



Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-008

April 7, 2012

A Method for Fast, High-Precision Characterization of Synthetic Biology Devices

Jacob Beal, Ron Weiss, Fusun Yaman, Noah
Davidsohn, and Aaron Adler

A Method for Fast, High-Precision Characterization of Synthetic Biology Devices

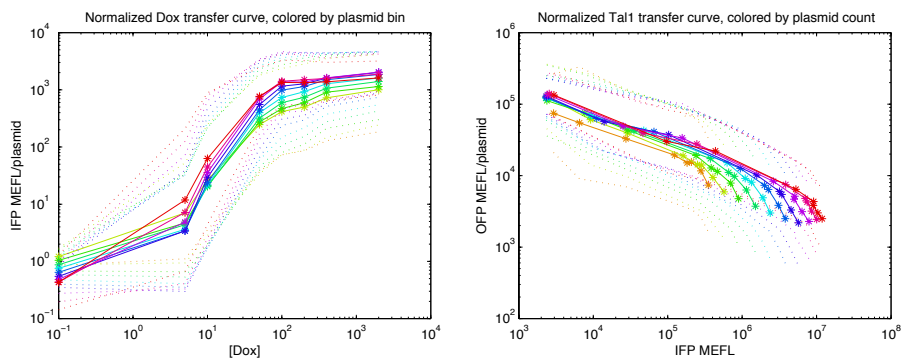
Jacob Beal¹, Ron Weiss², Fusun Yaman¹, Noah Davidsohn²,
and Aaron Adler¹

¹Raytheon BBN Technologies and ²MIT

April 6, 2012

Abstract

Engineering biological systems with predictable behavior is a foundational goal of synthetic biology. To accomplish this, it is important to accurately characterize the behavior of biological devices. Prior characterization efforts, however, have generally not yielded enough high-quality information to enable compositional design. In the TASBE (A Tool-Chain to Accelerate Synthetic Biological Engineering) project we have developed a new characterization technique capable of producing such data. This document describes the techniques we have developed, along with examples of their application, so that the techniques can be accurately used by others.



Work partially sponsored by DARPA; the views and conclusions contained in this document are those of the authors and not DARPA or the U.S. Government.

Contents

1	Motivation and Overview	3
1.1	Requirements for DNA Part Characterization	3
1.2	Comparison with Prior Techniques	5
2	Method	5
2.1	Constructs	6
2.2	Fluorescent Protein Selection	7
2.3	Experimental Protocols	8
2.4	Analysis	9
2.4.1	Compensation for Autofluorescence and Spectral Overlap	10
2.4.2	Segmentation into Bins	10
2.4.3	Computation of Bin Statistics	11
2.4.4	Normalization to Obtain Per-Plasmid Behavior	12
2.5	TASBE Analysis Service	16
3	Best Practices	16
4	Usage Example	17
4.1	Fluorescent Protein Selection Example	17
4.2	Example Construct	21
4.3	Example Analysis	21

1 Motivation and Overview

DNA part characterization has foundational significance in the field of synthetic biology. From the very inception of the field, the vision has been one of standardized parts that allow for predictable design, after the fashion of electronic parts (e.g., TTL part data sheets). There have been two great obstacles in realizing this vision:

1. obtaining accurate measurements of relevant chemical properties within individual cells, and
2. predicting (with a certainty) the behavior of a DNA component when it is used in a novel design.

The second obstacle is predicated on our ability to make progress on the first; it is not possible to predict what behaviors should be measured from a novel system if it is not possible to obtain accurate measurements of the reference system needed to make the prediction.

During the TASBE (A Tool-Chain to Accelerate Synthetic Biological Engineering) project, just as many prior researchers have, we had difficulty acquiring high-precision characterization data. Without sufficient part characterization, the ability to use parts in different contexts is greatly diminished, causing each additional use of an existing part to frequently involve long, difficult, and costly debugging and laboratory experimentation.

The TASBE project's tight coupling between wet-lab work and high-level design tools has provided us with a clear set of requirements for DNA part characterization. Through wet-lab experimentation, we determined that prior methods were insufficient for the requirements of predictive design. This document describes the TASBE characterization process (Figure 1) that enables the construction of a library of biological computing devices with input/output relation characterization sufficient to enable predictive design. The remainder of this section describes the requirements for successful DNA part characterization (Section 1.1) and a comparison to prior methods (Section 1.2). Section 2 describes the TASBE characterization method. A set of best practices are described in Section 3, and an example is provided in Section 4.

1.1 Requirements for DNA Part Characterization

Our work in TASBE has shown us that, with regards to DNA part characterization, any type of predictive design will need at least:

- Large numbers of single-cell measurements (as opposed to population average values),
- Measurements of the level of part output signal(s) across the full dynamic range of levels of part input signal(s),
- Data to determine the per-copy effect of the construct, and

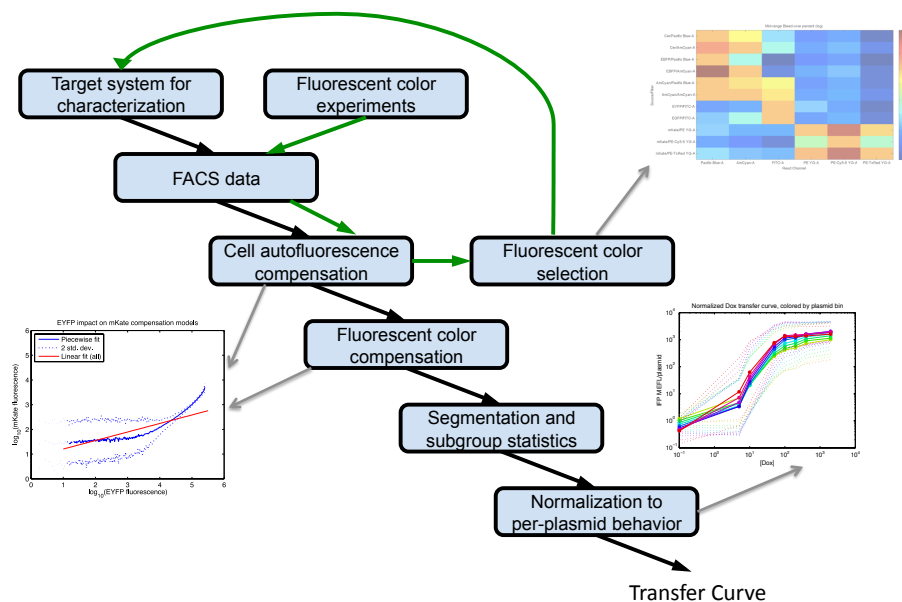


Figure 1: A visual summary of the steps in the TASBE characterization process. Details of these steps can be found in Section 2.

- The statistical distribution of single-cell output levels for each input level, in order to estimate the variability of behavior.

