

[0309] In certain embodiments, the host cell is an algal cell such as, Chlamydomonas (e.g., C. Reinhardtii) and Phormidium (P. sp ATCC29409).

[0310] In other embodiments, the host cell is a prokaryotic cell. Suitable prokaryotic cells include gram positive, gram negative, and gram-variable bacterial cells. The host cell may be a species of, but not limited to: Agrobacterium, Alcaligenes, Anabanaea, Anacystis, Acinetobacter, Acidothermus, Arthrobacter, Azobacter, Bacillus, Bifidobacterium, Brevibacterium, Butyribrio, Buchnera, Campstris, Camplyobacter, Clostridium, Corynebacterium, Chromatium, Coprococcus, Escherichia, Enterococcus, Enterobacter, Erwinia, Fusobacterium, Faecalibacterium, Francisella, Flavobacterium, Geobacillus, Haemophilus, Helicobacter, Klebsiella, Lactobacillus, Lactococcus, Llyobacter, Micrococcus, Microbacterium, Mesorhizobium, Methylobacterium, Mycobacterium, Neisseria, Pantoea,
Pseudomonas, Prochlorococcus, Rhodobacter, Rhodopseudomonas, Rhodopseudomonas, Roseburia, Rhodospirillum, Rhodococcus, Scenedesmus, Streptomyces, Streptococcus, Synecococcus, Saccharomonospora, Saccharopolyspora, Staphylococcus, Serratia, Salmonella, Shigella, Thermoanaerobacterium, Tropheryma, Tularenis, Temecula, Thermosynechococcus, Thermococcus, Ureaplasma, Xanthomonas, Xylella, Yersinia, and Zymomonas. In some embodiments, the host cell is Corynebacterium glutamicum.

[0311] In some embodiments, the bacterial host strain is an industrial strain. Numerous bacterial industrial strains are known and suitable in the methods and compositions described herein.

[0312] In some embodiments, the bacterial host cell is of the Agrobacterium species (e.g., A. radiobacter, A. rhizogenes, A. rubi), the Arthrobacter species (e.g., A. aurescens, A. citreus, A. globiformis, A. hydrocarboglutamicus, A. myscoren, A. nicotianae, A. paraffineus, A. protophomia, A. roseoparaffinum, A. sultureus, A. ureafaciens), the Bacillus species (e.g., B. thuringiensis, B. anthracis, B. megaterium, B. subtilis, B. lentus, B. circulars, B. pumilus, B. lautus, B. coagulans, B. brevis, B. firmus, B. alkalophilus, B. licheniformis, B. clausii, B. stearothermophilus, B. halodurans and B. amyloliquefaciens. In particular embodiments, the host cell will be an industrial Bacillus strain including but not limited to B. subtilis, B. pumilus, B. licheniformis, B. megaterium, B. clausii, B. stearothermophilus and B. amyloliquefaciens. In some embodiments, the host cell will be an industrial Clostridium species (e.g., C. acetobutyllicum, C. tetani E88, C. lituseburensce, C. saccharobutyllicum, C. perfringens, C. beijerinckii). In some embodiments, the host cell will be an industrial Corynebacterium species (e.g., C. glutamicum, C. acetoacidophilum). In some embodiments, the host cell will be an industrial Escherichia species (e.g., E. coli). In some embodiments, the host cell will be an industrial Erwinia species (e.g., E. uredovora, E. carotovora, E. ananas, E. herbicola, E. punctata, E. terreus). In some embodiments, the host cell will be an industrial Pantoea species (e.g., P. citrea, P. agglomerans). In some embodiments, the host cell will be an industrial Pseudomonas species, (e.g., P. putida, P. aeruginosa, P. mevalonii). In some embodiments, the host cell will be an industrial Streptococcus species (e.g., S. equisimiles, S. pyogenes, S. uberis). In some embodiments, the host cell will be an industrial Streptomyces species (e.g., S. ambofaciens, S. achromogenes, S. avermitilis, S. coelicolor, S. aureofaciens, S. aureus, S. fungidicus, S. griseus, S. lividans). In some embodiments, the host cell will be an industrial Zymomonas species (e.g., Z. mobilis, Z. lipolytica), and the like.
The present disclosure is also suitable for use with a variety of animal cell types, including mammalian cells, for example, human (including 293, WI38, PER.C6 and Bowes melanoma cells), mouse (including 3T3, NS0, NS1, Sp2/0), hamster (CHO, BHK), monkey (COS, FRhL, Vero), and hybridoma cell lines.

In various embodiments, strains that may be used in the practice of the disclosure including both prokaryotic and eukaryotic strains, are readily accessible to the public from a number of culture collections such as American Type Culture Collection (ATCC), Deutsche Sammlung von Mikroorganismen and Zellkulturen GmbH (DSM), Centraalbureau Voor Schimmelcultures (CBS), and Agricultural Research Service Patent Culture Collection, Northern Regional Research Center (NRRL).

In some embodiments, the methods of the present disclosure are also applicable to multicellular organisms. For example, the platform could be used for improving the performance of crops. The organisms can comprise a plurality of plants such as Gramineae, Fagaceae, Poaceae, Agrostis, Phleum, Dactylis, Sorgum, Setaria, Zea, Oryza, Triticum, Secale, Avena, Hordeum, Saccharum, Poa, Festuca, Stenotaphrum, Cynodon, Coix, Olyreae, Phareae, Compositae or Leguminosae. For example, the plants can be corn, rice, soybean, cotton, wheat, rye, oats, barley, pea, beans, lentil, peanut, yam bean, cowpeas, velvet beans, clover, alfalfa, lupine, vetch, lotus, sweet clover, wisteria, sweet pea, sorghum, millet, sunflower, canola or the like. Similarly, the organisms can include a plurality of animals such as non-human mammals, fish, insects, or the like.

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

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

In some embodiments, a first step in the genetic design methods of the present disclosure is to obtain an initial genetic diversity pool population with a plurality of sequence variations from which a new host genome may be reconstructed.
[0318] In some embodiments, a subsequent step in the genetic design methods taught herein is to use one or more of the aforementioned HTP molecular tool sets (e.g. SNP swapping or promoter swapping or transposon mutagenesis) to construct HTP genetic design libraries, which then function as drivers of the genomic engineering process, by providing libraries of particular genomic alterations for testing in a host cell.

Harnessing Diversity Pools From Existing Wild-type Strains

[0319] In some embodiments, the present disclosure teaches methods for identifying the sequence diversity present among microbes of a given wild-type population. Therefore, a diversity pool can be a given number \( n \) of wild-type microbes utilized for analysis, with the microbes’ genomes representing the “diversity pool.”

[0320] In some embodiments, the diversity pools can be the result of existing diversity present in the natural genetic variation among the wild-type microbes. This variation may result from strain variants of a given host cell or may be the result of the microbes being different species entirely. Genetic variations can include any differences in the genetic sequence of the strains, whether naturally occurring or not. In some embodiments, genetic variations can include SNPs swaps, PRO swaps, Start/Stop Codon swaps, or STOP swaps, among others.

Harnessing Diversity Pools From Existing Industrial Strain Variants

[0321] In other embodiments of the present disclosure, diversity pools are strain variants created during traditional strain improvement processes (e.g., one or more host organism strains generated via random mutation and selected for improved yields over the years). Thus, in some embodiments, the diversity pool or host organisms can comprise a collection of historical production strains.

[0322] In particular aspects, a diversity pool may be an original parent microbial strain (S1) with a “baseline” genetic sequence at a particular time point (S1Gen1) and then any number of subsequent offspring strains (S2, S3, S4, S5, etc., generalizable to S2-n) that were derived/developed from the S1 strain and that have a different genome (S2-nGen2-n), in relation to the baseline genome of S1.

[0323] For example, in some embodiments, the present disclosure teaches sequencing the microbial genomes in a diversity pool to identify the SNP’s present in each strain. In one embodiment, the strains of the diversity pool are historical microbial production strains. Thus, a
diversity pool of the present disclosure can include for example, an industrial base strain, and one or more mutated industrial strains produced via traditional strain improvement programs.

[0324] Once all SNPs in the diversity pool are identified, the present disclosure teaches methods of SNP swapping and screening methods to delineate (i.e. quantify and characterize) the effects (e.g. creation of a phenotype of interest) of SNPs individually and in groups. Thus, as aforementioned, an initial step in the taught platform can be to obtain an initial genetic diversity pool population with a plurality of sequence variations, e.g. SNPs. Then, a subsequent step in the taught platform can be to use one or more of the aforementioned HTP molecular tool sets (e.g. SNP swapping) to construct HTP genetic design libraries, which then function as drivers of the genomic engineering process, by providing libraries of particular genomic alterations for testing in a microbe.

[0325] In some embodiments, the SNP swapping methods of the present disclosure comprise the step of introducing one or more SNPs identified in a mutated strain (e.g., a strain from amongst S_{2,nGen{2,n}}) to a base strain (S_{1Gen1}) or wild-type strain.

[0326] In other embodiments, the SNP swapping methods of the present disclosure comprise the step of removing one or more SNPs identified in a mutated strain (e.g., a strain from amongst S_{2,nGen{2,n}}).

Creating Diversity Pools via Mutagenesis

[0327] In some embodiments, the mutations of interest in a given diversity pool population of cells can be artificially generated by any means for mutating strains, including mutagenic chemicals, or radiation. The term “mutagenizing” is used herein to refer to a method for inducing one or more genetic modifications in cellular nucleic acid material.

[0328] The term “genetic modification” refers to any alteration of DNA. Representative gene modifications include nucleotide insertions, deletions, substitutions, and combinations thereof, and can be as small as a single base or as large as tens of thousands of bases. Thus, the term “genetic modification” encompasses inversions of a nucleotide sequence and other chromosomal rearrangements, whereby the position or orientation of DNA comprising a region of a chromosome is altered. A chromosomal rearrangement can comprise an intrachromosomal rearrangement or an interchromosomal rearrangement.
[0329] In one embodiment, the mutagenizing methods employed in the presently claimed subject matter are substantially random such that a genetic modification can occur at any available nucleotide position within the nucleic acid material to be mutagenized. Stated another way, in one embodiment, the mutagenizing does not show a preference or increased frequency of occurrence at particular nucleotide sequences.

[0330] The methods of the disclosure can employ any mutagenic agent including, but not limited to: ultraviolet light, X-ray radiation, gamma radiation, N-ethyl-N-nitrosourea (ENU), methylnitrosourea (MNU), procarbazine (PRC), triethylene melamine (TEM), acrylamide monomer (AA), chlorambucil (CHL), melphalan (MLP), cyclophosphamide (CPP), diethyl sulfate (DES), ethyl methane sulfonate (EMS), methyl methane sulfonate (MMS), 6-mercaptopurine (6-MP), mitomycin-C (MMC), N-methyl-N′-nitro-N-nitrosoguanidine (MNNG), H.O, and urethane (UR) (See e.g., Rinchik, 1991; Marker et al., 1997; and Russell, 1990). Additional mutagenic agents are well known to persons having skill in the art, including those described in http://www.iephb.nw.ru/~spirov/hazard/mutagen_lst.html.

[0331] The term “mutagenizing” also encompasses a method for altering (e.g., by targeted mutation) or modulating a cell function, to thereby enhance a rate, quality, or extent of mutagenesis. For example, a cell can be altered or modulated to thereby be dysfunctional or deficient in DNA repair, mutagen metabolism, mutagen sensitivity, genomic stability, or combinations thereof. Thus, disruption of gene functions that normally maintain genomic stability can be used to enhance mutagenesis. Representative targets of disruption include, but are not limited to DNA ligase I (Bentley et al., 2002) and casein kinase I (U.S. Pat. No. 6,060,296).

[0332] In some embodiments, site-specific mutagenesis (e.g., primer-directed mutagenesis using a commercially available kit such as the Transformer Site Directed mutagenesis kit (Clontech)) is used to make a plurality of changes throughout a nucleic acid sequence in order to generate nucleic acid encoding a cleavage enzyme of the present disclosure.

[0333] The frequency of genetic modification upon exposure to one or more mutagenic agents can be modulated by varying dose and/or repetititon of treatment, and can be tailored for a particular application.

[0334] Thus, in some embodiments, “mutagenesis” as used herein comprises all techniques known in the art for inducing mutations, including error-prone PCR mutagenesis, oligonucleotide-directed
mutagenesis, site-directed mutagenesis, and iterative sequence recombination by any of the techniques described herein.

**Single Locus Mutations to Generate Diversity**

[0335] In some embodiments, the present disclosure teaches mutating cell populations by introducing, deleting, or replacing selected portions of genomic DNA. Thus, in some embodiments, the present disclosure teaches methods for targeting mutations to a specific locus. In other embodiments, the present disclosure teaches the use of gene editing technologies such as ZFNs, TALENS, or CRISPR, to selectively edit target DNA regions.

[0336] In other embodiments, the present disclosure teaches mutating selected DNA regions outside of the host organism, and then inserting the mutated sequence back into the host organism. For example, in some embodiments, the present disclosure teaches mutating native or synthetic promoters to produce a range of promoter variants with various expression properties (see promoter ladder *infra*). In other embodiments, the present disclosure is compatible with single gene optimization techniques, such as ProSAR (Fox *et al.* 2007. “Improving catalytic function by ProSAR-driven enzyme evolution.” Nature Biotechnology Vol 25 (3) 338-343, incorporated by reference herein).

[0337] In some embodiments, the selected regions of DNA are produced *in vitro* via gene shuffling of natural variants, or shuffling with synthetic oligos, plasmid-plasmid recombination, virus plasmid recombination, virus-virus recombination. In other embodiments, the genomic regions are produced via error-prone PCR (see e.g., Figure 1).

[0338] In some embodiments, generating mutations in selected genetic regions is accomplished by “reassembly PCR.” Briefly, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of a nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length. Each of the segments is amplified with a set of such primers. The PCR products are then “reassembled” according to assembly protocols. In brief, in an assembly protocol, the PCR products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTP’s) and appropriate buffer salts in the
absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes.

[0339] In some embodiments of the disclosure, mutated DNA regions, such as those discussed above, are enriched for mutant sequences so that the multiple mutant spectrum, i.e. possible combinations of mutations, is more efficiently sampled. In some embodiments, mutated sequences are identified via a mutS protein affinity matrix (Wagner et al., Nucleic Acids Res. 23(19):3944-3948 (1995); Su et al., Proc. Natl. Acad. Sci. (U.S.A.), 83:5057-5061(1986)) with a preferred step of amplifying the affinity-purified material in vitro prior to an assembly reaction. This amplified material is then put into an assembly or reassembly PCR reaction as described in later portions of this application.

Promoter Ladders

[0340] Promoters regulate the rate at which genes are transcribed and can influence transcription in a variety of ways. Constitutive promoters, for example, direct the transcription of their associated genes at a constant rate regardless of the internal or external cellular conditions, while regulatable promoters increase or decrease the rate at which a gene is transcribed depending on the internal and/or the external cellular conditions, e.g. growth rate, temperature, responses to specific environmental chemicals, and the like. Promoters can be isolated from their normal cellular contexts and engineered to regulate the expression of virtually any gene, enabling the effective modification of cellular growth, product yield and/or other phenotypes of interest.

[0341] In some embodiments, the present disclosure teaches methods for producing promoter ladder libraries for use in downstream genetic design methods. For example, in some embodiments, the present disclosure teaches methods of identifying one or more promoters and/or generating variants of one or more promoters within a host cell, which exhibit a range of expression strengths, or superior regulatory properties. A particular combination of these identified and/or generated promoters can be grouped together as a promoter ladder, which is explained in more detail below.

[0342] In some embodiments, the present disclosure teaches the use of promoter ladders. In some embodiments, the promoter ladders of the present disclosure comprise promoters exhibiting a continuous range of expression profiles. For example, in some embodiments, promoter ladders are created by: identifying natural, native, or wild-type promoters that exhibit a range of expression
strengths in response to a stimuli, or through constitutive expression (see e.g., Figure 12 and Figures 17-19). These identified promoters can be grouped together as a promoter ladder.

[0343] In other embodiments, the present disclosure teaches the creation of promoter ladders exhibiting a range of expression profiles across different conditions. For example, in some embodiments, the present disclosure teaches creating a ladder of promoters with expression peaks spread throughout the different stages of a fermentation (see e.g., Figure 17). In other embodiments, the present disclosure teaches creating a ladder of promoters with different expression peak dynamics in response to a specific stimulus (see e.g., Figure 18). Persons skilled in the art will recognize that the regulatory promoter ladders of the present disclosure can be representative of any one or more regulatory profiles.

[0344] In some embodiments, the promoter ladders of the present disclosure are designed to perturb gene expression in a predictable manner across a continuous range of responses. In some embodiments, the continuous nature of a promoter ladder confers strain improvement programs with additional predictive power. For example, in some embodiments, swapping promoters or termination sequences of a selected metabolic pathway can produce a host cell performance curve, which identifies the most optimum expression ratio or profile; producing a strain in which the targeted gene is no longer a limiting factor for a particular reaction or genetic cascade, while also avoiding unnecessary over expression or misexpression under inappropriate circumstances. In some embodiments, promoter ladders are created by: identifying natural, native, or wild-type promoters exhibiting the desired profiles. In other embodiments, the promoter ladders are created by mutating naturally occurring promoters to derive multiple mutated promoter sequences. Each of these mutated promoters is tested for effect on target gene expression. In some embodiments, the edited promoters are tested for expression activity across a variety of conditions, such that each promoter variant’s activity is documented/characterized/annotated and stored in a database. The resulting edited promoter variants are subsequently organized into promoter ladders arranged based on the strength of their expression (e.g., with highly expressing variants near the top, and attenuated expression near the bottom, therefore leading to the term “ladder”).

[0345] In some embodiments, the present disclosure teaches promoter ladders that are a combination of identified naturally occurring promoters and mutated variant promoters.
[0346] In some embodiments, the present disclosure teaches methods of identifying natural, native, or wild-type promoters that satisfied both of the following criteria: 1) represented a ladder of constitutive promoters; and 2) could be encoded by short DNA sequences, ideally less than 100 base pairs. In some embodiments, constitutive promoters of the present disclosure exhibit constant gene expression across two selected growth conditions (typically compared among conditions experienced during industrial cultivation). In some embodiments, the promoters of the present disclosure will consist of a ~60 base pair core promoter, and a 5’ UTR between 26- and 40 base pairs in length.

[0347] In some embodiments, one or more of the aforementioned identified naturally occurring promoter sequences are chosen for gene editing. In some embodiments, the natural promoters are edited via any of the mutation methods described supra. In other embodiments, the promoters of the present disclosure are edited by synthesizing new promoter variants with the desired sequence.


[0349] A non-exhaustive list of the promoters of the present disclosure is provided in the below Table 1. Each of the promoter sequences can be referred to as a heterologous promoter or heterologous promoter polynucleotide.

<table>
<thead>
<tr>
<th>SEQ ID No.</th>
<th>Promoter Short Name</th>
<th>Promoter Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P1</td>
<td>Pcg0007_lib_39</td>
</tr>
<tr>
<td>2</td>
<td>P2</td>
<td>Pcg0007</td>
</tr>
<tr>
<td>3</td>
<td>P3</td>
<td>Pcg1860</td>
</tr>
<tr>
<td>4</td>
<td>P4</td>
<td>Pcg0755</td>
</tr>
<tr>
<td>5</td>
<td>P5</td>
<td>Pcg0007_265</td>
</tr>
<tr>
<td>6</td>
<td>P6</td>
<td>Pcg3381</td>
</tr>
<tr>
<td>---</td>
<td>----</td>
<td>---------</td>
</tr>
<tr>
<td>7</td>
<td>P7</td>
<td>Pcg0007_119</td>
</tr>
<tr>
<td>8</td>
<td>P8</td>
<td>Pcg3121</td>
</tr>
</tbody>
</table>

[0350] In some embodiments, the promoters of the present disclosure exhibit at least 100%, 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 79%, 78%, 77%, 76%, or 75% sequence identity with a promoter from the above table 1.

