Supplementary Material

Accurate Predictions of Genetic Circuit Behavior from Part Characterization and Modular Composition

Noah Davidsohn^{*}, Jacob Beal^{*}, Samira Kiani, Aaron Adler, Fusun Yaman, Yinqing Li, Zhen Xie, Ron Weiss

Contents

1	Abbreviations	2
2	Repressor-Promoter Design	2
3	Fluorescent Protein Conversion/Variation	2
4	Estimation of Cell Division Rate	3
5	Prediction Model Details 5.1 Modeling Alternatives and Extensions	$\frac{4}{7}$
6	Details of Dose-Response Assays and Cascade Predictions	7
7	Comparison with Hill Function Models	18
8	Plasmid Sequences	19
9	Additional Experimental Methods 9.1 Feed-Forward Circuit Transfections 9.2 Vector Details	22 22 22
10	Plasmid Copy Variance in Co-transfection	24
11	Computation of Sample Statistics	25
12	Cross-Batch Compensation	26
13	Internal Cross-Validation	27

^{*}Equal Contribution



Supplementary Figure 1: Hybrid promoter architecture (highlighted in blue is the TATA box)

1 Abbreviations

Throughout this supplemental information, the following abbreviations are used for fluorescent proteins:

- IFP: Input Fluorescent Protein. In this manuscript, IFP is always EBFP2.
- OFP: Output Fluorescent Protein. In this manuscript, OFP is always EYFP.
- CFP: Constitutive Fluorescent Protein. In this manuscript, CFP is always mKate.

2 Repressor-Promoter Design

We designed hybrid promoters for use in our circuits that take as an input both an activator protein and a repressor protein (Supplementary Figure 1). There are five binding sites for VP16Gal4 activator to turn on expression, followed by a minimal promoter surrounded by two repressor binding domains. The repressor binding domains were placed to interfere maximally with transcriptional binding proteins associated with the minimal promoter. We constitutively express VP16Gal4 in all circuits described in the manuscript, and therefore we can treat this hybrid complex as a repressor, even though this regulatory relation is capable of more complex behavior when both inputs are varied. Sequences are given in Supplementary Section 8.

3 Fluorescent Protein Conversion/Variation

Our approach requires high-precision quantitative data, and as a result, we need to carry out additional experiments and perform special processing of the raw flow cytometry data. We require that the flow cytometry readings are commensurate despite the fact that multiple fluorescent proteins are involved. Note that data on protein excitation dynamics and emission spectrum is not sufficient for our purposes, as the goal is to obtain equivalence between proteins as expressed under control of equivalent promoters in the cellular context. Because expression of particular proteins may be affected by the context, conversion coefficients must be determined empirically through controls.

Measurements obtained from flow cytometry are not immediately usable: they are incommensurate and arbitrary, depending both on the instrument's configuration and measurement settings and also on the behavior of each particular fluorescent protein under the characterization protocol. Unit calibration (Beal et al., 2012) maps all measurements to the same standardized MEFL units (Molecules of Equivalent FLuorescein) (SpheroTech, 2001) and compensates for systematic variation between data sets.

At low fluorescence levels, autofluorescence is a significant component of the observed signal. We therefore begin by subtracting expected autofluorescence (computed as a normal distribution fit to flow cytometry data from HEK293 cells transfected for 72 hours with blank plasmid pExp_EMPTY). We then compensate linearly for spectral overlap, as determined for each fluorescent protein from flow cytometry data from HEK293 cells transfected for 72 hours with a single fluorescent plasmid, pExp_CAG:mKate, pExp_CAG:EBFP2, or pExp_CAG:EYFP per the protocol described in (Beal et al., 2012). Note that the spectral overlap between our selected fluorescent proteins is small, but significant when trying to obtain accurate measurements of fluorescent proteins with orders of magnitude difference between their expression levels.

Next, we convert compensated EBFP and mKate arbitrary units to equivalent EYFP arbitrary units. This is a linear conversion with the scale computed from flow cytometry data from HEK293 cells transfected for 72 hours with parallel constitutive expression of plasmids for all three fluorescent proteins (pExp_CAG:mKate, pExp_CAG:EBFP2, pExp_CAG:EYFP), per the protocol described in (Beal et al., 2012). Finally, we convert EYFP arbitrary units to MEFL using flow cytometry data of a sample of SpheroTech RCP-30-5A beads (SpheroTech, 2001).

4 Estimation of Cell Division Rate

The expected number of cell divisions is estimated from time series data using a model in which cells begin evenly distributed across a cell division cycle of length λ hours. Transfected plasmids enter the nucleus during cell division and produce fluorescent protein at rate p. We also add an initial delay d, based on the observation that there is an initial period before any cell begins to express distinguishable amounts of fluorescent protein, which we conjecture is primarily due to the routing of DNA during lipofection (Lechardeur et al., 2005). Finally, only f fraction of the cells are successfully transfected and will ever express fluorescence.

Under this model, the fraction of distinguishable fluorescently expressing cells at time t is:

$$a(t) = \begin{cases} 0 & \text{if } t \le d \\ f \cdot \frac{t-d}{\lambda} & \text{if } d < t \le d + \lambda \\ f & \text{if } d + \lambda < t \end{cases}$$

We model discrete evolution in steps of Δt , letting the time $\tau = 0$ be the point when a cell begins to distinguishably fluoresce. We designate the expression level for a cell with respect to this relative starting time as $e_0(\tau)$. By definition, there is no expression initially, so $e_0(\tau \leq 0) = 0$. From this initial state, we then compute forward with the discrete production and decay/dilution model.

