Precise determination of input-output mapping for multimodal gene circuits using data from transient transfection

Author(s):
Stelzer, Christoph; Benenson, Yaakov

Publication date:
2020-11-30

Permanent link:
https://doi.org/10.3929/ethz-b-000466577

Originally published in:
PLoS Computational Biology 16(11), https://doi.org/10.1371/journal.pcbi.1008389

Funding acknowledgement:
281490 - Synthetic regulatory circuits for programmable control of cell physiology (EC)

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Precise determination of input-output mapping for multimodal gene circuits using data from transient transfection

Christoph Stelzer, Yaakov Benenson*

Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, Basel 4058, Switzerland

* kobi.benenson@bsse.ethz.ch

OPEN ACCESS

Citation: Stelzer C, Benenson Y (2020) Precise determination of input-output mapping for multimodal gene circuits using data from transient transfection. PLoS Comput Biol 16(11): e1008389. https://doi.org/10.1371/journal.pcbi.1008389

Editor: Lingchong You, Duke University, UNITED STATES

Received: April 23, 2020

Accepted: September 23, 2020

Published: November 30, 2020

Data Availability Statement: The code is available on GitHub, https://github.com/benensonlab/PFAFF. The data are available on http://hdl.handle.net/20.500.11850/446198.

Funding: The research was funded by European Research Council Starting Grant StG 281490 and Swiss National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests to declare.

Abstract

The mapping of molecular inputs to their molecular outputs (input/output, I/O mapping) is an important characteristic of gene circuits, both natural and synthetic. Experimental determination of such mappings for synthetic circuits is best performed using stably integrated genetic constructs. In mammalian cells, stable integration of complex circuits is a time-consuming process that hampers rapid characterization of multiple circuit variants. On the other hand, transient transfection is quick. However, it is an extremely noisy process and it is unclear whether the obtained data have any relevance to the input/output mapping of a circuit obtained in the case of a stable integration. Here we describe a data processing workflow, Peakfinder algorithm for flow cytometry data (PFAFF), that allows extracting precise input/output mapping from single-cell protein expression data gathered by flow cytometry after a transient transfection. The workflow builds on the numerically-proven observation that the multivariate modes of input and output expression of multi-channel flow cytometry datasets, pre-binned by the expression level of an independent transfection reporter gene, harbor cells with circuit gene copy numbers distributions that depend deterministically on the properties of a bin. We validate our method by simulating flow cytometry data for seven multi-node circuit architectures, including a complex bi-modal circuit, under stable integration and transient transfection scenarios. The workflow applied to the simulated transient transfection data results in similar conclusions to those reached with simulated stable integration data. This indicates that the input/output mapping derived from transient transfection data using our method is an excellent approximation of the ground truth. Thus, the method allows to determine input/output mapping of complex gene network using noisy transient transfection data.

Author summary

One of the key features of a gene circuit is its input/output behavior. A few earlier publications attempted to develop methods to extract this behavior using transient transfection of circuit components in mammalian cells. However, the hitherto developed methods are only suitable for circuit with monomodal output distribution. Moreover, the relationship

between the extracted I/O mapping and the "ground truth" that would have obtained with stably-integrated circuits, has not been addressed. Here we explore cell populations easily identifiable in flow cytometry data, namely, the peaks of fluorescent readout distribution in cells binned by the common expression value of the transfection reporter, or marker, gene. Using numerical simulations, we find that the distribution of circuit copy number in these cells deterministically depends on marker fluorescence in the noise-dependent manner. Moreover, we find that this is true also in the case of bi-modal output distribution. Using the peaks of input and output distributions, we are able to reconstruct the I/O mapping of the circuit and relate it to the I/O mapping of the stably-integrated circuit. The reconstruction is enabled by a new computational method we call *PFAFF*. The method is extensively validated with forward-simulated flow cytometry data from stable and transient transfections, with up to seven different circuits. The results show excellent correlation between the I/O behavior extracted by PFAFF from simulated transient transfection data, and the data simulated for stably integrated circuit.

Introduction

Many synthetic gene circuits fall into the category of information-processing systems that convert molecular inputs to molecular outputs according to a specific relationship [1], often called a "program". A typical design-build-test cycle of a synthetic gene circuit requires that an input/output (I/O) relationship be characterized in order to confirm circuit function. Direct characterization is possible when both the input(s) and the output(s) can be measured simultaneously in single cells. Using fluorescent reporters, it is possible to obtain the collection of single-cell data points of the type [*input*; *output*], including for natural regulatory pathways, either by direct observation using staining, or by creating synthetic analogs of natural circuits furnished with fluorescent reporters [2–8]. It has emerged that the output forms a distribution at a single cell level for each input [8–10], resulting in a two-dimensional probability distribution for the entire I/O relationship, rather than a curve, due to cell-to-cell variation in parameter values. Nevertheless, after averaging, these noisy data sets usually collapse to Hill functions or to multimodal, two-value functions [11].

Characterization of a circuit that is stably integrated in a cell genome or on replicating fixed-copy episomal vectors is usually straightforward, provided that the inputs and the outputs can be measured. Thus, till now most of characterized input/output behaviors were obtained in bacteria or yeast, where genome manipulation is relatively facile. However, obtaining such "ground truth" information in mammalian cells has lagged behind, because it is still very labor-intensive to establish stably integrated multi-gene circuits. Further, properly executed characterization requires multiple accompanying control circuits to serve as baseline, thus requiring that not one but multiple stable cell lines be developed. Even though technologies such as transposon [12,13] and viral delivery [14,15], targeted integration via Zinc finger nucleases (ZFNs) [16], transcription activator-like effector nucleases (TALENs) [17] or clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 [18,19] are available today, they are still time consuming even in simple cases and become more challenging with the increase in circuit size. Integration locus-specific effects further complicate the characterization.

Transient transfection of gene circuits is a widespread alternative to stably-integrated circuit characterization in mammalian cells [20–24]. Multiple plasmids, each carrying a single gene, can be co-delivered, leading to correlated gene copy numbers in individual cells. The

expression of gene products in dividing cell cultures typically reaches quasi-steady-state two to three days post transfection, and decreases on days four to six due to plasmid dilution [24,25]. The advantage of the transient transfection is that the genome integration-specific effects can be ignored; likewise, secondary effects that often result from having a few genes close to each other on the genome do not play a role because each gene is encoded on a separate plasmid. On the other hand, transient transfections are extremely noisy due to large copy number variation (1–150 transcriptionally-active gene copies per cell [6]), which makes direct interpretation of the resulting datasets impossible. Accordingly, the standard analysis applied to transient transfection data is at the cell population level, with average values of inputs and outputs reported for entire cell populations (see, Schreiber et al. [26] as a representative case). This works sufficiently well for logic gene circuits that are often characterized at the extremes of their input values. Progress towards deriving continuous input/output relationship using transient transfection data has been made in the past [6,27–29]. However, these methods were designed to extract monomodal input/output curves and are thus unsuitable for bi- or multimodal circuits. Moreover, there has been very little computational or experimental validation of these results, in particular, how they compare to stably-integrated systems, and to what the different input/output curves correspond.

We sought to develop a data analysis strategy that would determine input/output relationships from transient transfection data and be applicable to all steady-state networks, including those with bi-modal or bi-stable behavior. We also sought to understand what exactly constitutes a comparable "stable integration" scenario for the information extracted from raw transient transfection data. Accordingly, we first investigate the gene copy number distributions in cell populations that are easily identifiable in flow cytometry-like datasets. We address the question numerically and find a number of important reproducible trends that make it possible to draw reliable and interpretable conclusions from data obtained in transient transfections, and map them back to their stable-integration counterparts. In order to validate the method, we perform *in-silico* experiments by simulating flow cytometry data expected in a transient transfection using dynamic circuit models. At the same time, we use the exact same models and parameter values to simulate the input/output relationship for the case of stable genomic integration. With this approach, we are able to evaluate whether our workflow, when applied to transient transfection data, results in an input/output behavior that is similar to the input/output behavior one would expect for a stable integration.

As benchmarks, we focus on three-node gene network motifs that have been extensively studied earlier [30,31]. We find excellent correspondence between the results of our processing pipeline and the ground truth of the stable integration. Importantly, we are able to capture multi-value, bi-modal responses. Therefore, the method described here can be used to analyze transient transfection data and draw conclusions about the underlying input/output mapping in complex gene circuits, without the need to construct stable cell lines.

Results

Statistical framework for transient transfection

In what follows, we define a gene circuit as a set of N genes

$$\boldsymbol{g} = [g_1, g_2, \ldots, g_N] \tag{1}$$

and their corresponding gene products

$$\boldsymbol{G} = [G_1, G_2, \ldots, G_N] \tag{2}$$

in which a subset of components

$$\boldsymbol{I} = [I_1, I_2, \ldots, I_Q] \subset \boldsymbol{G} \tag{3}$$

is defined as *input* and a subset of components

$$\boldsymbol{O} = [O_1, O_2, \ldots, O_P] \subset \boldsymbol{G} \tag{4}$$

is defined as *output*.

Further, consider a cell that harbors a gene circuit, either in a stably-integrated or transiently-delivered fashion, such that a gene g_i is present in k_i copies in that cell, and the entire set of copy numbers is a vector

$$\boldsymbol{k} = [k_1, k_2, \ldots, k_N] \tag{5}$$

Hereafter, we consider only interactions between circuit components that have been intentionally engineered (*i.e.*, chromatin-related effects do not interfere with circuit function in the stable case), and assume that the biochemical parameters describing individual interactions do not change between stably integrated and transiently-delivered components. Even though individual cells in a population of stable clones may behave differently, *e.g.*, through stochastic effects [32], we expect the aggregate statistics of different clones containing identical circuit copy number to be similar to the aggregate statistics of cells transiently transfected with the same circuit copy number. Therefore, when considering stable clones, we imply an idealized "averaged" clone in which the integrated circuit is governed by the same parameters as the transiently transfected circuit. It then follows that if we apply the same input $\boldsymbol{I}$ to a population of cells that all harbor the circuit with the copy number vector $\boldsymbol{k}$, and allow the cells to arrive at a steady state in the stable case and to the quasi-steady state in the transient case (see S1 Text "*In-silico* time-courses"), then the outputs $\boldsymbol{O}$ will form the same statistical distribution, which can be mono- or multimodal [33], in both cases. Reporting the distribution of $\boldsymbol{O}$ for various inputs $\boldsymbol{I}$ would conclude the characterization of a stably-integrated circuit, because all cells harbor the exact same vector $\boldsymbol{k}$, which can be engineered or experimentally determined *post factum* after clonal isolation.

In the transient transfection experiment, while the values of $\boldsymbol{I}$ and $\boldsymbol{O}$ could be collected for individual cells, the underlying values of $\boldsymbol{k}$ are unknown because the process of transient delivery is extremely noisy. The only way to derive useful data from transient transfections is to deduce, at least for a subset of cells, their $\boldsymbol{k}$ values, and group together data from cells with similar values of $\boldsymbol{k}$. If this can be accomplished, the input and output values measured in these cells will be similar to the values one would have obtained with a circuit stably integrated at $\boldsymbol{k}$ copies. Below, we develop a statistical description of a transient co-transfection process, which leads us to identify cells residing in binned modes of input and output distributions as cells for which the copy number vector $\boldsymbol{k}$ can be estimated.

We start with the statistical description of a multi-plasmid co-transfection of N constitutively expressed and mutually independent genes $g_1, g_2, \ldots, g_N$, generating (fluorescent) protein products $O_1, O_2, \ldots, O_N$. Note that there is no input in this system, so every protein product can be called an "output". Available data [6] suggest that experimentally-observed distributions of a protein level expressed from a constitutive promoter are lognormal. The mean of the distribution is proportional to the gene copy number k_i with the promoter-dependent global proportionality coefficient β_i being independent of k_i; the standard deviation σ_i of the log-transformed protein level distribution may, in principle, depend on a copy number, but we assume it to be constant in the following equations. Let us define a random variable Y_i as the

log-transformed protein output of the gene g_i.

$$Y_i = ln\ O_i \tag{6}$$

Y_i is distributed normally and its mean/mode μ_i, and standard deviation, σ_i, relate as follows to the mean of the underlying pre-transformed distribution

$$E[O_i] = \beta_i k_i = exp\left(\mu_i + \frac{\sigma_i^2}{2}\right) \tag{7}$$

and therefore

$$\mu_i = ln(\beta_i k_i) - \frac{\sigma_i^2}{2} \tag{8}$$

The conditional probability density function (*pdf*) of Y_i given a gene copy number k_i and parameter β_i is then described by

$$p(Y_i|k_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(-\frac{\left(Y_i - ln(\beta_i k_i) + \frac{\sigma_i^2}{2}\right)^2}{2\sigma_i^2}\right) \tag{9}$$

For a vector of gene copy numbers $\boldsymbol{k} = [k_1,k_2,\ldots,k_N]$, a conditional multivariate *pdf* of the log-transformed protein expression values $\boldsymbol{Y} = [Y_1,Y_2,\ldots,Y_N]$, provided that each gene generates its own protein output independently of each other, is described by a multivariate normal distribution without covariances (for simplicity, we assume σ_i to be the same for all genes and use a symbol σ in what follows):

$$p_Y(\boldsymbol{Y}|\boldsymbol{k}) = p_Y([Y_1, Y_2, \ldots, Y_N]|\boldsymbol{k}) = \prod_{i=1}^{N} p(Y_i|k_i) = \frac{1}{\sigma^N\sqrt{2\pi^N}} exp\left(-\sum_{i=1}^{N} \frac{\left(Y_i - ln(\beta_i k_i) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}\right) \tag{10}$$

To describe the distribution of gene copy numbers in a transient transfection, we introduce an independent parameter m that we call "multiplicity of transfection". Indeed, there is no experimental data that concerns the probability distribution of genes in a co-transfection, and it likely depends on the exact transfection protocol. Therefore, we make a baseline assumption about the *pdf* of the gene copy number vector $\boldsymbol{k}$ as a multivariate normal distribution without covariance that depends on the multiplicity of transfection. The standard deviation of each gene copy number distribution scales linearly with multiplicity, with the scaling factor $\boldsymbol{\varepsilon}$.
To account for gene combinations that deviate from an equimolar ratio, a parameter a_i describes the relative abundance of a gene. In this case, one gene is assigned as the "reference" with $a_i = 1$.

