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], Yiduo Wang[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 77005, 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 multiple 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 genetic screening platform that combines long- and short-read next-generation sequencing (NGS) modalities to quantitatively assess pools of constructs of arbitrary length containing diverse part compositions. We show that CLASSIC can measure expression profiles of >10^5 drug-inducible gene circuit designs (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 accurate 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. They could also facilitate the development of data-driven models that are capable of inferring context-specific part function or forward predicting the behavior of novel circuit designs. HT screening approaches that utilize short-read 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, high-depth functional profiling of libraries of DNA constructs long enough (>1 kb) to encode entire circuits can be costly and technically challenging. Methods have emerged that permit multiplexed analysis of long constructs by appending short barcode index sequences that facilitate amplicon-based readout by short-read NGS[25]. This approach has been used in array formats in which each construct is

91 assembled and barcoded separately, or through nested assembly schemes[25,26], limiting library
92 design flexibility and assay throughput. More recently, random barcoding and indexing via long-
93 read single-molecule real-time (SMRT) sequencing (PacBio) has been used to characterize li-
94 braries of variants ranging from 10^2 to 10^4 constructs, mostly for individual ORFs of up to 4 kb in
95 length[27-29].

96 To perform indexed multiplexing for construct libraries containing combinations of genetic
97 parts at larger scales, we devised an approach that combines long-read nanopore and short-read
98 Illumina NGS to analyze construct libraries generated via pooled genetic part assembly (**Fig. 1A**).
99 In our scheme, libraries are generated via pooled assemblies of genetic parts that incorporate
100 semi-random barcodes. Nanopore sequencing is used to rapidly and inexpensively assess part
101 composition and create a composition-to-barcode index. Libraries can then be introduced into
102 cells, binned based on expression phenotype, and analyzed by short-read (Illumina) amplicon
103 sequencing to produce a barcode-to-phenotype index. A map that matches construct composition
104 to phenotype can then be revealed by comparing the two indices. Using this technique, which we
105 refer to as CLASSIC (combining long- and short-range sequencing to investigate genetic com-
106 plexity), it is possible to obtain high-depth phenotypic expression data for large libraries (>10^5) of
107 genetic part compositions of arbitrary length using standard phenotypic selection or flow sorting
108 experiments.

109 In order to configure CLASSIC to quantitatively profile libraries of gene circuits with diverse
110 part combinations in human cells (**Fig. 1B**), we devised a custom hierarchical golden-gate[30] clon-
111 ing scheme in which library diversity is programmed through a series of pooled DNA assembly
112 steps. Diversified pools of 5', coding, and 3' gene elements are first generated through assembly
113 of input part fragments (level 0 to 1), and subsequently combined to create single-gene expres-
114 sion unit (EU) pools (level 1 to 2). EU pools are then combined with a plasmid-encoded barcode
115 pool to yield multi-EU circuit libraries (level 2 to 3). Following nanopore sequencing and data
116 analysis (see **Methods**), circuit libraries are genomically integrated at single copy into a HEK293T

117 cell line harboring a custom “landing pad” cassette (HEK293T-LP)[31,32] (see **Methods**). Library-
118 integrated cells are then flow-sorted based on circuit expression output, followed by Illumina NGS
119 to quantitate bin distributions of circuit-associated barcodes[14,15] (see **Methods**).

120 To evaluate the quantitative accuracy of this approach, we constructed and measured
121 expression for a 384-member library in which mRuby expression was modulated by variable com-
122 binations of genetic parts: a set of 8 constitutively active promoters frequently used in mammalian
123 genetic engineering[33,34]; 6 variable Kozak sequences capable of tuning expression by modulating
124 translational initiation rate[35]; and 8 common mammalian transcriptional terminator se-
125 quences[20,36,37] (**Fig. 2A left**). The library was cloned by combining promoter, Kozak, and termina-
126 tor input pools with a single mRuby ORF to generate a diversified EU pool, which was then com-
127 bined with barcode plasmids to create the indexed library (**Fig. 2A, right**). Nanopore sequencing
128 of the library demonstrated complete, part-balanced coverage of EU design space (**Fig. 2B top**)
129 with proportional assortment of unique barcodes, indicating an absence of systematic bias for
130 both EU assembly and barcode indexing steps (**Fig. 2B bottom**).

131 Following integration into HEK293T-LP cells, the EU library was flow-sorted into 10 bins
132 based on mRuby expression (414,407 cells) and EU-associated barcodes were measured by
133 Illumina NGS (**Fig. 2C left**). We aggregated barcodes that mapped to the same EU to infer a
134 mean expression and distribution for each library composition (**Fig. 2B bottom**; see **Methods**).
135 To compare our NGS-derived measurements to “ground-truth” values, we made direct measure-
136 ments on 15 randomly sorted and clonally expanded library members using flow cytometry. Ge-
137 ometric mean values for these isolates showed excellent agreement with corresponding CLAS-
138 SIC-derived values (*MAE*=0.07), with residual errors mostly falling within an experimentally de-
139 termined range of clonal expression heterogeneity (*ERCH*; see **Methods**) (**Fig. 2C**). Our analysis
140 demonstrated that the precision of our measurements was enhanced by mapping multiple unique
141 barcodes to each composition, leading to greater accuracy than one-to-one mapping. To validate
142 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[38], 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**), and showed superior performance to a first-principles mechanistic model. Interestingly, part compositions that were shown by both the RF regression models 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)[39-41] 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

169 expression[39]. Transcriptional induction occurs upon addition of the small molecule 4-hydroxyta-
170 moxifen (4-OHT), which binds to a mutant version of the human estrogen receptor (ERT2)[42] ap-
171 pended to the synTF, facilitating its translocation from the cytoplasm into the nucleus to activate
172 reporter transcription. As we and others have demonstrated[39,43-45], identifying designs that are
173 optimized for high fold-change (HFC) (**Fig. 3A, right**) expression can be challenging in eukaryotic
174 systems; basal and induced expression levels must be respectively minimized and maximized
175 through fine-tuning of both transcriptional regulatory features of the locus and the molecular prop-
176 erties of the synTF, thereby ensuring robust expression of a transgene exclusively in the presence
177 of inducer.

178 To create a composition-to-function map for an inducible synTF circuit, we compiled a 10-
179 feature design space consisting of genetic part categories that we hypothesized could be im-
180 portant for achieving HFC behavior. This included 4 different TAs[46-48], 3 ZF affinities[39], and a set
181 of 4 intrinsically disordered protein (IDP) domains[49-51] that have been shown to facilitate liquid-
182 liquid phase condensation of nuclear-localized TFs[49,50]. We also varied transcriptional regulatory
183 features of the circuit, including 4 promoters and 4 terminators in the synTF coding EU, as well
184 as 3 core promoters[52-54] and different numbers of BMs (2, 4, 8, or 12) in the reporter EU. Addi-
185 tionally, we varied the spacing between the EUs (0, 250, or 500 bp) along with their 5'-to-3' orien-
186 tation (**Fig. 3B, left**) to yield an overall circuit design space of 165,888 compositions (**Fig. 3B,**
187 **left**). This library was constructed in 3 steps by first assembling protein domain parts to create a
188 level 1 synTF ORF pool, then conducting parallel assemblies to generate level 2 pools of synTF
189 coding and reporter EUs, and finally combining EU and barcode pools into the level 3 destination
190 vector (**Fig. 3B right**).

191 Nanopore sequencing of the pooled library yielded barcode assignments for 95.3% of total
192 compositions (**Figs. 3C**), with a mean of 8.4 barcodes for each circuit (**Fig. 3C bottom left**). We
193 integrated the library into HEK293T-LP cells and sorted un-induced and 4-OHT-induced popula-
194 tions (8.6 and $15.7 \cdot 10^6$ cells, respectively) separately into 8 bins based on eGFP fluorescence

195 (**Fig. 3D left**). Following analysis by Illumina NGS, we matched a total of 121,292 (73% of design
196 space) compositions to barcodes that were detected in both the sorted populations (mean bar-
197 codes circuit^{-1}=2.25) (**Fig. 3C**). We then used barcode bin distributions to compute basal, induced,
198 and fold-change expression values for each variant. CLASSIC-derived fold-change values
199 demonstrated excellent overall agreement (*MAE*=0.145) (**Fig. 3D top middle**) with those of ran-
200 dom isolates (n=40) (**Figs. 3D bottom**), and values for both basal and induced expression
201 demonstrated comparably high percent similarities (**Fig. 3D top right**). These results confirm that
202 CLASSIC retains quantitative measurement accuracy when scaled to orders-of-magnitude larger
203 and more complex libraries.