To simplify our equations, we name the discrete dilution rate for a given half-life as δ , which can be found by solving:

$$\frac{1}{2} = (1-\delta)^{\lfloor \frac{\lambda}{\Delta t} \rfloor}$$
$$\delta = 1 - \frac{1}{2}^{\lceil \frac{\Delta t}{\lambda} \rceil}$$

to produce the equation:

Given these definitions, the expression level for any given cell at discrete time τ may be computed recursively as production plus decay of the expression level at the previous time step:

$$e_0(\tau) = p \cdot \Delta t + e_0(\tau - \Delta t) \cdot (1 - \delta)$$

The expected expression model for the population may then be found by summing over the distribution of starting times:

$$e(t) = \frac{1}{\lambda} \sum_{x=0}^{\lambda} e_0(t - x - d)$$

We fit this model to minimize squared error with respect to the collection of data points from the constitutive expression dynamics data (main text Figure 2b and 2c) with Δt equal to 1 hour. The result is a

division time λ of 19.8 hours and an initial delay d of 15.3 hours, for an expected 2.4 divisions per expressing cell over the 72-hour experimental period $(\frac{72-d}{\lambda} - \frac{1}{2})$. To find the appropriate length for simulation, we find the expected beginning of fluorescent expression as $T_0 = d + \frac{\lambda}{2} = 25.2$ hours (the initial delay plus half a division cycle, giving the time when half of the cells that will ultimately measurably express fluorescent protein have begun their expression).

The length of time for a simulation to run is determined by the initial delay and the division rate. With the parameters calculated, the expected total protein expression time (the average amount of time a cell expresses a heterologous protein over the course of the experiment accounting for transfection delay) is $T = 72 - T_0 = 46.7$ hours over the course of a 72 hour transfection experiment.

This estimate of cell division rate (and thus the λ dilution parameter) is not tightly bound by the time series data. However, perturbations in the value of λ have only a minor impact on the accuracy of our predictions (Supplementary Table 1).

Cascade	1 division	2 divisions	2.4 divisions	3 divisions	4 divisions	5 divisions
TAL14-TAL21	1.23x	1.21x	1.20x	1.19x	1.18x	1.17x
TAL14-LmrA	1.81x	1.80x	1.81x	1.80x	1.78x	1.74x
TAL21-TAL14	1.22x	1.21x	1.30x	1.32x	1.40x	1.45x
TAL21-LmrA	1.55x	1.56x	1.56x	1.54x	1.45x	1.41x
LmrA-TAL14	1.89x	1.82x	1.75x	1.72x	1.71x	1.66x
LmrA-TAL21	1.71x	1.69x	1.74x	1.80x	1.73x	1.66x

Supplementary Table 1: Prediction error vs. expected number of cell divisions

To validate our calculated division rate, we also measured cell division rates in a separate experiment. Cells were seeded in 24 well dishes and transfected with pExp_CAG:mKate (see Section 8) at time t = 0. Cells were then trypsinized as described previously over the next several days and counted using a hemocytometer. Measurements for each well were taken in quadruplicate due to the high variance in measurement using a hemocytometer. Each time point was taken in duplicate. We found the division rate to vary from 20-30 hours (Supplementary Figure 2), which is similar to the division rate that we found through analytical fit of the constitutively expressed fluorescence protein.

5 Prediction Model Details

For each repressor/promoter pair we define a production rate function P(I, t, C) and a loss rate δ , where I is the input expression level, t is the current time, and C is constitutive fluorescence (an observable variable used to approximate the relative number of circuit copies). The production rate function P(I, t, C) and loss rate are computed from the dosage-response assay for the repressor/promoter pair, the estimated cell division rate λ , and the duration T of expected active expression per cell. Our models take into account three elements of incremental change in protein concentration:

- The concentration of any protein is subject to proportional loss from decay and from dilution due to cell growth. The proteins we use here are believed to be highly stable on the time scale of our experiments, so we incorporate only dilution rate, with a half-life of λ .¹
- The plasmids we use do not replicate in HEK293 cells, so at each cell division the number of plasmids (and hence production rate) is expected to halve.

¹Note that if a protein is not stable, this would be reflected in an effective decrease in λ for that species. As noted above in Section 4, a change in λ would have little effect on prediction quality.





Supplementary Figure 2: Measured cell growth using hemocytometer

• The relationship between concentration of regulatory species and rate of production per plasmid is assumed to be time-invariant.

Note that because the expression dynamics data shows no significant lag between constitutive and activator-driven expression, we do not include any independent effect of changing rtTa or Gal4 activation in our model.

The form of these functions will vary slightly depending on whether a continuous or discrete time model is chosen. Because the functions we are dealing with are relatively smooth, there should be no significant difference between the results produced by either approach. We have chosen to use a discrete model, with constant size steps of Δt such that our rate functions will produce difference equations rather than differential equations. This simplifies the interpolation and extrapolation that will be used for the production model.

Given these assumptions, for the loss rate we use the discrete incremental rate δ computed in Supplementary Section 4. For the production rate, time independence means there should be a linear relation between observed expression in the dosage-response assay and P(I, t, C) for any given time point. We thus encode P(I, t, C) as a product of three factors:

$$P(I, t, C) = \alpha \cdot \phi(t) \cdot X(I, C)$$

where X(I, C) is an expression level interpolated or extrapolated from the dosage-response assay, $\phi(t)$ is the relative circuit copy number at time t, and α is a proportionality constant that matches the production model with the observed values of the dosage-response assay at 72 hours for a discrete production model over duration T with half-life λ and time-step Δt . Figure 3 shows an example of such a production function at different times.

The formula for $\phi(t)$ is a relative count which begins at 1 and halves at each expected division:

$$\phi(t) = \frac{1}{2}^{\lfloor t/\lambda \rfloor}$$

Note that although $\phi(t)$ and C both relate to circuit copy numbers, they measure different quantities: C the relative circuit copy, as indicated by its proportional relationship to constitutive fluorescence at a snapshot in time, while $\phi(t)$ models how any value of C changes over time.