The reason for these requirements is that much of the interaction between parts takes place within individual cells, rather than averaged across the whole population (except for special case systems involving intercellular communication). Thus, we need to measure the behavior of single cells. Cells exhibit a high degree of behavioral variability within a population, so to control for variability we acquire a large number of single-cell measurements.

Finally, when evaluating DNA parts in a digital logic context, it is typically required for parts to have an approximately sigmoidal response with well-defined high and low input and output signal levels. Different DNA parts have different output signal levels and make their high-to-low transitions at different input signal levels. Thus predictable composition requires knowledge of the full input/output relation, particularly when working with parts whose transition between low and high expression has a slope that is not very steep.

Additionally we need to know the per-copy (per-plasmid) behavior of the constructs in order to predict overall signal levels, since the number of copies of a system in a cell can vary greatly both by design (e.g., multiple copies, plasmids with different mean copy numbers) and through natural cell dynamics (e.g., copy number variation, transient transfection).

1.2 Comparison with Prior Techniques

Although others have recognized this same set of requirements, many prior efforts at characterization, not driven by a tight coupling with high-level design efforts, have gathered data that is insufficient for predictive design. For example, the variable strength library of promoters generated by the Collins lab [3] were only characterized for high and low expression level. The BioBricks specification sheet produced by Canton et al. [1] gathered data across the full dynamic range, but the transfer curve reports only population averages (the standard deviations shown are differences in population mean), as does the ongoing BIOFAB project [2]. Elowitz et al. [5] successfully predicted a circuit result, but the result was for an integrated, feedback circuit; it remains unclear how well the result generalizes. As far as we can tell, the data sheets produced by Imperial CSynBI [4] do not have the needed full-range transfer curves. Prior work by Weiss [6] does not calculate the transfer functions on a per-copy basis. None of these previous efforts satisfied all four requirements above, and thus cannot produce the kind of characterization data that is necessary for predictable part composition.

2 Method

The TASBE characterization process is based on Fluorescence Activated Cell Sorting (FACS), a standard flow cytometry technique that allows large numbers of single-cell fluorescence measurements to be obtained quickly. These single cell measurements capture variation in device behavior, a necessity when composing devices within a cell.

The TASBE characterization process produces high-precision per-plasmid behavior data by adding three new elements:

- A constitutive fluorescent protein that allows measurement of the number of copies of a system in the cell (introduced by the high variability in number of transfected or transformed plasmids),¹
- Fluorescent protein/read channel screening to ensure $< 1\%$ bleed-over² between colors, and
- Multi-dimensional data segmentation that greatly increases the signal-to-noise ratio in systems with a variable plasmid count.

The aim of this section is to describe our characterization process in sufficient detail so that it may be implemented in any laboratory for any organism. To date, the method has been completely validated using mammalian cells and partially validated using *E. coli*.

¹Estimating plasmid count from constitutive fluorescent protein is not new, but using it as a part of characterization to obtain an input/output transfer curve is new.

²Because of spectral overlap of the read channels a color might be picked up in multiple channels. Bleed-over occurs when existence of one color is picked up on a channel that is not designated as the read channel of that color.

2.1 Constructs

Characterization of a device is done with a system composed of the following constructs:

- The device itself.
- A construct for externally inducing expression of each input to the device that is not directly controlled by experimental conditions (e.g., transcription factors).
- An *Input Fluorescent Protein* (IFP) for each input to the device that is not directly controllable. The IFP expression must be directly controlled by the same construct (e.g., promoter) that is controlling the input. This will be used to measure the input signal levels in each cell.
- An *Output Fluorescent Protein* (OFP) for each output of the device, whose expression is directly controlled by that output. This will be used to measure the output signal levels in each cell.
- A *Constitutive Fluorescent Protein* (CFP) that is expressed constitutively at a high level. This will be used to measure the number of copies of the system contained in each cell.

Figure 2 gives examples of how a characterization system can be instantiated, for a small-molecule sensor, repressor, and hybrid promoter.

In some organisms, it is possible to introduce multiple species of plasmid at closely correlated numbers (e.g., cotransfection via lipofection in mammalian cells). For such organisms, it is recommended that initially each functional unit

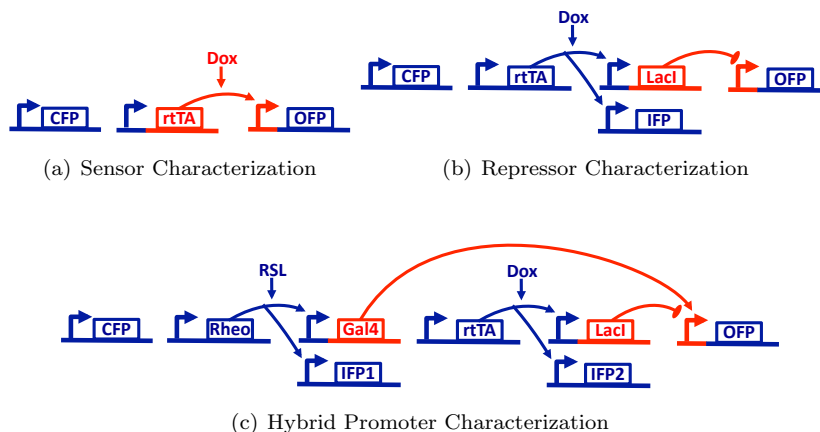


Figure 2: A characterization system embeds the device to be characterized (red) with supporting constructs (blue) that control an input, measure input and output, and measure the number of copies of the system.

in the system be placed on its own plasmid, and the plasmids be cotransfected or cotransformed. Thus, for example, a repressor characterization would be a 5-plasmid system (in Figure 2(b), CFP, rtTA, LacI, IFP, and OFP would be on separate plasmids).

A key advantage of separation onto multiple plasmids is eliminating the problems that may arise from larger constructs, such as interaction between adjacent functional units or low transfection yield. Note that if the plasmids can reproduce in the cell line (e.g., bacterial cells), it is important for them to have different origins of replication so that their relative copy counts do not drift.

2.2 Fluorescent Protein Selection

It is critically important to have extremely low spectral overlap when performing characterization. The reason is that, for any characterization, there will be some conditions where some proteins are highly expressed and others will have very low expression. When the spectral overlap between fluorescent proteins leads to more than about 1% bleed-over of signal from one channel to another, even well-calibrated fluorescence compensation is often inadequate to produce good quantitative data.

In selecting fluorescent proteins for IFP, OFP, and CFP, therefore, it is necessary to begin by screening for a low-bleed-over combination of proteins and laser/filter pairs (“channels”) used for FACS measurements.

To achieve this goal we have outlined a seven step procedure below. We first construct a bleed-over matrix (Steps 1-6) and then pick the best set of colors (Step 7) using this matrix.

