DISSERTATION

DATA ANALYSIS AND PREDICTIVE MODELING FOR SYNTHETIC AND NATURALLY OCCURRING BIOLOGICAL SWITCHES

Submitted by

Katherine A. Schaumberg

Graduate Degree Program in Bioengineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2016

Doctoral Committee:

Advisor: Ashok Prasad
Co-Advisor: June Medford

Patrick Shipman
Mauricio Antunes
Diego Krapf

ABSTRACT

DATA ANALYSIS AND PREDICTIVE MODELING FOR SYNTHETIC AND

NATURALLY OCCURRING BIOLOGICAL SWITCHES

Biological switches are biochemical network motifs responsible for determining the chemical state of cells, and are a key part of every biological system. The impact of these biological switches on cell behavior is broad. For example, many diseases such as cancer are thought to be caused by a misregulation of the bio-chemical state in a cell or group of cells. Also cell fates in differentiating stem cells are controlled by biological switches. Because of their general importance the synthetic biology community has also constructed synthetic biological switches in living organisms. While there are different kinds of possible switches, in my thesis I study switches capable of stably generating two unique molecular states, also called bi-stable switches. Here these switches are studied from two perspectives. In Chapters 1-4 I present theoretical and experimental work on analysis of specific circuits that act like biological switches. In Chapter 5 I employ a data mining perspective to identify gene expression signatures of switches that are sensitive to cytotoxic cancer drugs.

This dissertation starts with a computational analysis of the effect of leaky promoter expression on bi-stable biological switches. In several biological and synthetic systems gene transcription is never completely off, even when repressed. This residual expression is referred to here as leaky expression. Bi-stable systems would be expected to have some amount of leaky expression in their off state. However, the impact of leaky expression on the functioning and properties of biological switches has not been well studied. To help fill this gap we conducted a theoretical analysis of leaky expression's effect on biological switches. Two switches, a positive feedback and negative inhibition-based switch were

studied. We found that the different circuit topologies showed different advantages in terms of their ability to handle leaky expression.

Next this dissertation describes work done in collaboration with the Medford lab at Colorado State University, to construct and characterize a library of genetic plant parts. These parts would later be used in construction of perhaps the first synthetic bi-stable toggle switch in a plant. As part of this study, experiments were designed and conducted for finding the nature of the experimental noise associated with the assays used to test these plant parts. A mathematical normalization was developed to estimate quantitative information on the performance of each part. Validation experiments were done to assess the usefulness of this method for predicting the behavior of stably transformed plants from higher throughput transient assays. In the end a library of over one hundred quantitatively characterized plant parts in both Arabidopsis and Sorghum was constructed. The quantitative parameters of this library of genetic parts were then used in combination with a probabilistic bootstrap method we developed to predict optimal part combinations for construction of a bi-stable switch in *Arabidopsis*.

The dissertation concludes with a study of biological networks in cancer cells from a data mining perspective. A large amount of data exists in the public domain on the sensitivity of cancer cell lines to cytotoxic drugs. Some cancers appear to be in a “sensitive state” while others are in a “resistant state”. We would like to be able to know the gene expression signatures of these two states in order to predict cancer drug sensitivity from gene expression data. As a first step towards this goal we assessed the repeatability of predictions between the two standard databases of cancer cell lines, the NCI60 and the GDSC. This lead to identification of a preprocessing method needed to combine data from multiple databases. This was then followed up with the development of a comparative analysis platform. This platform was used to test the accuracy of models designed to predict drug sensitivity, when different model construction methods were used.

ACKNOWLEDGEMENTS

I have been blessed to have such wonderful people helping me along my PhD journey. First I would like to thank my Adviser Ashok Prasad for his insight and supportive feedback without which I would not be where I am at today. I would also like to thank my committee: June Medford my co-adviser, Patrick Shipman, Mauricio Antunes and Diego Krapf for their support, advice and guidance. Each one of them contributed unique and invaluable perspectives which I hope to take with me as I grow in my scientific career.

I would like to thank my colleagues and fellow students. I would not have learned nearly as much without the great questions and helpful ideas that came up in our discussions. I would especially like to thank Elaheh Alizadeh, Wenlong Xu, Chintan Joshi and Samy Lyons.

Chapter three of this dissertation represents joint work with other co-authors. I would like to thank my co-authors, as well as D. McCarthy for valuable help in preparing the manuscript. Financial support from US Department of Energy, Advanced Research Projects Agency-Energy 2012 Grant No. DE-AR0000311 and US Department of Defense, Defense Threat Reduction Agency Grant No. W911NF-09-10526 is gratefully acknowledged.

For chapter four I would like to thank the Gustafson group for creating a supportive environment without which I could not have done the work presented in this chapter. I would also like to thank them for the edits to the chapter and invaluable feedback on the project. I would like to acknowledge the financial support from a Colorado State University Cancer Supercluster Seed Grant.

I would like to thank Emily Kiwimagi, Gary Kiwimagi and Allison Humphries for their help in editing portions of this dissertation. Finally, I would also like thank my husband Chris Schaumberg for being supportive of my scientific endeavors. I find with him, my life and research are more meaningful.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2

Chapter 4

Chapter 5

Appendix B

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix B

CHAPTER 1

Introduction

1.1 Mathematical Models can lead to Discoveries in Biology

Computational and mathematical methods are transforming biology and modern medicine. Mathematical ideas and computational methods developed and documented from the 1800s to the present are being applied to biological problems today. This has and is leading to wide variety of new discoveries in biological regulatory systems. New insights into methods for treatment of diseases are being discovered with the aid of mathematical models. In the field of synthetic biology, predictions of biological circuit function *in silico* has and is being implemented in labs across the world [1] [2] [3] [4].

In one striking example of a new discovery using computational and mathematical methods, the Barkai group discovered a previously unknown molecular relationship involved in eye development. This molecular relationship in the Drosophila eye was found after building a model attempting to describe molecular interactions in eye development [1]. This model was unable to reproduce the molecular patterns observed in the developing eye. This led to asking about which additional relationship would lead to the observed pattern formation. After identifying the needed relationship, experimental studies were conducted to test whether it actually did exist, and the experiments discovered the previously unknown molecular relationship [1].

Another example can be found in the development of the three drug treatment of human immunodeficiency virus, HIV [5]. A model of HIV infection was constructed, that helped

in uncovering viral dynamics during the long dormancy period between infection and acquisition of AIDS [5]. From an approximation of the virus behavior in the human body, different methods of drug treatment were explored. In the end, the study helped to propose and establish the aggressive early three drug treatment currently in use for the control of HIV [5].

Methods for cancer treatment have also been explored via computational and mathematical modeling [6] [7]. Mechanisms of different cancers have been mathematically modeled to gain further insight into the process [8] [9]. Large data repositories have been mined using computational methods to pull out differences between tumor types and identify molecules and mutations of interest in these complex biological systems [6] [10]. Pathway vulnerabilities have been discovered and are being investigated for the purpose of developing new treatments for cancer [10].

In the field of synthetic biology, mathematical models have been developed based on mechanistic information of the underlying biological processes. These models and the accompanying quantification of the biological processes, have given us new insights into basic gene regulatory processes. Mathematical models can illuminate parts of the biological process that would not have been apparent in its absence. One example of this is the constraint that a finite number of ribosomes places on translational activity. This was modeled quantitatively using computational methods developed for queuing theory of job processing on central processing units, CPUs [11] [12]. Another example is found in the identification of what is called "gene bursting" [13] [14], or bursty gene transcription. This was modeled using stochastic differential equations describing the differences in transcription and translation rates [14].

1.2 The Law of Mass Action and Hill Equations

One of the early historical developments in chemistry is still used today in the application of computational and mathematical methods to biochemical phenomena. This is the Law of Mass Action (LMA), which was documented in 1862 when a chemist and mathematician published the first paper describing it. The LMA is a mathematical rule that describes the progression of a chemical reaction over time under certain ideal conditions [15]. The LMA has been used to model genetic systems of biochemical molecules [15]. It has even been used as a starting point for modeling more complex relationships. These more complex models have led to other rate laws such as the Michaelis-Menten rate law [15].

LMA can be paraphrased as: a chemical reaction's product concentration changes proportionally to the product of the concentration of its reactants, raised to the power of their stoichiometries. This law assumes a well-mixed system with high molecular numbers. With this law whole systems of chemical reactions can be translated into sets of ordinary differential equations, ODEs. A variety of software has been developed for automating the generation of the ODE systems from biochemical reaction information, such as BioNetGen, SimBiology toolbox for MATLAB and Copasi [16] [17] [18]. ODEs have been numerically approximated to predict *in silico* the behavior of complex biological systems for complex signaling pathways such as those found in cancer [8]. Also, particular parameters from the Michaelis-Menten rate law have become a way to report the efficiency of various enzymes [15].

Hill equations, named after Archibald Vivian Hill, were first used by Hill to describe a drug's effect on biological responses in 1910 [19]. This equation has grown immensely in its applications. Hill equations can model the input-output response of systems whose behavior is

either hyperbolic or sigmodal. Examples of such systems include: cooperative binding of transcription factors [20], the cooperative effect of the lacI regulation of transcription [21] and hemoglobin's transport of oxygen [22]. It is common place in the mathematical description of biochemical systems that the LMA and Hill equations are used together to describe the system. It has also been shown experimentally that both LMA and Hill equations serve as good mathematical models for the appropriate processes when compared to experimental data [2] [23].

1.3 Need for Quantitative Parameters

Using LMA, Hill functions and other rate laws such as Michaelis-Menten kinetics, we can write systems of ordinary differential equations (ODEs) that describe the time progression of the concentrations of all the various species in the system. However, making predictions using these ODEs requires knowledge about the parameters used in the equations. For reactions governed by LMA, these involve estimates of the on and off rates, or at least the equilibrium constants. For Hill equations or Michaelis-Menten rate laws, this involves estimation of the rate law from experiment. Often times ODEs describing biological systems have unknown parameters, and experiments to find such parameters can be challenging. If the general behavior of the system does not change within the range of possible parameter values, knowing the exact parameters will not change the predictions from the system. However, sometimes a small change in a parameter can make a noticeable difference in the system. An example of this can be found when an ODE system undergoes a bifurcation. A bifurcation happens when fixed points within the system change their stability and/or their number. One example is when a system goes from being mono-stable to bi-stable. A mono-stable system has

one stable fixed point, which means no matter the history of the system the steady state value will always be the same. In contrast, a bi-stable system has two stable fixed points allowing the system to have one of two steady state values depending on the history of the system. Bi-stable systems also show history dependent behavior, called hysteresis. When working with biological systems capable of being bi-stable or mono-stable within the range of acceptable parameters, knowing quantitative information about the parameters, or system wide stability properties, is an important part of modeling these biological systems.

1.4 Synergy between Computation and Experiment

As quantitative experiments become more sophisticated in biology, there is an increasing use of mathematical modeling to help understand and even predict biological phenomena. There are emerging interdisciplinary fields of study that have incorporated computational prediction with experimental verification to both understand and engineer the state of biological organisms. Many of these fields are adapting engineering approaches for work flow and designing of biological systems. For example, Synthetic Biology has used the idea of building quantitatively characterized modular molecular parts, analogous to parts used to build electronic microprocessor such as transistors or resistors. The molecular parts can then be assembled into functional units much like many smaller pieces are assembled to make circuit boards. This type of design could then be taken further to build more and more complex systems. Construction of electronic devices such as microprocessors is often optimized through a design process involving computation simulation and prediction of performance, followed by construction and testing of new designs leading back to more computation predictions. The same workflow of quantitative measurement for the characterization of genetic parts, followed

by computational design and experiments for testing and validation, further followed by improved computational design and improved experimental implementation, can also be applied to designing biological systems. This idea of using the synergy between computational models to predict the behavior of novel synthetic designs has driven the construction of libraries such as BioBricks, providing quantitatively characterized genetic parts useful for building more complex circuits [24]. These synthetically designed molecular parts have been used to build and investigate properties of biological systems [24].

1.5 Biological Switch Properties

Biological switches, as defined here, are genetic circuits that control the molecular state of a cell. Synthetically designed parts are often used to toggle the concentration of a molecule of interest between a low and high state. Biological switches have been built in the past and are found throughout nature both in eukaryotic and prokaryotic organisms. Two examples of naturally occurring switches can be found in the lactose system and sonic hedgehog-based systems [25] [21]. These switches play key roles in many everyday cell functions. A few particular single-cell-based-switches have been built synthetically [2] [23] [26]. It can be said that at the minimum a switch requires a threshold separating two states. We can call these the off and the on states. If the states are distinguished by low and high concentrations of a protein X, then a sharply sigmoidal input-output curve for protein X could be a kind of switch. One way to generate this kind of sigmoidality, called ultrasensitivity, requires oppositely acting phosphorylation and dephosphorylation loops [3]. Other processes in biology, especially involving cooperativity, can also give rise to ultrasensitive responses [22] [27] [28]. However, these types of circuits suffer from the drawback that the system could settle into intermediate

states, i.e. neither off nor on. A more complicated switch is one that can exist in only two stable and well-separated states. Such switches are called bi-stable, and they can be made in many ways, usually involving positive and negative inhibition loops [25] [21] [29] [30].

One consequence of bi-stable switches is the observation that over a cell population this can lead to a bimodal response, *i.e.* distinct populations of cells that are either off or on. Bimodal distributions (i.e. two peaks in the distribution) have been observed in both experimental (*e.g.* fluorescent cell population in flow cytometry experiment) and computational (*e.g.* equilibrium probability distribution in a stochastic system) experiments [2] [23] [21] [31]. However, it should be noted that bi-stability is not the only way a bi-modal response can be achieved. Cooperative binding between regulator proteins can also produce a bi-modal response under some circumstances [20].

It has been postulated that for all known naturally occurring switches, bi-stability requires cooperativity [2]. However, cooperativity is not always needed for bi-stability. Using synthetically designed parts, a non-cooperative sequestration system was built and has been shown to create a bi-stable system [2]. Another characteristic of biological switches is in how they affect the noise within the systems biological process of interest. In particular, positive and negative inhibition loops have been shown to magnify or diminish this intrinsic noise [4] [32] [33]. Intrinsic noise in biological switches can also play an important role in organism fitness [34] [35].

1.6 Gaps in our Knowledge and How this Dissertation Addresses These Gaps

A major part of this dissertation is focused on the design of feedback-based bi-stable systems. During the practical, experimental implementation and testing of the designs, we

realized that many promoters used in synthetic biology applications, or found in nature, were naturally “leaky”, *i.e.* showed some low level of expression. However, at that time there was no theoretical analysis of how leaky expression affected the important practical properties of biological switches. I carried out this theoretical analysis, which forms Chapter 2 of this thesis. This work is still unpublished, but was presented as a poster in the q-bio conference in August 2014 [36].

While bi-stable biological switches have been built in the past [2] [23] [26] they are confined to single cells. In particular, no quantitatively characterized bi-stable synthetic gene circuits have been constructed in plants due to the challenges posed by plant biology. Chapters 3 & 4 presents our work, in a joint collaboration between the Prasad group and the Medford group, to characterize genetic parts in plants. Development of more quantitatively characterized parts useful for assembly of larger system in plants has a plethora of applications. One could imagine allowing plants to reach a set biomass before turning on a synthetically designed switch to start creating biofuels as suggested in this perspective on plant synthetic biology [37]. Also, control over the embryonic state of plants for plant transformation could be achieved through controlled expression of a master regulator (*e.g.* morphogen) [38]. Plants have always been an important part of our world from food to moving carbon dioxide to oxygen. Gaining predictable or even tunable control over plant function will impact the way we live.

For our work in the world of plants described in Chapter 3, we had to overcome several obstacles. First, we needed to work with plants such as Arabidopsis and Sorghum that take weeks to months to complete their life cycle. Protoplasts, plant cells without their cell wall, can

be transiently transformed in an assay taking two days, helping us overcome this challenge of time. However, using transiently transformed protoplasts also increased the amount of experimental noise. This noise was around 2 orders of magnitude. Also, the data was collected in relative luciferase units, whose properties can change from lab to lab. The experimental noise and non-physical units made the task of creating a library of characterized plant parts for construction of a bi-stable switch a challenge. We met this challenge by running additional wet lab experiments to move our system to physical units of molecule number as well as to tease out the nature of the experimental noise. We then created a mathematical model to describe how the noise was affecting our system. This led to the development of a normalization scheme that decreased the noise enough to find statistically significant differences between the different parts in the library within and across two plant species. As the end-circuit will be stably transformed into plants, the last challenge was to see if the transiently transformed protoplasts would give a reasonable approximation of the genetic part behavior observed in stable transformed plants then isolated to form protoplasts. Comparison of protoplasts with plant data required changes in the normalization scheme, which we implemented. We were ultimately successful in tackling these challenges and with the help of these methods we now have the largest-to-date quantitatively characterized parts library for plants. This work was recently published in Nature Methods [39].

Chapter 4 then goes on to describe how we used this quantitatively characterized plant part library to predict ideal part combination for construction of a negative inhibition based toggle switch. Chapter 4 also goes through our predictions for a positive feedback switch. For the positive feedback switch we did not have a library of plant parts but instead a few

preliminary experiments probing the behavior of the full circuit in Arabidopsis. With this information we strove to identify ideal properties each part should have when designing a positive based feedback system. To do this we constructed a mathematical model to describe the system, followed by narrowing down the region of parameter space where the preliminary plant system could exist. This then led to predictions in how to change the preliminary plant parts to make them more suitable for generating a positive feedback bi-stable switch.

In Chapters 2-4 of my thesis I use methods that could be described as bottom-up, in the sense that they describe work for the construction of a quantitatively accurate model at the level of individual genes and proteins. In the last chapter I move to using a different suite of mathematical and computational methods that have been developed to look at biological systems from the top-down. In other words, Chapter 5 takes the top down approach in looking for genetic markers for drug resistance in large-scale gene expression data of cancer cell lines. These genetic markers may form part of naturally occurring biological switches that turn on drug resistance processes in cancer cells. Cancer can be thought of as the miss-regulation of the biochemical state of a cell or groups of cells. If we can understand the biochemical state of the cancer perhaps we can predict which drug or treatment would be best to treat the disease in different individuals. Several types of data have been collected to learn about this biochemical state. Among them is microarray data describing the transcriptome of different cell lines derived from different tumor types. These cell lines have also been used in drug screens looking to see how the sensitivity to each drug changes across cells lines. If we can predict cell line sensitivity to a drug, and we can collect a tumor biopsy from an individual with cancer that is comparable to one of the cell lines, then we can use this data to start to predict which drug to

give which patient. However, as there are many computational methods available for predicting drug sensitivity, the challenge then becomes to identify which computational method to use. A few model comparison studies have been done to assess model performance in identifying key biochemical states useful for prediction of drug treatment [40] [6]. One of these studies used a crowd-sourcing approach to attract many different groups to submit predictive models for breast cancer given the same data set [6]. Although there are many databases storing data useful for prediction of drug sensitivity, two well-known databases that are used to predict drug sensitivity for cancer cell lines come from: (1) a repository of data from 60 cancer cell lines maintained by the National Cancer Institute, called the NCI60, and (2) the Genomics of Drug Sensitivity in Cancer (GDSC) database maintained by the Wellcome Trust [41] [42]. However, there has been no study that compares predictions between these two databases to assess the reliability of the data and predictions. We fill this gap by running a systematic comparative study of different models, presented in Chapter 5 using the combined set of cell lines in the NCI60 and the GDSC and two different cancer types, lung and bladder. We found that validation of normalizing methods is imperative when using cell line microarray and drug data across databases. It was interesting that drug sensitivity data collected for very different assays was comparable. We also found data that suggested that some gene filtration methods can be damaging to the overall predictive power across many model types.

From synthetic biology to modern medicine, our understanding of the molecular state of cells and organisms is being revolutionized by incorporating computational methods synergistically with wet lab experimental study. This dissertation adds to this growing body of interdisciplinary work by addressing challenges and gaps in our knowledge. To summarize again

in brief: Chapter 2 addresses the previously uncharacterized effect of leaky expression on biological switches. Chapter 3 describes the development of the largest quantitatively characterize plant part library to date. Chapter 4 develops methods for predicting *in silico* ideal molecular part combinations and properties for building feedback based biological switches. Chapter 5 then concludes this work with a comparative analysis of computational methods used to identify bio-markers involved in setting the molecular state for different cancer cell lines.

REFERENCES

[1] A. Gavish, A. Shwartz, A. Weizman, E. Schejter, B.-Z. Shilo and N. Barkai, "Periodic patterning of the Drosophila eye is stabilized by the diffusible activator Scabrous," *Nature Communications,* vol. 7, no. 10461, 2016.

[2] D. Chen and A. Arkin, "Sequestration-based bistability enables tuning of the switching boundaries and design of a latch," *Molecular Systems Biology,* vol. 8, no. 620, 2012.

[3] A. Goldbeter and D. Koshland, "Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects.," *The Journal of Biological Chemistry,* vol. 259, no. 23, pp. 14441-14447, 1984.

[4] G. Hornung and N. Barkai, "Noise Propagation and Signaling Sensitivity in Biological Networks: A Role for Positive Feedback," *PLOS,* vol. 4, no. 1, 2008.

[5] A. Perelson and P. Nelson, "Mathematical Analysis of HIV-1 Dynamics in Vivo," *SIAM Rev.,* vol. 41, no. 1, pp. 3-44, 1999.

[6] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. Collins, D. Gallahan, D. Singer, J. Saez-Rodrigue, S. Kaski, J. Gray and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology,* vol. 32, pp. 1202-1212, 2014.

[7] J. Zhao, X.-S. Zhang and S. Zhang, "Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs," *CPT: Pharmacometrics and Systems Pharmacology,* vol. 3, no. 2, p. e102, 2014.

[8] M. Birtwistle, M. Hatakeyama, N. Yumoto, B. Ogunnaike, J. Hoek and B. Kholodenko, "Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses," *Molecular Systems Biology,* vol. 3, no. 144, pp. 1744-4292, 2007.

[9] P. Altrock, L. Liu and F. Michor, "The mathematics of cancer: integrating quantitative models," *Nature Reviews Cancer,* vol. 15, pp. 730-745, 2015.

[10] J. Young, M. Peyton, H. S. Kim, E. McMillan, J. Minna, M. White and E. Marcotte, "Computational discovery of pathway-level genetic vulnerabilities in non-small-cell lung cancer," *Bioinformatics,* 2016.

[11] D. Chu and T. v. d. Haar, "The architecture of eukaryotic translation," *Nucleic Acids Research,* vol. 40, pp. 10098-10106, 2012.

[12] W. Mather, J. Hasty, L. Tsimring and R. Williams, "Translational Cross Talk in Gene Networks," *Biophysical Journal,* vol. 104, pp. 2564-2572, 2013.

[13] K. B. Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg and S. Itzkovitz, "Bursty Gene Expression in the Intact Mammalian Liver," *Molecular Cell,* vol. 58, no. 1, pp. 147-156, 2015.

[14] N. Kumar, A. Singh and R. Kulkarni, "Transcriptional Bursting in Gene Expression: Analytical Results for General Stochastic Models," *PLOS,* vol. 11, no. 10, p. e1004292, 2015.

[15] E. Voit, H. Martens and S. Omholt, "150 Years of the Mass Action Law," *PLOS,* vol. 11, no. 1, 2011.

[16] J. Faeder, M. Blinov and W. Hlavacek, "Rule-based modeling of biochemical systems with BioNetGen," in *Systems Biology*, vol. 500, Humana Press, 2009, pp. 113-167.

[17] *MATLAB and SymBiology Toolbox Release 2014b,* Natick, Massachusetts, United States: The MathWorks, Inc., 2012.

[18] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes and U. Kummer, "COPASI—a COmplex PAthway SImulator," *Bioinformatics,* vol. 22, no. 24, pp. 3067-3074, 2006.

[19] R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz and A. Tosaki, "The Hill equation and the origin of quantitative," *Arch. Hist. Exact Sci.,* vol. 66, pp. 427-438, 2012.

[20] P. Gutierrez, D. Monteoliva and L. Diambra, "Cooperative Binding of Transcription Factors Promotes Bimodal Gene Expression Response," *PLOS one,* vol. 7, no. 9, 2012.

[21] E. Ozbudak, M. Thattai, H. Lim, B. Shraiman and A. Oudenaarden, "Multistability in the lactose utilization network of *Escherichia coli*," *Nature Letters,* vol. 427, pp. 737-740, 2004.

[22] J. Berg, J. Tymoczko and L. Stryer, "Section10.2 Hemoglobin Transports Oxygen Efficiently by Binding Oxygen Cooperatively," in *Biochemistry. 5th edition.*, New York, W H Freeman, 2002.

[23] T. Gardner, C. Cantor and J. Collins, "Construction of a genetic toggle switch in *Escherichia coli*," *Nature,* pp. 339-342, 2000.

[24] C. D. Smolke, "Building outside of the box: iGEM and the BioBricks Foundation," *Nature Biotechnology,* vol. 27, pp. 1099-1102, 2009.

[25] K. Lai, M. Robertson and D. Schaffer, "The Sonic Hedgehog Signaling System as a Bistable Genetic Switch," *Biophysical Journal,* vol. 86, pp. 2748-2757, 2004.

[26] K. Müller, R. Engesser, S. Metzger, S. Schulz, M. Kämpf, M. Busacker, T. Steinberg, P. Tomakidi, M. Ehrbar, F. Nagy, J. Timmer, M. Zubriggen and W. Weber, "A red/far-red light-responsive bi-stable toggle switch to control gene expression in mammalian cells.," *Nucleic Acids Res.,* vol. 7, 2013.

[27] X. Xu, Y.-L. Sun and T. Hoey, "Cooperative DNA Binding and Sequence-Selective Recognition Conferred by the STAT Amino-Terminal Domain," *Science,* vol. 273, no. 5276, pp. 794-797, 1996.

[28] F. JE and H. SH, "Ultrasensitivity part II: multisite phosphorylation, stoichiometric inhibitors, and positive feedback.," *Trends Biochem Sci.,* vol. 39, no. 11, pp. 556-569, 2014.

[29] S. Harris and A. Levine, "The p53 pathway: positive and negative inhibition loops," *Oncogene,* vol. 24, pp. 2899-2908, 2005.

[30] O. Weiner, P. Neilsen, G. Prestwich, M. Kirschner, L. Cantley and H. Bourne, "A PtdInsP3- and Rho GTPase-mediated positive feedback loop regulates neutrophil polarity," *Nature Cell Biology,* vol. 4, pp. 509-513, 2002.

[31] J. Wang, C. Li and E. Wang, "Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network," *PNAS,* vol. 107, no. 18, pp. 8195-8200, 2010.

[32] I. Lestas, G. Vinnicombe and J. Paulsson, "Fundamental limits on the suppression of molecular fluctuations," *Nature,* vol. 467, no. 7312, pp. 174-178, 2010.

[33] Y. Dublanche, K. Michalodimitrakis, N. Ku¨mmerer, M. Foglierini and L. Serrano, "Noise in transcription negative inhibition loops: simulation and experimental analysis," *Molecular Systems Biology,* vol. 1, 2006.

[34] A. Eldar and M. Elowitz, "Functional roles for noise in genetic circuits," *Nature,* vol. 467, pp. 167-173, 2010.

[35] M. Kittisopikul and G. Süel, "Biological role of noise encoded in a genetic network motif," *PNAS,* vol. 107, no. 30, pp. 13300-13305, 2010.

[36] K. Schaumberg, A. Torres, M. Po, T. Kassaw, C. Zalewski, M. Antunes, J. Medford and A. Prasad, "Practical Properties of Biological Switches," in *QBio Conferance*, Santa Fe, New Mexico, USA, 2014.

[37] J. Medford and A. Prasad, "Plant synthetic biology takes root," *Science,* vol. 346, no. 6206, pp. 162-163, 2014.

[38] J. Zuo, Q.-W. Niu, G. Frugis and N.-H. Chua, "The WUSCHEL gene promotes vegetative-to-embryonic transition in Arabidopsis," *The Plant Journal,* vol. 30, no. 3, 2002.

[39] K. Schaumberg, M. Antunes, T. Kassaw, W. Xu, C. Zalewski, J. Medford and A. Prasad, "Quantitative characterization of genetic parts and circuits for plant synthetic biology," *Nature Methods,* vol. 13, no. 1, pp. 94-100, 2016.

[40] R. Braun and S. Shah, "Network Methods for Pathway Analysis of Genomic Data," *arXiv preprint 1411.1993,* 2015.

[41] R. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer,* vol. 6, pp. 813-823, 2006.

[42] W. Yang, J. Soares, P. Greninger, E. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. Smith, R. Thompson, S. Ramaswamy, A. Futreal, D. Haber, M. Stratton, C. Benes, U. McDermott and M. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research,* vol. 41, pp. 955-961, 2012.

CHAPTER 2

An investigation of leaky expression's effect on biological feedback circuits

2.1 Introduction

Gene expression is controlled by promoters that are specialized DNA sequences usually situated upstream of a gene's coding region. When the gene is being expressed at a high level we say that the promoter is in an "ON" state, and when it is not expressed or expressed at a low level we say that it is in an "OFF" state. One common way in which this change in expression level is achieved is through repressors (proteins that repress or prevent the expression of a gene from a promoter that is otherwise constitutively expressed). Another common way this change in expression occurs is with a promoter whose gene expression can be activated when an activator protein is bound to it. In both cases however, the promoter is often never completely "OFF" when fully repressed or not activated. A possible exception to this behavior can be found in synthetic gene circuits which contain DNA invertases [1] [2] and natural processes using recombinases, such as sporulation in *B. subtilis* [3], whose physical change in DNA structure may allow for a zero leakage system. However, for many systems controlled by repressors or activators, there is low level activity, which can be called "leaky expression". In particular, several inducible gene expression systems allowing for time dependent control of gene expression have been said to be "leaky" or even "notoriously leaky" [4] [5]. Another example of leaky expression can be found in dCas9 activation of endogenous genes [6]. The body of work characterizing the effect of this leaky expression on other properties of the genetic systems is small but does include a recent comparison of positive and

negative inhibition in how leaky expression affects the noise within these systems [7]. Also there has been work done to describe how the dynamic range of small molecule inducers can be tuned in prokaryotes via modulation of intracellular receptors without notably changing the amount of leaky expression [8]. However, more work is needed to look into the effect of leaky expression on the ability of a system to be deterministically bi-stable and the practical properties within the bi-stable regions of different circuit topologies.

The work presented here demonstrates how leaky expression plays a role in the balance needed for bi-stability as well as functional roles (*i.e.* practical properties) of switches. Practical properties of switches investigated here are: fold change, FC, and basal level, BL, of the switch. FC is the highest level molecular expression divided by the lowest level molecular expression of a switch. BL is the low level molecular expression of a switch. Some researchers use the terms leaky expression and basal expression interchangeably. Here, however, we will define BL as the state of the circuit which has the lowest gene expression. Leaky expression then describes the lowest possible expression level of a particular molecule. This is often the "Off" state of a promoter. To better explain the difference between BL and leaky expression we can look to a rate equation for a repressible promoter. We include leaky expression by adding a constant term, as seen below.

$$\frac{dx}{dt} = \alpha + \frac{\beta}{1+R} - x \qquad 2.1$$

An example of how leaky expression is different from BL can be found in the rate equation 2.1, which describes a hypothetical genetic system regulating the expression of the molecule *x*. This equation has three terms on the right hand side that describe the expression rate of the molecule, *x*, over time, *t*. The first term α represents leaky expression; it describes the amount

of gene expression when the system is fully repressed. The second term, $\frac{\beta}{1+R}$, represents how the rate of expression is changed given a repressor, *R,* where β is a constant. The third term, *-x*, describes the degradation of the molecule *x*. The leaky expression of the system, α, is different than the BL of the system, which depends on the steady state maximum expression of *R*. In other words, it is possible for the BL to be larger than the leaky expression for this system. Even though these properties of bi-stable switches are important in fields such as synthetic biology to stem cell research, few have explored the role leaky expression has on these properties for different circuit topologies. Comparing different circuit topologies in their effectiveness to handle leaky expression should give us insight on how to design better circuits in synthetic biology, while providing a platform for understanding what role leaky expression may play in nature.

As listed below, this study takes a computational approach to investigate how the practical properties of these circuits are affected by leaky expression. Three questions were asked.

1) How much leaky expression can a bi-stable system withstand? In other words, can leaky expression abrogate bi-stability?
2) How does the FC of a bi-stable system, i.e. the ratio between the "high" state and the "low" state (*i.e.* BL), change with respect to leaky expression values?
3) How does the low state (BL) in these bi-stable systems change with respect to leaky expression?

We ask question one because we are interested in the genetic systems' ability to be a bi-stable system in the presence of leaky expression. We ask question two because a switch with a large difference between its low and high molecular states would prove useful in many synthetic

biology applications in need of a large change in a molecule of interest. We ask question three because when integrating these genetic circuits into natural systems, the low molecular state must often lie below the threshold of an endogenous molecule or there can be no "true" off state for the switch. Although there are other practical properties of biological switches, these three are readily useful for our *in silico* design of biological switches in coming chapters and as the need arises other practical properties could be explored from the algorithms developed here.

2.2 Methods

2.2.1 System Layout

Bi-stable biological switches are commonly based on either negative inhibition or positive feedback. We therefore chose to study both system topologies. We modeled the negative inhibition system following Gardner et. al. [9] and the positive feedback system after Chen and Arkin [10]. The negative inhibition system is a four promoter, two repressors (R_1 and R_2) and two inducers (I_1 and I_2) genetic construct. The positive-feedback switch (denoted as the positive feedback system) is a three promoter, one activator (B), two inducers (I_1 and I_2) and one inhibitor (A) genetic construct. Figure 2.1 a and b illustrate the layout of each switch.

Why would one expect the positive and negative inhibition systems described in Figure 2.1 to act like switches? For the negative inhibition system, when the repression of R_1 and R_2 is balanced, as described by Gardner et. al. [9], then as R_1 increases it keeps R_2 in check allowing for a stable high R_1 and low R_2 state. However, when R_2 increases it keeps R_1 in check allowing for a stable high R_2 and low R_1 state. The ability to have two stable states is the definition of a bi-stable system. The ability to exist in a high molecular state and a low molecular state makes

this negative inhibition system a switch. The positive feedback system is also able to create a bi-stable system when the molecule regulation is balanced [10]. When *B* is at a low expression state *A* can keep *B* in check. However, when *B* is at a high expression state *A* will have little effect on *B*. Unless of course, *A* is induced to have a high expression level, which should reset the expression of *B*. Note that while we have assumed that *A* is an inhibiting protein, it could also be a microRNA.

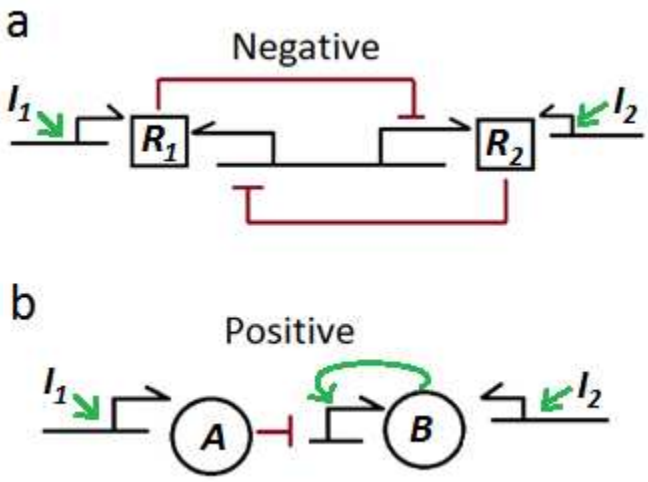

Figure 2.1 System Layouts for Negative and Positive Biological Switches. Diagram of the positive and negative inhibition-based switches explored in this chapter. Panel **a,** the diagram of the negative inhibition switch depicts a four promoter based genetic circuit. Each of the two repressible promoters drives the expression of either repressors R_1 or R_2. Also the two inducible promoters drive the expression of the two repressors. Each inducible promoter can be activated by one small molecule inducer, either I_1 or I_2. Panel **b**, the diagram of the positive inhibition switch depicts a three promoter based genetic circuit also with two inducible promoters driving either the expression of an activator protein, *B* or an inhibitor, *A* (which is thought to bind to either the protein *B* or *B*'s mRNA transcript). Each inducible promoter can also be activated by one small molecule inducer, either I_1 or I_2. The key difference between this positive feedback system and the negative inhibition system is in the third promoter of the positive feedback system. This third promoter is driving the expression of its own activator. Green arrows represent the upregulation of the molecule at the end of the arrow by the molecule at the beginning of the arrow. Red blunt arrows represent the down regulation of the molecule at the end of the arrow by the molecule at the beginning of the arrow, with the exception of the blunt arrow coming from *A* which could represent either inhibition by a micro RNA or a protein capable of sequestration of *B*.

2.2.2 Model Design

To model the negative inhibition system, Hill equations were used to describe the effect of the transcription factors on the production of R_1 and R_2. Leaky expression was added as a zero order process and degradation was added as a first order process. These equations are very similar to those employed by Gardner et. al. [9].

$$\frac{dR_1}{dt} = \alpha_1 + \frac{\beta_1}{1 + \left(\frac{R_2}{k_1}\right)^{n1}} - d_1 R_1 + f(I_1) \quad 2.2$$

$$\frac{dR_2}{dt} = \alpha_2 + \frac{\beta_2}{1 + \left(\frac{R_1}{k_2}\right)^{n2}} - d_2 R_2 + g(I_2) \quad 2.3$$

Where: R_1 and R_2 represent the expression of the repressors R_1 and R_2. α_n $(n = 1,2)$ represents the leaky expression of repressor "*n*" (*i.e.* 1 or 2) from both inducible and repressible promoters. Using the same notation, β_n helps to set the max expression of repressor "*n*" (*i.e.* 1 or 2) ; k_n is concentration of repressor "*n*" (*i.e.* 1 or 2) needed to bring the max expression, $\alpha_n + \beta_n$, to $\alpha_n + \frac{\beta_n}{2}$, or the concentration of repressor needed to bring the max expression to half its value, ($\frac{\beta_n}{2}$), assuming $\beta_n \gg \alpha_n$, n_n is the Hill coefficient of the repressor "*n*" (*i.e.* 1 or 2) input-output function, d_n is the degradation coefficient for repressor "*n*" (*i.e.* 1 or 2) and $f(I_1)$ and $g(I_2)$ are unknown functions for the inducer's impact on the circuit. The exact relationship defining the effect the inducers have on the rate equation is not needed as we are only looking at the stability of the system without addition of the inducers. What we are assuming is that when the inducer is present, it is enough to switch the state of the system. The difference in equations 2.2 and 2.3, used here for the negative inhibition system and those used by Gardner et. al. [9], is in the addition of leaky expression terms for each repressor (e.g. α_1 and α_2). These

terms assume that no matter the state of the system there exists some expression from those promoters.