The value of X(I, C) varies depending on both the input I and the relative circuit copy number, as indicated by CFP intensity C. The distribution of C should be uniform from sample to sample, while values



(a) TAL21 Production at t=1 hour (b) TAL21 Production at t=20 hours (c) TAL21 Production at t=46 hours

Supplementary Figure 3: Examples of a production function, in this case for TAL21, at different times. Colored lines indicate production function for different levels of C, colored as in main text Figure 2. Production rates decrease over time as the set of plasmids becomes diluted by cell division.

for I are dependent on both C and inducer dosage. We thus chose to simplify the problem of approximating X(I,C) by segmenting into quantized "strips" by level of C. We thus segment dose-response data into logarithmic "bins" by CFP level at 10 bins per decade, as described in Supplementary Section 11. The value of X(I,C) is then interpolated or extrapolated for each quantized level of C from the mean value of each bin and induction level from a dosage-response assay. We begin by finding the bin that is closest to the CFP value currently being predicted for, and selecting the set of IFP and OFP bin means for that CFP bin at each induction level. These values may be viewed as forming a curve that relates IFP and OFP values. When the value of I is between the minimum and maximum observed values, we determine X(I, C) by linear interpolation between the closest adjacent IFP/OFP value pairs, using the ratio of distances from IFP values to determine the contribution of each OFP value to the estimate provided for X(I, C). When I is below the minimum or above the maximum observed IFP value, we instead determine X(I,C) by extrapolation. For this, we use a second-order polynomial fit to the nine lowest inductions for low IFP extrapolation and the nine highest inductions for high IFP extrapolation. A second order polynomial approximation was chosen as a simple approximation of a curve with few degrees of freedom, to allow curvature while minimizing the chance of overfitting. The number of inductions used for extrapolation was arrived at by incrementally extending the number of data points used until the set of curves for each set of extrapolations was relatively smooth.

In the case of the second repressor in the cascade, we multiply the IFP values used for calculation of X(I, C) by three, because the plasmids for this repressor are transfected at one-third the DNA concentration of the TRE-containing plasmids used to gather IFP data for the dosage response assay and the first repressor. This means that production of the second repressor should be three times lower relative to activation of the first repressor during the initial transient, when much of the repressor is produced for this system, and thus, using a linear first approximation, three times the input is expected to be required to have the same effect.

Finally, the proportionality constant α is computed by normalizing against unit production over the duration T. The following formula totals the production, assuming a production of 1 per unit time, adjusting for the relative copy number, and incorporating the loss function to model the dilution that will occur until the end time.

$$\alpha^{-1} = \sum_{i=1}^{\lfloor \frac{T}{\Delta t} \rfloor} 1 \cdot \phi(i\Delta t) \cdot (1-\delta)^{(\lfloor \frac{T}{\Delta t} \rfloor - i)}$$

Simulation of circuit dynamics proceeds in discrete time steps. Specifically:

- 1. All expression levels E_i are initialized to zero, i.e., $E_i(0) = 0$.
- 2. For each induced input $E_i(t)$ at CFP level C, with a final value of I_i , we compute a constant rate of expression: $E_i(t + \Delta t) = \alpha \cdot \phi(t) \cdot I_i \lfloor \frac{T}{\Delta t} \rfloor + (1 \delta) \cdot E_i(t)$
- 3. For each device whose output $E_i(t)$ is regulated by input $E_j(t)$, we compute $E_i(t) = P(E_j(t), t, C) + (1 \delta) \cdot E_i(t \Delta t)$

Note that this current formulation supports only combinational or feed-forward circuits, because extrapolation of $P(E_j(t), t, C)$ is not well defined if $E_j(t)$ is zero. Thus, we ensure that $E_j(t)$ is always defined by computing from the same time, rather than the prior time. This incremental shift means the simulation uses temporal ordering rather than absolute time, but should generally have minimal impact on the simulation of combinational or feed-forward circuits with small time delays. A simulation that uses proper time, where $P(E_j(t), t, C)$ is used to compute $E_t(t + \Delta t)$, would allow prediction of feedback circuits and circuits with long delays, but requires the ability to determine a value for P(0, t, C).

5.1 Modeling Alternatives and Extensions

Although we have chosen to implement prediction with EQuIP rate models using difference equations, two key alternative modeling frameworks deserve discussion, particularly with respect to extending EQuIP to prediction of genetic circuits with complex dynamics.

First, for any discrete difference equation model there is an analogous continuous differential equation model. For smooth and slowly changing functions, as in the devices considered in this manuscript, differential and difference equations do not produce significantly different results. In systems exhibiting a high degree of nonlinearity or where behavior is strongly determined by interactions within brief critical period, however, differential equations are generally more accurate. In such circumstances, it would be preferable to make predictions using EQuIP rate models implemented with differential equations.

Here, it is important to note that EQuIP's representational commitment is to the set of production and loss rate models, not to their instantiation as difference equations. Thus, it is just as possible to make EQuIP predictions using differential equations, following standard mathematical transforms. Examples of systems where this would be desirable include "weak" oscillators with parameters that are near a phase transition from the desired behavior to a qualitative loss of function, or other feedback circuits with a high degree of sensitivity in their behavior.

At a finer level of detail, chemical reaction networks are not continuous, but inherently stochastic in their execution, since their action transpires through individual molecular interactions. When each reaction involves large numbers of molecules with slow loss rates, as is expected to be the case for the systems considered in this manuscript, the law of large numbers ensures that stochastic and bulk models produce closely similar results. In many biological systems, of course, these assumptions do not hold, and stochastic effects can produce highly divergent populations. Examples include biological noise generators and circuits designed to detect or produce very low concentrations of target chemicals.

EQuIP's rate models cannot be directly mapped to stochastic chemical reaction networks, as these typically require a number of rate constants that are difficult to observe experimentally, and certainly cannot be determined from the time-series and dose-response experiments we have specified for EQuIP. Stochastic effects can also, however, be modeled using probability distributions. Thus, it will likely be possible to extend EQuIP to systems with significant stochastic effects by considering the observed distributions in addition to observed mean values.