**Terminator Ladders**

[0351] In some embodiments, the present disclosure teaches methods of improving genetically engineered host strains by providing one or more transcriptional termination sequences at a position 3′ to the end of the RNA encoding element. In some embodiments, the present disclosure teaches that the addition of termination sequences improves the efficiency of RNA transcription of a selected gene in the genetically engineered host. In other embodiments, the present disclosure teaches that the addition of termination sequences reduces the efficiency of RNA transcription of a selected gene in the genetically engineered host. Thus in some embodiments, the terminator ladders of the present disclosure comprises a series of terminator sequences exhibiting a range of transcription efficiencies (e.g., one weak terminator, one average terminator, and one strong promoter).

[0352] A transcriptional termination sequence may be any nucleotide sequence, which when placed transcriptionally downstream of a nucleotide sequence encoding an open reading frame, causes the end of transcription of the open reading frame. Such sequences are known in the art and may be of prokaryotic, eukaryotic or phage origin. Examples of terminator sequences include, but are not limited to, PTH-terminator, pET-T7 terminator, T3-T7 terminator, pBR322-P4 terminator, vesicular stomatitis virus terminator, rrnB-T1 terminator, rrnC terminator, TTade transcriptional terminator, and yeast-recognized termination sequences, such as Matα (α-factor) transcription terminator, native α-factor transcription termination sequence, ADR1 transcription termination sequence, ADH2 transcription termination sequence, and GAPD transcription.
termination sequence. A non-exhaustive listing of transcriptional terminator sequences may be found in the iGEM registry, which is available at: http://partsregistry.org/Terminators/Catalog.

[0353] In some embodiments, transcriptional termination sequences may be polymerase-specific or nonspecific, however, transcriptional terminators selected for use in the present embodiments should form a ‘functional combination’ with the selected promoter, meaning that the terminator sequence should be capable of terminating transcription by the type of RNA polymerase initiating at the promoter. For example, in some embodiments, the present disclosure teaches a eukaryotic RNA pol II promoter and eukaryotic RNA pol II terminators, a T7 promoter and T7 terminators, a T3 promoter and T3 terminators, a yeast-recognized promoter and yeast-recognized termination sequences, etc., would generally form a functional combination. The identity of the transcriptional termination sequences used may also be selected based on the efficiency with which transcription is terminated from a given promoter. For example, a heterologous transcriptional terminator sequence may be provided transcriptionally downstream of the RNA encoding element to achieve a termination efficiency of at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% from a given promoter.

[0354] In some embodiments, efficiency of RNA transcription from the engineered expression construct can be improved by providing nucleic acid sequence forms a secondary structure comprising two or more hairpins at a position 3' to the end of the RNA encoding element. Not wishing to be bound by a particular theory, the secondary structure destabilizes the transcription elongation complex and leads to the polymerase becoming dissociated from the DNA template, thereby minimizing unproductive transcription of non-functional sequence and increasing transcription of the desired RNA. Accordingly, a termination sequence may be provided that forms a secondary structure comprising two or more adjacent hairpins. Generally, a hairpin can be formed by a palindromic nucleotide sequence that can fold back on itself to form a paired stem region whose arms are connected by a single stranded loop. In some embodiments, the termination sequence comprises 2, 3, 4, 5, 6, 7, 8, 9, 10 or more adjacent hairpins. In some embodiments, the adjacent hairpins are separated by 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, or 15 unpaired nucleotides. In some embodiments, a hairpin stem comprises 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more base pairs in length. In certain embodiments, a hairpin stem is 12 to 30 base pairs in length. In certain embodiments, the
termination sequence comprises two or more medium-sized hairpins having stem region comprising about 9 to 25 base pairs. In some embodiments, the hairpin comprises a loop-forming region of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides. In some embodiments, the loop-forming region comprises 4-8 nucleotides. Not wishing to be bound by a particular theory, stability of the secondary structure can be correlated with termination efficiency. Hairpin stability is determined by its length, the number of mismatches or bulges it contains and the base composition of the paired region. Pairings between guanine and cytosine have three hydrogen bonds and are more stable compared to adenine-thymine pairings, which have only two. The G/C content of a hairpin-forming palindromic nucleotide sequence can be at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90% or more. In some embodiments, the G/C content of a hairpin-forming palindromic nucleotide sequence is at least 80%. In some embodiments, the termination sequence is derived from one or more transcriptional terminator sequences of prokaryotic, eukaryotic or phage origin. In some embodiments, a nucleotide sequence encoding a series of 4, 5, 6, 7, 8, 9, 10 or more adenines (A) are provided 3’ to the termination sequence.

[0355] In some embodiments, the present disclosure teaches the use of a series of tandem termination sequences. In some embodiments, the first transcriptional terminator sequence of a series of 2, 3, 4, 5, 6, 7, or more may be placed directly 3’ to the final nucleotide of the dsRNA encoding element or at a distance of at least 1-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50, 50-100, 100-150, 150-200, 200-300, 300-400, 400-500, 500-1,000 or more nucleotides 3’ to the final nucleotide of the dsRNA encoding element. The number of nucleotides between tandem transcriptional terminator sequences may be varied, for example, transcriptional terminator sequences may be separated by 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50 or more nucleotides. In some embodiments, the transcriptional terminator sequences may be selected based on their predicted secondary structure as determined by a structure prediction algorithm. Structural prediction programs are well known in the art and include, for example, CLC Main Workbench.

[0356] Persons having skill in the art will recognize that the methods of the present disclosure are compatible with any termination sequence. In some embodiments, the present disclosure teaches use of annotated Corynebacterium glutamicum terminators as disclosed in from Pfeifer-Sancar et al. 2013. “Comprehensive analysis of the Corynebacterium glutamicum transcriptome using an improved RNaseq technique” Pfeifer-Sancar et al. BMC Genomics 2013, 14:888). In other
embodiments, the present disclosure teaches use of transcriptional terminator sequences found in the iGEM registry, which is available at: http://partsregistry.org/Terminators/Catalog. A non-exhaustive listing of transcriptional terminator sequences of the present disclosure is provided in Table 1.1 below.

**Table 1.1.** Non-exhaustive list of termination sequences of the present disclosure.

<table>
<thead>
<tr>
<th>E. coli</th>
<th>Description</th>
<th>Direction</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBa_B0010</td>
<td>T1 from E. coli rnb</td>
<td>Forward</td>
<td>80</td>
</tr>
<tr>
<td>BBa_B0012</td>
<td>TE from coliphageT7</td>
<td>Forward</td>
<td>41</td>
</tr>
<tr>
<td>BBa_B0013</td>
<td>TE from coliphage T7 (+/-)</td>
<td>Forward</td>
<td>47</td>
</tr>
<tr>
<td>BBa_B0015</td>
<td>double terminator (B0010-B0012)</td>
<td>Forward</td>
<td>129</td>
</tr>
<tr>
<td>BBa_B0017</td>
<td>double terminator (B0010-B0010)</td>
<td>Forward</td>
<td>168</td>
</tr>
<tr>
<td>BBa_B0053</td>
<td>Terminator (His)</td>
<td>Forward</td>
<td>72</td>
</tr>
<tr>
<td>BBa_B0055</td>
<td>-- No description --</td>
<td></td>
<td>78</td>
</tr>
<tr>
<td>BBa_B1002</td>
<td>Terminator (artificial, small, %T~&lt;85%)</td>
<td>Forward</td>
<td>34</td>
</tr>
<tr>
<td>BBa_B1003</td>
<td>Terminator (artificial, small, %T~&lt;80%)</td>
<td>Forward</td>
<td>34</td>
</tr>
<tr>
<td>BBa_B1004</td>
<td>Terminator (artificial, small, %T~&lt;55%)</td>
<td>Forward</td>
<td>34</td>
</tr>
<tr>
<td>BBa_B1005</td>
<td>Terminator (artificial, small, %T~&lt;25%)</td>
<td>Forward</td>
<td>34</td>
</tr>
<tr>
<td>BBa_B1006</td>
<td>Terminator (artificial, large, %T~&lt;90%)</td>
<td>Forward</td>
<td>39</td>
</tr>
<tr>
<td>BBa_B1010</td>
<td>Terminator (artificial, large, %T~&lt;10)</td>
<td>Forward</td>
<td>40</td>
</tr>
<tr>
<td>BBa_111013</td>
<td>Modification of biobricks part BBA_B0015</td>
<td></td>
<td>129</td>
</tr>
<tr>
<td>BBa_151003</td>
<td>-- No description --</td>
<td></td>
<td>110</td>
</tr>
<tr>
<td>BBa_161048</td>
<td>[rnpB-T1] Terminator</td>
<td>Forward</td>
<td>113</td>
</tr>
<tr>
<td>BBa_K1392970</td>
<td>Terminator+Tet Promoter+T4 Endolysin</td>
<td></td>
<td>623</td>
</tr>
<tr>
<td>BBa_K1486001</td>
<td>Arabinose promoter + CpxR</td>
<td>Forward</td>
<td>1924</td>
</tr>
<tr>
<td>ID</td>
<td>Description</td>
<td>Direction</td>
<td>Length</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>BBa_K1486005</td>
<td>Arabinose promoter + sfGFP-CpxR [Cterm]</td>
<td>Forward</td>
<td>2668</td>
</tr>
<tr>
<td>BBa_K1486009</td>
<td>CspR &amp; Split IFP1.4 [Nterm + Nterm]</td>
<td>Forward</td>
<td>3726</td>
</tr>
<tr>
<td>BBa_K7800000</td>
<td>Terminator for Bacillus subtilis</td>
<td>Forward</td>
<td>54</td>
</tr>
<tr>
<td>BBa_K864501</td>
<td>T22, P22 late terminator</td>
<td>Forward</td>
<td>42</td>
</tr>
<tr>
<td>BBa_K864600</td>
<td>T0 (21 imm) transcriptional terminator</td>
<td>Forward</td>
<td>52</td>
</tr>
<tr>
<td>BBa_K864601</td>
<td>Lambda t1 transcriptional terminator</td>
<td>Forward</td>
<td></td>
</tr>
<tr>
<td>BBa_B0011</td>
<td>LuxICDABEG (+/-)</td>
<td>Bidirectional</td>
<td>46</td>
</tr>
<tr>
<td>BBa_B0014</td>
<td>double terminator (B0012-B0011)</td>
<td>Bidirectional</td>
<td>95</td>
</tr>
<tr>
<td>BBa_B0021</td>
<td>LuxICDABEG (+/-), reversed</td>
<td>Bidirectional</td>
<td>46</td>
</tr>
<tr>
<td>BBa_B0024</td>
<td>double terminator (B0012-B0011), reversed</td>
<td>Bidirectional</td>
<td>95</td>
</tr>
<tr>
<td>BBa_B0050</td>
<td>Terminator (pBR322, +/-)</td>
<td>Bidirectional</td>
<td>33</td>
</tr>
<tr>
<td>BBa_B0051</td>
<td>Terminator (yciA/tolA, +/-)</td>
<td>Bidirectional</td>
<td>35</td>
</tr>
<tr>
<td>BBa_B1001</td>
<td>Terminator (artificial, small, %T~90)</td>
<td>Bidirectional</td>
<td>34</td>
</tr>
<tr>
<td>BBa_B1007</td>
<td>Terminator (artificial, large, %T~80)</td>
<td>Bidirectional</td>
<td>40</td>
</tr>
<tr>
<td>BBa_B1008</td>
<td>Terminator (artificial, large, %T~70)</td>
<td>Bidirectional</td>
<td>40</td>
</tr>
<tr>
<td>BBa_B1009</td>
<td>Terminator (artificial, large, %T~40%)</td>
<td>Bidirectional</td>
<td>40</td>
</tr>
<tr>
<td>BBa_K187025</td>
<td>terminator in pAB, BioBytes plasmid</td>
<td>Bidirectional</td>
<td>60</td>
</tr>
<tr>
<td>BBa_K259006</td>
<td>GFP-Terminator</td>
<td>Bidirectional</td>
<td>823</td>
</tr>
<tr>
<td>BBa_B0020</td>
<td>Terminator (Reverse B0010)</td>
<td>Reverse</td>
<td>82</td>
</tr>
<tr>
<td>BBa_B0022</td>
<td>TE from coliphageT7, reversed</td>
<td>Reverse</td>
<td>41</td>
</tr>
<tr>
<td>BBa_B0023</td>
<td>TE from coliphage T7, reversed</td>
<td>Reverse</td>
<td>47</td>
</tr>
<tr>
<td>BBa_B0025</td>
<td>double terminator (B0015), reversed</td>
<td>Reverse</td>
<td>129</td>
</tr>
<tr>
<td>BBa_B0052</td>
<td>Terminator (rmC)</td>
<td>Forward</td>
<td>41</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Direction</td>
<td>Length</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>BBa_B0060</td>
<td>Terminator (Reverse B0050)</td>
<td>Bidirectional</td>
<td>33</td>
</tr>
<tr>
<td>BBa_B0061</td>
<td>Terminator (Reverse B0051)</td>
<td>Bidirectional</td>
<td>35</td>
</tr>
<tr>
<td>BBa_B0063</td>
<td>Terminator (Reverse B0053)</td>
<td>Reverse</td>
<td>72</td>
</tr>
<tr>
<td><strong>Yeast and other Eukaryotes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBa_J63002</td>
<td>ADH1 terminator from S. cerevisiae</td>
<td>Forward</td>
<td>225</td>
</tr>
<tr>
<td>BBa_K110012</td>
<td>STE2 terminator</td>
<td>Forward</td>
<td>123</td>
</tr>
<tr>
<td>BBa_K1462070</td>
<td>cyc1</td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>BBa_K1486025</td>
<td>ADH1 Terminator</td>
<td>Forward</td>
<td>188</td>
</tr>
<tr>
<td>BBa_K392003</td>
<td>yeast ADH1 terminator</td>
<td></td>
<td>129</td>
</tr>
<tr>
<td>BBa_K801011</td>
<td>TEF1 yeast terminator</td>
<td></td>
<td>507</td>
</tr>
<tr>
<td>BBa_K801012</td>
<td>ADH1 yeast terminator</td>
<td></td>
<td>349</td>
</tr>
<tr>
<td>BBa_Y1015</td>
<td>CycE1</td>
<td></td>
<td>252</td>
</tr>
<tr>
<td>BBa_J52016</td>
<td>eukaryotic -- derived from SV40 early poly A signal sequence</td>
<td>Forward</td>
<td>238</td>
</tr>
<tr>
<td>BBa_J63002</td>
<td>ADH1 terminator from S. cerevisiae</td>
<td>Forward</td>
<td>225</td>
</tr>
<tr>
<td>BBa_K110012</td>
<td>STE2 terminator</td>
<td>Forward</td>
<td>123</td>
</tr>
<tr>
<td>BBa_K1159307</td>
<td>35S Terminator of Cauliflower Mosaic Virus (CaMV)</td>
<td></td>
<td>217</td>
</tr>
<tr>
<td>BBa_K1462070</td>
<td>cyc1</td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>BBa_K1484215</td>
<td>nopaline synthase terminator</td>
<td></td>
<td>293</td>
</tr>
<tr>
<td>BBa_K1486025</td>
<td>ADH1 Terminator</td>
<td>Forward</td>
<td>188</td>
</tr>
<tr>
<td>BBa_K392003</td>
<td>yeast ADH1 terminator</td>
<td></td>
<td>129</td>
</tr>
<tr>
<td>BBa_K404108</td>
<td>hGH terminator</td>
<td></td>
<td>481</td>
</tr>
<tr>
<td>BBa_K404116</td>
<td>hGH_[AAV2]-right-ITR</td>
<td></td>
<td>632</td>
</tr>
<tr>
<td>BBa_K678012</td>
<td>SV40 poly A, terminator for mammalian cells</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>BBa_K678018</td>
<td>hGH poly A, terminator for mammalian cells</td>
<td>635</td>
<td></td>
</tr>
<tr>
<td>BBa_K678019</td>
<td>BGH poly A, mammalian terminator</td>
<td>233</td>
<td></td>
</tr>
<tr>
<td>BBa_K678036</td>
<td>trpC terminator for Aspergillus nidulans</td>
<td>759</td>
<td></td>
</tr>
<tr>
<td>BBa_K678037</td>
<td>T1-motni, terminator for Aspergillus niger</td>
<td>1006</td>
<td></td>
</tr>
<tr>
<td>BBa_K678038</td>
<td>T2-motni, terminator for Aspergillus niger</td>
<td>990</td>
<td></td>
</tr>
<tr>
<td>BBa_K678039</td>
<td>T3-motni, terminator for Aspergillus niger</td>
<td>889</td>
<td></td>
</tr>
<tr>
<td>BBa_K801011</td>
<td>TEF1 yeast terminator</td>
<td>507</td>
<td></td>
</tr>
<tr>
<td>BBa_K801012</td>
<td>ADH1 yeast terminator</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>BBa_Y1015</td>
<td>CycE1</td>
<td>252</td>
<td></td>
</tr>
</tbody>
</table>

*Corynebacterium*

<table>
<thead>
<tr>
<th>Terminator</th>
<th>Terminat or Start</th>
<th>Terminator End</th>
<th>strand</th>
<th>Transcript End</th>
<th>DNA Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg0001</td>
<td>1628</td>
<td>1647</td>
<td>+</td>
<td>loop</td>
<td>SEQ ID NO: 9</td>
</tr>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg0007</td>
<td>7504</td>
<td>7529</td>
<td>+</td>
<td>stem 1</td>
<td>SEQ ID NO: 10</td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg0371</td>
<td>322229</td>
<td>322252</td>
<td>+</td>
<td>stem 1</td>
<td>SEQ ID NO: 11</td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg0480</td>
<td>421697</td>
<td>421720</td>
<td>-</td>
<td>stem 1</td>
<td>SEQ ID NO: 12</td>
</tr>
<tr>
<td>T4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg0494</td>
<td>436587</td>
<td>436608</td>
<td>+</td>
<td>loop</td>
<td>SEQ ID NO: 13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis-driven Diversity Pools and Hill Climbing

[0357] The HTP genomic engineering methods of the present disclosure do not require prior genetic knowledge in order to achieve significant gains in host cell performance. Indeed, the disclosure teaches methods of generating diversity pools via several functionally agnostic approaches, including random mutagenesis, and identification of genetic diversity among pre-existing host cell variants (e.g., such as the comparison between a wild type host cell and an industrial variant).

[0358] In some embodiments however, the disclosure also teaches hypothesis-driven methods of designing genetic diversity mutations that will be used for downstream HTP engineering. That is, in some embodiments, the present disclosure teaches the directed design of selected mutations. In some embodiments, the directed mutations are incorporated into the engineering libraries of the present disclosure (e.g., SNP swap, PRO swap, or STOP swap).

[0359] In some embodiments, the present disclosure teaches the creation of directed mutations based on gene annotation, hypothesized (or confirmed) gene function, or location within a genome. The diversity pools of the present disclosure may include mutations in genes hypothesized to be involved in a specific metabolic or genetic pathway associated in the literature with increased performance of a host cell. In other embodiments, the diversity pool of the present disclosure may also include mutations to genes present in an operon associated with improved host performance. In yet other embodiments, the diversity pool of the present disclosure may also include mutations to genes based on algorithmic predicted function, or other gene annotation.
In some embodiments, the present disclosure teaches a “shell” based approach for prioritizing the targets of hypothesis-driven mutations. The shell metaphor for target prioritization is based on the hypothesis that only a handful of primary genes are responsible for most of a particular aspect of a host cell’s performance (e.g., production of a single biomolecule). These primary genes are located at the core of the shell, followed by secondary effect genes in the second layer, tertiary effects in the third shell, and… etc. For example, in one embodiment the core of the shell might comprise genes encoding critical biosynthetic enzymes within a selected metabolic pathway (e.g., production of citric acid). Genes located on the second shell might comprise genes encoding for other enzymes within the biosynthetic pathway responsible for product diversion or feedback signaling. Third tier genes under this illustrative metaphor would likely comprise regulatory genes responsible for modulating expression of the biosynthetic pathway, or for regulating general carbon flux within the host cell.

