TITLE

Ultra-high throughput mapping of genetic design space

AUTHORS

Ronan W. O'Connell[1,3,*], Kshitij Rai[1,4,*], Trenton C. Piepergerdes[1,3], Kian D. Samra[1], Jack A. Wilson[1], Shujian Lin[1], Thomas H. Zhang[1], Eduardo M. Ramos[1], Andrew Sun[1], Bryce Kille[5], Kristen D. Curry[5], Jason W. Rocks[6], Todd J. Treangen[5], Pankaj Mehta[6,7,8], and Caleb J. Bashor[1,2,§]

AFFILIATIONS

[1]Department of Bioengineering, Rice University, Houston, TX 77030, USA
[2]Department of Biosciences, Rice University, Houston, TX 77030, USA
[3]Graduate Program in Bioengineering, Rice University, Houston, TX 77030, USA
[4]Graduate Program in Systems, Synthetic and Physical Biology, Rice University, Houston, TX 77030, USA
[5]Department of Computer Science, Rice University, Houston, TX 77030, USA
[6]Department of Physics, Boston University, Boston, MA 02215, USA
[7]Biological Design Center, Boston University, Boston, MA 02215, USA
[8]Faculty of Computing and Data Science, Boston University, Boston, MA 02215, USA
[§]Correspondence. Email: caleb.bashor@rice.edu
*These authors contributed equally to this work.

ABSTRACT

Massively parallel genetic screens have been used to map sequence-to-function relationships for a variety of genetic elements. However, because these approaches only interrogate short sequences, it remains challenging to perform high throughput (HT) assays on constructs containing combinations of sequence elements arranged across multi-kb length scales. Overcoming this barrier could accelerate synthetic biology; by screening diverse gene circuit designs, "composition-to-function" mappings could be created that reveal genetic part composability rules and enable rapid identification of behavior-optimized variants. Here, we introduce CLASSIC, a generalizable genetic screening platform that combines long- and short-read next-generation sequencing (NGS) modalities to quantitatively assess pooled libraries of DNA constructs of arbitrary length. We show that CLASSIC can measure expression profiles of >10^5 drug-inducible gene circuit designs (ranging from 6-9 kb) in a single experiment in human cells. Using statistical inference and machine learning (ML) approaches, we demonstrate that data obtained with CLASSIC enables predictive modeling of an entire circuit design landscape, offering critical insight into underlying design principles. Our work shows that by expanding the throughput and understanding gained with each design-build-test-learn (DBTL) cycle, CLASSIC dramatically augments the pace and scale of synthetic biology and establishes an experimental basis for data-driven design of complex genetic systems.

MAIN TEXT

Synthetic gene circuits are constructed by assembling DNA-encoded genetic parts into multi-gene programs that perform computational tasks in living cells[1-3]. Over the past two decades, gene circuits have emerged as important models for understanding native gene regulation[4] and have been used to create powerful biotechnologies by enabling user-defined control over cellular behavior[5-7]. Despite this progress, the design of quantitatively precise circuit behavior remains challenging. Regulatory interactions within a circuit must be carefully tuned, often through multiple iterative DBTL cycles, before part compositions that support a desired circuit behavior are identified[8]. Additionally, since genetic parts must work in close physical proximity to one another, as well as within a crowded intracellular environment[9], incidental molecular coupling can occur between parts and with host cell regulatory machinery[10-12]. Because these context-dependent interactions are difficult to predict, they can confound model-driven circuit design, further extending the number of DBTL cycles required to achieve a target behavior[13].

One potential strategy for increasing the pace of gene circuit engineering is to expand the number of circuits tested in each cycle by performing HT functional screens on pooled circuit libraries. By profiling an entire circuit design landscape in a single experiment, such an approach could enable rapid identification of circuit variants with desired behaviors and facilitate the development of data-driven models capable of inferring context-specific part function and forward predicting the behavior of novel circuit designs. HT screening approaches that utilize NGS as a readout[14-18] have been used to generate detailed sequence-to-function mappings for multiple genetic part classes, including promoters[19], terminators[20], transcription factors (TFs)[21,22], nucleic acid switches[23], and

receptors[24]. However, the ability to functionally profile libraries of DNA constructs long enough (>1kb) to encode entire circuits is limited by current NGS technology; short-read platforms (e.g., Illumina) can generate high-depth data, but their read length is limited to DNA fragments or PCR amplicons of <500 bp[25], while long-read platforms (e.g., nanopore) can read >1 kb but are either too expensive or too error prone[26] to achieve the depth needed to profile larger libraries.

Methods have emerged that permit multiplex analysis of long constructs by appending short barcode index sequences to facilitate amplicon-based readout by short-read NGS[27]. However, this approach is predominantly confined to array formats where each construct is assembled and barcoded separately, or to nested assembly schemes[27,28], limiting library design flexibility and assay throughput. To perform indexed multiplexing for long constructs at a greater scale, we devised an approach that leverages the strengths of both long- and short-read NGS to analyze construct libraries generated via pooled genetic part assembly (**Fig. 1A**). In our scheme, libraries are generated by pooled parts assemblies that incorporate semi-random barcodes; a composition-to-barcode index is then created using long-read (nanopore) sequencing. The library is then introduced into cells, binned based on expression phenotype, and analyzed by short-read (Illumina) amplicon sequencing to produce a barcode-to-phenotype index. A map that matches construct composition to phenotype can then be revealed by comparing the two indices. Using this technique, which we refer to as CLASSIC (combining long- and short-range sequencing to investigate genetic complexity), it is possible to obtain high-depth phenotypic expression data for large libraries of DNA constructs of arbitrary length using standard phenotypic selection or flow sorting experiments. Further, due to the random

assignment of barcodes to assembled constructs, each variant in a CLASSIC library is associated with multiple unique barcodes that generate independent phenotypic measurements, leading to greater accuracy than a one-to-one construct-to-barcode library.

In order to configure CLASSIC to quantitatively profile libraries of gene circuits with diverse part combinations in human cells (**Fig. 1B)**, we devised a custom hierarchical golden-gate[29] cloning scheme in which library diversity is programmed through a series of pooled DNA assembly steps**.** Diversified pools of 5', coding, and 3' gene elements are first generated through assembly of input part fragments (level 0 to 1), and subsequently combined to create single-gene expression unit (EU) pools (level 1 to 2). EU pools are then combined with a plasmid-encoded barcode pool to yield multi-EU circuit libraries (level 2 to 3). Following nanopore sequencing and data analysis (see **Methods**), circuit libraries are genomically integrated at single copy into a HEK293T cell line harboring a custom "landing pad" cassette (HEK293T-LP)[30,31] (see **Methods**). Library-integrated cells are then flow-sorted based on circuit expression output, followed by Illumina NGS to quantitate bin distributions of circuit-associated barcodes[14,15] (see **Methods**).

To evaluate the quantitative accuracy of this approach, we constructed and measured expression for a 384-member library in which mRuby expression was modulated by variable combinations of genetic parts: a set of 8 constitutively active promoters frequently used in mammalian genetic engineering[32,33]; 6 variable Kozak sequences capable of tuning expression by modulating translational initiation rate[34]; and 8 common mammalian transcriptional terminator sequences[20,35,36] (**Fig. 2A, left**). The library was cloned by combining promoter, Kozak, and terminator input pools with a single mRuby ORF to generate a diversified EU pool, which was then combined with barcode plasmids to create the

construct index (**Fig. 2A, right**). Nanopore sequencing of the library demonstrated complete, part-balanced coverage of EU design space (**Fig. 2B, top**) with proportional assortment of unique barcodes, indicating an absence of systematic bias for both EU assembly and barcode indexing steps (**Fig. 2B, bottom**).

Following integration into HEK293T-LP cells, the EU library was flow-sorted into 10 bins based on mRuby expression (414,407 cells) and EU-associated barcodes were measured by Illumina NGS (**Fig. 2C, left**). We aggregated barcodes that mapped to the same EU to infer a mean expression and distribution for each library composition (**Fig. 2C, bottom;** see **Methods**). To compare our NGS-derived measurements to "ground-truth" expression values, we made direct measurements on 15 randomly sorted and clonally expanded library members using flow cytometry. Geometric mean values for these isolates showed excellent agreement with corresponding CLASSIC-derived values (*MAE*=0.065), with residual errors mostly falling within an experimentally determined range of clonal expression heterogeneity (*ERCH*; see **Methods**) (**Fig. 2C, right**). To validate the precision of CLASSIC, we compared data for separate sorting experiments of the same 384-member library (**Fig. 2D, top**) as well as independently integrated and sorted libraries (**Fig. 2D, bottom**). Both showed high correlation (R^2=0.99 and 0.97, respectively), demonstrating that CLASSIC is highly repeatable between both technical and biological replicates.