1. Select a set of candidate proteins and a set of candidate channels for each protein.
2. For each protein under consideration, create a plasmid with only high constitutive expression of that protein.
3. Culture one set of cells for each protein, transfected/transformed with only the plasmid for that protein, along with a negative control (for computing autofluorescence). Ideally the negative control should be cells that have been transfected with a mock (non-expressing) plasmid, but cells that have not been transfected may be an acceptable substitute.
4. When cells are strongly expressing fluorescence, measure fluorescence via FACS on all channels under consideration.
5. For each channel c_i , compute autofluorescence mean $a_{\mu,i}$ and standard deviation $a_{\sigma,i}$. These should be arithmetic, since noise here is expected to be dominated by additive sensor noise.
6. For each protein / channel combination (p, c_i) , compute the bleed-over b_{ij} with each other channel c_j as follows:

- (a) Let m_x be the measured fluorescence on channel c_x .
 - (b) Select only those cells with an m_i far from both autofluorescence and saturation (e.g., in the $10^{3.5}$ to $10^{4.5}$ range).
 - (c) The estimated bleed-over b_{ij} is the arithmetic mean m_j/m_i for this subset of cells.
7. Select a set of protein / channel combinations (p, c_i) such that:
- Mean fluorescence of p on c_i is at least two decades above $a_{\mu,i} + 2 \cdot a_{\sigma,i}$, and
 - $b_{ij} < 0.01$ for all other channels in the set.

Following this procedure should allow for the selection of a set of IFP, OFP, and CFP proteins and FACS settings that will allow good quantitative characterization data to be gathered even when some expression is near autofluorescence levels. In mammalian cells, we have found the following combination to be effective:

- EBFP2, measured with a 405 nm laser and a 450/50 filter,
- EYFP, measured with a 488 nm laser and a 530/30 filter, and
- mKate, measured with a 561 nm laser and a 610/20 filter,

with iRFP likely to provide a good fourth fluorescent protein.

2.3 Experimental Protocols

Experimental protocol will mainly be determined by the cell strain being used as the chassis for the device being characterized. Moreover, there is not yet a “standard” set of conditions under which devices are expected to be characterized.

Our basic protocol framework is:

1. Introduce characterization system plasmid or plasmids into competent cells. If using multiple species of plasmid, use a protocol that results in correlated plasmid numbers (e.g., lipofection, but not nucleofection).
2. Culture cells in your preferred conditions.
3. Simultaneously, culture under the same conditions controls for the unmodified strain and for each fluorescent protein (e.g., a 3 color characterization system has a total of 4 controls). The fluorescent protein controls should contain nothing but a constitutive expression of that protein, using the same promoter used for CFP in the characterization system.
4. To measure expression dynamics:
 - (a) Select an evenly distributed set of sampling times, based on expected expression dynamics, as specified below in Best Practices (Section 3).

- (b) Cultivate six colonies of cells for each time at which you plan to sample: three at high induction, three without induction. Cells should not be reused after FACS measurement.
 - (c) Acquire FACS data from three sets of high-induction cells and three sets of non-induced cells at each sampling time.
5. To measure an input/output transfer curve:
- (a) Select a set of induction levels which are somewhat uniformly distributed on the log-scale, as specified below in Best Practices (Section 3).
 - (b) Cultivate three colonies of cells at each induction level, changing media as necessary to ensure cells are kept healthy and consistently induced.
 - (c) Take FACS data when OFP is expected to near its maximum or minimum, whichever takes longer, as determined by expression dynamics experiments.

FACS data should be taken uncompensated: FACS software typically does not do a good job handling multi-laser compensation or compensation for autofluorescence, so compensation will be performed afterwards during analysis. FACS data should, however, be filtered to exclude events likely from debris rather than cells, but should not be filtered to exclude non-functional cells (that will be done later). Also run fluorescent beads for calibration, e.g., Spherotek rainbow calibration particles, to ensure that your FACS units will be able to be converted to MEFLs during later processing.

Based on our own experiences, we recommend the following for mammalian cells from the HEK 293 FT cell line:

- Culture in DMEM medium (CellGro), supplemented with 10% FBS (PAA Laboratories), 2mM L-Glutamine (CellGro), 100x Strep/pen (CellGro), 100x Non-Essential amino acids (NEAA) (HyClone), and 10,000x Fungin (Invivogen).
- Passage cells with 0.05% Trypsin.
- Cotransfect plasmids via lipofection. Optionally select for plasmid using a resistance marker (e.g., with 2ug/ml of puromycin (Invivogen) for 2-4 days or until control cells that did not contain puromycin resistance are dead.)
- Take FACS data at 72 hours, after resuspending cells in the appropriate media, such as 1xPBS that does not contain calcium or magnesium.

2.4 Analysis

Once FACS data has been acquired, it can be analyzed in order to extract per-plasmid transfer curves relating device input level to output expression level.

These can then be further processed as desired to extract other properties, such as Hill equation models or chemical kinetics.

There are four stages to data analysis:

1. Compensate for autofluorescence and spectral overlap.
2. Segment data into *bins* by both induction and CFP.
3. Compute statistics of each data points in each bin.
4. Normalize by CFP to obtain per-plasmid behavior.

2.4.1 Compensation for Autofluorescence and Spectral Overlap

The first stage of data processing is to map observed fluorescence levels to MEFLs. This is a three-stage process for each sample:

- Subtract mean autofluorescence $a_{\mu,i}$ from each channel's measurement m_i .
- Each measurement consists of $m_i = f_i + \sum_{j \neq i} b_{ji} f_j$, where f_i is the fluorescence from the protein that channel i is intended to measure and b_{ji} is the bleed-over from channel j to channel i . Solve simultaneous equations to obtain the set of f_i .
- Translate f_i from relative fluorescence to MEFLs using standard FACS bead-based calibration.

Note that many compensation methods (including those typically built into FACS machines) will not perform complete compensation of this sort, typically either compensating only within a particular laser or omitting autofluorescence compensation. It is also not safe to assume that a fluorescent protein will emit only on adjacent channels (many have surprising outlying lobes), or that a protein will emit precisely how the manufacturer's spectrum indicates that it will emit (as the cellular context may be sufficiently different). These distortions may have a major effect on your measured values, particularly for low fluorescence values.

2.4.2 Segmentation into Bins

The next step is to segment data into bins, simultaneously by induction conditions and logarithmic CFP intervals, and to throw away likely invalid data:

- Discard all samples with CFP below $2 \cdot \sigma$, where σ is the standard deviation of autofluorescence on the CFP channel. These are expected to be cells that failed to receive plasmids or where the number of plasmids is excessively low.