$$p_g(\boldsymbol{k}|m) = p_g([k_1, k_2, \ldots, k_N]|m) = \frac{1}{(\varepsilon m)^N\sqrt{2\pi^N}\prod_{i=1}^{N} a_i} exp\left(-\sum_{i=1}^{N} \frac{(k_i - a_i m)^2}{2(\varepsilon a_i m)^2}\right) \tag{11}$$

Lastly, m itself can be distributed non-uniformly according to its *pdf* $p(m)$. Distributions such as Poisson [34], Gamma [35], lognormal [36] or even a combination of them [37], have been used to describe the process of DNA or viral vector delivery to cells. For transient

lipofection of DNA, lognormal distributions approximate experimental data well, and therefore

$$p(m) = \mathcal{LN}(\mu_{\mathrm{m}}, \sigma_m) = \frac{1}{m\sigma_m\sqrt{2\pi}} exp\left(-\frac{(ln(m) - \mu_{\mathrm{m}})^2}{2\sigma_m{}^2}\right) \tag{12}$$

One of the genes and its protein product is assigned the role of, respectively, a reference gene and a reference protein (sometimes called "transfection marker"); let us assume it is k_1, with gene product O_1 and its log-transformed counterpart Y_1. Thus, by definition $a_1 = 1$. To derive the conditional marginal *pdf* $p(Y_i|Y_1)$, which is the probability to find the value Y_i of the log-transformed protein O_i expression in a cell in which the log-transformed reference protein expression equals Y_1, we first drop irrelevant variables from Eq 10 to evaluate joint probability distribution of log-transformed protein levels $[Y_i,Y_1]$ given the underlying gene copy numbers $[k_i,k_1]$:

$$p([Y_i, Y_1]|[k_i, k_1]) = \frac{1}{\sigma^2\sqrt{2\pi^2}} exp\left(-\frac{\left(Y_i - ln(\beta_i k_i) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2} - \frac{\left(Y_1 - ln(\beta_1 k_1) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}\right) \tag{13}$$

In turn the gene copy numbers are conditionally dependent on the multiplicity parameter m:

$$p([k_i, k_1]|m) = \frac{1}{a_i(\varepsilon m)^2\sqrt{2\pi^2}} exp\left(-\frac{(k_i - a_i m)^2}{2(\varepsilon a_i m)^2} - \frac{(k_1 - m)^2}{2(\varepsilon m)^2}\right) \tag{14}$$

The global joint probability distribution $p([Y_i,Y_1])$ of Y_i and Y_1 over all values of $[k_i,k_1]$ is obtained by integrating over all values of $[k_i,k_1]$:

$$p([Y_i, Y_1]) = \iint p([Y_i, Y_1]|[k_i, k_1])p([k_i, k_1]m)p(m)d[k_i, k_1]dm \tag{15}$$

It is customary, as already done earlier [6,38], to bin cells that share the same Y_1, the log-transformed value of O_1, because this is the only readout independent of the other components, as it is a self-contained gene expressed from a constitutive promoter. We follow this approach here: cells binned according to their Y_1 value will exhibit certain log-transformed distributions of the other proteins $Y_2,\ldots,Y_N$. Knowing the joint *pdf* (Eq 15), one can derive the conditional probability of Y_i given Y_1 (that is, the *pdf* of Y_i among cells that express Y_1 log-transformed copies of the reference protein), as follows:

$$p(Y_i|Y_1) = \frac{p([Y_i, Y_1])}{p(Y_1)} = \frac{\iint p([Y_i, Y_1]|[k_i, k_1])p([k_i, k_1]|m)p(m)d[k_i, k_1]dm}{\iint p(Y_1|k_1)p(k_1|m)p(m)dk_1 dm} \tag{16}$$

The mode of this distribution, *i.e.*, the most probable value of Y_i given the reference Y_1, can be found by solving the equation

$$\frac{dp(Y_i|Y_1)}{dY_i} = 0 \tag{17}$$

Let us denote this most probable value as $Y_i^{MODE}(Y_1)$. The value of $Y_i^{MODE}(Y_1)$ can be determined experimentally as a mode of Y_i distribution after binning the cells according to their Y_1 value. The equation may have more than one solution, corresponding to multimodal probability density function from Eq 16.

This bring us to the most relevant question of this section: What is the distribution of the gene copy number k_i for the cells that reside in the mode(s) of Y_i, and what is the most probable value of k_i? To answer this question, we evaluate the conditional probability $p(k_i|Y_i)$

according to Bayes' theorem:

$$p(k_i|Y_i) = \frac{p(Y_i|k_i)p(k_i)}{p(Y_i)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\left(Y_i - ln(\beta_i k_i) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}\right) \int p(k_i|m)p(m)dm}{\iint p(Y_i|k_i)p(k_i|m)p(m)dk_i dm} \quad (18)$$

In order to find the most probable copy number $k_i^{MODE}(Y_i^{MODE}(Y_1))$ we solve the equation (or equations, when $Y_i^{MODE}(Y_1)$ takes more than one value)

$$\frac{dp(k_i|Y_i^{MODE}(Y_1))}{dk_i} = 0 \quad (19)$$

Knowing the most probable gene copy numbers in the cells residing in the modes of log-transformed protein distributions allows us to correlate the data to what might be obtained in cells with stably integrated constructs harboring similar gene copy numbers.

Numerical analysis of transient co-transfection of constitutively expressed genes

An analytical solution of Eq 19 does not exist, and we solve it using numerical simulations. To this end, we performed *in-silico* simulations of a transient co-transfection containing multiple ($N = 5$) independent genetic constructs (Methods). The change in protein expression over time, $\dot{O}_i$, of each gene g_i can be described by an ordinary differential equation (ODE) with kinetic parameter $\bar{b}_i$, gene copy number k_i and degradation rate δ_i

$$\dot{O}_i = \bar{b}_i k_i - \delta_i O_i \quad (20)$$

In the steady state, i.e. $\dot{O}_i = 0$, the steady-state level of O_i is proportional to k_i with the global coefficient of proportionality $\beta_i = \bar{b}_i/\delta_i$ and identical to Eq 7:

$$O_i^{s.s.} = \beta_i k_i \quad (21)$$

Iterating multiple times to simulate multiple single cells j ($1 \leq j \leq C$, where C is the total number of simulated "cells"), we draw the multiplicity m_j from a lognormal distribution (Eq 12) with parameters that roughly fit experimental data (see below) and initialize gene copy number vectors

$$\boldsymbol{k_j} = [k_{j1}, k_{j2}, \ldots, k_{jN}] \quad (22)$$

according to Eq 11 with pre-set parameters (Methods). To create log-normal protein distributions given $\boldsymbol{k_j}$ according to Eq 10, for each k_{ji} in $\boldsymbol{k_j}$, local proportionality factor b_{ji} is drawn from a log-normal distribution:

$$\mathcal{LN}\left(\ln(\beta_i) - \frac{\sigma^2}{2}, \sigma\right) \quad (23)$$

with fixed β_i values (Methods, S1 Table); the values of σ are fixed for a given simulation run and systematically varied between 0.00 and 0.32 in different runs. A value

$$O_{ji} = k_{ji} b_{ji} \quad (24)$$

is the level of protein O_i in cell j (S1 Fig).

The generated *in-silico* dataset (S2A Fig) is similar to a flow cytometry dataset what one would obtain in a transient co-transfection experiment of constitutively-driven genes. In order

to confirm that the parameters, and in particular the values of σ are realistic, we transiently co-transfected five plasmids, each expressing constitutively different fluorescent protein (O_1: SBFP2, O_2: Cerulean, O_3: Citrine, O_4: mCherry and O_5: iRFP; Methods) (S2B Fig). We find that the standard deviation of the log-transformed protein expression distribution in cells pre-binned on similar values of the reference protein, which we denote σ^*,

$$\sigma^* = \sigma_{Y_i|Y_1} \tag{25}$$

depends on Y_1, and indeed ranges between 0.1–0.3 (S2C Fig). Higher values of $\sigma^*$are observed as very low Y_1 values, and they plateau towards 0.1 for larger Y_1. Accordingly, the range of σ values used in the simulations, and given that $\sigma<\sigma^*$, constitutes a realistic range for the gene expression variability due to "intrinsic noise".

Next, we simulate transient co-transfections using two different gene ratios; (i) equimolar and (ii) a ratio of 1.0:1.3:0.8:0.5:0.4, the latter following some fine-tuning in a parallel experimental project (manuscript under preparation), to generate a joint *pdf p(**Y**)* (Figs 1A, S3A, S3B, S4A and S4B). We use these datasets to solve Eqs 17 and 19 numerically, that is, determine the $Y_i^{MODE}(Y_1)$ and $k_i^{MODE}(Y_i^{MODE}(Y_1))$. To do so, we bin cells that share a log-transformed reference protein value Y_1, evaluate the conditional *pdf* $p(Y_i|Y_1)$ and, first, determine numerically the value of $Y_i^{MODE}(Y_1)$. Second, we retrospectively look up the values of k_i in cells whose Y_i and Y_1 expression levels lie in the vicinity of a vector $[Y_i^{MODE}(Y_1); Y_1]$. The empirical distribution of k_i (Figs 1B, S3C and S4C) is $p(k_i|Y_i^{MODE}(Y_1))$ from Eq 19, and the mode of the gene copy number distribution, $k_i^{MODE}(Y_i^{MODE}(Y_1))$, or $k_i^{MODE}(Y_1)$, is determined numerically (Figs 1C, S3D and S4D, Methods).

According to Eq 7 there is a simple, linear relation between O_i and the gene copy number k_i, linking them via the global coefficient of proportionality β_i. The global coefficient of proportionality can be determined experimentally using *e.g.*, a calibrated Western blot to measure the absolute amount of protein and calibrated qPCR to measure absolute mean internalized gene

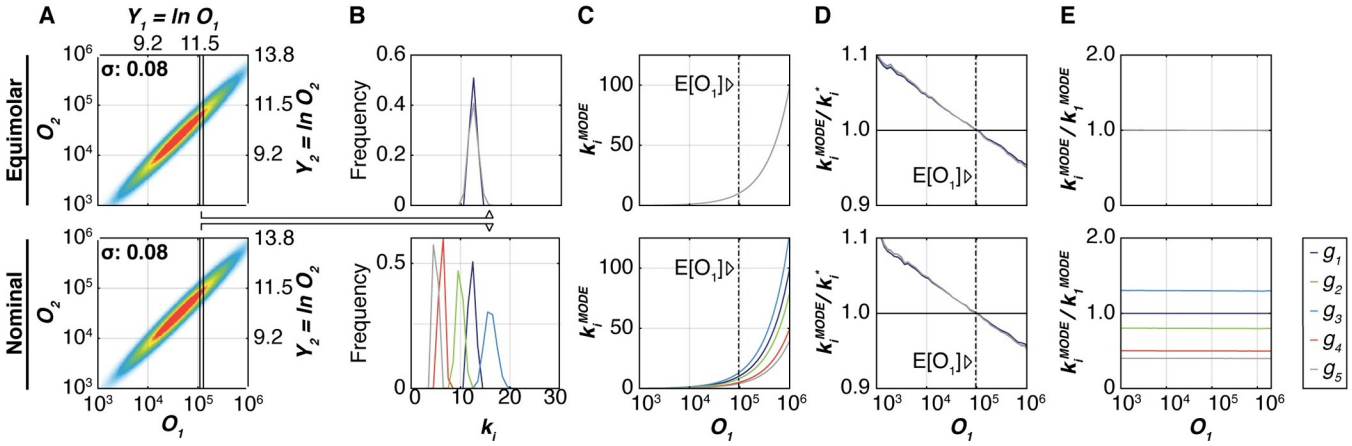

Fig 1. *In-silico* simulation of multiple gene co-transfections. *In-silico* simulations shown for two cases with parameter settings $\sigma = 0.08$, $\varepsilon = 0.04$ and $m = 10$: equimolar (a_1:a_2:a_3:a_4:a_5 = 1.0: 1.0: 1.0: 1.0: 1.0; top row) and "nominal" (a_1:a_2:a_3:a_4:a_5 = 1.0: 1.3: 0.8: 0.5: 0.4; bottom row) ratio of gene mix. **(A)** Density plots show the amount of expressed proteins O_2 versus O_1. Solid black lines indicate the edges of an example bin for the transfection reference protein O_1. **(B)** The plots show the distribution of gene copy numbers in cells whose O_1 values fall into the bin shown in panel A. Copy number distributions corresponding to different genes g_i are shown using different colors (legend on the very right of the figure). In the equimolar case, gene copy number distributions overlap, while in the "nominal" case they are separated. **(C)** The modes of the copy number distributions are plotted versus the median signal of the transfection reference protein O_1 in all bins (colored lines). The dash-dotted line marks the mean ($E[O_1]$) of the global O_1 distribution. **(D)** For each bin of the transfection reference protein O_1, the modes of the gene copy number distributions from each gene g_i are determined numerically. The ratio of the numerically-determined mode of the copy number distribution k_i^{MODE} and the anticipated copy number k_i^* are computed and plotted versus the corresponding O_1 values in individual bins. The global mean of O_1 is shown with a dash-dotted line. **(E)** Ratios of gene copy number modes k_i^{MODE} relative to the gene copy number mode of the transfection reference protein k_1^{MODE}, as a function of the O_1 median value in the bins.