The EU library data allowed us to systematically analyze relative contributions of individual parts and part categories to mRuby expression magnitude; promoters showed the broadest range of expression modulation, followed by Kozaks, and then terminators (**Fig. 2E**). To test whether we could use an ML model to predict expression level based

on part composition, we trained a random forest (RF) regression model[37], which we selected due to its predictive power and ability to capture non-linear relationships, as well as its relative interpretability over other black-box models such as neural networks. We used an 80:20 train:test split with part categories as features to successfully demonstrate that the model has high accuracy (R^2=0.96) and assigns feature importance to part categories in a manner consistent with their observed effect on expression (**Fig. 2F**, bottom right). Interestingly, part compositions that were shown by the RF regression model to have the highest absolute error (>0.025) were determined to all contain the same terminator, T8, paired with strong promoters (hEF1a1, hEF1a2, and CMV) and Kozaks, indicating that pairing weak and strong transcriptional control elements can yield synergistic effects that confound expression level prediction (**Fig. 2G**). These results demonstrate that data gathered with CLASSIC can be used in conjunction with ML-based modeling to benchmark and accurately predict context-specific part function.

We next tested whether CLASSIC could be scaled to higher throughput (>10^5) to quantitatively profile a high-dimensional gene circuit design space. For this purpose, we selected a small molecule drug-inducible circuit consisting of two EUs (**Fig. 3A**): one encoding a constitutively expressed synthetic Cys2-His2 zinc-finger (ZF)-based transcription factor (synTF)[38-40] with appended transcriptional activation domains (ADs), and the other encoding a reporter gene harboring cognate synTF binding motifs (BMs) located upstream of a minimal promoter driving eGFP expression[38]. Transcriptional induction occurs upon addition of the small molecule 4-hydroxytamoxifen (4-OHT), which binds to a mutant version of the human estrogen receptor (ERT2)[41] appended to the synTF, facilitating its translocation from the cytoplasm into the nucleus to activate reporter

transcription. As we and others have demonstrated[38,42-44], identifying designs that are optimized for high fold-change (HFC) (**Fig. 3A, right**) expression can be challenging in eukaryotic systems; basal and induced expression levels must be respectively minimized and maximized through fine-tuning of both transcriptional regulatory features of the locus and the molecular properties of the synTF, thereby ensuring robust expression of a transgene exclusively in the presence of inducer.

To create a composition-to-function map for an inducible synTF circuit, we composed a 10-feature design space consisting of genetic part categories that we hypothesized could be important for achieving HFC behavior. This included 4 different TAs[45-47], 3 ZF affinities[38], and a set of 4 intrinsically disordered protein (IDP) domains[48-50] that have been shown to facilitate liquid-liquid phase condensation of nuclear-localized TFs[48,49]. We also varied transcriptional regulatory features of the circuit, including 4 promoters and 4 terminators in the synTF coding EU, as well as 3 core promoters[51-53] and 2, 4, 8, or 12 BMs in the reporter EU.. Additionally, we varied the spacing between the EUs (0, 250, or 500 bases) along with their 5'-to-3' orientation (**Fig. 3B, left**) to yield an overall design space of 165,888 compositions (**Fig. 3B, left**). This library was constructed in 3 steps by first assembling protein domain parts to create a level 1 synTF ORF pool, then conducting parallel assemblies to generate level 2 pools of synTF coding and reporter EUs, and finally combining EU and barcode pools into the level 3 destination vector (**Fig. 3B, right**).

Nanopore sequencing of the pooled library yielded barcode assignments for 95.3% of total compositions (**Figs. 3C**), with a mean of 8.4 barcodes for each circuit (**Fig. 3C, bottom left**). We integrated the library into HEK293T-LP cells and sorted un-induced and 4-OHT-induced populations (8.6 and 15.7 mil cells, respectively) separately into 8 bins

based on eGFP fluorescence (**Fig. 3D, left**). Following analysis by Illumina NGS, we matched a total of 121,292 (73% of design space) compositions to barcodes detected in both the sorted populations (mean number of barcodes per circuit=2.25) (**Fig. 3C**). We then used barcode bin distributions to compute basal, induced, and fold-change expression values for each variant. CLASSIC-derived fold-change values demonstrated excellent overall agreement (*MAE*=0.145) (**Fig. 3D, top middle**) with those of random isolates (n=40) (**Figs. 3D, bottom**), and values for both basal and induced expression demonstrated comparably high percent similarities (**Fig. 3D, top right**). These results confirm that CLASSIC retains quantitative measurement accuracy when scaled to orders-of-magnitude larger and more complex libraries.

We analyzed the distribution of CLASSIC-measured values across a 2-dimensional behavior space (basal vs. induced eGFP expression), examining the relative density of compositions in 3 regions of interest: low basal (<500 AU), high induced (>70,000 AU), and HFC expression (>25x fold-change) (**Fig. 4A**). We observed a higher proportion of library members in the low basal than in the high induced region (~4.3x), and ~50% of compositions exhibited fold-change values of <3x, while a smaller fraction of the library (~8%) fell within the HFC region. Since our data set omits 27% of our overall design space, the data suggest that over 3,000 HFC circuits are not represented, potentially limiting our ability to understand design principles for circuits in that region. Therefore, we asked whether we could create a complete mapping of design space by using an ML model to predict the behavior of unmeasured compositions. We trained two least-squares boosted RF regression models with our basal and induced data sets respectively, using randomized proportional down-sampling (see **Methods**) to normalize circuit

representation across expression bins (**Fig. 4A, middle**). Evaluation of the 80:20 train:test split indicated strong predictive power for the model, with train/test R^2 values of 0.85/0.75 for the basal and 0.84/0.81 for induced data sets. Feature importance scores were highest for part categories related to synTF function (AD, ZF affinity), the synTF expression promoter, and the reporter gene core promoter.

Using the RF models, we predicted both basal and induced expression values for the 44,596 unmeasured part compositions, and also fit values for the 121,292 measured compositions, yielding a behavior map of our entire 165,888-member design space (**Fig. 4A, right**). While a comparison between RF-modeled and CLASSIC-measured distributions revealed similar global features, we observed a high absolute model error (>2) for compositions measured at the periphery of behavior space (3% of variants, including many in the HFC region), potentially resulting from either CLASSIC measurement error or poor model prediction in these regions (**Fig. 4A, right, inset**). To validate the predictive power of our model, we LP-integrated unmeasured and measured compositions from across behavior space (**Fig. 4B**). Flow cytometry measurements of the resulting cell lines showed close agreement with both predicted values for unmeasured (*MAE*= 0.17) and measured compositions (*MAE*= 0.19), most notably for compositions with the highest predicted fold-change values in the design space (50-100x). Additionally, we demonstrated that cell lines corresponding to configurations with high-error measurements previously observed in the behavior space periphery showed close agreement with model predictions, indicating that CLASSIC measurement outliers are accurately adjusted by the model (**Fig. 4B, HFC circuits**).

With a quantitatively accurate functional mapping of synTF circuit design space in-hand, we asked whether our data set could provide insight into the design rules underlying circuit behavior. First, we compared the frequency of part usage between low basal, high induced, and HFC regions (**Fig. 4C**). We observed distinct part usage for compositions from each region, with highly asymmetric usage in specific categories, including those associated with the synTF protein (AD, IDP, and ZF affinity), the synTF expression promoter, the number of BMs, core promoter identity, and 5'-to-3' EU orientation. For example, VPR, a strong AD, is used extensively in the high induced region, but is nearly absent amongst low basal circuits in favor of the weaker AD, p65, while intermediate-strength VP16 and VP64 are enriched amongst HFC circuits. By the same token, lower and medium activity core promoters (miniTK and ybTATA) are excluded from high activity circuits, while low basal activity circuits exclude mCMV (high activity core promoter), and HFC circuits utilize a mix of ybTATA and mCMV.

We next assessed whether part usage co-varies between categories by computing cross-category mutual information (MI) for each behavior space region (**Fig. 4D**). We saw "coupling" between categories in all 3 regions, with the highest degree observed in the HFC region. The strongest interactions were between AD and the following categories: ZF affinity, core promoter, and synTF expression promoter. To better understand the role that this apparent interdependent part usage plays in optimizing HFC circuit function, we performed UMAP-assisted K-means clustering[54] on modeled HFC compositions using the 7 part categories with highest MI coupling (**Fig. 4E**). Inspection of the resulting projection revealed 2 distinct clusters that were differentiated by their specific deployment of parts: compositions in cluster A were enriched for lower affinity synTFs fused to stronger ADs

(e.g., VPR), while those in cluster B favor higher affinity synTFs, medium activity ADs (VP16, VP64) and stronger core promoters.

Our CLASSIC-enabled analysis of part composition patterns supports a set of design principles for programming HFC behavior that emphasizes a careful balance between minimizing basal expression while enhancing induced expression (**Fig. 4F**). Part usage in categories associated with circuit locus design favors reporter-to-coding EU orientation, extended (500 bp) spacer regions, and moderate synTF expression; features that are also found in the low basal expression region of behavior space and thus likely contribute to minimizing leaky reporter expression in the absence of inducer. On the other hand, high BM valency, higher-activity core promoters, and preference for terminator T7, features also found in the high induced region, potentially maximize reporter output in the presence of nuclear-localized synTF. As our MI and clustering analyses demonstrated, usage of parts comprising the synTF protein employ two parallel molecular solutions—matching strong activators with weak affinity (cluster A) or medium activators with strong affinity (cluster B)—that likely constrain specific activity of the synTF to a regime that minimizes background expression while still producing strong induction.