- Discard all samples with any fluorescence measure below 0, as they will cause problems with geometric statistics later.³
- Discard all samples with CFP fluorescence above f_{max} , set to be less than the expected saturation point for the FACS detectors.
- For a logarithmic partition, $\pi_C = \{[a, b]\}$, place all samples with $a < CFP \leq b$ into bin $[a, b]$. Suggested bin width is $\log_{10}b/a \leq 0.25$.
- Discard all bins with less than a minimum m samples (suggested $m = 100$), as being statistically invalid.

Bins are logarithmic because expression noise is expected to be mainly multiplicative rather than additive. Binning is used because of the expected cell-to-cell variation in number of copies of the system; the differences can be caused by natural variation in copy number or transient transfection. As noted above, good characterization data requires knowing the per-plasmid behavior of the system. The observed behavior depends on the number of copies of the interacting parts (Figure 3), because the concentrations of chemicals produced from different copies add together. Consider the typical case of a single promoter/regulator interaction: when there are more copies of the promoter and the gene it regulates (a fluorescent protein in our case), then the same concentration of regulator will produce more fluorescence. Likewise, when there are more copies of the regulator gene, it takes less inducer to produce the same concentration of regulatory protein. What this means is that cells with different numbers of copies of a circuit may have radically different behaviors under the same conditions. An example of where this effect shows up strongly is lipofection of mammalian cells, where the number of copies of the plasmid that enter each cell typically range over two to three orders of magnitude.

Due to this variation in observed behavior, it is important to segment data not just by induction condition, but also by number of copies of the system, as indicated by CFP. The relation between observed CFP expression and number of copies is not, however, linear, as will be explained in detail below in Section 2.4.4.

2.4.3 Computation of Bin Statistics

Within each bin, compute the **geometric** mean and standard deviation for each fluorescent value for all data points within that bin. Geometric mean and standard deviation can be taken by computing the ordinary mean and standard deviation on a log scale:

$$\begin{aligned}\mu_g(x_i) &= e^{\mu(\ln x_i)} \\ \sigma_g(x_i) &= e^{\sigma(\ln x_i)}\end{aligned}$$

³Note that this will cause some good data to be discarded and some distortion in measurements near the autofluorescence floor. Improved estimators are under development, aimed towards a simultaneous joint estimator of autofluorescence, spectral overlap, and expression variance.

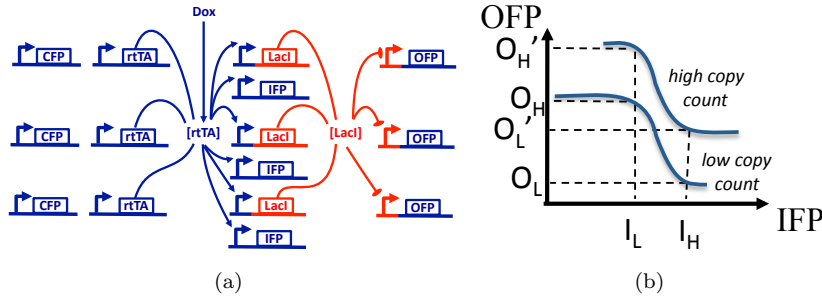


Figure 3: When there are multiple copies of a characterization system, the chemical concentrations superpose (a). This affects the observed device input and output behavior differently (b). For the input, more copies means that less inducer is needed to get the same input concentration. On the output, the response to a given input concentration is multiplied by the number of copies.

It is important to use geometric rather than arithmetic statistics because expression noise of fluorescent proteins is typically multiplicative rather than additive. Using arithmetic means will result in statistics that are skewed upwards by high outliers.

Remember that to compute error ranges, the geometric mean is multiplied and divided by the geometric standard deviation, not added and subtracted. For example, a range of two standard deviations is:

$$[\mu_g / \sigma_g^2, \mu_g \sigma_g^2]$$

2.4.4 Normalization to Obtain Per-Plasmid Behavior

Transforming bin statistics into per-plasmid behavior takes two steps:

- Estimate mean number of plasmids from mean CFP.
- Divide mean OFP by estimated mean plasmids.

Note that the standard deviations and the IFP mean are *not* transformed. The standard deviations are not transformed because they are multiplicative and not additive; an increased number of plasmids may actually tighten the observed standard deviation. The IFP mean is not transformed because the devices respond to the chemical concentration created by the combined production of all inputs, not the per-plasmid production of inputs.

Estimating Plasmid Count The expression of MEFLs from a single plasmid can be determined in a number of different ways. For example, in mammalian cells with transient transfection, plasmids do not replicate and so, as the cells divide, the number of plasmids will decrease, until eventually all cells have either

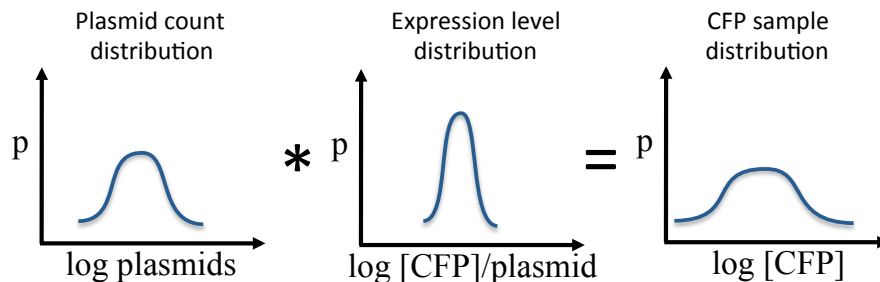


Figure 4: The distribution of observed CFP samples (CFP) is the convolution of the distribution of plasmid count (L) by the distribution of expression levels (V). Since these are typically both gaussian, it means that samples in any given bin are biased toward the mean level of CFP.

one or zero plasmids, at which point expression can be measured.^{4,5} Let the mean expression for a reference constitutive promoter (e.g., the CAG promoter in mammalian cells) from a single plasmid be called E .

Estimating the number of plasmids from the CFP in a bin will use this measure, but also must compensate for sampling bias caused by the underlying plasmid count distribution. A naïve estimate of the mean number of plasmids in a bin can be taken simply by taking the mean CFP from samples in the bin and dividing by E MEFLs/plasmid. This gives a reasonable estimate for bins near the mean overall CFP level, but as bins move away from the center, the estimate will be more and more incorrect.

The reason for this is that the distribution of CFP is the *convolution* of the distribution of plasmid count and the distribution of expression levels (Figure 4), and expression noise is typically larger than the range of a bin. When both of these distributions are gaussian in shape (as is often the case) and the variation in expression levels is relatively independent for each fluorescent protein, this means that in any given bin the samples are biased toward cells closer to the mean number of plasmids, simply because there are so many more of those cells. This is illustrated in Figure 5.