6 Details of Dose-Response Assays and Cascade Predictions

Supplementary Figure 4 shows a summary of prediction quality with a scatter graph of predicted vs. experimental MEFL for all cascades, inductions, and CFP levels. The graph includes only low-extrapolation data

	Error for Population Mean	Error for Population Std. Dev.
TAL14-TAL21	1.41x	$1.09 \mathrm{x}$
TAL14-LmrA	1.37x	$1.62 \mathrm{x}$
TAL21-TAL14	1.53x	$1.24\mathrm{x}$
TAL21-LmrA	1.40x	1.33x
LmrA-TAL14	1.34x	$1.60 \mathrm{x}$
LmrA-TAL21	1.55x	$1.40 \mathrm{x}$
Total	1.43x	1.38x
Maximum	2.13x	$1.99 \mathrm{x}$

Supplementary Table 2: Prediction error for population statistics of cascades.

points—that is, those where no more than 10% of the production steps in the prediction simulation involved extrapolations (high or low, either repressor).

Constitutive fluorescence histograms for dose-response assays (main text Figure 2) and cascade experiments (Supplementary Figure 5 through 10) show the mean number of samples per bin across all samples. Extrapolation in dose-response assays (main text Figure 2) is shown out to the minimum and maximum levels used in any prediction for any data point in the entire range.

Prediction curves with full details for all six cascades are shown in Supplementary Figure 5 through 10. In these graphs, any data point where more than 10% of the production steps in the prediction simulation involved extrapolations (high or low, either repressor) is shown in grey and with a diamond for the prediction rather than a circle. Importantly, note that although the extrapolation method is clearly suboptimal, the impact on predictions is relatively minor.

Supplementary Figure 11 and 12 show the details of prediction error and number of extrapolations required for the six cascades. In general, prediction is worse for high copy number and high levels of induction. Note that the number of extrapolations is typically highest for high levels of induction: this is driven primarily by extrapolation on the upper end, since the output range of each input/output curve goes to higher MEFL than the input range. Typically, extrapolations at the low end come from the beginning of the simulation, when only a small fraction of the ultimate expression has been produced, and have a smaller effect on predicted output values.

Predicting expression for a range of different CFP bins also implicitly predicts the distribution of expression across the population of active cells. Supplementary Figure 13 compares the predicted and experimental geometric mean and standard deviation of all cells in the predicted band of CFP levels. Given the accuracy of prediction for individual CFP levels, and the fact that error tends to be highest where the number of cells per CFP bin is lower, these prediction are quite accurate—in fact, more accurate than the predictions for individual CFP levels. Table 2 gives the mean prediction error for population statistics for each cascade, as well as overall accuracy. Even the maximum error on any population statistic for any cascade and condition is only 2.1x.

Supplementary Figure 14 shows the predicted and observed differences in expression level for the 0 nM Dox and 2000 nM Dox induction conditions, for each bin. Supplementary Figure 15 shows a summary of prediction quality with a scatter graph of predicted vs. experimental MEFL for all feed-forward circuits, inductions, and CFP levels.

Findally, prediction curves with full details for all three feed-forward circuits are shown in Supplementary Figure 16 through 18. Graphs use the same display conventions as for the csacade prediction figures.



(d) TAL21-LmrA Prediction Scatter (e) LmrA-TAL14 Prediction Scatter (f) LmrA-TAL21 Prediction Scatter

Supplementary Figure 4: Summary scatter graph of predicted mean vs. experimental mean for all cascades, inductions, and CFP levels. For each observed input condition, a corresponding simulated output was generated, and is compared to the observed experimental output corresponding to that input. Note that the same predicted and/or observed outputs may be generated by multiple inputs, because of the variation in circuit response due to relatively copy number, as indicated by CFP level. Experimental standard deviation of means between replicates is indicated by vertical line on each point.



Supplementary Figure 5: TAL14-TAL21 cascade. Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 6: TAL14-LmrA cascade Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 7: TAL21-TAL14 cascade Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 8: TAL21-LmrA cascade Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 9: LmrA-TAL14 cascade Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 10: LmrA-TAL21 cascade Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 4.



Supplementary Figure 11: Error and number of extrapolations vs. induction and constitutive fluorescence for TAL14-TAL21, TAL14-LmrA, and TAL21-TAL14 cascades. "Zero error" regions are considered invalid predictions due to containing excessive extrapolation. Color indicates fold-error and \log_{10} number of extrapolations, respectively. Predictions are same as in Supplementary Figure 4.



Supplementary Figure 12: Error and number of extrapolations vs. induction and constitutive fluorescence for TAL14-TAL21, TAL14-LmrA, and TAL21-TAL14 cascades. "Zero error" regions are considered invalid predictions due to containing excessive extrapolation. Color indicates fold-error and \log_{10} number of extrapolations, respectively. Predictions are same as in Supplementary Figure 4.



Supplementary Figure 13: Population distribution of fluorescence, comparing predicted and experimental geometric mean and standard deviation over all cells in active population. Population predictions were generated by weighted statistics of CFP-subpopulation predictions, using weights from experimental observation of CFP distribution, and compared to population statistics of experimental observations falling into the selected CFP range.



Supplementary Figure 14: Ratio between uninduced and full induced expression for each cascade and relative copy number, as indicated by CFP level. Predicted ratios (circles) are lower than experimental ratios (pluses), but show a close and similar pattern.



(a) TAL21-TAL14 Prediction Scatter (b) TAL21-LmrA Prediction Scatter (c) LmrA-TAL14 Prediction Scatter

Supplementary Figure 15: Summary scatter graph of predicted mean vs. experimental mean for all cascades, inductions, and CFP levels. For each observed input condition, a corresponding simulated output was generated, and is compared to the observed experimental output corresponding to that input. Note that the same predicted and/or observed outputs may be generated by multiple inputs, because of the variation in circuit response due to relatively copy number, as indicated by CFP level. Experimental standard deviation of means between replicates is indicated by vertical line on each point.