Here, we have established the feasibility of combining long- and short-read NGS modalities to perform massively parallel quantitative profiling of multi-kb length-scale constructs in human cells. CLASSIC holds potential as a generalizable method for exploring the emergence of function from genetic composition across a spectrum of organizational scales and phylogenetic contexts, including for viruses[55], bacterial operons[56], and chromatin domains[56-58]. As we show in this piece, by enabling HT profiling of diverse combinations of genetic parts, CLASSIC significantly expands the scope of inquiry for synthetic

biology projects; it reduces the time and cost required to identify behavior-optimized part compositions and establishes guidelines for part composability. As demonstrated by our identification and analysis of HFC circuit variants, the design rules revealed by CLASSIC may be non-intuitive and challenging to capture using biophysical modeling alone. Furthermore, since CLASSIC leverages extant, broadly-accessible molecular cloning and experimental analysis pipelines, we anticipate it will increase the pace and scale of genetic design for diverse synthetic biology applications across a range of organismal hosts, including the development of bioproduction strains[59] and multigenic cell therapy programs[5].

Finally, we showed that data acquired using CLASSIC can be used to train ML models to accurately make predictions for out-of-sample and edge-case circuit behavior. While extensive recent work has used ML approaches to develop sequence-to-function models for various classes of genetic parts[23,60-62] our work serves as a starting point for developing AI-based models of gene circuit function that use part compositions as learned features. While our current work has focused on mapping a design space of 10^5 compositions, it may be possible to create predictive models for more complex circuits with far more expansive design spaces by using data acquired with CLASSIC to train high capacity deep-learning algorithms (e.g., transformers) which require much larger datasets than currently exist. Such approaches could work in black-box fashion, without the incorporation of regulatory or biophysical priors, or synergistically with existing mechanistic frameworks to create interpretable models that provide deeper insights into genetic design.

METHODS

Plasmid library construction

Circuit libraries were cloned using a custom hierarchical golden gate[29] assembly scheme, which enables rapid, modular cloning of complex gene circuits starting from individual genetic part sequences. Briefly, input DNA fragments are amplified from genomic or commercial DNA sources and cloned into kanamycin resistant sub-part entry vectors using BpiI (Thermofisher) (*level 0*). Sequence-verified sub-part plasmids are then used as inputs for assembly into carbenicillin-resistant entry vectors using BsaI-v2 HF (NEB) to yield genetic part plasmids (i.e., promoters, ORFs, terminators) (*level 1*). EUs are then constructed by assembling promoter, ORF, and terminator part plasmids into a kanamycin-resistant entry vector using Esp3I (Thermofisher) (*level 2*). EUs are then combined into multi-unit arrays by assembling into a carbenicillin- or spectinomycin-resistant destination vectors using BpiI (Thermofisher) (*level 3*).

In this system, barcode pools are incorporated into library assemblies at level 2 and 3. Level 2 barcoded EU pools were generated by first constructing a destination vector carrying a ccdB placeholder expression cassette downstream of the BFP stop codon. A semi-degenerate 18 bp barcode oligo pool (IDT)[63] was polymerase-extended and cloned in place of the ccdB cassette using BsaI-v2 (NEB) in a 20 µL golden gate reaction: 2 min at 37 °C and 5 min at 16 °C for a number of cycles equal to 10x the number of input fragments, followed by a 30 min digestion step at 37 °C and subsequent sequential 15 min denaturation steps at 65 °C and 80 °C. Resulting assemblies were purified using a miniprep column (Epoch Life Sciences), electroporated (BioRad) into 100 µL NEB 10-Beta electrocompetent *E. coli* (NEB), and plated onto a custom 30 in x 24 in LB_{Kan} agar plate, yielding ~120M colonies. Plates were grown at 37 °C for 16 h, at which time

colonies were scraped into LB_{Kan} (20 mL), incubated for 30 min at 37 °C, and the plasmid library extracted via miniprep (Qiagen). For large-scale libraries level 3 barcode pools were generated in order to ensure a unique barcode-to-circuit mapping. To create level 3 barcode pools, a semi-degenerate 17 bp oligo pool (IDT) was polymerase-extended and cloned via golden gate assembly into a SpectR destination vector carrying a placeholder ccdB cassette using PaqCI (NEB) following the assembly protocol above. Resulting assemblies were transformed into 100 µL homemade chemically competent[64,65] Stbl3 *E. coli* and plated onto a 10 cm LB_{Spect} agar plate. Plates were grown at 37 °C for 20 h to yield approximately 15,000 colonies, followed by colony scraping and plasmid DNA extraction, as described above.

Part pools (levels 1), EU pools (level 2) and circuit libraries (level 3) were constructed by combining input plasmids at 50 fmol per part category (15 µL total volume) using the above-described cycling protocol. Transformations varied by library size: level 2 and 3 assemblies for the 384-member library (**Fig. 2**) were transformed into 100 µL of chemically competent *E. coli* (DH5a and Stbl3, cells/mL) and respectively plated on LB_{Kan} and LB_{Carb} agar plates for 12-16 h (37 °C) to yield ~15,000 colonies. Colonies were scraped into 13 mL LB_{Kan} or LB_{Carb} and miniprepped using a Qiagen kit. Level 1 and 2 assemblies for the 166k-member library (**Fig. 3**) were respectively transformed using 100 µL and 400 µL of chemically competent *E. coli* (Stbl3, cells/mL) and grown on LB_{Carb} and LB_{Kan} agar plates at 37 °C for 12-16 h to yield ~5,000 and ~20,000 colonies, respectively. Level 3 166k-member assemblies were purified using a miniprep column (Epoch Life Sciences), electroporated (BioRad) into 100 µL 10-beta electrocompetent cells (NEB), and then plated on 10 15 cm LB_{Spect} agar plates for 16-20 h (37 °C) to yield ~3M colonies.

Long-read plasmid sequencing

To generate indices linking assembled constructs to their associated DNA barcodes, we used Oxford Nanopore Technology (ONT) long-range sequencing[26,66]. Libraries were prepared for long-read sequencing by digesting 2 µg of a level 3 plasmid library using Esp3I (Thermofisher) and purifying linearized fragments using a 0.5x volume of magnetic beads (Omega Bio-Tek) by volume. Nanopore sequencing adapters were added to the linearised pool using the LSK-112 genomic ligation kit (ONT) and sequenced using on a minION device (ONT) equipped with an R10 flow cell (FLO-MIN112). Basecalling was performed using Guppy (ONT, super high accuracy mode) running on a GPU (Nvidia RTX 3090). Composition and barcode assignment was performed using WIMPY (what's in my pot, y'all), a custom Matlab analysis pipeline that imports fastq files from Guppy, indexes the reads to a constant region in the level 3 plasmid backbone and filters them based on length to remove incomplete assemblies and non-library fragments. WIMPY then determines composition for filtered reads by identifying and assigning genetic parts through a combination of general Smith-Waterman alignments[67] and localized containment searches. This is done by splitting the reference sequence into 6-10bp "tiles" with a 1bp stride, after which nanopore reads are queried for the number of tiles contained for each part reference sequence. Reads containing >3% of tiles for a reference sequence are assigned while reads with more than one reference assignment are discarded. Barcode sequences are then determined by aligning the region downstream of the BFP EU to a degenerate reference sequence using a custom alignment matrix.

Cell culture

HEK293T cells (ATCC® CRL-11268™) used in this study were cultured under humidity control at 37 °C with 5% CO_2 in media containing Dulbecco's modified Eagle medium (DMEM) with high glucose (Gibco, 12100061) supplemented with 10% Fetal Bovine Serum (FBS; GeminiBio, 900-108), 50 units/ml penicillin, 50 µg/ml streptomycin (Pen Strep; Gibco, 15070063), 2 mM L-Alanyl-L-Glutamine (Caisson labs, GLL02), referred to hereafter as complete DMEM. HEK293T-LP cells with and without integrated libraries were maintained in DMEM supplemented with 50 µg/mL Hygromycin B (Sigma, H3274) and 1 µg/mL Puromycin (Sigma, P8833), respectively.

Single-copy library integration

To establish a landing pad (LP) cell line, low passage HEK293T cells were co-transfected with a linearized repair template comprising a YFP-HygR-expression cassette containing a BxB1 attP recognition site (pROC079) and a dual Cas9 and gRNA expression vector targeting the human *AAVS1* locus[68]. Genomic integration events were selected using 50 µg/mL Hygromycin B, and YFP+ cells were flow sorted to isolate clones (WOLF, NanoCellect). Clones were tested for integration competency and the presence of the intact LP cassette was subsequently confirmed by PCR. The resulting LP cell line enables efficient single-copy integration of attP-containing vectors into the genome via activity of the serine recombinase BxB1, which is expressed on a co-transfected plasmid. Cells harboring successfully integrated vectors are selected using Puromycin, yielding homogeneous engineered cell populations within ~10 days.