204 We analyzed the distribution of CLASSIC-measured values across a 2-dimensional be-
205 havior space (basal vs. induced eGFP expression), examining the relative density of compositions
206 in 3 regions of interest: low basal (<500 AU), high induced (>70,000 AU), and HFC expression
207 (>25x fold-change) (**Fig. 4A**). We observed a greater proportion of library members in the low
208 basal than in the high induced region (~4.3x), and ~50% of compositions exhibited fold-change
209 values of <3x, while a smaller fraction of the library (~8%) fell within the HFC region. Since in-
210 complete measurement of our design space (27% was unmeasured) may limit our insight into
211 circuit design principles—particularly for the HFC region—we asked whether we could create a
212 complete mapping of design space by using an ML model to predict the behavior of unmeasured
213 compositions. We trained two least-squares boosted RF regression models[55] with our basal and
214 induced data sets, respectively, using randomized proportional down-sampling (see **Methods**) to
215 normalize circuit representation across expression bins (**Fig. 4A middle**). Evaluation of the 80:20
216 train:test split indicated stronger predictive power for these RF models (train/test r^2 values of
217 0.88/0.83 for the basal and 0.92/0.87 for induced data sets) over regression models with linear
218 and quadratic interaction terms where parts are independently parameterized to fit the data
219 (train/test r^2 = 0.53/0.52 and 0.64/0.63 respectively). This suggests the existence of higher order
220 non-linear interactions between part categories that are better captured by the RF model. RF

221 feature importance scores were highest for part categories related to synTF function (AD, ZF
222 affinity), the synTF expression promoter, and the reporter gene core promoter.

223 Using the RF models, we predicted basal and induced expression values for the 44,596
224 unmeasured part compositions, and the 121,292 measured compositions, yielding a behavior
225 map of our entire 165,888-member design space (**Fig. 4A right**). While a comparison between
226 RF-modeled and CLASSIC-measured distributions revealed similar global features, we observed
227 a high absolute model error (>2) for compositions measured at the periphery of behavior space
228 (~3% of variants, including many in the HFC region), potentially resulting from either CLASSIC
229 measurement error or poor model prediction in these regions (**Fig. 4A right inset**). To experi-
230 mentally validate that our RF models accurately reflect the behavior landscape of our circuit de-
231 sign space, we LP-integrated unmeasured (n=15) and measured (n=81) compositions from
232 across behavior space (**Fig. 4B**). Flow cytometry measurements of the resulting cell lines showed
233 close agreement with both predicted values for both unmeasured (*MAE*=0.16) and measured
234 compositions (*MAE*=0.15), allowing us to validate numerous compositions with the highest pre-
235 dicted fold-change values in the design space (50-100x) (**Fig. 4B**). Additionally, we demonstrated
236 that cell lines harboring circuit configurations with high-error measurements that were previously
237 observed in the behavior space periphery showed close agreement with RF predictions, indicating
238 that CLASSIC measurement outliers are accurately adjusted by the models.

239 With a quantitatively accurate functional mapping of synTF circuit design space in-hand,
240 we asked whether our dataset could provide insight into the design rules underlying circuit behav-
241 ior. First, we compared the frequency of part usage between low basal, high induced, and HFC
242 regions (**Fig. 4C**). We observed distinct part usage for compositions from each region, with highly
243 asymmetric usage in specific categories, including those associated with the synTF protein (AD,
244 IDP, and ZF affinity), the synTF expression promoter, the number of BMs, core promoter identity,
245 and 5'-to-3' EU orientation. For example, VPR, a strong AD, is used extensively in the high- in-
246 duced region, but is nearly absent amongst low basal circuits in favor of the weaker AD, p65,

247 while intermediate-strength VP16 and VP64 are enriched amongst HFC circuits. By the same
248 token, lower and medium activity core promoters (miniTK and ybTATA) are excluded from high
249 activity circuits, while low basal activity circuits exclude mCMV (high activity core promoter), and
250 HFC circuits utilize a mix of ybTATA and mCMV.

251 We next assessed whether part usage co-varies between categories by computing cross-
252 category mutual information (MI) for each behavior space region (**Fig. 4D**). We observed "cou-
253 pling" between categories in all 3 regions, with the highest degree in the HFC region. The strong-
254 est observed interactions were for AD with ZF affinity and the synTF expression promoter. To
255 better understand the role that this apparent part usage interdependence plays in optimizing HFC
256 behavior, we performed UMAP-assisted K-means clustering[56] on modeled HFC compositions us-
257 ing the 7 part categories with highest MI coupling (**Fig. 4E**). Inspection of the resulting projection
258 revealed three distinct clusters that were differentiated by their specific deployment of parts. Com-
259 positions in cluster A were enriched for higher expression of lower affinity synTFs harboring all
260 four ADs (including p65 and VPR), while those in cluster B and C favor medium expression of
261 higher affinity synTFs with moderate activity ADs (VP16, VP64).

262 Our CLASSIC-enabled analysis of part composition patterns supports a set of design prin-
263 ciples for programming HFC behavior that emphasizes a careful balance between minimizing
264 basal expression and enhancing induced expression (**Fig. 4F**). Part usage in categories associ-
265 ated with circuit locus design favors reporter-to-coding EU orientation, extended (500 bp) spacer
266 regions, and moderate synTF expression; features that are also found in the low basal expression
267 region of behavior space and thus likely contribute to minimizing leaky reporter expression in the
268 absence of inducer. On the other hand, high BM valency, higher-activity core promoters, and
269 preference for terminator T7—features also enriched in the high induced region—potentially max-
270 imize reporter output in the presence of nuclear-localized synTF. As our MI and clustering anal-
271 yses demonstrated, usage of parts comprising the synTF protein employ parallel molecular solu-
272 tions—matching strong expression with low affinity (cluster A) or medium-strength activators with

273 high affinity (clusters B and C)—that likely constrain specific activity of the synTF to a regime that
274 minimizes background expression while still producing strong induction.

275 Here, we have established the feasibility of combining long- and short-read NGS modali-
276 ties to perform massively parallel quantitative profiling of multi-kb length-scale genetic part as-
277 semblies in human cells. CLASSIC holds potential as an approach for exploring the emergence
278 of function from genetic composition across a range of organizational scales and phylogenetic
279 contexts, including for viruses[57], bacterial operons[58], and chromatin domains[58-60]. As we show in
280 this piece, by enabling HT profiling of diverse combinations of genetic parts, our approach signif-
281 icantly expands the scope of inquiry for synthetic biology projects; it reduces the time and cost
282 required to identify behavior-optimized part compositions and enables discovery of guidelines for
283 part composability. As demonstrated by our identification and analysis of HFC circuit variants, the
284 design rules revealed by CLASSIC may be non-intuitive and challenging to capture using bio-
285 physical modeling alone. Furthermore, since CLASSIC leverages extant, broadly-accessible mo-
286 lecular cloning and experimental analysis pipelines, we anticipate it will increase the pace and
287 scale of genetic design for diverse synthetic biology applications across a spectrum of organismal
288 hosts, including the development of bioproduction strains[61] and multigenic cell therapy programs[5],
289 as well as other applications that require probing complex genetic design spaces to achieve multi-
290 dimensional phenotypic optima.

291 Finally, we showed that data acquired using CLASSIC can be used to train ML models to
292 accurately make predictions for out-of-sample and edge-case circuit behavior. While extensive
293 recent work has used ML approaches to develop sequence-to-function models for various classes
294 of genetic parts[23,62-64] our work serves as a starting point for developing AI-based models of gene
295 circuit function that use part compositions as learned features. While our current work has focused
296 on mapping a design space of 10^5 compositions, it may be possible to create predictive models
297 for more complex circuits with far more expansive design spaces by using data acquired with
298 CLASSIC to train high capacity deep-learning algorithms (e.g., transformers) which require much

299 larger datasets than currently exist. Such approaches could work in black-box fashion, without
300 the incorporation of regulatory or biophysical priors, or synergistically with existing mechanistic
301 frameworks to create interpretable models that provide deeper insights into genetic design.
302

303 **METHODS**

304 ***Plasmid library construction***

305 Circuit libraries were cloned using a custom hierarchical golden gate[30] assembly scheme,
306 which enables rapid, modular cloning of complex gene circuits starting from individual genetic part
307 sequences. Briefly, input DNA fragments are amplified from genomic or commercial DNA sources
308 and cloned into kanamycin resistant sub-part entry vectors using BpiI (Thermofisher) (*level 0*).
309 Sequence-verified sub-part plasmids are then used as inputs for assembly into carbenicillin-re-
310 sistant entry vectors using BsaI-v2 HF (NEB) to yield genetic part plasmids (i.e., promoters, ORFs,
311 terminators) (*level 1*). EUs are then constructed by assembling promoter, ORF, and terminator
312 part plasmids into a kanamycin-resistant entry vector using Esp3I (Thermofisher) (*level 2*). EUs
313 are then combined into multi-unit arrays by assembling into a carbenicillin– or spectinomycin-
314 resistant destination vectors using BpiI (Thermofisher) (*level 3*).