The positive feedback system equations were constructed by assuming zero order leaky expression and first order degradation for both *A* (an inhibitor illustrated in Fig. 2.1) and *B* (an activator illustrated in Fig 2.1). Unbound *B*, B_u, is then calculated assuming *A* binds to *B* faster and independently of degradation and leaky expression. B_u is then assumed to affect *B*'s production via a positive Hill equation. This system of equations is very similar to those employed by Chen and Arkin [10].

$$\frac{dA}{dt} = \alpha_1 - d_1 A + f(I_1) \tag{2.4}$$

$$\frac{dB}{dt} = \alpha_2 + \frac{\beta_2 {B_u}^{n_2}}{1 + \left(\frac{B_u}{k_2}\right)^{n_2}} - d_2 B + g(I_2) \tag{2.5}$$

$$B_u = \frac{1}{2}\left((B - A - k_d) + \sqrt{(B + A + k_d)^2 - 4BA}\right) \tag{2.6}$$

Where *A* represents the expression level of a regulatory molecule, such as a micro RNA or repressor protein, *B* represents the activator responsible for the positive feedback of the positive feedback promoter, α_1 is the leaky expression of *A*, α_2 is the leaky expression of *B*, d_1 is the degradation coefficient for *A*, d_2 is the degradation coefficient for *B*, B_u is the amount of *B* not bound by *A*, β_2 scales the maximum production of the positive feedback promoter, k_2 helps to scale the effect of B_u's max positive feedback effect. k_d is the dissociation constant for *A* binding to *B*, n_2 represents the hill coefficient or the activating hill function. Finally $f(I_1)$ and $g(I_2)$ are unknown functions for the inducer's impact on the circuit. The exact relationship defining the effect the inducers have on the rate equation is not needed as we are only looking

at the stability of the system without addition of the inducers. What we are assuming is that when the inducer is present it is enough to switch the state of the system.

The differences in equations 2.4-6, used here for the positive feedback system and those used by Chen and Arkin are in the cooperativity component and in the addition of a rate equation for *A*. The cooperativity component, n_2, is to account for possible non-linear behavior of the positive feedback promoter. The leaky expression terms follow the same reasoning as in the negative inhibition system: no matter the state of the system there exists some expression of *A* and *B*. The Gardner et. al. and Chen and Arkin equations have been shown to represent their corresponding biological system using wet lab experimental data [9] [10] and should therefore be a productive place to begin our study.

It is noteworthy in the derivation of B_u found in Box 2.1 that a difference in time scales is assumed as B_u is calculated independently from the rest of the system. There are more sophisticated methods for which to simplify a system given a difference in time scales such as scaling the system by the slowest time [11]. However, we chose to follow Chen and Arkin due to their use of a simple and commonly used in biological modeling approach along with our desire to start with a tested simple model for the system. Also as the delay due to translation is not considered in these equations, hence whether A is a micro RNA or a protein will not affect how it is modeled. As our analysis at this point is only considering steady state behavior and models without delay have been shown to be able to represent steady state behavior [9] [10], I believe this is a reasonable assumption. If we were to look at the stochastic or time-dependent nature of the system, delay may play more of a role.

Box 2.1 Derivation of B_u. This box describes Chen and Arkin's derivation of B_u.

$$A + B \xrightarrow{f} A{:}\,B$$

$$A{:}\,B \xrightarrow{b} A + B$$

Law of Mass Action

$$\frac{dA{:}\,B}{dt} = fAB - bA{:}\,B$$

Assume Steady State and Let $\frac{b}{f} = k_d$

$$AB = A{:}\,Bk_d$$

Use Conservation Rules to Substitute in $\mathrm{A} = A_T - A{:}\,B$ and $\mathrm{A{:}\,B} = B_T - B$

$$BA_T - B(B_T - B) = (B_T - B)k_d$$

Solve for *B* and Notice +/- Must Equal + to have Biologically Relevant Values for *B*

$$B = 0.5\left[B_T - A_T - k_d + \sqrt{(B_T + A_T + k_d)^2 - 4B_TA_T}\right]$$

Assume B_T and A_T are the Instantaneous Values of B and A in the Full Rate Equation

Assume B = unbound *B* (i.e. B_u)

$$B_u = 0.5\left[B - \mathrm{A} - k_d + \sqrt{(B + A + k_d)^2 - 4BA}\right]$$

These equations were non-dimensionalized to make them easier to work with. As all possible behavior in the dimensional equations can also be observed with the non-dimensional equations, we will not lose information by non-dimensionalizing.

$$\frac{dr_1}{d\tau} = L_1 + \frac{X_1}{1 + r_2{}^{n_1}} - r_1 + \tilde{f}(I_1) \qquad 2.7$$

$$\frac{dr_2}{d\tau} = L_2 + \frac{X_2}{1 + (r_1)^{n_2}} - D_R r_2 + \tilde{g}(I_2) \qquad 2.8$$

Where: r_1, r_2 and τ are dimensionless and related to the dimensional terms R_1, R_2 and t by the following relationships: $r_1 = \frac{R_1}{k_2}$, $r_2 = \frac{R_2}{k_1}$, $\tau = \frac{t}{d_1}$. Also, the dimensionless parameters are combinations of the dimensional parameters as follows: $L_1 = \frac{\alpha_1}{k_2 d_1}$, $L_2 = \frac{\alpha_2}{k_1 d_1}$, $X_1 = \frac{\beta_1}{k_2 d_1}$, $X_2 = \frac{\beta_2}{k_1 d_1}$ and $D_R = \frac{d_2}{d_1}$. Finally, $\tilde{f}(I_1)$ and $\tilde{g}(I_2)$ are unknown functions for the inducer's impact on the circuit.

$$\frac{da}{d\tau} = 1 - a + \tilde{f}(I_1) \qquad 2.9$$

$$\frac{db}{d\tau} = L_2 + \frac{X\, {b_u}^{n_2}}{1 + {b_u}^{n_2}} - D_R b + \tilde{g}(I_2) \qquad 2.10$$

$$b_u = \frac{1}{2}\left((b - aL_1 - T) + \sqrt{(b + aL_1 + T)^2 - 4abL_1}\right) \qquad 2.11$$

Where: a, b and τ are dimensionless and related to the dimensional terms A, B and t by the following relationships: $a = \frac{A}{k_2}$, $b = \frac{B}{k_2}$, $\tau = \frac{t}{d_1}$. Also, the dimensionless parameters are combinations of the dimensional parameters as follows: $L_1 = \frac{\alpha_1}{k_2 d_1}$, $L_2 = \frac{\alpha_2}{k_2 d_1}$, $X = \frac{\beta_2}{d_1 {k_2}^{(n_2-1)}}$, $T = \frac{k_d}{k_2}$ and $D_R = \frac{d_2}{d_1}$. Finally, $\tilde{f}(I_1)$ and $\tilde{g}(I_2)$ are unknown functions for the inducer's impact on the circuit. Note that in both circuits the exact relationship defining the effect the inducers have on the rate equation is not needed as we are only looking at the stability of the system without addition of the inducers. As pointed out before, we are assuming that when the inducer is present it is enough to switch the state of the system.

The non-dimensionalized negative inhibition system is presented by equations 2.7 and 2.8. The non-dimensionalized positive feedback system is presented by equations 2.9-11. Three facts to be aware of when comparing data collected in these non-dimensionalized systems are as follows: (1) When the scaled "leaky expression" terms L_1 and L_2 in both systems are scaled equivalently they can be compared directly without returning to dimensionalized space. (2) The values of r_1 and r_2 are scaled by different *k* values. It is only when k_1=k_2 for the negative inhibition system that $r_1$and r_2 non-dimensional terms are directly comparable. On the other hand, for the positive feedback system *a* and *b* are scaled by the same *k* values. Therefore, their non-dimensional terms are always comparable. (3) Assuming *k* values across systems are

equivalent, r_1, r_2, a and b values should be directly comparable. Note: If the k values are not equivalent, the relationships found in the non-dimensional space would need to be mapped back to the dimensionalized space for a particular system whose k parameters are known. Even if this is the case, the analysis done here provides a platform from which these systems can be investigated for many different systems found in many different organisms. To increase readability from this point on we will work with the non-dimensionalized systems unless otherwise noted.

2.2.3 Parameterization of Model

What are biological relevant parameter values for each of these non-dimensional parameters in the two systems? This is a challenging question to answer in the absence of experimental data for the systems in question. However, databases such as Harvard Bio Numbers [12] have started to store information on general estimates for parameters such as stable protein degradation, transcription and translations rates. We also have the advantage of having developed one of the first quantitative plant part libraries (as described in Chapter 3) giving more information to draw on when determining these estimates. We decided to use these resources to find a range or starting point for each parameter in our models even though the parameters are estimated in different organisms. As we are only looking for biological relevant parameter values, this will give us a sufficient starting point. Although outside the scope of this study, it should be noted that when building or observing these systems in a particular organism a more defined range of parameters should be obtained. Details on the estimate of each non-dimensionalized parameter can be found in Table 2.1.

Table 2.1 Descriptions of Parameter Estimations.

Negative			Positive		
Param-eter	Value	Reasoning	Param-eter	Value	Reasoning
L_1	unknown	plan to vary	L_1	Unknown	plan to vary
L_2	unknown	plan to vary	L_2	Unknown	plan to vary
D_R	1	If both R_1 and R_2 are stable proteins, then we assume the ratio of their degradation rates will be around 1.	D_R	0.02 ,1	If A is a micro RNA or protein and B is a protein, we can look for an estimate of micro RNA and protein degradation in the literature using Harvard Bio Numbers [12]. 0.05/min is in the range of estimated micro RNA degradation [13]. A particular stable protein degradation was estimated to be about 0.05/hour [14] giving us a starting point for this estimate ($0.05hr^{-1}/0.05min^{-1} \sim = 0.02$). On the other hand, if both *A* and *B* are proteins as presented in the Chen and Arkin design of this system [10], the ratio of their degradation rates we will assume to be around 1.
n_1 and n_2	2-4	Assuming we are working with transcription factors similar to those used in our library of plant parts developed in Ch. 3, we will start by assuming the amount of achievable sigmodality is low, 2, to the max sigmodality observed in our library of 4.	n_2	2-4	Assuming we are working with transcription factors similar to those used in our library of plant parts developed in Ch. 3, we will start by assuming the amount of achievable sigmodality is low, 2, to the max sigmodality observed in our library,4.
X_1 and X_2	0.01-100	From our library of plant parts developed in Ch. 2 we get a range of [0.01-34]. As we are comparing to a system where less is known about each parameter we should also vary these parameters with vigor increasing this range to [0.01-100].	X	3×10^{-7}, 100	To get a rough idea of this value, Harvard Bio Numbers [12] was referenced. We found: - A translation rate of 18 amino acids, aa, per-sec and a transcription rate of 54 nucleotides, nt, per-sec [21]. -Also "the median length of the proteins annotated among Eukaryotes (361 aa)" [15]. -Given 3 nt per aa, this gives us 1083 nt in the coding region (1083nt*(1sec/54nt) = 20sec)+(361aa * (1sec/18aa)=20sec) = 40sec/protein. This goes to $b_2/d_1 = 30$, assuming d_1 is around $0.5/min = 8.3 \times 10^{-4}/sec$. -For k_2 a bottom limits would be 1 molecule and a top limit would be the max number of molecules produced. Ron Milo in 2013 [16] estimated a total number of proteins per cell to be 2–4 million proteins per cubic micro. A rough estimate of plant cell size is 25 cubic microns [17]. 4*25 = 100 million molecules. Assuming $n_2 = 2$ This gives us a range for X as [3×10^{-7}, 30]. There are many unknowns. Therefore we will vary this parameter with vigor increasing the range to [3×10^{-7}-100].
			T	$0.001\text{-}10^8$	At this point estimating the binding rate and dissociation rate of different molecule types is quite challenging. So we will start with $k_d = 1$, assuming it is just as likely to bind and unbind, and then vary this parameter. k_2 as estimated above lives between [1 and 10^8]. There are many unknowns. Therefore we will vary this parameter with vigor increasing the range to $0.001\text{-}10^8$.

Note: These parameters come from many different organisms giving us only a general idea of what is biologically relevant. An important step when incorporating this information into a particular organism, such as Arabidopsis, would be to find the range of parameter values relevant for that organism.

The uncertainty for each parameter is large. As we are working with systems capable of being either bi-stable or mono-stable, we will need to explore this larger parameter space to determine which regions are bi-stable and which are mono-stable. Remember bi-stable systems are capable of existing in one of two steady state values depending on their history, where mono-stable systems have only one steady state value. Transitioning from mono-stable to bi-stable is illustrated in the nullclines and phase diagram seen in Figure 2.2.

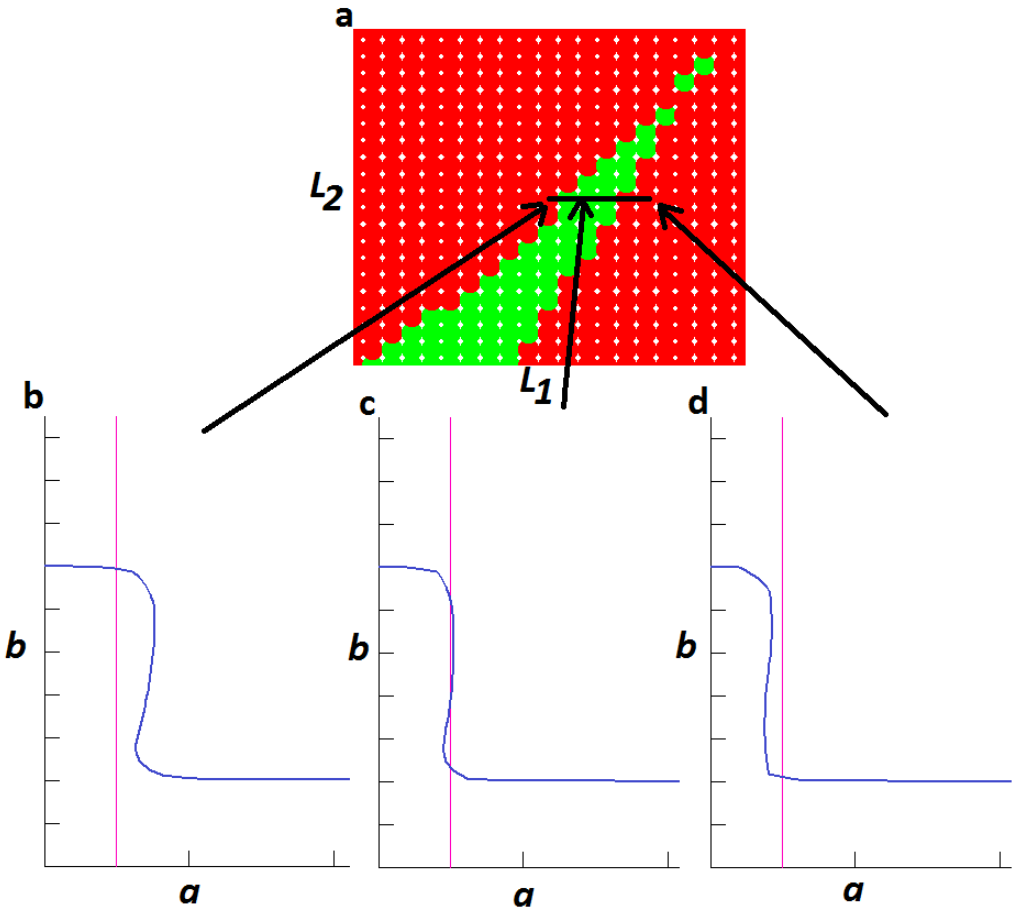

Figure 2.2 Illustration of Nullcines changing Across Phase Diagram. Panel **a** is a phase diagram created for the positive feedback system using parameters outlined in Table 2.2. The phase diagram depicts the bi-stable region in the positive feedback system. The green area represents the area of leaky expression parameter space which is bi-stable. The red is correspondingly, mono-stable. The x-axis represents different L_1 values. The y-axis represents different L_2 values. The arrows connecting panels **b**, **c** and **d** to panel **a**, indicate where the nullcline diagrams come from in terms of parameter space. Panels **b-d** are plots of the nullclines in the non-dimensionalized system. The x-axis represents different values of *a* where the y axis represents

different values of *b*. The blue line represents the *b* nullcline. The *b* nullcline is combinations of *a* and *b* values where the rate equation describing *b* equal zero (i.e. the amount of *b* is not changing). The magenta line represent the *a* nullcline. The points at which neither the amount of *a* nor the amount of *b* is changing are indicated when the blue and magenta lines cross. These are called fixed points. When two stable fixed points exist the system is said to be bi-stable. Panels **b** and **d** only have one fixed point which is stable making them mono-stable systems. As the arrows indicate, panels **b** and **d** come from the red mono-stable area in the phase diagram of panel **a**. Panel **c** has three fixed points two of which are stable making this a bi-stable system. The arrow from panel **c** indicates that this system maps back to part of the green bi-stable area in the phase diagram of panel **a**. Stability of these fixed points was determined during the numerical simulations of these systems but could also be calculated analytically by linearizing about the fixed point.

Again as Table 2.1 depicts a large range of uncertainty for each parameter, we decided to divide our analysis into three sections. First we varied L_1 and L_2 while keeping all other parameters fixed giving us a slice of parameter space existing in $\mathbb{R}^2$ (two dimensional real coordinate space). This provided our first look at the effect leaky expression has on the system. Table 2.2 describes the parameter values used for this first look. Second, we varied these parameters by doubling and halving their values (except for *n*'s which we varied to be 2, 3 or 4) to see if the trends we found in our first look held. Thirdly, a larger parameter range analysis was conducted, where we looked to see if any large deviation occurred in the effect leaky expression had on the bi-stable system. The initial parameter values can be found in Table 2.2 and the larger range parameters can be found in Table 2.3.

Table 2.2 Parameter Values. Initial parameter values used for the ODE modeling of the systems.

Negative		Positive	
Parameter	Value	Parameter	Value
L_1	plan to vary	L_1	plan to vary
L_2	plan to vary	L_2	plan to vary

D_R	1	D_R	1
n_1 and n_2	3	n_2	3
X_1 and X_2	2	X	10
		T	1

Table 2.3 Large Range Parameter Values. Parameter values used in the larger range parameter analysis. Note: As all X = 0.01 combinations in the positive feedback system showed no sign of bi-stability we did not vary this parameter further. This is also true for X_1 and X_2.

Negative			Positive
Parameter	Value	Parameter	Value
L_1	plan to vary	L_1	plan to vary
L_2	plan to vary	L_2	plan to vary
D_R	1/3,1,3	D_R	1/3,1,3
n_1 and n_2	2,3,4	n_2	2,3,4
X_1 and X_2	0.01,1,10,100	X	0.01,1,10,100
		T	0.001,0.1,1,10,10^3,10^5

2.2.4 Exploration of practical properties via deterministic modeling

MATLAB was used to simulate the non-dimensionalized ODE systems. These simulations were used to investigate the following three preliminary questions about switch properties: (1) How much leaky expression can a bi-stable system withstand before losing its bi-stable nature? (2) How does the ratio between the bi-stable states (FC) change with respect to leaky expression values? (3) How does the low state (BL) in these bi-stable systems change with respect to leaky expression?

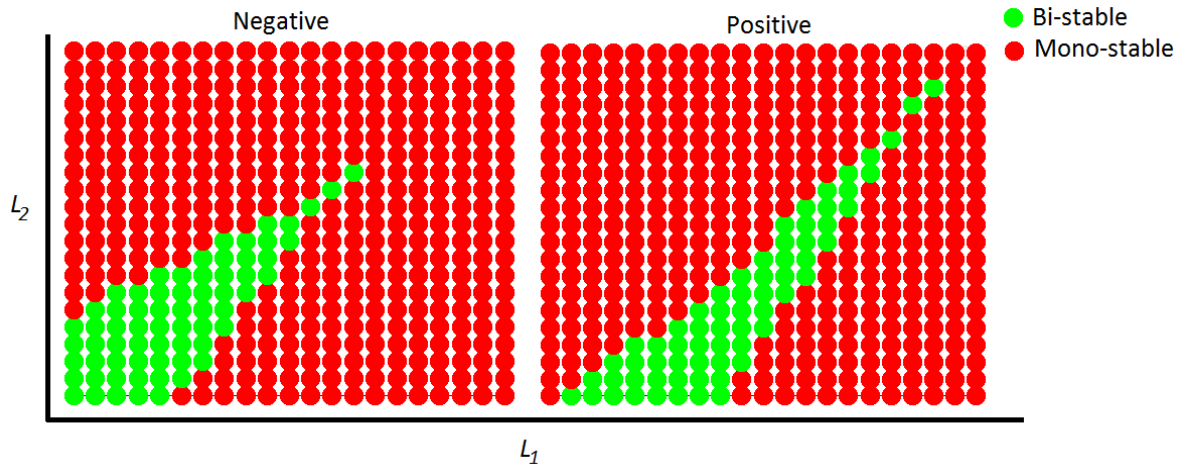

Figure 2.3: Bi-stable Region. Phase diagrams depicting the bi-stable region in both the negative and positive feedback systems. The green area represents the slice of leaky expression parameter space which is bi-stable. The red is correspondingly, mono-stable. The x-axis represents different L_1 values. The y-axis represents different L_2 values. Note: These images were made using the automated program described in Table 2.4 followed by centering the bi-stable area in the figures generated. These images are also here to illustrate the shape of the bi-stable region on not the relative size between the negative and positive systems.

To investigate question 1 we started with the parameters in Table 2.2. As no range for $L_1$and L_2 was defined we varied these parameters from near zero to a max value determined by the size of the bi-stable space with twenty one evenly spaced points. Figure 2.3 illustrates this bi-stable region for both the positive and negative inhibition systems. The size of parameter space to sample was determined by dynamically incrementing the max value for $L_1$and L_2 until the border of the $L_1$and L_2 defined space was mono-stable and contained a sample of the complete bi-stable region. The method associated with this incrementing of the max $L_1$and L_2 values can be found in Table 2.4. The area of the bi-stable region was then estimated from the simulated data. This method of incrementing the max parameter values is only possible because in terms of leaky expression the bi-stable space is finite for both systems in all parameters combinations sampled.

To determine whether a parameter combination was bi-stable, both non-dimensionalized ODE systems (positive and negative) were simulated from two initial conditions. The first initial condition consisted of a high value of *r2* or *b* and 0 for *r1* or *a*. The second initial condition consisted of a high value for r_1 and 0 for r_2, *b* and *a*. The two steady state values of *r2* or *b* were then stored for each simulation starting at the different initial conditions. If the two steady state values were different from each other the parameter combination was classified as bi-stable. Steady states were said to be different from each other if the following conditions were met: (1) The difference between the low and high steady state values was greater than 0.001. Taking note of the non-dimensional relationships for *r2* and *b* ($r_2 = \frac{R_2}{k_2}, b = \frac{B}{k_2}$) a change smaller than 0.001 would mean a change smaller than 0.001*k_2 in number of molecules. (2) The FC between the two states must be greater than 1.1. Given a particular experimental system, one could fine tune these conditions appropriately. For this general analysis however, these conditions will provide a platform from which to investigate the effect of leaky expression on these two different topologies. (Note: A tolerance of 0.9 was set if FC was less than 1. That is to say if the ODE solver numerical approximation error produced a ratio between the high state (steady state value produced by initial conditions set to find a high state) and the low state (steady state value produced by initial conditions set to find the low state) was less than 1 but greater than 0.9, the point in parameter space was stored as mono-stable.)

To look into questions 2 and 3, the low and high steady state values were stored for the bi-stability analysis by writing them out to file for each parameter combination. This allowed for creation of heat maps describing how the low state, and FC changed within the bi-stable region.

We can calculate the FC directly here because the scaling terms from the non-dimensionalization cancel when calculating the ratio we are calling FC. However, the non-dimensionalized BL (*i.e.* low state) is only comparable between systems when their scaling terms are equivalent. Note: From this point on the phrase leaky expression defined bi-stable area will always refer to bi-stable area in the L_1 and L_2 parameter space. BL will represent the amount of r_2 or *b* in the low state as defined in equations 2.7-11. FC is the same in the dimensional and non-dimensional systems as the scaling terms cancel when calculating the ratio between the low and high states.

Table 2.4 Method for incrementing the max value of max $L_1$and L_2. Method used to determine the max $L_1$and L_2 such that the bi-stable region was well sampled for each parameter combination.

Step #	Brief description
1	Start with a parameter range of 0.0001-1 for both $L_1$and L_2. Change this range into a vector of 21 evenly spaced $L_1$and L_2 values. For all possible combinations of these values determine if the parameter combination is mono-stable or bi-stable.
2	Choose the initial increment by which the max $L_1$and L_2will be changed. This is done by starting with an increment of 0 and increasing it by 1 order of magnitude if the bi-stable region was not full sampled for both $L_1$and L_2. It was possible for us to do this independently for both $L_1$and L_2, as max $L_1$is only associated with the right border and L_2 is only associated with the top border of the bi-stable region described in Figure 2.3.
3	Change the size of the increment to be 25% less than its value until the bi-stable region is one increment change away from no-longer being completely sampled.

Using the methods described here we can now compare the following three properties: (1) the size of the leaky expression defined bi-stable space between the positive and negative inhibition systems, (2) how FC and BL change within this bi-stable region and (3) how the BL and FC values themselves compare between the positive and negative inhibition systems. This will give us the first look at how leaky expression affects these systems. We can then vary these parameters by doubling and halving their values (except for *n*'s which we varied to be 2, 3 or 4)

to see if the trends we found remain the same when the other parameters are varied. Finally, we can run an even larger parameter range study to see if any large deviation occurred in the effect leaky expression has on these bi-stable systems.

2.3 Results and Discussion

2.3.1 Positive feedback system is generally more robust against leaky expression, and can achieve higher fold-change, but has higher average basal expression values.

This study set out to investigate 3 questions in regards to how leaky expression affects the properties of the positive and negative inhibition systems. (1) How much leaky expression can a bi-stable system withstand? (2) How does the ratio between the bi-stable states (FC) change with respect to leaky expression values? (3) How does the low molecular state (BL) in these bi-stable systems change with respect to leaky expression? Due to the large biologically relevant parameter space we chose to approach the simulations in three sections.

Our first section consisted of simulating our systems from the parameters described in Table 2.2. We found that the bi-stable region was larger for the positive feedback system with an area of 31.3 compared to the negative inhibition system's area of 0.2. Remember this bi-stable region may need to be mapped back to dimensional space if k scaling factors for a, b, r_1 and r_2 are not equivalent. However, when estimates for the k scaling factors are obtained for particular genetic parts this data can be mapped back for a more precise comparison of the behavior. For now, the assumption that the k scaling factors are equivalent allows for a general comparison between the systems. A larger bi-stable area suggests that the positive feedback switch is more robust against increases in leaky expression.

We also found that the average FC within the bi-stable area was 9.0 x 10^3 for the positive feedback system compared to 5.0 for the negative inhibition system, a 3-order of magnitude difference. Last but not least, the average BL within this area was 2.3 for the positive compared to 0.5 for the negative inhibition system, a one order of magnitude difference. These results are illustrated in Figure 2.4. Also as seen in Figure 2.4, the standard deviations for these average properties are large indicating that there is a wide range of behavior within these bi-stable regions.

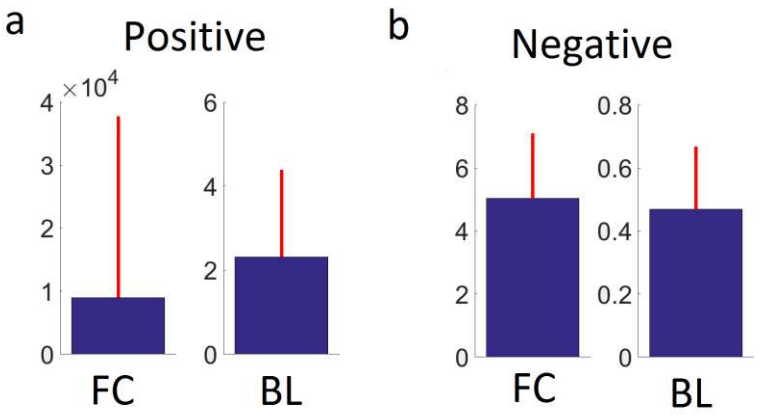

Figure 2.4: Average Behavior. These bar graphs illustrated the predicted mean behavior within this bi-stable area of FC, fold change, and BL, the non-dimensionalized low state. The average value for fold change is larger in the positive feedback system at 9.0 x 10^3 compared to 5.0. BL expression for the negative inhibition system is lower than the positive, at 0.5 compared to 2.3. The error bars are + 1 standard deviation. The x-axis is categorical representing either FC or BL. The y-axis represents how much FC or BL. Panel a represents values for the negative inhibition system whereas panel b represents values for the positive feedback system. (Note: we are working within the non-dimensionalized space.)

To get a view of the predicted behavior within this bi-stable region we created heat maps of the FC and BL values within this bi-stable region (Figure 2.5 a and b). We can see that the FC appears to decrease in both systems as the leaky expression of r_2 or b goes up. We can also see that the BL appears to change inversely to FC (*i.e.* as FC goes up BL goes down). To get an idea

of the magnitude of these changes we created histograms for the FC and BL values within this bi-stable region, as illustrated in Figure 2.5 c and d.

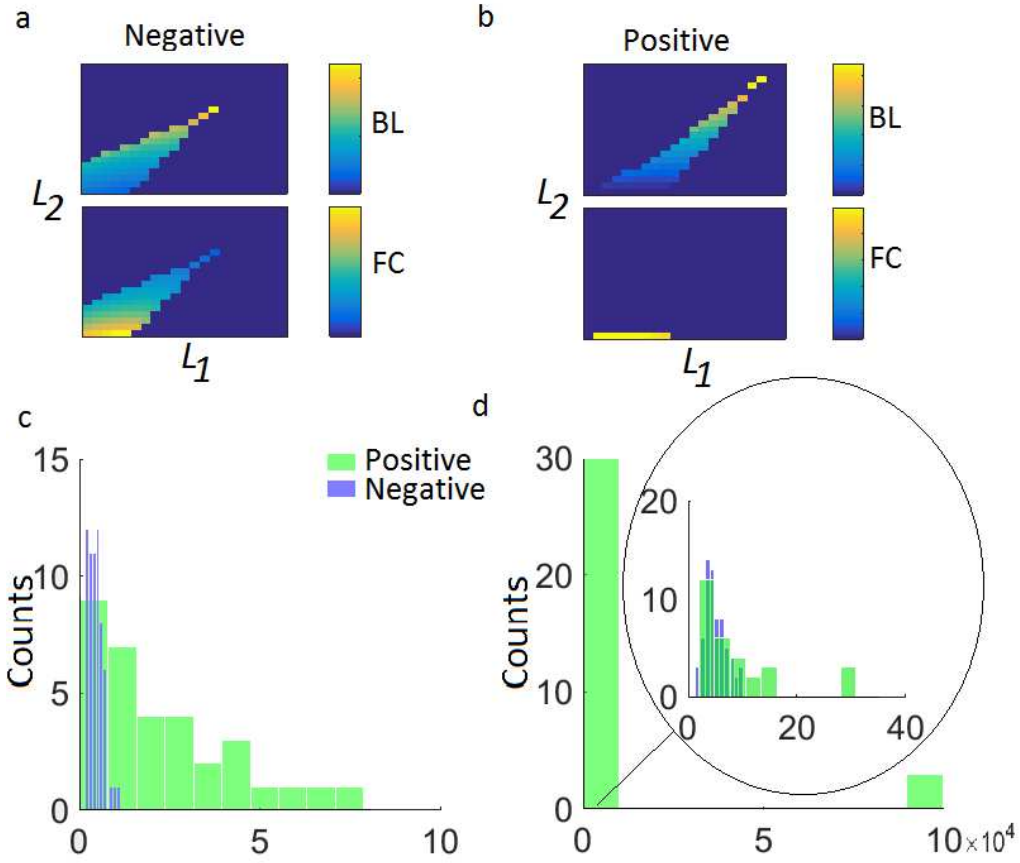

Figure 2.5: Histograms and Heat maps for estimated parameters. Change in key biological features of these switches change within the bi-stable region. Panels a and b describe how the BL, basal level and FC values change within the bi-stable regions. The x-axis for these plots represents L_1 values whereas the y-axis represents L_2 values. Panel c and d describe the distribution of BL and FC within the bi-stable region. The x-axis represents either BL or FC values. The y-axis represents the number of parameter combinations with those FC or BL values (*i.e.* counts). The inset graph in panel d is a histogram of the low cluster of positive feedback system FC values with all the negative inhibition system FC values. The x-axis of this inset graph represents FC where the y-axis represents counts.

In Figure 2.5 d we can see the positive feedback system appears to have two "clusters" of FC values: a low cluster around that of the negative inhibition system and a high cluster around 3-orders of magnitude greater than that of the negative inhibition system. On the other hand, the

negative inhibition system has a tight grouping of low BL values compared to that of the positive feedback system, as illustrated in Figure 2.5 c.

These analyses lead to the following conclusions. The positive feedback system has more potential for higher FC between the states compared to the negative inhibition system. However, the negative inhibition system has more control of the BL of the system. Even though the positive feedback system has a larger tolerance to leaky expression (in terms of its bi-stable space), fold change values within this region are only high for a small region of this bi-stable space where the leaky expression of the bi-stable feedback molecule, *B*, is low.

2.3.2 Parameter Sensitivity

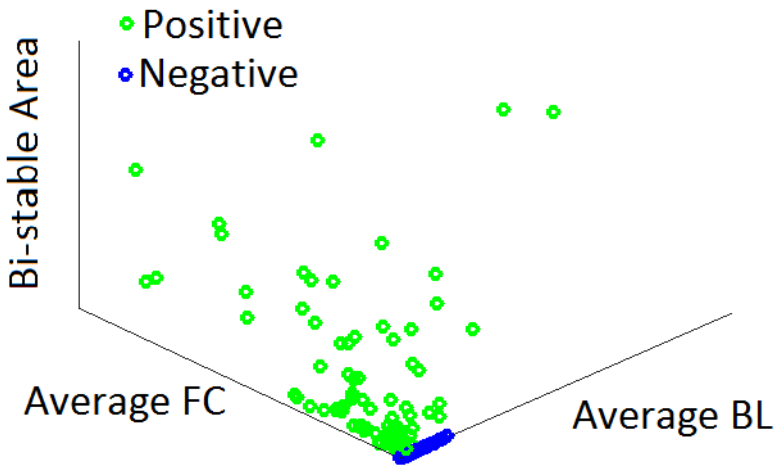

Figure 2.6: Average behavior in parameter sensitivity analysis. Average FC, fold change, and BL, basal level, results for each bi-stable area explored in the parameter sensitivity analysis for parameters represented in Table 2.2. Green points represent the positive feedback system behavior. Blue points represent the negative inhibition system behavior. The x-axis represents the average BL. The y-axis represents the average FC. The z-axis represents the non-dimensionalized leaky expression defined bi-stable area.

All of the analyses up to this point have been for the parameters described in Table 2.2. How dependent are the observations found for the Table 2.2 parameters on the other parameters in the system? Varying these parameters becomes a logical next step. Parameters

were varied to half and double their initial value presented in Table 2.2, with the exception of *n*'s that are set to be 2, 3 or 4. All possible combinations of these parameter configurations were explored. Looking at the leaky expression defined bi-stable areas and corresponding average values of FC and BL within those bi-stable areas we can see in Figure 2.6 that for many parameter combinations the positive feedback system has a larger leaky expression defined bi-stable area, a higher average FC, and a higher BL.

For parameters in Table 2.2, the leaky expression defined bi-stable area for the positive feedback system was larger than that of the negative inhibition system. However, this was not always true for all parameters tested in our parameter sensitivity analysis. In fact, the maximum bi-stable area for the negative inhibition system is 1.1 and the minimum bi-stable area for the positive feedback system is 0.4, if excluding parameter combinations that did not have a bi-stable area. However, as illustrated in Figure 2.5, many of the positive feedback system parameters had a larger bi-stable area compared to the negative inhibition system. Again for parameters in Table 2.2, there was a higher regime of fold change values above 3 orders of magnitude compared to the negative inhibition system. However, in our parameter sensitivity analysis as illustrated in Figure 2.7 a and b, even though the higher regime of FC values for most parameter combination exists, they do not always maintain a three order of magnitude difference when compared to the negative inhibition system. Finally for our parameters in Table 2.2 the positive feedback system did not have as tight of range in BL expression compared to the negative inhibition system, assuming both systems are scaled equivalently in terms of the their non-dimensionalized concentrations. In our parameter sensitivity analysis the positive feedback system also did not have as tight of control over the BL expression. This is illustrated

in Figure 2.7 c and d. This result is also supported by the maximum values of BL expression being around 70 for the positive feedback system compared to around 2 for the negative inhibition system.

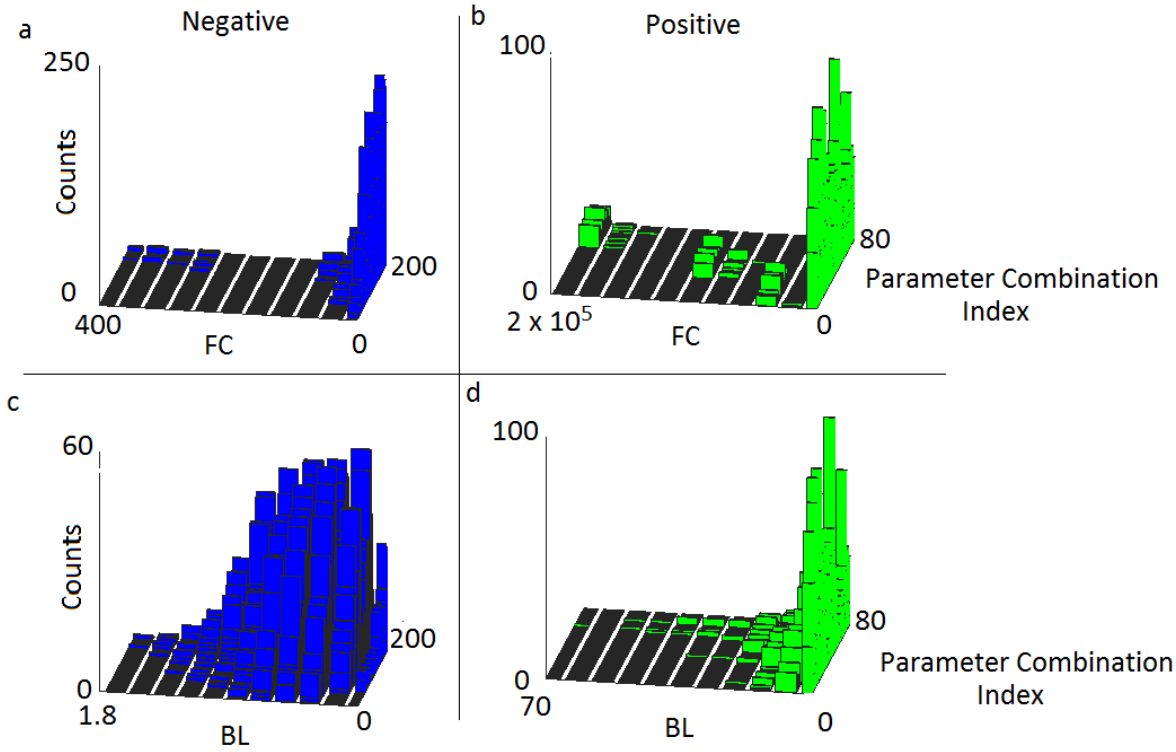

Figure 2.7: Histograms of each parameter combination data. These are histograms for the values within each bi-stable region for each parameter combination. The green histograms are for the positive feedback system whereas the blue histograms are for the negative inhibition system. Panels a and b are for FC values whereas panels c and d are for BL values. The z-axis for each panel represents the number of sampled points for each bar (*i.e.* counts). The x-axis for each panel represents either FC or BL values. The y-axis for each panel represents the parameter combinations being tested. Over 80 different parameter combinations were tested for the positive feedback system and over 200 different parameter combinations were tested for the negative inhibition system.