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

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

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

Cell Culture and Fermentation

[0364] Cells of the present disclosure can be cultured in conventional nutrient media modified as appropriate for any desired biosynthetic reactions or selections. In some embodiments, the present disclosure teaches culture in inducing media for activating promoters. In some embodiments, the present disclosure teaches media with selection agents, including selection agents of transformants (e.g., antibiotics), or selection of organisms suited to grow under inhibiting conditions (e.g., high ethanol conditions). In some embodiments, the present disclosure teaches growing cell cultures in media optimized for cell growth. In other embodiments, the present disclosure teaches growing cell cultures in media optimized for product yield. In some embodiments, the present disclosure teaches growing cultures in media capable of inducing cell growth and also contains the necessary precursors for final product production (e.g., high levels of sugars for ethanol production).


[0366] The culture medium to be used must in a suitable manner satisfy the demands of the respective strains. Descriptions of culture media for various microorganisms are present in the “Manual of Methods for General Bacteriology” of the American Society for Bacteriology (Washington D.C., USA, 1981).

[0367] The present disclosure furthermore provides a process for fermentative preparation of a product of interest, comprising the steps of: a) culturing a microorganism according to the present disclosure in a suitable medium, resulting in a fermentation broth; and b) concentrating the product of interest in the fermentation broth of a) and/or in the cells of the microorganism.

[0368] In some embodiments, the present disclosure teaches that the microorganisms produced may be cultured continuously—as described, for example, in WO 05/021772—or discontinuously in a batch process (batch cultivation) or in a fed-batch or repeated fed-batch process for the purpose of producing the desired organic-chemical compound. A summary of a general nature about known cultivation methods is available in the textbook by Chmiel (Bioprozesstechnik. 1: Einführung in die Bioverfahrenstechnik (Gustav Fischer Verlag, Stuttgart, 1991)) or in the textbook by Storhas (Bioreaktoren and periphere Einrichtungen (Vieweg Verlag, Braunschweig/Wiesbaden, 1994)).

[0369] In some embodiments, the cells of the present disclosure are grown under batch or continuous fermentations conditions.

[0370] Classical batch fermentation is a closed system, wherein the compositions of the medium is set at the beginning of the fermentation and is not subject to artificial alternations during the fermentation. A variation of the batch system is a fed-batch fermentation which also finds use in the present disclosure. In this variation, the substrate is added in increments as the fermentation
progresses. Fed-batch systems are useful when catabolite repression is likely to inhibit the metabolism of the cells and where it is desirable to have limited amounts of substrate in the medium. Batch and fed-batch fermentations are common and well known in the art.

[0371] Continuous fermentation is a system where a defined fermentation medium is added continuously to a bioreactor and an equal amount of conditioned medium is removed simultaneously for processing and harvesting of desired biomolecule products of interest. In some embodiments, continuous fermentation generally maintains the cultures at a constant high density where cells are primarily in log phase growth. In some embodiments, continuous fermentation generally maintains the cultures at a stationary or late log/stationary, phase growth. Continuous fermentation systems strive to maintain steady state growth conditions.

[0372] Methods for modulating nutrients and growth factors for continuous fermentation processes as well as techniques for maximizing the rate of product formation are well known in the art of industrial microbiology.

[0373] For example, a non-limiting list of carbon sources for the cultures of the present disclosure include, sugars and carbohydrates such as, for example, glucose, sucrose, lactose, fructose, maltose, molasses, sucrose-containing solutions from sugar beet or sugar cane processing, starch, starch hydrolysate, and cellulose; oils and fats such as, for example, soybean oil, sunflower oil, groundnut oil and coconut fat; fatty acids such as, for example, palmitic acid, stearic acid, and linoleic acid; alcohols such as, for example, glycerol, methanol, and ethanol; and organic acids such as, for example, acetic acid or lactic acid.

[0374] A non-limiting list of the nitrogen sources for the cultures of the present disclosure include, organic nitrogen-containing compounds such as peptones, yeast extract, meat extract, malt extract, corn steep liquor, soybean flour, and urea; or inorganic compounds such as ammonium sulfate, ammonium chloride, ammonium phosphate, ammonium carbonate, and ammonium nitrate. The nitrogen sources can be used individually or as a mixture.

[0375] A non-limiting list of the possible phosphorus sources for the cultures of the present disclosure include, phosphoric acid, potassium dihydrogen phosphate or dipotassium hydrogen phosphate or the corresponding sodium-containing salts.
[0376] The culture medium may additionally comprise salts, for example in the form of chlorides or sulfates of metals such as, for example, sodium, potassium, magnesium, calcium and iron, such as, for example, magnesium sulfate or iron sulfate, which are necessary for growth.

[0377] Finally, essential growth factors such as amino acids, for example homoserine and vitamins, for example thiamine, biotin or pantothenic acid, may be employed in addition to the abovementioned substances.

[0378] In some embodiments, the pH of the culture can be controlled by any acid or base, or buffer salt, including, but not limited to sodium hydroxide, potassium hydroxide, ammonia, or aqueous ammonia; or acidic compounds such as phosphoric acid or sulfuric acid in a suitable manner. In some embodiments, the pH is generally adjusted to a value of from 6.0 to 8.5, preferably 6.5 to 8.

[0379] In some embodiments, the cultures of the present disclosure may include an anti-foaming agent such as, for example, fatty acid polyglycol esters. In some embodiments the cultures of the present disclosure are modified to stabilize the plasmids of the cultures by adding suitable selective substances such as, for example, antibiotics.

[0380] In some embodiments, the culture is carried out under aerobic conditions. In order to maintain these conditions, oxygen or oxygen-containing gas mixtures such as, for example, air are introduced into the culture. It is likewise possible to use liquids enriched with hydrogen peroxide. The fermentation is carried out, where appropriate, at elevated pressure, for example at an elevated pressure of from 0.03 to 0.2 MPa. The temperature of the culture is normally from 20°C to 45°C and preferably from 25°C to 40°C, particularly preferably from 30°C to 37°C. In batch or fed-batch processes, the cultivation is preferably continued until an amount of the desired product of interest (e.g. an organic-chemical compound) sufficient for being recovered has formed. This aim can normally be achieved within 10 hours to 160 hours. In continuous processes, longer cultivation times are possible. The activity of the microorganisms results in a concentration (accumulation) of the product of interest in the fermentation medium and/or in the cells of the microorganisms.

[0381] In some embodiments, the culture is carried out under anaerobic conditions.

**Screening**
[0382] In some embodiments, the present disclosure teaches high-throughput initial screenings. In other embodiments, the present disclosure also teaches robust tank-based validations of performance data (see Figure 4B).

[0383] In some embodiments, the high-throughput screening process is designed to predict performance of strains in bioreactors. As previously described, culture conditions are selected to be suitable for the organism and reflective of bioreactor conditions. Individual colonies are picked and transferred into 96 well plates and incubated for a suitable amount of time. Cells are subsequently transferred to new 96 well plates for additional seed cultures, or to production cultures. Cultures are incubated for varying lengths of time, where multiple measurements may be made. These may include measurements of product, biomass or other characteristics that predict performance of strains in bioreactors. High-throughput culture results are used to predict bioreactor performance.

[0384] In some embodiments, the tank-based performance validation is used to confirm performance of strains isolated by high throughput screening. Fermentation processes/conditions are obtained from client sites. Candidate strains are screened using bench scale fermentation reactors (e.g., reactors disclosed in Table 3 of the present disclosure) for relevant strain performance characteristics such as productivity or yield.

**Product Recovery and Quantification**

[0385] Methods for screening for the production of products of interest are known to those of skill in the art and are discussed throughout the present specification. Such methods may be employed when screening the strains of the disclosure.

[0386] In some embodiments, the present disclosure teaches methods of improving strains designed to produce non-secreted intracellular products. For example, the present disclosure teaches methods of improving the robustness, yield, efficiency, or overall desirability of cell cultures producing intracellular enzymes, oils, pharmaceuticals, or other valuable small molecules or peptides. The recovery or isolation of non-secreted intracellular products can be achieved by lysis and recovery techniques that are well known in the art, including those described herein.

[0387] For example, in some embodiments, cells of the present disclosure can be harvested by centrifugation, filtration, settling, or other method. Harvested cells are then disrupted by any
convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, or other methods, which are well known to those skilled in the art.

[0388] The resulting product of interest, e.g. a polypeptide, may be recovered/isolated and optionally purified by any of a number of methods known in the art. For example, a product polypeptide may be isolated from the nutrient medium by conventional procedures including, but not limited to: centrifugation, filtration, extraction, spray-drying, evaporation, chromatography (e.g., ion exchange, affinity, hydrophobic interaction, chromatofocusing, and size exclusion), or precipitation. Finally, high performance liquid chromatography (HPLC) can be employed in the final purification steps. (See for example Purification of intracellular protein as described in Parry et al., 2001, Biochem. J. 353:117, and Hong et al., 2007, Appl. Microbiol. Biotechnol. 73:1331, both incorporated herein by reference).


[0390] In some embodiments, the present disclosure teaches the methods of improving strains designed to produce secreted products. For example, the present disclosure teaches methods of improving the robustness, yield, efficiency, or overall desirability of cell cultures producing valuable small molecules or peptides.

[0391] In some embodiments, immunological methods may be used to detect and/or purify secreted or non-secreted products produced by the cells of the present disclosure. In one example approach, antibody raised against a product molecule (e.g., against an insulin polypeptide or an immunogenic fragment thereof) using conventional methods is immobilized on beads, mixed with cell culture media under conditions in which the endoglucanase is bound, and precipitated. In some
embodiments, the present disclosure teaches the use of enzyme-linked immunosorbent assays (ELISA).

[0392] In other related embodiments, immunochromatography is used, as disclosed in U.S. Pat. No. 5,591,645, U.S. Pat. No. 4,855,240, U.S. Pat. No. 4,435,504, U.S. Pat. No. 4,980,298, and Se-Hwan Paek, et al., “Development of rapid One-Step Immunochromatographic assay, Methods”, 22, 53-60, 2000), each of which are incorporated by reference herein. A general immunochromatography detects a specimen by using two antibodies. A first antibody exists in a test solution or at a portion at an end of a test piece in an approximately rectangular shape made from a porous membrane, where the test solution is dropped. This antibody is labeled with latex particles or gold colloidal particles (this antibody will be called as a labeled antibody hereinafter). When the dropped test solution includes a specimen to be detected, the labeled antibody recognizes the specimen so as to be bonded with the specimen. A complex of the specimen and labeled antibody flows by capillarity toward an absorber, which is made from a filter paper and attached to an end opposite to the end having included the labeled antibody. During the flow, the complex of the specimen and labeled antibody is recognized and caught by a second antibody (it will be called as a tapping antibody hereinafter) existing at the middle of the porous membrane and, as a result of this, the complex appears at a detection part on the porous membrane as a visible signal and is detected.

[0393] In some embodiments, the screening methods of the present disclosure are based on photometric detection techniques (absorption, fluorescence). For example, in some embodiments, detection may be based on the presence of a fluorophore detector such as GFP bound to an antibody. In other embodiments, the photometric detection may be based on the accumulation on the desired product from the cell culture. In some embodiments, the product may be detectable via UV of the culture or extracts from the culture.

[0394] Persons having skill in the art will recognize that the methods of the present disclosure are compatible with host cells producing any desirable biomolecule product of interest. Table 2 below presents a non-limiting list of the product categories, biomolecules, and host cells, included within the scope of the present disclosure. These examples are provided for illustrative purposes, and are not meant to limit the applicability of the presently disclosed technology in any way.

Table 2. – A non-limiting list of the host cells and products of interest of the present disclosure.
<table>
<thead>
<tr>
<th>Product category</th>
<th>Products</th>
<th>Host category</th>
<th>Hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acids</td>
<td>Lysine</td>
<td>Bacteria</td>
<td><em>Corynebacterium glutamicum</em></td>
</tr>
<tr>
<td>Amino acids</td>
<td>Methionine</td>
<td>Bacteria</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>Amino acids</td>
<td>MSG</td>
<td>Bacteria</td>
<td><em>Corynebacterium glutamicum</em></td>
</tr>
<tr>
<td>Amino acids</td>
<td>Threonine</td>
<td>Bacteria</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>Amino acids</td>
<td>Threonine</td>
<td>Bacteria</td>
<td><em>Corynebacterium glutamicum</em></td>
</tr>
<tr>
<td>Amino acids</td>
<td>Tryptophan</td>
<td>Bacteria</td>
<td><em>Corynebacterium glutamicum</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Filamentous fungi</td>
<td><em>Trichoderma reesei</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Fungi</td>
<td><em>Myceliopathora thermophila (C1)</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Filamentous fungi</td>
<td><em>Aspergillus oryzae</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Filamentous fungi</td>
<td><em>Aspergillus niger</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Bacteria</td>
<td><em>Bacillus subtilis</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Bacteria</td>
<td><em>Bacillus licheniformis</em></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Enzymes (11)</td>
<td>Bacteria</td>
<td><em>Bacillus clausii</em></td>
</tr>
<tr>
<td>Flavor &amp;</td>
<td>Agarwood</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td><strong>Product category</strong></td>
<td><strong>Products</strong></td>
<td><strong>Host category</strong></td>
<td><strong>Hosts</strong></td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------</td>
<td>------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Fragrance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Ambrox</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Nootkatone</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Patchouli oil</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Saffron</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Sandalwood oil</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Valencene</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Flavor &amp; Fragrance</td>
<td>Vanillin</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Food</td>
<td>CoQ10/Ubiqunol</td>
<td>Yeast</td>
<td><em>Schizosaccharomyces pombe</em></td>
</tr>
<tr>
<td>Food</td>
<td>Omega 3 fatty acids</td>
<td>Microalgae</td>
<td><em>Schizochytrium</em></td>
</tr>
<tr>
<td>Product category</td>
<td>Products</td>
<td>Host category</td>
<td>Hosts</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Food</td>
<td>Omega 6 fatty acids</td>
<td>Microalgae</td>
<td><em>Schizochytrium</em></td>
</tr>
<tr>
<td>Food</td>
<td>Vitamin B12</td>
<td>Bacteria</td>
<td><em>Propionibacterium freudenreichii</em></td>
</tr>
<tr>
<td>Food</td>
<td>Vitamin B2</td>
<td>Filamentous fungi</td>
<td><em>Ashbya gossypii</em></td>
</tr>
<tr>
<td>Food</td>
<td>Vitamin B2</td>
<td>Bacteria</td>
<td><em>Bacillus subtilis</em></td>
</tr>
<tr>
<td>Food</td>
<td>Erythritol</td>
<td>Yeast-like fungi</td>
<td><em>Torula coraline</em></td>
</tr>
<tr>
<td>Food</td>
<td>Erythritol</td>
<td>Yeast-like fungi</td>
<td><em>Pseudozyma tsukubaensis</em></td>
</tr>
<tr>
<td>Food</td>
<td>Erythritol</td>
<td>Yeast-like fungi</td>
<td><em>Moniliella pollinis</em></td>
</tr>
<tr>
<td>Food</td>
<td>Steviol glycosides</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>Hydrocolloids</td>
<td>Diutan gum</td>
<td>Bacteria</td>
<td><em>Sphingomonas sp</em></td>
</tr>
<tr>
<td>Hydrocolloids</td>
<td>Gellan gum</td>
<td>Bacteria</td>
<td><em>Sphingomonas elodea</em></td>
</tr>
<tr>
<td>Hydrocolloids</td>
<td>Xanthan gum</td>
<td>Bacteria</td>
<td><em>Xanthomonas campestris</em></td>
</tr>
<tr>
<td>Intermediates</td>
<td>1,3-PDO</td>
<td>Bacteria</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>Intermediates</td>
<td>1,4-BDO</td>
<td>Bacteria</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>Intermediates</td>
<td>Butadiene</td>
<td>Bacteria</td>
<td><em>Cupriavidus necator</em></td>
</tr>
<tr>
<td>Intermediates</td>
<td>n-butanol</td>
<td>Bacteria (obligate anaerobe)</td>
<td><em>Clostridium acetobutylicum</em></td>
</tr>
<tr>
<td>Product category</td>
<td>Products</td>
<td>Host category</td>
<td>Hosts</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------</td>
<td>------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Organic acids</td>
<td>Citric acid</td>
<td>Filamentous fungi</td>
<td><em>Aspergillus niger</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>Citric acid</td>
<td>Yeast</td>
<td><em>Pichia guilliermondii</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>Gluconic acid</td>
<td>Filamentous fungi</td>
<td><em>Aspergillus niger</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>Itaconic acid</td>
<td>Filamentous fungi</td>
<td><em>Aspergillus terreus</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>Lactic acid</td>
<td>Bacteria</td>
<td><em>Lactobacillus</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>Lactic acid</td>
<td>Bacteria</td>
<td><em>Geobacillus thermoglucosidasius</em></td>
</tr>
<tr>
<td>Organic acids</td>
<td>LCDAs - DDDA</td>
<td>Yeast</td>
<td><em>Candida</em></td>
</tr>
<tr>
<td>Polyketides/Ag</td>
<td>Spinosad</td>
<td>Bacteria</td>
<td><em>Saccharopolyspora spinosa</em></td>
</tr>
<tr>
<td>Polyketides/Ag</td>
<td>Spinetoram</td>
<td>Bacteria</td>
<td><em>Saccharopolyspora spinosa</em></td>
</tr>
</tbody>
</table>

**Selection Criteria and Goals**

[0395] The selection criteria applied to the methods of the present disclosure will vary with the specific goals of the strain improvement program. The present disclosure may be adapted to meet any program goals. For example, in some embodiments, the program goal may be to maximize single batch yields of reactions with no immediate time limits. In other embodiments, the program goal may be to rebalance biosynthetic yields to produce a specific product, or to produce a particular ratio of products. In other embodiments, the program goal may be to modify the chemical structure of a product, such as lengthening the carbon chain of a polymer. In some embodiments, the program goal may be to improve performance characteristics such as yield, titer, productivity, by-product elimination, tolerance to process excursions, optimal growth temperature and growth rate. In some embodiments, the program goal is improved host performance as
measured by volumetric productivity, specific productivity, yield or titre, of a product of interest produced by a microbe.

[0396] In other embodiments, the program goal may be to optimize synthesis efficiency of a commercial strain in terms of final product yield per quantity of inputs (e.g., total amount of ethanol produced per pound of sucrose). In other embodiments, the program goal may be to optimize synthesis speed, as measured for example in terms of batch completion rates, or yield rates in continuous culturing systems. In other embodiments, the program goal may be to increase strain resistance to a particular phage, or otherwise increase strain vigor/robustness under culture conditions.

[0397] In some embodiments, strain improvement projects may be subject to more than one goal. In some embodiments, the goal of the strain project may hinge on quality, reliability, or overall profitability. In some embodiments, the present disclosure teaches methods of associated selected mutations or groups of mutations with one or more of the strain properties described above.

[0398] Persons having ordinary skill in the art will recognize how to tailor strain selection criteria to meet the particular project goal. For example, selections of a strain’s single batch max yield at reaction saturation may be appropriate for identifying strains with high single batch yields. Selection based on consistency in yield across a range of temperatures and conditions may be appropriate for identifying strains with increased robustness and reliability.

[0399] In some embodiments, the selection criteria for the initial high-throughput phase and the tank-based validation will be identical. In other embodiments, tank-based selection may operate under additional and/or different selection criteria. For example, in some embodiments, high-throughput strain selection might be based on single batch reaction completion yields, while tank-based selection may be expanded to include selections based on yields for reaction speed.

**Sequencing**

[0400] In some embodiments, the present disclosure teaches whole-genome sequencing of the organisms described herein. In other embodiments, the present disclosure also teaches sequencing of plasmids, PCR products, and other oligos as quality controls to the methods of the present disclosure. Sequencing methods for large and small projects are well known to those in the art.
[0401] In some embodiments, any high-throughput technique for sequencing nucleic acids can be used in the methods of the disclosure. In some embodiments, the present disclosure teaches whole genome sequencing. In other embodiments, the present disclosure teaches amplicon sequencing ultra deep sequencing to identify genetic variations. In some embodiments, the present disclosure also teaches novel methods for library preparation, including tagmentation (see WO/2016/073690). DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary; sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing; 454 sequencing; allele specific hybridization to a library of labeled oligonucleotide probes; sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation; real time monitoring of the incorporation of labeled nucleotides during a polymerization step; polony sequencing; and SOLiD sequencing.