Supplementary Figure 16: TAL21-TAL14 feed-forward circuit. Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 15.



Supplementary Figure 17: TAL21-LmrA feed-forward circuit. Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 15.



Supplementary Figure 18: LmrA-TAL14 feed-forward circuits. Experimental observations shown as pluses, predictions shown as circles for $\leq 10\%$ extrapolation and diamonds for > 10%. Relative copy number, as indicated by CFP level, is shown by color. Predictions are generated as described in Supplementary Figure 15.

7 Comparison with Hill Function Models

To compare the full EQuIP method with partial implementations, we fit a Hill function model against dosage-response curves produced for each device by partial implementation of EQuIP data processing. Hill functions are chosen as they are a widely used and well-grounded model for transcriptional regulation. In particular, we consider the following four models:

- 1. **Population a.u.:** This model uses arithmetic means of the raw arbitrary unit data from the flow cytometer, equivalent to population data captured wth a fluorimeter.
- 2. **Population MEFL:** This model uses the arithmetic means of the flow cytometer data after it has been converted from arbitrary units to calibrated MEFL, equivalent to fluorimeter data calibrated to absolute measurements.
- 3. Active Population MEFL: This model exploits the per-cell measurements provided by flow cytometry: constitutive fluorescence is used to identify the population of successfully transfected cells, and geometric means are used rather than population means (better fitting the log-normal variation of expression).
- 4. **Binned:** This model uses exact same binned-distribution data as EQuIP, but continues to fit it a Hill function model rather than using the piecewise model in EQuIP.

We implement the population Hill function models using the following differential equation for regulated production:

$$\frac{\delta E_i}{\delta t} = \alpha_i \cdot \phi(t) \cdot \frac{1 + k_i^{-1} \cdot \left(\frac{E_i}{d_i}\right)^{h_i}}{1 + \left(\frac{E_i}{d_i}\right)^{h_i}} - \lambda E_i$$

The binned ODE model adds an additional g_i term that controls how closely the outputs from adjacent bins are packed:

$$\frac{\delta E_i}{\delta t} = \alpha_i \cdot \phi(t) \cdot P^{g_i} \cdot \frac{1 + k_i^{-1} \cdot (\frac{E_i}{d_i})^{h_i}}{1 + (\frac{E_i}{d_i})^{h_i}} - \lambda E_i$$



Supplementary Figure 19: Mean squared error for best fit of Hill function models, showing good fits in all cases.

Input production is modeled with a differential equation equivalent to the difference equation given in Section 5:

$$\frac{\delta E_i}{\delta t} = \alpha_i \cdot \phi(t) - \lambda E_i$$

and cascade models compose device rates functions equivalently as well.

The parameters α_i , g_i , k_i , d_i , and h_i (plus g_i for binned ODEs) are determined for each device and each model by a non-linear parameter fit using the MATLAB implementation of the Nelder-Mead simplex algorithm (Lagarias et al., 1998). To ensure a good fit, we ran fitting five times per model with different initial values and used only the model with the best overall fit. Fit error and values for the individual fits varied significantly, likely due to the fact that Hill function models are under-constrained with respect to the observed data. In all cases, however, at least one fit found a low mean-squared-error, as shown in Figure 19, providing qualitatively good fits to the curves, as shown in Figure 20. The actual parameters of the Hill function for these best fits vary highly from model to model, as shown in Figure 21, again likely due to under-constraint. This emphasizes the point that, just because a model fits well does not mean that the model is necessarily correct.

In evaluating the quality of prediction methods, we are concerned not only with accuracy but also precision. We thus measure both the mean prediction error across the six cascades for each method, but also the 95% confidence envelope. Mean error is computed as geometric mean of the mean-fold errors for all induction levels for each cascades. The 95% confidence intervals are estimated as the geometric mean error and two geometric standard deviations of error (i.e., $\mu_q \cdot \sigma_a^2$).

Figure 22 gives the mean error in predicting the population means of each of the six cascades for each of the evaluated methods. Note that EQuIP is the only method that provides consistently good predictions for all cascades. Significantly, for the baseline model of population data in arbitrary units (which is equivalent to typical methods used in prior work), the 95% confidence envelope is nearly 30-fold error, meaning that although some individual systems might happen to yield a reasonable match between predicted and observed results, the range between the highest and lowest likely observation for any given prediction approaches 1000-fold, meaning the baseline model provides virtually no reliable information.

8 Plasmid Sequences

All plasmids used in this paper have been deposited with Addgene. Supplementary Table 3 shows the identifiers of all plasmids.



Supplementary Figure 20: Comparison between data and model for Hill function fits of characterization data.

Model	h	d	k	α	g
TAL14					
Pop. a.u.	$5.01e{+1}$	3.97e-1	$2.74e{+4}$	$1.95e{+}2$	$2.51e{+1}$
Pop. MEFL	6.46e-1	3.30e + 3	$1.10e{+4}$	$1.07e{+}4$	$1.25e{+1}$
Active MEFL	$5.15e{+1}$	3.57e + 6	$2.51e{+1}$	$2.56e{+}2$	$1.03e{+1}$
Binned Hill	$2.17e{+}0$	$3.10e{+}0$	1.54e + 6	7.36e + 5	1.40e + 0
TAL21					
Pop. a.u.	$1.01e{+}0$	5.49e + 2	$1.27e{+}4$	2.45e-1	2.46e + 0
Pop. MEFL	8.89e-1	7.52e + 6	$3.88e{+7}$	5.98e + 2	2.23e-1
Active MEFL	5.34e-1	$6.85e{+}4$	3.37e + 3	1.31e + 3	$1.47e{+1}$
Binned Hill	6.68e-1	$5.49e{+4}$	$1.08e{+7}$	$2.72e{+1}$	$1.28e{+}0$
LmrA					
Pop. a.u.	$1.01e{+}0$	2.30e+2	1.41e + 5	3.67e-1	1.66e + 0
Pop. MEFL	5.87e-1	1.05e+5	4.70e + 3	5.48e + 2	$1.98e{+1}$
Active MEFL	4.12e-1	2.42e+4	$1.27e{+}4$	2.47e+2	$1.81e{+1}$
Binned Hill	8.53e-1	9.88e-1	5.33e-1	$1.52e{+}0$	$1.33e{+}0$

Supplementary Figure 21: Hill function best-fit values.