2.3.3 The larger parameter range

We ran a large range parameter sensitivity analysis to look for any breaks in trends we had observed in the earlier analysis. All combinations of each parameter's more extreme points as well as a few intermediate points (as described in Table 2.3) were examined. A further trend

in the data was observed with the positive feedback system's parameter *T*. Remember *T* is the non-dimensionalized parameter representing the relationship $= \frac{k_d}{k_2}$. Figure 2.8 illustrates how for T << 1 (i.e 0.001) the bi-stable region stretches to very large numbers.

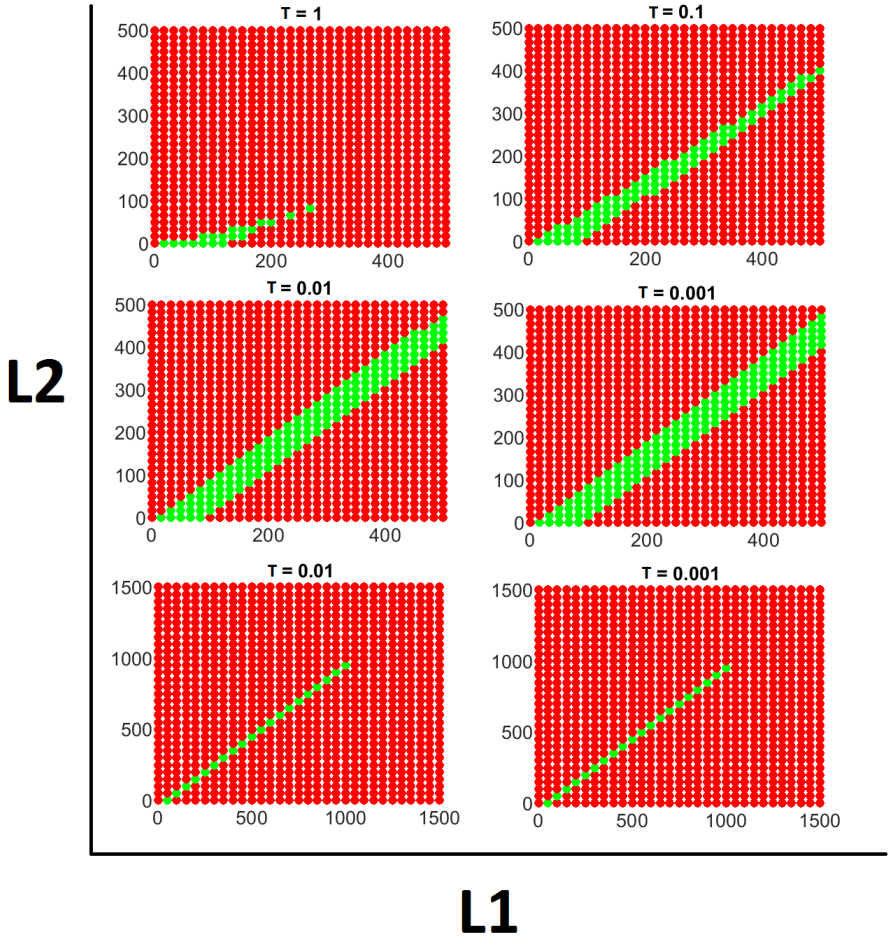

Figure 2.8: Parameter *T*'s effect on the Leaky Expression defined bi-stable region. L_1 and L_2 are the non-dimensionalized leaky expression terms. The green area represents the bi-stable region whereas the red area represents the mono-stable region of parameter space. The x-axis represents L_1 values. The y-axis represents L_2 values. The parameter space for the plots where T = 0.01 and T = 0.001 was sample twice, first for a region of parameter space comparable to the T = 1 and T = 0.1 plots, and second for a larger parameter space showing how the bi-stable region stretched to large values of L_1 and L_2.

Figure 2.9 illustrates the three ideas that have been prevalent throughout this study: (1) the positive feedback system has higher tolerance to leaky expression in terms of its bi-stable

area. (2) The positive feedback system still reaches much larger average values for FC compared to the negative inhibition system. (3) Average BL is still more tightly controlled for the negative inhibition system compared the positive feedback system. These results do show more overlap between the systems which is not surprising considering the range of each parameter in this large parameter sensitivity analysis.

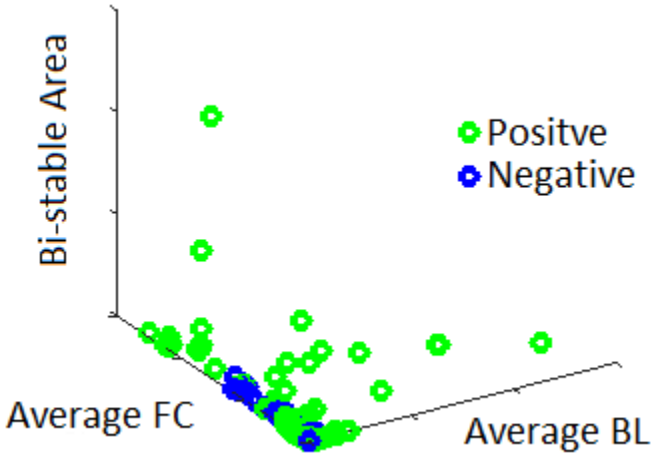

Figure 2.9: Large range average behavior in parameter sensitivity analysis. Average FC, fold change, and BL, basal level, results for each non-dimensionalized leaky expression defined bi-stable area explored in the parameter sensitivity analysis for parameters represented in Table 2.3. Green points represent the positive feedback system behavior. Blue points represent the negative inhibition system behavior. The x-axis represents the average BL. The y-axis represents the average FC. The z-axis represents the leaky expression defined bi-stable area.

Thus our analysis shows that the positive feedback system generally outperforms the negative inhibition system when faced with high levels of leaky expression from the promoter. However, the positive feedback system does suffer from the disadvantage that the levels of basal expression can be much higher than that of the negative inhibition system. Synthetic biologists constructing switches that require a low basal expression in the off state may still prefer the negative inhibition-based architecture. Positive feedback-based switches appear to be (anecdotally) more common in real systems. There are many examples of naturally occurring

bi-stable switches that have been found or hypothesized to exist in nature [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34]. Although negative inhibition systems are contained within the natural switches references here, all of these natural switches contain some form of positive feedback. Our discovery of their increased robustness against leaky promoters in terms of deterministically defined bi-stable area, as well as being able to achieve a higher fold-change, may be one reason why this is so.

REFERENCES

[1] T. S. Moon, E. Clarke, E. Groban, A. Tamsir, R. Clark, M. Eames, T. Kortemme and C. Voigt, "Construction of a genetic multiplexer to toggle between chemosensory pathways in Escherichia coli," *J Mol Biol,* vol. 406, no. 2, pp. 215-227, 2011.

[2] J. Fernandez-Rodriguez, L. Yang, T. Gorochowski, D. B. Gordon and C. Voigt, "Memory and Combinatorial Logic Based on DNA Inversions: Dynamics and Evolutionary Stability," *ACS Synth. Biol.,* vol. 4, no. 12, pp. 1361-1372, 2015.

[3] J. Lengeler, "Genes Can Aslo Be Turned on and off by DNA Recombination," in *Biology of the Prokaryotes*, Abingdon, Oxon, United Kingdom, Marston Book Services Ltd, 1999, pp. 457-458.

[4] P. Hoppe, D. Coutu and T. Schroeder, "Single-cell technologies sharpen up mammalian stem cell research," *Nature Cell Biology,* vol. 16, pp. 919-927, 2014.

[5] G. Rosano and E. Ceccarelli, "Recombinant protein expression in Escherichia coli: advances and challenges," *Front Microbiol.,* vol. 5, no. 172, 2014.

[6] D. Balboa, JereWeltner, S. Eurola, R. Trokovic, KirmoWartiovaara and T. Otonkoski, "Conditionally Stabilized dCas9 Activator for Controlling Gene Expression in Human Cell Reprogramming and Differentiation," *Stem Cell Reports,* vol. 5, pp. 448-459, 2015.

[7] A. Ochab-Marcinek and M. Tabaka, "Transcriptional leakage versus noise: A simple mechanism of conversion between binary and graded response in autoregulated genes," *Physical Review E,* vol. 91, 2015.

[8] B. Wang, M. Barahona and M. Buck, "Amplification of small molecule-inducible gene expression via tuning of intracellular receptor densities," *Nucleic Acids Research,* vol. 1, 2015.

[9] T. Gardner, C. Cantor and J. Collins, "Construction of a genetic toggle switch in Escherichia coli," *Nature,* pp. 339-342, 2000.

[10] D. Chen and A. Arkin, "Sequestration-based bistability enables tuning of the switching boundaries and design of a latch," *Molecular Systems Biology,* vol. 8, no. 620, 2012.

[11] S. Jayanthi and D. D. Vecchio, "Tuning Genetic Clocks Employing DNA Binding Sites," *PLOS,* vol. 7, no. 7, p. e41019, 2012.

[12] R. Milo1, P. Jorgensen, U. Moran, G. Weber and M. Springer, "BioNumbers—the database of key numbers in molecular and cell biology," *Nucleic Acids Research,* vol. 38, no. 1, pp. D750-D753, 2010.

[13] Z. Whichard, A. Motter, P. Stein and S. Corey, "Slowly Produced MicroRNAs Control Protein Levels," *The Journal of Biochemistry,* vol. 286, no. 6, pp. 4742-4748, 2011.

[14] C. Trötschel, S. Albaum and A. Poetsch, "Proteome turnover in bacteria: current status for Corynebacterium glutamicum and related bacteria.," *Microb Biotechnol. ,* vol. 6, pp. 708-719, 2013.

[15] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Research,* vol. 33, no. 10, pp. 3390-3400, 2005.

[16] R. Milo, "What is the total number of protein molecules per cell volume? A call to rethink some published values," *Bioessays,* vol. 35, no. 12, pp. 1050-1055, 2013.

[17] F.-H. Wu, S.-C. Shen, L.-Y. Lee, S.-H. Lee, M. Chan and C.-S. Lin, "Tape-Arabidopsis Sandwich - a simpler Arabidopsis protoplast isolation method," *Plant Methods,* vol. 5, no. 16, 2009.

[18] M. Acar, A. Becskei and A. v. Oudenaarden, "Enhancement of cellular memory by reducing stochastic transitions," *Nature Letters,* vol. 435, pp. 228-232, 2005.

[19] A. Arkin, J. Ross and H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage l-Infected Escherichia coli Cells," *Genetics,* vol. 149, pp. 1633-1648, 1998.

[20] S. Agrawal, C. Archer and D. V. Schaffer, "Computational Models of the Notch Network Elucidate Mechanisms of Context-dependent Signaling," *PLOS,* vol. 5, no. 5, p. e1000390, 2009.

[21] C. Bagowski and J. J. Ferrell, "Bistability in the JNK cascade," *Current Biology,* vol. 11, pp. 1176-1182, 2001.

[22] T. Blauwkamp and A. Ninfa, "Physiological role of the GlnK signal transduction protein of Escherichia coli: survival of nitrogen starvation," *Molecular Microbiology,* vol. 46, no. 1, pp. 203-214, 2002.

[23] J. Das, M. Ho, J. Zikherman, C. Govern, M. Yang, A. Weiss, A. Chakraborty and J. Roose, "Digital Signaling and Hysteresis Characterize Ras Activation in Lymphoid Cells," *Cell,* vol. 136, pp. 337-351, 2009.

[24] J. J. Ferrell and E. Machleder, "The Biochemical Basis of an All-or-None Cell Fate Switch in Xenopus Oocytes," *Science,* vol. 280, 1998.

[25] J. Gavin-Smyth, Y.-C. Wang, I. Butler and E. Ferguson, "A Genetic Network Conferring Canalization to a Bistable Patterning System in Drosophila," *Current Biology,* vol. 23, pp. 2296-2302, 2013.

[26] A. D. Hernday, B. Braaten and D. Low, "The Mechanism by which DNA Adenine Methylase and PapI Activate the Pap Epigenetic Switch," *Molecular Cell,* vol. 12, pp. 947-957, 2003.

[27] T. Hong, J. Xing, L. Li and J. Tyson, "A Mathematical Model for the Reciprocal Differentiation of T Helper 17 Cells and Induced Regulatory T Cells," *PLOS,* vol. 7, no. 7, p. e1002122, 2011.

[28] K. Lai, M. Robertson and D. Schaffer, "The Sonic Hedgehog Signaling System as a Bistable Genetic Switch," *Biophysical Journal,* vol. 86, pp. 2748-2757, 2004.

[29] S. Legewie, N. Bluthgen and H. Herzel, "Mathematical Modeling Identifies Inhibitors of Apoptosis as Mediators of Positive Feedback and Bistability," *PLOS,* vol. 2, no. 9, pp. 1061-1073, 2006.

[30] Y. Li, Y. Li, H. Zhang and Y. Chen, "MicroRNA-Mediated Positive Feedback Loop and Optimized Bistable Switch in a Cancer Network Involving miR-17-92," *PLOS,* vol. 6, no. 10, 2011.

[31] H. Maamar and D. Dubnau, "Bistability in the Bacillus subtilis K-state (competence) system requires a positive feedback loop," *Mol Microbiol,* vol. 56, no. 3, 2005.

[32] E. Ozbudak, M. Thattai, H. Lim, B. Shraiman and A. v. Oudenaarden, "Multistability in the lactose utilization network of Escherichia coli," *Nature Letters,* vol. 427, pp. 737-740, 2004.

[33] J. Pomerening, E. Sontag and J. J. Ferrell, "Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2," *Nature Cell Biology,* vol. 5, no. 4, pp. 346-351, 2003.

[34] N. Trunnell, A. Poon, S. Y. Kim and J. J. Ferrell, "Ultrasensitivity in the regulation of Cdc25C by Cdk1," *Mol Cell,* vol. 41, no. 3, pp. 263-274, 2011.

CHAPTER 3

Quantitative characterization of genetic parts and circuits for plant synthetic biology[1]

3.1 Introduction

Plant synthetic biology promises immense technological benefits, with hope of developing a sustainable bio-based economy through enabling the predictive design of synthetic gene circuits. These circuits are built from quantitatively characterized genetic parts. This characterization presents a significant barrier for plants because of the time required for stable transformation. We describe a method for rapid quantitative characterization of genetic plant parts using transient expression in protoplasts and dual luciferase outputs. We observed experimental variability in transient assays, and developed a mathematical model to describe, and statistical normalization methods to account for, this variability, allowing extraction of quantitative parameters. We characterized over 120 synthetic parts in Arabidopsis and validated our method by comparing transient expression with stably transformed plants. We further tested over 100 synthetic parts in sorghum (*Sorghum bicolor*) protoplasts, showing that

[1] I am a co-first author on this work published in Nature Methods in 2016 [1]. This work is presented here in its entirety over Chapter 3, Appendix A and B to maintain intellectual coherence with permission of the Nature Methods Journal and the Colorado State University Graduate School. Individual contributions are as presented in the Author Contributions. **Author Contributions:** Katherine A. Schaumberg designed and performed experiments and a significant part of the data analysis, and contributed to writing the paper. Mauricio S. Antunes designed and performed experiments and contributed to writing the paper. Tessema K. Kassaw engineered many of the constructs, designed and performed experiments and contributed to writing the paper. Christopher S. Zalewski designed and performed experiments and contributed to writing the paper. Wenlong Xu performed data analysis, developed the camera correction method and contributed to writing the paper. June I. Medford designed experiments, contributed to writing the paper, and supervised the overall project. Ashok Prasad designed data analysis, supervised the computational part of the project, and contributed to and supervised the writing of the paper. All authors contributed to editing the paper and read the final version.

our method works in diverse plant groups. Our approach enables tunable gene circuits to be built in complex eukaryotic organisms.

Synthetic Biology promises to bring new understanding of living organisms while allowing design of predictable biological function. To date, all quantitatively defined gene circuits have been produced in unicellular organisms (bacteria, yeast) or cells in culture [2] [3] [30] [5], raising a question as to whether predictable synthetic gene circuits can be produced in multicellular organisms. Sexual reproduction in multicellular organisms proceeds through meiosis, and plants include development into gametophyte and sporophyte generations, creating a challenge for synthetic genetic circuits. Yet, plant gene circuits with predictable and tunable function could have profound applications towards sustainable life on earth. For example, such circuits could be used to control biofuel production or for optimal production of plant-based biomaterials.

The ability to design and produce such synthetic gene circuits in plants requires a deep understanding of plant biology and rigorous quantitative data. The concerns for producing quantitative predictable functions are myriad. Plants develop continuously, move regulatory molecules between cells and tissues, and control aspects of differentiation by positional information with inputs from their local environment [6]. The challenges are further: develop synthetic genetic parts that are orthogonal, *i.e.*, independent of endogenous regulation [7] [8], and methods for their rapid prototyping to enable rational and predictable design of synthetic circuits. We addressed these challenges by developing principles for rational engineering of synthetic plant genetic parts, along with an experimental and mathematical framework for their quantitative testing.

To produce quantitatively defined and orthogonal genetic parts for plants, we designed synthetic repressors and repressible promoters. Our synthetic repressors consist of well-characterized DNA-binding domains from non-plant proteins and modular repressor motifs from plants. Repressible plant promoters are engineered by rationally inserting DNA sequences recognized by the above repressors into promoters that naturally drive constitutive gene expression in plants (*e.g.*, CaMV35S).

Quantification of genetic parts' input-output characteristics is essential for building circuits with predictable function. However, quantitative analysis of a large number of stably integrated plant genetic parts (*e.g.*, promoters, terminators, UTRs) would require years of work. Current methods for rapid analysis of gene expression in plants, such as particle bombardment, *Agrobacterium* infiltration, VIGS (virus-induced gene silencing)-based systems, and protoplasts [9] [10] [11] [12] could be used. However, most of these methods either lack high throughput capabilities or are difficult to quantify.

To overcome these issues, we scaled up a transient expression assay using plant protoplasts to allow medium throughput (96-well plate) testing of our designed repressible promoters and repressors. To allow simultaneous quantification of both repressor levels and repressible promoter activity, we used dual luciferase outputs. We combined this assay with a rigorous mathematical analysis accounting for significant stochastic factors, thereby developing quantitative analyses for plant genetic parts. Here, we describe our methodology and demonstrate its use to quantitatively characterize over 200 new synthetic plant promoter-repressor pairs. We further show that these parts can be computationally selected and assembled to produce a tunable function *in planta*.

3.2 Results

3.2.1 Building synthetic plant components

To develop our quantitative test system for genetic parts and circuits, we constructed synthetic transcriptional repressor proteins and cognate repressible promoters. To aid in orthogonal function, we designed many synthetic components with previously described elements from bacteria, yeast, and plant viruses. Our synthetic plant transcriptional repressors consist of translational fusions between the yeast Gal4 [13] or the bacterial LexA [14] DNA-binding domains (DBD), shown to function in plants, and known transcriptional repression domains from Arabidopsis proteins (EAR, OFP, BRD) [15] [16] [17] [18]. In addition to the native OFP1 repressor domain, we designed a synthetic repressor domain based on previously characterized sequences and functions of OFP proteins, and designated this OFPx (Appendix A).

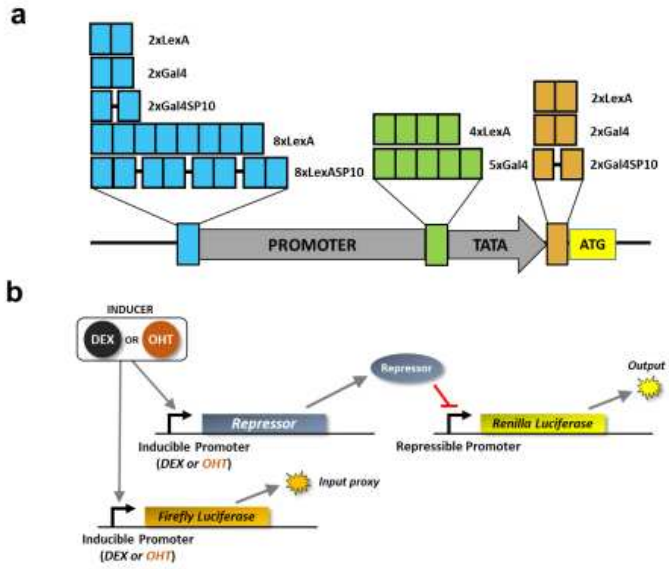

Figure 3.1 Design of synthetic repressible promoters and genetic circuit architecture. (**a**) Diagram of synthetic repressible promoter design containing repressor binding sites (operators) placed upstream (*cyan*), downstream (*orange*), or just upstream of the TATA-box (*green*) in constitutive promoter scaffolds. Number (2x, 4x, 5x, 8x), spacing between binding sites (*SP10*, 10-nt spacer represented by horizontal black bar connecting binding sites), and type of binding sites (Gal4 or LexA) are indicated. (**b**) Genetic circuit architecture used for testing promoter-

repressor combinations in protoplasts. An external inducer (DEX or OHT) activates transcription of a repressor protein as well as Firefly luciferase (F-luc) through the same promoter. The repressor protein represses the constitutively active repressible promoter directing transcription of *Renilla* luciferase (R-luc).

Constitutively active, repressible synthetic promoters (cognates to synthetic repressors) were produced using a scaffold of previously characterized promoters known to drive constitutive plant gene expression [Cauliflower Mosaic Virus (*CaMV35S*), Figwort Mosaic Virus (*FMV*), and Nopaline Synthase (*NOS*)] [19] [20] [21]. To make these promoters repressible, we inserted multiple copies of DNA elements recognized by Gal4 or LexA DBDs at various positions in the CaMV35S, FMV, and NOS promoters (Fig. 3.1a and Table B.1). Our goal was to produce various binding levels for the repressor proteins, and hence tunable repression of promoter activity. We predict these designed promoters will direct constitutive expression of a downstream gene in the absence of their cognate synthetic repressors.

3.2.2 Quantitative testing of plant parts in Arabidopsis

To quantitatively measure the input-output function of our repressors and repressible promoters, we constructed a simple genetic circuit (Fig. 3.1b). With this genetic device, each synthetic promoter is linked to *Renilla* luciferase (R-luc) to provide a quantitative readout of the promoter's behavior. We then modulated the cognate synthetic repressors' expression levels with one of two previously characterized externally applied inducers, dexamethasone (DEX) or 4-hydroxytamoxifen (OHT) [22] [23]. To simultaneously quantify the repressor levels, we added a second copy of the inducible promoter to a second reporter, firefly luciferase (F-luc). Hence, F-luc serves as a proxy for the amount of repressor, as the concentrations of F-luc and repressor

should be proportional to each other. Using dual luciferase reporters, we simultaneously measured repressor production and quantitative repressible promoter function.

Repressible promoters and cognate repressors were cloned into a single plasmid in 128 different pairwise combinations and transiently expressed in Arabidopsis leaf protoplasts (Fig. B.1). We varied repressor levels by changing the concentration of DEX or OHT, and measured both F-luc and R-luc activity in the same sample with a commercially available dual-luciferase assay (Promega Co.) and single photon ICCD Camera (Stanford Photonics, Inc.). To increase assay throughput, we modified an Arabidopsis leaf protoplast assay (as presented in Appendix A) to allow testing promoter-repressor combinations, with multiple inducer concentrations, in a 96-well plate format.

With increasing inducer concentrations, we expect increasing F-luc levels (input, repressor concentration), coupled with decreasing R-luc levels (output, promoter activity). Initial results showed the expected trend, but had large variability (noise) between transient assays (Fig. 3.2a). Because accurate quantitative characterization of genetic parts requires high reproducibility and comparability across components and experiments, we investigated the noise's source(s).

3.2.3 Analysis of stochastic and experimental variability

To determine whether data are comparable across genetic circuits, we examined the basal level of F-luc luminescence (*i.e.*, without inducer added). For all circuits controlled by a given inducer (DEX or OHT), basal F-luc levels should be the same. Basal F-luc data indicate higher variability between plasmids than between technical replicates (Fig. 3.2a). This suggests that the protoplast assays are subject to variability from different batches of protoplasts,

transformation efficiency, and different genetic circuits, among other possible effects. We discovered that one source of noise in our data was systematic, and related to our luciferase image collection method. To correct for these imaging errors, we developed a geometric method (Appendix B.1, B.2 and Figs. B.2 and 3). Next, we addressed the sources of variability and developed methodology to mitigate their effects.

From the experimental design, we expect three major contributors to noise in our system.

1) Within-plate variation, *i.e.,* random variation arising from 96-well plate assay procedures, such as pipetting.
2) Between-transformation variation, *i.e.,* variation arising from processes involved in protoplast transformation, such as variations in transformation efficiency from different plasmids.
3) Between-batch variation, *i.e.,* variation arising from processes involved in preparing each protoplast batch. This includes variations in the leaf tissue health, different types of cells isolated, shear stress effects during protoplast pipetting and centrifugation, and slight variations from enzymatic supplies.

We designed experiments to isolate and quantify these three potential sources of noise. We used a test plasmid (beta plasmid) with all elements found in our promoter-repressor genetic device, except the repressor, and transformed protoplasts with either a DEX-inducible or OHT-inducible gene circuit. We collected luminescence data with no inducer added, and repeated the experiment using protoplasts batches prepared on three different days. For each circuit, in the absence of any noise, all wells should display identical F-luc and R-luc luminescence.

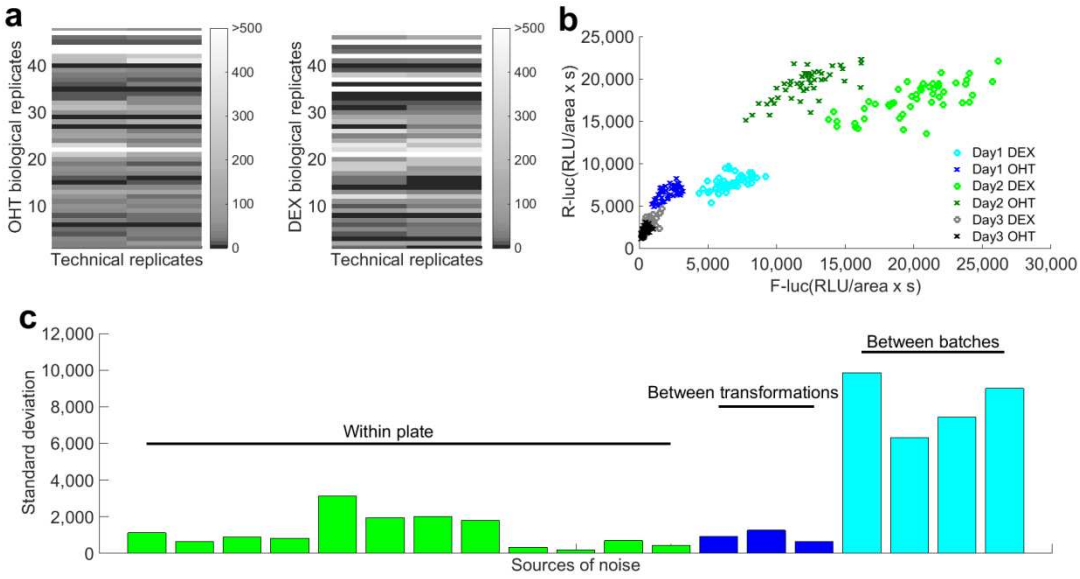

Figure 3.2 Analysis of noise in the protoplast data. (**a**) Greyscale heat map represents measured F-luc luminescence intensities with no inducer present (*i.e.*, basal level expression) for two technical (horizontal axis) and different biological (vertical axis) replicates. OHT- and DEX-inducible promoter-repressor pairs are plotted separately. (**b**) Scatter plot of R-luc and F-luc luminescence for a repressible promoter without its repressor (beta plasmid), measured on different days. *Open circles*, DEX-inducible promoter; *small x's*, OHT-inducible promoter. The same repressible promoter was used in both circuits. (**c**) Standard deviation of the three different noise sources: within a 96-well plate (12 samples), between different transformations (3 samples), and between different batches of protoplasts (4 samples).

The predominant source of noise originates from distinct batches of protoplasts prepared on different days (Fig. 3.2b). Luminescence values of each batch form distinct and well-separated clusters, and variations within each cluster are smaller than those between clusters. Plotting the standard deviation from each source (Fig. 3.2c) confirms that the most variance comes from different days' preparation (batches) of protoplasts, which we define as a 'batch effect'. A plot of the average R-luc vs. F-luc luminescence displays strong linear trends, as expected if the batch effect can be represented by a random multiplicative factor that is the same for both F-luc and R-luc (Fig. 3.3a and Appendix B.3).

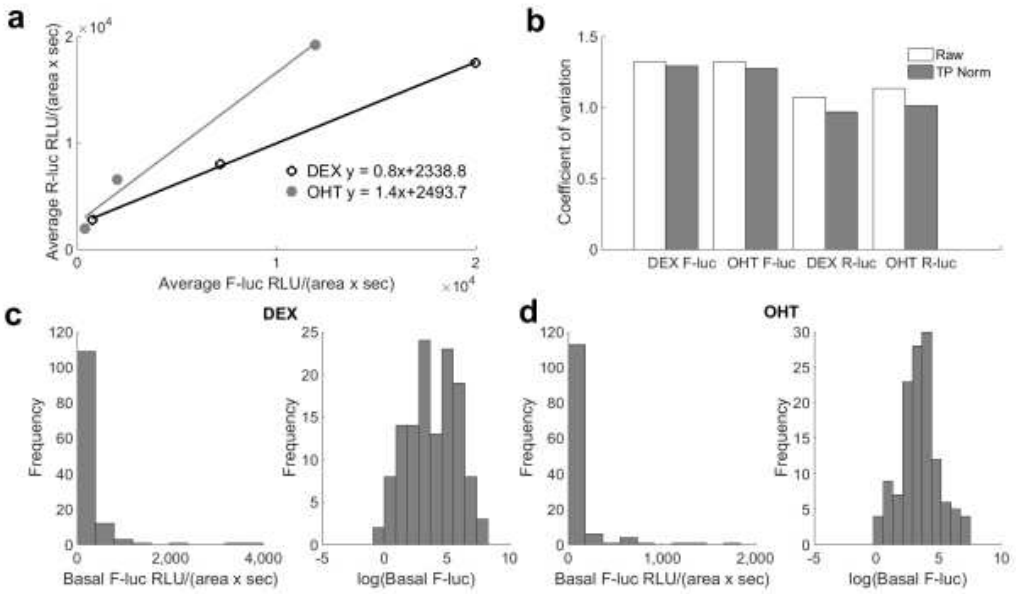

Figure 3.3 Analysis of the variation from different protoplast batches. (**a**) Plot of average R-luc and F-luc luminescence values for DEX- (*open circles*) and OHT-inducible (*closed circles*) beta plasmids from different days. Lines represent linear fits. (**b**) F-luc and R-luc coefficient of variation from beta plasmids, without (*Raw*) and with (*TP Norm*) normalization by the total protein content in the well. (**c, d**) Histogram of all basal (*i.e.*, no added inducer) F-luc luminescence values, plotted as RLU/(area x sec) (*left graph*) and on a log scale (*right graph*), for DEX-based plasmids (**c**) and OHT-based plasmids (**d**).

While our data clearly show that variation from protoplast batches is the greatest noise source, we have not identified its origin. Because our protoplasts are prepared from leaves, each preparation could contain distinct compositions of differentiated leaf cells (*e.g.*, mesophyll, palisade parenchyma, bundle sheath) produced from plants that experience micro-climatic variations. While all protoplasts are pooled and treated equally, our data represent a bulk measurement of protoplast populations in an individual well. As such, different protoplast populations may be represented in each preparation, giving rise to a batch effect. Our data do show our experiments are carefully performed, as the within-plate variation has an approximately normal distribution, with a small standard deviation. Differences arising from

intrinsic plasmid properties were also found to be minimal. Based on the above analysis, we constructed a quantitative model to determine the input-output characteristics of the promoter-repressor pairs. In accordance with the experimental findings, our quantitative model incorporates the batch variability as a multiplicative factor, and takes the experimental setup into account.

3.2.4 Mathematical model and normalization of batch effect

Our experiments are performed in 96-well plates, and each plasmid is tested with a different level of the inducer (DEX or OHT) in each well, with the corresponding F-luc and R-luc luminescence measured. In all cases, the indices *ij* refer to the *j*-th well from those wells with the *i*-th plasmid. Concentrations are expressed in molecules per well and RLU represent Relative Luciferase Units/(area x sec). We first describe the experimental data with symbols as follows:

1. Repressor concentration = R_{ij}
2. R-luc concentration = $[Rluc]_{ij}$
3. F-luc concentration = $[Fluc]_{ij}$
4. R-luc luminescence in RLU = L_{ij}
5. F-luc luminescence in RLU = F_{ij}

We want to determine the quantitative relationship between the repressor concentration and the R-luc expression, controlled by the constitutively active repressible promoter. It is standard to assume that this relationship is represented by a Hill function [30] [24]. Hence, for a single plasmid in a protoplast we can write:

$$[Rluc]_{ij} = \frac{\beta_i}{1+\left(\frac{R_{ij}}{K_i}\right)^{n_i}} \quad 3.1$$

Here, β_i represents the maximal expression of the R-luc protein with no repressor (*i.e.*, promoter strength), while K_i is the repressor concentration required for half-maximal expression of R-luc, and n_i is the Hill coefficient. In order for this equation to correspond with the experimental data, we need to make two transformations. First, we need to scale up to the entire well. Second, we need to express this relationship in terms of luminescence, which is what we experimentally measure, instead of concentration. To scale up to the entire well, we multiply both sides of this equation by $\alpha_i N_{ij}$, where α_i represents the batch variability factor and N_{ij} represents the total number of plasmids in the *j*-th well of the *i*-th plasmid. To transform to luminescence units, denoted by L_{ij} for R-luc and F_{ij} for F-luc, we use the fact that luminescence is proportional to concentration (Fig. B.4 and Appendix B.4), and multiply both sides by the corresponding proportionality factor. This gives the following equation:

$$L_{ij} = \frac{B_i}{1 + \left(\frac{F_{ij}}{H_i}\right)^{n_i}} \quad 3.2$$

Here, $F_{ij} = C_1 \alpha_i N_{ij} \tilde{C} R_{ij}$, where C_1 is the concentration-luminescence proportionality factor and $\tilde{C}$ is a proportionality factor between the repressor concentration and F-luc concentration (both of which are controlled by the same promoter).

The parameter H_i represents the half-maximal whole-well R-luc luminescence, while the parameter B_i represents the whole-well maximal R-luc luminescence of the *i*-th plasmid. Thus a

best-fit estimate of B_i is given by $B_i = C_2\langle N_{ij}\rangle_j \alpha_i \beta_i$, where the *j*-subscript on the angled brackets indicates a mean over the *j*-well index, *i.e.*, over the wells associated with the *i*-th plasmid. Similarly, $H_i = C_1\tilde{C}\langle N_{ij}\rangle_j \alpha_i K_i$.

The unknown multiplicative batch-effect, α, complicates comparison of the repressible promoter strength, β, between plasmids. Thus, we considered normalization methods that remove or reduce the effect of this parameter. We first tested whether the batch effect can be removed by normalization with the total protein content per well, but this resulted only in a minor difference between the variability of raw versus normalized data from different batches (Fig. 3.3b), thus failing to account for the batch effect.

Next, we hypothesized that the batch variability factor, α_i, is related to the protoplasts' preparation. Thus, $\alpha_i N_{ij}$ describes the plasmid copy number in viable protoplasts in the *j*-th well of the *i*-th plasmid. The basal luminescence distribution in the absence of inducer, *i.e.*, $F_{i1} = C_1 \alpha_i N_{i1} \gamma_i$, where γ_i is the basal expression of the inducible promoter (*j* = 1 corresponds to wells with no inducer), is proportional to the distribution of $\alpha_i N_{ij}$, and is fit well by a log-normal distribution (Fig. 3.3c,d), further supporting a multiplicative source for the batch variation. Now, we define a normalization factor:

$$\lambda_i = \frac{\langle F_{i1}\rangle_r}{\langle F_{i1}\rangle_{ir}} = \frac{\langle N_{i1}\rangle_r \alpha_i}{\langle N_{i1}\rangle_{ir}\langle \alpha_i\rangle_{ir}} \approx \frac{\alpha_i}{\langle \alpha_i\rangle_{ir}} \qquad 3.3$$

Here, the mean is taken over the subscripts on the angled brackets (*r* refers to technical replicates). We have assumed that the distribution of N_{i1} and α_i are independent of each other, and that $\langle N_{i1}\rangle_r \approx \langle N_{i1}\rangle_{ir}$, *i.e.*, there is little variation between the means in each transformation. Both assumptions are reasonable. We divide F_{ij} and L_{ij} for each well by this

normalization factor, which replaces the varying batch effect factor α_i with a constant, *i.e.*, its mean $\langle \alpha_i \rangle_{ir}$.

The preceding analysis predicts that the variability in the F-luc luminescence at each value of inducer should arise mainly from the variability between different protoplast batches and should therefore be significantly reduced by the normalization factor, λ_i. In agreement with this, for each inducer value, the normalized data have a significantly smaller coefficient of variation than the raw data (Fig. 3.4a,b), showing that the normalization method does indeed reduce or even eliminate the batch effect. Tests of the normalization method against simulated data (Appendix B.5 and Fig. B.5) show that it reduces the variability in estimating the parameters B_i and H_i, and makes them more comparable across plasmids.