https://doi.org/10.1371/journal.pcbi.1008389.g001

copy numbers. For the transfection reference protein, we introduce the variable k_1^* that corresponds to the gene copy number that one would "naïvely" anticipate to be the most probable value leading to a particular Y_1, given β_1:

$$k_1^*(Y_1) = \frac{exp(Y_1)}{\beta_1} \tag{26}$$

For the other log-transformed outputs Y_i the naïvely anticipated copy number in cells that express certain level of the reference protein Y_1, is defined by a similar relationship:

$$k_i^*(Y_1) = \frac{exp(Y_1)}{\beta_1} \cdot a_i \tag{27}$$

Given that the value of Y_1 and β_1 are the only "knowable" parameters, it is of interest to ask how the actual copy numbers relate to these anticipated values. Using our simulated datasets, we compute the ratio between numerically found $k_i^{MODE}(Y_1)$ and the anticipated copy number $k_i^*(Y_1)$ from Eq 27 (Fig 1D) as a function of Y_1. We find that the deviation from the anticipated value is a decreasing monotonous function of Y_1 with the following properties: (1) The deviation is always positive for values of $Y_1 < \ln E[O_1]$; (2) the deviation is essentially zero when $Y_1 = \ln E[O_1]$, and (3) it is negative for $Y_1 > \ln E[O_1]$. Further, the absolute magnitude of the deviation increases with increasing σ (S3E and S4E Figs). However, for all noise levels, the deviation is zero at the global mean of the O_1 distribution, $E[O_1]$, and

$$\frac{k_i^{MODE}(ln(E[O_1]))}{k_i^*(ln(E[O_1]))} \approx 1 \tag{28}$$

We further analyzed the ratio of the modes of gene copy number distributions k_i^{MODE} to k_1^{MODE} in cells that reside in the close vicinity of the log-transformed expression vector $[Y_1, Y_2^{MODE}(Y_1), Y_3^{MODE}(Y_1), \ldots, Y_N^{MODE}(Y_1)]$. The ratio stays constant for almost the entire range of Y_1 values (Figs 1E, S3F and S4F). Since the naïve estimate and the numerical mode of the absolute gene copy number coincide at the global mean of the O_1, both the relative and the absolute abundance of the gene copy numbers can be deduced with high certainty in cells that express O_1 around its global mean. This is true regardless of the chosen distribution of m and β_i. Simulations that employ Poisson, Gamma or lognormal distributions show a strikingly similar effect (S5 Fig). Appropriate experimental techniques allow measuring both the protein copy number [39] and the gene copy [6] number in the cells residing at the global mean of O_1, making it possible to determine the value of β_1 experimentally and therefore extrapolate directly to the ground truth expected in the stable cell line with the similar gene copy number.

Numerical analysis of transient co-transfection of non-trivial gene circuits

Next, we consider the case when the same genes, apart from the transfection reference protein gene g_1, encode a set of genes interacting in a circuit. Depending on the circuit, log-transformed distribution Y_i of protein O_i in cells pre-binned on the value of Y_1 may exhibit mono-, bi- or multimodality. We may consider the joint probability distribution of the vector of independent constitutive genes and their gene products, $\boldsymbol{k}\cdot\boldsymbol{Y}$, as a baseline state of any circuit. When the genes are interconnected (not including the reference gene g_1 and its log-transformed product Y_1), this baseline distribution is transformed because the values of $\boldsymbol{Y}$ are no longer independent. However, the values of $\boldsymbol{k}$ remain the same, because they represent the exact same underlying process of DNA delivery, and only the $\boldsymbol{Y}$ values change relative to the

independent, constitutive values. We hypothesize that despite the fact that the values of $\boldsymbol{Y}$ are no longer independent of k_j for $i \neq j$, $\boldsymbol{k}$ vectors corresponding to the (possible multiple) multivariate modes of $\boldsymbol{Y}|Y_1$, would not deviate far from the $\boldsymbol{k}$ vectors obtained in the case of independent co-transfection. We further hypothesize that this deviation will decrease as the noise in the system increases to biologically-plausible levels.

To test these hypotheses, we simulated two three-node gene circuit architectures (currently being investigated experimentally in a related project, see S6 Fig for the experimental results of the fan-out circuit); a simple monomodal *fan-out* circuit (*FO*; Fig 2A) and a non-trivial

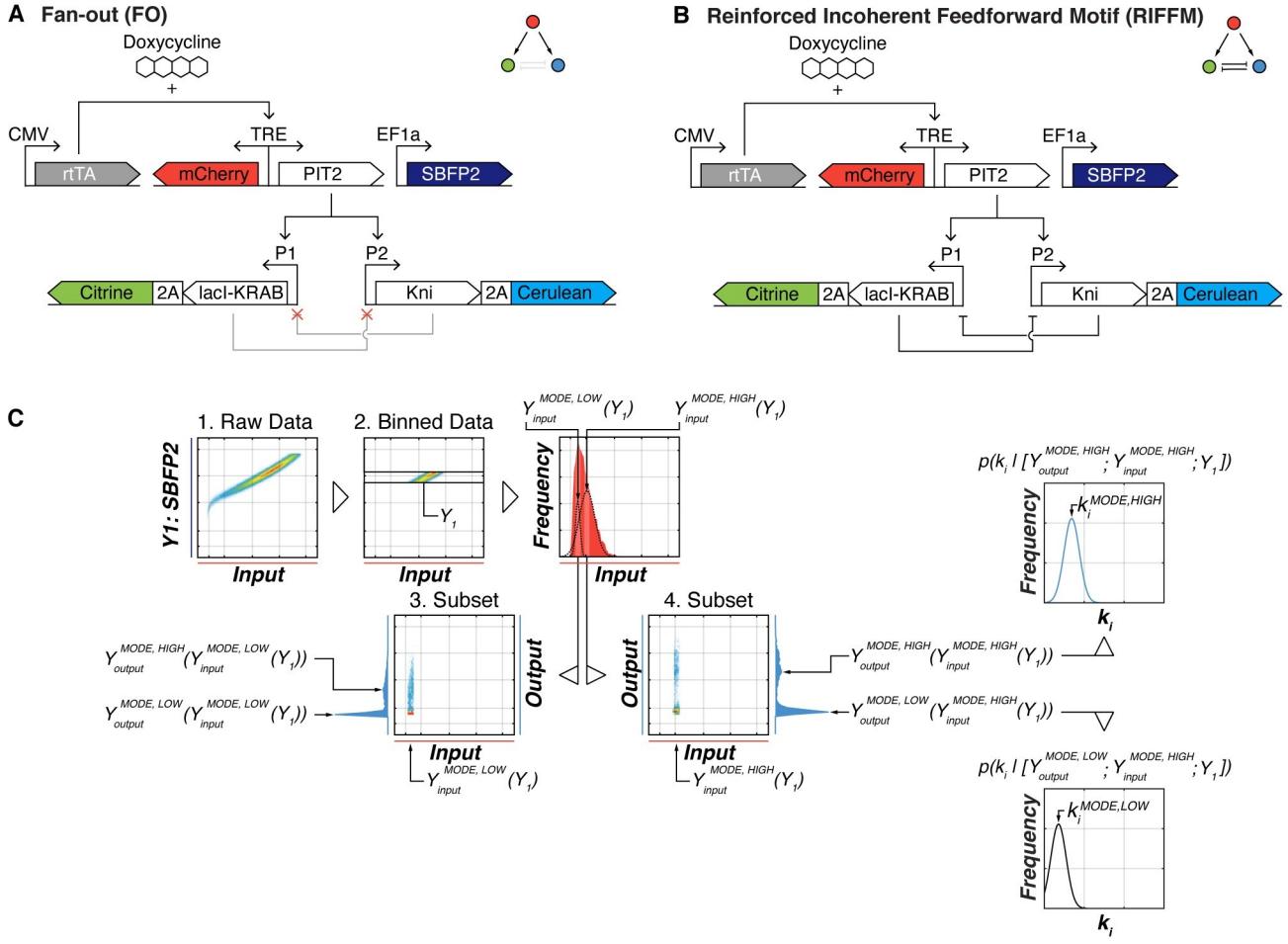

Fig 2. Schematics of gene circuits FO and RIFFM and analysis outline. (**A**)| Monomodal/fan-out (FO) and (**B**) bi-modal (RIFFM) gene circuit. The circuits are composed of five independent genes. Constitutively-expressed transcription factor rtTA co-induces PIT2 and the fluorescent protein mCherry in a Doxycycline-dependent fashion. PIT2 in turn activates two promoters P1 and P2, which express a transcriptional repressor lacI-KRAB and Kni, respectively, co-expressed via P2A linkers with Citrine and Cerulean fluorescent reporters. In the FO circuit the repressors do not interact due to mutated promoter binding sites, while in the RIFFM circuit they repress each other, establishing a mutual inhibition. (**C**) Graphical illustration of all steps to find the modes within our simulated data sets. The raw flow cytometry like data (**1**) is binned by the transfection marker (i.e. Y_1: SBFP2). The binned data is isolated and subsequent analysis is done only on this subset (**2**). Afterwards the distribution of the log-transformed input signal (mCherry) within the binned subset is determined and at least one Gaussian is fitted to the distribution. A narrow bin(s) around the mode(s) (black arrows) of the fitted distribution(s) is determined (pink bars), thus obtaining a subset of the originally-binned dataset. The subset around the peaks ($Y_{input}^{MODE,LOW}(Y_1)$ and $Y_{input}^{MODE,HIGH}(Y_1)$) are now analyzed individually. For illustrative purposes, we show the input (mCherry) and output (Cerulean) signal of the subset around $Y_{input}^{MODE,LOW}(Y_1)$ (**3**) and $Y_{input}^{MODE,HIGH}(Y_1)$ (**4**). The modes of the log-transformed output distributions are identified using a similar peak finding procedure as for the input signal (mCherry). The output histograms and their modes (black arrows) are shown on the sides of plot. Within these modes, we look at the *pdf* of the gene copy number distributions for all circuit genes as well as the transfection reference gene and identify the vectors $\boldsymbol{k}^{\boldsymbol{MODE,HIGH}} = [k_1^{MODE,HIGH}, \ldots, k_i^{MODE,HIGH}, \ldots, k_N^{MODE,HIGH}]$ and $\boldsymbol{k}^{\boldsymbol{MODE,LOW}} = [k_1^{MODE,LOW}, \ldots, k_i^{MODE,LOW}, \ldots, k_N^{MODE,LOW}]$, respectively.

https://doi.org/10.1371/journal.pcbi.1008389.g002

pitchfork bifurcation circuit, also known as *reinforced incoherent feed forward motif RIFFM* [4,30,40,41] (Fig 2B). The input to the circuit is a transcriptional activator PIT2 [42], whose level is tuned by Doxycycline via an bi-directional TRE promoter that also drives a fluorescent protein mCherry as a proxy for input expression. The first PIT2 target promoter (P1) drives the *D. melanogaster* derived transcriptional repressor Knirps (*kni*) and translationally-linked fluorescent protein Cerulean, constituting the first circuit output. The second PIT2 target promoter (P2) drives the transcriptional repressor LacI fused to a KRAB domain and translationally linked to a fluorescent protein Citrine, representing the second circuit output. In *RIFFM* circuit, *kni* is able to repress P2 while LacI is able to repress P1; in *FO*, the mutual repression is eliminated via mutations.

We built mechanistic kinetic models of the circuits *FO* and *RIFFM* (S2 Text "Simple Fan-Out Model" and S3 Text "Detailed Models") and simulated the flow cytometry dataset for multiple transiently transfected cells j ($1 \leq j \leq C$, where C is the total number of simulated "cells"). As above, every gene is encoded on a separate plasmid. We also compared this to a single-plasmid setup with all five genes are located on a single DNA backbone, but saw only a marginal difference in outcomes (S7 Fig). The multiplicity of transfection and the gene copy numbers are simulated as above (S1A Fig); the gene copy numbers become the initial conditions for running a simulation. Differently from that case of constitutive co-transfection, we directly simulate circuit dynamics governed by kinetic parameters $\boldsymbol{p}$; to simulate the effects of intrinsic gene expression noise, the parameters that govern protein translation rates are sampled independently from a lognormal distributions with nominal parameter values π_i and preset "noise" levels ranging, as above, from 0.00 to 0.32:

$$\mathcal{LN}\left(\ln(\pi_i) - \frac{\sigma^2}{2}, \sigma\right) \tag{29}$$

For every cell j, the drawn parameter values p_{ij} are used in a dynamic simulation ran to a steady state, with the simulated steady state input and output protein levels corresponding to the readouts from that cell.

First, we simulate mono- and bi-modal gene circuits for a single Doxycycline/input level. Similar to the data analysis above, we bin the cells according to Y_1 value. In the bin, we first focus on the input protein Y_{input} and identify its mode. Importantly, in the general case the distribution of $Y_{input}|Y_1$ can be bi-modal, leading to two numerically-found values $Y_{input}^{MODE,HIGH}(Y_1)$ and $Y_{input}^{MODE,LOW}(Y_1)$. In this case, we consider separately the cells residing close to the expression vectors $[Y_{input}^{MODE,HIGH}(Y_1); Y_1]$ and $[Y_{input}^{MODE,LOW}(Y_1); Y_1]$. Next, for every circuit output we consider the distributions $Y_{output}|Y_{input}^{MODE,HIGH}(Y_1)$ and $Y_{output}|Y_{input}^{MODE,LOW}(Y_1)$. These distributions can also be multimodal; in what follows we assume they are bi-modal. We denote the modes of the output distribution corresponding to the high mode of the input $Y_{output}^{MODE,HIGH}(Y_{input}^{MODE,HIGH}(Y_1))$ and $Y_{output}^{MODE,LOW}(Y_{input}^{MODE,HIGH}(Y_1))$, and use similar notation for the output modes corresponding to the low mode of the input, if the latter is present. Lastly, we consider all cells in the vicinity of the expression vector $[Y_{output}^{MODE,HIGH}(Y_{input}^{MODE,HIGH}(Y_1));\ Y_{input}^{MODE,HIGH}(Y_1);\ Y_1]$ and $[Y_{output}^{MODE,LOW}(Y_{input}^{MODE,HIGH}(Y_1));\ Y_{input}^{MODE,HIGH}(Y_1);\ Y_1]$ and evaluate the copy number distribution of every circuit gene as well as the reference gene. These are monomodal distributions, with the modes denoted respectively as $k_i^{MODE,HIGH}$ and $k_i^{MODE,LOW}$ (see Fig 2C for schematic description of the process). These numerically evaluated values are then compared to the naively anticipated values calculated according to the Eq 27. Note that in these simulations, the transfection reference gene expression is modelled explicitly as a transcription/translation/degradation

cascade with corresponding kinetic parameters; the value of β_1 for Eq 27 is calculated according to Eqs 20 and 21.