Next we will discuss the procedure that will correct the sampling bias effect. We use the notation:

μ the geometric mean of a distribution

σ^2 the geometric variance of a distribution (σ is the standard deviation of the distribution)

Going back to Figure 4 we need the mean and variance for the three distributions, L : plasmid counts, V : Expression level and CFP : Observed CFP. Some

⁴Typically we do not wait for this point to collect data, but in order to determine the expression from a single plasmid we need to collect data when the cells only have one plasmid.

⁵Assuming appropriate degradation tags or computational models are used to cancel out the inherited levels.

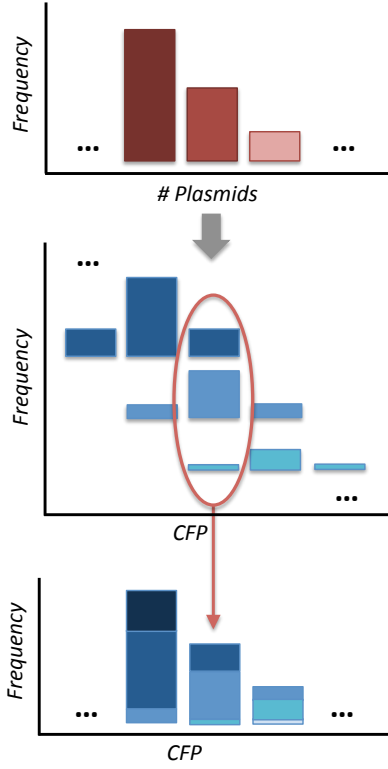


Figure 5: The frequency of different plasmid counts varies (top). The expression of CFP from these plasmid counts is noisy and larger than the bin range (middle). Thus the observed frequencies of CFP levels are biased toward cells closer to the mean number of plasmids.

of these statistics are computed from the data and others need to be estimated:

- The geometric mean and variance of CFP, μ_{CFP} and σ_{CFP}^2 , are directly computed from the data.
- The geometric mean expression level, μ_V , is E .
- We can estimate the geometric variance of expression level, σ_V^2 , by computing the observed (non-CFP) expression geometric variance in each bin.
- The geometric mean plasmid count is estimated as: $\mu_L = \mu_{CFP}/\mu_V$ (using the convolution theorem).
- The geometric variance of plasmid count is estimated as: $\sigma_L = \left| \sqrt{\sigma_{CFP}^2 - \sigma_V^2} \right|$.

Once we have the three distributions, L , V and CFP , we can apply the Bayes Theorem to compute the expected plasmid count in a cell given an observed CFP

level. We choose to operate on the discrete level for these computations thus we partition the space of plasmid counts and observed CFP levels into bins. Let π_L be a partition of the space of plasmid counts into bins $[x, y]$. Then

- Probability of a cell expressing CFP in the range $[a, b]$ given that it contains plasmids in the range $[x, y]$ may be approximated for small ranges $[x, y]$ by $p(CFP \in [a, b] | L \in [x, y]) = \int_a^b N(\mu_V \cdot (x + y)/2, \sigma_V)$, where N is a geometric normal distribution.
- Probability of a cell containing plasmids in the range $[x, y]$ given that it has CFP levels in the range $[a, b]$ is estimated by:

$$p(L \in [x, y] | CFP \in [a, b]) = \frac{p(CFP \in [a, b] | L \in [x, y]) \cdot p(L \in [x, y])}{\sum_{[x, y] \in \pi_L} p(CFP \in [a, b] | L \in [x, y]) \cdot p(L \in [x, y])}$$

- Expected number of plasmids in a cell given that it has CFP levels in the range $[a, b]$ is:

$$E(L | CFP \in [a, b]) = \sum_{[x, y] \in \pi_L} \mu([x, y]) \cdot p(L \in [x, y] | CFP \in [a, b])$$

Details of the derivation for estimates For readers who are interested in the mathematical derivation of the estimates, the details are as follows: Given the distributions for V and CFP , we can calculate the underlying plasmid distribution (μ_L, σ_L) using the convolution theorem adapted for geometric mean and variance:

$$\begin{aligned} CFP &= V * CFP \\ \mu_{CFP} &= \mu_V * \mu_L \\ \mu_L &= \mu_{CFP} / \mu_V \\ \sigma_{CFP}^2 &= \sigma_V^2 + \sigma_L^2 \\ \sigma_L &= \left| \sqrt{\sigma_{CFP}^2 - \sigma_V^2} \right| \end{aligned}$$

Letting π_L be a partition of the space of plasmid counts into bins $[x, y]$, the distribution of plasmids contributing to each CFP bin $[a, b]$ can be estimated using Bayes' law:

$$p(L \in [x, y] | CFP \in [a, b]) = \frac{p(CFP \in [a, b] | L \in [x, y]) \cdot p(L \in [x, y])}{p(CFP \in [a, b])}$$

$$p(L \in [x, y] | CFP \in [a, b]) = \frac{p(CFP \in [a, b] | L \in [x, y]) \cdot p(L \in [x, y])}{\sum_{[x, y] \in \pi_L} p(CFP \in [a, b] | L \in [x, y]) \cdot p(L \in [x, y])}$$

where the $p(CFP \in [a, b] | L \in [x, y])$ term is taken from the observed expression variation model and the $p(L \in [x, y])$ term is taken from the plasmid distribution model.

Finally, the mapping from mean CFP level to mean plasmid count can be produced by taking the expectation of plasmid counts:

$$E(L|CFP \in [a, b]) = \sum_{[x,y] \in \pi_L} \mu([x, y]) \cdot p(L \in [x, y]|CFP \in [a, b])$$

2.5 TASBE Analysis Service

The TASBE Analysis Service is a collection of free web-based software intended to make using this characterization method simple, by providing implementations of all of the necessary data processing. The TASBE Analysis Service always implements the most updated versions of all the analysis protocols described above. As an online tool, the service will be continually upgraded as estimators, compensation techniques, filters, etc. continue to improve. This software also provides data curation tools to allow users to manage their data and update analyses as the techniques continue to improve. The software will be available starting April 2012 from <http://synthetic-biology.bbn.com/>

See also Section 4 below.

3 Best Practices

For ensuring that results can be useful and trustworthy, we recommend the following best practices:

- All constructs other than the device being characterized should be in the same direction on the same strand of DNA.
- Fluorescent proteins should have < 1% bleed-over.
- Controls should include fluorescent beads for FACS calibration, the strain with a mock (non-expressing) transfection (for autofluorescence), and constitutive expression of each fluorescent protein—one control per protein (for fluorescence calibration). An unmodified strain may be an acceptable alternative for measuring autofluorescence.
- Consistent experimental protocols, e.g., time, laser power, amount of DNA, should be used for all devices intended to be interconnected.
- Induction levels should be evenly logarithmically distributed, with at least three induction levels per decade, across at least three decades (e.g., 1 nM, 2 nM, 5 nM, 10 nM, 20 nM, 50 nM, 100 nM, 200 nM, 500 nM, 1000 nM). There should also be a non-induced case (e.g., 0 nM), except for sensors where this is not possible (e.g., induction by temperature).
- Time sequence data for expression dynamics should be timed based on the time for approximately full expression of the constitutive protein (T_C). They should be evenly distributed, both before and after T_C , including

at least ten samples, starting no later than $0.25T_C$ and ending no earlier than $3T_C$.