Supplementary Figure 22: Comparison of mean prediction error for each cascade for Hill function models and EQuIP.

Nickname	Purpose	Addgene plasmid #
pExp_EMPTY	Blank control plasmid	51812
pExp_CAG:mKate	constitutive fluorescence	51789
pExp_CAG:EBFP2	constitutive fluorescence	51790
pExp_CAG:EYFP	constitutive fluorescence	51791
pExp_CAG:rtTa3-T2A-VP16Gal4;	Dox induction and Gal4 activation	51792
pExp_TRE:EBFP2	Input fluorescence	51793
pExp_TRE:TAL14	Input repressor	51794
pExp_TRE:TAL21	Input repressor	51795
pExp_TRE:LmrA	Input repressor	51796
pExp_UAS-TAL14:EYFP	Output fluorescence	51797
pExp_UAS-TAL21:EYFP	Output fluorescence	51798
pExp_UAS-LmrA:EYFP	Output fluorescence	51799
pExp_UAS-TAL14:TAL21	Second stage of cascade	51800
pExp_UAS-TAL14:LmrA	Second stage of cascade	51802
pExp_UAS-TAL21:TAL14	Second stage of cascade	51801
pExp_UAS-TAL21:LmrA	Second stage of cascade	51803
pExp_UAS-LmrA:TAL14	Second stage of cascade	51805
pExp_UAS-LmrA:TAL21	Second stage of cascade	51804

Supplementary Table 3: Addgene plasmid identifiers for all plasmids used

9 Additional Experimental Methods

9.1 Feed-Forward Circuit Transfections

Feed-forward circuits were assayed in a separate experiment with slightly different methods, reproduced here in full. Cells were seeded in 24 well plates at 300,000/well concentration a day before transfection so that they reach 80-90% confluency for the transfections. On the day of transfection, the media was changed and 500 ml of fresh media containing doxycycline (logarithmic concentration from 0 to 2000 ng/ml) was added to each well. A 96 well plate was used to prepare the mixture of transfection reagents. DNA comprising the transcription units of the circuits (seven plasmid DNAs: total 650 ng) were diluted to 50 ng/ul (except for those carrying TRE promoter that were diluted to 150 ng/ul). 63 ul of DMEM media (without serum and supplements) was added to each well of a 96 well plate. Subsequently, 1 ul of each diluted plasmid DNA was added to DMEM (7 ul DNA total, corresponding to 7 transfected plasmid DNAs and 650 ng total) and the DNA/media mix pipetted immediately once, using a multi-channel pipettor. 2 ul metafectone pro reagent was added to the DNA/DMEM mixture in each well of a 96 well plate and pipetted once immediately. The reagent mixture was incubated for 25 minutes to allow the formation of reagent/DNA complex assembly. The mixtures were then added drop-wise to cells in 24 well plates. Media was changed one day post-transfection and fresh media containing doxycycline was added to the cells.

9.2 Vector Details

The transfections carried out for characterization of the repressors were six plasmid co-transfections with one plasmid being a blank empty vector. The vectors used were: pExp_CAG:rtTa3-T2A-VP16Gal4; pExp_TRE:EBFP2; pExp_TRE:rep; pExp_UAS-rep:EYFP; pExp_CAG:mKate; pExp_EMPTY. The two TRE-containing vectors had 3x the amount of DNA as the other vectors during the transfection. In other words, for one well of a 24 well dish the DNA for the constructs above were used in the amounts shown in Supplementary Table 4.

Similarly for the cascades, the transfection vectors used were: pExp_CAG:rtTa3-T2A-VP16Gal4; pExp_TRE:EBFP2; pExp_TRE:rep1; pExp_UAS-rep1:rep2; pExp_UAS-rep2:EYFP; pExp_CAG:mKate. Here there is not a need for an empty vector, as there are six plasmids in the circuit. The DNA ratios remain the same

Plasmid Name	DNA amount (ng)
pExp_CAG:rtTa3-T2A-VP16Gal4	50
$pExp_TRE:EBFP2$	150
pExp_TRE:rep	150
pExp_UAS-rep:EYFP	50
pExp_CAG:mKate	50
pExp_EMPTY	50
Total DNA per well	500

Supplementary Table 4: Plasmid DNA amounts for characterization transfections

Plasmid Name	DNA amount (ng)
pExp_CAG:rtTa3-T2A-VP16Gal4	50
pExp_TRE:EBFP2	150
pExp_TRE:rep1	150
pExp_UAS-rep1:rep2	50
pExp_UAS-rep2:EYFP	50
pExp_CAG:mKate	50
Total DNA per well	500

Supplementary Table 5: Plasmid DNA amounts for cascade transfections

Plasmid Name	DNA amount (ng)
pExp_CAG:rtTa3-T2A-VP16Gal4	50
pExp_TRE:EBFP2	150
pExp_TRE:rep1	150
pExp_TRE:rep2	150
pExp_UAS-rep1:rep2	50
pExp_UAS-rep2:EYFP	50
pExp_CAG:mKate	50
Total DNA per well	650

Supplementary Table 6: Plasmid DNA amounts for feed-forward circuit transfections

as above, with the TRE-containing vectors having 3x the DNA of the other constructs (amounts shown in Supplementary Table 5). This is because the concentration of the first repressor is limited by the degree to which the TRE promoters are activated; using a higher concentration raises the range of expression and allows stronger repression to be observed and better signal matching. Feed-forward circuit transfections were the same, with addition of another TRE vector (amounts shown in Supplementary Table 6).