The analysis of the simulated data reveals the following: for the monomodal FO circuit, the behavior of the copy number modes of the input and the output genes is quantitatively identical to what is observed in the simulation of multiple constitutive gene co-transfection (Fig 3A–3E). The bi-modal circuit (Fig 3F) shows its bi-modal behavior at the lower intensities of the transfection reference protein O_1 (10^3–10^5). In this range, distributions of gene copy numbers for the high and low modes of output expression are slightly diverging (Figs 3G–3K). They are, however, almost fully overlapping, and their modes differ only by a few percent respectively upwards or downwards relative to the monomodal case, despite large difference in the corresponding protein modes. We quantify the divergence in gene copy number modes between high and low protein output modes by introducing a metric

$$\Delta\tilde{k}_i = \tilde{k}_i^{HIGH} - \tilde{k}_i^{LOW} = \frac{k_i^{MODE,HIGH}}{\hat{k}_1^{MODE}} - \frac{k_i^{MODE,LOW}}{\hat{k}_1^{MODE}} \tag{30}$$

with $\hat{k}_1^{MODE}$ being the mode of the transfection marker copy number distribution found in the constitutive co-transfection simulation (Fig 1). We observe a steady increase in $\Delta\tilde{k}_i$ upon an increase in noise level σ (Fig 3L), however, it is less that 10% for realistic levels of noise. Furthermore, we quantify bi-modality-dependent deviations of gene copy number ratios between the high and low output modes and introduce the metric

$$\Delta\phi_i = \phi_i^{HIGH} - \phi_i^{LOW} = \frac{k_i^{MODE,HIGH}}{k_1^{MODE,HIGH}} - \frac{k_i^{MODE,LOW}}{k_1^{MODE,LOW}} \tag{31}$$

Unlike $\Delta\tilde{k}_i$, this deviation of gene copy number ratios, $\Delta\phi_i$, decreases with an increase in the noise level σ (Fig 3M). These observations confirm our hypothesis that even in multimodal circuits, the cells that share the same amount of the reference protein, also share very similar gene copy numbers, both in absolute and especially, in relative terms (S8–S10 Figs). Moreover, deviations from the nominal gene copy number ratio decrease with the increase of noise levels σ.

The rationale for extracting input/output relationship from transient transfection data

The analyses above suggest a workflow for analyzing and deducing input/output relationships from transiently-transfected circuits. To summarize the findings so far, we show that cells, which express a certain level of reference protein O_1 and reside at multivariate modes of log-transformed input and output expression, harbor both input and output genes (or plasmids) with the following properties: (1) even for multimodal outputs with large differences in protein level modes, the distribution of the input and output genes copy numbers corresponding to the different output protein modes are almost overlapping, with the copy number modes varying by about 10% for biologically-realistic noise values, and thus can be treated as the same copy number for all practical purposes; (2) the copy number distribution modes' ratio almost exactly corresponds to the nominal ratio used in a transfection for all reference protein bin values, for both low and high output modes, and the deviation from the nominal ratio decreases with increased noise; (3) the distribution modes' absolute values exactly match the naïve anticipation (Eq 27) when the reference protein is expressed at the level $E[O_1]$; (4) the distribution modes' absolute values (for both high and low output modes) deviate from the naïve expectation in a predictable linear fashion as a function of log-transformed reference protein level Y_1,

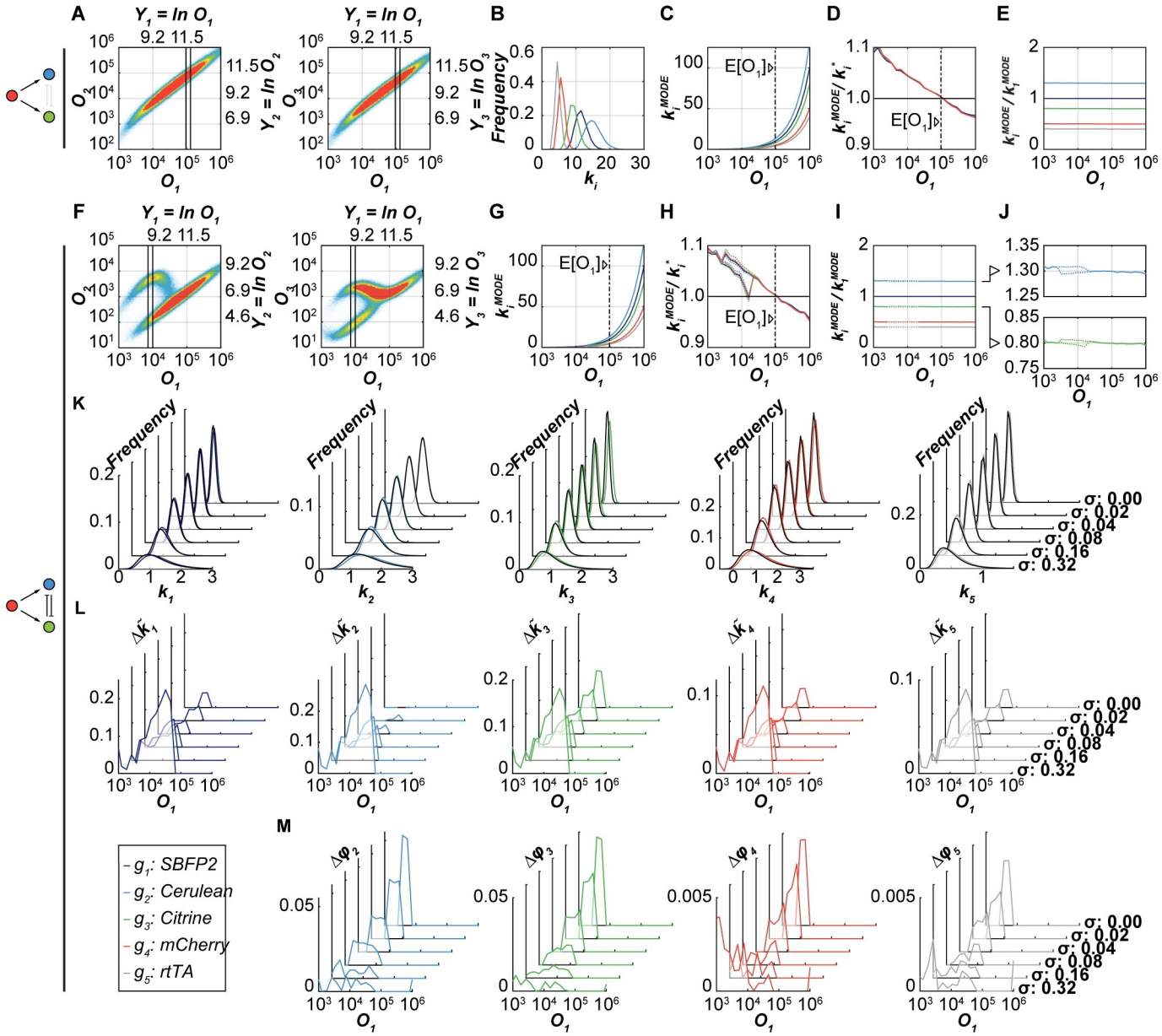

Fig 3. *In-silico* simulation of transiently-transfected gene circuits FO and RIFFM and the copy number analyses at various σ noise levels. *In-silico* simulations (global parameters $\sigma = 0.08$, $\varepsilon = 0.04$, $\mu_m = 1.4979$, $\sigma_m = 1.2686$ and $a_1{:}a_2{:}a_3{:}a_4{:}a_5 = 1.0{:}\ 1.3{:}\ 0.8{:}\ 0.5{:}\ 0.4$) shown for two circuits: **A-E** monomodal/fan-out (FO) and **F-M** bi-modal (RIFFM). **(A)** Raw data of the simulated transiently transfected FO circuit. The amount of expressed proteins O_2 (left: Cerulean) or O_3 (right: Citrine) versus O_1 (transfection marker: SBFP2) is shown as a density plot. Solid lines indicate the edges of a transfection marker bin. **(B)** Gene copy number distributions of cells binned by a particular value of log-transformed O_1 in a bin shown in panel **A**. **(C)** The modes of the copy number distributions, k_i^{MODE}, are plotted versus the median signal of the transfection reference protein O_1 for all bins (colored lines). The dash-dotted line marks the mean ($E[O_1]$) of the global O_1 distribution. **(D)** The ratio of the numerically-determined mode of the copy number distribution k_i^{MODE} and the anticipated copy number k_i^* plotted versus the O_1 values in individual bins. The global mean of O_1 is shown with a dash-dotted line. **(E)** Modes of the gene copy number distributions k_i^{MODE} normalized by the mode of the transfection reference gene k_1^{MODE} are shown as a function of O_1 values in individual bins. **(F)** Raw data of the simulated RIFFM circuit with bi-modal output O_2 (Cerulean; left) or O_3 (Citrine; right). The black lines indicate a bin within the bimodal range. **(G)** The modes of the copy number distributions, $k_i^{MODE,HIGH}$ and $k_i^{MODE,LOW}$, are plotted versus the median signal of the transfection reference protein O_1 of all bins (colored lines). Dashed segments indicate the range of O_1 in which the high and low modes do not coincide; the black dash-dotted line indicates the mean ($E[O_1]$) of the global O_1 distribution. **(H)** The ratios of the numerically-determined mode of the copy number distribution $k_i^{MODE,HIGH}$ and $k_i^{MODE,LOW}$, and the anticipated copy number k_i^* are plotted versus the corresponding O_1 values in individual bins. Dashed segments indicate the range of O_1 in which the ratios corresponding to high and low modes do not coincide. The global mean of O_1 is shown with a straight dash-dotted line. **(I)** Modes of the gene copy number distributions $k_i^{MODE,HIGH}$ and $k_i^{MODE,LOW}$, normalized by the mode of the transfection reference gene k_1^{MODE} are shown as a function of O_1 values in individual bins. Dashed lines indicate the range of O_1 where the values do not coincide. **(J)** The ratios k_i^{MODE}/k_1^{MODE} depicted in panel **I** are shown in greater detail for the outputs O_2 (Cerulean; top) and O_3 (Citrine; bottom). **(K)** Fitted gene copy

number distributions of the indicated bin in **F** are shown for all genes (left to right) and noise levels σ. Black curves indicate the fitted distributions to the low output mode, k_i^{LOW}, and colored curves the fitted distributions of the high output mode k_i^{HIGH}. **(L)** Divergence in gene copy number modes $\Delta\tilde{k}_i$ between high and low protein output modes normalized to the mode of copy number distribution of the transient co-transfection for all noise levels σ as a function of O_1 values in individual bins. **(M)** Difference of copy number modes' ratios $\Delta\phi_i$ in the high and low protein output modes for all noise levels σ as a function of O_1 values in individual bins.

https://doi.org/10.1371/journal.pcbi.1008389.g003

with the magnitude of the deviation increasing with the overall noise level. However, because the actual noise level and thus the degree of deviation can be quantified experimentally, even in cells that lie away from $E[O_1]$ the copy numbers can be estimated not only in relative but also in absolute terms. We show that this holds for a case of a complex circuit that generates bi-modal output distribution, with only slight deviations of the copy number modes from the expectation. Accordingly, by analyzing the input and output values in the cells that reside in the multivariate modes of circuit inputs' and outputs' distributions (after binning by the log-transformed reference protein value Y_1), we should be able to extract the information about the input/output response of the circuit that is comparable to the stable cell line harboring the circuit at the copy number derived from Y_1 according to Eq 26 and corrected by the measure of deviation that depends on the noise level.

Overview of the workflow validation procedure

To validate the workflow suggested above, we simulate transient transfection and stable integration for *FO* and *RIFFM* circuits for a wide range in circuit input levels using the exact same ODE model, with the input modulated via varying Dox level; otherwise the parameter randomization is performed as described above according to Eq 29. To simulate a stable integration dataset, we initialize a copy number vector $\boldsymbol{k}$ such that the ratio between individual genes corresponds to the nominal ratio of the transient transfection, and the absolute copy number of the reference gene is set to different fixed values corresponding to the bins used for transient transfection data analysis (S11 Fig; Methods). After the datasets are simulated, we extract input/output relationships corresponding to various bins of log-transformed values of the (transfection) reference protein Y_1 from the simulated transient transfection data, as described in detail in the next section. This is compared to the results of the stable integration simulation performed for the copy numbers that correspond to those reference protein levels. The simulation of the stable integration scenario generates an input/output "cloud", as has also been demonstrated experimentally [8,43,44]. The cloud can be used "as is" for the purpose of comparison, or it can also be processed via output mode identification for different input levels and building averaged curves. The process is illustrated schematically in S12 Fig.