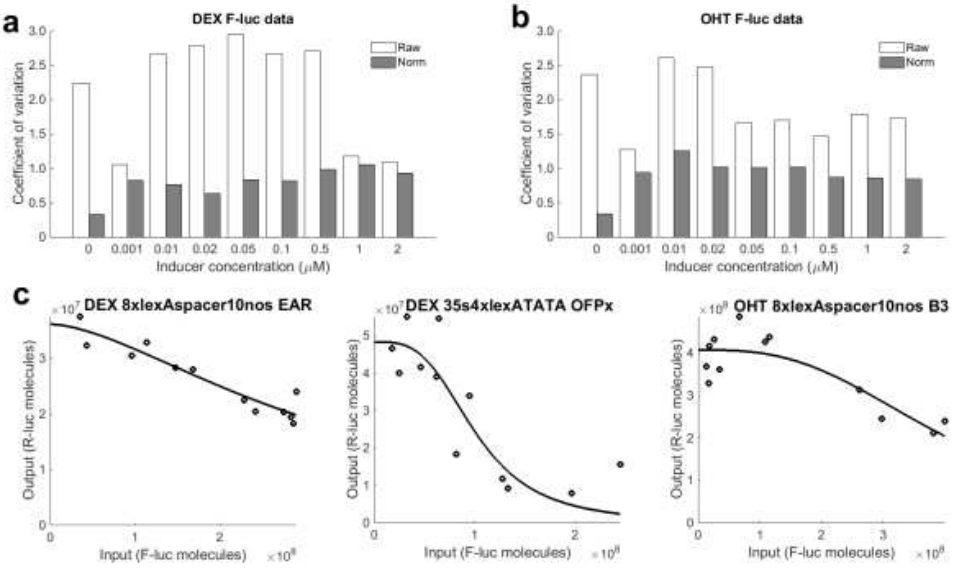

Figure 3.4 Effect of normalization on the Arabidopsis dataset. (a, b) Coefficient of Variation (COV) of experimental F-luc luminescence values for different inducer levels, in DEX-inducible plasmids (a) and in OHT-inducible plasmids (b). COV of normalized data (*Norm*) is significantly reduced and more uniform across inducer levels [F-test (two-sample F-test for equal variances), *P*-value = 0.05]. (c) Representative data and curve fits of some of the best performing promoter-repressor pairs. These promoter-repressor pairs were among those that satisfied

established criteria for a functional pair, *i.e.*, luminescence above the threshold, fold change greater than 1.3, and Hill coefficient between biologically reasonable limits (*e.g.* 0 and 6).

We fit the normalized data using a common nonlinear least squares package in Matlab (http://www.mathworks.com/products/matlab/) and implemented further quality control steps to remove protoplast assays that appeared to have failed, or promoters that did not function (Appendix B.6 and Fig. B.6). From 128 gene circuits tested in Arabidopsis protoplasts, 42 met all the functionality criteria. Figure 3.4c shows representative fits of some of our best repressor and promoter pairs, and the quantitative parameterization allows comparison of different circuits. We found that our synthetic repressor motifs worked but some appear to favor different pairings; for example, LexA appeared to favor placement with an OFP motif.

3.2.5 Validating the model in a different plant family: sorghum

To test our method's generality, we quantitatively characterized promoter-repressor pairs in another species in a diverse plant grouping, the monocots. Over 100 synthetic promoter-repressor pairs for monocots were constructed (Appendix A), and characterized in sorghum protoplasts using a similar protocol used for Arabidopsis. In designing sorghum components, we used an intron in the 5'-UTR to help assure function in monocots [25]. As in Arabidopsis, the basal F-luc expression shows significant difference across constructs (Fig. 3.5a), indicating substantial batch variability. We found the F-luc luminescence at every inducer level was log-normally distributed, as in Arabidopsis. These results suggest that the multiplicative model for luminescence proposed in Equation 3.1 and 2 also applies to sorghum protoplasts. We found for sorghum protoplasts, as with Arabidopsis, the normalization scheme leads to a substantial decrease in the coefficient of variation for noisy data (Fig. 3.5b).

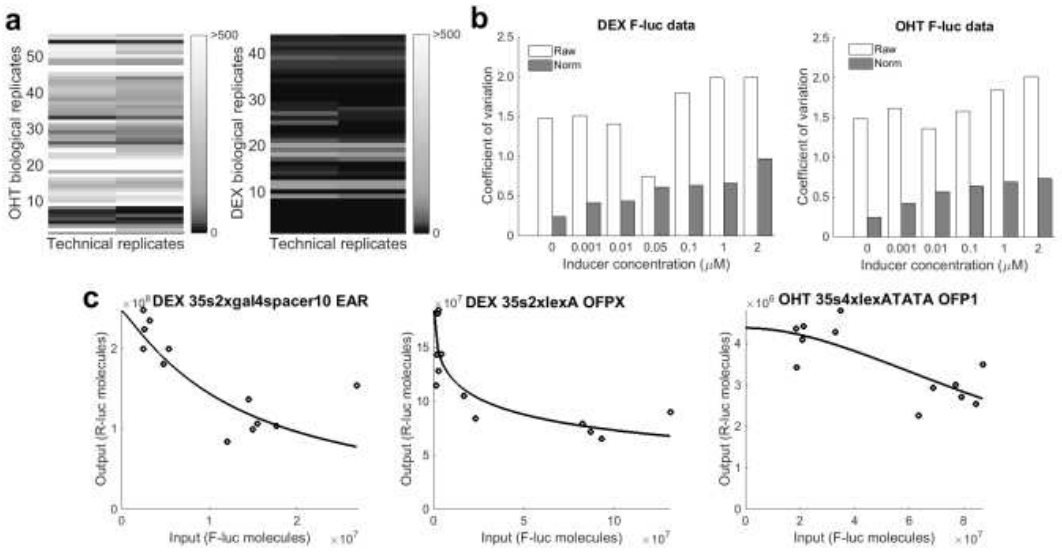

Figure 3.5 Effect of normalization on the sorghum dataset. (**a**) Greyscale heat map represents measured F-luc luminescence intensities with no inducer present (*i.e.*, basal level expression) for two technical (horizontal axis) and different biological (vertical axis) replicates. OHT- and DEX-inducible promoter-repressor pairs are plotted separately. (**b**) COV of experimental F-luc luminescence values for different inducer levels, in DEX-inducible plasmids (*left*) and in OHT-inducible plasmids (*right*). COV of normalized data (*Norm*) is significantly reduced for all inducer concentrations (except 0.05 μM, where there is a non-significant decrease) and more uniform across inducer levels (F-test, *P*-value = 0.05). (**c**) Representative data and curve fits of some of the best performing promoter-repressor pairs. These promoter-repressor pairs were among those that satisfied established criteria for a functional pair, *i.e.*, luminescence above the threshold, fold change greater than 1.3, Hill coefficient between biologically reasonable limits.

Using our method in sorghum protoplasts, of the 112 gene circuits tested, 41 met all the criteria (Fig. 3.5c and Appendix B.6). We also found analogous patterns in the data, suggesting that the design principles for building synthetic genetic components in plants may apply across species (see Discussion).

3.2.6 Validating predictions with stably transformed plants

Our transient assays and the derived model should provide quantitative data to predict the performance of a given genetic circuit in stably transformed plants. Therefore, we

compared the performance of promoter-repressor pairs in transient assays with that in stably transformed Arabidopsis plants (Fig. 3.6a). Transgenic Arabidopsis plants were produced with three different promoter-repressor genetic circuits: DEX 35S2xLexA EAR, OHT nos2xGal4 EAR, and DEX 2xGalNos EAR (Appendix A). At least 20 independent transgenic lines for 11 distinct promoter-repressor genetic circuits were analyzed and screened by Mendelian segregation for one copy of the introduced transgene (Table B.2). Plants from one transgenic line each for three of the genetic circuits (DEX 35S2xLexA EAR, OHT nos2xGal4 EAR, and DEX 2xGalNos EAR) were used as a source for protoplasts with the gene circuit stably integrated (Fig. 3.6a).

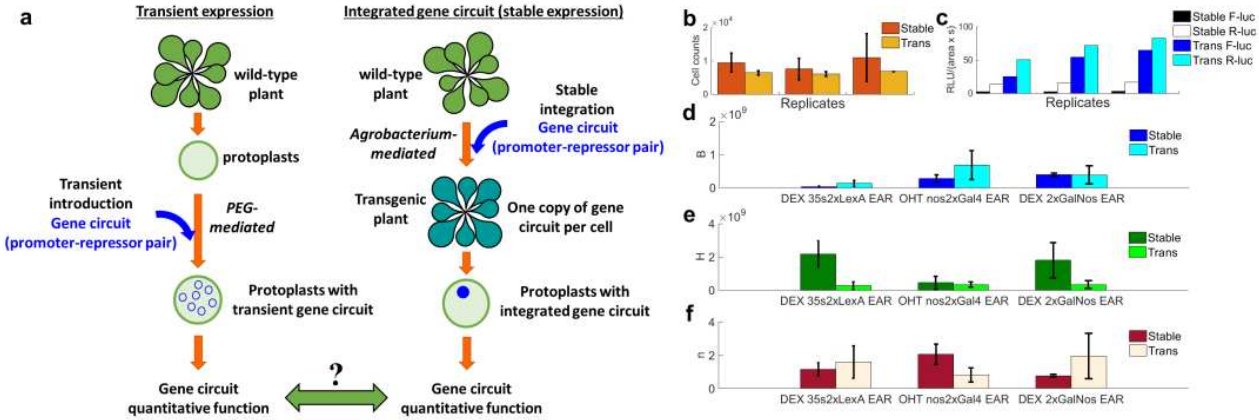

Figure 3.6 Experimental design and validation of predictions in stably transformed plants. (a) Schematic of experimental design comparing quantitative function of synthetic promoters and repressors analyzed with transient expression (*left*) or stable transgenic expression (*right*). For transient expression, the genetic circuit is introduced into protoplasts via PEG-mediated transformation. For stable expression, the genetic circuit is first integrated into the plant's chromosome via Agrobacterium-mediated transformation. Protoplasts are then prepared from the stably transformed plants. Quantitative parameters for the promoter-repressor pair are compared from transient and stably transformed processes. (b) Protoplast cell counts at the time of luciferase imaging from the transient assay (*Trans*) and from the stable integration (*Stable*), for 2-5 replicates. (c) F-luc and R-luc luminescence values from stable transformants or transient assays of the same replicates shown in b. (d) Estimates of the promoter strength parameter *B* for stably transformed plants (*Plant*) and transient expression (*Trans*) in protoplasts for three promoter-repressor pairs. (e) Estimates of the half-maximal repressor

expression *H*, and (f) estimates of the Hill coefficient *n*, for the same three pairs. Error bars are standard deviations. Significance was determined using a two-sample t-test.

We then compared the quantitative function (behavior) of our gene circuits in transient assays to their behavior when stably integrated into the genome. Protoplasts were prepared from the transgenic lines containing a specific promoter-repressor gene circuit; this same promoter-repressor gene circuit was transiently introduced into protoplasts from wildtype plants (Fig. 3.6a). We induced repressor expression in both protoplast sets and determined the parameters describing their quantitative function. Protoplasts prepared from plants with stably integrated gene circuits consistently produced less F-luc and R-luc luminescence (Fig. 3.6b). One explanation for this difference is that circuit copy number varies between each protoplast set, with the transient assays likely containing more. To account for the variation, we normalized the data to correct for these differences (Appendix B.7 and 8 and Fig. B.7).

We then fit the data to the Hill function of Equation 3.1 and 2 above, and compared the calculated parameters for each gene circuit between stable and transient assays (Fig. 3.6c-e). Relative promoter strengths (*B*) predicted by the transient assays match those found in stable plants (Fig. 3.6c). For all three parameters, error bars of the estimates typically overlap, and for these estimates the differences were found not to be significant (*P*-value = 0.1). We further scrutinized our data by carrying out a bootstrapping statistical analysis (Appendix B.9). This analysis confirmed the agreement between the results for the transient assay and stable transgenic plants (Fig. B.8). Moreover, it is known that Agrobacterium-mediated transformation (used to produce transgenic plants) produces random integrations in the plant chromosome. Despite this additional variability, the broad agreement we found in quantitative parameters

between protoplasts from stable and transient assays suggests that our approach provides reliable prediction for functional synthetic plant parts.

3.3 Discussion

Our detailed analysis of promoter-repressor pairs in isolated plant cells provides a basis to quantitatively define relationships between genetic elements, an essential first step towards producing predictable genetic circuits in multicellular organisms such as plants. Our results showed that quantitative data obtained from a rapid transient protoplast assay, when combined with rigorous noise analysis and mathematical modeling, allows fast and quantitative parameter estimation of synthetic gene parts. We demonstrated that it is possible to reliably assess repressor strength using the suite of experimental methods, and these quantitative measures were shown to be valid in both eudicots (Arabidopsis) and monocots (sorghum). The results support our mathematical model as a rational depiction of quantitative experimental data. By comparing our quantitative characteristics with synthetic promoter architecture, we can formulate the first design principles for constructing synthetic gene components for plants. Interestingly, we found commonality in these principles, suggesting they are general for both eudicots and monocots.

In designing our synthetic elements we use and expand upon concepts developed by others. First, our synthetic repressors were produced using well-known DNA-binding domains combined with modular repressor domains from plant genes. The success of this design suggests a path to produce other synthetic components such as activators. We designed our synthetic repressible promoters using well-characterized promoters as scaffolds, into which we placed DNA elements for our repressors. Quantitative analyses (Appendix B.10) suggested that

the CaMV35S promoter forms the best scaffold, even though it is not the strongest constitutive promoter in either data set.

Our data also suggest that position of the repressor binding site affects the maximum expression of the synthetic promoter in the same manner for both Arabidopsis and sorghum (Fig.B.9 and Appendix B.10). Specifically, repressor-binding sites that are positioned near the TATA box decrease the maximum strength of the promoter. One explanation is that the bound repressor, even at low concentration in our "Off" state, leads to steric exclusion of RNA polymerase. The data further suggest synergies between the DNA-binding domains and the scaffold into which these are inserted. For example, in Arabidopsis, LexA performed better when paired with a CaMV35S scaffold, whereas Gal4 performed better in a NOS promoter scaffold (though less substantially). We also found synergies between the DNA-binding domains and repressor motifs. In Arabidopsis, the LexA DNA-binding domain performed better with an OFP motif. The synthetic repressor built with Gal4 showed improvement (though less substantially) with B3. In contrast, the EAR repressor motif worked well with both LexA and Gal4 DNA-binding domains. We were not able to fully determine the impact of spacer sequences between the repressor binding sites; further studies are needed to determine their effect.

In conclusion, our detailed analysis of synthetic promoter-repressor pairs in isolated plant cells provides a basis to quantitatively define relationships between genetic components, an essential first step towards engineering tunable function in multicellular organisms such as plants. The procedures described here are immediately applicable for the development of comprehensive quantitatively characterized libraries of synthetic plant gene parts, in principle

for any plant species. The quantitative parameters of each promoter-repressor pair can be then used for *in silico* suitability testing of its use in more complex genetic circuits, such as a genetic toggle switch and feedback circuits.

REFERENCES

[1] K. A. Schaumberg, M. S. Antunes, T. K. Kassaw, W. Xu, C. S. Zalewski, J. I. Medford and A. Prasad, "Quantitative characterization of genetic parts and circuits for plant synthetic biology," *Nature Methods,* vol. 13, pp. 94-100, 2016.

[2] S. e. a. Kiani, "CRISPR transcriptional repression devices and layered circuits in mammalian cells," *Nat. Methods,* vol. 11, pp. 723-726, 2014.

[3] K. e. a. Rinaudo, "A universal RNAi-based logic evaluator that operates in mammalian cells," *Nat. Biotechnol.,* vol. 25, pp. 795-801, 2007.

[4] T. S. Gardner, C. R. Cantor and J. J. Collins, "Construction of a genetic toggle switch in Escherichia coli," *Nature,* vol. 403, pp. 339-342, 2000.

[5] L. You, R. S. Cox, R. Weiss and F. H. Arnold, "Programmed population control by cell-cell communication and regulated killing," *Nature,* vol. 428, pp. 868-871, 2004.

[6] T. A. Steeves and I. M. Sussex, Patterns in Plant Development, 2nd edn. ed., New York: Cambridge University Press, 1989.

[7] J. B. Lucks, L. Qi, W. R. Whitaker and A. P. Arkin, "Toward scalable parts families for predictable design of biological circuits," *Curr. Opin. Microbiol.,* vol. 11, pp. 567-573, 2008.

[8] A. L. Slusarczyk, A. Lin and R. Weiss, "Foundations for the design and implementation of synthetic genetic circuits," *Nat. Rev. Genet.,* vol. 13, pp. 406-420, 2012.

[9] J. Kim, P. G. Klein and J. E. Mullet, "Synthesis and turnover of photosystem II reaction center protein D1. Ribosome pausing increases during chloroplast development," *J. Biol. Chem. ,* vol. 269, pp. 17918-17923, 1994.

[10] T. e. a. Asai, "MAP kinase signalling cascade in Arabidopsis innate immunity," *Nature,* vol. 415, pp. 977-983, 2002.

[11] H. W. e. a. Mewes, "Overview of the yeast genome," *Nature,* vol. 387, pp. 7-65, 1997.

[12] T. M. Klein, E. D. Wolf, R. Wu and J. C. Sanford, "High-velocity microprojectiles for delivering nucleic acids into living cells," *Nature,* vol. 327, pp. 70-73, 1987.

[13] E. Giniger, S. M. Varnum and M. Ptashne, "Specific DNA-binding of Gal4, a positive regulatory protein of yeast," *Cell,* vol. 40, pp. 767-774, 1985.

[14] M. Schnarr, P. Oertelbuchheit, M. Kazmaier and M. Grangerschnarr, "DNA-binding properties of the LexA repressor," *Biochimie ,* vol. 73, pp. 423-431, 1991.

[15] S. Wang, Y. Chang, J. Guo and J.-G. Chen, "Arabidopsis Ovate Family Protein 1 is a transcriptional repressor that suppresses cell elongation," *Plant J.,* vol. 50, pp. 858-872, 2007.

[16] S. e. a. Wang, "Arabidopsis ovate family proteins, a novel transcriptional repressor family, control multiple aspects of plant growth and development," *PloS one,* vol. 6, p. e23896, 2011.

[17] M. Ikeda and M. Ohme-Takagi, "A novel group of transcriptional repressors in Arabidopsis," *Plant Cell Physiol.,* vol. 50, pp. 970-975, 2009.

[18] M. Ohta, K. Matsui, K. Hiratsu, H. Shinshi and M. Ohme-Takagi, "Repression domains of class II ERF transcriptional repressors share an essential motif for active repression," *Plant Cell,* vol. 13, pp. 1959-1968, 2001.

[19] J. Odell, F. Nagy and N. Chua, "Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter," *Nature,* vol. 313, pp. 810-812, 1985.

[20] M. Sanger, S. Daubert and R. Goodman, "Characteristics of a strong promoter from figwort mosaic virus: comparison with the analogous 35S promoter from cauliflower mosaic virus and the regulated mannopine synthase promoter," *Plant Mol. Biol.,* vol. 14, pp. 433-443, 1990.

[21] C. Shaw, G. Carter and M. Watson, "A functional map of the nopaline synthase promoter," *Nucleic Acids Res.,* vol. 12, pp. 7831-7846, 1984.

[22] T. Aoyama and N. Chua, "A glucocorticoid-mediated transcriptional induction system in transgenic plants," *Plant J.,* vol. 11, pp. 605-612, 1997.

[23] J. Zuo, Q. Niu and N. Chua, "An estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants," *Plant J.,* vol. 24, pp. 265-273, 2000.

[24] U. Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits., Boca Raton, FL: Chapman & Hall/CRC, 2007.

[25] D. Mascarenhas, I. Mettler, D. Pierce and H. Lowe, "Intron-mediated enhancement of heterologous gene expression in maize," *Plant. Mol. Biol.,* vol. 15, pp. 913-920, 1990.

CHAPTER 4

The Computational Design of Two Different Bi-stable Switches

4.1 Introduction

This chapter describes work done for computational prediction of the properties of bi-stable switches based on both negative inhibition as well as on positive feedback. Computational predictions for the negative inhibition systems were made using the library of plant parts constructed as part of the work presented in Chapter 3. As discussed in Chapter 1, two repressible promoters can be utilized to make a genetic toggle switch. We used combinations of these repressible promoters, and using the quantitative parameters that we estimated in Chapter 3, we simulated the possible operation of the combined circuit. We defined the signature of bi-stability, followed by identification of parts with the ideal properties to generate this bi-stability. To deal with the high level of noise in the system, we designed a novel bootstrap method to estimate the probability of a particular combination existing in the bi-stable region of parameter space.

Unlike the negative inhibition system, we did not measure the characteristics of the individual parts of the positive feedback system, a version of which was built by the Medford lab. Therefore, for the positive feedback system, an ODE-based parameter sensitivity analysis was carried out to predict ideal part properties. First a set of equations was constructed to simulate the system *in silico*. Then a non-dimensionalization was performed to reduce the dimensionality of the problem. Next, ideal parameter values were identified. This was followed by the development of hypotheses on how each parameter is connected to the particular

genetic parts. Finally, a list of ideal part properties was identified for the positive feedback system design.

4.2 Negative Inhibition System Methods

4.2.1 Circuit Design

A bi-stable switch is a genetic circuit capable of being set to two different states with the ability to stay in either state given no outside influence. Such a circuit was first synthetically built by Gardner et al. in the form of a dual repression circuit [1]. Our version of this circuit (described in Figure 4.1) followed Gardner et al. approach in having two different repressors that repress each other [98].

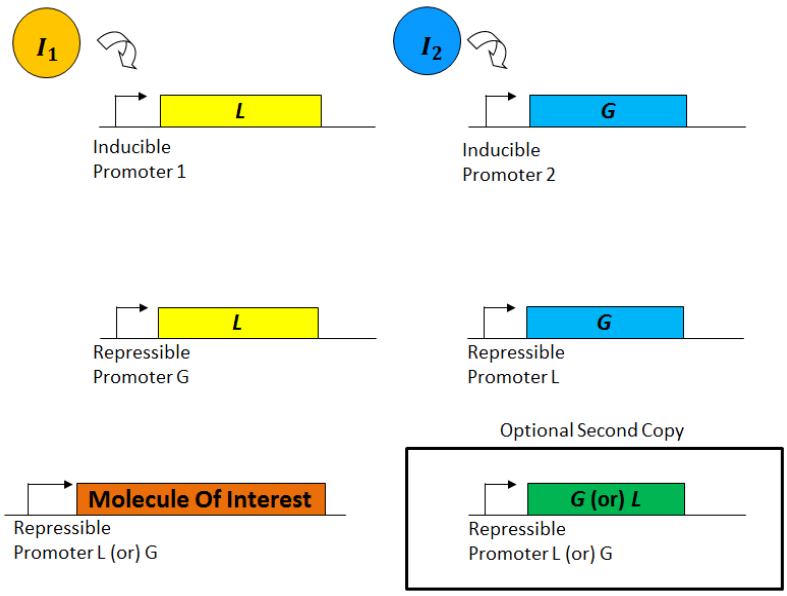

Figure 4.1 Illustration of the Negative inhibition circuit. I_1 represents inducer 1 and I_2 represents inducer 2. The inducible promoter 1 is driving the expression of repressor *L* (LexA repressor). The inducible promoter 2 is driving the expression of a distinct repressor *G* (Gal4 repressor). Repressible promoter G represents the promoter that is repressible by *G* and is driving the expression of *L*. Repressible promotor L represents the promoter that is repressible by *L* and is driving the expression of *G*. Either repressible promoter (L or G) can be used to drive the expression of the "Molecule of Interest", which could be a reporter molecule such as luciferase. The optional second copy promoter could either be repressible promotor L driving *G* expression or repressible promotor G driving *L* expression. This second copy is optional and its inclusion will be based on the computation results presented in this chapter.

We plan to build this circuit in Arabidopsis using our library of parts characterized in plants discussed in Chapter 3. This library contains over 100 promoter-repressor pairs tested in *Arabidopsis*. These 100 promoter-repressor pairs make use of 8 types of repressors. Each of these 8 repressor types is built with one of two DNA binding domains: the yeast Gal4 binding domain or the bacterial LexA binding domain. Yeast and bacterial binding domains were used to maintain orthogonality when incorporating these constructs into *Arabidopsis*. More information on the construction of these genetic circuits can be found in Appendix B. As the dual repressor circuit will require one LexA-based repressor combined with one Gal4-based repressor we shall call the repressors Gal4 and LexA from this point on.

The dual repressor circuit will ideally be robust against environmental noise but responsive to key inducers meant to switch between states. This leads to the question, what promoter-repressor characteristics would give us the best chance of having a bi-stable circuit?

4.2.2 Creation of the non-dimensional phase diagram

To look for the most promising bi-stable candidates we used the same set of ordinary differential equations, ODEs, that we discussed in Chapter 2 and which were employed by Gardner et al. [1], in the following form. Note: As we are not including a leaky expression terms as we did in Chapter 2, we are assuming the leaky expression is effectively zero compared to the fully repressed state of the circuit.

$$\frac{dG}{dt} = \frac{B_G}{1 + \left(\frac{L}{H_G}\right)^{n_g}} - D_G G \tag{4.1}$$

$$\frac{dL}{dt} = \frac{B_L}{1 + \left(\frac{G}{H_L}\right)^{n_l}} - D_L L \tag{4.2}$$

Where: *G* and *L* are the concentrations of the repressors (Gal4 and LexA). *t* represents time. B_G is maximum expression over time of the Gal4 repressible promoter. B_L is maximum expression over time of the LexA repressible promoter. H_G is the repressibility of the promoter driving Gal4 and H_L is the repressibility of the promoter driving LexA. By repressibility we mean the relative amount of repressor needed to bring the expression level to half of its relative max expression. D_G is a first order degradation rate describing the degradation of the Gal4 repressor. D_L is a first order degradation rate describing the degradation of the LexA repressor. n_g represents the Hill coefficient of Gal4 repressor effect on the system. n_l represents the Hill coefficient of LexA repressor effect on the system. We can non-dimensionalize these equations to make them easier to work with.

$$\frac{dg}{d\tau} = \frac{X_g}{1 + (l)^{n_g}} - g \tag{4.3}$$

$$\frac{dl}{d\tau} = \frac{X_l}{1 + (g)^{n_l}} - Dl \tag{4.4}$$

Where: *g*, *l* and *τ* are dimensionless and are related to the concentrations and time: *G*, *L* and *t* by the following relationships: $g = \frac{G}{H_L}$, $l = \frac{L}{H_G}$, $\tau = \frac{t}{D_G}$. The other dimensionless parameters are related to the dimensional parameters via the following relationships: $X_g = \frac{B_G}{H_L D_G}$, $X_l = \frac{B_L}{H_G D_G}$, $D = \frac{D_L}{D_G}$.

We then used a MATLAB ODE solver to simulate the system starting at different initial conditions for *g* and *l* (one initial condition with high *g* and low *l* and the other initial condition with high *l* and low *g*), recording the steady state concentration of each simulation. If the different initial conditions came to two different steady states, the parameter combination was

stored as bi-stable, whereas if only one steady state was found, the parameter combination was stored as mono-stable. A tolerance of high *I* minus low *I* greater than one was set to determine if steady states were different from one another. Although this difference of one is arbitrary, it is sufficient to capture the desired trend in stability throughout parameter space, as illustrated in Figure 4.2. With this information we created phase diagrams for different parameter values to see how the bi-stable space was affected (Figure 4.2).

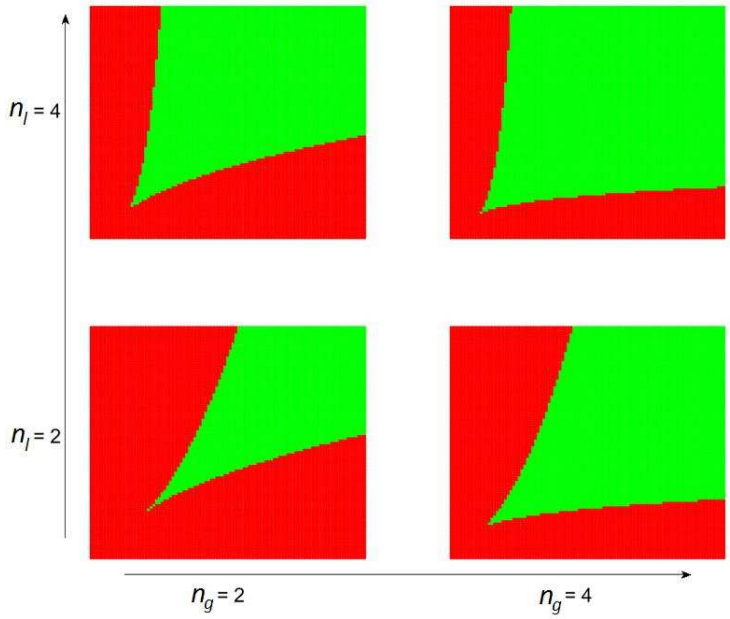

Figure 4.2 Phase Diagrams. Red areas indicate mono-stable parameter combinations. Green areas indicate bi-stable parameter combinations. Each plot represents X_g being varied along the x-axis and X_l being varied along the y-axis. The Hill coefficients n_g and n_I values are varied between plots.

The phase diagrams show the regions where parameter combinations lead to bi-stable behavior and where they lead to mono-stable behavior. Two key observations are as follows: (1) Increasing the Hill coefficient, *n*, increases the size of the bi-stable region and (2) larger and equivalent values for parameters X_g and X_l are needed to push the system into the bi-stable region. Thus, strong and balanced promoters are required for a bi-stable toggle switch. Although in principle, the balanced system may be achieved by decreasing the stability of the

protein controlled by the stronger promoter, increasing the strength of the weaker promoter will result in a system that is predicted to be farther into the bi-stable region.

4.2.3 Selecting the parts for switch construction

Table 4.1 The seven promoter-repressor pairs that meet the criteria. Constitutive Element: a sequence of DNA necessary for creating promoters with constitutive expression. Repressor Binding Type: a fragment of DNA whose sequence allows particular repressors to bind. Repressor motif: the domain in our protein-based repressors that is required to repress transcription. Number of Binding Site: the number of places the repressor can bind to the promoter. Other Binding Site Properties: include if the sites were near the TATA box or if a larger amount of spacer DNA was used between the binding sites compared to the other promoters. Order: how each genetic domain was positioned with respect to the other domains. Figure 3.1a contains a visual representation of the positioning of these pieces.

Constitutive Element	Repressor Binding Type	Repressor Motif	Number of Binding Sites	Other Binding Site Properties	Order
35s	LexA	EAR	1		35s LexA
35s	LexA	OFPx	4	near TATA box	35s 4xLexA
35s	LexA	OFP1	4	near TATA box	35s 4xLexA
Nos	LexA	B3	8	with spacer DNA	8xLexA Nos
Nos	LexA	OFPX	2		2xLexA Nos
Nos	Gal4	EAR	2		Nos 2xGal4
Nos	Gal4	EAR	2		2xGal4 Nos

As presented in Chapter 3 we developed a library of quantitatively characterized promoter-repressor pairs. We then produced phase diagrams which described the effect of the different promoter-repressor properties on the size of the bi-stable region as illustrated in Figure 4.2. With this information, we developed further criteria based on the information we gathered from our phase diagram analysis to select from our library the most promising promoter-repressor pairs. The criteria are as follows:

- one experimental replicate must result in the Hill coefficient n>2 for the promoter-repressor pair

- this experimental replicate must also show a P value < 0.1 for the least squares parameter estimate of the statistical hypothesis test that $n > 0$.
- one experimental replicate must result in a fold change > 2

Seven out of over one hundred promoter-repressor pairs in our library meet these criteria. These seven promoter-repressor pairs are listed in Table 4.1.

As two different repressor binding types are required for construction of this switch, we need to choose one LexA and one Gal4 promoter-repressor pair for each combination. Choosing promoter-repressor pairs from Table 4.1 gives us a total of 10 possible combinations for the dual repressor circuit. The task then became to predict which of these 10 combinations has the highest probability of generating a bi-stable circuit.

4.2.4 Finding probability of being bi-stable: a bootstrap method

We developed a method for determining the probability of producing a bi-stable system for each combination using the quantitative information collected as part of our plant part library presented in Chapter 3. This quantitative information included 3 parameter values for each promoter-repressor pair being tested, *B*, *H* and *n*. *B* describes the relative max strength for each promoter. This would be comparable to $\frac{B_G}{D_G}$ or $\frac{B_L}{D_L}$ in our model described in equations 4.3 and 4. Remeber *H* describes the repressibility of each promoter. This would be comparable to H_G or H_L in our model. *n* is then the Hill coefficient for the repressor's effect on the promoter. This would be comparable to n_g or n_l in our model. However, the estimates we made for the values of these parameters are limited in their accuracy due to noise in the data. In order to account for noisy data, we thought that if we could understand the error distribution of these parameter estimates, we could predict how much of this distribution lies in the bi-stable region,

giving us a probability of being bi-stable. We then developed a bootstrap-based approach to make this prediction. A reference for bootstrap methods can be found in [2].

Table 4.2 Outline of bootstrap-based method for predicting the bi-stability of each combination.

Step #	Brief Description
1	Estimate the mean, μ, and standard deviation, σ, of the log of the parameters for each promoter-repressor pair.
2	Generate a random set by sampling 1,000 *B* and *H* values from a normal distribution given the mean, μ, and standard deviation, σ, for log *B* values and μ and σ for log *H* values.
3	For each of the 8 combinations, exponentiate each of these randomly sampled values and calculate all possible X_g and X_l values from the sampled *B*s and *H*s. This gives a vector of length 1 x 10^6 for both X_g and X_l.
4	Separately for n = 2, 3 or 4 determine what percentage of the X_g, X_l data points lie in the bi-stable region using the non-dimensionalized phase diagram.
5	Rank combos by their estimated probability of being in the bi-stable region.
6	Repeat steps 3 and 4 after adding the corresponding mean(X) values to each of the 1 x 10^6 points we have for X_g and X_l distributions. This was done to approximate the probability for combinations with 2 copies of one repressible promoter paired with 1 copy of the other (*i.e.* 1:2 and 2:1 combinations). Different copy numbers were considered as adding additional copies of the various repressor-promoter pairs would be within reason considering our experimental time line for the project.

Development of the method for predicting the probability of being bi-stable, given the quantitative information from our library of plant parts, started by assuming a log normal error distribution for the parameter estimates *B* and *H*. The reasoning behind the log normal assumption is due to the following: The bootstrap sampling performed for the error estimates in Chapter 3, for estimating *B* and *H*, yields a distribution resembling a log-normal distribution (illustrated in Appendix B Figure B.8 for *B*, *H* and *n*). Another reason influencing this choice were the log normal distributions of basal firefly luciferase data illustrated in Figure 3.3, considering that both *B* and *H* have units of molecules and the firefly luciferase data is linearly proportional

to molecular number. We can estimate the mean and standard deviation for the assumed log normal error distribution for six out of the seven promoter-repressor pairs that met our criteria. However, the seventh promoter-repressor pair, 35s 4xLexA OFPx, was only tested once in our library of characterized plant parts. We therefore have no replicate measurements from which to estimate a mean and standard deviation.

As presented in Chapter 3, the stable transformation experiments suggested that we require values for *n* in the range of 2-4. This provides the information needed to develop our method for predicting the probability of being bi-stable for each combination as outlined in Table 4.2. This method was applied to each combination whose error distribution could be estimated.

4.3 Negative inhibition system Results and Discussion

4.3.1 Combination strengths

Table 4.3 Strengths of 10 our combinations.

Promoter-Repressor Pair 1	Promoter-Repressor Pair 2	Strengths
Nos 2xGal4 EAR	35s LexA EAR	These are one of the first constructs tested in 2011 and have continued to show the promise of bi-stability since then.
2xGal4 Nos EAR	35s LexA EAR	
Nos 2xGal4 EAR	35s 4xLexA OFPX	Show the highest cooperativity values of any pair.
2xGal4 Nos EAR	35s 4xLexA OFPX	
Nos 2xGal4 EAR	35s 4xLexA OFP1	Are closely related to the 35s 4xLexA OFPX construct but do not rank among the higher preforming combinations in terms of their probability of being bi-stable.
2xGal4 Nos EAR	35s 4xLexA OFP1	
Nos 2xGal4 EAR	8xLexA Nos B3	
2xGal4 Nos EAR	8xLexA Nos B3	

Nos 2xGal4 EAR	2xLexA Nos OFPX	These are the top four in terms of their probability of being bi-stable.
2xGal4 Nos EAR	2xLexA Nos OFPX	

The ranking of 8 out of 10 possible combinations is made apparent in Figure 4.3. This whole process was repeated 1,000 times to assess the variation in the estimation of these probabilities. The red error bars on the plots in Figure 4.3 are the standard deviations of these 1,000 repeats. Given this analysis, a summary of all 10 combinations strengths can be found in Table 4.3. Several of these constructs are currently being tested in Arabidopsis for the presence of a bi-stable response.

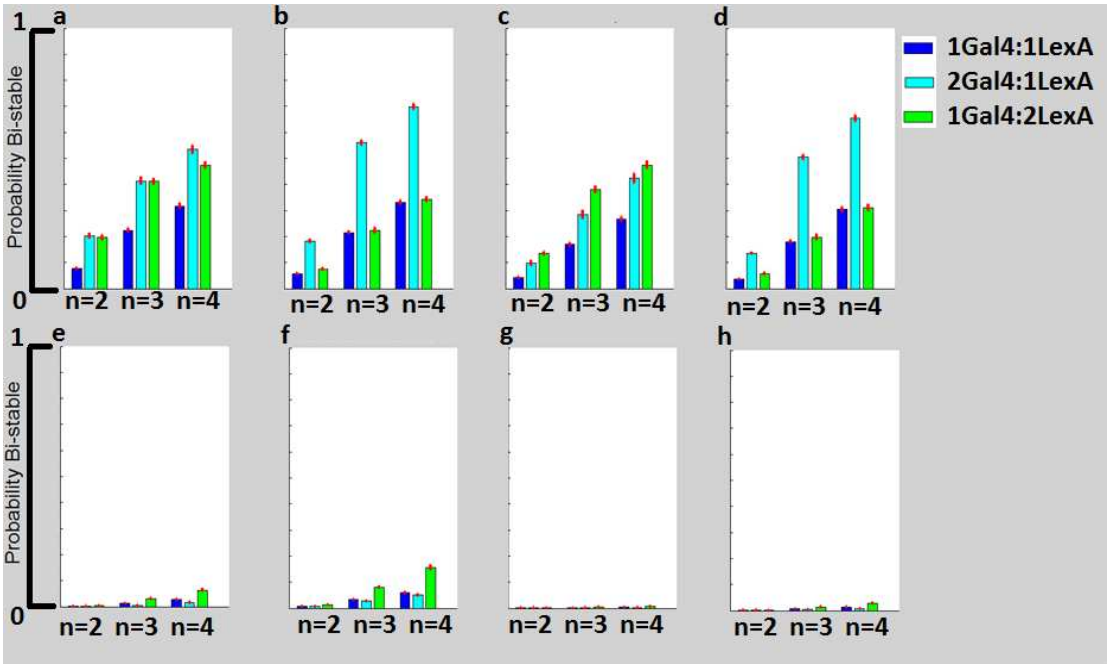

Figure 4.3 Estimated Bi-stable Probability. The combinations are represented in the following panels: (a) 2xGal4 Nos EAR with 2x LexA Nos OFPX (b) Nos 2xGal4 EAR with 2x LexA Nos OFPX (c) 2xGal4 Nos EAR with 8x LexA Nos B3 (d) Nos 2xGal4 EAR with 8x LexA Nos B3 (e) 2xGal4 Nos EAR with 35s 2xLexA EAR (f) Nos 2xGal4 EAR with 35s 2xLexA EAR (g) 2xGal4 Nos EAR with 35s 4xLexA OFP1 (h) Nos 2xGal4 EAR with 35s 4xLexA OFP1. Blue bars represent 1 copy of the Gal4-based repressible promoter paired with 1 copy of the LexA-based repressible promoter. The cyan bars represent 2 copies of the Gal4-based repressible promoter paired with 1 copy of the LexA-based repressible promoter. The green pars represent 1 copy of the Gal4-based repressible promoter paired with 2 copies of the LexA-based repressible promoter. The red error bars are +/- 1

standard deviation for the 1,000 replicated simulations. The x-axis represents categorical labels representing the assumed *n* value for each parameter combination. The y-axis represents the calculated probability of the different parameter combinations existing in the bi-stable region.