To integrate individual constructs, 125k cells were plated into a 24-well plate one day prior to transfection. The well was co-transfected with 250 ng of a level 3 vector containing a construct of interest and 125 ng of BxB1 expression plasmid (Addgene #51271) using JetPrime (VWR). Two days after transfection, cells were passaged into complete DMEM containing 1 µg/mL Puromycin and selected for 10 days to achieve homogenous expression. Library integration was performed by co-transfecting HEK293T-LP with 250 ng of plasmid library and 125 ng of BxB1 per 200k cells (384-member library = 1M cells transfected; 166k-member library = 100M cells transfected) using JetPrime. Media exchange was performed after 8 h and cells were expanded for an additional 40 h, split into five culture dishes, and grown for 6-8 days under Puromycin selection. Cells were passaged 1:3 and cultured for an additional 5 days under puromycin selection, and then combined in complete DMEM for flow sorting (10M cells/mL).

To determine the clonal variability within LP-integrated cell populations, we integrated a constitutively expressed mCherry EU following the protocol described above. We then randomly sorted 23 clones from the population (WOLF, NanoCellect) and measured their mCherry expression. We calculated the MAE for this set and used this value to define an expected error range from clonal heterogeneity (ERCH).

Flow sorting

To prepare libraries for flow sorting, cells were lifted using TrypLE (Gibco), washed with PBS, and resuspended in complete DMEM (10M cells/mL). For the 384-member library, the mRuby expression distribution was sorted into 10 evenly log-spaced bins on a Sony MA900 set to purity mode. The number of cells collected for each bin was

proportional to the % of the mRuby distribution expressed in that bin, with a target of ~125k cells sorted in the most populous bin [e.g., 2,252 cells were collected for bin 1 (0.53% of library), 126,562 for bin 4 (29.78% of library)]. For the 166k-member library, cells were split and either grown in DMEM containing 1μM 4-OHT (Sigma Aldrich) ("induced") or without 4-OHT ("uninduced") for 72 hrs. The eGFP expression distributions for both conditions were then sorted separately into 8 evenly log-spaced bins. Cells were sorted into bins in proportion to the relative abundance of the library in each bin, with the most populous bin set at 1.5M cells. For both libraries, sorted cells were plated into 96-, 24- or 6-well plates to achieve a plating density of 10-30%, grown under puromycin selection for 2 days, passaged and washed with PBS, and grown for an additional 3 days. Cells were then lifted using TrypLE and total RNA was extracted using the Takara RNA plus kit (Takara). 5 μL of mRNA was converted to cDNA (Verso, Life Technologies).

Short-read (Illumina) sequencing

To prepare sorted libraries for short-read sequencing, barcode regions were PCR amplified from cDNA using Phantamax polymerase (Vazyme) and custom bin-specific primer sets (IDT). The resulting amplicon pool was extracted from a 2% agarose gel. A second PCR step added Illumina sequencing adapters (i5 and i7) and sample-specific sequencing barcodes using Phantamax and custom primer sets (IDT), and the amplified product was extracted from a 2% agarose gel. Purified amplicons from each bin were pooled at equimolar concentrations and sequenced using an Illumina MiSeq (kit v2, 300 cycles) or NovaSeq6000 (Sp v1.5 kit, 300 cycles) for the 384- and 166k-member libraries, respectively. Illumina data were analyzed using a custom analysis pipeline (Matlab).

Briefly, fastq files are imported and split according to their sample-specific sequencing barcode. Barcode sequencing data is converted to average expression for all compositions/variants as follows: (1) Counts for each barcode in each sample/bin are normalized by the percentage of the library distribution in that bin to account for differences in read depth across bins. (2) A weighted average of number of reads in each bin for each barcode is then calculated, to assign an average expression level (using a bin → expression conversion table) to each barcode. (3) Barcode expression levels are then cross-referenced with the Nanopore indices to assign average expression values for all barcodes associated with a given variant in a n-by-y cell array, where n is the number of library variants and y is the (variable) number of barcodes for each variant. (4) For each variant, a gaussian kernel density estimation is performed over the log-normalized expression values from all barcodes associated with each variant, using the matlab function ksdensity. The expression value corresponding to the kernel peak is assigned as the mean expression value for each variant.

Random forest regression

For the 384-member EU library, a bootstrap aggregation (bag) RF model was constructed using the matlab function *templatetree*. The promoter, Kozak and terminator were used as categorical input variables for each EU, with log(mRuby) expression (AU) as the predicted output. The data was split 80:20 (307 EUs and 77 EUs) for training and testing, respectively. The number of trees was chosen based on 10-fold cross validated RMSE loss as a function of increasing number of trees using the matlab function *fitrensemble*. The trained RF model was then validated on the test set (**Fig. 2F**).

Identification of genetic part interactions was determined by the absolute log error between predicted and observed mRuby expression.

For the 166k-member circuit library, two independent least-squares boosted RF models were used to predict basal and induced expression for experimentally unmapped circuit members using the same matlab functions mentioned above. Data used for RF training were obtained by first filtering the dataset and keeping points that had 10 or more barcode reads. The space was then split into 8 equally log-spaced regions and, where available, 1000 & 200 data points were randomly sampled from each bin to create the training and test sets, respectively. In bins that had fewer than 1200 data points, all the available points were taken and randomly split into 80:20 for training and testing. The inputs to the model were the 10 categorical variables representing the parameters tuned in the library (e.g., SynTF affinity, #BM etc.), while the output was the log(eGFP) expression (AU) for basal or induced expression values. Following hyperparameter optimization for number of trees, least squares boosting algorithm learn rate, and maximum decisions splits, model structure and performance were as follows: basal, 15 trees, max number of splits=31 max number of splits, learn rate 0.263, R^2_{train}=0.85, R^2_{test}=0.75 and; induced, 75 trees, max number of splits=41, learn rate=0.1586, R^2_{train}=0.84, R^2_{test}=0.81.

Mutual information calculation

For a given pair of features in a region of design space, mutual information was computed using the following formula:

$$\sum_{x,y} p\,(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$$

where x and y are 2 different features of the library, and p(x, y) represents the joint frequency for specific x and y combinations divided by the total number of points in the dataset. p(x) and p(y) represent the fractional frequency of the corresponding parts by themselves.

UMAP clustering

For UMAP dimensional reduction the *clusterevaluation* function in Matlab was used to carry out a gap test to identify the optimal number of clusters in the data, followed by K-means clustering to identify clusters.

ACKNOWLEDGEMENTS

We thank Oleg Igoshin, Ankit Patel, Yashwanth Lagisetty, Satpreet Singh, and members of the Bashor lab for helpful discussions. This work was supported by grants from NIH R01 EB029483 (C.J.B.), NIH R01 EB032272 (C.J.B.), ONR N00014-21-1-4006 (C.J.B.), and funding from the Robert J. Kleberg Jr. and Helen C. Kleberg Foundation (C.J.B.). R.W.O. was supported by a graduate fellowship from the American Heart Association (917746). B.K. was supported by a NLM Training Program in Biomedical Informatics and Data Science fellowship (T15LM007093-31) and by NIH grant P01-AI15299901. K.D.C. was supported by NSF EF-2126387 and the Ken Kennedy Institute Computational Science & Engineering Recruiting Fellowship.

AUTHOR CONTRIBUTIONS

R.W.O., K.R., and C.J.B. conceived of the study, R.W.O. and K.R. carried out the experiments and developed the analysis software, with assistance from T.C.P., K.D.S., J.A.W., S.L., T.H.Z., E.M.R., and A.S. R.W.O., and T.C.P. developed the modular cloning scheme and landing pad cell line, with assistance from S.L. B.K., K.C., and T.J.T. helped develop the barcoding scheme and analysis software. R.W.O., K.R., T.C.P., J.W.R., P.M, and C.J.B. analyzed the data. C.J.B. supervised the study. R.W.O., K.R., and C.J.B wrote the manuscript, with input from all authors.