315 In this system, barcode pools are incorporated into library assemblies at level 2 and 3.
316 Level 2 barcoded EU pools were generated by first constructing a destination vector carrying a
317 ccdB placeholder expression cassette downstream of the BFP stop codon. A semi-degenerate
318 18 bp barcode oligo pool (IDT)[65] was polymerase-extended and cloned in place of the ccdB cas-
319 sette using BsaI-v2 (NEB) in a 20 µL golden gate reaction: 2 min at 37 °C and 5 min at 16 °C for
320 a number of cycles equal to 10x the number of input fragments, followed by a 30 min digestion
321 step at 37 °C and subsequent sequential 15 min denaturation steps at 65 °C and 80 °C. Resulting
322 assemblies were purified using a miniprep column (Epoch Life Sciences), electroporated (BioRad)
323 into 100 µL NEB 10-Beta electrocompetent *E. coli* (NEB), and plated onto a custom 30 in x 24 in
324 LB_{Kan} agar plate, yielding ~120M colonies. Plates were grown at 37 °C for 16 h, at which time

325 colonies were scraped into LB_{Kan} (20 mL), incubated for 30 min at 37 °C, and the plasmid library
326 extracted via miniprep (Qiagen). For large-scale libraries level 3 barcode pools were generated
327 in order to ensure a unique barcode-to-circuit mapping. To create level 3 barcode pools, a semi-
328 degenerate 18 bp oligo pool (IDT) was polymerase-extended and cloned via golden gate assem-
329 bly into a SpectR destination vector carrying a placeholder ccdB cassette using PaqCI (NEB)
330 following the assembly protocol above. Resulting assemblies were transformed into 100 µL home-
331 made chemically competent[66,67] Stbl3 *E. coli* and plated onto a 10 cm LB_{Spect} agar plate. Plates
332 were grown at 37 °C for 20 h to yield approximately 15,000 colonies, followed by colony scraping
333 and plasmid DNA extraction, as described above.

334 Part pools (levels 1), EU pools (level 2) and circuit libraries (level 3) were constructed by
335 combining input plasmids at 50 fmol per part category (15 µL total volume) using the above-
336 described cycling protocol. Transformations varied by library size: level 2 and 3 assemblies for
337 the 384-member library (**Fig. 2**) were transformed into 100 µL of chemically competent *E. coli*
338 (DH5a and Stbl3, cells/mL) and respectively plated on LB_{Kan} and LB_{Carb} agar plates for 12-16 h
339 (37 °C) to yield ~15,000 colonies. Colonies were scraped into 13 mL LB_{Kan} or LB_{Carb} and mini-
340 prepped using a Qiagen kit. Level 1 and 2 assemblies for the 166k-member library (**Fig. 3**) were
341 respectively transformed using 100 µL and 400 µL of chemically competent *E. coli* (Stbl3,
342 cells/mL) and grown on LB_{Carb} and LB_{Kan} agar plates at 37 °C for 12-16 h to yield ~5,000 and
343 ~20,000 colonies, respectively. Level 3 166k-member assemblies were purified using a miniprep
344 column (Epoch Life Sciences), electroporated (BioRad) into 100 µL 10-beta electrocompetent
345 cells (NEB), and then plated on 10 15 cm LB_{Spect} agar plates for 16-20 h (37 °C) to yield ~3M
346 colonies.

347

348 ***Long-read plasmid sequencing***

349 To generate indices linking assembled constructs to their associated DNA barcodes, we
350 used Oxford Nanopore Technology (ONT) long-read sequencing[68,69]. Libraries were prepared for

351 long-read sequencing by digesting 2 µg of a level 3 plasmid library using Esp3I (Thermofisher)
352 and purifying linearized fragments using a 0.5x volume of magnetic beads (Omega Bio-Tek) by
353 volume. Nanopore sequencing adapters were added to the linearized DNA pool using the LSK-
354 112 genomic ligation kit (ONT) and sequenced using on a 14inion device (ONT) equipped with
355 an R10 flow cell (FLO-MIN112). Basecalling was performed using Guppy (ONT, super high ac-
356 curacy mode) running on a GPU (Nvidia RTX 3090). Composition and barcode assignment was
357 performed using WIMPY (what's in my pot, y'all), a custom Matlab analysis pipeline composed of
358 six discrete modules that converts basecalled Fastq files into a part composition array and bar-
359 code sequence for each Nanopore read. Briefly, WIMPY first calls Fastqall (1) to import fastq files
360 from Guppy and Bowtiles (2) to index the reads to a constant region in the level 3 plasmid back-
361 bone. In this study, the puromycin resistance cassette (PuroR) was designated as the constant
362 region, and after indexing to PuroR, reads were filtered to remove incomplete assemblies and
363 non-library fragments. WIMPY then calls Tilepin (3), which uses localized containment searching
364 (aka tiling) to identify important genetic landmarks, such as ORFs, to define the different sections
365 of the sequence (e.g., reporter array, synthetic transcription factor coding region, or barcoded
366 BFP expression unit). Chophat (4) then truncates the previously identified sections into smaller
367 arrays to aid later identification of individual parts. WIMPY then calls Viscount (5a), which deter-
368 mines compositions for filtered reads by identifying and assigning genetic parts through a combi-
369 nation of general Smith-Waterman alignments[70] and localized containment searches. This is done
370 by splitting the reference sequence into 6-10bp "tiles" with a 1bp stride, after which nanopore
371 reads are queried for the number of tiles contained for each part reference sequence. Reads
372 containing >3% of tiles for a reference sequence are assigned while reads with more than one
373 reference assignment are discarded. Barcode sequences for each read are concurrently deter-
374 mined by Barcoat (5b), in which the sequence section downstream of the BFP EU is aligned to a
375 degenerate reference sequence using a custom alignment matrix.
376

377 ***Cell culture***

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

386

387 ***Single-copy library integration***

388 To establish a landing pad (LP) cell line, low passage HEK293T cells were co-transfected
389 a dual Cas9 and gRNA expression vector (pX330-T2) targeting the human *AAVS1* locus[71] and a
390 PmeI-linearized (NEB) repair template comprising a constitutively expressed YFP-HygR-expres-
391 sion cassette with a BxB1 *attP* recognition site flanked by 800bp homology arms for the human
392 *AAVS1* locus (pROC079). Genomic integration events were selected using 50 µg/mL Hygromycin
393 B, and YFP+ cells were flow sorted to isolate clones (WOLF, NanoCellect). Clones were ex-
394 panded to 24-well plates and tested for integration competency by integration of an *attB*-contain-
395 ing destination vector and subsequent flow cytometry and the presence of the intact LP cassette
396 at the *AAVS1* locus by PCR. gDNA was extracted from each clonal population (Omega Bio-Tek)
397 and 50 ng of gDNA was used as the template in a 25 µL reaction along with 0.25 µL of Q5 poly-
398 merase (NEB), 0.5 µL of 10mM dNTPs, 5 µL of 5X Q5 polymerase buffer, and 1 µL of each 1 mM
399 genotyping. Primers were specific to the native AAVS1 locus, Landing Pad, or destination vector.
400 The thermocycling protocol was 2 min at 98 °C, then 30 cycles of 98 °C for 20 s, 65 °C for 20 s,
401 72 °C for 1 min and then a final step of 72 °C for 5 min. We identified a clone (LP79 #36) that

402 demonstrated BxB1-dependent, single copy integration of vectors harboring an *attB* recognition
403 site and used this cell line for all following experiments.

404 To integrate individual constructs, 125k cells were plated into a 24-well plate one day prior
405 to transfection. The well was co-transfected with 250 ng of a level 3 vector containing a construct
406 of interest and 125 ng of BxB1 expression plasmid (Addgene #51271) using JetPrime (VWR).
407 pCAG-NLS-HA-Bxb1 was a gift from Pawel Pelczar (Addgene plasmid #51271)[72]. Two days after
408 transfection, cells were passaged into complete DMEM containing 1 µg/mL Puromycin and se-
409 lected for 10 days to achieve homogenous expression. Library integration was performed by co-
410 transfecting HEK293T-LP with 250 ng of plasmid library and 125 ng of BxB1 per 200k cells (384-
411 member library = 1M cells transfected; 166k-member library = 100M cells transfected) using
412 JetPrime. Media exchange was performed after 8 h and cells were expanded for an additional 40
413 h, split into five culture dishes, and grown for 6-8 days under Puromycin selection. Cells were
414 passaged 1:3 and cultured for an additional 5 days under puromycin selection, and then combined
415 in complete DMEM for flow sorting (10M cells/mL).