4.4 Positive feedback system Methods

4.4.1 Why create a positive feedback system?

It was desirable to construct a positive feedback-based switch for plants in addition to the negative inhibition-based switch discussed earlier in this chapter. Chapter 2 goes into more detail on advantages of a positive feedback system over a negative inhibition system. Also theoretically, positive feedback systems can produce a bi-stable system using only one feedback promoter, which is less complicated then the double repressor system with requires two promoters that inhibit each other. Figure 4.4 describes a case of such a positive feedback system. This relatively simple circuit topology is where we decided to start construction of our positive feedback-based system for plants. To augment the design and construction of the genetic parts we performed a computational analysis of the switch described in Figure 4.4 with the purpose of defining ideal properties for each part. We hope this work will set up the foundation for further synergistic computational and experimental work in design of this system.

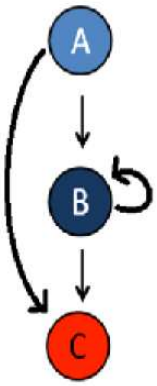

Figure 4.4 Illustration of Positive Feedback circuit. The "A" layer represents an inducible promoter driving the expression of a protein-based activator molecule, also called *A*. The layer "B" represents the positive feedback layer in the genetic network. This layer consists of a promoter driving the expression of the protein-based activator molecule *B* which in turn activates its own transcription. "C" represents the reporter layer in the genetic network. This layer consists of a promoter driving the expression of a reporter, *C*, such as green fluorescent protein or a luciferase molecule that can be monitored to assess the activity of the circuit. The arrows represent upregulation of the layer at the end of the arrow by the layer at the beginning of the arrow.

4.4.2 Model Creation

Positive feedback based bi-stable genetic circuits have been built in eukaryotic organisms via a three-layered system [3]. We wanted to construct a similar yet bi-stable three-layered system with the circuit topology illustrated in Figure 4.4. We first created a mathematical framework from which to predict ideal part properties. The simplest mathematical model we could think of was used in development of this framework. This model consisted of activating Hill equations for the transcription factors effect on each promoter. First order degradation was assumed for each transcription factor and reporter. We also assumed a zero order leaky expression term.

$$\frac{dA}{dt} = a_1 - d_1 A \tag{4.5}$$

$$\frac{dB}{dt} = a_2 + \frac{b_1 B^{n_1}}{1 + \left(\frac{B}{k_1}\right)^{n_1}} + \frac{b_2 A^{n_2}}{1 + \left(\frac{A}{k_2}\right)^{n_2}} - d_2 B \tag{4.6}$$

$$\frac{dC}{dt} = a_3 + \frac{a_3 B^{n_3}}{1 + \left(\frac{B}{k_3}\right)^{n_3}} + \frac{b_4 A^{n_4}}{1 + \left(\frac{A}{k_4}\right)^{n_4}} - d_3 C \tag{4.7}$$

Where: a_1 and a_2 represent the lowest possible level of expression of the transcription factors *A* and *B* respectively. a_3 represents the lowest possible level of expression of the reporter

protein, *C*. d_1, d_2 and d_3 represent the degradation coefficients for the proteins *A*, *B* and *C* respectively. *t* represents time. b_1, k_1 and n_1 are the parameters of the Hill function representing the positive feedback transcription factor's (*B*'s) activation of its own expression. b_2, k_2 and n_2 are the parameters of the Hill function representing the effect of the transcription factor *A* on the expression of *B*. b_3, k_3 and n_3 are the parameters of the Hill function representing the effect of the transcription factor *B* on the expression of *C*. b_4, k_4 and n_4 are the parameters of the Hill function representing the effect of the transcription factor *A* on the expression of *C*. The model presented in equations 4.5-7 was used to describe the effect different aspects of the genetic parts will have on the behavior of the circuit.

4.4.3 Positive Feedback: Non-Dimensional Phase Diagrams

To make this system of equations easier to work with, equation 4.7 was removed under the assumption that the promoter driving the expression of *C* is not being saturated by transcription factors. In other words we are assuming the promoter driving the expression of *C* will not become saturated during the low state expression of *B*. Also, although outside of the scope of this first model, it should be noted that downstream components may affect the stability of upstream states in a process referred to as retroactivity [4]. Therefore, by writing the equations as we have, we are also assuming any retroactivity effects are negligible. Also assuming steady state, this now two-equation system (equations 4.5 and 6), can be non-dimensionalized to further simplify the analysis to one equation.

$$0 = X\frac{\tilde{B}^{n_1}}{1+\tilde{B}^{n_1}} + L - \tilde{B} \qquad 4.8$$

Where: $\widetilde{A}$ and $\widetilde{B}$ are related to *A* and *B* via the following relationships: $\widetilde{A} = \frac{A}{k_2}$ and $\widetilde{B} = \frac{B}{k_1}$. The dimensionless parameters *X* and *L* are related to the dimensional parameters via the following relationships: $X = \frac{b_1 {k_1}^{(n_1-1)}}{d_2}$, $L = \frac{a_2}{k_1 d_2} + \frac{b_2 {k_2}^{n_2}}{k_1 d_2} \times \frac{\widetilde{A}^{n_2}}{1+\widetilde{A}^{n_2}}$ and $\tilde{A} = \frac{a_1}{k_2 d_1}$.

At steady state this non-dimensional equation 4.8 gives us a parameter space containing only three parameters, n_1, *L* and *X.* This space describes all possible behavior for the system as we have modeled it. This lower dimensional parameter space can be more easily visualized compared to higher dimensional space seen in equations 4.5-7. We can use equation 4.8 to ask the question; which parameter combinations gives us bi-stability? We can also visualize the non-dimensional phase diagram that encompasses the bi-stable region and ask the question: What effect does each parameter have on this bi-stable system? Finally we can relate changes in properties of different genetic parts with changes in different parameters, allowing for us to identify ideal part properties based on the mathematical analysis.

4.5 Positive feedback system: Results and Discussion

4.5.1 Parameters relationship to bi-stable space

Figure 4.5 illustrates how each dimensionless parameter in equation 4.8 is graphically related to the bi-stability of the system. As seen in Figure 4.5, when n_1 (the cooperativity of the transcription factor *B*'s effect on the system) increases, so does the size of the bi-stable region.

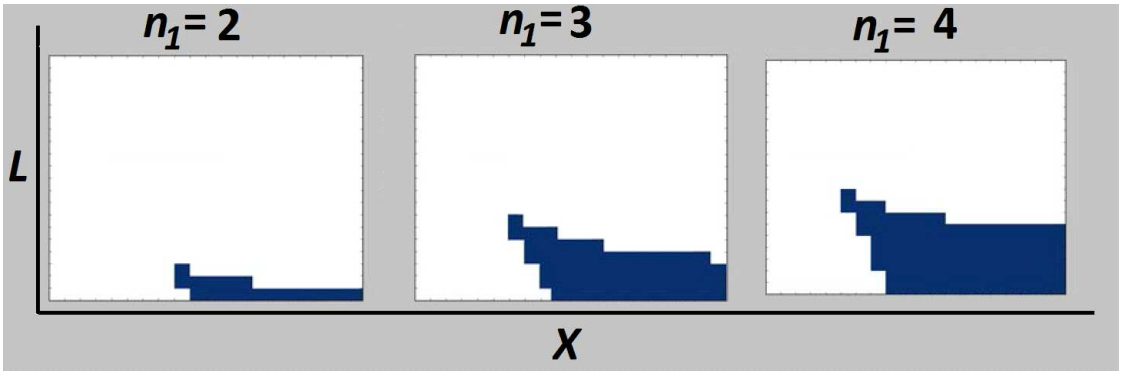

Figure 4.5 Phase Diagrams for the positive feedback system. The blue area represents bi-stable parameter combinations, whereas the white area represents mono-stable parameter combinations. The x-axis represents values of the dimensionless parameter *X*. The y-axis represents values of the dimensionless parameter *L*. n_1 is varied from left to right as indicated above each plot.

The bi-stable space exists for low values of *L*. There also appears to be a need for a balanced *X* value. As seen in equation 4.8, *X* helps to set the maximum expression of the positive feedback in the system. In other words, it scales the impact $\tilde{B}$ has on the circuit. Taking into account the effect of *X* on the circuit and this phase diagram analysis, we can see that too high an *X* value would correspond to a system that is always on, whereas too low an *X* value, a system that is always off. In other words, as we cross the bi-stable region, we move from a system on the left that is naturally always off to a system on the right that is naturally always on. Therefore, the ideal parameter value will be to keep *X* intermediate between high and low values.

We can now start to think about where early genetic parts built in the Medford lab place us in respect to the phase diagrams. In the Medford lab, early experimental data collected from these genetic parts suggested that the positive feedback promoter could be induced but had only one low expression level (*i.e.* steady state). As the experimental system suggests, we have a mono-stable system with a low steady state value; *X* is therefore confined to the left side of the bi-stable region. We did not draw any conclusions about *L* and n_1 from these data.

Given this starting region of parameter space, we can determine how each dimensionless parameter should be tuned to bring us closer or within the bi-stable region. *L* should be decreased to maximize our chances of existing in the bi-stable region. *X* should be increased as our current estimate for the value of *X* is lower than what is required for bi-stability. n_1 should be increased as this increases the size of the bi-stable region in both the *L* and *X* directions.

We can use the relationships between the dimensionless parameters and the dimensional parameters to identify what changes in the dimensional parameters will bring the system closer to the bi-stable region. To increase *X* we considered its definition, $\mathrm{X} = \frac{b_1 {k_1}^{(n_1-1)}}{d_2}$. To put this relationship into words, increasing b_1, k_1, n_1 or 1/ d_2 in combination or separately will increase the parameter *X*. To decrease *L* we consider its definition, $L = \frac{a_2}{k_1 d_2} + \frac{b_2 {k_2}^{n_2}}{k_1 d_2} \times \frac{\tilde{\mathrm{A}}^{n_2}}{1+\tilde{\mathrm{A}}^{n_2}}$. Using similar reasoning as we used for *X* we can identify dimensional parameters needed to be either increased or decreased to lower the value of the non-dimensional parameter *L*. Notice, increasing d_2 improves the value of *L* but does not improve *X.* However, increasing the value of k_1 improves the values of *L* and *X*. In fact, k_1 is the only parameter that has positive effects on both *L* and *X.*

4.5.2 Experimental part relationship to the parameters

Before we can identify ideal genetic part properties it is helpful to lay down a basic understanding of how each part is related to each parameter. This can be challenging as many parameter to part relationships are not known. However, this is an ideal challenge for us to tackle due to the synergistic benefit of our mathematical modeling in conjunction with

experimental study. We can use this benefit to provide a platform for hypothesizing and then testing what these parameter part relationship may be. We started by listing out the different experimental parts we have in our system: promoters, transcription factors (such as activators) and reporters. Next, we listed out the different parameter types we used: *b* type (*e.g.* b_1, b_2), d type (*e.g.* d_1, d_2), k type (*e.g.* k_1, k_2), n type (*e.g.* n_1, n_2) and *a* type (a_1 and a_2). Then, hypothesis were constructed about the parameter type to genetic part relationship should be.

The *b* type parameters are related with promoter strength since they appear in the numerator of the activating Hill function. Thus, they can be affected by changing the combined maximum strength of the promoters driving the particular transcription factors. This may be achieved by designing stronger transcription factors as well as stronger promoters. The *d* type parameters are all degradation parameters and hence can be affected by changing the thermodynamic stability of the different transcription factors or targeting the transcription factors for faster degradation via a degradation tag. The *k* type parameters can be affected by changing the number of binding sites a particular promoter has for its activator. As this assumption is less intuitive than the others, the following explanation may be helpful.

The parameter *k* has units of molecules and helps control at what point an activating transcription factor starts to saturate a particular promoter. Saturation is thought to happen when enough of a particular transcription factor is present that all the binding sites for that transcription factor on the promoter are filled. Therefore, further increase in the transcription factor concentration will have little to no effect. If we add more binding sites to the promoter, more transcription factors should be able to bind and the *k* type parameters should be increased. With this line of reasoning also comes the caveat that there will be a certain point at

which adding more binding sites does not increase promoter expression as the binding sites may be too far away from the promoter.

The Hill coefficient or *n* type parameters can be changed by changing cooperative binding effects between the transcription factors. A currently untested hypothesis on how to achieve this increase in cooperativity is to use or design transcription factors that have the ability to attract more transcription factors once bound to the promoter. If this hypothesis is valid we should see an increase in cooperativity by balancing the number of binding sites and perhaps positions of the binding sites with transcription factors capable of this attraction. Finally, the leaky expression or *a* type parameters can be changed by many mechanisms such as decreasing the stability of the corresponding transcription factor or by reducing the low level expression from a promoter.

Some of these changes are more easily achievable experimentally compared to others. Also in real systems it may be impossible to change parameters independently from one another, such as b_1 and k_1. With this in mind, the ideal part properties identified in these studies are as follows. The promoter driving *B* needs to have a sigmodal response to the transcription factor *B* which may be affected by changing the number of binding sites for *B*. The max strength of the feedback from the *B* needs to be increased which may be affected by changing the number of binding sides for *B*. The low level of *B* expression needs to be kept in check. This may be accomplished by changing d_1 but at the expense of also changing the max expression of *B*. Another option may come from changing the number of binding sites. Using the assumed parameter to part relationships it is quite likely there are many ways these ideal properties could be designed; however, one method that stands out here is adjusting the

number of binding sites. Remember with the dimensionless parameters L and X, k_1 was thought to increase X while decreasing L. If k_1 can be controlled by the number of binding sites this may be a key part property to optimize.

In the design of our positive feedback system we used computationally developed phase diagrams. We then estimated our position on these phase diagrams from qualitative results collected by the Medford lab for this positive feedback system. This allowed us to make creative suggestions to refine the design of the genetic parts in order to move to the desired area of the phase diagram. In the end we identified ideal part properties for design of our positive feedback system.

4.6 Conclusion

This study started with the goal of determining the ideal parts for two different types of bi-stable systems. The first system was based on a negative inhibition circuit that required a combination of genetic parts from our previously developed library. We developed part selection criteria and a method for predicting the probability of creating a bi-stable system with different part combinations; this leads to determining the strengths and weakness of each part combination.

The second system was based on positive feedback. As we did not have access to a part library for building the positive feedback system in plants, we wanted to know if we could identify key properties that the different parts should have. To do this we constructed a system of ODE's to describe the effect different plant genetic parts could have on the system. Non-dimensionalizing these ODEs allowed us to reduce the parameter space to 3 dimensions. This in turn allowed for identification of the mathematical parameter's effect in the system. Finally, we

were able to suggest ideal part properties for the positive feedback system based on connections between the parameters in the model and the genetic parts. Altogether these predictions provide a platform for construction of plant synthetic gene circuits that are hoped to lead to the first multicellular bi-stable switches.

REFERENCES

[1] T. Gardner, C. Cantor and J. Collins, "Constuction of a genetic toggle switch in Escherichia coli," *Nature,* vol. 403, pp. 339-342, 2000.

[2] F. J, "Bootstrapping Regression Models," in *Applied Regression Analysis and Generalized Linear Models*, SAGE Publications, 2008, pp. 587-595.

[3] C. M. Ajo-Franklin, D. A. Drubin, J. A. Eskin, E. P. Gee, D. Landgraf, I. Phillips and P. A. Silver, "Rational design of memory in eukaryotic cells," *Genes and Dev.,* vol. 21, pp. 2271-2276, 2007.

[4] D. Vecchio, A. Ninfa and E. Sontag, "Modular cell biology: retroactivity and insulation," *Molecular Systems Biology,* vol. 4, no. 161, 2008.

CHAPTER 5

Identification of Transcriptomic Trends in Cancer Cell Lines

5.1 Introduction

Cancer is one of the leading causes of mortality worldwide, and over half a million people died of cancer in the United States in 2015 [1]. Many cancer patients are treated with chemotherapy. While there are new targeted therapies available for a few cancers, many patients will encounter broad range chemotherapy drugs, such as cytotoxic drugs, as one of the common lines of treatment [2]. However, there are a large number of drugs available and patient response to drugs is very heterogeneous [3] [4]. Because every patient and every cancer is different, this naturally raises the question: can we tailor cancer drugs to the specific cancer of a patient, and develop individualized medicine? In order to do this, we need to find signatures of sensitivity or resistance to a drug, i.e. signatures of those biological switches that cells can turn on to escape from a chemotherapy drug. Because these biological circuits are genetic, we need to look for the molecular signatures that results from the underlying genetic circuitry. Some of these signatures can be found in the transcriptome.

Transcriptomic data is publically available for cancer cell lines in many databases, including the NCI60 and GDSC. The NCI60 was established in 1990 as a panel of 60 cell lines to screen existing and potential drugs for cancer treatment [5]. The NCI60 database is part of the Developmental Therapeutics Program at the National Cancer Institute. The GDSC database contains a panel of over 600 cell lines used to screen drugs for cancer treatment [6]. Its main reference was published in 2013 and is maintained by the Cancer Genome Project at Wellcome

Trust Sanger Institute and the Center for Molecular Therapeutics at Massachusetts General Hospital Cancer Center [6]. Both databases contain transcriptomic data in the form of micro-array analysis at the native state of the cell lines (*i.e.* transcriptomic state of the cell before drug treatment).

Micro-array analysis is done by extracting mRNA from the cell lines. This mRNA is then reverse transcribed to cDNA that is labeled with fluorescence. The labeled cDNA then is allowed to hybridize with probes on a microchip. These probes consist of complementary sequences to many known genetic transcripts. The chip is then washed to remove unbound cDNA and the fluorescence is measured for each probe. There are often several probes on a chip corresponding to one gene target; these are called probe sets. Preprocessing methods, such as RMA (Robust Multi-array Average) [7], are applied to these probe sets to give a signal value for each gene represented on a micro array chip.

The purpose of RMA, which may be among the most widely used preprocessing methods, is to reduce the experimental variation within a chip as there are differences in probe sets used and the affinities of those probe sets for their targets. This method was developed using data collected on Affymetrix Gene Chips [7]. It has been shown that when compared with other methods, RMA can reduce more of the variation while maintaining comparatively little bias [7]. Even after RMA processing, there exists experimental variation from data collected on different chips and different experiments. Methods, such as COMBAT ('Combating' batch effect when combining batches of gene expression microarray data), have been developed to reduce the further variation that exists when combining data across different experiments from multiple chips [8].

The micro array data found in the NCI60 and GDSC databases is coupled with drug sensitivity data for each of the cell lines. There also has been much research over the years on the power of these micro-array assays to predict drug sensitivity [9]. This work includes the development of gene filtration methods and models for prediction of drug sensitivity [10] [11] [12]. However, there has been less work done in comparing different methods and repeatability between methods for construction of predictive models for these data. For example questions like the following are still open:

1) Which gene filtration method and model type are best for predicting drug response?
2) Are the best performing methods and models different for different drugs?
3) How do the predictions change if the database used (*i.e.* NCI60 or GDSC) changes?

DREAM projects are challenges posed to the public often aimed at solving biological problems [13]. They often pose a challenge and accept entries attempting to meet that challenge over a set time range. Recent work published in response to one of the DREAM projects has begun to look at the predictive power of this transcriptomic data for breast cancer cell lines while running a comparative model analysis for different model types [14]. Here we expand on this idea by running a comparative analysis of Databases, Models and gene filtration methods focusing on 13 chemotherapeutics tested in both the NCI60 and GDSC databases. This led us to assess the repeatability of data across the two databases and create a data set of accuracy scores for over 1,000 database, drug, gene filtration and model type combinations.

5.2 Repeatability Between the Databases

The two major repositories of cancer line data, the NCI60 and the GDSC, have data on many of the same cell lines, which were collected by different groups at different points of

time. This gives us an opportunity to test whether the data in the two databases give comparable results, i.e. whether the results of a statistical analysis from one database is reproducible when taking data from the other database. Reproducibility of results across databases is important for science, and would give us confidence in the quality of the data maintained. The question becomes how to compare databases for repeatability. For the NCI60 and GDSC there are two different types of data we need to compare for repeatability, micro-array data and the drug sensitivity data. The 35 matching cell lines between the NCI60 and GDSC (*i.e.* cell lines that were tested in both databases) present an ideal place from which to assess this repeatability.

We started by comparing the micro-array data. Before we could begin to compare these databases we needed to make sure we were using the same probe sets in both the NCI60 and GDSC dataset by only keeping probes that were present in both data sets. We also needed to reduce the known variability of working with micro-array data collected in different samples. Common techniques used to reduce this variability are RMA and COMBAT (as discussed in the introduction of this chapter). Although not included in this study, there are other methods available for normalizing micro-array data, such as FROZEN [15].

For the purposes of this comparison we tried many different methods of preprocessing the data, using the techniques mentioned above, followed by clustering of the matching cell lines. Four of the preprocessing methods are described in Table 5.1. We used an Euclidean distance agglomerative hierarchical clustering algorithm implemented in R through the *fastcluster* package for clustering matching cell lines based on their micro-array data [16]. This algorithm clusters cell lines according to their Euclidean distance assuming a space where each

probe measurement is a dimension. We then chose the method that clustered the most matching cell lines together. Matching cell lines were said to cluster together when there were no other cell lines between them in terms of their Euclidean distance.

Table 5.1 Preprocessing Methods for Combining Databases. Four preprocessing methods used to test whether if microarray data from different databases could be combined.

Method Number	Details of Method	Results of Method
1	1. Download .cell files from NCI60 and GDSC databases 2. Combine data before RMA and COMBAT processing 3. Run through clustering algorithm	One pair of matching cell lines clustered together
2	1. Download .cell files from NCI60 and GDSC databases 2. RMA and COMBAT processes NCI60 and GDSC data separately 3. Run through clustering algorithm	zero pairs of matching cell lines clustered together
3	1. Download .cell files from NCI60 and GDSC databases 2. Combine data before RMA and COMBAT processing 3. Center data by probe value means and divide by probe value standard deviations for each probe across all cell lines in combined NCI60 and GDSC data set 4. Run through clustering algorithm	Zero pairs of matching cell lines clustered together
4	1. Download .cell files from NCI60 and GDSC databases 2. RMA and COMBAT processes NCI60 and GDSC data separately 3. Center data by probe value means and divide by probe value standard deviations for each probe across all cell lines in combined NCI60 and GDSC data set 4. Run through clustering algorithm	26 pairs of matching cell lines cluster together

Figure 5.1 illustrates the results from each of these four methods. We chose to move forward with preprocessing method 4 as it was by far the one with the most matching cell lines clustered correctly. Without further study we cannot say why this method performed the best, however, one possible reason could be in how COMBAT handles the batch effect. COMBAT was developed by testing data collected on different chips of the same microchip array type. However, the NCI60 and GDSC are collected not only between many different chips but the

NCI60 contain many different array types; here we chose to use the NCI60's Affymetrix HG-U133A array data [5] whereas for the GDSC we used HT-HGU122A Affymetrix whole genome array data [6]. Perhaps it is the differences in these arrays that make method 4 the best. Altogether this brings up an important point for these studies across databases: It is imperative to have a control for which to check repeatability between the databases to ensure usability of the data when running multi-database studies.

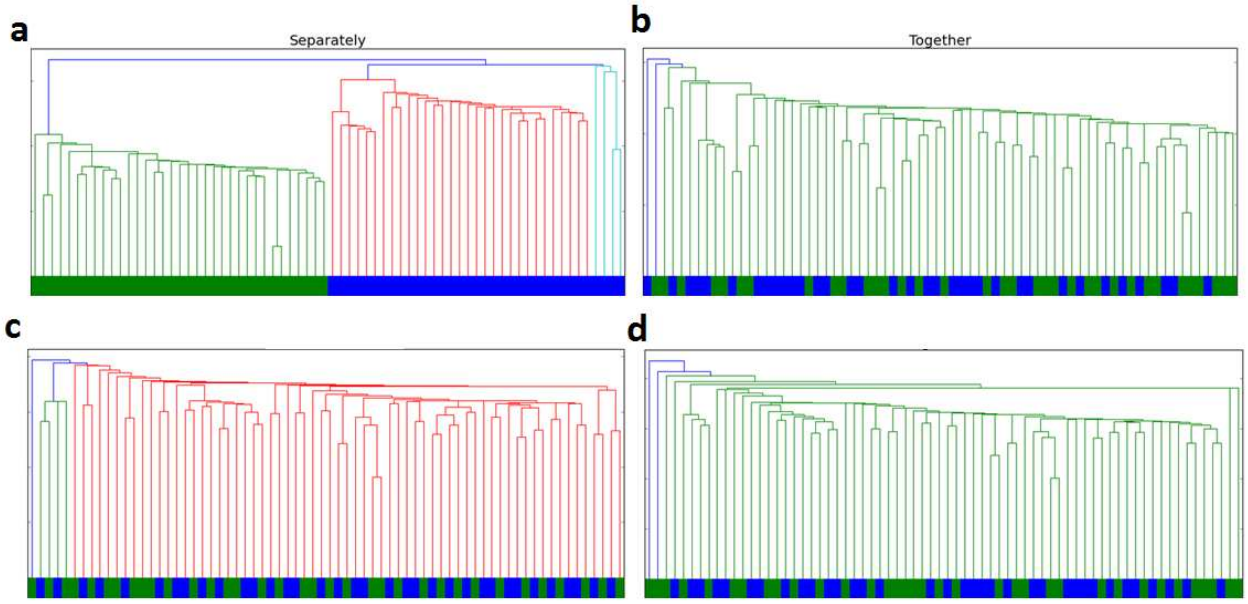

Figure 5.1 Cell Line Clustering. Dendogram outputs of the fastcluster algorithm on top of a blue and green bar graphs indicating the database type of each cell line (i.e. NCI60=green, GDSC-blue). a) Data clustered after preprocessing method 1. As we can see in the bar graph the cell lines are clustering by database. b) Preprocessing method 2, (c) preprocessing method 3 and (d) preprocessing method 4.

Next, we wanted to assess the repeatability of the drug sensitivity data between these databases. Both the NCI60 and GDSC report growth inhibition scores as defined in equation 5.1. One way to conceptualize this score is that scores between 0 and 1 represent retardation in the growth of the cells (*i.e.* cell are dividing more slowly but are still growing the population size), while scores < 0 represent a decrease in the initial size of the population. We chose to use the

GI50 for our study. The GI50 is the log of drug concentration needed to bring the growth inhibition score to 50% of its initial value [5] [6].

$$GI = \frac{Ti-Tz}{C-Tz} \times 100 \qquad 5.1$$

Where: GI = growth inhibition score, Tz = cell population when drug is added, Ti = cell population after incubation with drug and C = control: cell population after incubation with no drug.

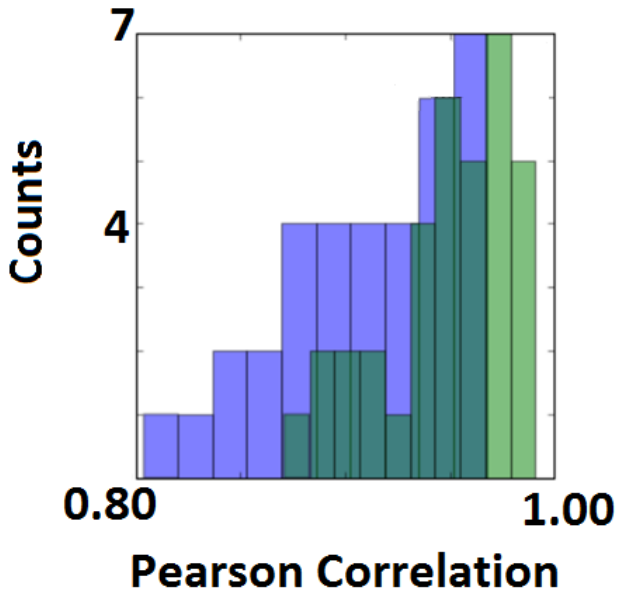

Figure 5.2 Matching Cell Line Drug Sensitivity Pearson Correlations. This figure shows a blue histogram of Pearson correlation coefficients between the GI50 values of 13 drugs for each of the matching cell lines in the NCI60 database against the GDSC database. The green histogram is obtained when extrapolated GI50 values have been removed.

Although the NCI60 and GDSC are both reporting the GI50, the assays they used to collect the cell population information are different. The NCI60 uses SRB staining, where cell population measure is based on the colorimetric dye, Sulforhodamine B, binding to amino acids

of fixed cells [17]. The GDSC on the other hand uses the Thermo Fisher Syto60 Fluorescent Nucleic Acid Stain [6]. As we are looking for a general score of drug sensitivity both of these assays should provide equivalent information. To assess the repeatability of the drug sensitivity data, we calculated the Pearson correlations for each matching cell line's GI50 values across all 13 drugs of interest, as represented by the blue histogram in Figure 5.2. However, we discovered that the GDSC database used extrapolated values when calculating the GI50 for certain cell line and drug combinations[2]. After removal of these extrapolated values we re-ran the comparison across the matching cell lines and found a general shift upwards in the Pearson correlations, as seen in the green histogram in Figure 5.2.

5.3 Comparative Analysis

To construct our comparative analysis of databases, drugs, models and gene filtration methods we started by creating a work flow depicted in Figure 5.3. The work flow was constructed to approach the comparative analysis with an unbiased systematic method for predictive model creation. As each model type has a random component, all models were created three times to access the repeatability of building each model. The work flow has four factors and several different measures of accuracy for each model.

Factor 1 represents which database to use. We decided to look at the two different databases in five different ways: (1) NCI60 only, (2) GDSC only, (3) NCI60 and GDSC combined, (4) Non-extrapolated GDSC only (*i.e.* we removed the cell lines for which the GI50 was based on

[2] The realization and removal of the extrapolated GI50 values was conducted by Joshua Mannheimer.

extrapolation) and (5) Non-extrapolated GDSC and NCI60 combined. This allowed us to look into the following questions:

- Do models constructed on NCI60 only and GDSC only give comparable results?
- Does the use of extrapolated values ever improve the models performance?
- Does combining the data sets improve model performance?

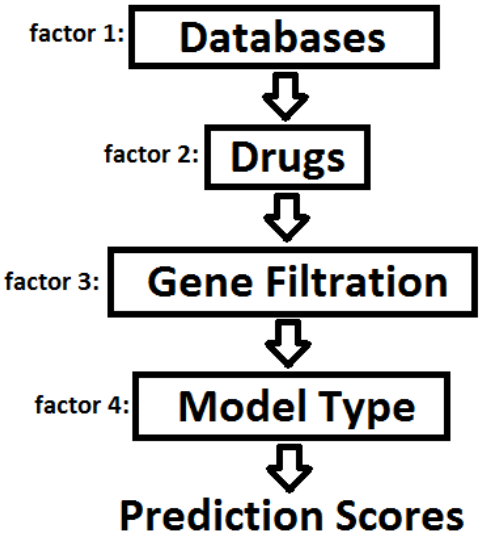

Figure 5.3 Prediction Score Generation Work Flow. This figure illustrates the work flow for generating the prediction scores for all database, drugs, gene filtration and model type combos. First one picks a database type. Then a drug is selected. This is followed by application of a chosen gene filtration methods to the data. Then a chosen model type is used to create predictions scores to assess each method of model construction. This was done for all possible combinations of the factors.

Factor 2 represents which drug to test. We chose to look at 13 different chemotherapeutic drugs tested in both the NCI60 and GDSC databases, Bleomycin, Bortezomib, Cisplatin, Cytarabine, Docetaxel, Doxorubicin, Etoposide, Gemcitabine, Methotrexate, Mitomycin, Paclitaxel, Vinblastine and Vorinostat. This allowed us to look into the question:

- Do models predict the behavior of chemotherapeutic drugs differently under different modeling conditions?

Factor 3 represents which gene filtration method to use. The gene filtration methods explored in this chapter are Differentially Expressed Genes, DEGs, [10] and COXEN [11]. The DEGs method filters based on how well the probe set data correlate with drug sensitivity data [10]. COXEN is based on filtering the micro-array data to only probes that correlate well with corresponding micro-array analysis from tumor biopsies of the same histotype [11]. To be more specific COXEN uses the probes that have the highest correlation of correlation scores [11]. Correlation of correlation scores can be broken down as follows: within target and reference probe values are correlated with other probe values within the same set [11]. These correlations are then correlated across the target and reference sets for each probe, hence a correlation of correlations score is determined for each probe [11]. Then COXEN probe values are filtered by choosing probes with high correlation of correlation scores, using a sample of the cell lines micro-array analysis as the target set and a specific tumor type micro-array analysis as the reference set [11]. Four methods of gene filtration were chosen: (1) no gene filtration, (2) COXEN with lung or bladder tumor samples, (3) Differentially Expressed Genes, DEG and (4) DEG then COXEN. This allowed us to look into the following questions:

- Do we see an improvement in model performance for any specific gene filtration method?
- Do we see an improvement in model performance for histologically similar cell lines when COXEN is performed with tumor samples?

Factor 4 represents the type of model to use. We looked at two different categories of models, regression and probabilistic type models. For our regression models we choose the following model types: (1) principal components regression, PCR, (2) partial least squares

regression, PLSR and (3) non-linear logistic regression (a neural net machine learning algorithm). For our probabilistic models we choose the following model types: (1) MiPP based linear discriminant analysis [18] and (2) a neural net classifier. This paved the way for questions such as:

- Do we see an increase in model performance for any a specific model type?
- Do we see a difference between the linear and non-linear models?

In order to look into all of these combinations we analyzed >1,000 models for each tumor type. To assess the repeatability of building each model, the whole algorithm was repeated three times for each combination. This systematic approach to model development would not work for all MiPP-based models, as MiPP-based models do not run fast enough to test all 20,000+ probes in around a month of simulation time. Thus MiPP-based models were not created when no gene filtration method was applied.

5.4 The Algorithm

5.4.1 Description of Algorithm

The algorithm was constructed in four different modular script types: Partition Lists, Gene Filtration, Create Model and Analyze Results. The modularity of these scripts allows for easy parallelization of the code when scripts are run on a cluster.

The Partition Lists script (partition_lists.py) was used to create partitions of the micro-array and drug sensitivity data, for each drug. A testing and a validating partition were created to allow models to be built using the testing partition and assessed using the validation partition. The Partition Lists script also insured that at least 1 cell line of the corresponding tumor type was in both the testing and validation partitions. The Gene Filtration script

(gene_flitration.py) was needed to run gene filtration on the testing partition. The Create Model scripts (create_model.R and create_model.py) were needed to create each model. Finally the Analyze Results script (analyze_results.py) was needed to calculate the model performance scores for each combination.

Model performance scores for the probabilistic models are: (1) the p value from the Binomial test. The Binomial test looks for deviation from an expected distribution of two possible outcomes. This test is used here to see if the number of predicted correctly drug sensitivity classes is different than a 50 percent successful Bernoulli experiment, (2) the percent correct and (3) the MiPP score. The MiPP score (misclassification penalized posterior) is a score that not only looks into the number of predicted correctly drug sensitivity classes but also considers how confident we were in each prediction by looking at the posterior probability used to make each prediction [18]. The performance scores for regression models are: (1) root mean squared error, RMSE (2) Pearson correlation and (3) REC. REC stands for regression error characteristic and is related to the error's cumulative distribution [19]. The formula we created for REC can be found in equation 5.2 below and takes into account what a random set of prediction scores would have been[3].

$$REC = aocRand - aocActual$$

$$aoc = MaxErrorValue - \sum P\left(\frac{\sqrt{(x-y)^2}}{\max(x)-\min(x)}\right) \qquad 5.2$$

[3] The development of our rec scores was conducted by Joshua Mannheimer.

Where, *x* and *y* are target and predicted values. *P(u)* is the estimated cumulative probability of the values <=*u* given all data. (Note: aoc stands for area over the cumulative error distribution curve).

With this set up we were able to create our data set of over 1,000 model prediction scores for two different tumors, specifically Lung and Bladder. To get a birds eye view we ran a four way ANOVA on the data set for each performance score. The ANOVA by design will look to see which factor or combination of factors is responsible for significant portions of the variance in the entire data set. Two confounding factors to take note of for this analysis are as follows: (1) If a significant portion of the variance is allotted to a signal factor and a combination of that factor with another factor, then the test cannot say if the significance is due to the signal factor alone. (2) When a combination of factors is said to contain a significant portion of the variance those factors are thought to interact. This interaction can be visualized in a marginal mean plot.

5.4.2 Description of Marginal Mean Plots

Marginal mean plots are plots of the mean behavior of one or more factors across all other factors. For example, Figure 5.4a contains a plot of the prediction score (binomial p values) marginal mean for two factors: Database and Model Type. Each point represents the mean across all other factors. In this figure we can see as the database changes so does the mean behavior of the different models. We can also see that for **different** databases **different** model types have on average lower Binomial p values; because of this difference the lines for the different models cross. The ANOVA tests to see if these crossing lines, “interactions”, are significant by looking to see if a significant portion of that variance belongs to the combination of these two factors. As seen in Table 5.2 the “Database:Model” factor combination is not

considered to hold a significant portion of the variance. However, it is still interesting that we can see a trend in the marginal means for the plots of Figure 5.4. Further study should investigate the distributions of scores that make up these means to see if any significant difference exists.

5.4.3 Trends in the Database Factor

For good performing classification models the trend should be low Binomial P values, high percent correct values and high MiPP score values. Using the key provided in the figure legend database levels: (0) NCI60 only, (3) Non-extrapolated GDSC only and (4) Non-extrapolated GDSC and NCI60 combined appear to not do as well in model performance compared to the (1) GDSC only and (2) GDSC and NCI60 combined. With a few exceptions for the (2) GDSC and NCI60 combined, this trend also appears to hold for the bladder data found in Figure 5.5. However, even though the database factor does hold a significant portion of the variance in all cases there exists confounding factors showing significant interactions (Tables 5.3 and 5.4). As this is a high level view, further studies may want to break down each marginal means individual scores to see if there is any fine detailed information we could gather from the distributions of these scores.

To continue with this high level view, we can turn to the impact of the factor databases on the regression models. Figure 5.6 and 5.7 show the marginal mean plots for the different predictions scores of these regression models. Tables 5.4 and 5.5 are the ANOVA results for the regression models. In these tables the factor databases, is again attributed a significant portion of the variance but there are more significant confounding interactions then we had seen with the classifications models. Therefore, again we only looked at the general trend to help identify

areas we would like to break down in future analysis. Good performing regression models should have low RMSE, high Pearson correlations and high REC scores. As there is less variation in R and REC between databases, we will base our bird's eye view on RMSE. However, it is interesting to take note of the differences presented between these scores' types. Perhaps further study could look into why there is less variation for these different scores. The general trend here for the database factor levels is a follows: (0) NCI60 only, (3) Non-extrapolated GDSC only and (4) Non-extrapolated GDSC and NCI60 combined appear to do better in RMSE model performance compared to the (1) GDSC only and (2) GDSC and NCI60 combined. Future directions should include a more fine-tuned break down of these interesting areas. However, even with the coarse grained high level view we can begin to see interesting trends between regression and probabilistic models. For example, probabilistic models do better on average when all data is present, including the extrapolated values in the GDSC. On the other hand, Regression models appear to do better when the extrapolated values are removed.