A set of control transfections was carried out with every experiment, where each fluorescent reporter was transfected individually as well as all together. The set consisted of five transfections: one individually for the vectors pExp_CAG:mKate, pExp_CAG:EBFP2, pExp_CAG:EYFP, a co-transfection of all three (mKate, EBFP2, and EYFP) and a transfection of pExp_EMPTY used to determine auto-fluorescence. These transfections included 100ng of active construct plus enough empty vector to keep the total DNA amount the same between these controls and the circuit transfections. e.g., 100ng of pExp_CAG:mKate and 400ng of pExp_EMPTY or 100ng pExp_CAG:mKate, 100ng pExp_CAG:EBFP2, 100ng pExp_CAG:EYFP, 200ng pExp_EMPTY.

Dosage-response and cascade samples were induced with 12 different concentrations of doxycycline: 0 nM, 1 nM, 2 nM, 5 nM, 10 nM, 20 nM, 50 nM, 100 nM, 200 nM, 500 nM, 1000 nM, and 2000 nM. For all dosage-response and cascade experiments, data was collected in triplicate.

Time series experiments were carried out similarly to dosage-response experiments. A complete char-



Supplementary Figure 23: Relative expression variance between constitutive fluorescent proteins decreases as intensity of any given protein increases.

acterization circuit was transfected with plasmid ratios as described above. Multiple transfections were performed over the course of three days, allowing data acquisition for multiple time points in the experiment to take place at the same physical time. There were a total of four distinct batches of data collection:

- 58 hours
- 16, 18, 20, 22, and 60 hours
- 12, 24, 42, and 63 hours
- 26, 28, 30, 48, 69, and 72 hours

Expression dynamics data for constitutive fluorescent protein expression was averaged over sixteen replicates: two replicates each of three single constitutive controls, three colors from the triple constitutive control, plus the constitutive mKate from the rtTa and Gal4 series. For the rtTa activation series, two replicates were collected from a TAL21 characterization circuit induced with 2000 nM Dox; for the Gal4 activation series, two replicates were collected from a TAL21 characterization circuit induced with 0 nM Dox. One exception was the absence of constitutive controls for 12 hours. Constitutive data for this time point comes only from the circuit transfections for the rtTa and Gal4 series.

10 Plasmid Copy Variance in Co-transfection

An important component that determines the accuracy of our approach is the degree to which there is variation in the relative copy number of cotransfected plasmids, where relative copy number is indicated by constitutive fluorescence as described in Supplementary Section 11 We chose multi-plasmid co-transfection over transfection of a single large plasmid because it is easier to model relative copy number variance than to understand DNA sequence adjacency interactions such as promoter interference. To the extent that the process of lipofection is understood, this variation appears to be typically small for the number of plasmids and the size of those plasmids used in our system.

We examine this issue by analyzing the relative variation of expression between cotransfected fluorescent proteins in a three constitutive fluorescent protein control. Supplementary Figure 23 shows the geometric standard deviation of expression levels of a second protein when binned logarithmically against the first, e.g. "mKate from EBFP2" shows the geometric standard deviations of mKate expression of flow cytometry events falling into the same logarithmic bin of EBFP2 expression levels. As expected, the level of variation decreases rapidly as constitutive expression rises (indicating higher numbers of plasmids and therefore greater applicability of the law of large numbers). This strong monotonic decrease and its consistency across color pairings indicate that any variation in relative plasmid copy number is likely to be unbiased and to not significantly affect the bin geometric means.

11 Computation of Sample Statistics

Within a sample, only some cells are successfully transfected, as indicated by their expression of a constitutive fluorescent protein (CFP), in this case mKate, above baseline levels. If transfection quality is poor or the event count in a sample is unexpectedly low, it is likely indicative of a protocol failure (e.g. contamination) that will degrade the quality of the data.

To evaluate sample quality, we fit the observed distribution of constitutive fluorescent protein levels in flow cytometry events from the sample to a bimodal Gaussian distribution on a logarithmic scale. Note that a bimodal Gaussian fit does not assign events to one component or another, but attributes a probability that each event was generated by one of the two components of the distribution.

A sample is then included only if:

- the distribution fits well as a bimodal Gaussian,
- the majority of data in the sample is expected to come from cells in the active Gaussian component (i.e., the component with the higher mean), and
- there are at least 100,000 events in the sample.

Applying this decision criteria, for dosage-response curves we discarded: TAL14: one replicate of 50 nM; TAL21: one full replicate, plus one replicate of 0 nM, 2 nM, 1000 nM, and 2000 nM. For the cascades we discarded: TAL14-TAL21: one replicate of 0 nM, 50 nM, 100 nM; LmrA-TAL21: one replicate of 10 nM, 200 nM, 500 nM. For expression dynamics, we discarded: one replicate of 0 nM at 58 hours.

Figure 24 shows an example of a bimodal Gaussian fit against the distribution of constitutive fluorescent protein for each sample in the LmrA dosage-response experiment. Note that the samples are highly consistent in their distribution. Note also that the breadth of the distribution is much higher than the expression variance shown in Figure 23, indicating a high range of variation in the number of circuit copies per cell.

When computing statistics over fluorescence data, note that it is important to use geometric statistics rather than the more conventional arithmetic statistics. This is because the variation of values in these systems is typically Gaussian on a logarithmic scale, rather than on a linear scale. Arithmetic statistics will thus tend to produce a poor estimate of the true distribution.

In our data analysis, when it is necessary to compute the mean of the active population (e.g., for expression dynamics or cross-batch compensation), the geometric mean from the active component of the bimodal Gaussian is used. When computing statistics with respect to circuit copy number, the set of events is segmented into logarithmic bins by CFP value at 10 bins per decade. We then compute geometric mean and standard deviation for each bin for the IFP and OFP events whose CFP value falls into that bin. The only bins that are used are those likely to contain high quality data (i.e., at least 100 events and low percentage attribution of events to the inactive component of the bimodal Gaussian): we select a consistent global threshold and use data from the bins encompassing CFP values from $10^{5.8}$ to $10^{7.9}$ MEFL.