The input/output relationship is generated for each level of log-transformed reference protein Y_1. We make use of the datasets simulated with different Doxycycline levels and thus different amounts of input expressed per gene copy. (Note that Doxycycline does not affect the gene copy number or the expression of the transfection reference protein, it is not a direct input and is not a part of the input/output relationships that we seek.) For a given Y_1 bin, a single transfection experiment simulation with a fixed Doxycycline value generates (for a bimodal case) at most four points on the input/output curve: $[Y_{input}^{MODE,LOW};\ Y_{output}^{MODE,LOW}(Y_{input}^{MODE,LOW})]$; $[Y_{input}^{MODE,LOW};\ Y_{output}^{MODE,HIGH}(Y_{input}^{MODE,LOW})]$; $[Y_{input}^{MODE,HIGH};\ Y_{output}^{MODE,LOW}(Y_{input}^{MODE,HIGH})]$; and $[Y_{input}^{MODE,HIGH};\ Y_{output}^{MODE,HIGH}(Y_{input}^{MODE,HIGH})]$. In most cases, the input will only exhibit a single mode that we denote for uniformity $Y_{input}^{MODE,HIGH}$ and thus only two points will be generated for a bimodal case, and one for a monomodal case. For this same reference protein bin, we repeat the procedure defined earlier (Fig 2C) for every Doxycycline level and generate multiple [*input*;

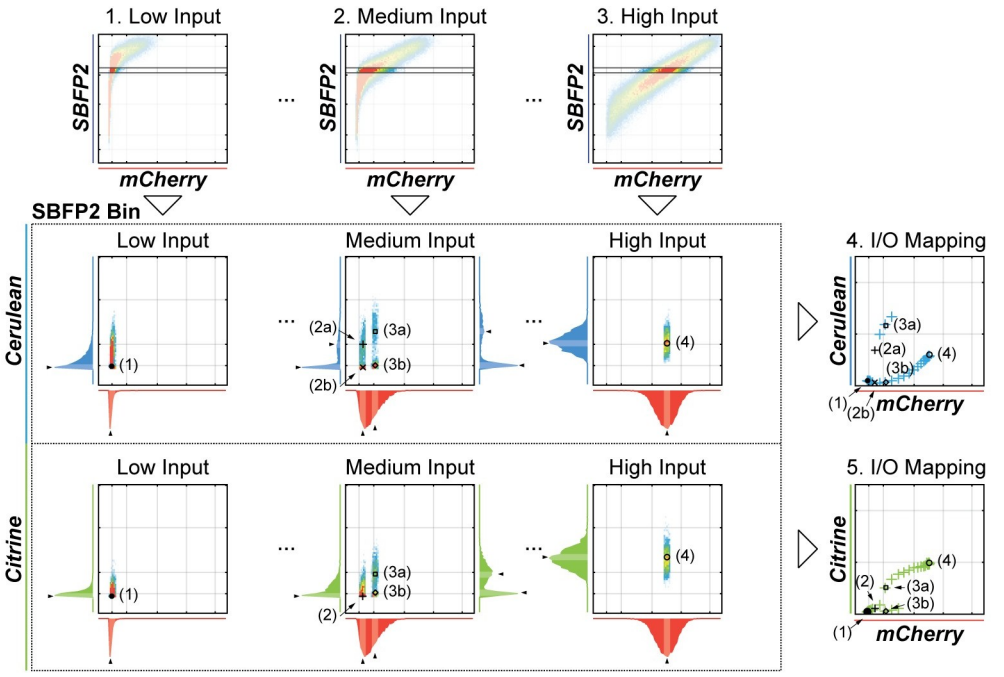

Fig 4. Peak Finder Algorithm For Flow cytometry (PFAFF) analysis strategy. The steps of finding peaks in distributions of a binned data as shown in Fig 2C are repeated for every input induction level (i.e. Doxycycline level; here three representative cases for low (**1**), medium (**2**) and high (**3**) input modulation levels are depicted) for each output (here, Cerulean and Citrine). Initially, a bin of the transfection reference protein (here, SBFP2) is determined and all downstream analyses only deals with cells residing in this bin. The plots in the dashed box show the example workflow of applying the peak finding to the individual input modulation levels. The density plots depict the binned data and the adjacent histograms show the distributions of their respective input and output proteins. Black wedges indicate the modes of the (convoluted) distributions and black markers indicate their location on the density plots. The [*input*; *output*] mode pairs, identified in this workflow (markers), derived from the raw data corresponding to the different input modulation level are plotted on the input/output mapping charts the output Cerulean (**4**) and Citrine (**5**).

https://doi.org/10.1371/journal.pcbi.1008389.g004

output] pairs that cover the entire input range. The procedure can be done for any desired reference protein bin, thus showing circuit behavior for different absolute gene copy number of its components (Fig 4).

The fact that every transient transfection performed with a certain Doxycycline level generates only up to four, and usually one or two, points on the input/output relationship curve is slightly counterintuitive because a flow cytometry plot would reveal wide distribution of the input values. However, this distribution results from the variability in the copy number of the input gene in the cells and is therefore irrelevant to the determination of the input/output relationship. In order to characterize an entire curve, there must be a practical way to modulate input expression per gene copy and repeat the experiment multiple times, every time with a different degree of modulation. This can be done with Doxycycline as in our case; when this is not feasible, one can mimic input modulation by systematically changing the relative dosage of a constitutive input-expressing gene, or use a series of constitutive promoters of varying strengths. In another observation, when extracting the modes, the high output modes corresponding to both the low and the high input modes fall on the same curve when plotted against the input values; the same is true for the low output modes (S13 Fig). This is not surprising, because the input value is the only determinant of the output. Therefore, we pool high and low output modes, respectively, and interpret them as the averaged input/output relationship of a circuit; when the behavior is bimodal, two curves are generated.

Validation using direct simulation data

We applied our data generation tool for transient transfections to FO and RIFFM circuit architectures and simulated 500,000 cells at twelve different Doxycycline input concentrations and six noise levels of σ. In Fig 5A we show representative examples of the raw data from transient

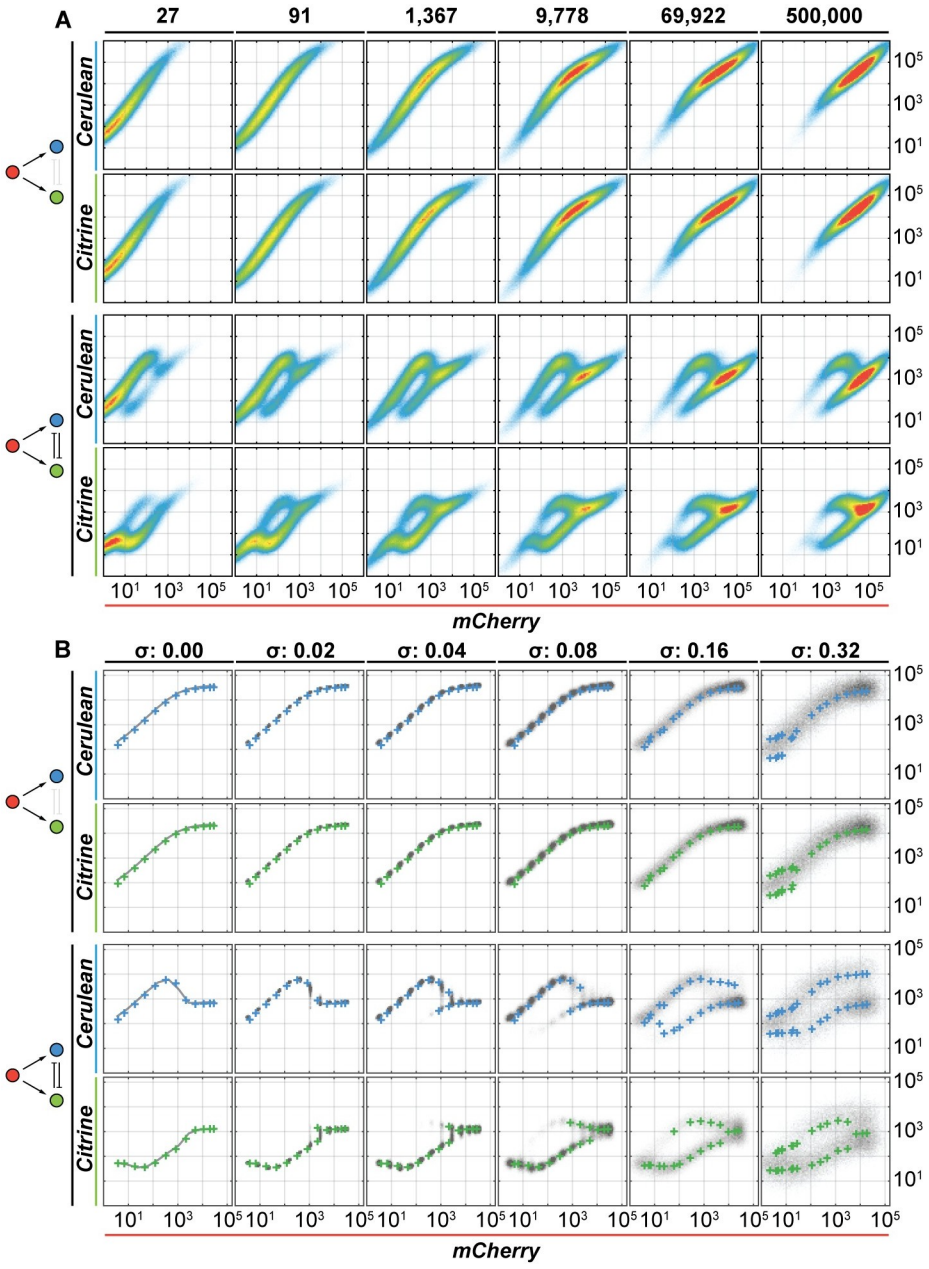

Fig 5. Results of PFAFF analysis on simulated FO and RIFFM flow cytometry data set. (A) Raw simulated transient transfection data for circuits FO and RIFFM at noise level $\sigma = 0.16$, modulated by different input levels of Doxycycline (columns). Each circuit (FO, RIFFM) is represented by two rows of charts depicting the input versus their respective output signals. **(B)** Simulated flow cytometry data set of the bin that lies at the transfection reference protein's global mean (transfection reference: SBFP2) at different gene expression noise levels σ (columns) ranging from 0.00–0.32. The plots show the input/output curves extracted by PFAFF (colored crosses; top row: Cerulean, bottom row: Citrine) atop of the simulated stably integrated circuit at the indicated gene expression noise level σ (grey density plot). FO undergoes an activation in both output colors. The RIFFM circuit shows a bi-modal behavior already at low noise levels for both output colors.

https://doi.org/10.1371/journal.pcbi.1008389.g005

circuit simulations at a single noise level ($\sigma = 0.16$) and various Doxycyline modulations. Note the shift in the scatter plots in response to Doxycycline increase. For each noise level, we extract the corresponding input/output relations of the data set with our analysis strategy that we call **P**eak**F**inder **A**nalysis **F**or **F**low cytometry, or PFAFF. The algorithm bins simulated cells according to the expression level of the transfection reference protein SBFP2 each bin containing an equal number of cells (9.5% of total population, 10 bins in total). Next, we determine the modes of log-transformed input and output protein distributions of cells residing in each bin for the different Doxycycline levels as described above, and build the input/output relationships corresponding to that bin (S14–S19 Figs). In the stable integration scenario, we build datasets that correspond to different fixed sets of gene copy numbers. Specifically, the copy numbers are set to correspond to the median copy numbers of the bins used to process the simulated transient transfection data (Methods). We simulate 5,000 cells per Doxycycline value and repeat this for twelve different Doxycycline values to cover the entire input range. This simulation is repeated for each σ. These data serve as the gold standard to evaluate the performance of our method for transient transfection data processing, by how well the input/output relationships match the stable integration simulation for matching bin/stable copy number.

In Fig 5B we show the stable integration simulation and the analysis results from PFAFF, the latter extracted from a transfection reference bin that lies close to the global mean of the reference protein (i.e. SBFP2 bin 5; see S14–S19 Figs for all bins, input and noise levels); the former simulated for the gene copy number that corresponds to this reference protein level. Note that the number of transfection reference bins does not influence the outcome (S20 Fig). We plot the stable integration outputs at various input levels as density plots in the background (grayscale). For the lowest noise case ($\sigma = 0.00$), the density plots for the stable integration collapse to curves, as expected. Gradually increasing σ leads to increasingly diffuse input/output relationships. Atop these density plots we superimpose the mode values extracted by PFAFF from the corresponding simulated transient transfection data and binned for the cells that express the same level of the transfection marker as the stable integration. The analysis suggests that the input/output relationship extracted using PFAFF superimposes with the input/output "cloud" simulated for the stable integration, when both reflect similar underlying absolute gene copy number.

In order to expand the number of circuits for analyses, we simulated another commonly studied circuit family–the type-1 incoherent feed-forward motifs. Our simulations include two versions of the I1-FFL (one for each repressor; I1-FFL1 and I1-FFL2; Fig 6A and 6B). We applied the same analyses as before to both I1-FFLs and show the comparison of input/output from stable integrated and transiently transfected circuits (Fig 6C; see S21–S26 Figs for all bins, inputs and noise levels). As is the case for other two circuits, there is excellent correspondence between the input/output curves extracted from simulated transient transfection data, and the input/output behavior of the comparable simulated stable case. This motivated us to expand the number of circuits by a coherent feed-forward, a negative feedback and lastly a positive feedback motif; all of them showing similar, excellent agreement (S27–S29 Figs).

The initial qualitative analysis uncovers excellent overlap between the input/output relationships found by PFAFF, and the input/output clouds from the corresponding stable integration simulations. Only at the highest simulated levels of σ, i.e. 0.32, the PFAFF algorithm has minor difficulties with extracting expected input/output relationships. Indeed, a σ of 0.32 is much larger than variations observed typically in nature [9,45]. To obtain a quantitative measure of the correspondence, we extracted modes from the log-transformed protein expression distributions of input (mCherry) and outputs (Cerulean and Citrine) from the stable integration data sets with the same peak finder algorithm that we employ in PFAFF (Methods).