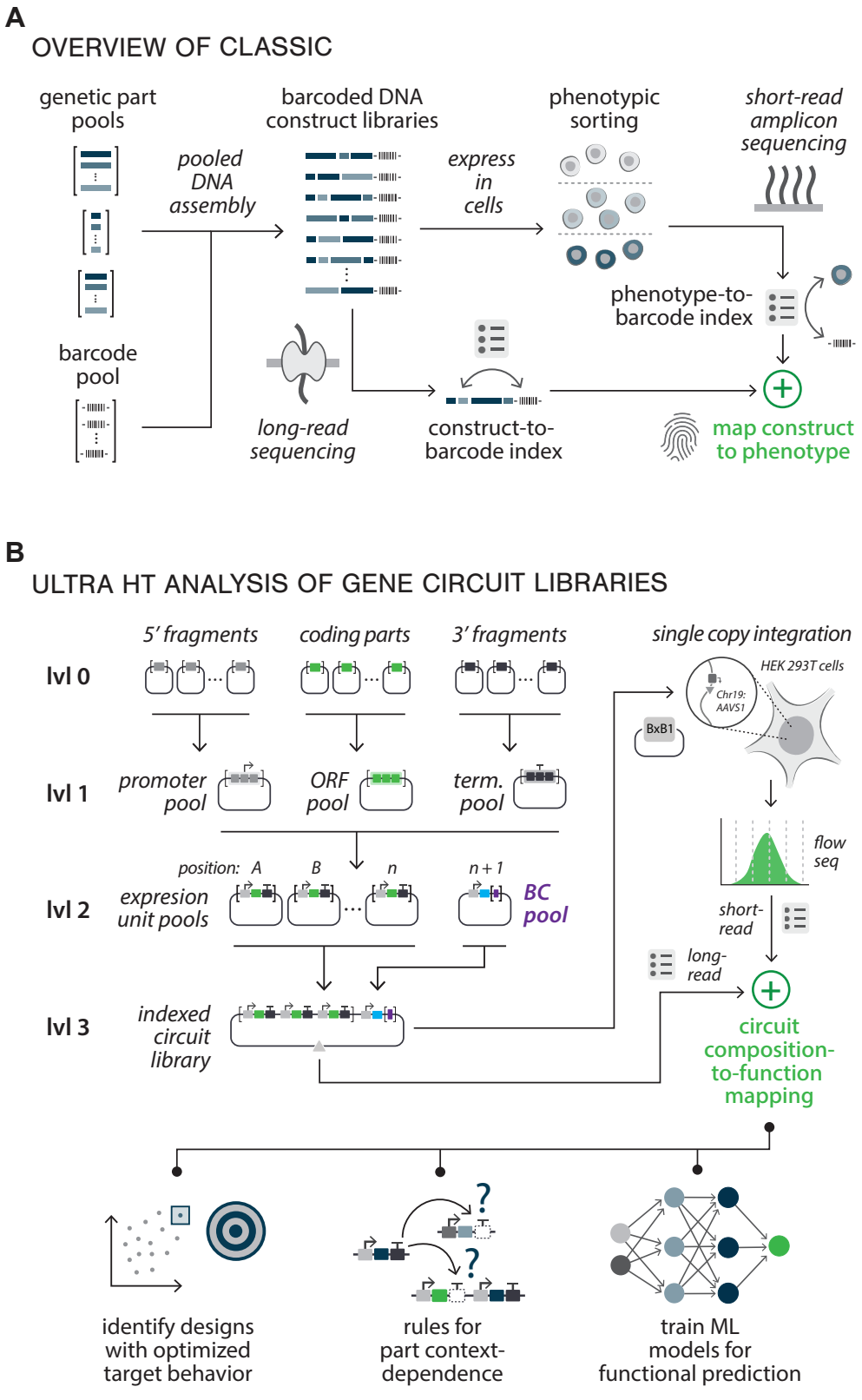


Figure 1. Using CLASSIC to systematically map the design space of complex genetic programs. (A) Overview of CLASSIC. Pooled assembly of genetic parts with DNA barcode sequences yields libraries of barcode-indexed constructs of arbitrary length and complexity. Long-read nanopore sequencing is used to create an index matching construct composition to an associated barcode. In parallel, libraries introduced into cells undergo sorting or selection to bin expression phenotypes. Barcode amplicons generated for each bin are subjected to short-read NGS to quantify expression phenotype, which is then mapped to construct composition via barcode indexes. **(B)** Application of CLASSIC to profile a synthetic gene circuit design space. Hierarchical golden gate assembly is used to compose libraries of multi-EU circuits with combinatorially varied part compositions and circuit designs. Sequence fragment pools for different parts categories from level 0 are combined to yield level 1 pools of promoters, open reading frames (ORF), and terminators (term.), which are then combined to yield level 2 EUs (square brackets: fragment/part pools). Barcode pools are combined with EU pools to create indexed multi-EU circuit libraries (level 3) that are integrated into HEK293T-LP cells placed at the *AAVS1* locus in chromosome 19 via expression of the BxB1 recombinase (top right). Library analysis by a combination of nanopore and flow-seq yields composition-to-function mapping.

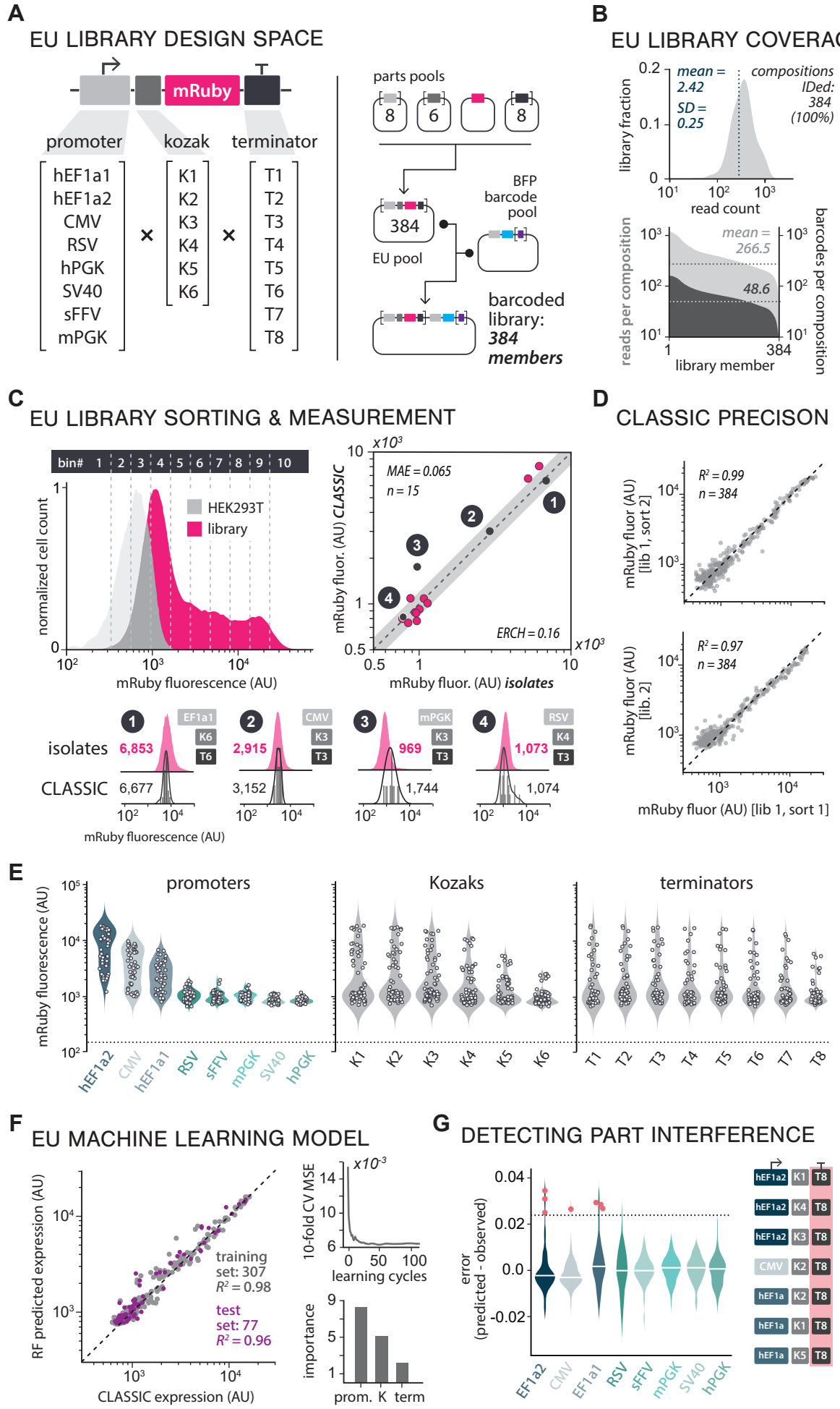


Figure 2. CLASSIC can quantitatively profile diverse compositions of genetic parts. (**A**) EU library design space. Left: the EU library comprises combinations of parts from 3 categories (promoters, Kozak sequences, and terminators). Right: Part pools are assembled (step 1) with an mRuby ORF to generate a 384-member EU pool, which is combined with a barcoded-BFP EU pool to generate the indexed EU library. (**B**) EU library assembly and indexing balance. Oxford nanopore sequencing was performed on the assembled library. Data were analyzed to assess library composition count (top) and composition/barcode balance [reads per composition (light grey), unique barcodes per composition (dark grey); data plotted in rank order of reads per composition] (bottom). (**C**) EU library expression quantification. Top left: the library was flow sorted into 10 equally log-spaced bins [empty HEK293T cells (grey histogram), library (pink histogram)]. Top right: residual for FACS-measured geometric mean fluorescence values for sorting-isolated clones (n=15) plotted against CLASSIC-derived values [ERCH (grey band), error range from clonal heterogeneity (see **materials and methods**)]; black dots, clonal isolates. Bottom: Representative data from 4 clonal isolates (pink solid) are shown with corresponding CLASSIC-computed distributions [kernel density (black line) calculated from normalized barcode read count (grey vertical bars)]; MAE, mean absolute error; AU, arbitrary fluorescence units. (**D**) CLASSIC precision. Correlation of CLASSIC-computed EU expression between technical replicate (top) and biological replicate (bottom) experiments. (**E**) Influence of part identity on expression. Violin plots of each part-specific distribution, ordered from strongest to weakest. Dotted line, HEK293T background expression mean. (**F**) RF analysis of EU behavior space. Left: RF modeled values plotted against CLASSIC measurements. Grey, training data; purple, test data. Top right: 10-fold cross-validated error (y

axis) vs number of learning cycles (x axis). CV, cross-validation; MSE, mean squared error. Bottom right: feature importance scores for each part category. (**G**) Analysis of part interference. Left: error (RF predicted – CLASSIC-observed) is plotted for all compositions associated with each promoter. White lines, mean values; red dots, >0.025 outliers. Right: part configurations for outliers. Red shaded area highlights commonality of terminator T8.