[0402] In one aspect of the disclosure, high-throughput methods of sequencing are employed that comprise a step of spatially isolating individual molecules on a solid surface where they are sequenced in parallel. Such solid surfaces may include nonporous surfaces (such as in Solexa sequencing, e.g. Bentley et al, Nature, 456: 53-59 (2008) or Complete Genomics sequencing, e.g. Drmanac et al, Science, 327: 78-81 (2010)), arrays of wells, which may include bead- or particle-bound templates (such as with 454, e.g. Margulies et al, Nature, 437: 376-380 (2005) or Ion Torrent sequencing, U.S. patent publication 2010/0137143 or 2010/0304982), micromachined membranes (such as with SMRT sequencing, e.g. Eid et al, Science, 323: 133-138 (2009)), or bead arrays (as with SOLiD sequencing or polony sequencing, e.g. Kim et al, Science, 316: 1481-1414 (2007)).

[0403] In another embodiment, the methods of the present disclosure comprise amplifying the isolated molecules either before or after they are spatially isolated on a solid surface. Prior amplification may comprise emulsion-based amplification, such as emulsion PCR, or rolling circle amplification. Also taught is Solexa-based sequencing where individual template molecules are spatially isolated on a solid surface, after which they are amplified in parallel by bridge PCR to form separate clonal populations, or clusters, and then sequenced, as described in Bentley et al (cited above) and in manufacturer's instructions (e.g. TruSeq™ Sample Preparation Kit and Data Sheet, Illumina, Inc., San Diego, Calif., 2010); and further in the following references: U.S. Pat. Nos. 6,090,592; 6,300,070; 7,115,400; and EP0972081B1; which are incorporated by reference.
[0404] In one embodiment, individual molecules disposed and amplified on a solid surface form clusters in a density of at least $10^5$ clusters per cm$^2$; or in a density of at least $5 \times 10^5$ per cm$^2$; or in a density of at least $10^6$ clusters per cm$^2$. In one embodiment, sequencing chemistries are employed having relatively high error rates. In such embodiments, the average quality scores produced by such chemistries are monotonically declining functions of sequence read lengths. In one embodiment, such decline corresponds to 0.5 percent of sequence reads have at least one error in positions 1-75; 1 percent of sequence reads have at least one error in positions 76-100; and 2 percent of sequence reads have at least one error in positions 101-125.

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

[0405] In some embodiments, the present disclosure teaches methods of predicting the effects of particular genetic alterations being incorporated into a given host strain. In further aspects, the disclosure provides methods for generating proposed genetic alterations that should be incorporated into a given host strain, in order for the host to possess a particular phenotypic trait or strain parameter. In given aspects, the disclosure provides predictive models that can be utilized to design novel host strains.

[0406] In some embodiments, the present disclosure teaches methods of analyzing the performance results of each round of screening and methods for generating new proposed genome-wide sequence modifications predicted to enhance strain performance in the following round of screening.

[0407] In some embodiments, the present disclosure teaches that the system generates proposed sequence modifications to host strains based on previous screening results. In some embodiments, the recommendations of the present system are based on the results from the immediately preceding screening. In other embodiments, the recommendations of the present system are based on the cumulative results of one or more of the preceding screenings.

[0408] In some embodiments, the recommendations of the present system are based on previously developed HTP genetic design libraries. For example, in some embodiments, the present system is designed to save results from previous screenings, and apply those results to a different project, in the same or different host organisms.
[0409] In other embodiments, the recommendations of the present system are based on scientific insights. For example, in some embodiments, the recommendations are based on known properties of genes (from sources such as annotated gene databases and the relevant literature), codon optimization, transcriptional slippage, uORFs, or other hypothesis driven sequence and host optimizations.

[0410] In some embodiments, the proposed sequence modifications to a host strain recommended by the system, or predictive model, are carried out by the utilization of one or more of the disclosed molecular tools sets comprising: (1) Promoter swaps, (2) SNP swaps, (3) Start/Stop codon exchanges, (4) Sequence optimization, (5) Stop swaps, (6) transposon mutagenesis, and (7) Epistasis mapping.

[0411] The HTP genetic engineering platform described herein is agnostic with respect to any particular microbe or phenotypic trait (e.g. production of a particular compound). That is, the platform and methods taught herein can be utilized with any host cell to engineer the host cell to have any desired phenotypic trait. Furthermore, the lessons learned from a given HTP genetic engineering process used to create one novel host cell, can be applied to any number of other host cells, as a result of the storage, characterization, and analysis of a myriad of process parameters that occurs during the taught methods.

[0412] As alluded to in the epistatic mapping section, it is possible to estimate the performance (a.k.a. score) of a hypothetical strain obtained by consolidating a collection of mutations from a HTP genetic design library into a particular background via some preferred predictive model. Given such a predictive model, it is possible to score and rank all hypothetical strains accessible to the mutation library via combinatorial consolidation. The below section outlines particular models utilized in the present HTP platform.

**Predictive Strain Design**

[0413] Described herein is an approach for predictive strain design, including: methods of describing genetic changes and strain performance, predicting strain performance based on the composition of changes in the strain, recommending candidate designs with high predicted performance, and filtering predictions to optimize for second-order considerations, e.g. similarity to existing strains, epistasis, or confidence in predictions.
Inputs to Strain Design Model

[0414] In one embodiment, for the sake of ease of illustration, input data may comprise two components: (1) sets of genetic changes and (2) relative strain performance. Those skilled in the art will recognize that this model can be readily extended to consider a wide variety of inputs, while keeping in mind the countervailing consideration of overfitting. In addition to genetic changes, some of the input parameters (independent variables) that can be adjusted are cell types (genus, species, strain, phylogenetic characterization, etc.) and process parameters (e.g., environmental conditions, handling equipment, modification techniques, etc.) under which fermentation is conducted with the cells.

[0415] The sets of genetic changes can come from the previously discussed collections of genetic perturbations termed HTP genetic design libraries. The relative strain performance can be assessed based upon any given parameter or phenotypic trait of interest (e.g. production of a compound, small molecule, or product of interest).

[0416] Cell types can be specified in general categories such as prokaryotic and eukaryotic systems, genus, species, strain, tissue cultures (vs. disperse cells), etc. Process parameters that can be adjusted include temperature, pressure, reactor configuration, and medium composition. Examples of reactor configuration include the volume of the reactor, whether the process is a batch or continuous, and, if continuous, the volumetric flow rate, etc. One can also specify the support structure, if any, on which the cells reside. Examples of medium composition include the concentrations of electrolytes, nutrients, waste products, acids, pH, and the like.

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

[0417] To create a predictive strain design model, genetic changes in strains of the same microbial species are first selected. The history of each genetic change is also provided (e.g., showing the most recent modification in this strain lineage – “last change”). Thus, comparing this strain’s performance to the performance of its parent represents a data point concerning the performance of the “last change” mutation.

Built Strain Performance Assessment
[0418] The goal of the taught model is to predict strain performance based on the composition of genetic changes introduced to the strain. To construct a standard for comparison, strain performance is computed relative to a common reference strain, by first calculating the median performance per strain, per assay plate. Relative performance is then computed as the difference in average performance between an engineered strain and the common reference strain within the same plate. Restricting the calculations to within-plate comparisons ensures that the samples under consideration all received the same experimental conditions.

[0419] Figure 10 shows an example in which the distribution of relative strain performances for the input data is under consideration. This was done in Corynebacterium. A relative performance of zero indicates that the engineered strain performed equally well to the in-plate base or “reference” strain. Of interest is the ability of the predictive model to identify the strains that are likely to perform significantly above zero. Further, and more generally, of interest is whether any given strain outperforms its parent by some criteria. In practice, the criteria can be a product titer meeting or exceeding some threshold above the parent level, though having a statistically significant difference from the parent in the desired direction could also be used instead or in addition. The role of the base or “reference” strain is simply to serve as an added normalization factor for making comparisons within or between plates.

[0420] A concept to keep in mind is that of differences between: parent strain and reference strain. The parent strain is the background that was used for a current round of mutagenesis. The reference strain is a control strain run in every plate to facilitate comparisons, especially between plates, and is typically the “base strain” as referenced above. But since the base strain (e.g., the wild-type or industrial strain being used to benchmark overall performance) is not necessarily a “base” in the sense of being a mutagenesis target in a given round of strain improvement, a more descriptive term is “reference strain.”

[0421] In summary, a base/reference strain is used to benchmark the performance of built strains, generally, while the parent strain is used to benchmark the performance of a specific genetic change in the relevant genetic background.

**Ranking the Performance of Built Strains with Linear Regression**

[0422] The goal of the disclosed model is to rank the performance of built strains, by describing relative strain performance, as a function of the composition of genetic changes introduced into
the built strains. As discussed throughout the disclosure, the various HTP genetic design libraries provide the repertoire of possible genetic changes (e.g., genetic perturbations/alterations) that are introduced into the engineered strains. Linear regression is the basis for the currently described exemplary predictive model.

[0423] Genetic changes and their effect on relative performance is then input for regression-based modeling. The strain performances are ranked relative to a common base strain, as a function of the composition of the genetic changes contained in the strain.

**Linear Regression to Characterize Built Strains**

[0424] Linear regression is an attractive method for the described HTP genomic engineering platform, because of the ease of implementation and interpretation. The resulting regression coefficients can be interpreted as the average increase or decrease in relative strain performance attributable to the presence of each genetic change.

[0425] For example, in some embodiments, this technique allows one to conclude that changing the original promoter to another promoter improves relative strain performance by approximately 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more units on average and is thus a potentially highly desirable change, in the absence of any negative epistatic interactions (note: the input is a unit-less normalized value).

[0426] The taught method therefore uses linear regression models to describe/characterize and rank built strains, which have various genetic perturbations introduced into their genomes from the various taught libraries.

**Predictive Design Modeling**

[0427] The linear regression model described above, which utilized data from constructed strains, can be used to make performance predictions for strains that haven’t yet been built.

[0428] The procedure can be summarized as follows: generate *in silico* all possible configurations of genetic changes → use the regression model to predict relative strain performance → order the candidate strain designs by performance. Thus, by utilizing the regression model to predict the performance of as-yet-unbuilt strains, the method allows for the production of higher performing strains, while simultaneously conducting fewer experiments.

**Generate Configurations**
[0429] When constructing a model to predict performance of as-yet-unbuilt strains, the first step is to produce a sequence of design candidates. This is done by fixing the total number of genetic changes in the strain, and then defining all possible combinations of genetic changes. For example, one can set the total number of potential genetic changes/perturbations to 29 (e.g. 29 possible SNPs, or 29 different promoters, or any combination thereof as long as the universe of genetic perturbations is 29) and then decide to design all possible 3-member combinations of the 29 potential genetic changes, which will result in 3,654 candidate strain designs.

[0430] To provide context to the aforementioned 3,654 candidate strains, consider that one can calculate the number of non-redundant groupings of size r from n possible members using \( \frac{n!}{((n - r)! \times r!)} \). If \( r = 3, n = 29 \) gives 3,654. Thus, if one designs all possible 3-member combinations of 29 potential changes the results is 3,654 candidate strains. The 29 potential genetic changes are present in the x-axis of Figure 14.

Predict Performance of New Strain Designs

[0431] Using the linear regression constructed above with the combinatorial configurations as input, one can then predict the expected relative performance of each candidate design. For example, the composition of changes for the top 100 predicted strain designs can be summarized in a 2-dimensional map, in which the x-axis lists the pool of potential genetic changes (29 possible genetic changes), and the y-axis shows the rank order. Black cells can be used to indicate the presence of a particular change in the candidate design, while white cells can be used to indicate the absence of that change. See, Figure 14.

[0432] Predictive accuracy should increase over time as new observations are used to iteratively retrain and refit the model. Results from a study by the inventors illustrate the methods by which the predictive model can be iteratively retrained and improved. The quality of model predictions can be assessed through several methods, including a correlation coefficient indicating the strength of association between the predicted and observed values, or the root-mean-square error, which is a measure of the average model error. Using a chosen metric for model evaluation, the system may define rules for when the model should be retrained.

[0433] A couple of unstated assumptions to the above model include: (1) there are no epistatic interactions; and (2) the genetic changes/perturbations utilized to build the predictive model were all made in the same background, as the proposed combinations of genetic changes.
Filtering for Second-order Features

[0434] The above illustrative example focused on linear regression predictions based on predicted host cell performance. In some embodiments, the present linear regression methods can also be applied to non-biomolecule factors, such as saturation biomass, resistance, or other measurable host cell features. Thus the methods of the present disclosure also teach in considering other features outside of predicted performance when prioritizing the candidates to build. Assuming there is additional relevant data, nonlinear terms are also included in the regression model.

Closeness with Existing Strains

[0435] Predicted strains that are similar to ones that have already been built could result in time and cost savings despite not being a top predicted candidate.

Diversity of Changes

[0436] When constructing the aforementioned models, one cannot be certain that genetic changes will truly be additive (as assumed by linear regression and mentioned as an assumption above) due to the presence of epistatic interactions. Therefore, knowledge of genetic change dissimilarity can be used to increase the likelihood of positive additivity. If one knows, for example, that the changes from the top ranked strain are on the same metabolic pathway and have similar performance characteristics, then that information could be used to select another top ranking strain with a dissimilar composition of changes. As described in the section above concerning epistasis mapping, the predicted best genetic changes may be filtered to restrict selection to mutations with sufficiently dissimilar response profiles. Alternatively, the linear regression may be a weighted least squares regression using the similarity matrix to weight predictions.

Diversity of Predicted Performance

[0437] Finally, one may choose to design strains with middling or poor predicted performance, in order to validate and subsequently improve the predictive models.

Iterative strain design optimization

[0438] In embodiments, the order placement engine 208 places a factory order to the factory 210 to manufacture microbial strains incorporating the top candidate mutations. In feedback-loop fashion, the results may be analyzed by the analysis equipment 214 to determine which microbes exhibit desired phenotypic properties (314). During the analysis phase, the modified strain cultures
are evaluated to determine their performance, i.e., their expression of desired phenotypic properties, including the ability to be produced at industrial scale. For example, the analysis phase uses, among other things, image data of plates to measure microbial colony growth as an indicator of colony health. The analysis equipment 214 is used to correlate genetic changes with phenotypic performance, and save the resulting genotype-phenotype correlation data in libraries, which may be stored in library 206, to inform future microbial production.

[0439] In particular, the candidate changes that actually result in sufficiently high measured performance may be added as rows in a database. In this manner, the best performing mutations are added to the predictive strain design model in a supervised machine learning fashion.

[0440] LIMS iterates the design/build/test/analyze cycle based on the correlations developed from previous factory runs. During a subsequent cycle, the analysis equipment 214 alone, or in conjunction with human operators, may select the best candidates as base strains for input back into input interface 202, using the correlation data to fine tune genetic modifications to achieve better phenotypic performance with finer granularity. In this manner, the laboratory information management system of embodiments of the disclosure implements a quality improvement feedback loop.

[0441] In sum, with reference to the flowchart of Figure 22 the iterative predictive strain design workflow may be described as follows:

- Generate a training set of input and output variables, e.g., genetic changes as inputs and performance features as outputs (3302). Generation may be performed by the analysis equipment 214 based upon previous genetic changes and the corresponding measured performance of the microbial strains incorporating those genetic changes.

- Develop an initial model (e.g., linear regression model) based upon a training set (3304). This may be performed by the analysis equipment 214.

- Generate design candidate strains (3306)

- In one embodiment, the analysis equipment 214 may fix the number of genetic changes to be made to a background strain, in the form of combinations of changes. To represent these changes, the analysis equipment 214 may provide to the interpreter 204 one or more DNA specification expressions representing those combinations of changes. (These genetic
changes or the microbial strains incorporating those changes may be referred to as “test inputs.”) The interpreter 204 interprets the one or more DNA specifications, and the execution engine 207 executes the DNA specifications to populate the DNA specification with resolved outputs representing the individual candidate design strains for those changes.

- Based upon the model, the analysis equipment 214 predicts expected performance of each candidate design strain (3308).

- The analysis equipment 214 selects a limited number of candidate designs, e.g., 100, with highest predicted performance (3310).

- As described elsewhere herein with respect to epistasis mapping, the analysis equipment 214 may account for second-order effects such as epistasis, by, e.g., filtering top designs for epistatic effects, or factoring epistasis into the predictive model.

- Build the filtered candidate strains (at the factory 210) based on the factory order generated by the order placement engine 208 (3312).

- The analysis equipment 214 measures the actual performance of the selected strains, selects a limited number of those selected strains based upon their superior actual performance (3314), and adds the design changes and their resulting performance to the predictive model (3316).

- The analysis equipment 214 then iterates back to generation of new design candidate strains (3306), and continues iterating until a stop condition is satisfied. The stop condition may comprise, for example, the measured performance of at least one microbial strain satisfying a performance metric, such as yield, growth rate, or titer.

[0442] In the example above, the iterative optimization of strain design employs feedback and linear regression to implement machine learning. In general, machine learning may be described as the optimization of performance criteria, e.g., parameters, techniques or other features, in the performance of an informational task (such as classification or regression) using a limited number of examples of labeled data, and then performing the same task on unknown data. In supervised machine learning such as that of the linear regression example above, the machine (e.g., a computing device) learns, for example, by identifying patterns, categories, statistical relationships,
or other attributes, exhibited by training data. The result of the learning is then used to predict whether new data will exhibit the same patterns, categories, statistical relationships or other attributes.

[0443] Embodiments of the disclosure may employ other supervised machine learning techniques when training data is available. In the absence of training data, embodiments may employ unsupervised machine learning. Alternatively, embodiments may employ semi-supervised machine learning, using a small amount of labeled data and a large amount of unlabeled data. Embodiments may also employ feature selection to select the subset of the most relevant features to optimize performance of the machine learning model. Depending upon the type of machine learning approach selected, as alternatives or in addition to linear regression, embodiments may employ for example, logistic regression, neural networks, support vector machines (SVMs), decision trees, hidden Markov models, Bayesian networks, Gram Schmidt, reinforcement-based learning, cluster-based learning including hierarchical clustering, genetic algorithms, and any other suitable learning machines known in the art. In particular, embodiments may employ logistic regression to provide probabilities of classification (e.g., classification of genes into different functional groups) along with the classifications themselves. See, e.g., Shevade, A simple and efficient algorithm for gene selection using sparse logistic regression, Bioinformatics, Vol. 19, No. 17 2003, pp. 2246-2253, Leng, et al., Classification using functional data analysis for temporal gene expression data, Bioinformatics, Vol. 22, No. 1, Oxford University Press (2006), pp. 68-76, all of which are incorporated by reference in their entirety herein.

[0444] Embodiments may employ graphics processing unit (GPU) accelerated architectures that have found increasing popularity in performing machine learning tasks, particularly in the form known as deep neural networks (DNN). Embodiments of the disclosure may employ GPU-based machine learning, such as that described in GPU-Based Deep Learning Inference: A Performance and Power Analysis, NVidia Whitepaper, November 2015, Dahl, et al., Multi-task Neural Networks for QSAR Predictions, Dept. of Computer Science, Univ. of Toronto, June 2014 (arXiv:1406.1231 [stat.ML]), all of which are incorporated by reference in their entirety herein. Machine learning techniques applicable to embodiments of the disclosure may also be found in, among other references, Libbrecht, et al., Machine learning applications in genetics and genomics, Nature Reviews: Genetics, Vol. 16, June 2015, Kashyap, et al., Big Data Analytics in Bioinformatics: A Machine Learning Perspective, Journal of Latex Class Files, Vol. 13, No. 9,
Iterative Predictive Strain Design: Example

[0445] The following provides an example application of the iterative predictive strain design workflow outlined above.

[0446] An initial set of training inputs and output variables was prepared. This set comprised 1864 unique engineered strains with defined genetic composition. Each strain contained between 5 and 15 engineered changes. A total of 336 unique genetic changes were present in the training.

[0447] An initial predictive computer model was developed. The implementation used a generalized linear model (Kernel Ridge Regression with 4th order polynomial kernel). The implementation models two distinct phenotypes (yield and productivity). These phenotypes were combined as weighted sum to obtain a single score for ranking, as shown below. Various model parameters, e.g. regularization factor, were tuned via k-fold cross validation over the designated training data.

[0448] The implementation does not incorporate any explicit analysis of interaction effects as described in the Epistasis Mapping section above. However, as those skilled in the art would understand, the implemented generalized linear model may capture interaction effects implicitly through the second, third and fourth order terms of the kernel.