The expected number of circuit copies per cell is not linear with respect to the CFP value, because other sources of expression variation interact with the transfection process to create a bias in the sources for cells in each bin, as shown in (Beal et al., 2012). The relationship is monotonic and consistent, however, so we



Supplementary Figure 24: Constitutive fluorescence exhibits a consistent wide bimodal gaussian distribution. A representative example is shown in (a): distribution of constitutive fluorescence for LmrA in the dosageresponse assay, showing number of samples vs. logarithmic bin center (solid lines), and fit against a bimodal gaussian model (dashed lines). One line per sample; colors correspond to induction levels, ranging from to red (lowest) purple (highest). (b) The model can be used to estimate the fraction of cells with a given constitutive fluorescence belonging to the successfully transfected population. Colors are the same as in (a).

can use CFP bins as a proxy for circuit copy number when making predictions without knowing the actual expected number of copies.

12 Cross-Batch Compensation

Certain variations between sample batches that may often be tolerated in experiments are not be acceptable for our stringent requirements. Such variations may come from many different sources, such as the state of source cells, variation in reagents, and imperfections in the culturing environment. Although such variations are generally small, they may be significant at the level of precision that we require. Fortunately, such variations are often systematic across entire batches of experimental samples or replicates, so it is possible to perform a first-order linear compensation to a chosen reference experiment. For the purposes of EQuIP, it is critical to have extremely high precision for dosage-response curves. Hence, for each fluorescent protein, we select a Dox induction level at which, given our circuit architecture and under ideal circumstances, behavior should be identical across all experiments: for CFP, Dox = 2000 nM; for IFP, Dox = 2000 nM; for OFP, Dox = 0 nM (since our TRE typically shows minimal leakage, at Dox = 0 nM, OFP should be constitutively activated by Gal4). The equivalence might, of course, be rendered invalid by some unexpected interaction, but is effectively cross-validated through the other 14 induction levels.

To perform the compensation for dosage-response curves, for each replicate of each experiment we compute the geometric mean of all events with at least 50% probability of generation by cells in the active component of the sample at the chosen induction level. The relative expression levels for the experiment are then taken as the geometric mean across replicates. Selecting reference points based on tightness of data, the relative scaling for CFP, IFP, and OFP for the three dosage-response assays is: a \log_{10} shift of: TAL14: 0.05, -0.54, -0.03; TAL21: -0.05, -0.70, 0.00; LmrA: 0.00, 0.00, -0.38.

For the cascades we compensate for CFP only, since IFP is taken as a given for purposes of generating the prediction and since there is no constitutive expression of OFP to calibrate from. For the cascades, the scaling values for CFP are: \log_{10} shifts of: TAL14-TAL21: 0.18, TAL14-LmrA: 0.03, TAL21-TAL14: -0.17,

13 Internal Cross-Validation

Given the myriad ways in which biological experimental protocols can go wrong, it is important to have means of distinguishing between experimental error and meaningful deviation from expected behavior. For this purpose, the large and continuous distribution of circuit copy numbers within a single sample resulting from transfection is in fact useful. Each sample provides a collection of highly variable but strongly correlated data points, whose behavior should vary systematically with circuit copy number. Our experiments are thus afforded a number of different means of internal cross-validation (both within a single sample and across samples), by considering each sample as a function of statistics over the bins into which the data has been segmented. We use this cross-validation to improve the precision of characterization by rejecting samples whose behavior is not internally consistent.

From our controls, above and beyond the qualitative validation which is typically provided, we apply the following quantitative tests:

- Peaks in the fluorescent bead control should fit to the linear peak model provided by the supplier to within 10%.
- Data sets taken at different times with the same flow cytometry settings must match every peak in their respective fluorescent bead controls with one another to within 10%.
- Autofluorescence data from cells transfected with blank plasmids should fit well to a Gaussian model on a linear scale.
- In the single color controls used to build spectral overlap compensation models, cells more than two standard deviations above autofluorescence should fit to within 10% of a linear function following subtraction of autofluorescence.
- In the multicolor control used to compute color conversion, the strongly expressing cells should have a linear relationship between colors. This is tested by validating that a round-trip translation between each pair of colors deviates in value by less than 10%.

Failure of any of these tests is rare, and indicates a problem in the execution of the specified experimental protocols.

Another internal cross-validation is the distribution of CFP values, which is expected to fall on a bimodal Gaussian and filtered accordingly as described above in Section 11. Finally, the incremental variation of time or induction affords another opportunity for cross-validation, since any phenomenon appearing in one sample should be expected to appear in incrementally varied samples, shifted both incrementally across circuit copy bins and in quantitative value. We expect a smooth and relatively regular variation across bins (such as seen in all of our experimental data presented in this paper), and any significant deviation from this indicates possible problems with experimental protocol.

References

- Beal, J.; Weiss, R.; Yaman, F.; Davidsohn, N.; Adler, A. A Method for Fast, High-Precision Characterization of Synthetic Biology Devices; 2012; Technical Report: MIT-CSAIL-TR-2012-008 http://hdl.handle.net/1721.1/69973.
- SpheroTech, Measuring Molecules of Equivalent Fluorescein (MEFL), PE (MEPE) and RPE-CY5 (MEPCY) using Sphero Rainbow Calibration Particles; 2001; SpheroTechnical Notes: STN-9, Rev C 071398.
- Lechardeur, D.; Verkman, A. S.; Lukacs, G. L. Advanced drug delivery reviews 2005, 57, 755–767, PMID: 15757759.

Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; Wright, P. E. SIAM Journal of Optimization 1998, 9, 112-147.