416 To determine the clonal variability within LP-integrated cell populations, we integrated a
417 constitutively expressed mCherry EU following the protocol described above. We then randomly
418 sorted 23 clones from the population (WOLF, NanoCellect) and measured their mCherry expres-
419 sion. We calculated the MAE for this set (*MAE = 0.08*) and used this value to define an expected
420 error range from clonal heterogeneity (ERCH), which represents the expected variability of cells
421 from an integrated population. For the 384-member library, the expression ERCH was set to 0.08
422 on either side of the expected value, and for the 166k-member library, the fold change ERCH was
423 calculated to be 0.16 on either side of the expected value.

424

425 ***Flow sorting***

426 To prepare libraries for flow sorting, cells were lifted using TrypLE (Gibco), washed twice
427 with PBS, and resuspended in complete DMEM (10M cells/mL). For the 384-member library, the

428 mRuby expression distribution was sorted into 10 evenly log-spaced bins on a Sony MA900 set
429 to purity mode. The number of cells collected for each bin was proportional to the percentage of
430 the mRuby distribution expressed in that bin, with a target of ~125k cells sorted in the most pop-
431 ulous bin [e.g., 2,252 cells were collected for bin 1 (0.53% of library), 126,562 for bin 4 (29.78%
432 of library)]. For the 166k-member library, cells were split and either grown in DMEM containing
433 1μM 4-OHT (Sigma Aldrich) ("induced") or without 4-OHT ("uninduced") for 72 h. The eGFP ex-
434 pression distributions for both conditions were then sorted separately into 8 evenly log-spaced
435 bins. Cells were sorted into bins in proportion to the relative abundance of the library in each bin,
436 with the most populous bin set at 1.5M cells. For both libraries, sorted cells were plated into 96-,
437 24- or 6-well plates to achieve a plating density of 10-30%, grown under puromycin selection for
438 2 days, passaged and washed with PBS, and grown for an additional 3 days. Cells were then
439 lifted using TrypLE and total RNA was extracted using the Takara RNA plus kit following manu-
440 facturer's instructions (Takara). 5 μL of total mRNA was converted to total cDNA using random
441 hexamers following manufacturer's instructions (Verso, Life Technologies).
442
443 ***Short-read (Illumina) sequencing***
444 To prepare sorted libraries for short-read sequencing, barcode regions were PCR ampli-
445 fied from cDNA in a 15 μL reaction with Phantamax polymerase (Vazyme) and custom bin-specific
446 primer sets (IDT) using the following thermocycling protocol: 2 min at 98 °C, then 12 cycles of
447 98 °C for 20 s, 60 °C for 10 s, 66 °C for 10 s, 72 °C for 20 s and then a final step of 72 °C for 2
448 min. The resulting amplicon pool was extracted from a 2% agarose gel. A second PCR step,
449 which used 10 cycles of the thermocycling protocol for PCR step 1, added Illumina sequencing
450 adapters (i5 and i7) and sample-specific sequencing barcodes using Phantamax and custom pri-
451 mer sets (IDT), and the amplified product was extracted from a 2% agarose gel. Purified ampli-
452 cons from each bin were pooled at equimolar concentrations and sequenced using an Illumina
453 MiSeq (kit v2, 300 cycles) or NovaSeq6000 (Sp v1.5 kit, 300 cycles) for the 384- and 166k-

454 member libraries, respectively. Illumina data were analyzed using a custom analysis pipeline
455 (Matlab). Briefly, fastq files are imported and split according to their sample-specific sequencing
456 barcode. Barcode sequencing data is converted to average expression for all compositions/vari-
457 ants as follows: (1) Counts for each barcode in each sample/bin are normalized by the percentage
458 of the library distribution in that bin to account for differences in read depth across bins. (2) A
459 weighted average of number of reads in each bin for each barcode is then calculated, to assign
460 an average expression level (using a bin → expression conversion table) to each barcode. (3)
461 Barcode expression levels are then cross-referenced with the Nanopore indices to assign average
462 expression values for all barcodes associated with a given variant in a n-by-y cell array, where n
463 is the number of library variants and y is the (variable) number of barcodes for each variant. (4)
464 For each variant, a gaussian kernel density estimation is performed over the log-normalized ex-
465 pression values from all barcodes associated with each variant, using the Matlab function
466 *ksdensity*. The expression value corresponding to the kernel peak is assigned as the mean ex-
467 pression value for each variant.
468
469 ***Experimental validation of CLASSIC measurements***
470 To assess the accuracy of CLASSIC measurements for both the 384-member and 166k-
471 member libraries, single cells were sorted from the integrated cell populations into 96-well plates
472 (MA900) and clonally expanded from across the mRuby and eGFP expression distributions, re-
473 spectively. Isolated clonal populations, which are referred to as 'isolates', were expanded to 24-
474 well plates and the fluorescence distribution was measured by flow cytometry (MA900). For the
475 384-member library, one 96-well plate of cells was sorted based on mRuby expression and 15
476 isolates were taken for validation. For the 166k-member library, two 96-well plates of cells were
477 sorted based on eGFP expression from the uninduced cell library to yield 40 isolates total. Fol-
478 lowing fluorescence measurement, gDNA was extracted from each cell population (Omega Bio-
479 Tek). The genetic variant in each isolate was identified by PCR-based amplification and Nanopore

480 sequencing of the Landing Pad locus using a forward primer in Puromycin resistance cassette
481 and a reverse primer in the BFP open reading frame. A set of 30 different forward primers, each
482 with a defined barcode, were paired with a common reverse primer to amplify the genetic variants
483 from each isolate and associate the genetic variant with the isolate during Nanopore sequencing.
484 50 ng of gDNA was used as the template in a 20 µL reaction along with 10 µL of 2X PhantaMax
485 Master Mix (Vazyme), 1 µL of each 1 mM primer, and brought up to 20 µL in nuclease-free H_2O.
486 The thermocycling protocol was 2 min at 98 °C, then 30 cycles of 98 °C for 20 s, 65 °C for 20 s,
487 72 °C for 5 min and then a final step of 72 °C for 10 min. Each PCR reaction was extracted from
488 a 1% agarose gel and column purified using QG buffer (Qiagen), following manufacture instruc-
489 tions. Purified amplicons were pooled and prepared for Nanopore sequencing using LSK-112
490 genomic ligation kit (ONT) and sequenced using on a 19inion device (ONT) equipped with an R10
491 flow cell (FLO-MIN112). Basecalling was performed using Guppy (ONT, super high accuracy
492 mode) running on a GPU (Nvidia RTX 3090). Composition and barcode assignment was per-
493 formed using WIMPY, as described above.
494
495 ***Random forest regression***
496 For the 384-member EU library, a bootstrap aggregation (bag) RF model was constructed
497 using the matlab function *templatetree*. The promoter, Kozak and terminator were used as cate-
498 gorical input variables for each EU, with log(mRuby) expression (AU) as the predicted output. The
499 data was split 80:20 (307 EUs and 77 EUs) for training and testing, respectively. The number of
500 trees was chosen based on 10-fold cross validated RMSE loss as a function of increasing number
501 of trees using the Matlab function *fitrensemble*. The trained RF model was then validated on the
502 test set (**Fig. 2F**). Identification of genetic part interactions was determined by the absolute log
503 error between predicted and observed mRuby expression.
504 For the 166k-member circuit library, two independent least-squares boosted RF models
505 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 12 or more barcode reads. The space was then split into 8 equally log-spaced regions. If the bin with the highest frequency of data contained more than 2x the number of points as the next most frequent bin, data was randomly downsampled from that bin to match the frequency of the next most frequent bin. This downsampled dataset was then randomly split 80:20 into a training and test set respectively. 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 decision splits, the model structures and performance were as follows: basal, 70 trees, max number of splits=51, learn rate = 0.1061, r^2_{train}=0.88, r^2_{test}=0.83 and induced, 100 trees, max number of splits=51, learn rate=0.1556, r^2_{train}=0.92, r^2_{test}=0.87.

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, part composition data for the seven genetic part categories with the greatest part usage asymmetry (promoter, AD, IDR, ZF affinity, BM #, core

promoter, and orientation) from the high fold change (>25-fold region) were used as inputs into the *clusterevaluation* function in Matlab. We carried out a gap test to identify the optimal number of clusters in the data, followed by K-means clustering using the *kmeans* function in Matlab to identify clusters. Cluster regions were generated by drawing a boundary around the outermost points corresponding to a particular cluster.