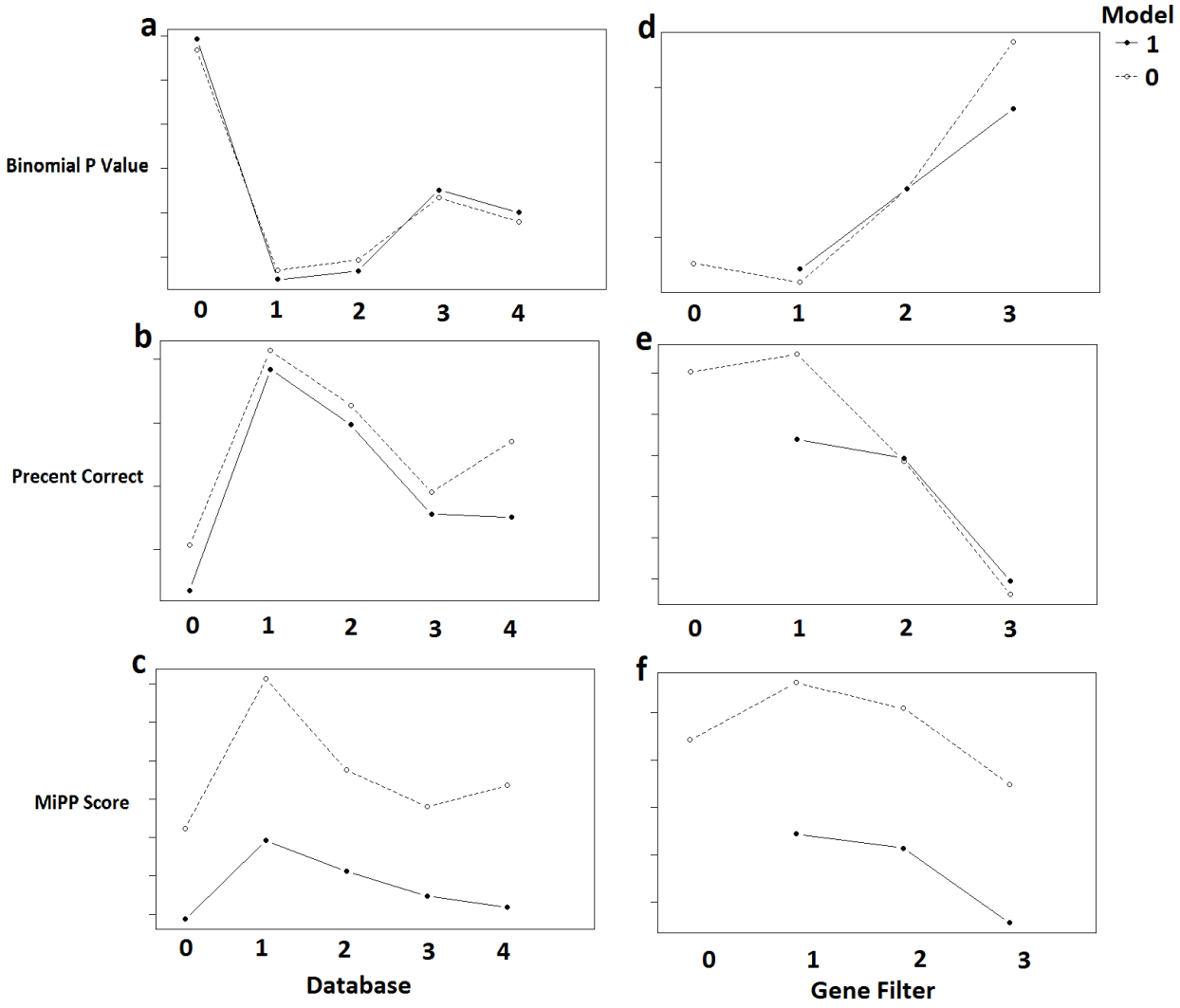

Figure 5.4 Classification Marginal Means From Lung ANOVA Analysis. This figure shows representative images of the marginal means for probabilistic models. The key is as follows:

Model	Database	Gene Filter
0 = neural net classifier	0 = NIC60	0 = no filtration
1 = MiPP	1 = GDSC	1 = COXEN
	2 = Combined	2 = DEG
	3 = non-extrapolated GDSC	3 = DEG then COXEN
	4 = non-extrapolated Combined	

Table 5.2 Four Way ANOVA Results for Lung Tumor based Classification Models. This table contains the output from three different ANOVA analyses. One for each of the following predictions scores: Binomial p value, Percent Correct and MiPP.

Lung	Binomial P Value		Percent Correct		MiPP Score	
Database	< 2e-16	***	1.09e-13	***	< 2e-16	***
Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
GeneFilter	1.37e-12	***	8.71e-12	***	2.1e-15	***
Model	0.431		0.341		< 2e-16	***
Database:Drug	1.23e-14	***	< 2e-16	***	< 2e-16	***
Database:GeneFilter	0.128		0.017	*	2.51e-4	***
Drug:GeneFilter	0.699		0.910		0.902	
Database:Model	0.321		0.854		0.0205	*
Drug:Model	0.211		0.634		0.0690	.
GeneFilter: Model	0.301		0.181		0.871	
Database:Drug:GeneFilter	0.725		0.925		0.694	
Database:Drug:Model	0.557		0.0963	.	0.0216	*
Database:GeneFilter:Model	0.572		0.705		0.912	
DrugLGeneFilter:Model	0.297		0.489		0.736	
Database:Drug:GeneFilter:Model	0.783		0.654		0.750	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

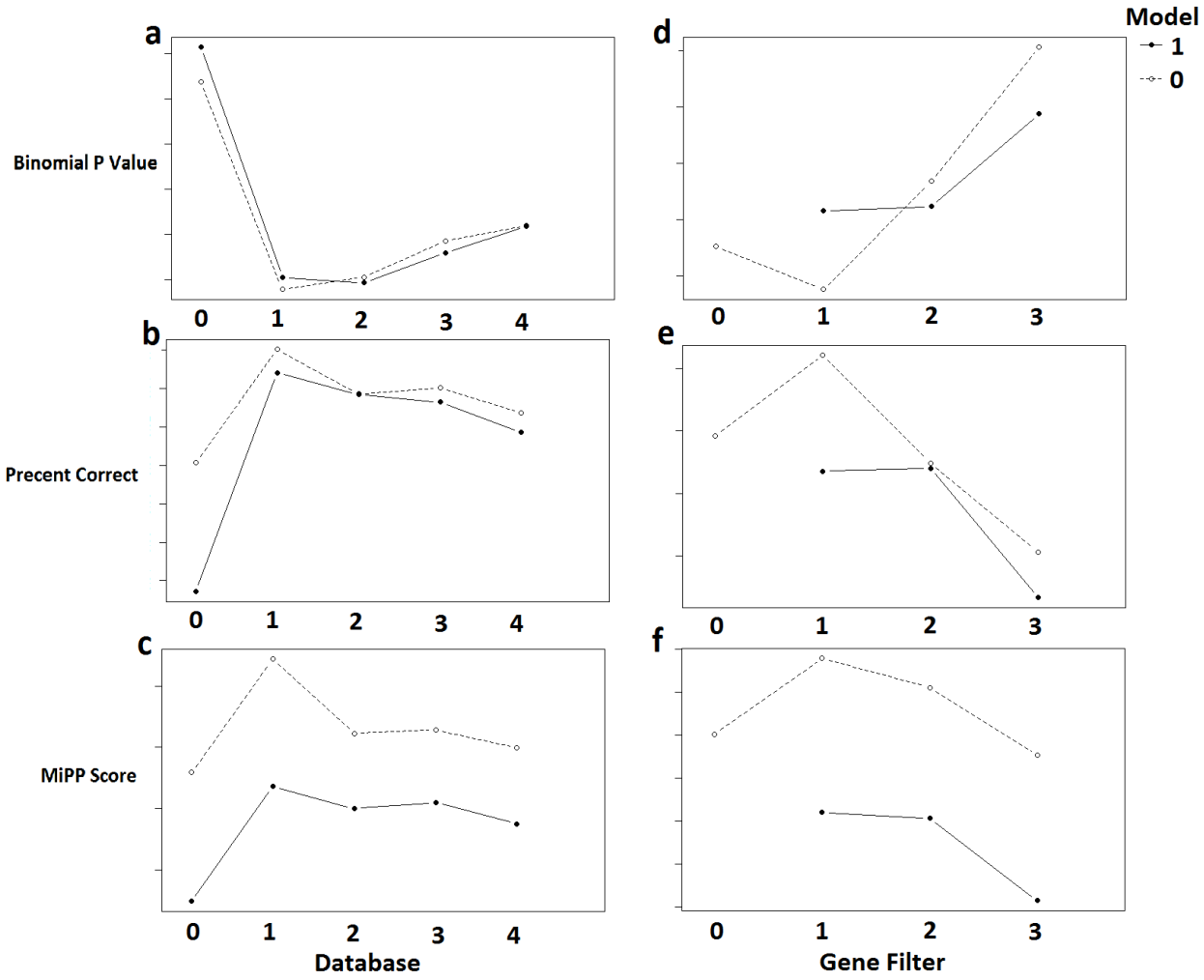

Figure 5.5 Classification Marginal Means From Bladder ANOVA Analysis. This figure show representative images of the marginal means for probabilistic models. The key is as follows:

Model	Database	Gene Filter
0 = neural net classifier	0 = NIC60	0 = no filtration
1 = MiPP	1 = GDSC	1 = COXEN
	2 = Combined	2 = DEG
	3 = non-extrapolated GDSC	3 = DEG then COXEN
	4 = non-extrapolated Combined	

Table 5.3 Four Way ANOVA Results for Bladder Tumor based Classification Models. This table contains the output from three different ANOVA analyses. One for each of the following predictions scores: Binomial p value, Percent Correct and MiPP.

Bladder	Binomial P Value		Percent Correct		MiPP Score	
Database	< 2e-16	***	< 2e-16	***	< 2e-16	***
Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
GeneFilter	4.07e-16	***	1.64e-11	***	6.92e-15	***
Model	0.793		0.00347	**	< 2e-16	***
Database:Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
Database:GeneFilter	0.276		0.534		0.0617	.
Drug:GeneFilter	0.0339	*	0.134		0.0707	.
Database:Model	0.163		7.35e-05	***	2.95e-4	***
Drug:Model	0.124		0.275		0.0456	*
GeneFilter: Model	0.00179	**	0.0531	.	0.633	
Database:Drug:GeneFilter	0.0326	*	0.184		0.173	
Database:Drug:Model	0.366		0.579		0.0270	*
Database:GeneFilter:Model	0.0893	.	0.368		0.465	
DrugLGeneFilter:Model	0.810		0.864		0.713	
Database:Drug:GeneFilter:Model	0.413		0.997		0.962	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

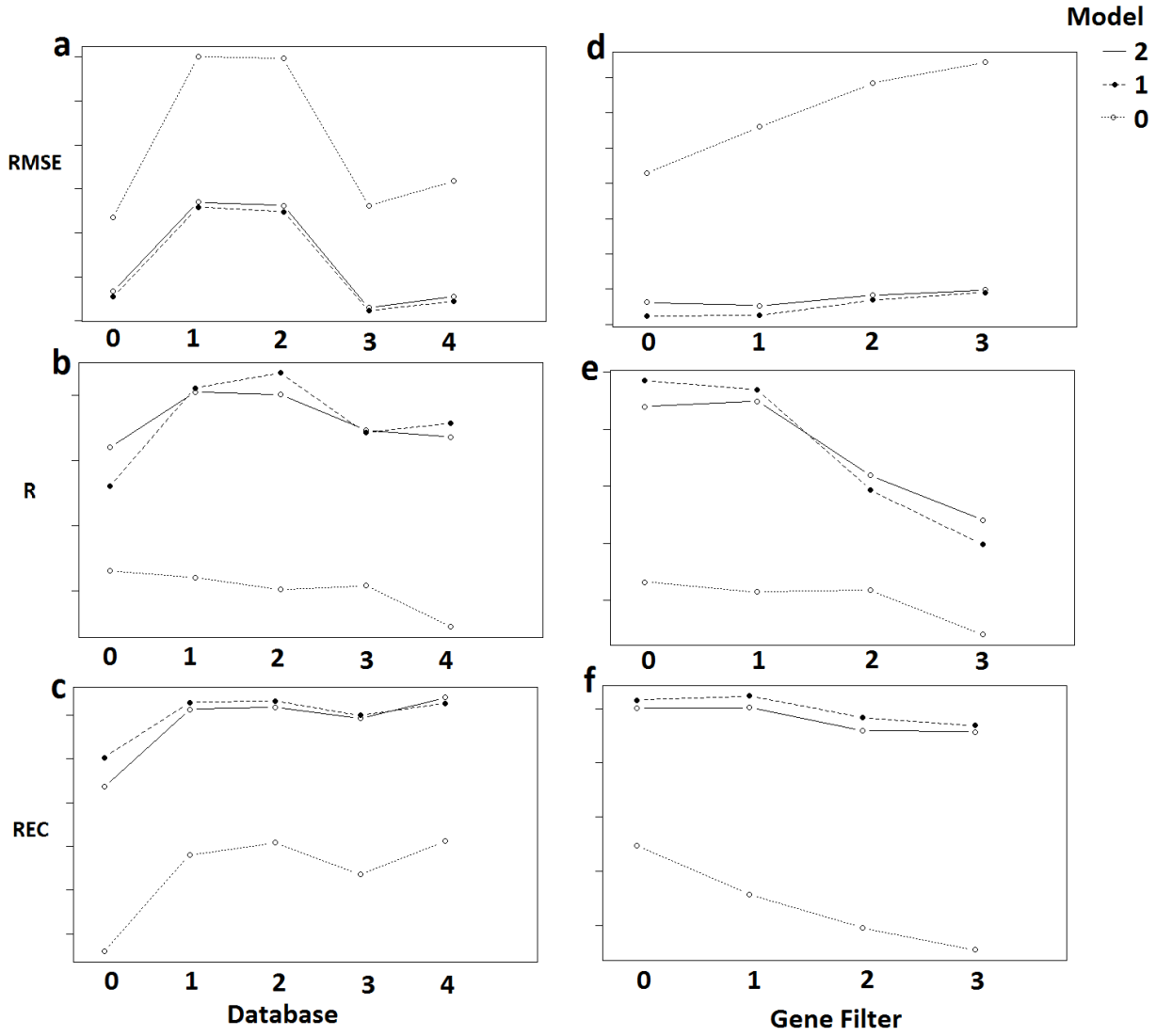

Figure 5.6 Regression Marginal Means From Lung ANOVA Analysis. This figure show representative images of the marginal means for regression models.

Model	Database	Gene Filter
0 = neural net	0 = NIC60	0 = no filtration
1 = PCR	1 = GDSC	1 = COXEN
2 = PLSR	2 = Combined	2 = DEG
	3 = non-extrapolated GDSC	3 = DEG then COXEN
	4 = non-extrapolated Combined	

Table 5.4 Four Way ANOVA Results for Lung Tumor based Regression Models. This table contains the output from three different ANOVA analysis. One for each of the following predictions scores: RMSE (root mean squared error), Pearson correlation and REC.

Lung	RMSE		Pearson Correlation		REC	
Database	< 2e-16	***	4.28e-07	***	< 2e-16	***
Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
GeneFilter	< 2e-16	***	< 2e-16	***	3.36e-12	***
Model	< 2e-16	***	< 2e-16	***	< 2e-16	***
Database:Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
Database:GeneFilter	0.085	.	0.433		0.00696	**
Drug:GeneFilter	0.127		0.0347	*	0.689	
Database:Model	< 2e-16	***	0.000145	***	0.0585	.
Drug:Model	< 2e-16	***	0.00350	**	0.0408	*
GeneFilter: Model	< 2e-16	***	2.44e-09	***	1.08e-4	***
Database:Drug:GeneFilter	0.999		0.747		0.997	
Database:Drug:Model	< 2e-16	***	0.287		0.00661	**
Database:GeneFilter:Model	4.98e-10	***	0.00121	**	0.0945	.
DrugLGeneFilter:Model	0.218		0.992		1.00	
Database:Drug:GeneFilter:Model	1.00		1.00		1.00	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

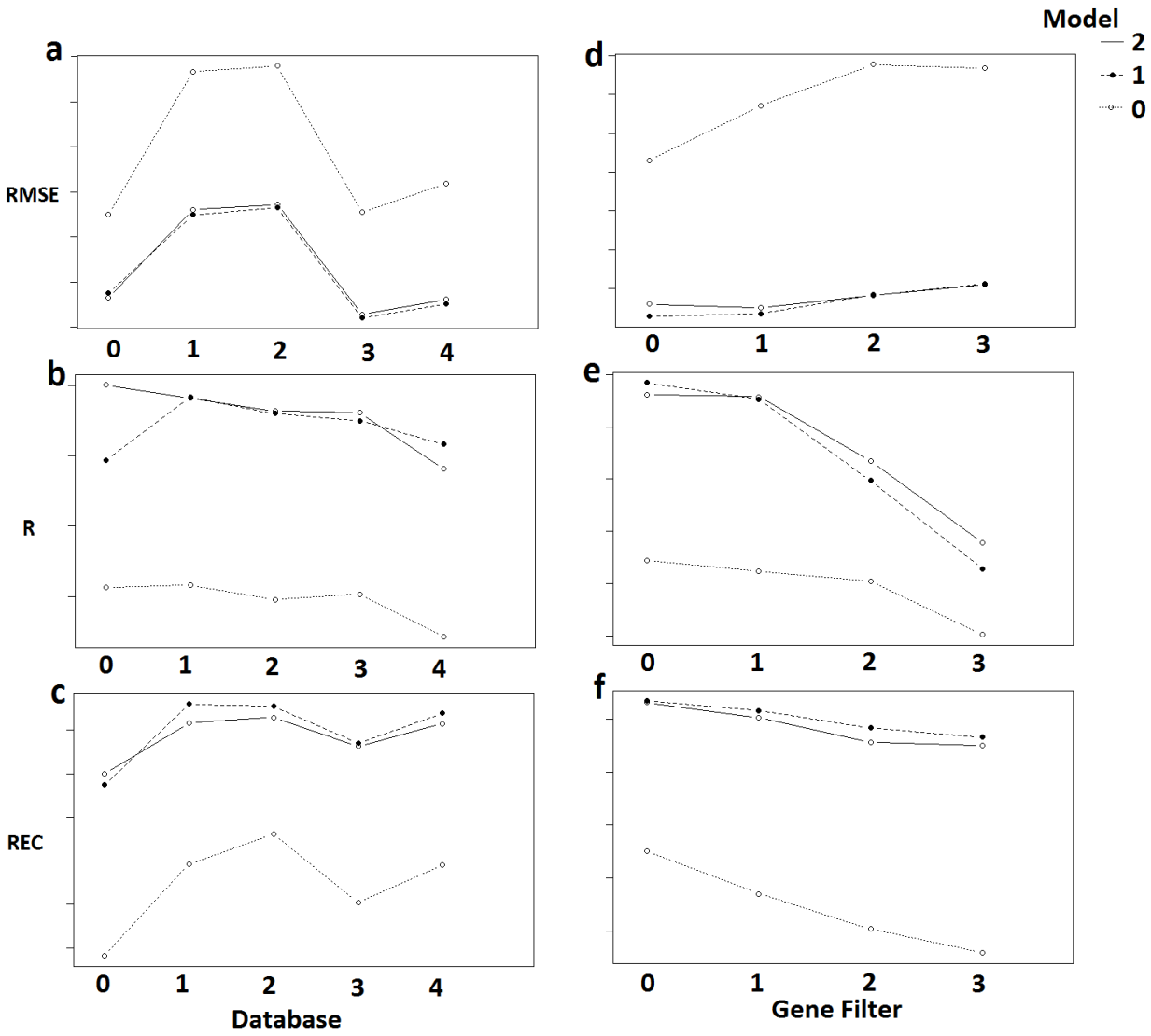

Figure 5.7 Regression Marginal Means From Bladder ANOVA Analysis. This figure show representative images of the marginal means for regression models.

Model	Database	Gene Filter
0 = neural net	0 = NIC60	0 = no filtration
1 = PCR	1 = GDSC	1 = COXEN
2 = PLSR	2 = Combined	2 = DEG
	3 = non-extrapolated GDSC	3 = DEG then COXEN
	4 = non-extrapolated Combined	

Table 5.5 Four Way ANOVA Results for Bladder Tumor based Regression Models. This table contains the output from three different ANOVA analysis. One for each of the following predictions scores: RMSE (root mean squared error), Pearson correlation and REC.

Bladder	RMSE		Pearson Correlation		REC	
Database	< 2e-16	***	4.83e-06	***	< 2e-16	***
Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
GeneFilter	< 2e-16	***	< 2e-16	***	< 2e-16	***
Model	< 2e-16	***	< 2e-16	***	< 2e-16	***
Database:Drug	< 2e-16	***	< 2e-16	***	< 2e-16	***
Database:GeneFilter	4.17e-06	***	0.00386	**	0.00089	***
Drug:GeneFilter	0.0622	.	1.59e-06	***	0.790	
Database:Model	< 2e-16	***	0.0318	*	0.00918	**
Drug:Model	< 2e-16	***	0.203		0.0555	.
GeneFilter: Model	< 2e-16	***	3.96e-11	***	0.00426	**
Database:Drug:GeneFilter	0.966		0.184		0.978	
Database:Drug:Model	< 2e-16	***	0.00574	**	0.259	
Database:GeneFilter:Model	1.00e-15	***	1.43e-4	***	0.0460	*
DrugLGeneFilter:Model	0.191		0.998		0.927	
Database:Drug:GeneFilter:Model	1.00		1.00		1.00	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

5.4.4 Lung Gene filtration Factor with No Confounding Factors

Another significant area highlighted by the ANOVA analysis can be seen in the lung tumor-based classification models. As seen in Table 5.2, the ANOVA analysis indicated a significant portion of the variance is attributed to the gene filtration factor. Furthermore, the significance of this factor is not confounded by significant interactions with the other factors in

the analysis (*e.g.* drug, database and model type). Considering this, we can ask the question: which gene filtration levels are significantly different from each other? To do this we ran an Honestly Significant Difference Tukey test whose results can be found in Table 5.6. These results indicate the significant difference is coming from the gene filtration level, "DEG then COXEN", as this is significantly different with every other level. From the marginal mean plots in Figure 5.4 the level, "DEG then COXEN", appears to not perform as well as the other levels (*e.g.* "COXEN" and "DEG"). This is surprising as this method was how the original COXEN paper performed their analysis [11]. One might have expected it to perform the best and not the worst as seen in Figure 5.4. However, we should point out that the level, "DEG then COXEN", represents the method with the least probes used for model creation. Considering this, further study of this data set should make sure the lack in performance is not due to the probe number alone. In other words, further studies should look to make sure the higher probe number models are **not** performing better due to non-meaningful connections found between the probes by chance.

Table 5.6 Tukey Results for Lung Classification Binomial P Value Gene filtration Results. This table illustrates the results of the Tukey test indicating the level "DEG then COXEN" gene filtration methods is significantly different from the other methods.

Group 1	Group 2	Reject for $\alpha = 0.05$
DEG then COXEN	COXEN	True
DEG then COXEN	DEG	True
DEG then COXEN	No gene filtration	True
DEG	COXEN	False
DEG	No gene filtration	False
COXEN	No gene filtration	False

The other predictive scores, such as MiPP, also suggest there may be a significant difference within the gene filtration factor; however, the ANOVA's using these other predictive scores show significant confounding interactions between the factors. Therefore, we cannot say if the significant difference across all scores is due to the gene filtration method alone. What we can do is look at the marginal means of these comparisons to see if the trend across gene filtration methods holds. Representative images in Figure 5.4 suggest that the trend does hold. We can also look at the classification bladder tumor-based models that show the same trend. However, unlike the lung tumor-based models, the bladder tumor-based models show confounding factors within all classification score types, as shown in Table 5.3. Though, again we can look at the marginal means to see if the trend of low performance for the level, "DEG then COXEN", gene filtration method holds. Representative images in Figure 5.5 suggest that the trend does hold. Further support for the poor performance in the level, "DEG then COXEN", gene filtration method comes from the regression models. Representative images (Fig. 5.6 and 7) of the marginal means in the regression models of both tumor types also indicate the level, "DEG then COXEN", as giving on average lower prediction accuracy scores.

5.5 Summary of Conclusions and Future Directions

In conclusion, we found that micro-array data across databases is comparable if normalized appropriately. This is a significant conclusion as it allows for use of both the NCI60 and the GDSC data, when treated appropriately, for predicting drug sensitivity. This in turn should improve the performance of drug sensitivity algorithms. We also found the nucleic acid stain used in the NCI60 for drug sensitivity measurements show Pearson correlations greater than 0.8 when compared to the same measurement taken for the GDSC using an amino acid type

stain. We also showed that as extrapolated values from the GDSC were removed, the overall trend in these Pearson correlations increased. Altogether this improved our ability to combine the databases.

For this large dataset we developed a parallelizable algorithm for model generation on a cluster. The output of this algorithm brought trends to light for different model building factors. We saw regression models appeared to do better when extrapolated data from the GDSC was removed, whereas classification models did better when this extrapolated data was included. We also found a trend in which gene filtration method was employed. For lung tumor-based models the gene filtration factor level, “DEG then COXEN”, was significantly different and showed a trend of poor behavior compared to all other gene filtration methods in terms the binomial p value. Although confounded by other factors, the trend of poor performance with, “DEG then COXEN”, was found more often than not to be present in all other accuracy scores for both regression and probabilistic methods.

Moving forward, the difference in behavior between the accuracy scores is worth investigating. It would be helpful to know if this difference is due to noise alone or if it is related to different aspects of model performance. Another future direction should include assessment of the question: Are the significant differences, such as seen within the gene filtration factor, due to removal of key probe information, or does the number of probes play a role? This study has put us a step closer to diving into the plethora of questions mentioned throughout this chapter. Altogether, this platform for model construction is aiding in the development and comparison of many model types, which in turn could be used to further the science of creating individualized treatments for cancer.

REFERENCES

[1] "Cancer Facts & Figures," American Cancer Society, Inc., Surveillance Research, 2015. [Online]. Available: http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2015/. [Accessed 1 May 2016].

[2] "Treatment Types," American Cancer Society, [Online]. Available: http://www.cancer.org/treatment/treatmentsandsideeffects/treatmenttypes/. [Accessed 1 May 2016].

[3] J. Barretina and et.al., "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* vol. 483, no. 7391, pp. 603-607, 2012.

[4] M. Garnett and etal., "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature,* vol. 483, no. 7391, pp. 570-575, 2012.

[5] R. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer,* vol. 6, pp. 813-823, 2006.

[6] W. Yang, J. Soares, P. Greninger, E. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. Smith, R. Thompson, S. Ramaswamy, A. Futreal, D. Haber, M. Stratton, C. Benes, U. McDermott and M. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research,* vol. 41, pp. 955-961, 2012.

[7] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostat,* vol. 4, no. 2, pp. 249-264, 2003.

[8] W. Johnson, A. Rabinovic and C. Li, "Adjusting batch effects in microarray expression data using Empirical Bayes methods," *Biostatistics,* vol. 8, no. 1, pp. 118-127, 2007.

[9] N. Kim, N. He and S. Yoon, "Cell line modeling for systems medicine in cancers (Review)," *International Journal of Oncology,* vol. 44, no. 2, pp. 371-376, 2014.

[10] S. Dudoit, Y. Yang, M. Callow and T. Speed, "STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN REPLICATED cDNA MICROARRAY EXPERIMENTS," *Statistica Sinica,* vol. 12, pp. 111-139, 2002.

[11] S. Smith, A. Baras, J. Lee and D. Theodorescu, "The COXEN Principle: Translating signatures of in vitro chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer," *Cancer Res,* vol. 70, no. 5, pp. 1753-1758, 2010.

[12] J. Zhao, X.-S. Zhang and S. Zhang, "Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs," *CPT: Pharmacometrics and Systems Pharmacology,* vol. 3, no. 2, p. e102, 2014.

[13] "DREAM CHALLENGES," Dream Challenges, 2006. [Online]. Available: http://dreamchallenges.org/. [Accessed 28 April 2016].

[14] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J. Collins, D. Gallahan, D. Singer, J. Saez-Rodrigue, S. Kaski, J. Gray and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology,* vol. 32, pp. 1202-1212, 2014.

[15] M. McCall, B. Bolstad and R. Irizarry, "Frozen robust multiarray analysis (fRMA)," *Biostatistics,* vol. 11, no. 2, pp. 242-253, 2010.

[16] D. Müllner, "fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python," *Journal of Statistical Software,* vol. 53, no. 9, pp. 1-18, 2013.

[17] V. Vicha and K. Kirtikara, "Sulforhodamine B colorimetric assay for cytotoxicity screening," *Nature Protocols,* vol. 1, pp. 1112-1116, 2006.

[18] M. Soukup, H. Cho and J. Lee, "Robust classification modeling on microarray," *Bioinformatics,* vol. 21, no. 1, pp. i423-i430, 2005.

[19] J. Bi and K. Bennett, "Regression Error Characteristic Curves," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

APPENDIX A[4]

Protocols for Quantitative characterization of genetic parts and circuits for plant synthetic biology

A.1 Plasmid Construction

Our transcriptional repressor proteins are built with two genetic components: a DNA binding (DB) and a repressor domain (RD). The DNA binding domains of the yeast Gal4 and the bacterial LexA transcription factors were used to create orthogonal repressor proteins for the plant synthetic circuits. The repressor domains we use are: Ethylene-responsive element binding factor-associated amphiphilic repression (EAR), plant-specific B3 repression domain (BRD), two variants of the Arabidopsis OVATE Family proteins (AtOFP1 and AtOFPx). AtOFPx represents a consensus sequence of the OVATE domains of the AtOFP family repressor proteins demonstrating the highest levels of repression [1]. Sequence optimized Gal4 and LexA DB, and the two OVATE RD, were synthesized as double stranded gBlocks (GeneArt/Life Technologies and IDT (Integrated DNA Technologies)). The synthetic repressor domains were fused in frame to one of two mentioned synthetic DNA binding domains using overlapping extension PCR with compatible *Bsa*I restriction enzyme sites built into the primers for downstream cloning. The small-sized EAR and B3 repressor domains were incorporated into reverse primers used to amplify DNA binding domains, creating in-frame C-terminal fusions. The hybrid products were

[4] Appendix A comprises the protocol supporting data that was published alongside the paper [10] in Nature Methods. It represents joint work and has been put here in its entirety to preserve the intellectual coherence of the project. In particular, the plasmid construction in section A.1 and the plant experiments in section A.7 was done by the co-authors from the Medford lab. The protoplast experiments detailed in were done by co-authors from the Medford lab, with assistance from Wenlong Xu and I of the Prasad lab. The image correction method in section A.4 was developed by Wenlong Xu.

sub-cloned and sequenced in pJET2.1 vector using pJET forward and reverse primers (Thermo Scientific). Two core repressor modules containing an upstream transcription block [2], estrogen inducible promoters, repressors and NOS terminator [3] were synthesized (GeneArt, Fig. B.1a,b). To interchange the different repressors in the module, the Golden Gate cloning method [4], using type II endonuclease *Bsa*I restriction sites, was used (Fig. B.1a,b). Repressor expression was controlled by two inducible promoters, 10xN1 and pOp6, which are 4-hydroxytamoxifen [5] (4-OHT) and dexamethasone [6] (DEX) inducible, respectively. Each of the inducible promoters was also designed to direct expression of the *Firefly luciferase* (F-luc) gene (Fig. B.1c,d). F-luc reporter gene serves as a proxy for quantifying the amount of repressor in the system.

The constitutively active repressible promoters were constructed by introducing DNA binding elements (operators) in the backbone of Cauliflower Mosaic Virus 35S (CaMV35S), Nopaline Synthase (NOS) and Figwort Mosaic Virus (FMV) promoters. The DNA binding elements containing two copies of Gal4, and two or eight copies of LexA, were synthesized as a gene block with appropriate restriction sites included (IDT). A series of repressible promoters was generated by varying the number of DNA binding elements, the spacing between each binding element and its position relative to the transcription start site. Plasmid backbone (Fig. B.1e) was used as a sub-cloning plasmid, from which promoter variants were made by adding DNA binding elements upstream and downstream of the promoters. Upstream of each promoter, DNA binding elements were cloned using *Bsa*I and *Hind*III restriction sites, whereas downstream of the promoters, elements were cloned using *Mlu*I and *Aat*II sites. Two CaMV35S based promoters with 4xLexA and 5xGal4 binding elements at the -32 position were

synthesized (GeneArt). For ease of cloning, a single repressible promoter module with two BsaI restriction enzyme sites flanking a repressible promoter was synthesized (Fig. B.1f). The synthetic module has a 5' transcription block, repressible promoter controlling expression of *Renilla* luciferase (R-luc), PEST domain [7], and transcriptional terminator (Fig. B.1f). The resulting promoter fragments from sub-cloning vector (Fig. B.1e) were digested with *Bsa*I and cloned into the repressible promoter module (Fig. B.1f) upstream of R-luc. We use R-luc to quantitatively determine the repressibility of the promoter upon repressor binding. In theory, a functional repressor-repressible promoter pair should show decreasing R-luc activity with increasing F-luc activity as a result of increasing inducer concentrations.

Two pBluescript SK^+ backbone plasmids (Fig. B.1c-g) containing F-luc under the control of two estrogen inducible systems, 10xN1 + NEV [5] and pOp6 + $LhGR_2$ [6], were used to assemble the expression cassette encoding the repressors and corresponding repressible promoters. Plasmid backbones were prepared by restriction digest with *Kpn*I and simultaneously dephosphorylated with alkaline phosphatase (FastAP, Thermo Scientific) to prevent self-ligation. The repressor and repressible promoter fragments were prepared by digesting with *Bsa*I and *Kpn*I. The two sticky *Bsa*I ends from the repressors and repressible promoters are compatible and the two external *Kpn*I sites were used for non-directional cloning into the vector backbone. Similarly, four beta plasmids, *i.e.*, containing repressible promoters directing R-luc expression without any repressors, were also constructed to monitor the maximum R-luc expression level (strength of the promoter). Electro-competent *E. coli* strain DH5α was used for all cloning purposes. Primers were synthesized by IDT. PCR reactions were performed using Herculase II fusion DNA polymerase (Agilent Technologies). All restriction enzymes were purchased from NEBioLabs and

Thermo Scientific. Plasmid preparations and gel extractions were conducted using Thermo Scientific GeneJET and Zymo Research miniprep and gel purification kits. All synthetic designs were sequence verified. DNA sequencing was provided by the Colorado State University Proteomics and Metabolomics Facility.

A.2 Protoplast Isolation and Transformation

Arabidopsis protoplast isolation and transformation were carried out according to the protocol described by Yoo *et al.* [8], with some modifications to allow higher throughput testing of synthetic components in 96-well plates. Wild-type *Arabidopsis thaliana* ecotype Columbia plants were grown in short days (10h light, 14h dark), and 20-25 leaves, approximately 4 cm in length, were used. In brief, leaves in W5 solution were cut into approximately 1 mm strips using a scalpel blade. Enzyme solution [0.4 M Mannitol, 20 mM KCl, 20 mM MES (pH 5.7), 1.5% Cellulase R-10 (Yakult Honsha), 0.4% Macerozyme R-10 (Yakult Honsha), 10 mM $CaCl_2$, 1 mg/ml BSA] was added, a slight vacuum was applied, and incubated at room temperature with gentle shaking (40 rpm) for 3 hours. Resulting protoplasts were filtered through a 70 µm cell strainer (BD Biosciences) and harvested by centrifugation at 600 *x g*. After two washes in W5 solution, the protoplasts were re-suspended in MMg solution, and the concentration adjusted to 2 x 10^5 protoplasts/ml. Protoplast transformation with plasmids of interest was performed in 15-ml conical centrifuge tubes by carefully mixing 50 µl of protoplasts (approx. 10,000 cells), 5 µl of plasmid DNA (1 µg/µl), and 55 µl of 40% PEG solution for one reaction. Larger-scale (14 reactions) transformations were used to allow testing of multiple concentrations of inducers. Transformed protoplasts were re-suspended in 200 µl of WI solution per reaction, and plated on black, clear-bottom, 96-well Costar assay plates (Corning), using a multi-channel pipette.

Inducers (4-OHT or dexamethasone) were added using a multi-channel pipette, and plates were incubated overnight in the dark, with gentle agitation (50 rpm). For transient assays aimed at validating our method in stable transgenic plants, initial experiments showed that protoplasts prepared from transgenic lines typically had reduced signal. Hence, we increased the number of cells per well when protoplasts were prepared from transgenic plants relative to the number of cells per well when protoplasts were prepared for transient expression assays.

A.3 Luciferase Imaging

All test plasmids used in this work had Firefly and *Renilla* luciferases as the measurable outputs. Therefore, we used the Dual-Luciferase Reporter Assay system (Promega) to lyse the protoplasts and provide both substrates for luciferase imaging. After overnight incubation of protoplasts with the inducers, cell lysis was carried out on the assay plates by removing 160 µl of supernatant from each well, followed by addition of 50 µl of 2x Passive Lysis Buffer, and incubation at room temperature for 30 minutes. Quantitative measurements of Firefly and *Renilla* luciferase expression were obtained by the addition of LAR II and Stop & Glo reagents, respectively, and imaged using a Stanford Photonics XR/Mega-10 ICCD Camera System and available Piper software (v. 2.6.17). Regions of interest, ROI's, are drawn around each well of a 96-well plate. Pixel intensity values for the first minute of collection time are summed and divided by the area of the ROI and time collected to give us the RLU/(area x sec) value in each well. The data then go through post-image correction (below). We remove assays that had obviously failed for reasons such as failure to show an increase in F-luc expression with the addition of inducers.

A.4 96-well plate post-image correction

First, we determine five primary systematical parameter values: 1) r_1 (Radius of well opening); 2) r_2 (Radius of well bottom); 3) h_1 (Height of the camera relative to the surface of the 96 wells); 4) h_2 (Depth of well); 5) V(Reaction reagent total volume). Second, we feed these parameters into the function $V(d)$ (Appendix B.1, B.2) to yield the secondary parameter d (depth of solution added in the well). Third, we substitute values of r_1, r_2 and h_1 into function A_v (Appendix B.1,B. 2) resulting in $A_v(s, D)$. Fourth, we substitute h_2 and d as the lower and upper integration limits, respectively, for the integration of $A_v(s)ds$, resulting in $V_{vtotal}(D)$. Here, the function is fully parameterized and the only input needed is D, the positional parameter corresponding to each well in this algorithm. We assume the well in the *i*-th row, *j*-th column from the top left corner of the microplate (as the origin of 96-well-plate plane) holds a coordinate of (x_{ij}, y_{ij}) and the projected camera center onto the plate holds a coordinate of (x, y). Then, D_{ij} for the well (*i*, *j*) can be calculated as $D_{ij} = \sqrt{(x_{ij} - x)^2 + (y_{ij} - y)^2}$. Substitute D_{ij} into $V_{vtotal}(D)$ to generate the total visible volume for well (*i*, *j*). The ratio of this total visible volume to V (total reagent volume) is then used for camera correction.