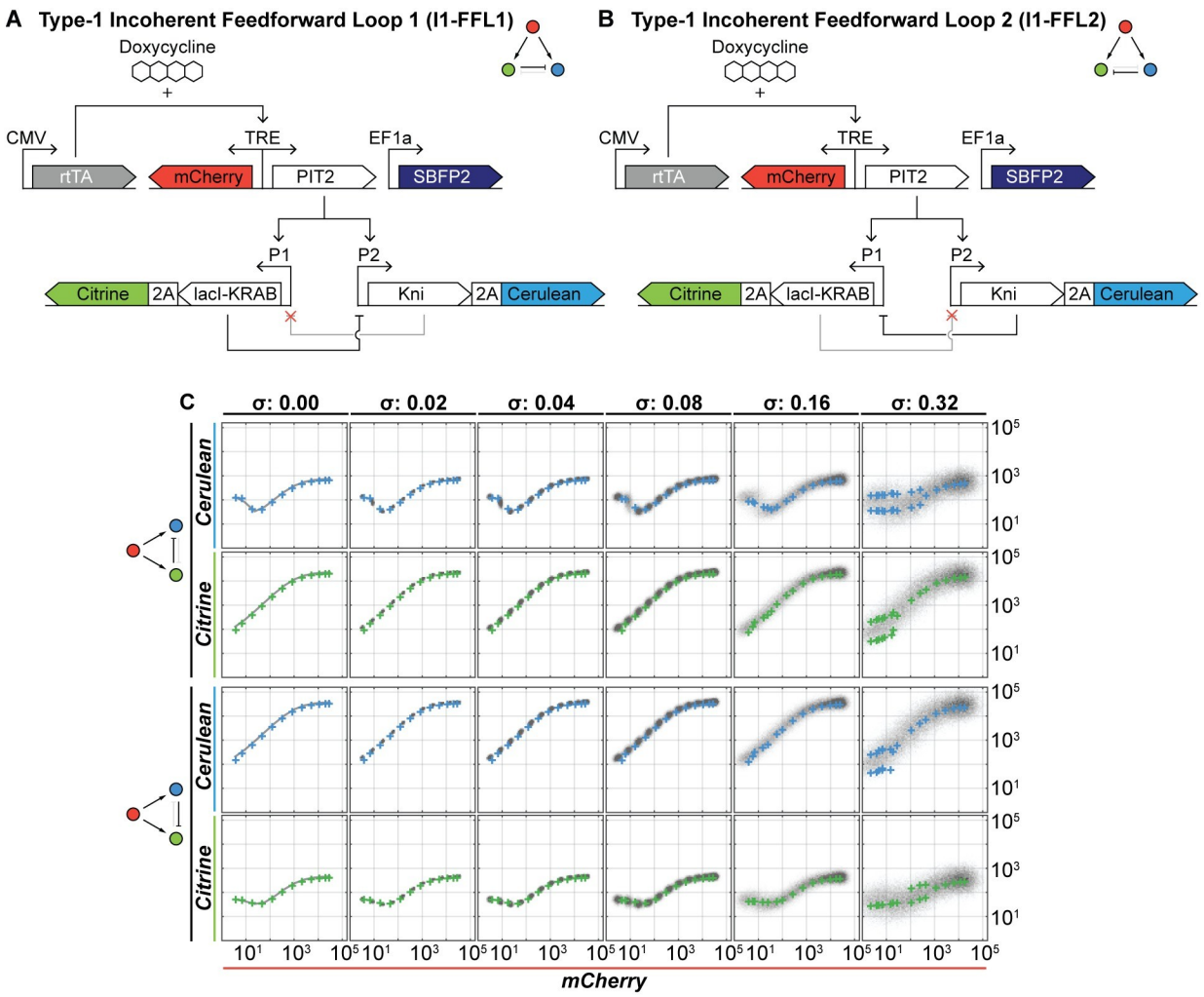

Fig 6. Results of PFAFF analysis on simulated I1-FFLs flow cytometry data sets. (A) and **(B)** Circuit architecture of two additional I1-FFLs. **(C)** Simulated flow cytometry data set of the bin that lies at the transfection reference protein's global mean (transfection reference: SBFP2) at different gene expression noise levels σ (columns) ranging from 0.00–0.32 is processed by PFAFF. The plots show the input/output curves extracted by PFAFF (colored crosses; top row: Cerulean, bottom row: Citrine) atop of the simulated stably integrated circuit at the indicated gene expression noise level σ (grey density plot). Both I1-FFLs show adaptive behaviors in their respective outputs.

https://doi.org/10.1371/journal.pcbi.1008389.g006

We correlated the obtained modes with the modes that were found by PFAFF in the transient transfection case (Fig 7). In the pooled modes from all data sets, meaning all external input levels and bins for each noise level, we find a high correlation between the modes from both simulation scenarios for all expression noise levels (mean of Pearson correlation coefficient $\rho > 0.91 \pm 0.01$ for output modes; σ: 0.00–0.32).

Discussion

In this study we show that transient transfection data can be used to extract input/output relationships of gene circuits that are comparable to the data that would have been obtained with stably-integrated circuits. Our findings reveal that it is sufficient to focus on small subsets of transiently transfected cells that lie at multivariate modes of input and output expression, post-binning on a transfection reference protein, otherwise known as a "transfection marker". We prove numerically that cells in these modes harbor distributions of circuit genes with the

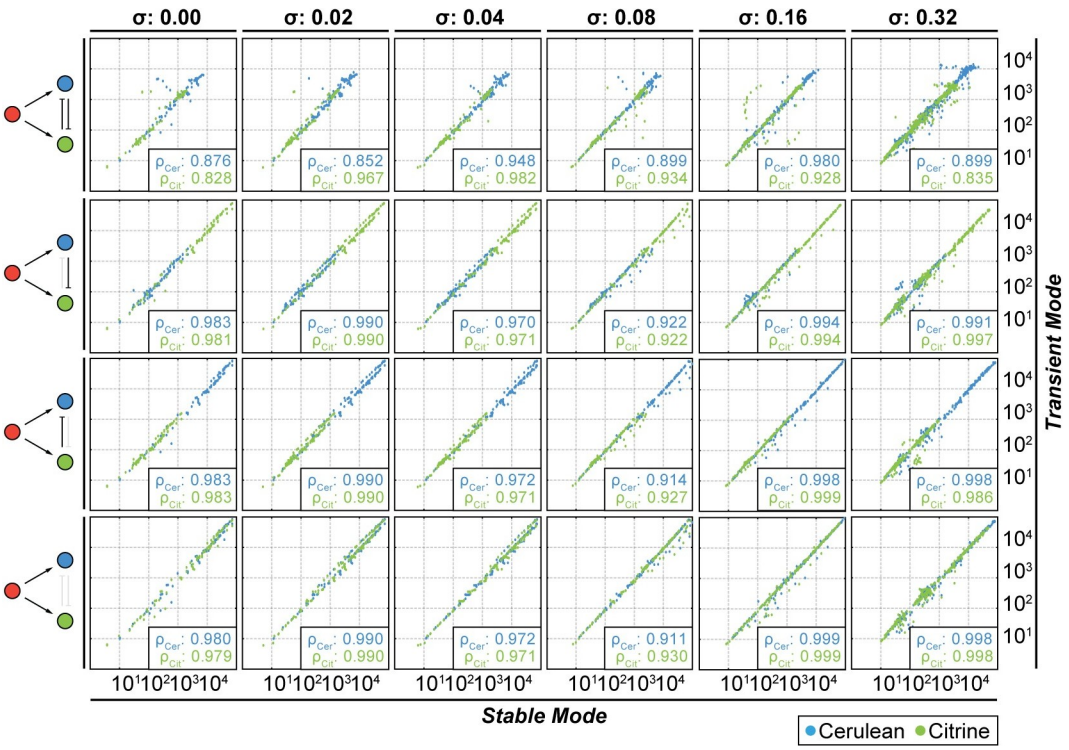

Fig 7. Quantitative analysis of the stable integration and transient transfection data sets. Correlation between extracted output modes of stable integration simulation and PFAFF results from transient transfection simulations. The mode values for both output colors (Cerulean and Citrine) for all input modulator levels and for all bins/stable copy number sets are plotted, and Pearson correlation coefficient (ρ_{Cer} or ρ_{Cit}) is shown for each plot.

https://doi.org/10.1371/journal.pcbi.1008389.g007

following properties: (1) the modes' absolute values can be deterministically deduced from the observed level of the transfection reference protein expression, the noise level in the experimental data, and the distance from the global mean of the transfection reference protein; (2) the modes' ratios are identical to the gene or plasmid ratios used in the transient transfection experiments, for all values of the reference protein expression. Moreover, in the case of multimodal data, large differences in protein expression do not translate into significant differences between the underlying gene copy number distributions, and for all practical purposes the absolute and relative copy numbers can be considered identical for the different protein expression modes. Interestingly, the cells that belong to the bin that lies close to the global mean of the reference protein expression, harbor gene copy numbers whose modes' absolute **and** relative values correspond exactly to what one would expect from the naive expectation, namely the ratio between the knowable expressed protein level and the knowable global coefficient of proportionality between the protein and gene copy number. The detailed understanding of the gene copy number behavior in the multivariate protein expression modes that can be identified in the experimental data, provides a degree of confidence in relevance of extracted data to the ground truth behavior of the same circuit when stably integrated in a cell. This confidence is confirmed by the direct *in-silico* validation experiments, where both types of data are directly simulated, the PFAFF workflow analysis is applied to the data from simulated transient transfections, and the results are compared to, and are shown to reproduce, the ground truth.

While transient transfections are often valued as a tool to rapidly analyze genetic circuit behavior, they are rarely used to draw fine-tuned conclusions about the input/output

relationship of corresponding stably integrated circuits, more so for multimodal circuits. This is likely due to various pitfalls in existing analysis methods, the most prominent being the insufficient treatment of multimodal systems and the lack of conclusive analysis of the underlying gene copy number distributions in identifiable cell populations. Our analysis strategy allows a thorough comparison of input/output relationships from both scenarios and results in an excellent agreement between them. This will play an important role in gene circuit design and characterization, as it alleviates the need to generate multiple stable cell lines.

Materials and methods

Cloning

Standard cloning techniques were used to clone all plasmids. We used *E. coli* DH5a and DH10B as the cloning strains, cultured in LB Broth Miller Difco (BD; Cat. no. 244610) and Ampicillin (100ug/ml, Sigma-Aldrich; Cat. no. A0166-5G) as selection medium.

Cell culture and reagents

All experiments were done with HEK 293 cells (Life Technologies) and were grown at 37˚C, 5% CO_2 in complete medium (DMEM (Thermo Fischer; Cat no. 11965092) supplemented with 10% fetal bovine serum (FBS; Sigma-Aldrich; Cat. no. F9665) and 1% Penicillin/Streptomycin (Sigma-Aldrich; Cat. no. P4333)). They were sub-cultured by seeding 10^6 cells into T75 flasks every 3–4 days.

Transfection

One day prior transfection, cells were passed through a 40um cell strainer (Falcon; Cat. No 352340) and counted with Bio Rad TC10. In each well (uncoated 6-well plates, Thermo Scientific Nunc; Cat. No. 2020–10) 300,000 cells were seeded and incubated for another 24 hours. On the day of transfection DNA was diluted in 250ul Opti-MEM I Reduced Serum (Gibco, Life Technologies Cat no. 31985–962) and mixed with a 244ul Opti-MEM I/6ul Lipofectamin 2000 Transfection Reagent (Thermo Fischer; Cat. no. 11668019). After a 20 minutes' incubation step at room temperature, the transfection mix was added drop wise to the wells. The cells were incubated for another 72 hours before being measured by flow cytometry.

Flow cytometry

All samples were measured with a BD LSR Fortessa cell analyzer. The medium was removed and cells were incubated with 300ul StemPro Accutase Cell Dissociation Reagent (Thermo Fischer; Cat. no. A1110501) at 37˚C, 5% CO_2 for 10 minutes. Reporter specific combinations were used to measure all four fluorescent proteins independently, but still providing a setup of little bleed over. In particular, we used for: SBFP2 a 405nm laser with 445/15, Cerulean 445nm laser with 473/10, Citrine 488nm laser with 542/27, mCherry 561nm laser with 610/20 and iRFP 640nm laser with 710/50 emission filter sets. We used the same PMTs (FSC: 350, SSC: 350, SBFP2: 220, Cerulean: 242, Citrine: 220, mCherry: 245, iRFP: 460) throughout all measurements and controlled for consistency of the instrument by using SPHERO RainBow Calibration particles (Cat no. 559123, BD).

Co-transfection experiment

We experimentally co-transfected five different fluorescent protein genes (SBFP2 [46], Cerulean [47], Citrine [48], mCherry [49], iRFP [50]; S30 Fig), individually driven by an Ef1a promoter and analyzed them via flow cytometry. The amount of transfected DNA (ng) was

adjusted according to each plasmid's size (nominal ratio: SBFP2: Cerulean: Citrine: mCherry: iRFP = 504: 634: 400: 239: 249). We collected more than 1,000,000 events and stringently gated the live population (~750,000 cells). This experiment created a five-dimensional distribution of fluorescent values.

Fan-out gene circuit experiment

We transfected five gene cassettes on individual plasmids as depicted in Fig 2A into HEK293 cells and activated the circuit through the addition of Doxycycline at eight different input modulation levels (0nM, 0.90nM, 3.15nM, 0.01uM, 0.05uM, 0.13uM, 0.45uM and 1.35uM). After 72h post-transfection we analyzed the induced cells using flow cytometry and collected more than 1,000,000 events per replicate ($n = 3$). The obtained data were subjected to our analysis pipeline as outlined in section Data Analysis.

Model

Co-transfection model. We generated the model of multiple constitutively expressed genes using a steady state approximation (Eqs 7 and 21). Once the gene copy number k_i and expression parameter β_i are determined, the protein output O_i is computed as described in S1 Fig.

Circuit models. ODE circuit models were created with Simbiology, a MathWorks MATLAB 2018b package. Each molecular interaction was modeled according to the law of mass action (S3 Text "Detailed Models" and parameter values in S2 Table). This includes binding and unbinding of a transcription factor to inducible promoters, transcription of mRNAs and translation into proteins. All circuits have the same underlying interaction map. We created four different topologies by inactivating the translation reaction of the respective repressor mRNAs. Therefore, we generated *in silico* "knock outs" with minimal changes to the model.

Parameters

Expression rate parameters. The vector of global proportionality coefficients $\boldsymbol{\beta}$ used in the simulation of constitutive gene co-transfection is a measure for the conversion of gene copy numbers to the number of proteins. In our circuit models, this coefficient is derived from expression and degradation rate constants. We adjusted the values of β_i according to the maximum expression levels of our circuit models to obtain similar and biologically feasible amounts (S1 Table).

In dynamic circuit models, binding/unbinding and transcription rates were either fitted from experimental data (Manuscript in preparation), literature values, or set arbitrarily at biologically feasible values. The translation rates π_i are based on previously-reported values [51] and were adjusted according to the length of each protein (S3 Text "Detailed Models").