Data analysis and statistics

All statistical analyses and graphical displays were performed in Matlab (2022a). Pearson's r^2 was used to calculate correlations between CLASSIC replicate experiments, CLASSIC expression measurements and model predictions, and individual expression measurements and their model predictions. The MAE was calculated as the mean of the absolute differences of the $\log_{10}$ transformed expression between every pair of isolate and CLASSIC measurements. The MAE of the $\log_{10}$ transformed fold change values was computed between CLASSIC measurements and individual measurements as well as model predictions and individual measurements.

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. P.M. and J.W.R were supported by NIH R35GM119461 (P.M.)

AUTHOR CONTRIBUTIONS

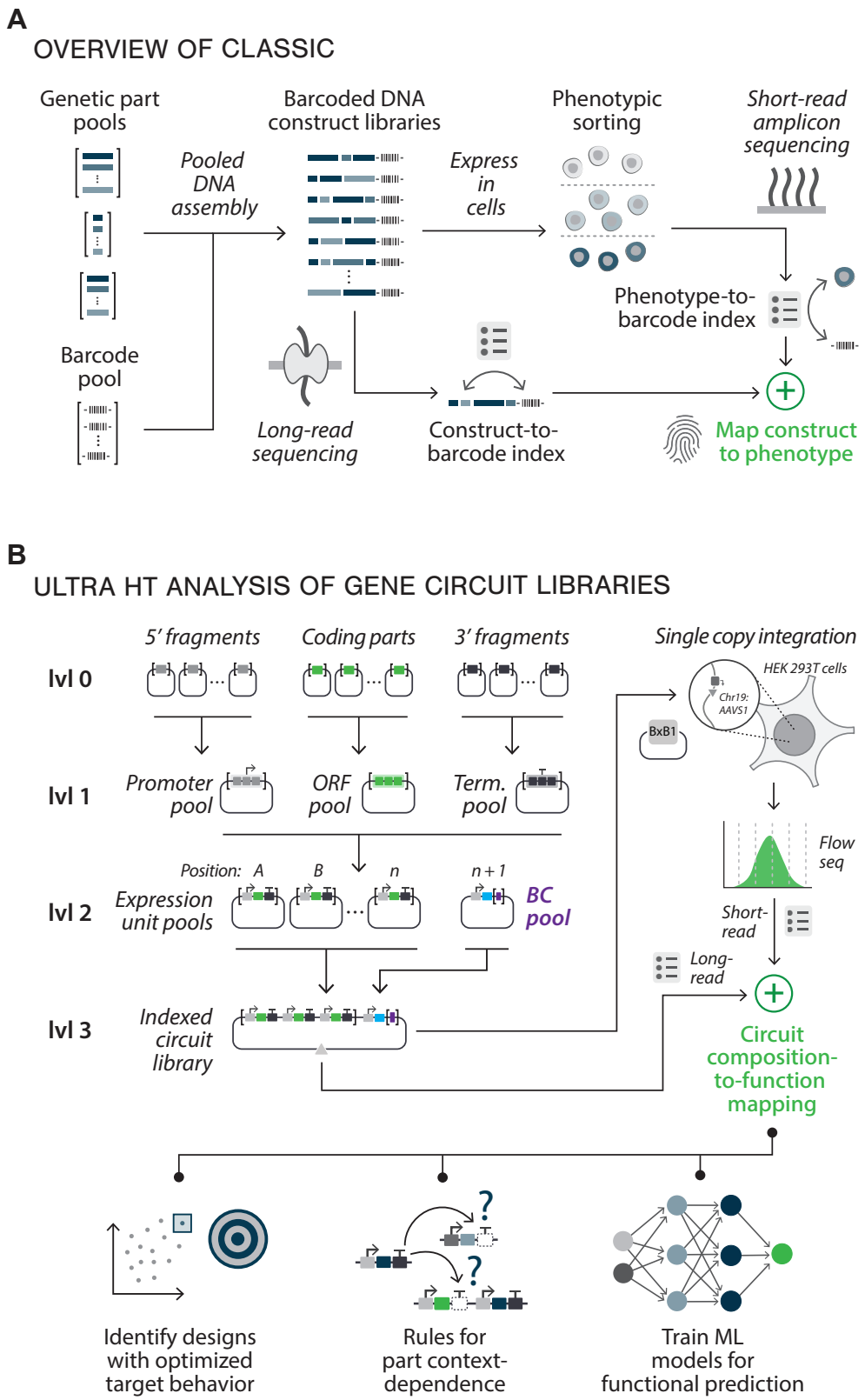
R.W.O., K.R., and C.J.B. conceived of the study, R.W.O. and K.R. carried out the experiments and developed the analysis software, with assistance from T.C.P., Y.W., K.D.S., J.A.W., S.L., T.H.Z., E.M.R., and A.S. R.W.O., and T.C.P. developed the modular cloning scheme and landing pad cell line, with assistance 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., Y.W., 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.

CORRESPONDING AUTHOR

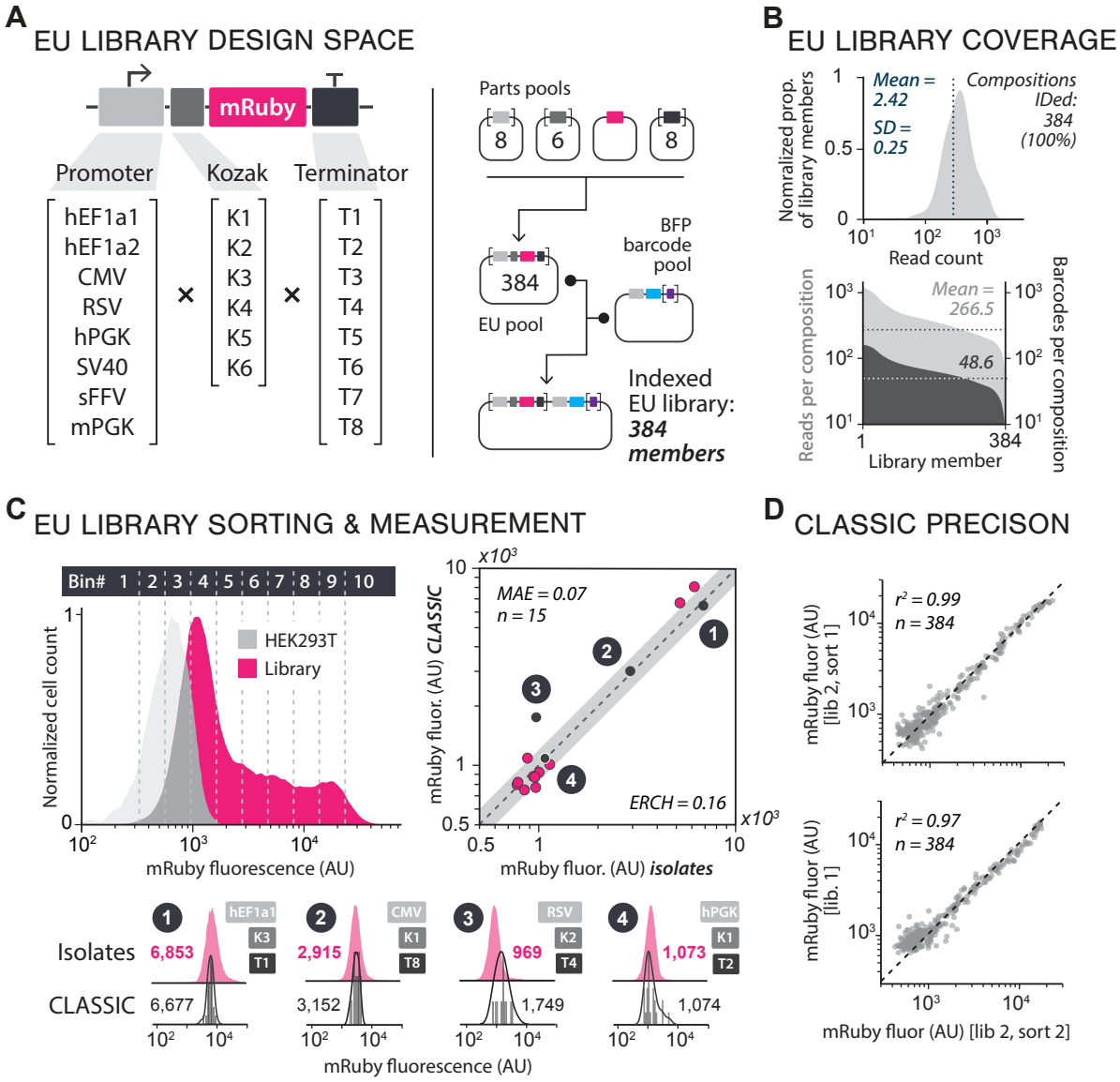
Correspondence and requests for materials should be addressed to Caleb J. Bashor.