[0449] The model is trained against the training set. After training, a significant quality fitting of the yield model to the training data can be demonstrated.

[0450] Candidate strains are then generated. This embodiments includes a serial build constraint associated with the introduction of new genetic changes to a parent strain. Here, candidates are not considered simply as a function of the desired number of changes. Instead, the analysis equipment 214 selects, as a starting point, a collection of previously designed strains known to have high performance metrics (“seed strains”). The analysis equipment 214 individually applies genetic changes to each of the seed strains. The introduced genetic changes do not include those already present in the seed strain. For various technical, biological or other reasons, certain mutations are explicitly required, or explicitly excluded.
[0451] Based upon the model, the analysis equipment 214 predicted the performance of candidate strain designs. The analysis equipment 214 ranks candidates from “best” to “worst” based on predicted performance with respect to two phenotypes of interest (yield and productivity). Specifically, the analysis equipment 214 uses a weighted sum to score a candidate strain:

\[ \text{Score} = 0.8 \times \frac{\text{yield}}{\text{max(yields)}} + 0.2 \times \frac{\text{prod}}{\text{max(prods)}}, \]

where yield represents predicted yield for the candidate strain,
max(yields) represents the maximum yield over all candidate strains,
prod represents productivity for the candidate strain, and
max(prods) represents the maximum yield over all candidate strains.

[0453] The analysis equipment 214 generates a final set of recommendations from the ranked list of candidates by imposing both capacity constraints and operational constraints. In some embodiments, the capacity limit can be set at a given number, such as 48 computer-generated candidate design strains.

[0454] The trained model (described above) can be used to predict the expected performance (for yield and productivity) of each candidate strain. The analysis equipment 214 can rank the candidate strains using the scoring function given above. Capacity and operational constraints can be then applied to yield a filtered set of 48 candidate strains. Filtered candidate strains are then built (at the factory 210) based on a factory order generated by the order placement engine 208 (3312). The order can be based upon DNA specifications corresponding to the candidate strains.

[0455] In practice, the build process has an expected failure rate whereby a random set of strains is not built.

[0456] The analysis equipment 214 can also be used to measure the actual yield and productivity performance of the selected strains. The analysis equipment 214 can evaluate the model and recommended strains based on three criteria: model accuracy; improvement in strain performance; and equivalence (or improvement) to human expert-generated designs.

[0457] The yield and productivity phenotypes can be measured for recommended strains and compared to the values predicted by the model.
Next, the analysis equipment 214 computes percentage performance change from the parent strain for each of the recommended strains.

Predictive accuracy can be assessed through several methods, including a correlation coefficient indicating the strength of association between the predicted and observed values, or the root-mean-square error, which is a measure of the average model error. Over many rounds of experimentation, model predictions may drift, and new genetic changes may be added to the training inputs to improve predictive accuracy. For this example, design changes and their resulting performance were added to the predictive model (3316).

Genomic design and engineering as a service

In embodiments of the disclosure, the LIMS system software 3210 of Figure 21 may be implemented in a cloud computing system 3202 of Figure 21, to enable multiple users to design and build microbial strains according to embodiments of the present disclosure. Figure 21 illustrates a cloud computing environment 3204 according to embodiments of the present disclosure. Client computers 3206, such as those illustrated in Figure 21, access the LIMS system via a network 3208, such as the Internet. In embodiments, the LIMS system application software 3210 resides in the cloud computing system 3202. The LIMS system may employ one or more computing systems using one or more processors, of the type illustrated in Figure 21. The cloud computing system itself includes a network interface 3212 to interface the LIMS system applications 3210 to the client computers 3206 via the network 3208. The network interface 3212 may include an application programming interface (API) to enable client applications at the client computers 3206 to access the LIMS system software 3210. In particular, through the API, client computers 3206 may access components of the LIMS system 200, including without limitation the software running the input interface 202, the interpreter 204, the execution engine 207, the order placement engine 208, the factory 210, as well as test equipment 212 and analysis equipment 214. A software as a service (SaaS) software module 3214 offers the LIMS system software 3210 as a service to the client computers 3206. A cloud management module 3216 manages access to the LIMS system 3210 by the client computers 3206. The cloud management module 3216 may enable a cloud architecture that employs multitenant applications, virtualization or other architectures known in the art to serve multiple users.
Genomic Automation

[0461] Automation of the methods of the present disclosure enables high-throughput phenotypic screening and identification of target products from multiple test strain variants simultaneously.

[0462] The aforementioned genomic engineering predictive modeling platform is premised upon the fact that hundreds and thousands of mutant strains are constructed in a high-throughput fashion. The robotic and computer systems described below are the structural mechanisms by which such a high-throughput process can be carried out.

[0463] In some embodiments, the present disclosure teaches methods of improving host cell productivities, or rehabilitating industrial strains. As part of this process, the present disclosure teaches methods of assembling DNA, building new strains, screening cultures in plates, and screening cultures in models for tank fermentation. In some embodiments, the present disclosure teaches that one or more of the aforementioned methods of creating and testing new host strains is aided by automated robotics.

HTP Robotic Systems

[0464] In some embodiments, the automated methods of the disclosure comprise a robotic system. The systems outlined herein are generally directed to the use of 96- or 384-well microtiter plates, but as will be appreciated by those in the art, any number of different plates or configurations may be used. In addition, any or all of the steps outlined herein may be automated; thus, for example, the systems may be completely or partially automated.

[0465] In some embodiments, the automated systems of the present disclosure comprise one or more work modules. For example, in some embodiments, the automated system of the present disclosure comprises a DNA synthesis module, a vector cloning module, a strain transformation module, a screening module, and a sequencing module (see Figure 5).

[0466] As will be appreciated by those in the art, an automated system can include a wide variety of components, including, but not limited to: liquid handlers; one or more robotic arms; plate handlers for the positioning of microplates; plate sealers, plate piercers, automated lid handlers to remove and replace lids for wells on non-cross contamination plates; disposable tip assemblies for sample distribution with disposable tips; washable tip assemblies for sample distribution; 96 well
loading blocks; integrated thermal cyclers; cooled reagent racks; microtiter plate pipette positions (optionally cooled); stacking towers for plates and tips; magnetic bead processing stations; filtrations systems; plate shakers; barcode readers and applicators; and computer systems.

[0467] In some embodiments, the robotic systems of the present disclosure include automated liquid and particle handling enabling high-throughput pipetting to perform all the steps in the process of gene targeting and recombination applications. This includes liquid and particle manipulations such as aspiration, dispensing, mixing, diluting, washing, accurate volumetric transfers, retrieving and discarding of pipette tips; and repetitive pipetting of identical volumes for multiple deliveries from a single sample aspiration. These manipulations are cross-contamination-free liquid, particle, cell, and organism transfers. The instruments perform automated replication of microplate samples to filters, membranes, and/or daughter plates, high-density transfers, full-plate serial dilutions, and high capacity operation.

[0468] In some embodiments, the customized automated liquid handling system of the disclosure is a TECAN machine (e.g. a customized TECAN Freedom Evo).

[0469] In some embodiments, the automated systems of the present disclosure are compatible with platforms for multi-well plates, deep-well plates, square well plates, reagent troughs, test tubes, mini tubes, microfuge tubes, cryovials, filters, micro array chips, optic fibers, beads, agarose and acrylamide gels, and other solid-phase matrices or platforms are accommodated on an upgradeable modular deck. In some embodiments, the automated systems of the present disclosure contain at least one modular deck for multi-position work surfaces for placing source and output samples, reagents, sample and reagent dilution, assay plates, sample and reagent reservoirs, pipette tips, and an active tip-washing station.

[0470] In some embodiments, the automated systems of the present disclosure include high-throughput electroporation systems. In some embodiments, the high-throughput electroporation systems are capable of transforming cells in 96 or 384-well plates. In some embodiments, the high-throughput electroporation systems include VWR® High-throughput Electroporation Systems, BTX™, Bio-Rad® Gene Pulser MXcell™ or other multi-well electroporation system.
[0471] In some embodiments, the integrated thermal cycler and/or thermal regulators are used for stabilizing the temperature of heat exchangers such as controlled blocks or platforms to provide accurate temperature control of incubating samples from 0°C to 100°C.

[0472] In some embodiments, the automated systems of the present disclosure are compatible with interchangeable machine-heads (single or multi-channel) with single or multiple magnetic probes, affinity probes, replicators or pipetters, capable of robotically manipulating liquid, particles, cells, and multi-cellular organisms. Multi-well or multi-tube magnetic separators and filtration stations manipulate liquid, particles, cells, and organisms in single or multiple sample formats.

[0473] In some embodiments, the automated systems of the present disclosure are compatible with camera vision and/or spectrometer systems. Thus, in some embodiments, the automated systems of the present disclosure are capable of detecting and logging color and absorption changes in ongoing cellular cultures.

[0474] In some embodiments, the automated system of the present disclosure is designed to be flexible and adaptable with multiple hardware add-ons to allow the system to carry out multiple applications. The software program modules allow creation, modification, and running of methods. The system’s diagnostic modules allow setup, instrument alignment, and motor operations. The customized tools, labware, and liquid and particle transfer patterns allow different applications to be programmed and performed. The database allows method and parameter storage. Robotic and computer interfaces allow communication between instruments.

[0475] Thus, in some embodiments, the present disclosure teaches a high-throughput strain engineering platform, as depicted in Figures 15 and 16.

[0476] Persons having skill in the art will recognize the various robotic platforms capable of carrying out the HTP engineering methods of the present disclosure. Table 3 below provides a non-exclusive list of scientific equipment capable of carrying out each step of the HTP engineering steps of the present disclosure as described in Figures 15 and 16.

**Table 3** - Non-exclusive list of Scientific Equipment Compatible with the HTP engineering methods of the present disclosure.
<table>
<thead>
<tr>
<th>QC DNA parts</th>
<th>Equipment Type</th>
<th>Operation(s) performed</th>
<th>Compatible Equipment Make/Model/Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencer (sanger: Beckman)</td>
<td>fragment analyzers (capillary electrophoresis)</td>
<td>gel electrophoresis to confirm PCR products of appropriate size</td>
<td>Agilent Bioanalyzer, AATI Fragment Analyzer, or equivalents</td>
</tr>
<tr>
<td>NGS (next generation sequencing) instrument</td>
<td>Thermal cyclers</td>
<td>PCR amplification of DNA parts</td>
<td>Inheco Cycler, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents</td>
</tr>
<tr>
<td>nanodrop/plate reader</td>
<td>liquid handlers</td>
<td>Hitpicking (combining by transferring) primers/templates for PCR amplification of DNA parts</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Verifying sequence of parts/templates</td>
<td>Beckman Ceq-8000, Beckman GenomeLab™, or equivalents</td>
</tr>
<tr>
<td></td>
<td>Verifying sequence of parts/templates</td>
<td>Illumina MiSeq series sequences, illumina Hi-Seq, Ion torrent, pac bio or other equivalents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>assessing concentration of DNA samples</td>
<td>Molecular Devices SpectraMax M5, Tecan M1000, or equivalents.</td>
<td></td>
</tr>
<tr>
<td>Equipment Type</td>
<td>Operation(s) performed</td>
<td>Compatible Equipment Make/Model/Configuration</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>liquid handlers</td>
<td>Hitpicking (combining by transferring) DNA parts for assembly along with cloning vector, addition of reagents for assembly reaction/process</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Colony pickers</td>
<td>for inoculating colonies in liquid media</td>
<td>Scirobotics Pickolo, Molecular Devices QPix 420</td>
<td></td>
</tr>
<tr>
<td>liquid handlers</td>
<td>Hitpicking primers/templates, diluting samples</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Fragment analyzers (capillary electrophoresis)</td>
<td>gel electrophoresis to confirm assembled products of appropriate size</td>
<td>Agilent Bioanalyzer, AATI Fragment Analyzer</td>
<td></td>
</tr>
<tr>
<td>Sequencer (sanger: Beckman)</td>
<td>Verifying sequence of assembled plasmids</td>
<td>ABI3730 Thermo Fisher, Beckman Ceq-8000, Beckman GenomeLab™, or equivalents</td>
<td></td>
</tr>
<tr>
<td>NGS (next generation sequencing) instrument</td>
<td>Verifying sequence of assembled plasmids</td>
<td>Illumina MiSeq series sequences, illumina Hi-Seq, Ion torrent, pac bio or other equivalents</td>
<td></td>
</tr>
<tr>
<td>Equipment Type</td>
<td>Operation(s) performed</td>
<td>Compatible Equipment Make/Model/Configuration</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Prepare base strain and DNA assembly</td>
<td>centrifuge</td>
<td>spinning / pelleting cells</td>
<td>Beckman Avanti floor centrifuge, Hettich Centrifuge</td>
</tr>
<tr>
<td>Electroporators</td>
<td>electroporative transformation of cells</td>
<td>BTX Gemini X2, BIO-RAD MicroPulser Electroporator</td>
<td></td>
</tr>
<tr>
<td>Ballistic transformation</td>
<td>ballistic transformation of cells</td>
<td>BIO-RAD PDS1000</td>
<td></td>
</tr>
<tr>
<td>Incubators, thermal cyclers</td>
<td>for chemical transformation/heat shock</td>
<td>Inhoco Cycler, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>for combining DNA, cells, buffer</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Integrate DNA into genome of base strain</td>
<td>Colony pickers</td>
<td>for inoculating colonies in liquid media</td>
<td>Scirobotics Pickolo, Molecular Devices QPix 420</td>
</tr>
<tr>
<td>Equipment Type</td>
<td>Operation(s) performed</td>
<td>Compatible Equipment Make/Model/Configuration</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>For transferring cells onto Agar, transferring from culture plates to different culture plates (inoculation into other selective media)</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Platform shaker-incubators</td>
<td>incubation with shaking of microtiter plate cultures</td>
<td>Kuhner Shaker ISF4-X, Infors-hlt Multitron Pro</td>
<td></td>
</tr>
<tr>
<td>Colony pickers</td>
<td>for inoculating colonies in liquid media</td>
<td>Scirobtics Pickolo, Molecular Devices QPix 420</td>
<td></td>
</tr>
<tr>
<td>liquid handlers</td>
<td>Hitpicking primers/templates, diluting samples</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Thermal cyclers</td>
<td>cPCR verification of strains</td>
<td>Inheco Cycler, ABI 2720, ABI Proflex 384, ABI Veriti, or equivalents</td>
<td></td>
</tr>
<tr>
<td>Fragment analyzers (capillary electrophoresis)</td>
<td>gel electrophoresis to confirm cPCR products of appropriate size</td>
<td>Infors-hlt Multitron Pro, Kuhner Shaker ISF4-X</td>
<td></td>
</tr>
<tr>
<td>Sequencer (sanger: Beckman)</td>
<td>Sequence verification of introduced modification</td>
<td>Beckman Ceq-8000, Beckman GenomeLab™, or equivalents</td>
<td></td>
</tr>
<tr>
<td>NGS (next generation)</td>
<td>Sequence verification of introduced modification</td>
<td>Illumina MiSeq series sequences, illumina Hi-Seq, Ion</td>
<td></td>
</tr>
</tbody>
</table>

QC transformed strain
<table>
<thead>
<tr>
<th>Equipment Type</th>
<th>Operation(s) performed</th>
<th>Compatible Equipment Make/Model/Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequencing) instrument</td>
<td>For transferring from culture plates to different culture plates (inoculation into production media)</td>
<td>torrent, pac bio or other equivalents</td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>for inoculating colonies in liquid media</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td>Colony pickers</td>
<td>incubation with shaking of microtiter plate cultures</td>
<td>Scirobotics Pickolo, Molecular Devices QPix 420</td>
</tr>
<tr>
<td>Platform shaker-incubators</td>
<td></td>
<td>Kuhner Shaker ISF4-X, Infors-hht Multitron Pro</td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>For transferring from culture plates to different culture plates (inoculation into production media)</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td>Platform shaker-incubators</td>
<td>incubation with shaking of microtiter plate cultures</td>
<td>Kuhner Shaker ISF4-X, Infors-hht Multitron Pro</td>
</tr>
<tr>
<td>liquid dispensers</td>
<td>Dispense liquid culture media into microtiter plates</td>
<td>Well mate (Thermo), Benchcel2R (velocity 11), plateloc (velocity 11)</td>
</tr>
<tr>
<td>microplate labeler</td>
<td>apply barcoders to plates</td>
<td>Microplate labeler (a2+cab-agilent), benchcell 6R (velocity 11)</td>
</tr>
<tr>
<td>Equipment Type</td>
<td>Operation(s) performed</td>
<td>Compatible Equipment Make/Model/Configuration</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>For transferring from culture plates to different culture plates (inoculation into production media)</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td>Platform shaker-incubators</td>
<td>incubation with shaking of microtiter plate cultures</td>
<td>Kuhner Shaker ISF4-X, Infors-HT Multitron Pro</td>
</tr>
<tr>
<td>liquid dispensers</td>
<td>Dispense liquid culture media into multiple microtiter plates and seal plates</td>
<td>well mate (Thermo), Benchce12R (velocity 11), plateloc (velocity 11)</td>
</tr>
<tr>
<td>microplate labeler</td>
<td>Apply barcodes to plates</td>
<td>microplate labeler (a2+ cab-agilent), benchcell 6R (velocity11)</td>
</tr>
<tr>
<td>Liquid handlers</td>
<td>For processing culture broth for downstream analytical</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td>UHPLC, HPLC</td>
<td>quantitative analysis of precursor and target compounds</td>
<td>Agilent 1290 Series UHPLC and 1200 Series HPLC with UV and RI detectors, or equivalent; also any LC/MS</td>
</tr>
<tr>
<td>LC/MS</td>
<td>highly specific analysis of precursor and target compounds as well as side and degradation products</td>
<td>Agilent 6490 QQQ and 6550 QTOF coupled to 1290 Series UHPLC</td>
</tr>
<tr>
<td>Culture strains in flasks</td>
<td><strong>Equipment Type</strong></td>
<td><strong>Operation(s) performed</strong></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td>Spectrophotometer</td>
<td>Quantification of different compounds using spectrophotometer based assays</td>
</tr>
<tr>
<td>Fermenters:</td>
<td></td>
<td>incubation with shaking</td>
</tr>
<tr>
<td>Platform shakers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Generate product from strain</th>
<th><strong>Operation(s) performed</strong></th>
<th><strong>Compatible Equipment Make/Model/Configuration</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fermenters: DASGIPs (Eppendorf), BIO-FLOs (Sartorius-stedim)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluate performance</th>
<th><strong>Operation(s) performed</strong></th>
<th><strong>Compatible Equipment Make/Model/Configuration</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquid handlers</td>
<td>For transferring from culture plates to different culture plates (inoculation into production media)</td>
<td>Hamilton Microlab STAR, Labcyte Echo 550, Tecan EVO 200, Beckman Coulter Biomek FX, or equivalents</td>
</tr>
<tr>
<td>UHPLC, HPLC</td>
<td>quantitative analysis of precursor and target compounds</td>
<td>Agilent 1290 Series UHPLC and 1200 Series HPLC with UV and RI detectors, or equivalent; also any LC/MS</td>
</tr>
<tr>
<td>LC/MS</td>
<td>highly specific analysis of precursor and target compounds as well as side and degradation products</td>
<td>Agilent 6490 QQQ and 6550 QTOF coupled to 1290 Series UHPLC</td>
</tr>
<tr>
<td>Equipment Type</td>
<td>Operation(s) performed</td>
<td>Compatible Equipment Make/Model/Configuration</td>
</tr>
<tr>
<td>----------------</td>
<td>------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Flow cytometer</td>
<td>Characterize strain performance (measure viability)</td>
<td>BD Accuri, Millipore Guava</td>
</tr>
<tr>
<td>Spectrophotometer</td>
<td>Characterize strain performance (measure biomass)</td>
<td>Tecan M1000, Spectramax M5, or other equivalents</td>
</tr>
</tbody>
</table>

**Computer System Hardware**

[0477] Figure 23 illustrates an example of a computer system 800 that may be used to execute program code stored in a non-transitory computer readable medium (e.g., memory) in accordance with embodiments of the disclosure. The computer system includes an input/output subsystem 802, which may be used to interface with human users and/or other computer systems depending upon the application. The I/O subsystem 802 may include, e.g., a keyboard, mouse, graphical user interface, touchscreen, or other interfaces for input, and, e.g., an LED or other flat screen display, or other interfaces for output, including application program interfaces (APIs). Other elements of embodiments of the disclosure, such as the components of the LIMS system, may be implemented with a computer system like that of computer system 800.