Anticipated gene copy number k_i^*. In order to compute the values of anticipated gene copy numbers k_i^* from Eq 27, we first have to determine the Y_1 values. This is done by removing the top and bottom 0.1 percentile from the *in-silico* co-transfection and circuit simulations and binning the obtained data set into equally spaced bins (50 bins for co-transfection simulations; 25 bins for circuit simulations) according to the signal intensity of the transfection marker (O_1). Since the O_1 distribution within a bin is just a subset of the global O_1 distribution, we use the median of the log-transformed O_1 signal for each bin, which is then identified as Y_1 value. Together with the global proportionality coefficient β_1 (see above) and the abundance parameter a_i we can compute k_i^* according to Eq 27 for each bin of the respective data set.

In-silico flow cytometry simulations

Gene expression noise from lognormal distributions. Within our *in-silico* simulations, we introduce gene expression noise through the randomization of kinetic parameters. In the case of co-transfection simulations we randomize the vector of the proportionality coefficients $\boldsymbol{b}$ by drawing its individual values from a lognormal distribution $\mathcal{LN}(\mu_\beta, \sigma)$, with the mean μ_β being their coefficient of proportionality of the respective gene g_i (S1 Table) in log-space $\left(\mu_\beta = \ln(\beta_i) - \frac{\sigma^2}{2}\right)$ and the standard deviation σ being one of the six noise levels (0.00, 0.02, 0.04, 0.08, 0.16 or 0.32). In the case of circuit simulations, we introduce variability through the translation parameter vector $\boldsymbol{p}$. Likewise, it is drawn from a lognormal distribution $\mathcal{LN}(\mu_\pi, \sigma)$ with the mean being set according to S2 Table in log-space $\left(\mu_\pi = \ln(\pi_i) - \frac{\sigma^2}{2}\right)$ and the standard deviation being again one of the six noise levels.

Gene expression noise from Γ distributions. Gene expression noise is introduced by drawing individual values $\boldsymbol{b}$ of the proportionality coefficients from a Γ (gamma) distribution, $\Gamma(k,\theta)$. We chose the shape parameter k and the scaling parameter θ, so that the mean (β_i) and the variance of the distribution are the same as in the lognormal case. We simulated transient co-transfections at $\sigma = 0.08$.

Gene copy number/extrinsic noise. We introduce extrinsic noise in our transient transfection simulations by drawing gene copy numbers $\boldsymbol{k}$ from a five-dimensional multivariate normal distribution $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mean of the distribution, $\boldsymbol{\mu} = \boldsymbol{a}m$, depends on the multiplicity parameter m, which is drawn from a lognormal distribution $\mathcal{LN}(\mu_m, \sigma_m)(\mu_m = 1.4979,\ \sigma_m = 1.2686)$, and the abundance vector $\boldsymbol{a}$ (equimolar: a_1:a_2:a_3:a_4:a_5 = 1.0: 1.0: 1.0: 1.0: 1.0 or nominal: a_1:a_2:a_3:a_4:a_5 = 1.0: 1.3: 0.8: 0.5: 0.4). For our systematic comparison of gene copy number distributions (S5 Fig), we draw the multiplicity parameter m from a Poisson ($Pois(\lambda)$, $\lambda = 10$) and Γ distribution ($\Gamma(k,\theta)$, $k = 0.7436$, $\theta = 13.46$), respectively. The covariance $\Sigma = diag((\boldsymbol{a}m\varepsilon)^2)$ is diagonal matrix and depends on the abundance $\boldsymbol{a}$, the multiplicity m and the constant factor $\varepsilon = 0.04$. For every simulated cell, the multivariate distribution $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ changes along the lognormal distribution$\mathcal{LN}(\mu_m, \sigma_m)$ and in every iteration five gene copy numbers are drawn from it.

Single-plasmid simulations. The single-plasmid circuit contains all gene cassettes on a single entity. We achieve this by setting the constant factor ε to zero. Consequently, the covariance matrix of the five-dimensional multivariate distributions Σ also turns zero. The remaining simulation is performed as described below in section "Gene circuit simulations".

Co-transfection simulations. We simulated our simple model independently 5×10^6 times (C), randomizing both the parameters $\boldsymbol{b}$ and gene copy number $\boldsymbol{k}$, according to the description above. Each simulated run corresponds to a single cell and contains a randomized set of $\boldsymbol{b}$ and $\boldsymbol{k}$. We simulated the steady-state and the runs were stored in a single *.csv*-file that contained all information used to generate the data afterwards. This includes the individual parameters $\boldsymbol{b}$ and $\boldsymbol{k}$ for every cell as well as the output values. Thus, we obtained a data set that resembles a transient co-transfection experiment aided by the information of individual parameters. The simulations were performed in MATLAB 2018b.

```
program: Simulation Transient Co-transfection
   initialize abundance a, constant factor σ_e = 0.04
   for Noise-Level σ in [0.00, 0.02, 0.04, 0.08, 0.16, 0.32]
      for Cell j in 1:C
         draw m_j from LN(μ_m, σ_m), μ_m = 1.4979 and σ_m = 1.2686
         seed and draw k_j from N_N(μ, Σ), μ = a m_j, Σ = diag((a m_j σ_e)^2)
         draw b_j from LN(μ_β, σ), μ_β = ln(β) − σ²/2
         Model: O_j = k_j b_j
      end
```

```
10:    end
11:    save Simulation.csv
12: end program
```

Gene circuit simulations. *In-silico* simulations of flow cytometry data for our circuits requires a mathematical model (ODE) generated by MATLAB's Simbiology toolbox. The model was exported into the workspace and to decrease computational effort, we generated a SimFunction object. This function has five outputs: the number of fluorescent proteins (SBFP2, Cerulean, Citrine, mCherry) and transcription factor rtTA bound to Doxycycline at steady state (i.e. 1,500,000s). As the SimFunction's input serves a matrix wherein each column represents the gene copy number (k_1: pCS187, k_2: pCS171, k_3: pCS166, k_4: pCS200, k_5: pZ91), the Doxycycline level ***DOX*** ($Z = 12$ logarithmically spaced values from 10–500,000 molecules) and eight translation parameters $\boldsymbol{p}$, each for every protein ***O*** produced. Sets of gene copy numbers $\boldsymbol{k_j}$ and gene expression noise variations $\boldsymbol{p_j}$ were drawn as described above. The simulation input, output as well as all parameters used for each cell are stored in a table and saved as.*csv*-files for documentation and further analysis. The simulations were performed in MATLAB 2018b.

```
1: program: Simulation Transient Circuits
2:    initialize abundance a, constant factor σ_e = 0.04
3:       for Noise-Level σ in [0.00, 0.02, 0.04, 0.08, 0.16, 0.32]
4:          for Input-Level l in 1: Z
5:             for Cell j in 1: n
6:                draw m_ij from LN(μ_m, σ_m), μ_m = 1.4979 and σ_m = 1.2686
7:                seed and draw k_1j from N_N(μ, Σ), μ = a·m_ij, Σ = diag((a·m_ij·σ_e)^2)
8:                draw p_1j from LN(μ_π, σ), μ_π = ln(π) − σ^2/2
9:                initialize and simulate SimFunction-Model
10:            end
11:         end
12:      end
13:   save Simulation.csv
14: end program
```

Determine copy numbers for stably integrated gene circuits. We first simulated the transient transfection data set according to our initial parameters, which we drew from previous experiences. After binning the data set according to the protein output from our transfection marker O_1, we determined the mode of the gene copy number k_1 within each bin. The other values are derived from k_1 according to their abundance coefficients. These values serve as the gene copy number for stably integrated gene circuits.

Data analysis

Pre-processing of experimental flow cytometry data. Retrieved data from BD LSR Fortessa was recorded with BD FACS Diva Software. The resulting files were exported in *.fcs* format and loaded into FlowJo software [52]. There, compensation of individual fluorescent channels was performed, live population gated and exported as scaled values into *.csv*-files.

Bi-exponential transformation. Scaled FACS values were transformed into bi-exponential space when needed via formulas from Parks et al. [53] with parameters $M = 4.5$, $p = 2$, $T = 262144$ and $W = 0.401$:

$$S(X; W) = T * 10^{-\Delta}\left(10^{\Delta} - p^2 * 10^{-\frac{\Delta}{p}} + p^2 - 1\right)$$

where $\Delta = X - W$ *for* $X \geq W$ and $\Delta = W - X$ *else*.

Gene copy number distributions k_i in output peaks. After binning the data set according to the transfection marker (50 bins in case of the co-transfection simulations, 25 bins in case of

the circuit simulations), we fit Gaussians to the log-transformed values. We slice a window of ±0.15 $\log_{10}$ units around the mode(s) of the fitted distribution. Within this narrow window, we repeat the process for the remaining genes (co-transfection case: 1. Cerulean, 2. Citrine, 3. mCherry, 4. iRFP; circuit case: 1. mCherry, 2. Cerulean or Citrine). Since all parameters needed for the simulations are stored in an array, we can select all cells within that final slice and look up the gene copy numbers that were used to generate this subset of output data. The distributions of the gene copy numbers are then processed to discover their modes.

Peak Finder Algorithm for Flow cytometry (PFAFF). The software is available on GitHub (https://github.com/benensonlab/PFAFF). The repository contains the code, detailed S4 Text "PFAFF User Manual", S5 Text "Description of the example data set" and sample simulated data for running the analysis. User-provided data can also be analyzed according to the steps described in User Manual.

The algorithm's procedure starts by discarding the tails of the transfection control's distribution. Within this window (i.e. 2.5–97.5% of transfection control fluorescence intensity) the distribution is segmented into bins of equal number of events (i.e. ten bins). Each bin is analyzed sequentially and all values are transformed bi-exponentially. The input distribution (i.e. mCherry) is approximated by a histogram in bi-exponential space and Gaussians are fitted to it. A following set of rules determines the number of fitted Gaussians:

```
1: program Fit Gaussians to mCherry Distribution
2:       if Goodness-of-Fit for one Gaussian > 0.975 then
3:         save mode value
4:         exit
5:       else
6:         fit two Gaussians
7:         if Goodness-of-Fit for two Gaussians > 0.99 then
8:         save mode values
9:         else
10:         fit three Gaussians
11:         if distance between two peaks < 0.42 then
12:         go back to use two Gaussian fit
13:         exit
14:         elseif distance (mean closest-modes) to (two-Gaussian-Fit
modes) < 0.3
15:         save two-Gaussian-Fit mode and remaining three-Gaussian-
Fit mode
16:         end if
17:         save mode values
18:         end if
19:         if distance between the two modes < 0.75 then
20:         go back to use one Gaussian fit
21:         exit
22:         end if
23:       end if
24: end program
```

A window of ±0.1 bi-exponential units is sliced around the peaks' center. Within that subset of cells, distributions of the output colors (i.e. Cerulean and Citrine) are again approximated by histograms. Much like before a set of rules determines the number of Gaussians that are fitted to these distributions:

```
1: program Fit Gaussians to Cerulean or Citrine Distribution
2:       if Goodness-of-Fit for one Gaussian > 0.975 then
3:         save mode value
4:         exit
5:       else
```

```
6:         fit two Gaussians
7:         if Goodness-of-Fit for two Gaussians > 0.995 then
8:         save mode values
9:         else
10:         fit three Gaussians
11:         if distance between peaks with highest intensities < 0.9
then
12:         remove remaining peak from the data set
13:         fit one Gaussian for the highest peak
14:         if Goodness-of-Fit > = Goodness-of-Fit for two Gaussians
then
15:         save mode values
16:         else
17:         save mode values of two Gaussian-Fit
18:         end if
19:         if distance between the two modes < 0.75 then
20:         go back to use one Gaussian fit
21:         exit
22:         end if
23:       end if
24: end program
```

For each bin, we repeat this fitting procedure. All extracted modes are re-transformed into flow cytometry units and stored in a table. The output of this algorithm is saved as MATLAB workspaces, that contain variables for generating (weighted) input/output mappings. Furthermore, various plots are generated (density plots of (raw) data, individual fits to data distributions, weighted input/output mappings and weighted mean input/output mappings) and saved as individual files in the result folder (see provided manual for details).

Supporting information

S1 Text. In-silico time-courses.
(DOCX)

S2 Text. Simple Fan-Out Model.
(DOCX)

S3 Text. Detailed Models.
(DOCX)

S4 Text. PFAFF User Manual.
(DOCX)

S5 Text. Description of the example data set.
(DOCX)

S1 Fig. Parameter drawing workflow.
(TIF)

S2 Fig. Co-transfection experiment and in silico simulations.
(TIF)

S3 Fig. In-silico simulations of a transient co-transfection with equimolar plasmid ratio.
(TIF)

S4 Fig. In-silico simulations of a transient co-transfection with nominal plasmid ratio.
(TIF)

S5 Fig. In-silico simulation of transient co-transfections at various initial gene copy number and parameter distributions.
(TIF)

S6 Fig. PFAFF applied to experimental data of a FO circuit.
(TIF)

S7 Fig. Comparison of multi-plasmid and single-plasmid gene circuits.
(TIF)

S8 Fig. In-silico simulations of a transiently transfected monomodal circuit.
(TIF)

S9 Fig. In-silico simulations of a transiently transfected bi-modal circuit.
(TIF)

S10 Fig. Gene copy number ratios from the simulations of transiently transfected bi-modal circuit.
(TIF)

S11 Fig. In silico simulations of stable integrations and transient transfections.
(TIF)

S12 Fig. Workflow for simulation and analysis of genetic circuits.
(TIF)

S13 Fig. Concatenation of high and low input and output modes.
(TIF)

S14 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.00.
(TIF)

S15 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.02.
(TIF)

S16 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.04.
(TIF)

S17 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.08.
(TIF)

S18 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.16.
(TIF)

S19 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (RIFFM and FO) at intrinsic noise level 0.32.
(TIF)

S20 Fig. PFAFF output for various bin numbers.
(TIF)

S21 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.00.
(TIF)

S22 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.02.
(TIF)

S23 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.04.
(TIF)

S24 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.08.
(TIF)

S25 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.16.
(TIF)

S26 Fig. In-silico simulation of stably integrated and transiently transfected (PFAFF input/output) circuits (I1-FFL1 and I1-FFL2) at intrinsic noise level 0.32.
(TIF)

S27 Fig. Results of PFAFF analysis on simulated cFFL flow cytometry data sets.
(TIF)

S28 Fig. Results of PFAFF analysis on simulated negFB flow cytometry data sets.
(TIF)

S29 Fig. Results of PFAFF analysis on simulated posFB flow cytometry data sets.
(TIF)

S30 Fig. Maps of plasmids used in co-transfection experiment.
(TIF)

S31 Fig. (Quasi) steady states of a fan-out circuit at different protein degradation rates.
(TIF)

S1 Table. Model parameter values for co-transfection simulations.
(DOCX)

S2 Table. Model parameter values for tested circuit architectures (RIFFM, I1-FFL1/2, FO, cFFL, negFB, posFB).
(DOCX)

S3 Table. List of plasmids.
(DOCX)

S4 Table. List of primers.
(DOCX)

S5 Table. Model parameter values for simple fan-out circuit.
(DOCX)

Acknowledgments

Benenson group members for discussions. We thank Bart Deplancke, Alexander Stark and Gerald Stampfel for providing the *D.melanogaster* gene *kni*.