COMPETING INTERESTS

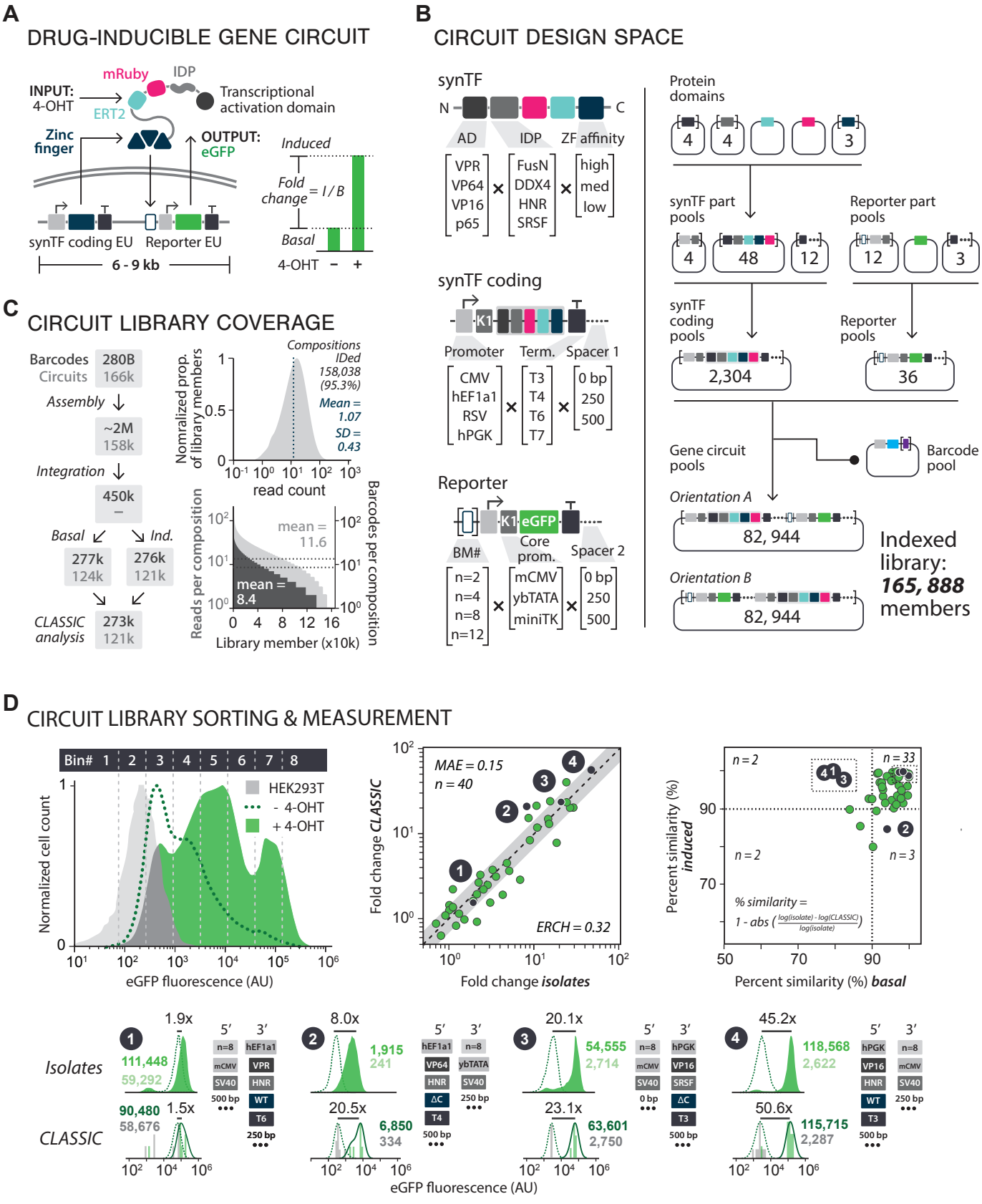
The authors declare no competing interests.


A
OVERVIEW OF CLASSIC
Genetic part pools
Barcoded DNA construct libraries
Phenotypic sorting
Short-read amplicon sequencing
Pooled DNA assembly
Express in cells
Phenotype-to-barcode index
Barcode pool
Long-read sequencing
Construct-to-barcode index
Map construct to phenotype
B
ULTRA HT ANALYSIS OF GENE CIRCUIT LIBRARIES
5′ fragments
Coding parts
3′ fragments
Single copy integration
lvl 0
HEK 293T cells
Chr19: AAVS1
BxB1
lvl 1
Promoter pool
ORF pool
Term. pool
Flow seq
Position: A
B
n
n + 1
lvl 2
Expression unit pools
BC pool
Short-read
Long-read
lvl 3
Indexed circuit library
Circuit composition-to-function mapping
Identify designs with optimized target behavior
Rules for part context-dependence
Train ML models for functional prediction

582 **Figure 1. Using CLASSIC to systematically map the design space of complex genetic pro-**
583 **grams.** (**A**) Overview of CLASSIC. Pooled assembly of genetic parts with DNA barcode se-
584 quences yields libraries of barcode-indexed constructs of arbitrary length and complexity. Long-
585 read nanopore sequencing is used to create an index matching construct composition to an as-
586 sociated barcode. In parallel, libraries introduced into cells undergo sorting or selection to bin
587 expression phenotypes. Barcode amplicons generated for each bin are subjected to short-read
588 NGS to quantify expression phenotype, which is then mapped to construct composition via bar-
589 code indexes. (**B**) Application of CLASSIC to profile a synthetic gene circuit design space. Hier-
590 archical golden gate assembly is used to compose libraries of multi-EU circuits with combinatori-
591 ally varied part compositions and circuit designs. Sequence fragment pools for different parts
592 categories from level 0 are combined to yield level 1 pools of promoters, open reading frames
593 (ORF), and terminators (term.), which are then combined to yield level 2 EUs (square brackets:
594 fragment/part pools). Barcode pools are combined with EU pools to create indexed multi-EU cir-
595 cuit libraries (level 3) that are integrated into HEK293T-LP cells placed at the *AAVS1* locus in
596 chromosome 19 via expression of the BxB1 recombinase (top right). Library analysis by a combi-
597 nation of nanopore and flow-seq yields composition-to-function mapping.
598
599
600
601
602
603
604
605


A
EU LIBRARY DESIGN SPACE
mRuby
Promoter
Kozak
Terminator
hEF1a1
hEF1a2
CMV
RSV
hPGK
SV40
sFFV
mPGK
K1
K2
K3
K4
K5
K6
T1
T2
T3
T4
T5
T6
T7
T8
Parts pools
BFP barcode pool
EU pool
Indexed EU library: 384 members
B
EU LIBRARY COVERAGE
Mean = 2.42
SD = 0.25
Compositions IDed: 384 (100%)
Nomralized prop. of library members
Read count
Mean = 266.5
48.6
Reads per composition
Barcodes per composition
Library member
C
EU LIBRARY SORTING & MEASUREMENT
Bin# 1 2 3 4 5 6 7 8 9 10
HEK293T
Library
Normalized cell count
mRuby fluorescence (AU)
MAE = 0.07
n = 15
ERCH = 0.16
mRuby fluor. (AU) CLASSIC
mRuby fluor. (AU) isolates
Isolates
CLASSIC
6,853
6,677
2,915
3,152
969
1,749
1,073
1,074
D
CLASSIC PRECISON
r² = 0.99
n = 384
r² = 0.97
n = 384
mRuby fluor (AU) [lib 2, sort 1]
mRuby fluor (AU) [lib. 1]
mRuby fluor (AU) [lib 2, sort 2]
E
EU EXPRESSION DISTRIBUTIONS
Promoters
Kozaks
Terminators
mRuby fluorescence (AU)
F
EU MACHINE LEARNING MODEL
RF predicted expression (AU)
CLASSIC expression (AU)
Training set: 307
r² = 0.98
Test set: 77
r² = 0.96
10-fold CV MSE
Learning cycles
Importance
Prom.
K
Term
G
DETECTING PART INTERFERENCE
Error (predicted - observed)