[0478] Program code may be stored in non-transitory media such as persistent storage in secondary memory 810 or main memory 808 or both. Main memory 808 may include volatile memory such as random access memory (RAM) or non-volatile memory such as read only memory (ROM), as well as different levels of cache memory for faster access to instructions and data. Secondary memory may include persistent storage such as solid state drives, hard disk drives or optical disks. One or more processors 804 reads program code from one or more non-transitory media and executes the code to enable the computer system to accomplish the methods performed by the embodiments herein. Those skilled in the art will understand that the processor(s) may ingest source code, and interpret or compile the source code into machine code that is understandable at the hardware gate level of the processor(s) 804. The processor(s) 804 may include graphics
processing units (GPUs) for handling computationally intensive tasks. Particularly in machine learning, one or more CPUs 804 may offload the processing of large quantities of data to one or more GPUs 804.

[0479] The processor(s) 804 may communicate with external networks via one or more communications interfaces 807, such as a network interface card, WiFi transceiver, etc. A bus 805 communicatively couples the I/O subsystem 802, the processor(s) 804, peripheral devices 806, communications interfaces 807, memory 808, and persistent storage 810. Embodiments of the disclosure are not limited to this representative architecture. Alternative embodiments may employ different arrangements and types of components, e.g., separate buses for input-output components and memory subsystems.

[0480] Those skilled in the art will understand that some or all of the elements of embodiments of the disclosure, and their accompanying operations, may be implemented wholly or partially by one or more computer systems including one or more processors and one or more memory systems like those of computer system 800. In particular, the elements of the LIMS system 200 and any robotics and other automated systems or devices described herein may be computer-implemented. Some elements and functionality may be implemented locally and others may be implemented in a distributed fashion over a network through different servers, e.g., in client-server fashion, for example. In particular, server-side operations may be made available to multiple clients in a software as a service (SaaS) fashion, as shown in Figure 21.

[0481] The term component in this context refers broadly to software, hardware, or firmware (or any combination thereof) component. Components are typically functional components that can generate useful data or other output using specified input(s). A component may or may not be self-contained. An application program (also called an “application”) may include one or more components, or a component can include one or more application programs.

[0482] Some embodiments include some, all, or none of the components along with other modules or application components. Still yet, various embodiments may incorporate two or more of these components into a single module and/or associate a portion of the functionality of one or more of these components with a different component.

[0483] The term “memory” can be any device or mechanism used for storing information. In accordance with some embodiments of the present disclosure, memory is intended to encompass
any type of, but is not limited to: volatile memory, nonvolatile memory, and dynamic memory. For example, memory can be random access memory, memory storage devices, optical memory devices, magnetic media, floppy disks, magnetic tapes, hard drives, SIMMs, SDRAM, DIMMs, RDRAM, DDR RAM, SODIMMS, erasable programmable read-only memories (EPROMs), electrically erasable programmable read-only memories (EEPROMs), compact disks, DVDs, and/or the like. In accordance with some embodiments, memory may include one or more disk drives, flash drives, databases, local cache memories, processor cache memories, relational databases, flat databases, servers, cloud based platforms, and/or the like. In addition, those of ordinary skill in the art will appreciate many additional devices and techniques for storing information can be used as memory.

[0484] Memory may be used to store instructions for running one or more applications or modules on a processor. For example, memory could be used in some embodiments to house all or some of the instructions needed to execute the functionality of one or more of the modules and/or applications disclosed in this application.

HTP Microbial Strain Engineering Based Upon Genetic Design Predictions: An Example Workflow

[0485] In some embodiments, the present disclosure teaches the directed engineering of new host organisms based on the recommendations of the computational analysis systems of the present disclosure.

[0486] In some embodiments, the present disclosure is compatible with all genetic design and cloning methods. That is, in some embodiments, the present disclosure teaches the use of traditional cloning techniques such as polymerase chain reaction, restriction enzyme digestions, ligation, homologous recombination, RT PCR, and others generally known in the art and are disclosed in for example: Sambrook *et al.* (2001) Molecular Cloning: A Laboratory Manual (3rd ed., Cold Spring Harbor Laboratory Press, Plainview, New York), incorporated herein by reference.

[0487] In some embodiments, the cloned sequences can include possibilities from any of the HTP genetic design libraries taught herein, for example: promoters from a promoter swap library, SNPs from a SNP swap library, start or stop codons from a start/stop codon exchange library, terminators
from a STOP swap library, sequence optimizations from a sequence optimization library or transposons from a transposon mutagenesis library.

[0488] Further, the exact sequence combinations that should be included in a particular construct can be informed by the epistatic mapping function.

[0489] In other embodiments, the cloned sequences can also include sequences based on rational design (hypothesis-driven) and/or sequences based on other sources, such as scientific publications.

[0490] In some embodiments, the present disclosure teaches methods of directed engineering, including the steps of i) generating custom-made SNP-specific DNA, ii) assembling SNP-specific plasmids, iii) transforming target host cells with SNP-specific DNA, and iv) looping out any selection markers (See Figure 2).

[0491] Figure 4A depicts the general workflow of the strain engineering methods of the present disclosure, including acquiring and assembling DNA, assembling vectors, transforming host cells and removing selection markers.

**Build Specific DNA Oligonucleotides**

[0492] In some embodiments, the present disclosure teaches inserting and/or replacing and/or altering and/or deleting a DNA segment of the host cell organism. In some aspects, the methods taught herein involve building an oligonucleotide of interest (*i.e.* a target DNA segment), that will be incorporated into the genome of a host organism. In some embodiments, the target DNA segments of the present disclosure can be obtained *via* any method known in the art, including: copying or cutting from a known template, mutation, or DNA synthesis. In some embodiments, the present disclosure is compatible with commercially available gene synthesis products for producing target DNA sequences (*e.g.*, GeneArt™, GeneMaker™, GenScript™, Anagen™, Blue Heron™, Entelechon™, GeNOsys, Inc., or Qiagen™).

[0493] In some embodiments, the target DNA segment is designed to incorporate a SNP into a selected DNA region of the host organism (*e.g.*, adding a beneficial SNP). In other embodiments, the DNA segment is designed to remove a SNP from the DNA of the host organisms (*e.g.*, removing a detrimental or neutral SNP).
In some embodiments, the oligonucleotides used in the inventive methods can be synthesized using any of the methods of enzymatic or chemical synthesis known in the art. The oligonucleotides may be synthesized on solid supports such as controlled pore glass (CPG), polystyrene beads, or membranes composed of thermoplastic polymers that may contain CPG. Oligonucleotides can also be synthesized on arrays, on a parallel microscale using microfluidics (Tian et al., Mol. BioSyst., 5, 714-722 (2009)), or known technologies that offer combinations of both (see Jacobsen et al., U.S. Pat. App. No. 2011/0172127).

Synthesis on arrays or through microfluidics offers an advantage over conventional solid support synthesis by reducing costs through lower reagent use. The scale required for gene synthesis is low, so the scale of oligonucleotide product synthesized from arrays or through microfluidics is acceptable. However, the synthesized oligonucleotides are of lesser quality than when using solid support synthesis (See Tian infra.; see also Staehler et al., U.S. Pat. App. No. 2010/0216648).

A great number of advances have been achieved in the traditional four-step phosphoramidite chemistry since it was first described in the 1980s (see for example, Sierzchala, et al. J. Am. Chem. Soc., 125, 13427-13441 (2003) using peroxo anion deprotection; Hayakawa et al., U.S. Pat. No. 6,040,439 for alternative protecting groups; Azhayev et al, Tetrahedron 57, 4977-4986 (2001) for universal supports; Kozlov et al., Nucleosides, Nucleotides, and Nucleic Acids, 24 (5-7), 1037-1041 (2005) for improved synthesis of longer oligonucleotides through the use of large-pore CPG; and Damha et al., NAR, 18, 3813-3821 (1990) for improved derivatization).

Regardless of the type of synthesis, the resulting oligonucleotides may then form the smaller building blocks for longer oligonucleotides. In some embodiments, smaller oligonucleotides can be joined together using protocols known in the art, such as polymerase chain assembly (PCA), ligase chain reaction (LCR), and thermodynamically balanced inside-out synthesis (TBIO) (see Czar et al. Trends in Biotechnology, 27, 63-71 (2009)). In PCA, oligonucleotides spanning the entire length of the desired longer product are annealed and extended in multiple cycles (typically about 55 cycles) to eventually achieve full-length product. LCR uses ligase enzyme to join two oligonucleotides that are both annealed to a third oligonucleotide. TBIO synthesis starts at the center of the desired product and is progressively
extended in both directions by using overlapping oligonucleotides that are homologous to the forward strand at the 5' end of the gene and against the reverse strand at the 3' end of the gene.

[0498] Another method of synthesizing a larger double stranded DNA fragment is to combine smaller oligonucleotides through top-strand PCR (TSP). In this method, a plurality of oligonucleotides spans the entire length of a desired product and contain overlapping regions to the adjacent oligonucleotide(s). Amplification can be performed with universal forward and reverse primers, and through multiple cycles of amplification a full-length double stranded DNA product is formed. This product can then undergo optional error correction and further amplification that results in the desired double stranded DNA fragment end product.

[0499] In one method of TSP, the set of smaller oligonucleotides that will be combined to form the full-length desired product are between 40-200 bases long and overlap each other by at least about 15-20 bases. For practical purposes, the overlap region should be at a minimum long enough to ensure specific annealing of oligonucleotides and have a high enough melting temperature (T_m) to anneal at the reaction temperature employed. The overlap can extend to the point where a given oligonucleotide is completely overlapped by adjacent oligonucleotides. The amount of overlap does not seem to have any effect on the quality of the final product. The first and last oligonucleotide building block in the assembly should contain binding sites for forward and reverse amplification primers. In one embodiment, the terminal end sequence of the first and last oligonucleotide contain the same sequence of complementarity to allow for the use of universal primers.

Assembling/Cloning Custom Plasmids

[0500] In some embodiments, the present disclosure teaches methods for constructing vectors capable of inserting desired target DNA sections (e.g. containing a particular SNP or transposon) into the genome of host organisms. In some embodiments, the present disclosure teaches methods of cloning vectors comprising the target DNA, homology arms, and at least one selection marker (see Figure 3).

[0501] In some embodiments, the present disclosure is compatible with any vector suited for transformation into the host organism. In some embodiments, the present disclosure teaches use of shuttle vectors compatible with a host cell. In one embodiment, a shuttle vector for use in the methods provided herein is a shuttle vector compatible with an E. coli and/or Corynebacterium
host cell. Shuttle vectors for use in the methods provided herein can comprise markers for selection and/or counter-selection as described herein. The markers can be any markers known in the art and/or provided herein. The shuttle vectors can further comprise any regulatory sequence(s) and/or sequences useful in the assembly of the shuttle vectors as known in the art. The shuttle vectors can further comprise any origins of replication that may be needed for propagation in a host cell as provided herein such as, for example, E. coli or C. glutamicum. The regulatory sequence can be any regulatory sequence known in the art or provided herein such as, for example, a promoter, start, stop, signal, secretion and/or termination sequence used by the genetic machinery of the host cell. In certain instances, the target DNA can be inserted into vectors, constructs or plasmids obtainable from any repository or catalogue product, such as a commercial vector (see e.g., DNA2.0 custom or GATEWAY® vectors). In certain instances, the target DNA can be inserted into vectors, constructs or plasmids obtainable from any repository or catalogue product, such as a commercial vector (see e.g., DNA2.0 custom or GATEWAY® vectors).


[0503] In some embodiments, the present disclosure teaches cloning vectors with at least one selection marker. Various selection marker genes are known in the art often encoding antibiotic resistance function for selection in prokaryotic (e.g., against ampicillin, kanamycin, tetracycline,
chloramphenicol, zeocin, spectinomycin/streptomycin) or eukaryotic cells (e.g. geneticin, neomycin, hygromycin, puromycin, blasticidin, zeocin) under selective pressure. Other marker systems allow for screening and identification of wanted or unwanted cells such as the well-known blue/white screening system used in bacteria to select positive clones in the presence of X-gal or fluorescent reporters such as green or red fluorescent proteins expressed in successfully transduced host cells. Another class of selection markers most of which are only functional in prokaryotic systems relates to counter selectable marker genes often also referred to as “death genes” which express toxic gene products that kill producer cells. Examples of such genes include sacB, rpsL(strA), tetAR, pheS, thyA, gata-1, or ccdB, the function of which is described in (Reyrat et al. 1998 “Counterselectable Markers: Untapped Tools for Bacterial Genetics and Pathogenesis.” Infect Immun. 66(9): 4011-4017).

**Protoplasting Methods**

[0504] In one embodiment, the methods and systems provided herein make use of the generation of protoplasts from filamentous fungal cells. Suitable procedures for preparation of protoplasts can be any known in the art including, for example, those described in EP 238,023 and Yelton et al. (1984, Proc. Natl. Acad. Sci. USA 81:1470-1474). In one embodiment, protoplasts are generated by treating a culture of filamentous fungal cells with one or more lytic enzymes or a mixture thereof. The lytic enzymes can be a beta-glucanase and/or a polygalacturonase. In one embodiment, the enzyme mixture for generating protoplasts is VinoTaste concentrate. Following enzymatic treatment, the protoplasts can be isolated using methods known in the art such as, for example, centrifugation.

[0505] The pre-cultivation and the actual protoplasting step can be varied to optimize the number of protoplasts and the transformation efficiency. For example, there can be variations of inoculum size, inoculum method, pre-cultivation media, pre-cultivation times, pre-cultivation temperatures, mixing conditions, washing buffer composition, dilution ratios, buffer composition during lytic enzyme treatment, the type and/or concentration of lytic enzyme used, the time of incubation with lytic enzyme, the protoplast washing procedures and/or buffers, the concentration of protoplasts and/or polynucleotide and/or transformation reagents during the actual transformation, the physical parameters during the transformation, the procedures following the transformation up to the obtained transformants.
[0506] Protoplasts can be resuspended in an osmotic stabilizing buffer. The composition of such buffers can vary depending on the species, application and needs. However, typically these buffers contain either an organic component like sucrose, citrate, mannitol or sorbitol between 0.5 and 2 M. More preferably between 0.75 and 1.5 M; most preferred is 1 M. Otherwise these buffers contain an inorganic osmotic stabilizing component like KCl, MgSO._sub.4_, NaCl or MgCl._sub.2_ in concentrations between 0.1 and 1.5 M. Preferably between 0.2 and 0.8 M; more preferably between 0.3 and 0.6 M, most preferably 0.4 M. The most preferred stabilizing buffers are STC (sorbitol, 0.8 M; CaCl._sub.2_, 25 mM; Tris, 25 mM; pH 8.0) or KCl-citrate (KCl, 0.3-0.6 M; citrate, 0.2% (w/v)). The protoplasts can be used in a concentration between $1 \times 10^5$ and $1 \times 10^{10}$ cells/ml. Preferably, the concentration is between $1 \times 10^6$ and $1 \times 10^9$; more preferably the concentration is between $1 \times 10^7$ and $5 \times 10^8$; most preferably the concentration is $1 \times 10^8$ cells/ml. DNA is used in a concentration between 0.01 and 10 ug; preferably between 0.1 and 5 ug, even more preferably between 0.25 and 2 ug; most preferably between 0.5 and 1 ug. To increase the efficiency of transfection carrier DNA (as salmon sperm DNA or non-coding vector DNA) may be added to the transformation mixture.

[0507] In one embodiment, following generation and subsequent isolation, the protoplasts are mixed with one or more cryoprotectants. The cryoprotectants can be glycols, dimethyl sulfoxide (DMSO), polyols, sugars, 2-Methyl-2,4-pentanediol (MPD), polyvinylpyrrolidone (PVP), methylcellulose, C-linked antifreeze glycoproteins (C-AFGP) or combinations thereof. Glycols for use as cryoprotectants in the methods and systems provided herein can be selected from ethylene glycol, propylene glycol, polypropylene glycol (PEG), glycerol, or combinations thereof. Polyols for use as cryoprotectants in the methods and systems provided herein can be selected from propane-1,2-diol, propane-1,3-diol, 1,1,1-tris-(hydroxymethyl)ethane (THME), and 2-ethyl-2-(hydroxymethyl)-propane-1,3-diol (EHMP), or combinations thereof. Sugars for use as cryoprotectants in the methods and systems provided herein can be selected from trehalose, sucrose, glucose, raffinose, dextrose or combinations thereof. In one embodiment, the protoplasts are mixed with DMSO. DMSO can be mixed with the protoplasts at a final concentration of at least, at most, less than, greater than, equal to, or about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 12.5%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, or 75% w/v or v/v. The protoplasts/cryoprotectant (e.g., DMSO) mixture can be distributed to microtiter plates prior to storage. The protoplast/cryoprotectant (e.g., DMSO) mixture can be stored at any
temperature provided herein for long-term storage (e.g., several hours, day(s), week(s), month(s), year(s)) as provided herein such as, for example -20°C or -80°C. In one embodiment, an additional cryoprotectant (e.g., PEG) is added to the protoplasts/DMSO mixture. In yet another embodiment, the additional cryoprotectant (e.g., PEG) is added to the protoplast/DMSO mixture prior to storage. The PEG can be any PEG provided herein and can be added at any concentration (e.g., w/v or v/v) as provided herein.

Protoplast Transformation Methods

[0508] In one embodiment, the methods and systems provided herein require the transfer of nucleic acids to protoplasts derived from filamentous fungal cells as described herein. In another embodiment, the transformation utilized by the methods and systems provided herein is high-throughput in nature and/or is partially or fully automated as described herein. Further to this embodiment, the transformation is performed by adding constructs or expression constructs as described herein to the wells of a microtiter plate followed by aliquoting protoplasts generated by the methods provided herein to each well of the microtiter plate. Suitable procedures for transformation/transfection of protoplasts can be any known in the art including, for example, those described in international patent applications PCT/NL99/00618, PCT/EP99/202516, Finkelstein and Ball (eds.), Biotechnology of filamentous fungi, technology and products, Butterworth-Heinemann (1992), Bennett and Lasure (eds.) More Gene Manipulations in fungi, Academic Press (1991), Turner, in: Puhlur (ed), Biotechnology, second completely revised edition, VHC (1992) protoplast fusion, and the Ca-PEG mediated protoplast transformation as described in EP635574B. Alternatively, transformation of the filamentous fungal host cells or protoplasts derived therefrom can also be performed by electroporation such as, for example, the electroporation described by Chakraborty and Kapoor, Nucleic Acids Res. 18:6737 (1990), Agrobacterium tumefaciens-mediated transformation, biolistic introduction of DNA such as, for example, as described in Christiansen et al., Curr. Genet. 29:100 102 (1995); Durand et al., Curr. Genet. 31:158 161 (1997); and Barcellos et al., Can. J. Microbiol. 44:1137 1141 (1998) or "magnetobiostic" transfection of cells such as, for example, described in U.S. Pat. Nos. 5,516,670 and 5,753,477. In one embodiment, the transformation procedure used in the methods and systems provided herein is one amendable to being high-throughput and/or automated as provided herein such as, for example, PEG mediated transformation.
Transformation of the protoplasts generated using the methods described herein can be facilitated through the use of any transformation reagent known in the art. Suitable transformation reagents can be selected from Polyethylene Glycol (PEG), FUGENE® HD (from Roche), Lipofectamine® or OLIGOFECTAMINE® (from Invitrogen), TRANSPASS®D1 (from New England Biolabs), LYPOVEC® or LIPOGEN® (from Invivogen). In one embodiment, PEG is the most preferred transformation/transfection reagent. PEG is available at different molecular weights and can be used at different concentrations. Preferably PEG 4000 is used between 10% and 60%, more preferably between 20% and 50%, most preferably at 30%. In one embodiment, the PEG is added to the protoplasts prior to storage as described herein.