Author Contributions

Conceptualization: Yaakov Benenson.

Data curation: Christoph Stelzer, Yaakov Benenson.

Formal analysis: Christoph Stelzer, Yaakov Benenson.

Funding acquisition: Yaakov Benenson.

Investigation: Christoph Stelzer, Yaakov Benenson.

Methodology: Christoph Stelzer, Yaakov Benenson.

Project administration: Yaakov Benenson.

Resources: Yaakov Benenson.

Software: Christoph Stelzer.

Supervision: Yaakov Benenson.

Validation: Christoph Stelzer, Yaakov Benenson.

Visualization: Christoph Stelzer, Yaakov Benenson.

Writing – original draft: Christoph Stelzer, Yaakov Benenson.

Writing – review & editing: Christoph Stelzer, Yaakov Benenson.

References

1. Benenson Y. Biomolecular computing systems: Principles, progress and potential. Nat Rev Genet. 2012; 13: 455–468. https://doi.org/10.1038/nrg3197 PMID: 22688678
2. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in Escherichia coli. Nature. 2000; 403: 339–42. https://doi.org/10.1038/35002131 PMID: 10659857
3. Nielsen AAKK, Der BS, Shin J, Vaidyanathan P, Paralanov V, Strychalski EA, et al. Genetic circuit design automation. Science (80-). 2016; 352: 53-+. https://doi.org/10.1126/science.aac7341 PMID: 27034378
4. Zhang Q, Bhattacharya S, Conolly RB, Clewell HJ, Kaminski NE, Andersen ME. Molecular signaling network motifs provide a mechanistic basis for cellular threshold responses. Environ Health Perspect. 2015; 122: 1261–1270. https://doi.org/10.1289/ehp.1408244 PMID: 25117432
5. Angelici B, Mailand E, Haefliger B, Benenson Y, Angelici B, Mailand E, et al. Synthetic Biology Platform for Sensing and Integrating Endogenous Transcriptional Inputs in Mammalian Cells Resource Synthetic Biology Platform for Sensing and Integrating Endogenous Transcriptional Inputs in Mammalian Cells. CellReports. 2016; 1–13. https://doi.org/10.1016/j.celrep.2016.07.061 PMID: 27545896
6. Bleris L, Xie Z, Glass D, Adadey A, Sontag E, Benenson Y. Synthetic incoherent feedforward circuits show adaptation to the amount of their genetic template. Mol Syst Biol. 2011; 7: 519. https://doi.org/10.1038/msb.2011.49 PMID: 21811230
7. Nevozhay D, Adams RM, Murphy KF, Josić K, Balázsi G. Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. Proc Natl Acad Sci U S A. 2009; 106: 5123–5128. https://doi.org/10.1073/pnas.0809901106 PMID: 19279212
8. Gregor T, Tank DW, Wieschaus EF, Bialek W. Probing the limits to positional information. Cell. 2007; 130: 153–164. https://doi.org/10.1016/j.cell.2007.05.025 PMID: 17632062
9. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science (80-). 2002; 297: 1183–1186. https://doi.org/10.1126/science.1070919 PMID: 12183631

10. Ozbudak EM, Thattal M, Lim HH, Shraiman BI, Van Oudenaarden A. Multistability in the lactose utilization network of Escherichia coli. Nature. 2004; 427: 737–740. https://doi.org/10.1038/nature02298 PMID: 14973486

11. Pedraza JH, Van Oudenaarden A. Noise propagations in gene networks. Science (80-). 2005; 307: 1965–1969. https://doi.org/10.1126/science.1109090 PMID: 15790857

12. Vigdal TJ, Kaufman CD, Izsvák Z, Voytas DF, Ivics Z. Common physical properties of DNA affecting target site selection of Sleeping Beauty and other Tc1/mariner transposable elements. J Mol Biol. 2002; 323: 441–452. https://doi.org/10.1016/s0022-2836(02)00991-9 PMID: 12381300

13. Wilson MH, Coates CJ, George AL. PiggyBac transposon-mediated gene transfer in human cells. Mol Ther. 2007; 15: 139–145. https://doi.org/10.1038/sj.mt.6300028 PMID: 17164785

14. Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, Hannenhalli S, et al. Genome-wide analysis of retroviral DNA integration. Nat Rev Microbiol. 2005; 3: 848–858. https://doi.org/10.1038/nrmicro1263 PMID: 16175173

15. Tratschin JD, Miller IL, Smith MG, Carter BJ. Adeno-associated virus vector for high-frequency integration, expression, and rescue of genes in mammalian cells. Mol Cell Biol. 1985; 5: 3251–3260. https://doi.org/10.1128/mcb.5.11.3251 PMID: 3018511

16. Gersbach CA, Gaj T, Gordley RM, Mercer AC, Barbas CF. Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. Nucleic Acids Res. 2011; 39: 7868–7878. https://doi.org/10.1093/nar/gkr421 PMID: 21653554

17. Hockemeyer D, Wang H, Kiani S, Lai CS, Gao Q, Cassady JP, et al. Genetic engineering of human pluripotent cells using TALE nucleases. Nat Biotechnol. 2011; 29: 731–734. https://doi.org/10.1038/nbt.1927 PMID: 21738127

18. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science (80-). 2013; 339: 823–826. https://doi.org/10.1126/science.1232033 PMID: 23287722

19. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science (80-). 2013; 339: 819–823. https://doi.org/10.1126/science.1231143 PMID: 23287718

20. Haefliger B, Prochazka L, Angelici B, Benenson Y. Precision multidimensional assay for high-throughput microRNA drug discovery. Nat Commun. 2016; 7. https://doi.org/10.1038/ncomms10709 PMID: 26880188

21. Prochazka L, Angelici B, Haefliger B, Benenson Y. Highly modular bow-tie gene circuits with programmable dynamic behaviour. Nat Commun. 2014; 5. https://doi.org/10.1038/ncomms5729 PMID: 25311543

22. Shimoga V, White JT, Li Y, Sontag E, Bleris L. Synthetic mammalian transgene negative autoregulation. Mol Syst Biol. 2013; 9. https://doi.org/10.1038/msb.2013.27 PMID: 23736683

23. Xie Z, Wroblewska L, Prochazka L, Weiss R, Benenson Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. Science (80-). 2011; 333: 1307–1311. https://doi.org/10.1126/science.1205527 PMID: 21885784

24. Lapique N, Benenson Y. Digital switching in a biosensor circuit via programmable timing of gene availability. Nat Chem Biol. 2014; 10: 1020–1027. https://doi.org/10.1038/nchembio.1680 PMID: 25306443

25. Recillas-Targa F. Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals. Molecular Biotechnology. 2006. pp. 337–354. https://doi.org/10.1385/MB:34:3:337 PMID: 17284781

26. Schreiber J, Arter M, Lapique N, Haefliger B, Benenson Y. Model-guided combinatorial optimization of complex synthetic gene networks. Mol Syst Biol. 2016; 12: 899. https://doi.org/10.15252/msb.20167265 PMID: 28031353

27. Davidsohn N, Beal J, Kiani S, Adler A, Yaman F, Li Y, et al. Accurate Predictions of Genetic Circuit Behavior from Part Characterization and Modular Composition. ACS Synth Biol. 2015; 4: 673–681. https://doi.org/10.1021/sb500263b PMID: 25369267

28. Stanton BC, Siciliano V, Ghodasara A, Wroblewska L, Clancy K, Trefzer AC, et al. Systematic transfer of prokaryotic sensors and circuits to mammalian cells. ACS Synth Biol. 2014; 3: 880–891. https://doi.org/10.1021/sb5002856 PMID: 25360681

29. Wang J, Isaacson SA, Belta C. Modeling Genetic Circuit Behavior in Transiently Transfected Mammalian Cells. ACS Synth Biol. 2019. https://doi.org/10.1021/acssynbio.8b00166 PMID: 30884948

30. Munteanu A, Cotterell J, Solé R V., Sharpe J. Design principles of stripe-forming motifs: The role of positive feedback. Sci Rep. 2014; 4. https://doi.org/10.1038/srep05003 PMID: 24830352

31. Schaerli Y, Munteanu A, Gili M, Cotterell J, Sharpe J, Isalan M. A unified design space of synthetic stripe-forming networks. Nat Commun. 2014; 5: 4905. https://doi.org/10.1038/ncomms5905 PMID: 25247316

32. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer D V. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell. 2005; 122: 169–182. https://doi.org/10.1016/j.cell.2005.06.006 PMID: 16051143

33. To TL, Maheshri N. Noise can induce bimodality in positive transcriptional feedback loops without bistability. Science (80-). 2010; 327: 1142–1145. https://doi.org/10.1126/science.1178962 PMID: 20185727

34. Ellis EL, Delbrück M. The growth of bacteriophage. J Gen Physiol. 1939; 22: 365–384. https://doi.org/10.1085/jgp.22.3.365 PMID: 19873108

35. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science (80-). 2010; 329: 533–538. https://doi.org/10.1126/science.1188308 PMID: 20671182

36. Beal J. Biochemical complexity drives log-normal variation in genetic expression. Eng Biol. 2017; 1: 55–60. https://doi.org/10.1049/enb.2017.0004

37. Mclean APF, Smolke CD, Salit M. Characterizing the Non-Normal Distribution of Flow Cytometry Measurements from Transiently Expressed Constructs in Mammalian Cells. 2016; 1–15. https://doi.org/10.1101/057950

38. Lillacci G, Benenson Y, Khammash M. Synthetic control systems for high performance gene expression in mammalian cells. Nucleic Acids Res. 2018; 46: 9855–9863. https://doi.org/10.1093/nar/gky795 PMID: 30203050

39. Gao XJ, Chong LS, Kim MS, Elowitz MB. Programmable protein circuits in living cells. Science (80-). 2018; 361: 1252–1258. https://doi.org/10.1126/science.aat5062 PMID: 30237357

40. Widder S, Schicho J, Schuster P. Dynamic patterns of gene regulation I: Simple two-gene systems. J Theor Biol. 2007; 246: 395–419. https://doi.org/10.1016/j.jtbi.2007.01.004 PMID: 17337276

41. Huang S, Guo YP, May G, Enver T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev Biol. 2007; 305: 695–713. https://doi.org/10.1016/j.ydbio.2007.02.036 PMID: 17412320

42. Fussenegger M, Morris RP, Fux C, Rimann M, Von Stockar B, Thompson CJ, et al. Streptogramin-based gene regulation systems for mammalian cells. Nat Biotechnol. 2000; 18: 1203–1208. https://doi.org/10.1038/81208 PMID: 11062442

43. Kim HD, O'Shea EK. A quantitative model of transcription factor-activated gene expression. Nat Struct Mol Biol. 2008; 15: 1192–1198. https://doi.org/10.1038/nsmb.1500 PMID: 18849996

44. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene Regulation at the Single-Cell Level. 2013; 1962: 1–5. https://doi.org/10.1126/science.1106914 PMID: 15790856

45. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res. 2005; 15: 1388–1392. https://doi.org/10.1101/gr.3820805 PMID: 16204192

46. Kremers GJ, Goedhart J, Van Den Heuvel DJ, Gerritsen HC, Gadella TWJ. Improved green and blue fluorescent proteins for expression in bacteria and mammalian cells. Biochemistry. 2007; 46: 3775–3783. https://doi.org/10.1021/bi0622874 PMID: 17323929

47. Rizzo MA, Springer GH, Granada B, Piston DW. An improved cyan fluorescent protein variant useful for FRET. Nat Biotechnol. 2004; 22: 445–449. https://doi.org/10.1038/nbt945 PMID: 14990965

48. Griesbeck O, Baird GS, Campbell RE, Zacharias DA, Tsien RY. Reducing the environmental sensitivity of yellow fluorescent protein. Mechanism and applications. J Biol Chem. 2001; 276: 29188–29194. https://doi.org/10.1074/jbc.M102815200 PMID: 11387331

49. Shaner NC, Campbell RE, Steinbach PA, Giepmans BNGG, Palmer AE, Tsien RY. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nat Biotechnol. 2004; 22: 1567–1572. https://doi.org/10.1038/nbt1037 PMID: 15558047

50. Filonov GS, Piatkevich KD, Ting LM, Zhang J, Kim K, Verkhusha V V. Bright and stable near-infrared fluorescent protein for in vivo imaging. Nat Biotechnol. 2011; 29: 757–761. https://doi.org/10.1038/nbt.1918 PMID: 21765402

51. Bostrom K, Wettesten M, Boren J, Bondjers G, Wiklund O, Olofsson SO. Pulse-chase studies of the synthesis and intracellular transport of apolipoprotein B-100 in Hep G2 cells. J Biol Chem. 1986; 261: 13800–13806. PMID: 3020051

52. Vallan C. Flow Cytometric Data Analysis with Flowjo. Cytom Part A. 2009; 75a: 720.

53. Parks DR, Roederer M, Moore WA. A new "logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytom Part A. 2006; 69: 541–551. https://doi.org/10.1002/cyto.a.20258 PMID: 16604519