A.5 Noise Estimation

The noise (Fig. 3.2c) was calculated as follows. For within-plate noise (source 1), we calculated the standard deviation of F-luc and R-luc luminescence for each plate independently. This gave us a measure of between-well noise for a single plasmid on a single plate. For between-transformation noise (source 2), we calculated the standard deviation between the mean R-luc values coming from the two different inducible genes (DEX- and OHT-inducible) on the same day (as R-luc expression is controlled by the same repressible promoter in both gene

circuits). Finally, we calculated the standard deviation of the mean luminescence between days for both F-luc and R-luc. This gave us a measure of the batch effect (source 3).

A.6 Data Analysis

Data is processed in the following steps using MATLAB.

(1) Camera-corrected F-luc RLU/(area x sec) and R-luc RLU/(area x sec) values, and inducer type and concentration are stored in different .csv files for each promoter tested.

(2) DEX and OHT data are separated.

(3) Fold Change (FC) values are calculated for each promoter. Promoters with a FC > 1.3 are stored for further analysis.

(4) Data from promoters that do not meet the threshold criteria are tagged and kept for further processing.

(5) RLU data are converted to Molecule Number per well via the RLU vs. concentration standard curves (Fig. B.4).

(6) The mean F-luc values at zero inducer concentration of those plasmids showing a FC > 1.3 are calculated for both DEX- and OHT-based systems.

(7) Values of the normalization factor λ_i are calculated using Equation 3.3 (Main Text) and the data values of F-luc and R-luc molecule number divided by this factor.

(8) Data are fit to the functional form given in Equation 3.2 with different initial conditions for the nonlinear fit (we used 4 different initial conditions to ensure convergence). If the fits converged to different minima, we chose the fit with the lowest *P*-value. The parameters of the fit are then stored, and can be used for further analysis.

A.7 Stably Transformed Plants

Select promoter-repressor pair genetic circuits (DEX-inducible LexA-EAR + 35S2xLexA, noted as *DEX 35S2xLexA EAR*; OHT-inducible LexA-EAR + NOS2xGal4, noted as *OHT NOS2xGal4 EAR*; DEX-inducible Gal4-EAR + 2xGal4NOS, noted as *DEX 2xGal4NOS EAR* in Fig. 3.6) were sub-cloned into pCAMBIA2300 plant transformation vector and stably transformed into *Arabidopsis thaliana* ecotype Columbia plants by *Agrobacterium* floral dip method [9]. We selected transgenic plants in kanamycin-containing media, and screened for F-luc expression (indicative of repressor expression) in the presence of the inducer (OHT or DEX) and luciferin. Plants not expressing F-luc were discarded, whereas F-luc expressing plants were allowed to set seed. Second generation (T1) plants were germinated in kanamycin-containing media and transgenic lines segregating 3:1 (resistant:sensitive), indicating one copy of the transgene, were used for further analysis.

REFERENCES

[1] S. e. a. Wang, "Arabidopsis ovate family proteins, a novel transcriptional repressor family, control multiple aspects of plant growth and development," *PLoS One,* vol. 6, p. e23896, 2011.

[2] M. Padidam and Y. Cao, "Elimination of transcriptional interference between tandem genes in plant cells," *Biotechniques ,* vol. 31, pp. 328-330, 2001.

[3] A. Depicker, S. Stachel, P. Dhaese, P. Zambryski and H. M. Goodman, "Nopaline synthase: transcript mapping and DNA sequence," *J. Mol. Appl. Genet.,* vol. 1, pp. 561-573, 1982.

[4] C. Engler, R. Gruetzner, R. Kandzia and S. Marillonnet, "Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes," *PloS one,* vol. 4, p. e5553, 2009.

[5] M. S. e. a. Antunes, "A synthetic de-greening gene circuit provides a reporting system that is remotely detectable and has a re-set capacity," *Plant Biotechnol. J.,* vol. 4, pp. 605-622, 2006.

[6] M. Samalova, B. Brzobohaty and I. Moore, " pOp6/LhGR: a stringently regulated and highly responsive dexamethasone-inducible gene expression system for tobacco," *Plant J.,* vol. 41, pp. 919-935, 2005.

[7] Y. e. a. Sakuma, "Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression," *Plant Cell,* vol. 18, pp. 1292-1309, 2006.

[8] S. Yoo, Y. Cho and J. Sheen, "Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis," *Nature Protoc.,* vol. 2, pp. 1565-1572 , 2007.

[9] S. J. Clough and A. F. Bent, "Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana," *Plant J.,* vol. 16, pp. 735-743, 1998.

[10] K. A. Schaumberg, M. S. Antunes, T. K. Kassaw, W. Xu, C. S. Zalewski, J. I. Medford and A. Prasad, "Quantitative characterization of genetic parts and circuits for plant synthetic biology," *Nature Methods,* vol. 13, pp. 94-100, 2016.

APPENDIX B[5]

Methods for Quantitative characterization of genetic parts and circuits for plant synthetic biology

B.1 Luminescence imaging correction

False-colored images of the protoplast luminescence collected from 96-well plates appeared to show a systematic difference based on well position. To measure the extent of this difference we designed a simple "flip-plate" experiment. Luminescence was collected for 5 minutes with well A1 in the top left hand corner and again for 5 minutes with well A1 in the bottom right hand corner. We used the first minute of each collection time to calculate the Relative Luminescence Units (RLU) for each well. We then calculated the percent change of F-luc (in RLU/(area x sec)) between the two values for each well. We repeated the experiment using purified recombinant F-luc protein diluted to give luminescence values in the range of our protoplast data. In both cases, we found substantial differences between the measured luminescence of the two positions for the outer wells. The graph in Fig. B.2b depicts the change in luminescence within the wells for one flip-plate experiment. The two measurements of the plate are superimposed, such that two measurements of the same well are plotted on top of each other (*i.e.*, A1 when imaged near the top of the camera's field of view, superimposed with

[5] Appendix B comprises of supporting data that was published alongside the paper [9] in Nature Methods. It represents joint work and has been put here in its entirety to preserve the intellectual coherence of the project. In particular, the experiments outlined in B.1 was conducted by co-authors from the Medford lab and myself. The camera correction method B.2, collection of simulated data B.5 and 8 were conducted by co-author Wenlong Xu in the Prasad lab. Also the Quantitative analysis of the design elements, B.10 was conducted by Wenlong Xu and myself.

A1 imaged near the bottom of the camera's field of view). As seen in the graph, luminescence values of the wells on the left-hand side of the plate were consistently lower than values measured on the right-hand side of the plate. This experiment was repeated three times with results showing an average maximum percent change of 36% with a standard deviation of $\pm$ 9%.

Since we imaged for five minutes (though only used the first minute of data for comparisons), we calculated the possible natural decrease of the luminescence signal during this time. Our data show that the F-luc signal decreased an average of 7% over a five-minute period (Fig. B.2a). Hence, we assumed that F-luc degradation would lead to a signal decrease of about 7%. We theorize that the most likely reason for these systematic imaging errors is that the camera does not pick up as many photons from wells that are farther away from its central axis when compared to wells that are closer to the central axis. This could be seen from the dark crescents in the images themselves (Fig. B.2c), suggesting a slight blocking effect by the non-transparent walls of the wells. The further away the well is from the projection of camera center on the plate, the larger the portion of its total volume is blocked by its wall, and consequently a smaller portion of the total F-luc luminescence is collected by the camera.

We developed a post-imaging mathematical correction method to correct the images based on a geometric calculation of this "missing volume" and physical measurements on our imaging system (details of the calculation are in Fig. B.3). We developed a formula for the percentage difference between the original luciferase level and the level registered by the camera for each of the wells on a plate, as a function of the distance between the geometric center of each well on the plate and the projection of camera center, given a chosen shelf height for the plate.

Another complication arises when the camera center does not coincide with the center of the 96-well plate. We estimated where the camera center lies from the pattern of percent changes for each well on the plate, since the wells closest to the camera center should have the minimum percent changes (zero if the camera center lies directly above any well). We corrected the luminescence data by using this formula to calculate the corrected luminescence of each well from its position on the plate based on the observed intensity. In addition, we built a frame for the 96-well plate that we used for all subsequent imaging in order to keep the plate center in a fixed position in relation to the camera center.

B.2 Image correction method

The formula derived below for estimating the imaging correction is based on the 96-well plate geometry (Fig. B.3).

1. Calculate the distance D between the center of the targeting well and projection of the camera center using similar triangles (Fig. B.3a):

$$\frac{h_1}{h_1 + h_2} = \frac{D - r_1}{D + l - r_2}$$

Yields,

$$l = \frac{h_2}{h_1}(D - r_1) - (r_1 - r_2)$$

Then we projected the upper edge of the well to the bottom along the sight line between the camera and its closest point on the upper edge.

l is the shift distance on the well bottom of the closest point along the sight line.

The visible portion of the bottom edge is the part enclosed by the projection of the upper edge and itself. We calculated the area of this portion as follows:

This portion can be separated into two parts, A_1 and A_2, by the connecting line between the two intersections of the two circles. A_1 and A_2 can then be calculated using the differences of their corresponding sectors and triangles (Fig. B.3b).

Before we calculate the areas, we need the lengths of y_1, y_2 and x via the equations listed as follows:

$$\mathrm{y}_1 + \mathrm{y}_2 = r_1 - r_2 + l$$

$$r_1^2 - y_1^2 = r_2^2 - y_2^2 = x^2$$

Yields,

$$\mathrm{y}_1 = \frac{r_1^2 - r_2^2 + (r_1 - r_2 + l)^2}{2(r_1 - r_2 + l)} = \frac{r_1^2 - r_2^2 + a^2}{2a}$$

$$\mathrm{y}_2 = a - \mathrm{y}_1 = \frac{a^2 - r_1^2 + r_2^2}{2a}$$

$$\mathrm{x} = \sqrt{r_1^2 - y_1^2}$$

With $a = r_1 - r_2 + l$;

Based on these equations, we calculated the central angles of these two sectors:

$$\alpha_1 = \arccos(\frac{\mathrm{y}_1}{\mathrm{r}_1})$$

$$\alpha_2 = \arccos(\frac{\mathrm{y}_2}{\mathrm{r}_2})$$

The areas of the two sectors can be expressed as:

$\frac{1}{2}\alpha_1 r_1^2$ and $\frac{1}{2}\alpha_2 r_2^2$

The two portions of the visible area on this plane can be calculated as:

$$A_1 = \frac{1}{2}\alpha_1 r_1^2 - y_1 x$$

$$A_2 = \frac{1}{2}\alpha_2 r_2^2 - y_2 x$$

The total visible area on the bottom is:

$$A_v = A_1 + A_2 = \frac{1}{2}\alpha_1 r_1^2 - \mathrm{y_1 x} + \frac{1}{2}\alpha_2 r_2^2 - \mathrm{y_2 x} = \frac{1}{2}\alpha_1 r_1^2 + \frac{1}{2}\alpha_2 r_2^2 - a\mathrm{x}$$

To get the visible volume from the visible area, we integrated from the bottom of the well to the liquid surface. Therefore, we needed to calculate the depth of the reagent inside the well. Taking note of the "imaginary cone" (Fig. B.3c), this integral can be set up using the three steps described below:

1) To calculate the height of the "imaginary cone" by similar triangles for the integration upper limit:

$$\frac{h_i}{h_i + h_2} = \frac{r_2}{r_1}$$

We can calculate h as:

$$h_i = \frac{r_2 h_2}{r_1 - r_2}$$

Also from another pair of similar triangles:

$$\frac{h_i}{h_i + d} = \frac{r_2}{r}$$

Results in:

$$r = \frac{h_i + d}{h_i} r_2 = \frac{\frac{r_2 h_2}{r_1 - r_2} + d}{\frac{r_2 h_2}{r_1 - r_2}} r_2 = \frac{r_2 h_2 + d(r_1 - r_2)}{h_2} = r_2 + \frac{d}{h_2}(r_1 - r_2)$$

Reagent volumes are derived from the protoplast transformation protocol, and we can employ this to calculate the depth using:

$$V(d) = \frac{1}{3}\pi r^2(h_i + d) - \frac{1}{3}\pi r_2^2 h_i = \frac{1}{3}\pi(r_2 + \frac{d}{h_2}(r_1 - r_2))^2(h_i + d) - \frac{1}{3}\pi r_2^2 h_i$$

2) Change h_2 in the expression for bottom visible area into a variable s, as the distance between the top circle and the current plane. This gives us the infinitesimal visible volume as:

$$A_v(s)ds$$

3) We integrate these elements from the bottom of the well to the surface of the liquid to get the total visible volume as:

$$V_{vtotal} = \int_{h_2}^{d} A_v(s)ds = V_{vtotal}(D)$$

B.3 Testing the sources of noise

We prepared protoplast transformations with one DEX-inducible gene circuit and one OHT-inducible gene circuit (enough for 48 wells each). Some wells were emptied and frozen for further studies. We collected luminescence data with no inducer added, and repeated the experiment on three different days with three different batches of protoplasts. In the absence of any noise, all wells should display identical F-luc and identical R-luc luminescence, with R-luc expression at its maximum (since no repressor should be present). Thus, variations between luminescence values from wells containing the same gene circuit on the same plate represent within-plate noise (the first source of noise). The difference between the mean R-luc luminescence measured from the DEX-inducible gene circuit and the OHT-inducible gene circuit in the same batch represents the between-transformation noise (the second source of noise).

Finally, the difference between the mean luminescence of the three batches represents the between-batch noise (the third source of noise).

Because the 'batch effect' is a random variation that affects the entire population of protoplasts in a batch, we can represent it mathematically by a random number α such that the observed luminescence in the *j*-th well of the *i*-th batch can be represented by $R_{ij} = \alpha_i B_R + \delta_{ij}$ and $F_{ij} = \alpha_i B_F + \delta'_{ij}$, for R-luc and F-luc luminescence, respectively. Here, B_R, B_F are the steady state number of luciferase molecules in the well in the absence of any noise for the R-luc and F-luc promoters, respectively; α_i is a random number that represents a multiplicative batch effect, while $\delta_{ij}, \delta'_{ij}$ are random variables that represent additive noise terms that could arise from the remaining noise sources. If we average the R-luc and the F-luc luminescence for each batch and plot them, we are plotting $\alpha_i B_R + \langle \delta_{ij} \rangle_j$ against $\alpha_i B_R + \langle \delta'_{ij} \rangle_j$ (where the subscript on the angled brackets indicates the index being averaged). If this plot is approximately linear, we can conclude that the batch effect is identical for both R-luc and F-luc, and dominates the additive noise terms.

B.4 Conversion of Luminescence Values to Physical Units

We experimentally characterized the function of our promoter-repressor pairs using luminescence from two types of luciferase. Luminescence values are typically reported in RLUs, or relative luciferase units. For our collection system (Stanford Photonics ICCD Camera), RLU is the sum of pixel intensity values within an area over collection time (*i.e.*, RLU/(area x sec)), and represents the activity of F-luc and R-luc for each protoplast sample. We converted RLUs to molecules of luciferase by quantifying the relationship between the luminescence and the luciferase activity using purified recombinant F-luc and R-luc. We plotted standard curves to

convert from RLU values to an absolute number of molecules for both F-luc and R-luc (Fig. B.4). Our standard curves are linear, with high R^2 values (0.97 and 0.96, respectively). We found that there is a difference in the number of molecules of R-luc or F-luc that generate the same RLU value. We used these standard curves with our image-corrected data to provide absolute molecule numbers for our mathematical analysis.

B.5 Testing the normalization scheme with simulated data

To generate simulated data, we first calculated single-plasmid data using Equation 3.1 with assumed parameter values. Then the single-plasmid data were multiplied by a normally distributed random number representing the number of plasmids in each well (N_{ij}), and another random number drawn from a lognormal distribution representing the batch effect factor (α_i). The latter was assumed smaller than one, based on our analysis described in the main text.

For simplicity, we set all the constants C_1, C_2 and $\tilde{C}$ to 1. We simulated 1,000 sets of data, consisting of six inducer levels and two technical replicates, similar to our experimental data. For each set we chose one value of α from a log normal distribution with a mean less than one. Because the lognormal distribution is unbounded in the positive infinity direction, we assumed a 95% cut-off for the distribution of α. To test the normalization scheme with different levels of noise, we increased the standard deviation to obtain a series of distributions with a decreasing population mean and increasing variance of α_i. Since each well in our experiments had approximately 10,000 protoplasts, we set this to be the mean of N_{ij} and simulated various levels of noise by changing the standard deviation of the normal distribution.

This procedure produced fits with an unreasonably high Hill coefficient at high levels of noise in the simulated data. We therefore imposed the criteria that the fitted Hill coefficient of the repressible promoter should lie between 0 and 6. Due to the high levels of noise we can artificially generate in the simulated data, there are also "bad fits" within $0 < n < 6$. These can be further characterized by unreasonably high fitted values of B which are far away from the well-formed distribution of most fitting results. We observed that the fitting results of each parameter form lognormal distributions similar to the assumed distribution of α. Therefore, we carried out logarithmic transformation to the fitted values of B and applied outlier tests following Peirce's method [1]. Specifically we used the R-code written by Dardis and Muller (https://r-forge.r-project.org/projects/peirce/), which extends the development of Peirce's method by S. M. Ross [2].

Fits that met the criterion of n and pass the outlier tests were deemed successful and this defined the Number of Successes (NOS) among the 1000 repeats carried out. Within these biologically feasible results, we compared the mean and standard deviation of the three fitted parameters, namely B, H and n in Equation 3.1 and 2. The variation in the parameters' magnitude is a measure of the effect of experimental noise on our estimates. We therefore plotted the estimated parameters coefficient of variation against the level of noise introduced in the simulated data (Fig. B.5). We found that our normalization procedure can indeed, as expected, reduce the coefficient of variation of the estimates of B and H between different log standard deviations of the alpha distribution, and thus make them more comparable. However, the estimates of the Hill coefficient n were not improved by our normalization.

B.6 Fitting the data and selecting plant gene circuits

We implemented a number of quality control steps to filter out the gene circuits whose behavior was questionable, as described below. First, we eliminated all gene circuits whose F-luc value with all inducer levels was below a designated threshold. These assays were assumed to have failed for various reasons and the data not usable. Our criteria for this threshold were: first, at least one well within the transformation had to produce a signal above the bottom 10% of our data. Second, our synthetic promoters had to demonstrate a reasonable fold-change when the repressor is produced. We defined a 'reasonable' change as having a fold-change of at least 1.3-fold in the output (R-luc), as calculated from the lowest amount of the repressor (average of three lowest F-luc values) to the highest amount of the repressor (average of three highest F-luc values). Third, we examined our data for biologically meaningful Hill coefficient (*n*) values, since our simulated results had shown that we had the most uncertainty in our estimates of this parameter. With these criteria, we selected the gene circuits whose fitted *n* values lay between 0 and 6. A Hill coefficient < 0 would indicate the repressor is acting more like an activator than a repressor, while a Hill coefficient of 6 or greater is implausible in our system. We then selected the best performing repressible promoter gene circuits from those that we assembled.

B.7 Normalization for comparison with stably transformed plants

A key difference in the mathematical description of protoplasts prepared from the stably transformed plants is the number of working circuits each protoplast contains. We used genetic segregation data to select for plants segregating for a single T-DNA insertion. In the T1 generation used in our study, heterozygous plants self-fertilized producing both homozygous

and heterozygous progeny in a 1:2 ratio. Hence, homozygous progeny contain two copies, and heterozygous plants contain one copy of the genetic circuit. However, for the transient protoplast assay, we expected that on average multiple copies of the plasmid would be found in each transformed protoplast, which was confirmed by our data (Fig. 3.6b,c). In the no-inducer treatment, R-luc luminescence levels are just over 4-fold smaller in protoplasts from stably transformed plants compared to transiently transformed protoplasts, despite the fact that the initial cell density of the former is five times greater than the latter. Since the parameter *B* in Equation 3.1 is proportional to the average number of viable plasmids $\langle\alpha_i\rangle\langle N_{ij}\rangle$ for the gene circuit, estimates of *B* from transient data are expected to be overestimates of *B* for stable gene circuit. In agreement with this expectation, tests on simulated data with varying levels of mean N_{ij} showed that B_i was systematically overestimated as the mean number of plasmids became larger (Fig. B.8). In order to correct for this overestimation, we normalized our stably transformed plant data with the mean of the distribution coming from the transient assay. In other words, we defined a normalization factor λ_i^* such that:

$$\lambda_i^* = \frac{\langle F_{i1}^s \rangle_r}{\langle F_{i1}^t \rangle_{ir}} = \frac{\langle N_{i1}^s \rangle_r \, \alpha_i}{\langle N_{i1}^t \rangle_{ir} \langle \alpha_i \rangle}$$

Here, the superscript *t* refers to the transient assay, and *s* refers to the stable transformation assay. The subscripts on the angled brackets indicate the index over which the average is being taken (*r* refers to technical replicates). Dividing the data by λ_i^* therefore accomplishes two goals. First, it replaces α_i by $\langle\alpha_i\rangle$. In addition, it multiplies each F-luc and R-luc value by the fraction by which the plasmids in a transient protoplast average well exceed those in assays

with protoplasts prepared from stably transformed plants (*i.e.*, the fraction $\frac{\langle N_{i1}^{t} \rangle_{ir}}{\langle N_{i1}^{s} \rangle_{r}}$). Tests on simulated data (Fig. B.7) show that the estimates of *B* and *H* obtained by this method are insensitive to changes in the mean of the plasmid number N_{ij} and therefore allow comparison of transient assays with stably transformed assays.

B.8 Testing the normalization factor λ_i^* with simulated data

As previously described, protoplasts prepared from stably transformed plants segregate for a single insertion of the gene circuit, whereas protoplasts prepared via transient assays contain multiple copies of the gene circuit. This leads to different multipliers found in parameter *B* in Equation 3.1 and hinders direct comparisons of the estimated parameter values between transient and stably transformed assays. In the main text, we proposed a normalization factor λ_i^* to correct this bias from plasmid numbers. We then tested if this normalization factor λ_i^* behaved as expected using simulated data (similar to Fig. B.5).

Due to the differences between the transient and stably transformed assays, the noise levels should be positively proportional to the mean numbers of plasmid in each protoplast. Therefore, the coefficient of variance (COV) can be assumed to be the same between the transient and stably transformed assays, whereas the absolute levels should be different. Hence, we varied the standard deviation and mean values at the same time and kept the COV the same (Fig. B.7). To observe the trend clearly, we carried out simulations with five decreasing absolute noise levels. We then normalized the simulated data at each level as discussed above. We applied the same fitting procedure and measurements to both the normalized dataset and its corresponding raw dataset. Similar to what we observed in Figure B.5, fitting of *n* is insensitive to the normalization we applied. As expected, decreases in mean

fitted values for *B* and *H* were observed for the raw data. In contrast, the mean fitted values for *B* and *H* in the normalized data were at similar levels across all five noise levels simulated. This shows our proposed normalization factor λ_i^* meets our expectation and makes different absolute levels comparable.

B.9 Bootstrapping data analysis of transient vs. stable transformants

Bootstrapping statistical analysis was carried out to generate mean values and confidence intervals for the predictions in stably transformed plants. Bootstrapping is a useful inference method when the underlying distribution of the data is not known or when the sample size is small [3]. Bootstrapping was used here to test whether the predictions (Fig. 3.6) would still be the same if the data had been sampled differently.

To generate the different sample sets, the original data set was randomly selected to form bootstrap sample sets in the following three steps.

1. We chose the appropriate number of bins to histogram F-luc values. The chosen number of bins was the largest number that yielded no bins with zero values in them. This was done to optimize the sampling of the data.
2. We chose the number of sample points to draw from each bin. This number was set to be one greater than the minimum number of points in any bin, to avoid drawing the same point an excessive number of times per sample. For example, if one bin in the F-luc histogram contained only one point, the maximum number of sample points that could be drawn from any bin was set as 2.
3. We made histograms of the F-luc data with the number of bins chosen in step 1, and the corresponding R-luc data were placed in a corresponding R-luc bin. Bootstrapped

samples were created by drawing the number of sample points fixed in step 2 from each bin, randomly and with replacement.

Five hundred bootstrap samples were created separately from data obtained in transient assays and from stable transgenic plants, for each gene circuit. Each bootstrap sample was fit using our standard procedure, and the parameters *B*, *H* and *n* estimated. This exercise produced a distribution of fitted values for each parameter. These distributions appear to have large outliers (Fig. B.8a). Outliers that were 3 standard deviations or greater away from the mean were identified and removed. Mean values and confidence intervals were then calculated from the remaining distribution. The lower and upper bound for each confidence interval were the 5% and 95% values from the final bootstrapped distribution, given a 90% confidence interval.

The results of the bootstrapping exercise are shown in Figure B.9b-d. To summarize these results:

- The predictions for *B* appeared to be in the same range and showed the same trend as the original fits.
- The predictions for *H* appeared to be in the same range and gave a similar, if not better, comparison between stable and transient data as in the original fits.
- The mean value for the predictions for *n* lay within at least a factor of 2.56 between the stable and transient data. However, the increased confidence intervals suggest that this parameter may be more difficult to recover, as suggested by the simulated data.

B.10 Quantitative analysis of design elements

B.10.1 Outline of Method

Each promoter-repressor gene circuit we constructed had seven design elements of our test system that we experimentally characterized. We analyzed our data for statistical patterns that associate design elements, and their combinations, with satisfactory performance and allowed derivation of design principles.

The seven design elements in our test genetic circuit are:

1) Inducible promoter (controlling repressor levels);
2) DNA-binding domain (*Gal4* or *LexA*);
3) Repressor motif;
4) Constitutive promoter scaffold;
5) Number of binding sites;
6) Location of binding sites;
7) Use of spacer DNA inserted between the binding sites.

Each design element can be used to divide the synthetic gene circuits into categories that reflect the choices of that design element. For example, there were two options for DNA-binding domain; hence, all promoter-repressor pairs can be divided into those that use Gal4 and those that use LexA DNA-binding domain. We compared repressible promoter performance between LexA-based and Gal4-based promoter-repressor pairs. To make this comparison, we defined six measures of promoter performance, two qualitative and four quantitative.

1) Success ratio (number of gene circuits with good fits/total number of promoters tested for each category);
2) Mean rank of gene-circuits ranked by fold-change (mean of the rank of all gene circuits in a category);
3) Fold-change of R-luciferase expression;
4) *B* parameter value (maximal expression of the repressible promoter);
5) *H* parameter value (amount of repressor needed to reduce expression by half);
6) *n* parameter value (sigmoidality of the input-output relation).

We specifically focused on two aspects:

1. Among the gene circuits that met the criteria for good promoter-repressor combinations, we examined the data for patterns that are associated with particular performance measures (see details and results of the ANOVA).

2. We sought differences between genetic components that met our criteria and those that did not meet our criteria.

B.10.2 Comparison among functional gene circuits

For the first aspect, we carried out an ANOVA analysis based on the four quantitative measures of promoter performance (items 3-6 above). Because of our limited data set, we assumed no multicollinearity. In this case, when analyzing the effect of one given design element, we assumed the effect of the remaining design elements averages out. For example, when analyzing the effect of the three distinct promoter scaffolds (CaMV35S, FMV, NOS), the effect of the other design elements (numbers and positions of DNA binding elements) average out. When possible, we further checked our assumption of no mulitcollinearity by dividing the

data set by element. For example, we separated data for DEX-inducible promoters from data for OHT-inducible promoters. Also, when possible, we ensured an observed trend correlated with known biological function. For example, the monocot promoter ZmUbi1 is known to be a very strong constitutive promoter [4], and our analysis shows that this promoter was found to have a large *B* value compared to CaMV35S, indicating a strong constitutive expression.

B.10.3 Details and results of the ANOVA

We carried out a one-way ANOVA on each of the design elements described above for the Arabidopsis normalized protoplast data. Fold-change and the parameter values of *B*, *H* and *n* were used to verify if there was a significant difference between the groups for each element. We used log(*B*) and log(*H*), since the bootstrap analysis showed that distributions of both *B* and *H* were approximately lognormal. We found that the inducer used had a significant impact on Fold-change and *n* value performance. This was expected because they represent two different biological systems, and because the normalization constant between the two inducible systems is also different. We therefore re-ran the one-way ANOVA for each of the six remaining design elements separately for DEX- and OHT-inducible systems (Table B.1). This analysis yielded two design elements to explore further: (i) the number of binding sites used, and (ii) the location of the binding sites.

We used a HSD (Honestly Significant Difference) Tukey test to perform a sequential comparison of all subgroups for these two design elements. We found a significant difference for the number of binding sites used between 2x and 4x for DEX-based circuits, and between 2x and 5x for OHT-based circuits. As 4x and 5x binding sites were only present in promoters with binding sites positioned just upstream of the TATA box, this significant difference could be due

to positioning of the binding site, rather than their copy number. As this trend in the *B* value was not seen with promoters containing 8x binding sites, the observed difference is likely due to positioning, and not binding site copy number. The HSD Tukey results for binding site position suggested that binding sites positioned just upstream of the TATA box have lower and significantly different *B* values compared to binding sites before or after the constitutive promoter scaffold (Fig. B.9a).

ANOVA results for the normalized sorghum data can be found in Table B.2. As we found in Arabidopsis, our results are most notable when there is significance with multiple genetic circuits (*e.g.*, DEX and OHT induction). The four elements that showed significance in both gene circuits are:

(i) The *B* value showed significance in terms of binding site copy number. However, similar to what we found in Arabidopsis, this difference seen in the binding site number is likely due to the positioning of the TATA box.

(ii) The *n* value showed significance in terms of binding site number. However, the HD Tukey results for the *n* value with the copy number showed different significant comparison between the DEX and OHT data.

(iii) The *B* value showed significance in terms of the position of the binding elements. The HD Tukey results were similar to those observed in Arabidopsis; promoters containing the binding sites just upstream of the TATA box have significantly lower *B* value compared to promoters with binding sites upstream or downstream of the constitutive scaffold (Fig. B.9b).

(iv) The *B* value showed significance in terms of the constitutive elements. The HD Tukey results suggested promoters based on the CaMV35S scaffold have significantly lower *B* values compared to ZmUbi1 promoters (Fig. B.9c).

Comparison between functional and non-functional circuits

For this comparison, we used our two qualitative metrics of performance, success ratio and mean rank. Success ratio was defined as the number of successful components over the total number of components tested. A successful component is one that displays the desirable behavior, in this case, transcriptional repression. We then ranked all the successful circuits according to their fold-change values. For each category, the mean rank was calculated as a non-parametric measurement of circuit performances. All results are outlined in the Tables B.1-4. We have summarized several of these comparisons below.

To facilitate pairwise comparisons, we used Fisher's exact test and Wilcoxon rank sum test to calculate *P*-values for the success ratio and mean rank, respectively. Fisher's exact test was chosen for success ratio due to its validity for small sample sizes and accuracy when sample size is large. The Wilcoxon rank sum test was chosen for mean rank as a non-parametric alternative for testing the null hypothesis that the medians of two samples are the same. The Wilcoxon rank sum test was applied directly to the fold-change data of the pair of interest. For ease of direct comparisons across different pairs for the bulk measurements, mean ranks listed in all the tables were calculated based on the ranks of the entire list of successful gene circuits. However, this caused no difference in the statistical test results. Due to our limited sample size, we chose a significance level of 0.1, which is higher than the conventional significance level of

0.05. In Tables B.3 and B.4, we used notations for different significance levels as * for *P*-value < 0.1, ** for *P*-value < 0.05 and *** for *P*-value < 0.01.

We applied this analysis to the Arabidopsis and sorghum datasets to identify possible design principles that correlated with function of synthetic components in plants. Supporting data are provided in Table B.3 for Arabidopsis circuits and Table B.4 for sorghum circuits.

In Arabidopsis, we found that the DEX-inducible promoter leads to statistically significantly better function than the OHT-inducible promoter, in terms of both success ratio and mean rank. This was also observed in sorghum. Of the three constitutively expressed scaffolds used in Arabidopsis circuits (CaMV 35S, FMV, NOS), FMV is typically regarded as the strongest constitutive promoter [5] [6] [7], and we found that it performed the worst in repressibility. The two DNA-binding domains, LexA and Gal4, worked similarly in terms of both mean rank and success ratio. Interestingly, some combinations of the constitutively expressed scaffolds and DNA-binding domains worked better than others. In particular, CaMV35S functioned statistically significantly better with LexA. Although not statistically significant, it is interesting to note that the NOS scaffold appeared to function better when paired with Gal4. Similarly, some combinations of repressor domains and DNA-binding domains worked better than others. Specifically in terms of mean rank, OFP1/OFPx worked significantly better with LexA. Although not significant there is an interesting trend of B3 working better with Gal4, and there is no obvious preference for EAR. Due to the limits on sample size, we were not able to draw general conclusions for other design elements (*i.e.*, the relative positions of binding elements and constitutive scaffolds, number of binding elements and presence of spacers between binding elements).

To address the genomic differences between monocots and dicots, different constitutive promoter scaffolds (namely ZmUbi1 and OsACT2) were used for sorghum circuits, except for CaMV 35S, which has been reported to also function in sorghum [8]. In sorghum, CaMV 35S was statistically the best scaffold in terms of success ratio and although not significant has a relatively better Mean Rank. Only five ZmUbi1-based circuits were characterized as repressible and none of the OsACT2 met the quality criteria. Although not statistically significant, it is interesting to note that Gal4 showed a trend to be favored in terms of the combinations between the DNA-binding domains and the repressor domain. Overall, two key conclusions we found in both Arabidopsis and sorghum data for function of our synthetic genetic circuit are: (i) CaMV 35S is the best performing constitutive scaffold. (ii) DEX-inducible promoter leads to statistically significantly better function than the OHT-inducible promoter.

Table B.1. Arabidopsis ANOVA results. A one-way ANOVA was carried out for each design element with each quantitative measure of promoter performance. Each row represents a different design element and each column represents a different measure. Only *P*-values < 0.1 are shown. Tests that did not show significance are marked by a dash (-).

Design Element	**DEX (Fold-change)**	**OHT (Fold-change)**	**DEX (*B*)**	**OHT (*B*)**	**DEX (*H*)**	**OHT (*H*)**	**DEX (*n*)**	**OHT (*n*)**
Repressors	-	-	-	*P* = 0.02	-	-	-	*P* < 0.01
Binding sites[a]	*P* = 0.08	-	*P* < 0.01	*P* = 0.05	*P* = 0.01	-	-	-
Constitutive scaffolds	P = 0.01	-	-	-	-	-	*P* = 0.08	-
Gal or Lex	-	-	*P* = 0.04	-	*P* = 0.08	-	-	-
Position[b]	*P* = 0.01	-	*P* < 0.01	*P* = 0.01	-	-	-	-
Spacers	-	-	-	-	-	-	-	-

[a]number of DNA binding elements inserted in the constitutive promoter scaffold;
[b]position of DNA binding elements (upstream of the scaffold, downstream in the 5'UTR, or just upstream of the TATA box).

Table B.2. Sorghum ANOVA results. A one-way ANOVA was carried out for each design element with each quantitative measure of promoter performance. Each row represents a different element and each column represents a different measure. Only *P*-values < 0.1 are shown. Tests that did not show significance are marked by a dash (-).

Design Element	DEX (Fold-change)	OHT (Fold-change)	DEX (*B*)	OHT (*B*)	DEX (*H*)	OHT (*H*)	DEX (*n*)	OHT (*n*)
Repressors	-	-	-		-	-	-	-
Binding sites[a]	-	-	$P < 0.01$	$P = 0.04$	$P = 0.095$	-	$P = 0.07$	$P = 0.01$
Constitutive scaffolds	-	-	$P < 0.01$	$P < 0.01$	-	-	-	-
Gal or Lex	-	-	-		-	-	-	-
Position[b]	$P = 0.08$	-	$P < 0.01$	$P = 0.02$	-	-	-	-
Spacers	-	-	-	$P = 0.03$	-	-	-	-

[a]number of DNA binding elements inserted in the constitutive promoter scaffold;
[b]position of DNA binding elements (upstream of the scaffold, downstream in the 5'UTR, or just upstream of the TATA box).