- Fluorescence values should be taken from at least 10,000 cells per induction level with a recommended cell count of 50,000 per induction level (after gating to remove likely non-cell FACS events).
- Each bin should have a constitutive fluorescence range of at most a quarter decade.
- Each bin used should contain at least 100 data points.
- Results should present the per-cell mean and standard deviation of each bin. Mean and standard deviation should be geometric.
- Experiments should be conducted in triplicate. This should be used to compute the mean and variance of both the per-cell means, the per-cell standard deviations, and the per-experiment variation.
- Plasmid sequences should be verified, preferably by sequencing, though restriction digest is an acceptable substitute when the sequences of component DNA parts have previously been verified. Constructs should not be made using error prone processes such as PCR; recommended assembly processes use restriction cloning or homologous recombination. Examples of recommended processes include BioBricks and Gateway/Gibson.
- Experiment records should include at least: strain of cells, DNA sequence of plasmid constructs (preferably encoded with SBOL⁶), the protocol used for introducing plasmids to cells, culturing conditions, and number of hours cultured before measurement.

4 Usage Example

4.1 Fluorescent Protein Selection Example

As described in Section 2.2, a first critical step is to determine which fluorescent colors are good candidates to use in circuits on a given FACS machine. This example uses a BD LSR II FACS machine with the following laser/filter combinations (“channels”):

Pacific Blue-A 405 nm laser with 450/50 filter,

AmCyan-A 405 nm laser with 510/50 filter,

FITC-A 488 nm laser with 530/50 filter,

PE YG-A 561 nm laser with 575/26 filter,

⁶<http://www.sbolstandard.org/>

PE-Cy5-5 YG-A 561 nm laser with 695/40 filter, and

PE-TxRed YG-A 561nm laser with 610/20 filter.

Cells cultures, each constitutively producing a different fluorescent protein of interest, are processed with the FACS machine. This example examines the following fluorescent proteins: Cerulean, EBFP2, AmCyan, EYFP, EGFP, and mKate.

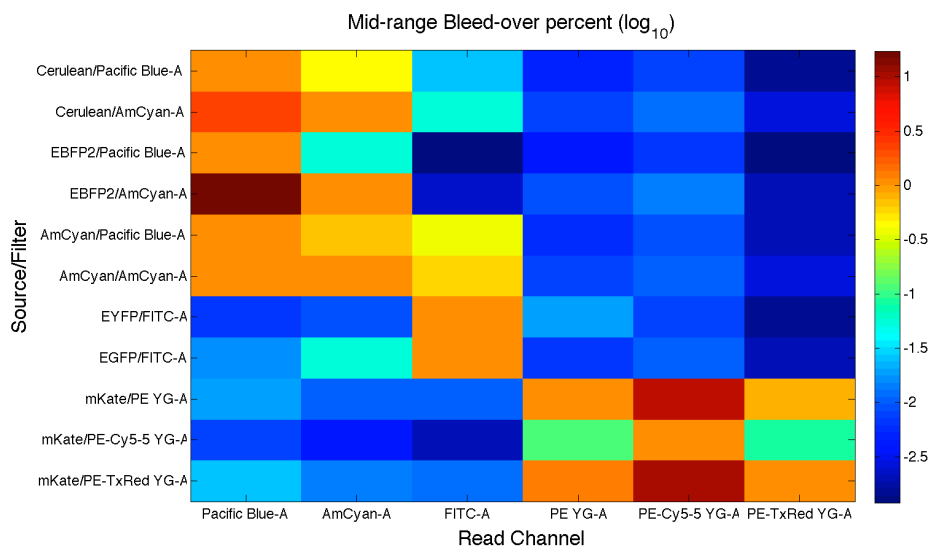


Figure 6: The heat map representing possible fluorescent proteins and their spectral overlap on the available read channels. Blue colors indicate less overlap while reds indicate significant overlap.

Figure 6 shows the resulting FACS data represented as a heat map. This can be used to identify good combinations of colors to use in characterization constructs. Each row of the heat map is a fluorescent protein and a target channel, for example the first row is the Cerulean fluorescent protein measured using the Pacific Blue-A channel. Each column in the heat map represents an available read channel on the FACS machine. As we want to minimize the spectral overlap between colors, the heat map plots the mid-range bleed-over percent between colors on a \log_{10} scale. Further, as shown in Figure 7, the bleed-over for the target channel is 100% or 1, which in \log_{10} is 0. Figure 8 highlights the EBFP2/AmCyan-A row. Here the AmCyan-A channel has a zero value. EBFP2 has high bleed-over on the Pacific Blue-A channel (dark red) but low bleed-over into the FITC-A, PE YG-A, and PE-TxRed YG-A channels (darker blues, $< 1\%$ bleed-over), and moderate bleed-over into the PE-Cy5-5 YG-A channel (lighter blue, $\sim 3\%$ bleed-over).

We are looking for three pairs of fluorescent proteins and channels that produce unique hot values on the read channels. An example of a bad combination

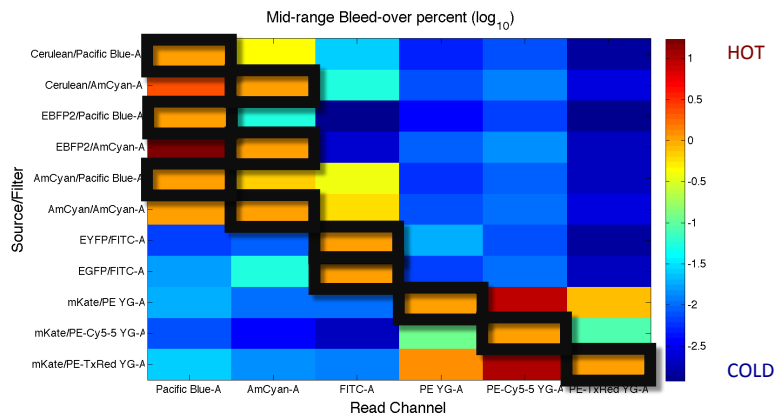


Figure 7: The values in each row are zero for the target channel (bleed-over is 100% or 1, and $\log_{10}1 = 0$).