606 **Figure 2. CLASSIC can quantitatively profile diverse compositions of genetic parts.** (**A**) EU
607 library design space. Left: the EU library comprises combinations of parts from 3 categories (pro-
608 moters, Kozak sequences, and terminators). Right: Part pools are assembled (step 1) with an
609 mRuby ORF to generate a 384-member EU pool, which is combined with a barcoded-BFP EU
610 pool to generate the indexed EU library. (**B**) EU library assembly and indexing balance. Oxford
611 nanopore sequencing was performed on the assembled library. Data were analyzed to assess
612 library composition count (top) and composition/barcode balance [reads per composition (light
613 grey), unique barcodes per composition (dark grey); data plotted in rank order of reads per com-
614 position] (bottom). (**C**) EU library expression quantification. Top left: the library was flow sorted
615 into 10 equally log-spaced bins [empty HEK293T cells (grey histogram), library (pink histogram)].
616 Top right: residual for FACS-measured geometric mean fluorescence values for sorting-isolated
617 clones (n=15) plotted against CLASSIC-derived values [ERCH (grey band), error range from
618 clonal heterogeneity (see **Methods**)]; black dots, clonal isolates. Bottom: Representative data
619 from 4 clonal isolates (pink solid) are shown with corresponding CLASSIC-computed distributions
620 [kernel density (black line) calculated from normalized barcode read count (grey vertical bars)];
621 MAE, mean absolute error; AU, arbitrary fluorescence units. (**D**) CLASSIC precision. Correlation
622 of CLASSIC-computed EU expression between technical replicate (top) and biological replicate
623 (bottom) experiments. (**E**) Influence of part identity on expression. Violin plots of each part-specific
624 distribution, ordered from strongest to weakest. Dotted line, HEK293T background expression
625 mean. (**F**) RF analysis of EU behavior space. Left: RF modeled values plotted against CLASSIC
626 measurements. Grey, training data; purple, test data. Top right: 10-fold cross-validated error (y
627 axis) vs number of learning cycles (x axis). CV, cross-validation; MSE, mean squared error. Bot-
628 tom right: feature importance scores for each part category. (**G**) Analysis of part interference. Left:
629 error [$\log_{10}$(RF predicted) – $\log_{10}$(CLASSIC-observed)] is plotted for all compositions associated
630 with each promoter. White lines, median values; red dots, outliers (error >0.025). Right: outlier
631 part configurations. Red shaded area highlights common terminator T8.


A
DRUG-INDUCIBLE GENE CIRCUIT
INPUT: 4-OHT
mRuby
IDP
ERT2
Transcriptional activation domain
Zinc finger
OUTPUT: eGFP
synTF coding EU
Reporter EU
6 - 9 kb
Induced
Fold change = I / B
Basal
4-OHT − +
B
CIRCUIT DESIGN SPACE
synTF
N
C
AD
IDP
ZF affinity
VPR VP64 VP16 p65
FusN DDX4 HNR SRSF
high med low
synTF coding
K1
Promoter
Term.
Spacer 1
CMV hEF1a1 RSV hPGK
T3 T4 T6 T7
0 bp 250 500
Reporter
K1
eGFP
BM#
Core prom.
Spacer 2
n=2 n=4 n=8 n=12
mCMV ybTATA miniTK
0 bp 250 500
Protein domains
synTF part pools
Reporter part pools
synTF coding pools
Reporter pools
Gene circuit pools
Barcode pool
Orientation A
82, 944
Orientation B
82, 944
Indexed library: 165, 888 members
C
CIRCUIT LIBRARY COVERAGE
Barcodes 280B
Circuits 166k
Assembly
~2M 158k
Integration
450k –
Basal
Ind.
277k 124k
276k 121k
CLASSIC analysis
273k 121k
Compositions IDed 158,038 (95.3%)
Mean = 1.07
SD = 0.43
Nomralized prop. of library members
read count
mean = 11.6
mean = 8.4
Reads per composition
Barcodes per composition
Library member (x10k)
D
CIRCUIT LIBRARY SORTING & MEASUREMENT
Bin# 1 2 3 4 5 6 7 8
HEK293T
- 4-OHT
+ 4-OHT
Normalized cell count
eGFP fluorescence (AU)
MAE = 0.15
n = 40
ERCH = 0.32
Fold change CLASSIC
Fold change isolates
n = 2
n = 33
n = 2
n = 3
% similarity = 1 - abs (log(isolate) - log(CLASSIC) / log(isolate))
Percent similarity (%) induced
Percent similarity (%) basal
Isolates
CLASSIC
1.9x
111,448
59,292
90,480
58,676
1.5x
8.0x
1,915
241
20.5x
6,850
334
20.1x
54,555
2,714
23.1x
63,601
2,750
45.2x
118,568
2,622
50.6x
115,715
2,287
5′ 3′
n=8 hEF1a1
mCMV VPR
SV40 HNR
WT
T6
500 bp
250 bp
hEF1a1 n=8
VP64 ybTATA
HNR SV40
ΔC
T4
500 bp
250 bp
hPGK n=8
mCMV VP16
SV40 SRSF
ΔC
T3
0 bp
500 bp
hPGK n=8
VP16 mCMV
HNR SV40
WT
T3
500 bp
250 bp
eGFP fluorescence (AU)

632 **Figure 3. Using CLASSIC to profile a synthetic gene circuit design landscape.** (**A**) Inducible
633 synTF circuit diagram. Left: The circuit contains two EUs: one codes for the synTF and the other
634 is an eGFP reporter. Without inducer, the synTF localizes to the cytoplasm. Upon addition of 4-
635 OHT (input), the synTF translocates into the nucleus to bind to and activate reporter eGFP ex-
636 pression (output). Right: Expression fold change is the ratio of expression levels in the presence
637 or absence of inducer. (**B**) synTF circuit design space. Left: SynTF diversity (arranged N-to-C
638 term): of 4 ADs, 4 IDPs, and 3 ZF affinity mutants (high: WT, med: ΔC, low: ΔC 6x). synTF coding
639 EU diversity: 4 constitutive promoters, 4 terminators. Reporter EU diversity: 4 BM number vari-
640 ants, 3 minimal core promoters. This unit contains a constant terminator. Both EUs have 3' spac-
641 ing sequences of 0bp, 250bp or 500bp downstream. Right: EUs assembled from input parts and
642 combined in 2 different 5'-to-3' orientations, to generate a combinatorial diversity of 165,888 pos-
643 sible circuit variants. (**C**) Balance of circuit library assembly and indexing. Left: data for ~121k out
644 of 166k variants (73%) were recovered by CLASSIC; compositions, light grey; barcodes, dark
645 grey. Right: Nanopore sequencing data of the library were analyzed to assess library composition
646 count (top) and composition/barcode balance [reads per composition (light grey), unique bar-
647 codes per composition (dark grey); data plotted in rank order of reads per composition] (bottom).
648 (**D**) Circuit library sorting and measurement. Top right: eGFP expression of circuit library in pres-
649 ence (solid green) and absence (dotted green line) of 4-OHT (top, center) is shown, along with
650 boundaries of flow sorting bins (vertical grey dotted lines); grey histogram, empty HEK293T-LP
651 cells. Top middle: Fold-change values for 40 clonal isolates plotted against CLASSIC-derived
652 values. Black dots, isolates displayed at the bottom of the panel; MAE, mean average error; grey
653 region, ERCH, error range of clonal heterogeneity AU, arbitrary fluorescence units. Top right: %
654 similarity between basal and induced values for CLASSIC and isolates. Bottom: Flow data from
655 4 isolates (green dotted line, uninduced; green solid, induced) are shown with corresponding
656 CLASSIC-computed distributions [kernel density (black line) calculated from normalized barcode
657 read count (grey vertical bars)]. Parts combinations corresponding to each index are shown.