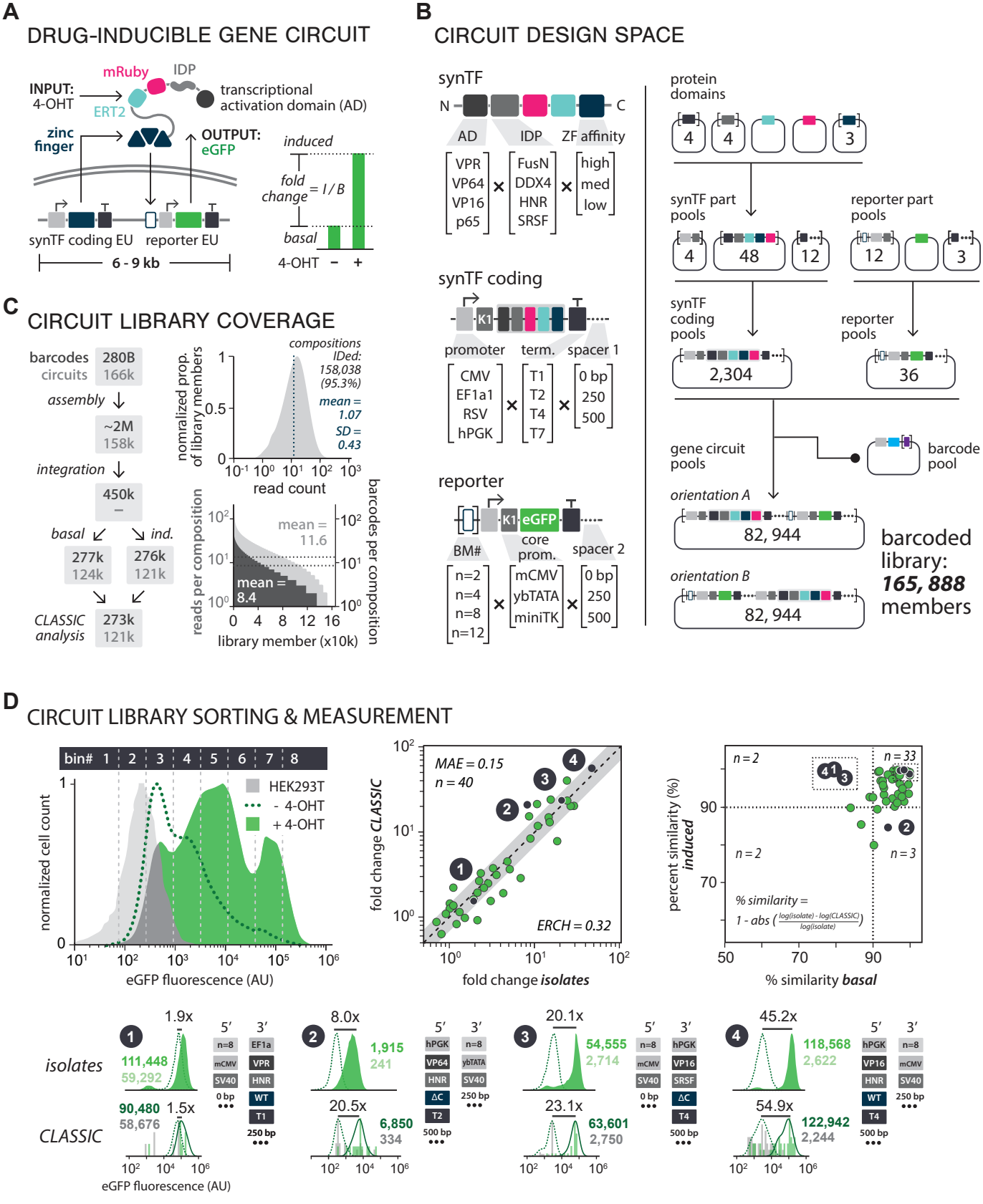
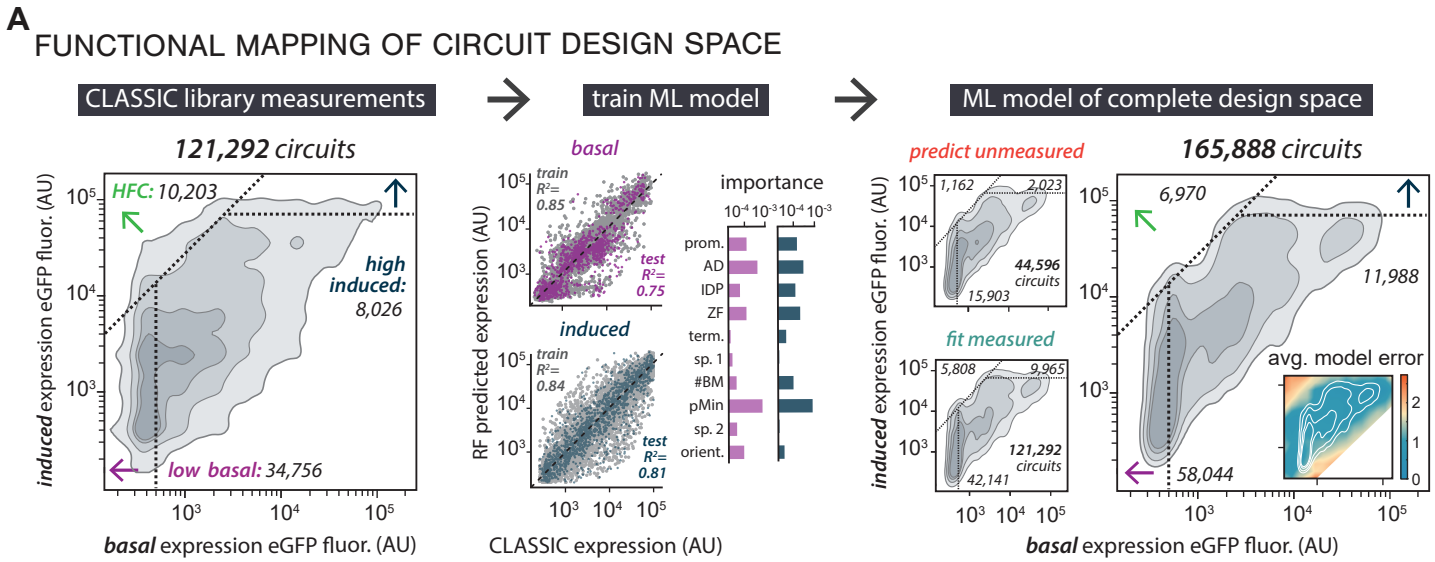
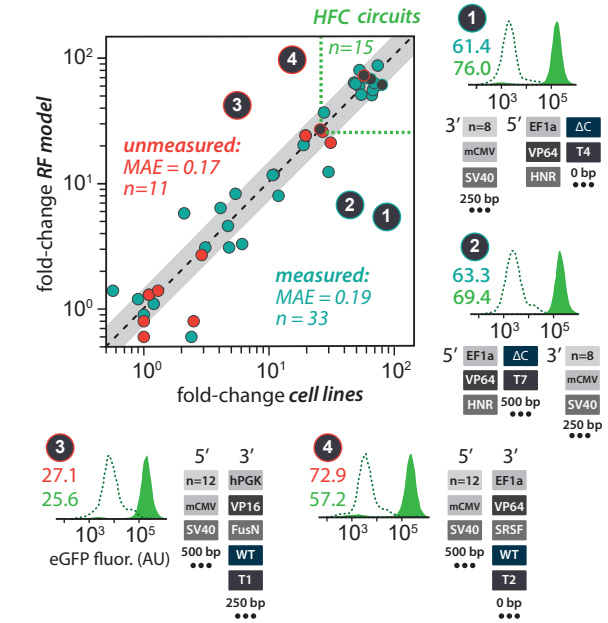
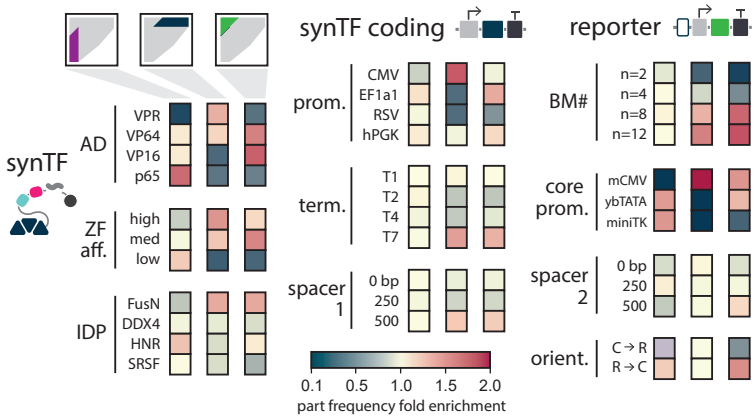