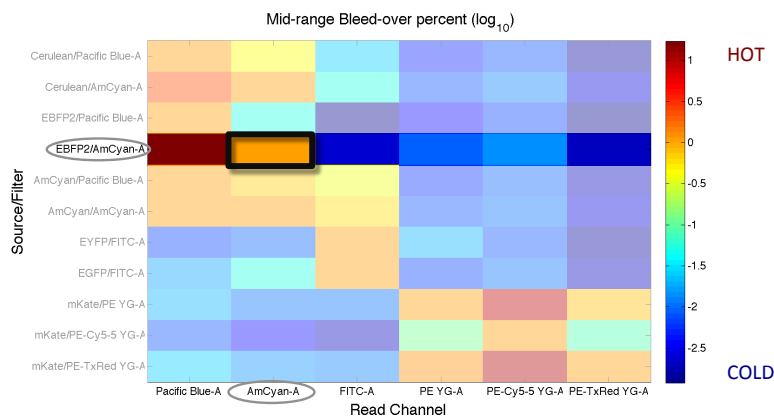


Figure 8: One possible source/channel combination is highlighted.

of fluorescent colors is shown in Figure 9. In this case, Cerulean and EBFP2 both produce high readings on the Pacific Blue-A and AmCyan-A channels. It would be impossible to separate the measurements on these channels into Cerulean and EBFP2 values.

Figure 10 shows the fluorescent proteins and channels we selected: 1) EBFP2 with the Pacific Blue-A channel, 2) EYFP with the FITC-A channel, and 3) mKate with the PE-TxRed YG-A channel. While any of the PE YG-A, PE-Cy5-5 YG-A, and PE-TxRed YG-A channels give high readings for mKate and low readings for EYFP and EBFP2, the selected channel gives a high reading for mKate and sufficiently low readings for the other colors. Our highest bleed-over is from mKate into the Pacific Blue-A channel ($\sim 2.3\%$) and FITC-A channel

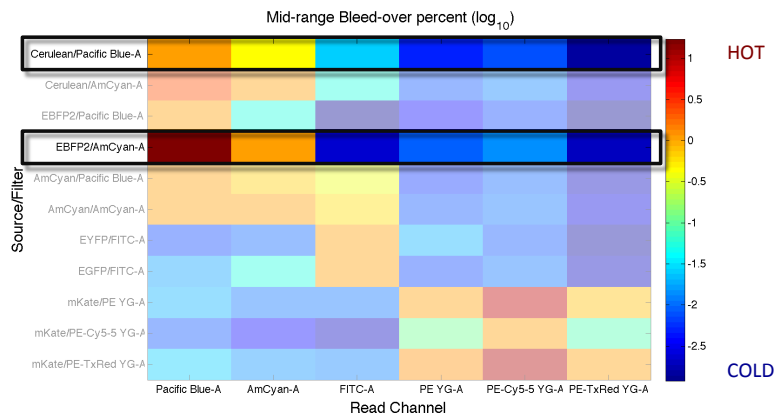


Figure 9: Two colors that would be a poor choice for use in the same characterization experiment. Cerulean and EBFP2 both have high readings on the Pacific Blue-A and AmCyan-A channels making them impossible to adequately separate.

(~1.3%). All of the other bleed-overs are well under 1%.

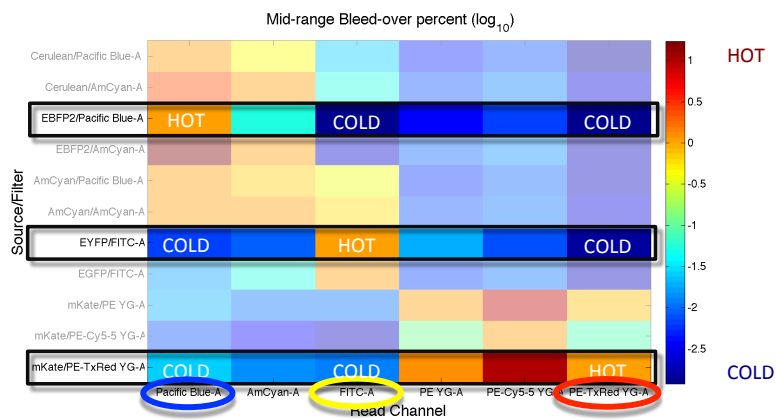


Figure 10: The fluorescent colors selected; the spectra of these colors do not overlap significantly making it possible to cleanly separate the data.

4.2 Example Construct

Note that this example does not conform to all best practices – do what you can with the data and equipment you have. Figure 11 shows the circuit construct used for characterization for the LacI repressor. In this example, mKate is the CFP, EBFP2 is the IFP, and EYFP is the OFP. The fluorescent color levels are measured using the “Pacific Blue-A,” “FITC-A,” and “PE-TxRed YG-A” channels.

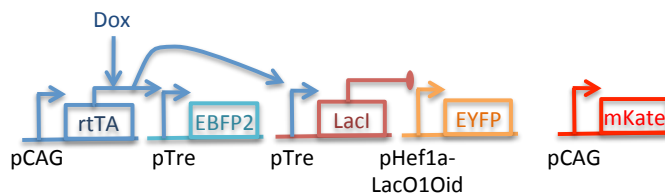


Figure 11: The circuit used to characterize LacI.

4.3 Example Analysis

Here we provide an example of the results produced by the TASBE Analysis Service, as applied to a characterization circuit for the Tal1 repressor, induced by doxycycline (Dox). More detailed supplemental material, including user manuals, will be available online at (<http://synthetic-biology.bbn.com/>) following release.

The service performs the following steps, as described in Section 2.4:

1. Compensate for autofluorescence and spectral overlap.
2. Segment data into bins by both induction and CFP.
3. Compute statistics of data points in each bin.
4. Convert setting-dependent FACS units to standardized MEFL units.
5. Estimate mean plasmid count per bin based on observed fluorescence distributions.
6. Normalize by plasmid count to obtain per-plasmid behavior.
7. Output the resulting data and graphs.

Typical previous results would have examined the mean fluorescence of all the cells for each induction level, failing to distinguish between behaviors for different plasmid counts. The dual segmentation of the TASBE method separates these, typically providing the following benefits:

- Tighter estimates of variation, since only comparable cells are being averaged together.

- A wider range of behavior revealed in transfer curves.
- Internal cross-validation of models by comparison of behavior in cells with different numbers of plasmids—either validating existing models or revealing new phenomena.

First we examine the doxycycline (Dox) transfer curve. Population means produce the curve shown in Figure 12(a), while the TASBE method produces the normalized curve shown in Figure 12(c). Figure 12(b) shows the non-normalized curve for better comparison with the population mean method.

Notice that with the TASBE method, it becomes clear that the value of the first Dox point at 10^{-1} is being driven primarily by autofluorescence rather than being actual leakage expression, as the non-normalized curve shows all different plasmid expression levels coming to a point. This leads to population averaging underestimating the high/low differential of the Dox curve by at least 5x, while the TASBE method finds the greater differential and clearly indicates that further experiments are necessary to properly quantify the low range of expression.