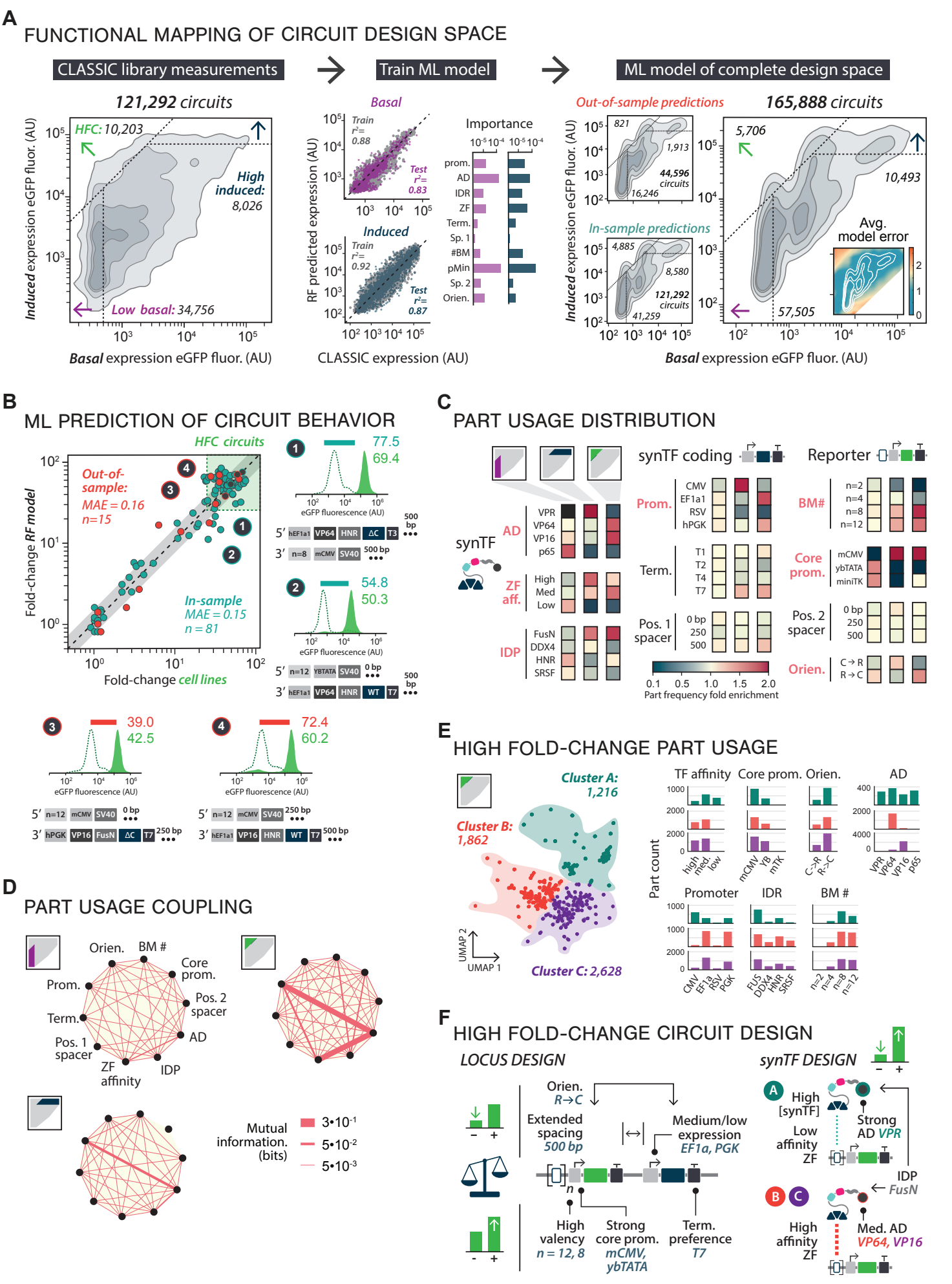

A
FUNCTIONAL MAPPING OF CIRCUIT DESIGN SPACE
CLASSIC library measurements
Train ML model
ML model of complete design space
121,292 circuits
HFC: 10,203
High induced: 8,026
Low basal: 34,756
Induced expression eGFP fluor. (AU)
Basal expression eGFP fluor. (AU)
Basal
Train r²= 0.88
Test r²= 0.83
Induced
Train r²= 0.92
Test r²= 0.87
RF predicted expression (AU)
CLASSIC expression (AU)
Importance
prom.
AD
IDR
ZF
Term.
Sp. 1
#BM
pMin
Sp. 2
Orien.
Out-of-sample predictions
821
1,913
44,596 circuits
16,246
In-sample predictions
4,885
8,580
121,292 circuits
41,259
165,888 circuits
5,706
10,493
57,505
Avg. model error
B
ML PREDICTION OF CIRCUIT BEHAVIOR
HFC circuits
Out-of-sample: MAE = 0.16 n=15
In-sample MAE = 0.15 n = 81
Fold-change RF model
Fold-change cell lines
77.5
69.4
54.8
50.3
39.0
42.5
72.4
60.2
eGFP fluorescence (AU)
5′ hEF1a1 VP64 HNR ΔC T3
3′ n=8 mCMV SV40
5′ n=12 YBTATA SV40
3′ hEF1a1 VP64 HNR WT T7
5′ n=12 mCMV SV40
3′ hPGK VP16 FusN ΔC T7
3′ hEF1a1 VP16 HNR WT T7
500 bp
0 bp
250 bp
C
PART USAGE DISTRIBUTION
synTF coding
Reporter
synTF
AD
VPR
VP64
VP16
p65
ZF aff.
High
Med
Low
IDP
FusN
DDX4
HNR
SRSF
Prom.
CMV
EF1a1
RSV
hPGK
Term.
T1
T2
T4
T7
Pos. 1 spacer
0 bp
250
500
BM#
n=2
n=4
n=8
n=12
Core prom.
mCMV
ybTATA
miniTK
Pos. 2 spacer
Orien.
C→R
R→C
0.1 0.5 1.0 1.5 2.0
Part frequency fold enrichment
D
PART USAGE COUPLING
Orien.
BM #
Core prom.
Prom.
Pos. 2 spacer
Term.
AD
Pos. 1 spacer
ZF affinity
IDP
Mutual information. (bits)
3•10⁻¹
5•10⁻²
5•10⁻³
E
HIGH FOLD-CHANGE PART USAGE
Cluster A: 1,216
Cluster B: 1,862
Cluster C: 2,628
UMAP 2
UMAP 1
TF affinity
Core prom.
Orien.
AD
Promoter
IDR
BM #
Part count
high
med.
low
mCMV
YB
mTK
C→R
R→C
VPR
VP64
VP16
p65
CMV
EF1a
RSV
PGK
FUS
DDX4
HNR
SRSF
n=2
n=4
n=8
n=12
F
HIGH FOLD-CHANGE CIRCUIT DESIGN
LOCUS DESIGN
synTF DESIGN
Orien. R→C
Extended spacing 500 bp
Medium/low expression EF1a, PGK
High valency n = 12, 8
Strong core prom. mCMV, ybTATA
Term. preference T7
High [synTF]
Low affinity ZF
Strong AD VPR
IDP FusN
High affinity ZF
Med. AD VP64, VP16

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 (Manhattan distance) between the $\log_{10}$(CLASSIC-derived measurements) and $\log_{10}$(RF-computed values). (**B**) Experimentally validating ML prediction of circuit function. Top left: Fold-change values for cell lines from unmeasured configurations (red) and measured configurations (teal) 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; green box, HFC behavior region. Right and bottom: Flow cytometry analysis of cell line behavior (green dotted line, basal; solid green, induced) and associated circuit compositions (below each plot). Horizontal bars indicate RF predicted fold change. 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 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).
26 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).
27 Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874-882, doi:10.1038/s41588-018-0122-z (2018).
28 Castellanos-Rueda, R. *et al.* speedingCARs: accelerating the engineering of CAR T cells by signaling domain shuffling and single-cell sequencing. *Nat Commun* **13**, 6555, doi:10.1038/s41467-022-34141-8 (2022).
29 Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *mSystems* **3**, doi:10.1128/mSystems.00143-17 (2018).
30 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).
31 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).
32 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).
33 Norrman, K. *et al.* Quantitative comparison of constitutive promoters in human ES cells. *PLoS One* **5**, e12413, doi:10.1371/journal.pone.0012413 (2010).
34 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).
35 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).
36 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).
37 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).
38 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).
39 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).

40 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).
41 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).
42 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).
43 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).
44 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).
45 Muldoon, J. J. *et al.* Model-guided design of mammalian genetic programs. *Sci Adv* **7**, doi:10.1126/sciadv.abe9375 (2021).
46 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).
47 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).
48 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).
49 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).
50 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).
51 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).
52 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).
53 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).
54 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).
55 Mehta, P. *et al.* A high-bias, low-variance introduction to Machine Learning for physicists. *Phys Rep* **810**, 1-124, doi:10.1016/j.physrep.2019.03.001 (2019).
56 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>.
57 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).
58 Brophy, J. A. & Voigt, C. A. Principles of genetic circuit design. *Nat Methods* **11**, 508-520, doi:10.1038/nmeth.2926 (2014).
59 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).
60 Pinglay, S. *et al.* Synthetic regulatory reconstitution reveals principles of mammalian Hox cluster regulation. *Science* **377**, eabk2820, doi:10.1126/science.abk2820 (2022).

61 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).
62 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).
63 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).
64 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).
65 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).
66 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).
67 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).
68 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).
69 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).
70 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).
71 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).
72 Hermann, M. *et al.* Binary recombinase systems for high-resolution conditional mutagenesis. *Nucleic Acids Res* **42**, 3894-3907, doi:10.1093/nar/gkt1361 (2014).