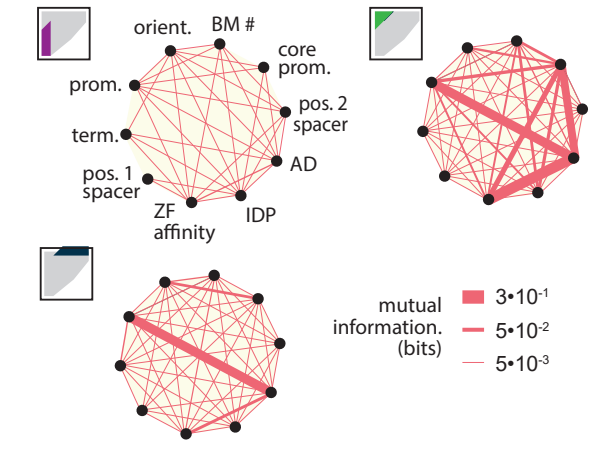
Figure 3. Using CLASSIC to profile a synthetic gene circuit design landscape. A) Inducible synTF circuit diagram. Left: The circuit contains two EUs: one codes for the synTF and the other is an eGFP reporter. Without inducer, the synTF localizes to the cytoplasm. Upon addition of 4-OHT (input), the synTF translocates into the nucleus to bind to and activate reporter eGFP expression (output). Right: Expression fold change is the ratio of expression levels in the presence or absence of inducer. **(B)** synTF circuit design space. Left: SynTF diversity (arranged N-to-C term): of 4 ADs, 4 IDPs, and 3 ZF affinities. synTF coding EU diversity: 4 constitutive promoters, 4 terminators. Reporter EU diversity: 4 BM number variants, 3 minimal core promoters. This unit contains a constant terminator. Both EUs have 3' spacing sequences of 0bp, 250bp or 500bp downstream. Right: EUs assembled from input parts and combined in 2 different 5'-to-3' orientations, to generate a combinatorial diversity of 165,888 possible circuit variants. **(C)** Balance of circuit library assembly and indexing. Left: data for ~121k out of 166k variants (73%) were recovered by CLASSIC; compositions, light grey; barcodes, dark grey. Right: Nanopore sequencing data of the library were analyzed to assess library composition count (top) and composition/barcode balance [reads per composition (light grey), unique barcodes per composition (dark grey); data plotted in rank order of reads per composition] (bottom). **(D)** Circuit library sorting and measurement. Top right: eGFP expression of circuit library in presence (solid green) and absence (dotted green line) of 4-OHT (top, center) is shown, along with boundaries of flow sorting bins (vertical grey dotted lines); grey histogram, empty HEK293T-LP cells. Top middle: Fold-change values for 40 clonal isolates plotted against CLASSIC-derived values. Black dots, isolates displayed at the bottom of the panel; MAE, mean average error; grey region, ERCH, error range of clonal heterogeneity

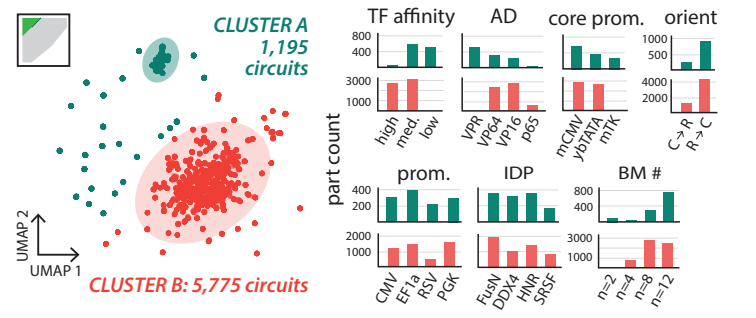
AU, arbitrary fluorescence units. Top right: % similarity between basal and induced values for CLASSIC and isolates. Bottom: Flow data from 4 isolates (green dotted line, uninduced; green solid, induced) are shown with corresponding CLASSIC-computed distributions [kernel density (black line) calculated from normalized barcode read count (grey vertical bars)]. Parts combinations corresponding to each index are shown.











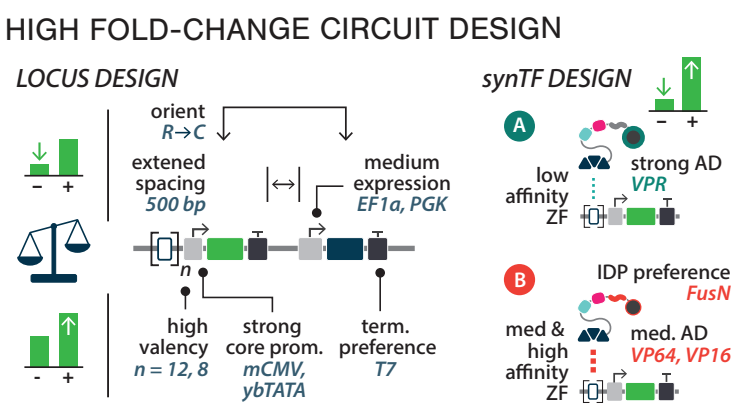


Figure 4. Analysis of CLASSIC library behavior space reveals gene circuit design rules. (A) ML model of inducible synTF circuit behavior space. Left: Basal and induced CLASSIC measurements were plotted as a contour plot [contours, light to dark: 97.5%, 90%, 70%, 50% of total measurements]. Highlighted regions bounded by dotted lines: low basal (<500 AU), purple arrow; high induction (>70k AU) blue; high fold change (HFC) (>25x, green). Values in the plot indicate the number of compositions in each region. Middle: RF of with 80:20 train:test split for basal and induced CLASSIC data [training data (grey) plotted against test data (purple, basal; blue, induced). Right: Contour plot of RF modeled design space for unmeasured compositions predicted by the RF model, measured circuits, and combined data (complete design space). Inset: average model error between the CLASSIC-derived measurements and RF-computed values. (**B**) Experimentally validating ML prediction of circuit function. Fold-change values for cell lines from unmeasured configurations (red) and measured configurations (green) were plotted against CLASSIC-derived values. Black dots, isolates displayed at the periphery of the panel; MAE, mean average error; grey region, ERCH, error range of clonal heterogeneity. AU, arbitrary fluorescence units. (**C**) Genetic part usage in highlighted regions of behavior space. Part fold enrichment is calculated by dividing observed part occurrence by expected part occurrence from a balanced library. Red text, categories with high asymmetry used for cluster analysis. (**D**) Mutual information between part categories in different regions of behavior space. MI between part categories is denoted by red line thickness. (**E**) Clustering analysis of HFC circuit designs. High asymmetry part categories were chosen for UMAP dimensional reduction followed by K-means clustering. Bar plots denote number of part occurrences within each cluster. **(F)** Strategies for engineering synTF circuits

with high fold change behavior involve combining design elements that maximize induction while limiting leaky basal expression.

REFERENCES

1 English, M. A., Gayet, R. V. & Collins, J. J. Designing Biological Circuits: Synthetic Biology Within the Operon Model and Beyond. *Annu Rev Biochem* **90**, 221-244, doi:10.1146/annurev-biochem-013118-111914 (2021).

2 Mahata, B. *et al.* Compact engineered human transactivation modules enable potent and versatile synthetic transcriptional control. *bioRxiv*, 2022.2003.2021.485228, doi:10.1101/2022.03.21.485228 (2022).

3 Slusarczyk, A. L., Lin, A. & Weiss, R. Foundations for the design and implementation of synthetic genetic circuits. *Nat Rev Genet* **13**, 406-420, doi:10.1038/nrg3227 (2012).

4 Bashor, C. J. & Collins, J. J. Understanding Biological Regulation Through Synthetic Biology. *Annu Rev Biophys* **47**, 399-423, doi:10.1146/annurev-biophys-070816-033903 (2018).

5 Bashor, C. J., Hilton, I. B., Bandukwala, H., Smith, D. M. & Veiseh, O. Engineering the next generation of cell-based therapeutics. *Nat Rev Drug Discov*, doi:10.1038/s41573-022-00476-6 (2022).

6 Beitz, A. M., Oakes, C. G. & Galloway, K. E. Synthetic gene circuits as tools for drug discovery. *Trends Biotechnol* **40**, 210-225, doi:10.1016/j.tibtech.2021.06.007 (2022).

7 Kitada, T., DiAndreth, B., Teague, B. & Weiss, R. Programming gene and engineered-cell therapies with synthetic biology. *Science* **359**, doi:10.1126/science.aad1067 (2018).

8 Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat Rev Microbiol* **12**, 381-390, doi:10.1038/nrmicro3239 (2014).

9 Yeung, E. *et al.* Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Syst* **5**, 11-24 e12, doi:10.1016/j.cels.2017.06.001 (2017).