The averaging of unlike cells together also causes population averaging to overestimate cell-to-cell variation of behavior. Population averaging shows that at high levels of Dox induction, the standard deviation of cells is approximately 4x, while the TASBE method shows that it actually averages closer to 2x.

The difference in the transfer curve of the Tal1 repressor is even more marked, since both the input and output are affected by plasmid count. Population means produce the curve shown in Figure 13(a), while the TASBE method produces the normalized curve shown in Figure 13(c) and Figure 13(b) shows the non-normalized curve.

With population averaging, Tal1 appears to have wildly different input and output ranges, while the TASBE method shows that per-plasmid the ranges are actually quite similar. The high/low differential on the output is similar, but there is 5x more range in the input, as we would expect from the prior set of graphs. More importantly, however, the TASBE method's segmentation reveals that the behavior of the system is not uniform, but depends significantly on the plasmid count. The revealed phenomenon is highly structured, indicating systematic variation which should be tractable to model.

As before, expression variance is refined as well, in some cases tightening and in others actually revealing greater variation.

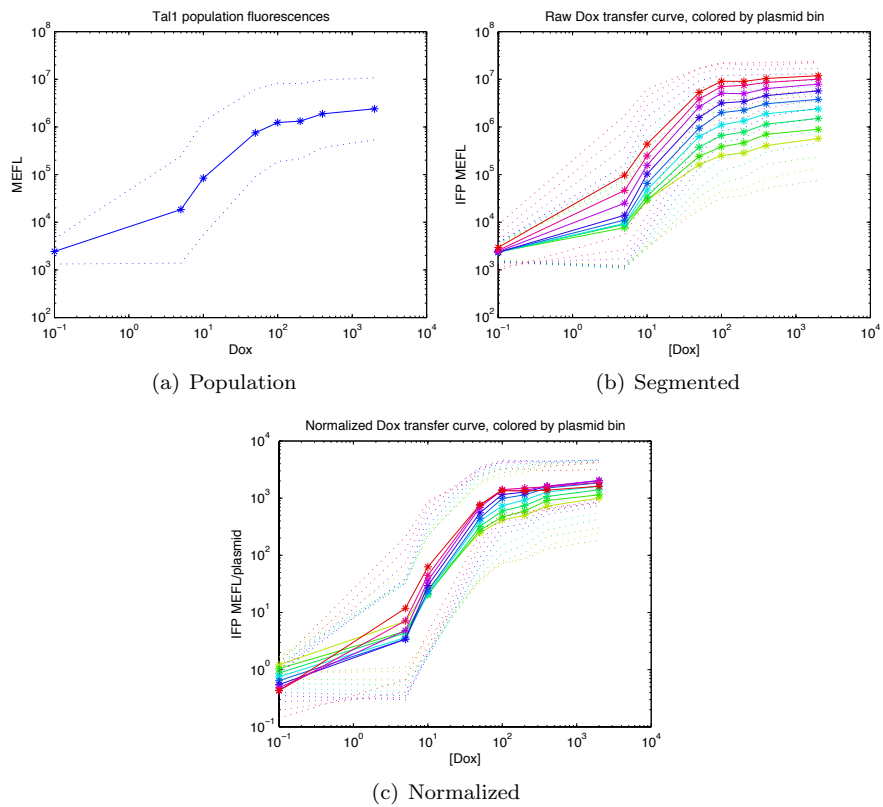


Figure 12: Tal1 population fluorescence data (a) vs. TASBE method segmented (b) and normalized (c) transfer curves. Notice the increase the steeper curve and larger difference in the expression magnitude.

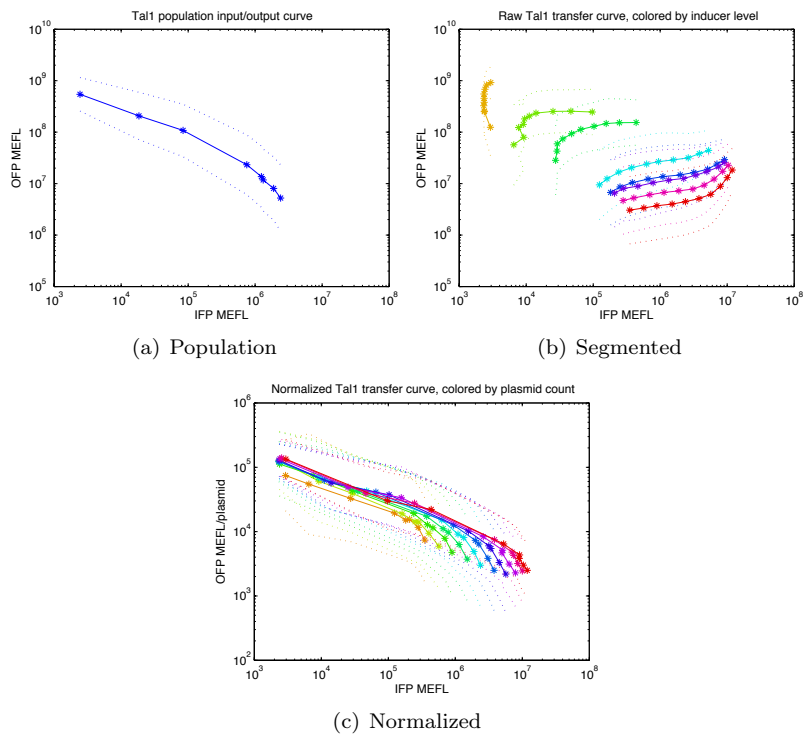


Figure 13: Tal1 population fluorescence data (a) vs. TASBE method segmented (b) and normalized (c) transfer curves. The population curve has 5x repression while the normalized curve has about 200x repression.

References

- [1] B. Canton, A. Labno, and D. Endy. Refinement and standardization of synthetic biological parts and devices. *Nature Biotechnology*, 26:787–93, July 2008.
- [2] Directors: Drew Endy, Adam Arkin. BIOFAB: International open facility advancing biotechnology. <http://biofab.org/>, Retrieved Jan 16th, 2012.
- [3] Tom Ellis, Xiao Wang, and James Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat Biotech*, 27(5):465–471, May 2009.
- [4] Centre for Synthetic Biology and Innovation (CSynBI). Part characterisation. <http://www3.imperial.ac.uk/syntheticbiology/parts>, January 2012.
- [5] Nitzan Rosenfeld, Jonathan W. Young, Uri Alon, Peter S. Swain, and Michael B. Elowitz. Accurate prediction of gene feedback circuit behavior from component properties. *Molecular systems biology*, 3(1), November 2007.
- [6] Ron Weiss. *Cellular Computation and Communications using Engineered Genetic Regulatory Networks*. PhD thesis, MIT, 2001.