**Transformation of Host Cells**

In some embodiments, the vectors of the present disclosure may be introduced into the host cells using any of a variety of techniques, including transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer (see Christie, P.J., and Gordon, J.E., 2014 “The Agrobacterium Ti Plasmids” Microbiol SPectr. 2014; 2(6); 10.1128). Particular methods include calcium phosphate transfection, DEAE-Dextran mediated transfection, lipofection, or electroporation (Davis, L., Dibner, M., Battey, I., 1986 “Basic Methods in Molecular Biology”). Other methods of transformation include for example, lithium acetate transformation and electroporation See, e.g., Gietz et al., Nucleic Acids Res. 27:69-74 (1992); Ito et al., J. Bacterol. 153:163-168 (1983); and Becker and Guarente, Methods in Enzymology 194:182-187 (1991). In some embodiments, transformed host cells are referred to as recombinant host strains.

In some embodiments, the present disclosure teaches high-throughput transformation of cells using the 96-well plate robotics platform and liquid handling machines of the present disclosure.

In some embodiments, the present disclosure teaches screening transformed cells with one or more selection markers as described above. In one such embodiment, cells transformed with a vector comprising a kanamycin resistance marker (KanR) are plated on media containing effective amounts of the kanamycin antibiotic. Colony forming units visible on kanamycin-laced media are presumed to have incorporated the vector cassette into their genome. Insertion of the desired sequences can be confirmed via PCR, restriction enzyme analysis, and/or sequencing of the relevant insertion site.
**Looping Out of Selected Sequences**

[0513] In some embodiments, the present disclosure teaches methods of looping out selected regions of DNA from the host organisms. The looping out method can be as described in Nakashima et al. 2014 “Bacterial Cellular Engineering by Genome Editing and Gene Silencing.” Int. J. Mol. Sci. 15(2), 2773-2793. In some embodiments, the present disclosure teaches looping out selection markers from positive transformants. Looping out deletion techniques are known in the art, and are described in (Tear et al. 2014 “Excision of Unstable Artificial Gene-Specific inverted Repeats Mediates Scar-Free Gene Deletions in Escherichia coli.” Appl. Biochem. Biotech. 175:1858-1867). The looping out methods used in the methods provided herein can be performed using single-crossover homologous recombination or double-crossover homologous recombination. In one embodiment, looping out of selected regions as described herein can entail using single-crossover homologous recombination as described herein.

[0514] First, loop out vectors are inserted into selected target regions within the genome of the host organism (e.g., via homologous recombination, CRISPR, or other gene editing technique). In one embodiment, single-crossover homologous recombination is used between a circular plasmid or vector and the host cell genome in order to loop-in the circular plasmid or vector such as depicted in Figure 3. The inserted vector can be designed with a sequence which is a direct repeat of an existing or introduced nearby host sequence, such that the direct repeats flank the region of DNA slated for looping and deletion. Once inserted, cells containing the loop out plasmid or vector can be counter selected for deletion of the selection region.

[0515] Persons having skill in the art will recognize that the description of the loop-out procedure represents but one illustrative method for deleting unwanted regions from a genome. Indeed the methods of the present disclosure are compatible with any method for genome deletions, including but not limited to gene editing via CRISPR, TALENS, FOK, or other endonucleases. Persons skilled in the art will also recognize the ability to replace unwanted regions of the genome via homologous recombination techniques.
EXAMPLES

[0516] The following examples are given for the purpose of illustrating various embodiments of the disclosure and are not meant to limit the present disclosure in any fashion. Changes therein and other uses which are encompassed within the spirit of the disclosure, as defined by the scope of the claims, will be recognized by those skilled in the art.

[0517] A brief table of contents is provided below solely for the purpose of assisting the reader. Nothing in this table of contents is meant to limit the scope of the examples or disclosure of the application.

**Table 4** Table of Contents For Example Section.

<table>
<thead>
<tr>
<th>Example #</th>
<th>Title</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance in <em>Saccharopolyspora spinosa</em></td>
<td>Describes an implementation of transposon mutagenesis techniques for improving the performance of a <em>Saccharopolyspora spinosa</em> strain producing spinosyns</td>
</tr>
<tr>
<td>2</td>
<td>HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance in <em>Escherichia coli</em></td>
<td>Describes an implementation of transposon mutagenesis techniques for creating a transposon mutagenesis library in <em>E. coli</em></td>
</tr>
</tbody>
</table>
Example 1 – HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance in *Saccharopolyspora*

[0518] This example describes a method to produce strain libraries by *in vivo* transposon mutagenesis in *S. spinosa*. Resulting libraries can be screened to identify strains that exhibit improved phenotypes (e.g. titer of a specific compound, such as spinosyn).

[0519] Strains can be further used in rounds of cyclical engineering or to decipher genotypes that contribute to strain performance. Strains in the library can also be used for consolidation with other strains having different genetic perturbations for creation of improved strains having increased production of one or more desired compounds.

[0520] Thus, the present disclosure describes a method of using an EZ-Tn5 Transposome system (Epicenter Bio) in *S. spinosa* to create a transposon mutagenesis microbial strain library. The transposase enzyme is first complexed with a DNA payload sequence flanked by mosaic element (ME) sequences and the resulting protein-DNA complex is then transformed into cells. This results in the random integration of the DNA payload into the organism’s genomic DNA.

[0521] Depending on the payload to be introduced, either Loss-of-Function (LoF) libraries or Gain-of-Function (GoF) libraries can be produced.

[0522] Loss-of-Function (LoF) transposon libraries – The sequence of the payload may be varied to elicit diverse phenotypic responses. In the basal case of a loss-of-function (LoF) library, this payload comprises a marker that allows for the selection of successful transposon integration events.

[0523] Random loss-of-function mutations can be made *in vivo* in a microorganism using an Tn5 transposase system (EZ-Tn5; EpiCentre®) to create a transposon mutagenesis library. The EZ-Tn5 transposase system is stable and can be introduced into living microorganisms by electroporation. Once in the cell, the transposon system is activated by Mg2+ in the host cell and the transposon is randomly inserted into the host’s genomic DNA.

[0524] Gain-of-Function (GoF) transposon libraries – To create GoF libraries, more complex incarnations of the genetic payload build upon the basal case, by incorporating additional features such as, for example, promoter elements or solubility tags (in this case, called Gain-of-Function solubility tag transposon), and counter-selectable markers to facilitate loop-out of a portion of the
payload containing the selectable marker, thus allowing serial transposon mutagenesis (in this case, called Gain-of-Function recyclable transposon). Together these implementations enable the creation of diverse libraries to improve a host phenotype.

[0525] Non-limiting exemplary constructs for transposons of the present disclosure are shown in Figure 25, and the sequences of representative Loss-of-Function (LoF) transposon, Gain-of-Function (GoF) transposon, Gain-of-Function recyclable transposon, and Gain-of-Function solubility tag transposon are provided as SEQ ID NO: 17, SEQ ID NO: 18, SEQ ID NO: 19, and SEQ ID NO: 20, respectively.

[0526] These transposons can be complexed with transposase and transformed into cells. The resulting cells will have random integration of the DNA payload, thus forming transposon mutagenesis microbial strain libraries. The libraries can be further screened according to the HTP procedure described herein and evaluated for phenotype improvements. Strains with desired phenotypes, due to the transposon integration, can be isolated for further characterization and further engineering, according to any method described in the present disclosure.

[0527] For example, LoF transposon libraries and GoF transposon libraries can be screened against the parent strains, and the performance data (titer of spinosyn) can be analyzed. Some of the new strains created in these libraries will have improved performance compared to the parent strain.

[0528] Methods described herein solve two main problems. First, even in a well studied organism, large portions of the genomic landscape remain poorly understood. It has also been noted that well-understood genetic elements may interact in unexpected ways. To this end, the present disclosure provides effective genetic engineering method for elicitation of phenotypic perturbations. Second, with slow growing or genetically recalcitrant organisms- especially those with large genomes- it maybe be time or cost prohibitive to perform targeted genetic perturbations on all possible genetic targets. The present disclosure provides an effective way to create strains with perturbed genome, which may lead to improved performance in producing a desired compound in the strain. Thus, the present disclosure addresses these problems, by a method for readily and randomly modulating genetic elements of host organisms using in vivo transposon mutagenesis. In this manner, strain libraries that harbor different mutations (gain-of-function and loss-of-function) can be made very quickly and can implicate new genetic targets to further improve a host’s phenotype.
Example 2 – HTP Genomic Engineering – Implementation of a Transposon Mutagenesis Library to Improve Strain Performance in *Escherichia coli*

[0529] Transposon mutagenesis may be performed to generate *E. coli* random strain libraries to improve strains. These strain libraries can be screened against a desired phenotype, such as tryptophan yield, to identify mutants with improved performance.

[0530] *E. coli* mutant libraries may be generated by applying the EZ-Tn5 transposon system. Briefly, the EZ-Tn5 transposase is incubated with payload DNA flanked by mosaic element sequences to complex EZ-Tn5 transposase with the DNA to form a transposome. The DNA/protein transposome complex is then transformed into *E. coli* through electroporation, and the EZ-Tn5 transposase catalyzes the random integration of the payload DNA into the *E. coli* genome, thus giving rise to a random library of strain variants.

[0531] The specific sequence of the payload DNA can further be varied to bias toward either loss of function (LoF) or gain of function (GoF) effects of the transposon insertion into the target genome. Loss of function can be accomplished through inclusion of an antibiotic selection marker in the DNA payload. The antibiotic marker allows for the selection of cells with a productive transposon insertion. The insertion of the DNA payload may disrupt the function of DNA into which it is inserted in various ways, including but not limited to disruption of an open reading frame that prevents translation of the disrupted gene.

[0532] Gain of function can be accomplished through the inclusion of an antibiotic marker and a strong promoter in the DNA payload. The antibiotic marker allows for the selection of cells with a productive transposon insertion. The insertion of the DNA payload may increase the expression of genes proximal to the insertion site through the action of the strong promoter.

[0533] Either loss of function or gain of function DNA payloads may further contain a counterselection marker in addition to a selection marker to enable marker recycling and thus further rounds of engineering.

[0534] The library of strain variants generated through this transposon mutagenesis can be screened against a desired phenotype. Strains can be cultivated and tested in high throughput to identify strains with an improved desired phenotype relative to the parent strain.
[0535] The improved stain variants can be subjected to additional rounds of cyclical engineering to further improve the desired phenotype (e.g. tryptophan yield). The additional rounds of engineering may consist of transposon mutagenesis or other library types described herein such as SNP Swap, PRO Swap, or random mutagenesis. The improved strains may also be consolidated with other strain variants exhibiting an improved phenotype to produce a further improved strain through the additive effect of distinct beneficial mutations.

[0536] These types of transformations reduce the cost involved in building high quality libraries for screening in cyclical engineering. Transposon mutagenesis applied to \textit{E. coli} enables the production of thousands of genome wide loss of function or gain of function mutants in a single reaction. An alternative method is to laboriously construct thousands of assigned plasmids to engineer strains through single crossover homologous recombination (SCHR). Another alternative method is to construct thousands of assigned linear fragments to engineer strains through lambda red recombineering. Both of these alternative methods are expensive as they require generating a unique DNA fragment for each mutant that contains the intended payload DNA and sequence homology that directs recombination to a specific location on the target genome. Conversely, transposon mutagenesis uses a single DNA payload an diversity is generated through random integration into the target genome.
NUMBERED EMBODIMENTS OF THE DISCLOSURE

[0537] Notwithstanding the appended claims, the disclosure sets for the following numbered embodiments:

Methods of using and creating a transposon mutagenesis library:

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

2. The HTP method of genomic engineering according to embodiment 1, wherein the transposon mutagenesis, comprises providing a transposase enzyme and a DNA payload sequence.
3. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the transposase enzyme and DNA payload sequence form a transposase-DNA payload complex.

4. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the transposon mutagenesis results in random insertion of a transposon into the genome of the plurality of microbes.

5. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the transposon mutagenesis causes a Loss-of-Function (LoF) phenotype.

6. The HTP method of genomic engineering according to any one of embodiments 1-4, wherein the transposon mutagenesis causes a Gain-of-Function (GoF) phenotype.

7. The HTP method of genomic engineering according to any one of embodiments 1-4 and 6, wherein the transposon mutagenesis inserts a DNA payload sequence that contains a Gain-of-Function (GoF) element into the genome.

8. The HTP method of genomic engineering according to embodiment 7, wherein the Gain-of Function element is selected from the group consisting of a promoter, a solubility tag element, and a counter-selectable marker.

9. The HTP method of genomic engineering according to any one of embodiments 1-5, wherein the transposon mutagenesis inserts a DNA payload complex that contains a Loss-of-Function (LoF) element.

10. The HTP method of genomic engineering according to embodiment 9, wherein the Loss-of-Function element is a marker.

11. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the transposon mutagenesis comprises transforming the plurality of microbes with at least two transposase-DNA payload complexes one of which contains a
Gain-of-Function (GoF) element and one of which contains a Loss-of-Function (LoF) element.

12. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the transposon mutagenesis uses the EZ-Tn5 transposon mutagenesis system.

13. The HTP method of genomic engineering according to any of the preceding embodiments, wherein the genome is perturbed by utilizing transposon mutagenesis and at least one of SNP swap, Promoter swap, Stop swap, sequence optimization, or any combination thereof.

14. A method for generating a transposon mutagenesis microbial strain library, comprising
   a) introducing a transposon into a population of microbial cells of one or more base microbial strains; and
   b) selecting for at least one microbial strain comprising a randomly integrated transposon, thereby creating an initial transposon mutagenesis microbial strain library, comprising a plurality of individual microbial strains with unique genetic variations found within each strain of the plurality of individual strains, wherein each of the unique genetic variations comprises one or more randomly integrated transposons.

15. The method of embodiment 14, further comprising:
   c) selecting a strain from the transposon mutagenesis microbial strain library that exhibits an increase in performance of a measured phenotypic variable compared to the phenotypic performance of the base microbial strain.

16. The method of any of embodiments 14-15, wherein the transposon is introduced into the base microbial strain using a complex of transposon and transposase protein which allows for in vivo transposition of the transposon into the genome of the base microbial strain.
17. The method of any of embodiments 14-16, wherein the transposase protein is derived from an EZ-Tn5 transposome system.

18. The method of any of embodiments 14-17, wherein the transposon is a Loss-of-Function (LoF) transposon or a Gain-of-Function (GoF) transposon.

19. The method of embodiment 18, wherein the Loss-of-Function transposon comprises a marker.

20. The method of embodiment 19, wherein the marker is a counter-selectable marker.

21. The method of embodiment 18, wherein the Gain-of-Function transposon comprises a solubility tag, a promoter, or a counter-selection marker.

22. A HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain, comprising the steps of:
   a. engineering the genome of a base microbial strain by transposon mutagenesis, to thereby create an initial transposon mutagenesis microbial strain library comprising a plurality of individual strains with unique genetic variations found within each strain of the plurality of individual strains, wherein each of the unique genetic variations comprises one or more transposons;
   b. screening and selecting individual microbial strains of the initial transposon mutagenesis microbial strain library for phenotypic performance improvements over a reference strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
   c. providing a subsequent plurality of microbial strains that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual strains screened in the preceding step, to thereby create a subsequent transposon mutagenesis microbial strain library;
   d. screening and selecting individual strains of the subsequent transposon mutagenesis microbial strain library for phenotypic performance improvements over the reference microbial strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
e. repeating steps c)-d) one or more times, in a linear or non-linear fashion, until a strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production microbial strain, wherein each subsequent iteration creates a new transposon mutagenesis microbial strain library, where each microbial strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual microbial strains of a preceding library.

23. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 22, wherein the subsequent transposon mutagenesis microbial strain library is a partial combinatorial library of the initial transposon mutagenesis microbial strain library.

24. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 22, wherein the subsequent transposon mutagenesis microbial strain library is a subset of a full combinatorial library of the initial transposon mutagenesis microbial strain library.

25. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 22 or embodiment 23, wherein the subsequent transposon mutagenesis microbial strain library is a partial combinatorial library of a preceding transposon mutagenesis microbial strain library.

26. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 22 or embodiment 24, wherein the subsequent transposon mutagenesis microbial strain library is a subset of a full combinatorial library of a preceding transposon mutagenesis microbial strain library.

27. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to any of embodiments 22-26, wherein steps c)-d) are repeated until the phenotypic performance of a microbial strain of a subsequent transposon mutagenesis microbial strain library exhibits at least a 10% increase in a
measured phenotypic variable compared to the phenotypic performance of the production microbial strain.

28. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to any of embodiments 22-27, wherein steps c)-d) are repeated until the phenotypic performance of a microbial strain of a subsequent transposon mutagenesis microbial strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production microbial strain.

29. The HTP transposon mutagenesis method for improving the phenotypic performance of a production strain according to any of embodiments 22-28, wherein the improved phenotypic performance of step e) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, increased or more efficient production of a product of interest, the product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic acid, synthetic compound, fuel, alcohol, primary extracellular metabolite, secondary extracellular metabolite, intracellular component molecule, and combinations thereof.

30. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to any of embodiments 22-29, wherein the transposon is a Loss-of-Function (LoF) transposon or a Gain-of-Function (GoF) transposon.

31. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 30, wherein the Loss-of-Function transposon contains a marker or a counter-selectable marker.

32. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to embodiment 30, wherein the Gain-of-Function transposon contains a promoter, a solubility tag, or a counter-selectable marker.
33. The HTP method of genomic engineering according to embodiment 9, wherein the marker is a counter-selectable marker.

[0538] The aforementioned methods in the numbered embodiments can be carried out in prokaryotes or eukaryotes. For example, the methods can be conducted in a host cell from the following genus: Agrobacterium, Alicyclobacillus, Anabaena, Anacystis, Acinetobacter, Acidothermus, Arthrobacter, Azobacter, Bacillus, Bifidobacterium, Brevibacterium, Butyribrio, Buchnera, Campestris, Camplyobacter, Clostridium, Corynebacterium, Chromatium, Coprococcus, Escherichia, Enterococcus, Enterobacter, Erwinia, Fusobacterium, Faecalibacterium, Francisella, Flavobacterium, Geobacillus, Haemophilus, Helicobacter, Klebsiella, Lactobacillus, Lactococcus, Ilyobacter, Micrococcus, Microbacterium, Mesorhizobium, Methylobacterium, Methylobacterium, Mycobacterium, Neisseria, Pantoea, Pseudomonas, Prochlorococcus, Rhodobacter, Rhodopseudomonas, Rhodopseudomonas, Roseburia, Rhodospirillum, Rhodococcus, Scenedesmus, Streptomyces, Streptococcus, Synecoccus, Saccharomonospora, Saccharopolyspora, Staphylococcus, Serratia, Salmonella, Shigella, Thermoanaerobacterium, Tropheryma, Tularensis, Temecula, Thermosynechococcus, Thermococcus, Ureaplasma, Xanthomonas, Xylella, Yersinia, and Zymomonas.

Table 5 – Sequences of the Disclosure

<table>
<thead>
<tr>
<th>SEQ ID NO:</th>
<th>Description</th>
<th>SEQ ID NO:</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Expression promoter derived from Peg0007_lib_39</td>
<td>11</td>
<td>cg0371 Terminator</td>
</tr>
<tr>
<td>2</td>
<td>Expression promoter derived from Peg0007</td>
<td>12</td>
<td>cg0480 Terminator</td>
</tr>
<tr>
<td>3</td>
<td>Expression promoter derived from Peg1860</td>
<td>13</td>
<td>cg0494 Terminator</td>
</tr>
<tr>
<td>4</td>
<td>Expression promoter derived from Peg0755</td>
<td>14</td>
<td>cg0564 Terminator</td>
</tr>
<tr>
<td>5</td>
<td>Expression promoter derived from Peg0007_265</td>
<td>15</td>
<td>cg0610 Terminator</td>
</tr>
<tr>
<td></td>
<td>Expression promoter derived from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6</td>
<td>Peg3381</td>
<td>16</td>
<td>cg0695 Terminator</td>
</tr>
<tr>
<td>7</td>
<td>Peg0007_119</td>
<td>17</td>
<td>Loss-of-Function (LoF) transposon</td>
</tr>
<tr>
<td>8</td>
<td>Peg3121</td>
<td>18</td>
<td>Gain-of-Function (GoF) transposon</td>
</tr>
<tr>
<td>9</td>
<td>cg0001 Terminator</td>
<td>19</td>
<td>Gain-of-Function recyclable transposon</td>
</tr>
<tr>
<td>10</td>
<td>cg0007 Terminator</td>
<td>20</td>
<td>Gain-of-Function solubility tag transposon</td>
</tr>
</tbody>
</table>
INeorporation By Reference

[0539] All references, articles, publications, patents, patent publications, and patent applications cited herein are incorporated by reference in their entireties for all purposes. However, mention of any reference, article, publication, patent, patent publication, and patent application cited herein is not, and should not be taken as an acknowledgment or any form of suggestion that they constitute valid prior art or form part of the common general knowledge in any country in the world.