10 Lou, C., Stanton, B., Chen, Y. J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* **30**, 1137-1142, doi:10.1038/nbt.2401 (2012).

11 Muller, I. E. *et al.* Gene networks that compensate for crosstalk with crosstalk. *Nat Commun* **10**, 4028, doi:10.1038/s41467-019-12021-y (2019).

12 Prindle, A. *et al.* Rapid and tunable post-translational coupling of genetic circuits. *Nature* **508**, 387-391, doi:10.1038/nature13238 (2014).

13 Shaw, W. M. *et al.* Engineering a Model Cell for Rational Tuning of GPCR Signaling. *Cell* **177**, 782-796 e727, doi:10.1016/j.cell.2019.02.023 (2019).

14 Kinney, J. B., Murugan, A., Callan, C. G., Jr. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**, 9158-9163, doi:10.1073/pnas.1004290107 (2010).

15 Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521-530, doi:10.1038/nbt.2205 (2012).

16 Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol Syst Biol* **10**, 748, doi:10.15252/msb.20145136 (2014).

17 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 14024-14029, doi:10.1073/pnas.1301301110 (2013).

18 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).

19 de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**, 56-65, doi:10.1038/s41587-019-0315-8 (2020).

20 Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91-106 e123, doi:10.1016/j.cell.2019.04.046 (2019).

21 Gera, T., Jonas, F., More, R. & Barkai, N. Evolution of binding preferences among whole-genome duplicated transcription factors. *Elife* **11**, doi:10.7554/eLife.73225 (2022).

22 DelRosso, N. *et al.* Large-scale mapping and systematic mutagenesis of human transcriptional effector domains. *bioRxiv*, 2022.2008.2026.505496, doi:10.1101/2022.08.26.505496 (2022).

23 Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable RNA switches. *Nat Commun* **11**, 5057, doi:10.1038/s41467-020-18677-1 (2020).

24 Jones, E. M. *et al.* Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife* **9**, doi:10.7554/eLife.54895 (2020).

25 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, doi:10.1038/nrg.2016.49 (2016).

26 De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat Rev Genet* **22**, 572-587, doi:10.1038/s41576-021-00367-3 (2021).

27 Zhou, Y. *et al.* Encoding Genetic Circuits with DNA Barcodes Paves the Way for Machine Learning-Assisted Metabolite Biosensor Response Curve Profiling in Yeast. *ACS Synth Biol* **11**, 977-989, doi:10.1021/acssynbio.1c00595 (2022).

28 Wong, A. S., Choi, G. C., Cheng, A. A., Purcell, O. & Lu, T. K. Massively parallel high-order combinatorial genetics in human cells. *Nat Biotechnol* **33**, 952-961, doi:10.1038/nbt.3326 (2015).

29 Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, e16765, doi:10.1371/journal.pone.0016765 (2011).

30 O'Gorman, S., Fox, D. T. & Wahl, G. M. Recombinase-mediated gene activation and site-specific integration in mammalian cells. *Science* **251**, 1351-1355, doi:10.1126/science.1900642 (1991).

31 Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res* **42**, 13440-13451, doi:10.1093/nar/gku1082 (2014).

32 Norrman, K. *et al.* Quantitative comparison of constitutive promoters in human ES cells. *PLoS One* **5**, e12413, doi:10.1371/journal.pone.0012413 (2010).

33 Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* **5**, e10611, doi:10.1371/journal.pone.0010611 (2010).

34 Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci U S A* **110**, 11284-11289, doi:10.1073/pnas.1305590110 (2013).

35 Petitclerc, D. *et al.* The effect of various introns and transcription terminators on the efficiency of expression vectors in various cultured cell lines and in the mammary gland of transgenic mice. *J Biotechnol* **40**, 169-178, doi:10.1016/0168-1656(95)00047-t (1995).

36 Wang, X. Y. *et al.* Enhanced Transgene Expression by Optimization of Poly A in Transfected CHO Cells. *Front Bioeng Biotechnol* **10**, 722722, doi:10.3389/fbioe.2022.722722 (2022).

37 Couronne, R., Probst, P. & Boulesteix, A. L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19**, 270, doi:10.1186/s12859-018-2264-5 (2018).

38 Khalil, A. S. *et al.* A synthetic biology framework for programming eukaryotic transcription functions. *Cell* **150**, 647-658, doi:10.1016/j.cell.2012.05.045 (2012).

39 Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* **4**, 1471-1501, doi:10.1038/nprot.2009.98 (2009).

40 Li, H. S. *et al.* Multidimensional control of therapeutic human cell function with synthetic gene circuits. *Science* **378**, 1227-1234, doi:10.1126/science.ade0156 (2022).

41 Feil, R., Wagner, J., Metzger, D. & Chambon, P. Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem Biophys Res Commun* **237**, 752-757, doi:10.1006/bbrc.1997.7124 (1997).

42 Bashor, C. J. *et al.* Complex signal processing in synthetic gene circuits using cooperative regulatory assemblies. *Science* **364**, 593-597, doi:10.1126/science.aau8287 (2019).

43 Donahue, P. S. *et al.* The COMET toolkit for composing customizable genetic programs in mammalian cells. *Nat Commun* **11**, 779, doi:10.1038/s41467-019-14147-5 (2020).

44 Muldoon, J. J. *et al.* Model-guided design of mammalian genetic programs. *Sci Adv* **7**, doi:10.1126/sciadv.abe9375 (2021).

45 Kabadi, A. M. & Gersbach, C. A. Engineering synthetic TALE and CRISPR/Cas9 transcription factors for regulating gene expression. *Methods* **69**, 188-197, doi:10.1016/j.ymeth.2014.06.014 (2014).

46 La Russa, M. F. & Qi, L. S. The New State of the Art: Cas9 for Gene Activation and Repression. *Mol Cell Biol* **35**, 3800-3809, doi:10.1128/MCB.00512-15 (2015).

47 Sadowski, I., Ma, J., Triezenberg, S. & Ptashne, M. GAL4-VP16 is an unusually potent transcriptional activator. *Nature* **335**, 563-564, doi:10.1038/335563a0 (1988).

48 Shin, Y. *et al.* Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets. *Cell* **168**, 159-171 e114, doi:10.1016/j.cell.2016.11.054 (2017).

49 Schneider, N. *et al.* Liquid-liquid phase separation of light-inducible transcription factors increases transcription activation in mammalian cells and mice. *Sci Adv* **7**, doi:10.1126/sciadv.abd3568 (2021).

50 Boeynaems, S. *et al.* Phase Separation in Biology and Disease; Current Perspectives and Open Questions. *J Mol Biol* **435**, 167971, doi:10.1016/j.jmb.2023.167971 (2023).

51 Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc Natl Acad Sci U S A* **89**, 5547-5551, doi:10.1073/pnas.89.12.5547 (1992).

52 Hansen, J. *et al.* Transplantation of prokaryotic two-component signaling pathways into mammalian cells. *Proc Natl Acad Sci U S A* **111**, 15705-15710, doi:10.1073/pnas.1406482111 (2014).

53 Ede, C., Chen, X., Lin, M. Y. & Chen, Y. Y. Quantitative Analyses of Core Promoters Enable Precise Engineering of Regulated Gene Expression in Mammalian Cells. *ACS Synth Biol* **5**, 395-404, doi:10.1021/acssynbio.5b00266 (2016).

54 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv180203426M>.

55 Wimmer, E., Mueller, S., Tumpey, T. M. & Taubenberger, J. K. Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nat Biotechnol* **27**, 1163-1172, doi:10.1038/nbt.1593 (2009).

56 Brophy, J. A. & Voigt, C. A. Principles of genetic circuit design. *Nat Methods* **11**, 508-520, doi:10.1038/nmeth.2926 (2014).

57 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

58 Pinglay, S. *et al.* Synthetic regulatory reconstitution reveals principles of mammalian Hox cluster regulation. *Science* **377**, eabk2820, doi:10.1126/science.abk2820 (2022).

59 Voigt, C. A. Synthetic biology 2020-2030: six commercially-available products that are changing our world. *Nat Commun* **11**, 6379, doi:10.1038/s41467-020-20122-2 (2020).

60 Valeri, J. A. *et al.* Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat Commun* **11**, 5058, doi:10.1038/s41467-020-18676-2 (2020).

61 Hollerer, S. *et al.* Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nat Commun* **11**, 3551, doi:10.1038/s41467-020-17222-4 (2020).

62 LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun* **13**, 5159, doi:10.1038/s41467-022-32829-5 (2022).

63 Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* **18**, 165-169, doi:10.1038/s41592-020-01041-y (2021).

64 Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proc Natl Acad Sci U S A* **86**, 2172-2175, doi:10.1073/pnas.86.7.2172 (1989).

65 Parrish, J. R. *et al.* High-throughput cloning of Campylobacter jejuni ORfs by in vivo recombination in Escherichia coli. *J Proteome Res* **3**, 582-586, doi:10.1021/pr0341134 (2004).

66 Currin, A. *et al.* Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth Biol (Oxf)* **4**, ysz025, doi:10.1093/synbio/ysz025 (2019).

67 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197, doi:10.1016/0022-2836(81)90087-5 (1981).

68 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).