CLAIMS

What is claimed is:

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

2. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis, comprises: providing a transposase enzyme and a DNA payload sequence.

3. The HTP method of genomic engineering according to claim 2, wherein the transposase enzyme and DNA payload sequence form a transposase-DNA payload complex.
4. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis results in random insertion of a transposon into the genome of the plurality of microbes.

5. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis causes a Loss-of-Function (LoF) phenotype.

6. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis causes a Gain-of-Function (GoF) phenotype.

7. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis inserts a DNA payload sequence that contains a Gain-of-Function (GoF) element into the genome.

8. The HTP method of genomic engineering according to claim 7, wherein the Gain-of Function element is selected from the group consisting of a promoter, a solubility tag element, and a counter-selectable marker.

9. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis inserts a DNA payload complex that contains a Loss-of-Function (LoF) element.

10. The HTP method of genomic engineering according to claim 9, wherein the Loss-of-Function element is a marker.

11. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis comprises transforming the plurality of microbes with at least two transposase-DNA payload complexes one of which contains a Gain-of-Function (GoF) element and one of which contains a Loss-of-Function (LoF) element.

12. The HTP method of genomic engineering according to claim 1, wherein the transposon mutagenesis uses the EZ-Tn5 transposon mutagenesis system.
13. The HTP method of genomic engineering according to claim 1, wherein the genome is perturbed by utilizing transposon mutagenesis and at least one of SNP swap, Promoter swap, Stop swap, sequence optimization, or any combination thereof.

14. The HTP method of genomic engineering according to claim 1, wherein the microbe is a prokaryote.

15. The HTP method of genomic engineering according to claim 1, wherein the microbe is from a genus selected from the group consisting of: Agrobacterium, Alicyclobacillus, Anabaena, Anaerostici, Acinetobacter, Acidothermus, Arthrobacter, Azobacter, Bacillus, Bifidobacterium, Brevibacterium, Butyribrio, Buchnera, Campestris, Campylobacter, Clostridium, Corynebacterium, Chromatium, Coprococcus, Escherichia, Enterococcus, Enterobacter, Erwinia, Fusobacterium, Faecalibacterium, Francisella, Flavobacterium, Geobacillus, Haemophilus, Helicobacter, Klebsiella, Lactobacillus, Lactococcus, Ilyobacter, Micrococcus, Microbacterium, Mesorhizobium, Methylobacterium, Mycobacterium, Mycoplasma, Pantoea, Pseudomonas, Prochlorococcus, Rhodobacter, Rhodopseudomonas, Rhodopseudomonas, Roseburia, Rhodospirillum, Rhodococcus, Scenedesmus, Streptomyces, Streptococcus, Synebacter, Saccharomonospora, Saccharopolyspora, Staphylococcus, Serratia, Salmonella, Shigella, Thermoanaerobacterium, Tropheryma, Tularensis, Temecula, Thermosynechococcus, Thermococcus, Ureaplasma, Xanthomonas, Xylella, Yersinia, and Zymomonas.

16. The HTP method of genomic engineering according to claim 1, wherein the microbe is *Saccharopolyspora spinosa*.

17. The HTP method of genomic engineering according to claim 1, wherein the microbe is *Escherichia coli*.
18. The HTP method of genomic engineering according to claim 1, wherein the microbe is a eukaryote.

19. A method for generating a transposon mutagenesis microbial strain library, comprising:
   a) introducing a transposon into a population of microbial cells of one or more base microbial strains; and
   b) selecting for at least one microbial strain comprising a randomly integrated transposon, thereby creating an initial transposon mutagenesis microbial strain library, comprising a plurality of individual microbial strains with unique genetic variations found within each strain of the plurality of individual strains, wherein each of the unique genetic variations comprises one or more randomly integrated transposons.

20. The method of claim 19, further comprising:
   c) selecting a strain from the transposon mutagenesis microbial strain library that exhibits an increase in performance of a measured phenotypic variable compared to the phenotypic performance of the base microbial strain.

21. The method of claim 19, wherein the transposon is introduced into the base microbial strain using a complex of transposon and transposase protein which allows for in vivo transposition of the transposon into the genome of the base microbial strain.

22. The method of claim 19, wherein the transposase protein is derived from an EZ-Tn5 transposome system.

23. The method of claim 19, wherein the transposon is a Loss-of-Function (LoF) transposon or a Gain-of-Function (GoF) transposon.

24. The method of claim 23, wherein the Loss-of-Function transposon comprises a marker.

25. The method of claim 24, wherein the marker is a counter-selectable marker.
26. The method of claim 23, wherein the Gain-of-Function transposon comprises a solubility tag, a promoter, or a counter-selection marker.

27. The method of claim 19, wherein the microbial strain is a prokaryote.


29. The method of claim 19, wherein the microbial strain is *Saccharopolyspora spinosa*.

30. The method of claim 19, wherein the microbial strain is *Escherichia coli*.

31. The method of claim 19, wherein the microbial strain is a eukaryote.

32. A HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain, comprising the steps of:
   a. engineering the genome of a base microbial strain by transposon mutagenesis, to thereby create an initial transposon mutagenesis microbial strain library comprising a
plurality of individual strains with unique genetic variations found within each strain of the plurality of individual strains, wherein each of the unique genetic variations comprises one or more transposons;
b. screening and selecting individual microbial strains of the initial transposon mutagenesis microbial strain library for phenotypic performance improvements over a reference strain, thereby identifying unique genetic variations that confer phenotypic performance improvements;
c. providing a subsequent plurality of microbial strains that each comprise a combination of unique genetic variations from the genetic variations present in at least two individual strains screened in the preceding step, to thereby create a subsequent transposon mutagenesis microbial strain library;
d. screening and selecting individual strains of the subsequent transposon mutagenesis microbial strain library for phenotypic performance improvements over the reference microbial strain, thereby identifying unique combinations of genetic variation that confer additional phenotypic performance improvements; and
e. repeating steps c)-d) one or more times, in a linear or non-linear fashion, until a strain exhibits a desired level of improved phenotypic performance compared to the phenotypic performance of the production microbial strain, wherein each subsequent iteration creates a new transposon mutagenesis microbial strain library, where each microbial strain in the new library comprises genetic variations that are a combination of genetic variations selected from amongst at least two individual microbial strains of a preceding library.

33. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the subsequent transposon mutagenesis microbial strain library is a partial combinatorial library of the initial transposon mutagenesis microbial strain library.

34. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the subsequent transposon
mutagenesis microbial strain library is a subset of a full combinatorial library of the initial transposon mutagenesis microbial strain library.

35. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the subsequent transposon mutagenesis microbial strain library is a partial combinatorial library of a preceding transposon mutagenesis microbial strain library.

36. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the subsequent transposon mutagenesis microbial strain library is a subset of a full combinatorial library of a preceding transposon mutagenesis microbial strain library.

37. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein steps c)-d) are repeated until the phenotypic performance of a microbial strain of a subsequent transposon mutagenesis microbial strain library exhibits at least a 10% increase in a measured phenotypic variable compared to the phenotypic performance of the production microbial strain.

38. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein steps c)-d) are repeated until the phenotypic performance of a microbial strain of a subsequent transposon mutagenesis microbial strain library exhibits at least a one-fold increase in a measured phenotypic variable compared to the phenotypic performance of the production microbial strain.

39. The HTP transposon mutagenesis method for improving the phenotypic performance of a production strain according to claim 32, wherein the improved phenotypic performance of step e) is selected from the group consisting of: volumetric productivity of a product of interest, specific productivity of a product of interest, yield of a product of interest, titer of a product of interest, increased or more efficient production of a product of interest, the product of interest selected from the group consisting of: a small molecule, enzyme, peptide, amino acid, organic
40. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the transposon is a Loss-of-Function (LoF) transposon or a Gain-of-Function (GoF) transposon.

41. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 40, wherein the Loss-of-Function transposon contains a marker or a counter-selectable marker.

42. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 40, wherein the Gain-of-Function transposon contains a promoter, a solubility tag, or a counter-selectable marker.

43. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the production microbial strain is a prokaryote.

44. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the production microbial strain is from a genus selected from the group consisting of: Agrobacterium, Alicyclobacillus, Anabaena, Anacystis, Acinetobacter, Acidothermus, Arthrobacter, Azobacter, Bacillus, Bifidobacterium, Brevibacterium, Butyribrio, Buchnera, Campestris, Camphylobacter, Clostridium, Corynebacterium, Chromatium, Coprococcus, Escherichia, Enterococcus, Enterobacter, Erwinia, Fusobacterium, Faecalibacterium, Francisella, Flavobacterium, Geobacillus, Haemophilus, Helicobacter, Klebsiella, Lactobacillus, Lactococcus, Ilyobacter, Micrococcus, Microbacterium, Mesorhizobium, Methylobacterium, Methylbacterium, Mycobacterium, Neisseria, Pantoea, Pseudomonas, Prochlorococcus, Rhodobacter, Rhodopseudomonas, Rhodospseudomonas, Roseburia, Rhodospirillum, Rhodococcus, Scenedesmus, Streptomyces, Streptococcus, Syneccocus, Saccharomonospora,
Saccharopolyspora, Staphylococcus, Serratia, Salmonella, Shigella, Thermoanaerobacterium, Tropheryma, Tularensis, Temecula, Thermosynechococcus, Thermococcus, Ureaplasma, Xanthomonas, Xylella, Yersinia, and Zymomonas.

45. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the production microbial strain is \textit{Saccharopolyspora spinosa}.

46. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the production microbial strain is \textit{Escherichia coli}.

47. The HTP transposon mutagenesis method for improving the phenotypic performance of a production microbial strain according to claim 32, wherein the production microbial strain is a eukaryote.

48. The HTP method of genomic engineering according to claim 9, wherein the marker is a counter-selectable marker.
FIGURE 1

Existing Gene Variants

Variant 1  Variant 2  Variant 3

Fragmented DNA

Chimeric Gene Variants
FIGURE 2

1. Generate and QC SNP Specific DNA
2. Assemble and QC SNPs and plasmid
3. Transform strain
4. Loop-out marker and QC final strain

100 SNPs to swap
FIGURE 3

1. Direct Repeat L
2. tDNAR
3. Insert Sequence
4. Genomic DNA
5. tDNAR
6. Transformation and Selection
7. Inserted DNA
8. Generate Insert DNA
9. Assemble Plasmid
10. Integrated into Target Genome
FIGURE 4A

Step 1 Build Vectors for Transformation - Instructions based on data from Step 3 or 4

Acquire and Build DNA

- Source DNA fragments from vendors or internal storage.
- Alternatively synthesize DNA fragments in house.
- Join smaller DNA fragments into larger inserted for vector cloning.
- Evaluate joined DNA inserts prior to cloning.

Assemble Vectors

- Stitch together DNA inserts into vectors for genomic insertion into host organisms.
- Evaluate vectors prior to transformation.

Go to 2

Step 2 Strain Transformation - Vectors from Step 1

Transform Vectors into Host Cell

- Transform host cell with DNA Vectors.
- Select for transformed cells and evaluate cassette insertion into genomic DNA.

Remove Selection Markers

- Loop out selection markers through counter selection.
- Evaluate insertion site to confirm deletion of loop out regions.

Go to 3
### FIGURE 4B

#### Step 3 Strain Evaluation in Multi-Well Plate - Test Strains from Step 2

<table>
<thead>
<tr>
<th>Culture Strains on Plate</th>
<th>Generate Product on Plate</th>
<th>Evaluate Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Transfer strains onto culture plate with growth media.</td>
<td>• Transfer aliquots from culture plate to production plate based on actual culture concentrations.</td>
<td>• Perform one or more assays to test performance (i.e. product concentration).</td>
</tr>
<tr>
<td>• Allow cultures to grow to predetermined density (e.g., OD at 600nm).</td>
<td>• Production plates include media and conditions designed for product generation.</td>
<td>• Real time evaluation possible.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Store selected cultures in cold storage.</td>
</tr>
</tbody>
</table>

#### Step 4 Strain Evaluation in Tank - Test Top Candidates from Step 3

<table>
<thead>
<tr>
<th>Culture Strains on Flasks</th>
<th>Generate Product in Tank</th>
<th>Evaluate Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Transfer strains into culture flask with growth media.</td>
<td>• Transfer aliquots from culture flask to production tank based on actual culture concentrations.</td>
<td>• Perform one or more assays to test performance (i.e. product concentration).</td>
</tr>
<tr>
<td>• Allow cultures to grow to predetermined density (e.g., OD at 600nm).</td>
<td>• Production tanks include media and conditions designed for product generation.</td>
<td>• Evaluation typically at mid-point and end-point of cultures.</td>
</tr>
<tr>
<td></td>
<td>• Production tanks may also include initial biomass growth period.</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 7

[Heatmap diagram showing correlations between SNPs and other variables such as pyc, lysA, zwf, etc.]

SNP 309
SNP 84
SNP 34
SNP 9
SNP Z
SNP 316
SNP 125
SNP 56
SNP 33
pyc
lysA
zwf

Correlation values range from 0.90 to -0.15.
FIGURE 9B

SNP Distribution

Total

SNPs unique to B

SNPs unique to C

SNPs in both B and C

332

46

153

133

Strain relationship

FIGURE 9A

C

B

A
FIGURE 11
Pathway Targets

PTS  ask  dddh
zwf  asd  lysA
dpi  dapA  lysE
tkt  dapB  hom
fhp  dapD  odx
ppc  cg0361  pck
pyc  dapE  icd
aspB  dapF

Zymergen Promoter Ladder

Multiple starting strains

Arbitrary expression units

PROswap Library

Combinatorialize successful hits
## FIGURE 15

### I. Build DNA

<table>
<thead>
<tr>
<th>Acquire and build DNA pieces</th>
<th>QC DNA parts</th>
<th>Generate DNA assembly</th>
<th>QC DNA assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source short DNA pieces from outside vendors and internal storage, and use to build parts to be assembled into larger DNA constructs</td>
<td>Evaluate DNA parts for quality by confirming size and/or sequence</td>
<td>Stitch together DNA parts using various methods to form assemblies</td>
<td>Evaluate assemblies for quality based on size and/or sequence</td>
</tr>
</tbody>
</table>

### II. Build Strain

<table>
<thead>
<tr>
<th>Prepare base strain and DNA assembly</th>
<th>Transform DNA into base strain</th>
<th>Integrate DNA into genome of base strain</th>
<th>QC transformed strain</th>
</tr>
</thead>
</table>
| Make the base strain competent to receive assembled DNA by growing them under special environmental conditions | Create holes in the cell membrane so that the assembled DNA can pass through it (e.g. electroporation) | Consists of 2 stages:  
  - Loop in: Incorporate entire DNA assembly into the genome  
  - Loop out: Remove extra pieces of the DNA assembly from the genome, leaving only the targeted genetic change behind | Confirm that the desired genetic changes are incorporated in base strain by using sizing and sequencing methods; Then bank transformed and QC’d strains (culture) in the freezer for transfer to Test |
III. Test in plate

Select and consolidate QC'd strains into test plate
- Transfer strains from liquid cultures or colony cultures into seed plates

Culture strains in seed plates
- Build strain biomass by shaking seed plates in culture media at an appropriate temperature

Generate product from strain
- Inoculate multiple culture conditions that model production conditions with seed cultures and grow cultures for 24-96 hours

Evaluate performance
- Perform one or more assays on samples from production media to test performance of the engineered strain (e.g. product concentration) then store in cold storage

IV. Test in tank

Acquire strains for evaluation
- Retrieve seed bank of strain from cold storage

Culture strains in flasks
- Grow in flasks and/or tanks to build biomass in seed media

Generate product from strain
- Generate product in feedback-controlled production environment inoculated with biomass from seed

Evaluate performance
- Process mid-point and end-point samples to evaluate performance
FIGURE 22

Generate training set
3302

Develop initial model
3304

Generate design candidate strains
3306

Predict expected performance
3308

Select strains with best predicted performance
3310

Manufacture
3312

Incorporate selected strains into model
3316

Select highest performing strains
3314
FIGURE 23

I/O 802

PROCESSOR 804

PERIPHERAL DEVICES 806

COMMUNICATION INTERFACES 807

MAIN MEMORY 808

SECONDARY MEMORY 810
FIGURE 24

Obtain measured performance data for mutations

Measure degree of similarity in performance among mutations

Select mutations having sufficiently dissimilar performance

Design microbial strain having selected mutations
FIGURE 25
Transposon

Loss of Function Transposon - example diagram

Gain of Function Transposon - example diagram

Gain of Function Recyclable Transposon - example diagram

‡ Selectable marker is modular and may be varied depending on the host organism
* GoF element is modular and may comprise a solubility tag, promoter, counter-selection, etc.
† These regions should be direct repeats to allow for recombination and loop-out
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

INV. C12N15/10
ADD.

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

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)

EPO-Internal, BIOSIS, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

Y

-----
---/---

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
  *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

*"Z" document member of the same patent family

Date of the actual completion of the international search
29 August 2018

Date of mailing of the international search report
11/09/2018

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax. (+31-70) 340-3016

Authorized officer
Weinberg, Suzanna

Form PCT/ISA210 (second sheet) (April 2005)
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>WO 02/29032 A2 (DIVERSA CORP [US]; SHORT JAY M [US]; FU PENGCHENG [US]; LATTERICH MART) 11 April 2002 (2002-04-11) the whole document In particular Figures 18, 21 and 22</td>
<td>1-48</td>
</tr>
<tr>
<td>A</td>
<td>WO 99/46387 A1 (DOW AGROSCIENCES LLC [US]; BALTZ RICHARD H [US]; BROUGHTON M CHRISTINE) 16 September 1999 (1999-09-16) the whole document</td>
<td>29,45</td>
</tr>
<tr>
<td></td>
<td>-/--</td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>Citation of document, with indication, where appropriate, of the relevant passages</td>
<td>Relevant to claim No.</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>A</td>
<td>JP 2003 061669 A (NAT INST OF ADV IND &amp; TECHNOL; JAPAN COOPERATION CT PETROLEUM) 4 March 2003 (2003-03-04) the whole document</td>
<td>1-48</td>
</tr>
<tr>
<td>Patent document cited in search report</td>
<td>Publication date</td>
<td>Patent family member(s)</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BR 112015010138 A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2890104 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CN 104903446 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 2917346 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2015533289 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 20150083895 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RU 2015119379 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SG 10201703558X A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SG 112015032331 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2015307873 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2018135041 A</td>
</tr>
<tr>
<td>WO 0229032 A2</td>
<td>11-04-2002</td>
<td>NONE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CN 108027849 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 20180012280 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2017100377 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 743305 B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2276064 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69835360 T2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DK 1717322 T3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 1717322 A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 2261373 A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES 2270505 T3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 4062366 B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2001508662 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 20000070258 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KR 20050043983 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6251674 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6352859 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6379964 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 380872 T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU 764737 B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BR 9909257 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2322449 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CN 1298447 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69937732 T2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 1062345 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES 2299240 T3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID 26682 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IL 1388241 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2002505881 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NZ 506624 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TW 1237058 B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6143526 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6274350 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6521406 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2004023343 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 9946387 A1</td>
</tr>
<tr>
<td>Patent document cited in search report</td>
<td>Publication date</td>
<td>Patent family member(s)</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>ZA</td>
<td></td>
<td>9901837 B</td>
</tr>
<tr>
<td>JP</td>
<td></td>
<td>JP 2003061669 A</td>
</tr>
</tbody>
</table>