Table B.3. Supporting data of design principles for Arabidopsis. Bulk measurements applied are success ratio and mean rank. For success ratio, the number inside the parentheses represents the actual number of successful gene circuits over the total number of gene circuits made, with calculated absolute ratio shown outside of the parentheses. *P*-values for mean rank are calculated using Wilcoxon rank sum test and *P*-values for success ratio using Fisher's exact test. a) Inducible promoters. Gene circuits with DEX-inducible promoter are statistically significantly better than the ones with OHT-inducible promoter in terms of both success ratio and mean rank. b) Constitutively expressing scaffold. Overall comparison shows FMV has the statistically significantly lowest success ratio and highest mean rank (statistically significant between FMV and CaMV35S). CaMV35S is the best scaffold in terms of both measurements, with the same performance compared to NOS in terms of success ratio. c) DNA binding domain (LexA/Gal4). LexA and Gal4 perform similarly in terms of both success ratio and mean rank. d) Combinations of constitutive promoter and DNA binding domain (LexA/Gal4). FMV works similarly with both LexA and Gal4 DNA binding elements. CaMV35S works the statistically significantly best with LexA, while there is no statistically significant difference between LexA and Gal4 for the NOS scaffold. e) Repressor domain. Due to high homology between OFP1 and OFPx, a new category is created as OFP1/OFPx by combining these two domains together. B3 works better with Gal4 than LexA in terms of mean rank (big difference but not statistically significant), OFP1/OFPx works statistically significantly better with LexA in terms of mean rank, and EAR has no statistically significant preference. There is no statistically significant advantage for any repressor domain in terms of overall comparisons. f) Positions of binding sites. "Upstream" stands for constitutively-expressing-scaffold-first gene circuit and "Downstream"

for DNA-binding-domain-first gene circuit. Four pairs of direct comparisons for relative order between constitutive promoter and DNA binding domain were made. None of them are statistically significant. g) Overall comparison for positions of binding site. Gene circuits with a spacer were excluded from this analysis. There is no statistically significant difference between "Upstream" and "TATA." "Downstream" performed worse than the other two. More specifically, "Upstream" is statistically significantly better than "Downstream" in terms of success ratio and "TATA" is statistically significantly better than "Downstream" in terms of mean rank. h) Effects of spacer. Eight pairs of direct comparisons were available for this analysis. None of them were significantly different. i) Overall comparison for effects of spacer. Gene circuits without a spacer are statistically significantly better than the ones with spacer in terms of success ratio. j) Effects of binding site copy number. No data for direct comparisons for effects of increase in number of binding sites (*e.g.*, 2 or 8 copies of LexA), as shown in f). The overall comparison shows (gene circuits with spacer excluded) 2 copies of LexA is statistically significantly better than 8 copies of LexA in terms of success ratio. 5xGal4 and 4xLexA are always associated with TATA, so these two are not included for the analysis on copy number of binding site. Notations for significance level: * for *P*-value < 0.1, ** for *P*-value < 0.05 and *** for *P*-value < 0.01.

a) Inducible promoters

	DEX	**OHT**	***P*-value**
Success ratio	0.44 (28/64)	0.22 (14/64)	0.014**
Mean rank	18.2	28.1	0.014**

b) Constitutively expressing scaffolds

1) Bulk measurements:

	CaMV35S	**FMV**	**Nos**
Success ratio	0.44 (14/32)	0.17 (8/48)	0.42 (20/48)
Mean rank	14	29.4	23.6

2) *P*-values:

Comparisons	**Fisher's Exact Test**	**Wilcoxon rank sum test**
CaMV35S vs FMV	0.011**	0.011**
CaMV35S vs NOS	1	0.020**
FMV vs NOS	0.013**	0.21

c) DNA Binding Domain (LexA/Gal4)

	Gal4	**LexA**	***P*-value**
Success ratio	0.31 (20/64)	0.34 (22/64)	0.85
Mean rank	23.3	19.9	0.38

d) Combinations of constitutive promoter and DNA Binding Domain (LexA/Gal4)

1) Bulk measurements:

Constitutive	**DNA Binding**	**Success Number**	**Mean Rank**

CaMV35S	LexA	0.63 (10/16)	9
	Gal4	0.25 (4/16)	26.5
Nos	LexA	0.33 (8/24)	28.9
	Gal4	0.5 (12/24)	20
FMV	LexA	0.17 (4/24)	29.3
	Gal4	0.17 (4/24)	29.5

2) *P*-value from Fisher's Exact Test:

		CaMV35S		**NOS**		**FMV**	
		LexA	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
CaMV35S	**LexA**	-	0.073*	0.114	0.53	0.0059***	0.0059***
	Gal4		-	0.734	0.19	0.69	0.69
NOS	**LexA**			-	0.38	0.32	0.32
	Gal4				-	0.030**	0.030**
FMV	**LexA**					-	1
	Gal4						-

3) *P*-value of Wilcoxon rank sum test:

		CaMV35S		**NOS**		**FMV**	
		LexA	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
CaMV35S	**LexA**	-	0.036**	3.2e-4***	0.019**	0.0040***	0.0080***
	Gal4		-	0.81	0.38	1	0.89
NOS	**LexA**			-	0.11	0.81	0.93
	Gal4				-	0.17	0.17
FMV	**LexA**					-	0.69
	Gal4						-

e) Repressor domain

1) Bulk measurements:

		B3	**EAR**	**OFP1**	**OFPx**	**OFP1/OFPx**
LexA	**Success**	0.38 (6/16)	0.31 (5/16)	0.31 (5/16)	0.38 (6/16)	0.34
	Mean Rank	24.2	19.4	16.4	19	17.8
Gal4	**Success**	0.31 (5/16)	0.56 (9/16)	0.25 (4/16)	0.13 (2/16)	0.19 (6/32)
	Mean Rank	12.2	24.1	28.8	36	31.2
Total	**Success**	0.34	0.44	0.28 (9/32)	0.25 (8/32)	0.27
	Mean Rank	18.7	22.4	21.9	23.3	22.6

2) *P*-values from Fisher's Exact Test for split comparisons:

		B3		**EAR**		**OFP1**		**OFPx**		**OFP1/OFPx**	
		LexA	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
B3	**LexA**	-	1	1	0.48	1	0.70	1	0.22	1	0.18
	Gal4		-	1	0.29	1	1	1	0.39	1	0.47
EAR	**LexA**			-	0.29	1	1	1	0.39	1	0.47
	Gal4				-	0.29	0.15	0.48	0.023**	0.22	0.019**

OFP1	**LexA**					-	1	1	0.39	-	-
	Gal4						-	0.70	0.65	-	-
OFPx	**LexA**							-	0.22	-	-
	Gal4								-	-	-
OFP1/ OFPx	**LexA**									-	0.26
	Gal4										-

3) *P*-values of Wilcoxon rank sum test for split comparisons:

		B3		**EAR**		**OFP1**		**OFPx**		**OFP1/OFPx**	
		Lex	**Gal4**	**LexA**	**Gal**	**Lex**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
B3	**Lex**	-	0.18	0.54	1	0.33	0.76	0.59	0.29	0.35	0.40
	Gal		-	0.31	0.15	0.69	0.016*	0.54	0.095*	0.51	0.0043***
EAR	**Lex**			-	0.61	0.84	0.29	0.93	0.095*	0.83	0.082*
	Gal				-	0.24	0.60	0.46	0.33	0.22	0.33
OFP1	**Lex**					-	0.19	0.93	0.19	-	-
	Gal						-	0.48	0.53	-	-
OFPx	**Lex**							-	0.29	-	-
	Gal								-	-	-
OFP1/ OFPx	**Lex**									-	0.078*
	Gal										-

4) *P*-values from Fisher's Exact Test for overall comparisons:

	B3	**EAR**	**OFP1**	**OFPx**	**OFP1/OFPx**
B3	-	0.61	0.79	0.59	0.48
EAR		-	0.30	0.19	0.11
OFP1			-	1	-
OFPx				-	-
OFP1/OFPx					-

5) *P*-values of Wilcoxon rank sum test for overall comparisons:

	B3	**EAR**	**OFP1**	**OFPx**	**OFP1/OFPx**
B3	-	0.49	0.54	0.49	0.42
EAR		-	0.92	0.92	1
OFP1			-	0.81	-
OFPx				-	-
OFP1/OFPx					-

f) Positions of Binding Site

1) Bulk measurements:

	Gal4			**LexA**				
Copy Number	**2**		**5**	**2**		**4**	**8**	
Relative Order	**Before**	**After**	**TATA**	**Before**	**After**	**TATA**	**Before**	**After**
DEX								
CaMV35S	-	1 (35)	2 (14.5)	4 (5.8)	-	4 (6.8)	-	-

FMV	2 (29.5)	0	-	2 (24.5)	-	-	-	1 (36)
NOS	3 (11)	3 (16.3)	-	-	3 (29.7)	-	-	0
OHT								
CaMV35S	-	0	1 (42)	1 (14)	-	1 (26)	-	-
FMV	2 (29.5)	0	-	0	-	-	-	1 (32)
NOS	1 (40)	2 (20)	-	-	2 (33)	-	-	0

2) *P*-values:

Before vs After	**Fisher's Exact Test**	**Wilcoxon rank sum test**
DEX_FMV_2xGal4	0.43	0.67
DEX_nos_2xGal4	1	0.70
OHT_FMV_2xGal4	0.43	0.67
OHT_nos_2xGal4	1	0.67

g) Overall comparison (excluding gene circuits with spacer) for Positions of Binding Site

1) Bulk measurements:

	Before	**After**	**TATA**
Success ratio	0.47 (15/32)	0.27 (13/48)	0.5 (8/16)
Mean rank	18.5	26.7	15.5

2) *P*-values

	Fisher's Exact Test	**Wilcoxon rank sum test**
Before vs After	0.095*	0.13
Before vs TATA	1	0.77
After vs TATA	0.13	0.039**

h) Effects of Spacer

1) Bulk measurements:

		Gal4		**LexA**	
Copy Number		**2**		**8**	
Spacer		**Yes**	**No**	**Yes**	**No**
DEX					
CaMV35S		-	-	-	-
FMV	**Success**	0 (0/4)	0 (0/4)	0 (0/4)	0.25 (1/4)
	Mean Rank	NA	NA	NA	36
NOS	**Success**	0.25 (1/4)	0.75 (3/4)	0.50 (2/4)	0 (0/4)
	Mean Rank	29	16.3	25.5	NA
OHT					
CaMV35S		-	-	-	-
FMV	**Success**	0 (0/4)	0 (0/4)	0 (0/4)	0.25 (1/4)
	Mean Rank	NA	NA	NA	32
NOS	**Success**	0.50 (2/4)	0.50 (2/4)	0.25 (1/4)	0 (0/4)
	Mean Rank	25	20	25	NA

2) *P*-values for available direct comparisons:

	Fisher's Exact Test	Wilcoxon rank sum test
DEX_FMV_2xGal4	1	NA
DEX_nos_2xGal4	0.49	0.50
OHT_FMV_2xGal4	1	NA
OHT_nos_2xGal4	1	0.67
DEX_FMV_8xLexA	1	NA
DEX_nos_8xLexA	0.43	NA
OHT_FMV_8xLexA	1	NA
OHT_nos_8xLexA	1	NA

i) Overall comparisons for Effects of Spacer

	With Spacer	Without Spacer	*P*-value
Success Ratio	0.19 (6/32)	0.38 (36/96)	0.054*
Mean Rank	25.8	20.8	0.36

j) Effects of binding site copy number

1) Bulk measurements:

	Success Ratio	Mean Rank
2xGal4	0.35 (14/40)	22.5
5xGal4	0.38 (3/8)	23.7
2xLexA	0.50 (12/24)	20.1
4xLexA	0.63 (5/8)	10.6
8xLexA	0.13 (2/16)	34

2) *P*-values:

	Fisher's Exact Test	Wilcoxon rank sum test
2xGal4 vs 5xGal4	1	0.86
2xLexA vs 4xLexA	0.69	0.19
2xLexA vs 8xLexA	0.020**	0.26
4xLexA vs 8xLexA	0.021**	0.095*

Table B.4. Supporting data of design principles for sorghum. Measurements are the success ratio and mean rank. For the success ratio, the number inside the parentheses is number of successful gene circuits over total number of gene circuits made, with calculated absolute number shown outside of the parentheses. *P*-values for mean rank are calculated using Wilcoxon rank sum test and *P*-values for success ratio is calculated using Fisher's exact test. **a) Inducible promoters.** Gene circuits with DEX inducible promoter are statistically significantly better than the ones with an OHT inducible promoter in terms of both the success ratio and mean rank. **b) Constitutive promoters.** Overall comparison shows CaMV35S is statistically significantly the best in terms of success ratio. None of the OSACT2-based gene circuits passed our criteria. **c) DNA binding domain (LexA/Gal4).** LexA and Gal4 perform similarly in terms of the success ratio and mean rank. **d) Combinations of constitutively expressing scaffold and**

DNA binding domain (LexA/Gal4). Other constitutively expressing scaffold were tested but have few gene circuits passing our criteria. CaMV 35S works statistically significantly the best with both LexA and Gal4 in terms of success ratio. **e) List of ZmUbi1-based gene circuits.** Except that there are more DEX gene circuits than OHT (3 versus 2), there is no other observation we can make for these ZmUbi1-based gene circuits. We did not include these five gene circuits in the following analyses. **f) Combinations of DNA binding domain (LexA/Gal4) and repressor domain combinations.** Due to high homology between OFP1 and OFPx, a new category is created as OFP1/OFPx by combining these two together. No statistically significant difference was observed for comparisons split into LexA and Gal4. In overall comparisons, the only statistically significant difference is between OFP1 and OFPx in terms of success ratio. **g) Positions of binding sites.** "Upstream" stands for constitutively-expressing-scaffold-first gene circuits and "Downstream" for DNA-binding-domain-first gene circuits. No direct comparison for TATA was available to make a solid conclusion. There are four pairs of direct comparisons available between "Upstream" and "Downstream." Only the DEX_2xLexA pair shows statistically significant difference in terms of mean rank. **h) Overall comparison for positions of binding site.** Gene circuits with a spacer were excluded from the analysis. The performance of relative position of "Upstream" for CaMV 35S-based gene circuits is nearly perfect in terms of success ratio. Except for the success ratio between "Upstream" and "Downstream," there is no statistically significant difference in other measurements. **i) Effects of spacer.** "Yes" stands for spacers present in the gene circuit, while "No" stands for no spacer in this gene circuit. From the available direct comparisons, having the spacer made the performance of each gene circuit worse in terms of both success ratio and mean rank compared to its corresponding gene circuits without the spacer in terms of the bulk measurements. Among these measurements, the only statistically significant difference is present for DEX_2xGal4 pair in terms of mean rank. **j) Overall comparison for effects of spacer.** The performances with or without spacer seem similar for all 35S-based gene circuits in terms of both success ratio and mean rank. Distinguishing between DEX and OHT-based gene circuits also suggests no statistically significant difference. **k) Effects of binding site copy number.** There are two pairs of direct comparisons as shown in **g**. These two pairs suggest that there is no difference on either measurement. 5xGal4 and 4xLexA gene circuits were always associated with TATA for binding site position, so these two types were not included for the overall comparison. We compared results for two and eight copies of LexA. The overall comparison shows 8xLexA performed worse than 2xLexA in terms of both success ratio and mean rank for bulk measurements, but these two differences are not statistically significant. Notations for significance level: * for *P*-value < 0.1, ** for *P*-value < 0.05 and *** for *P*-value < 0.01.

a) Inducible promoter

	DEX	**OHT**	***P*-value**
Success ratio	0.46 (26/56)	0.27 (15/56)	0.0065***
Mean rank	17.1	27.7	0.049**

b) Constitutively expressing scaffolds

1) Bulk measurements:

	35S	ZmUbi1	OSACT2
Success ratio	0.56 (36/64)	0.16 (5/32)	0 (0/16)
Mean rank	20.0	28.0	NA

2) *P*-values:

	Fisher's Exact Test	Wilcoxon rank sum test
35S vs ZmUbi1	1.6e-4***	0.20
35S vs OSACT2	2.5e-05***	NA
ZmUbi1 vs OSACT2	0.15	NA

c) DNA Binding Domain (LexA/Gal4)

	Gal4	LexA	*P*-value
Success ratio	0.36 (23/64)	0.38 (18/48)	1
Mean rank	18.9	23.7	0.20

d) Combinations of constitutively expressing scaffolds and DNA binding domain combinations (LexA/Gal4)

1) Bulk measurements:

	CaMV 35S		ZmUbi1		OSACT2	
	LexA	Gal4	LexA	Gal4	LexA	Gal4
Success ratio	0.50	0.63 (20/32)	0.125 (2/16)	0.19 (3/16)	-	0
Mean rank	22.5	18.1	33.5	24.3	NA	NA

2) *P*-values from Fisher's Exact Test:

		CaMV 35S		ZmUbi1		OSACT2	
		LexA	Gal4	LexA	Gal4	LexA	Gal4
CaMV 35S	**LexA**	-	0.45	0.013**	0.060*	-	3.0e-4***
	Gal4		-	0.0017***	0.0059***	-	2.0e-5***
ZmUbi1	**LexA**			-	1	-	0.48
	Gal4				-	-	0.23
OSACT2	**LexA**					-	-
	Gal4						-

3) *P*-values of Wilcoxon rank sum test:

		CaMV 35S		ZmUbi1	
		LexA	Gal4	LexA	Gal4
CaMV 35S	**LexA**	-	0.30	0.26	1
	Gal4		-	0.12	0.30
ZmUbi1	**LexA**			-	0.40
	Gal4				-

e) List of ZmUbi1-based gene circuits

DEX_2x_gal4_ZmUbi1_OFP1
DEX_2x_lexA_ZmUbi1_B3

DEX_8x_lexA_ZmUbi1_EAR
OHT_2x_gal4_ZmUbi1_OFPx
OHT_2x_gal4spacer10_ZmUbi1_B3

f) Combinations of DNA binding domain (LexA/Gal4) and repressor domain combinations (only 35S-based gene circuits included)

1) Bulk measurements:

		B3	EAR	OFP1	OFPx	OFP1/OFPx
LexA	**Success Ratio**	0.375	0.63 (5/8)	0.75 (6/8)	0.25 (2/8)	0.50 (8/16)
	Mean Rank	23	22	24.2	18	22.6
Gal4	**Success Ratio**	0.75 (6/8)	0.50 (4/8)	0.75 (6/8)	0.50 (4/8)	0.63 (10/16)
	Mean Rank	23.2	11.5	17.2	18.25	17.6
Overall	**Success Ratio**	0.56	0.56	0.75	0.38 (6/16)	0.56 (18/32)
	Mean Rank	23.1	17.3	20.7	18.2	19.8

2) *P*-values from Fisher's Exact Test for split comparisons:

		B3		EAR		OFP1		OFPx		OFP1/OFPx	
		LexA	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
B3	**Lex**	-	0.31	0.62	1	0.31	0.31	1	1	0.68	0.39
	Gal		-	1	0.61	1	1	0.13	0.61	0.39	0.67
EAR	**Lex**			-	1	1	1	0.31	1	0.68	1
	Gal				-	0.61	0.61	0.61	1	1	0.67
OFP1	**Lex**					-	1	0.13	0.61	-	-
	Gal						-	0.13	0.61	-	-
OFPx	**Lex**							-	0.61	-	-
	Gal								-	-	-
OFP1/ OFPx	**Lex**									-	0.72
	Gal										-

3) *P*-values of Wilcoxon rank sum test for split comparisons:

		B3		EAR		OFP1		OFPx		OFP1/OFPx	
		LexA	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**	**LexA**	**Gal4**
B3	**LexA**	-	1	1	0.40	0.71	0.55	0.80	0.63	0.92	0.47
	Gal4		-	0.93	0.26	0.94	0.39	0.86	0.61	1	0.37
EAR	**LexA**			-	0.41	0.93	0.79	0.86	0.90	0.83	0.77
	Gal4				-	0.17	0.35	0.53	0.49	0.15	0.30
OFP1	**LexA**					-	0.39	0.64	0.48	-	-
	Gal4						-	1	0.91	-	-
OFPx	**LexA**							-	1	-	-
	Gal4								-	-	-
OFP1/ OFPx	**LexA**									-	0.41
	Gal4										-

4) *P*-values from Fisher's Exact Test for overall comparisons:

	B3	EAR	OFP1	OFPx	OFP1/OFPx
B3	-	1	0.46	0.48	1
EAR		-	0.46	0.48	1
OFP1			-	0.073*	-
OFPx				-	-
OFP1/OFPx					-

5) *P*-values of Wilcoxon rank sum test for overall comparisons:

	B3	EAR	OFP1	OFPx	OFP1/OFPx
B3	-	0.44	0.70	0.46	0.52
EAR		-	0.55	0.78	0.55
OFP1			-	0.68	-
OFPx				-	-
OFP1/OFPx					-

g) Positions of binding site (only CaMV 35S-based gene circuits included)

1) Bulk measurements:

	Gal4			LexA				
Copy	2		5	2		4	8	
Positions	**Before**	**After**	**TATA**	**Before**	**After**	**TATA**	**Before**	**After**
DEX								
Success	1 (4/4)	0.25	0.75	1 (4/4)	0.75	0.50	-	0.75
Mean Rank	5.8	11	13	6	26	26	-	25.7
OHT								
Success	1 (4/4)	0.75	0.25	0.75	0 (0/4)	0.25	-	0 (0/4)
Mean Rank	22.3	22.3	41	31	NA	36	-	NA

2) *P*-values:

	Fisher's Exact Test	Wilcoxon rank sum test
DEX_2xGal4 Before vs After	0.14	0.40
OHT_2xGal4 Before vs After	1	0.86
DEX_2xLexA Before vs After	1	0.057*
OHT_2xLexA Before vs After	0.14	NA

h) Overall comparison (excluding gene circuits with spacer) for positions of binding site (only 35S-based gene circuits included)

1) Bulk measurements:

	After	Before	TATA
Success ratio	0.42 (10/24)	0.94 (15/16)	0.44 (7/16)
Mean Rank	23.3	15.3	24

2) *P*-values:

	Fisher's Exact Test	Wilcoxon rank sum test
Before vs After	9.1e-4***	0.10

Before vs TATA	0.14	0.14
After vs TATA	0.89	0.89

i) Effects of spacer (only 35S-based gene circuits included)

1) Bulk measurements:

Copy Number	**2xGal4**		**8x LexA**	
Spacer	**Yes**	**No**	**Yes**	**No**
DEX				
Success Ratio	0.75 (3/4)	1 (4/4)	-	-
Mean Rank	19.7	5.8	NA	NA
OHT				
Success Ratio	0.25 (1/4)	1 (4/4)	-	-
Mean Rank	32	22.3	NA	NA

2) *P*-values:

	Fisher's Exact Test	**Wilcoxon rank sum test**
DEX_2xGal4 Yes vs No	1	0.057*
OHT_2xGal4 Yes vs No	0.14	0.40

j) Overall comparisons for effects of spacer (only 35S-based gene circuits included)

		Without Spacer	**With Spacer**	***P*-value**
DEX	**Success Ratio**	0.71 (20/28)	0.75 (3/4)	1
	Mean Rank	15.2	19.7	0.39
OHT	**Success Ratio**	0.43 (12/28)	0.25 (1/4)	0.63
	Mean Rank	27.2	32	0.15
TOTAL	**Success Ratio**	0.57 (32/56)	0.5 (4/8)	0.72
	Mean Rank	19.7	22.8	0.50

k) Effects of binding site copy number (only 35S-based gene circuits without spacers included)

1) *P*-values for direct comparisons:

	Fisher's Exact Test	**Wilcoxon rank sum test**
DEX_LexA_35S 2x vs 8x	1	1
OHT_LexA_35S 2x vs 8x	1	NA

2) Bulk measurements for overall comparisons:

	Success Ratio	**Success Ratio**
2xGal4	0.75 (12/16)	15.8
5xGal4	0.50 (4/8)	20.0
2xLexA	0.63 (10/16)	19.5
4xLexA	0.38 (3/8)	29.3
8xLexA	0.38 (3/8)	25.7

3) *P*-values for overall comparisons:

	Fisher's Exact Test	Wilcoxon rank sum test
2xGal4 vs 5xGal4	0.36	0.86
2xLexA vs 4xLexA	0.39	0.29
2xLexA vs 8xLexA	0.39	0.57
4xLexA vs 8xLexA	1	0.70

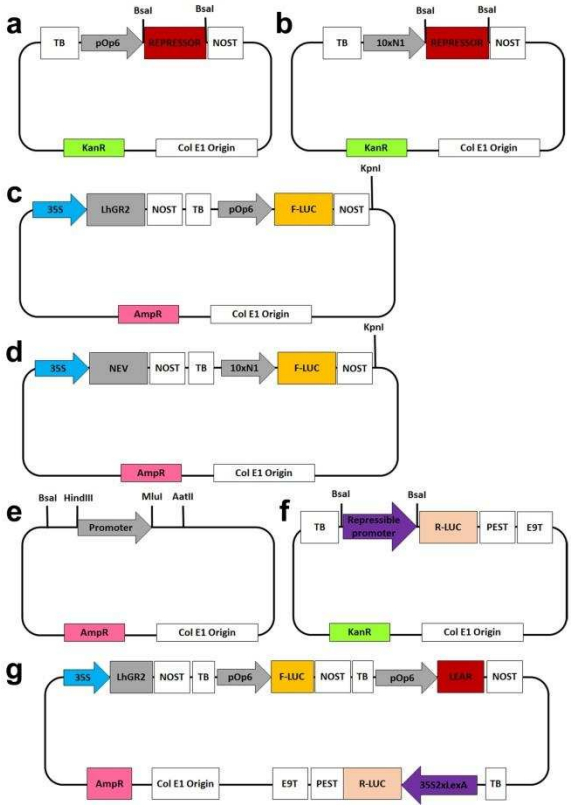

Figure B.1 Plasmids used to test repressors, repressible promoters, and promoter-repressor combinations in transient protoplast assays. (a) Repressor module used to assemble synthetic repressors under control of dexamethasone (DEX). (b) Repressor module used to assemble synthetic repressors under control of 4-hydroxytamoxifen (OHT). *Bsa*I restriction enzyme sites were included to exchange repressors. (c) Test plasmid used to assemble all DEX-inducible promoter-repressor combinations. (d) Test plasmid used to assemble all OHT-inducible promoter-repressor combinations. Repressors and repressible promoters were cloned into the test plasmids using the *Kpn*I restriction site. Both test plasmids contain Firefly luciferase (F-luc) expressed under control of one of two inducible promoters, pOp6 and 10xN1. F-luc is used as a proxy for the amount of repressors produced in the system. (e) Sub-cloning plasmid used to generate synthetic repressible promoters containing repressor binding sites upstream (*Bsa*I and *Hind*III) or downstream (*Mlu*I and *Aat*II) of the promoter. (f) Promoter module used to assemble repressible promoters controlling expression of the reporter gene, Renilla luciferase (R-luc). *Bsa*I restriction enzyme sites were included to exchange promoters. A PEST protein degradation sequence was added to R-luc to increase protein turnover and facilitate quantitative measurements of promoter repression. (g) Example of a complete DEX-inducible test plasmid used for protoplast assay. LEAR is a synthetic repressor composed of LexA DNA-binding domain and EAR repressor motif. 35S2xLexA is synthetic repressible promoter composed of constitutive CaMV 35S promoter and two copies of LexA binding elements placed downstream of the promoter scaffold. *LhGR2*, DEX-activated transcription factor; *NEV*, OHT-activated transcription factor; *NOST*, nopaline synthase terminator; *E9T*, pea rbcS-E9 terminator; *TB*, transcription block; *35S*, Cauliflower Mosaic Virus 35S promoter; *AmpR*, ampicillin resistance gene for

bacterial selection; *KanR*, kanamycin resistance gene for bacterial selection; *ColE1*, bacterial origin of replication.

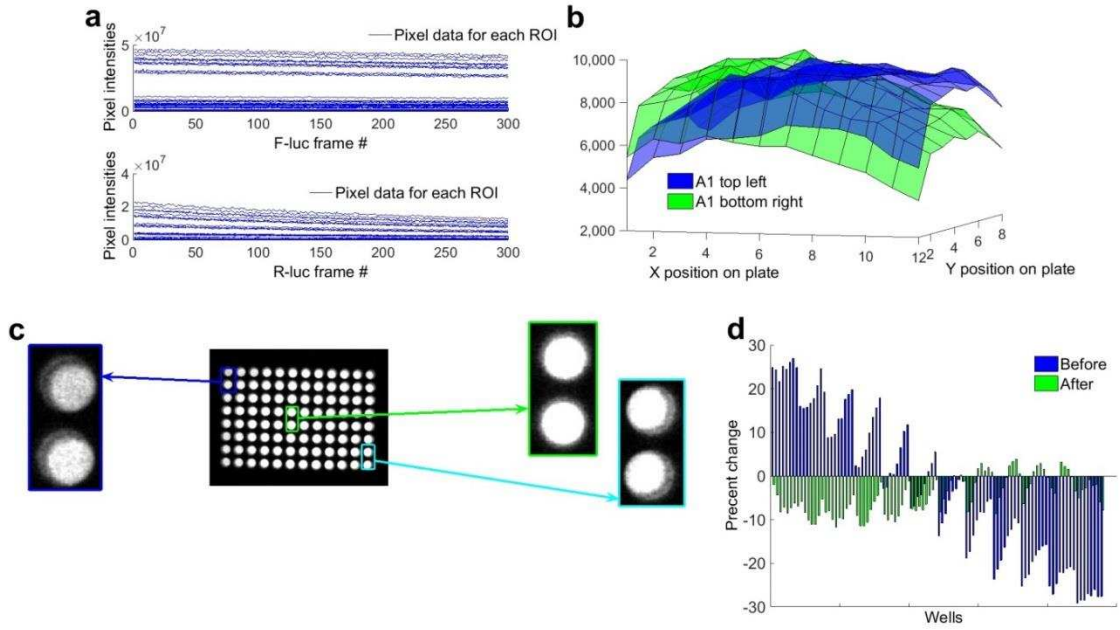

Figure B.2 Camera correction. (**a**) The camera collects one image (*i.e.*, frame) every 1/30 seconds. Each image represents the sum of pixel intensity within each well for every frame. Upper graph shows the F-luc signal is stable over time; lower graph shows the R-luc signal decays over the same time. (**b**) Representative graph showing the distribution of luciferase pixel intensity values RLU/(area x sec) for each well for both a plate imaged with well position A1 in the top left hand corner of the camera (*blue*), and the same plate with A1 in the bottom left hand corner of the camera (*green*). Data show that amount of luminescence recorded is influenced by the well position and changes on plate rotation. (**c**) Representative images of the luminescence of individual wells for one 96-well plate experiment. Wells at the edges of the plate (*blue and cyan outlines*) show "new-moon-shaped" occluded areas, whereas wells at the center of the plate (*green outline*) do not have these same occluded areas. (**d**) Percent change in the luminescence of the wells after rotation of the plate is shown for the original data (*blue*) and after imaging correction (*green*). Image correction removes almost all of the positional bias in the data.

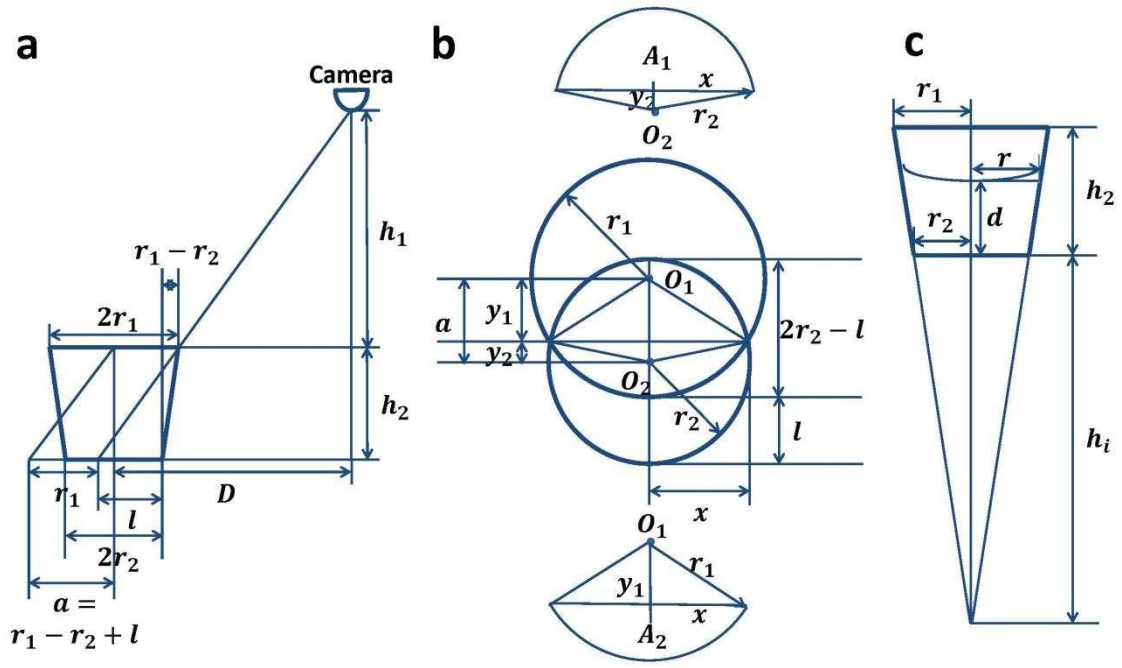

Figure B.3 Schematic geometric diagram of imaging correction method. (**a**) Side view of the optical system and the well in microplate of interest. Part of the well is blocked from the sight of the camera by the nontransparent wall. (**b**) Top view of the well of interest with the upper rim shifted to the bottom along the sight direction shown in **a**. The overlapping area of the two circles O_1 and O_2 is the visible part of the well bottom. (**c**) Side view of well in microplate as part of an imaginary cone. *r*, radius; *h*, height; *d*, diameter; *a*, area.

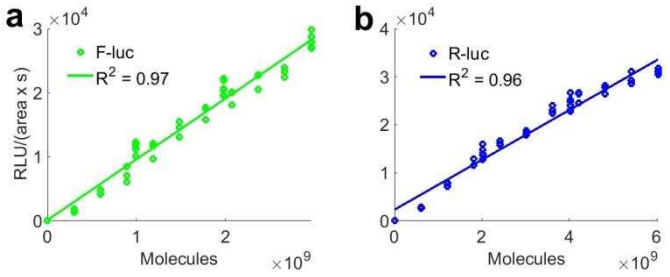

Figure B.4 Standard curves: luminescence to approximate number of molecules. Standard curve of luminescence produced in RLU/(area x sec) as a function of total number of molecules for F-luc (**a**) and R-luc (**b**). Lines represent best fits.

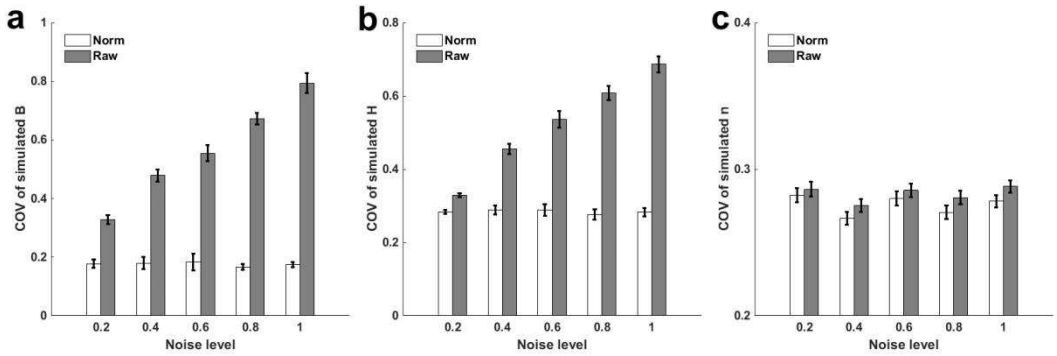

Figure B.5 Testing our normalization method with simulated and experimental repressor-repressible promoter data. (**a**) Coefficient of variation (COV) of the estimated parameter *B* with increasing noise levels in the distribution of the random multiplicative factor, . Non-normalized (*Raw*) data shows increasing COV, but the normalized data (*Norm*) is able to adjust for the increase in noise in , and shows no significant change in the COV. (**b**) COV of the estimated parameter *H* also increases with increasing noise for the raw data fits, but stays approximately constant for the normalized data. (**c**) COV of the estimates for *n* do not show a difference between the normalized and raw data.

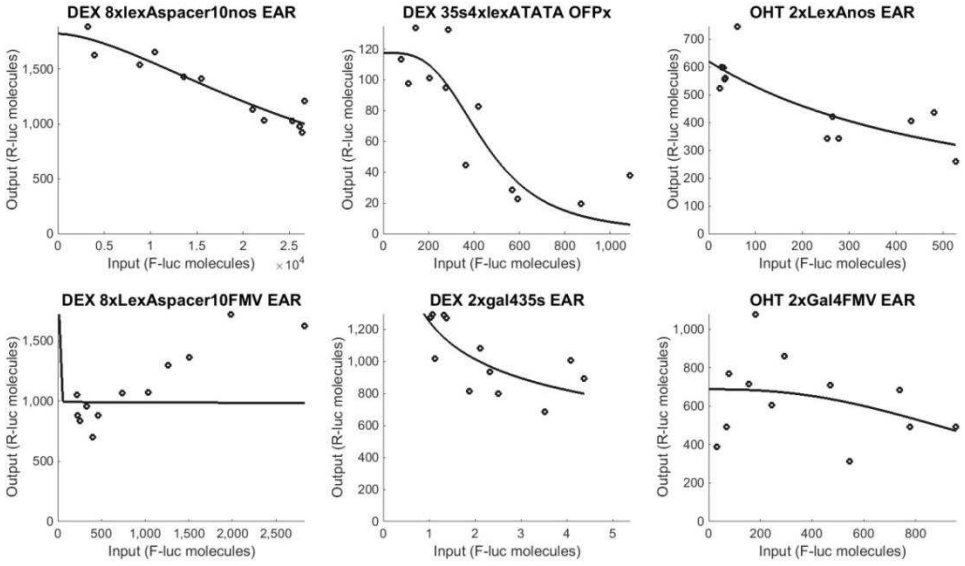

Figure B.6 Representative curve fits to non-normalized Arabidopsis data. Raw F-luc (*Input*) and R-luc (*Output*) luminescence values for six different promoter-repressor combinations, as indicated above the graphs. Solid lines represent fits to Hill function forms using the nonlinear least squares fitting package in MATLAB. Open circles represent experimental data.

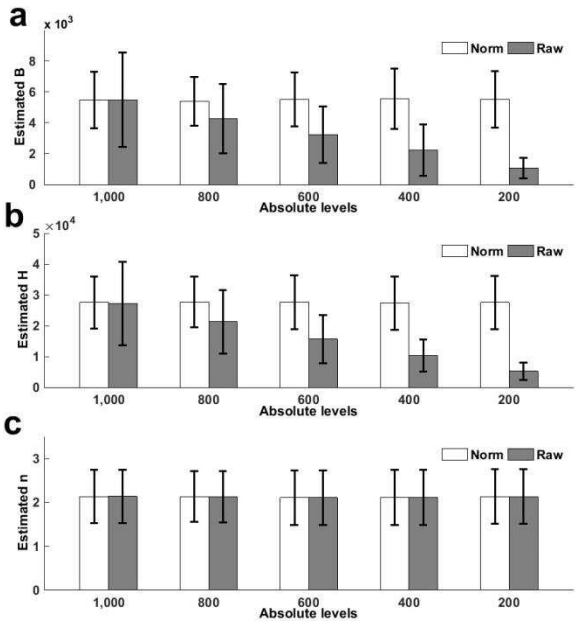

Figure B.7 Testing the normalization factor λ_i^* with simulated data. (**a**) Mean levels of estimated parameter *B* with increasing absolute levels in both mean and standard deviation of random number, N_{ij}. Raw data (*Raw*) show decreases in mean values, but the normalized data (*Norm*) show insensitivity to changes in absolute levels. (**b**) Mean levels of estimated parameter *H* also show decreases in raw data and remain constant with increasing absolute levels. (**c**) Mean levels of estimated parameter values of *n* do not show a difference between the normalized and raw data and across different absolute values.

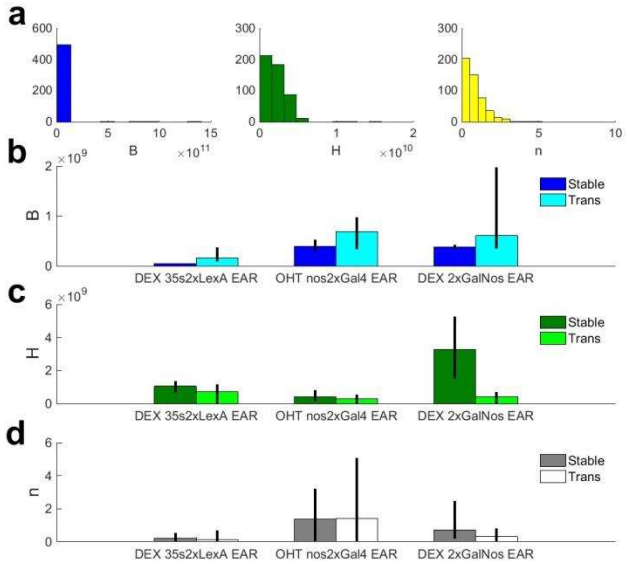

Figure B.8 Bootstrap results. (**a**) Distribution of parameter values (*B, H, n*) obtained from bootstrapping fits. (**b**) Comparison of bootstrapped estimates of the parameter *B* for protoplasts from the stably transformed plants (*Stable*) and protoplasts from transient expression (*Trans*) for three promoter-repressor pairs. (**c**) Comparison of bootstrapped

estimates of half-maximal expression, *H*. (**d**) Bootstrapped estimates of the Hill coefficient *n* are shown for the same three pairs.

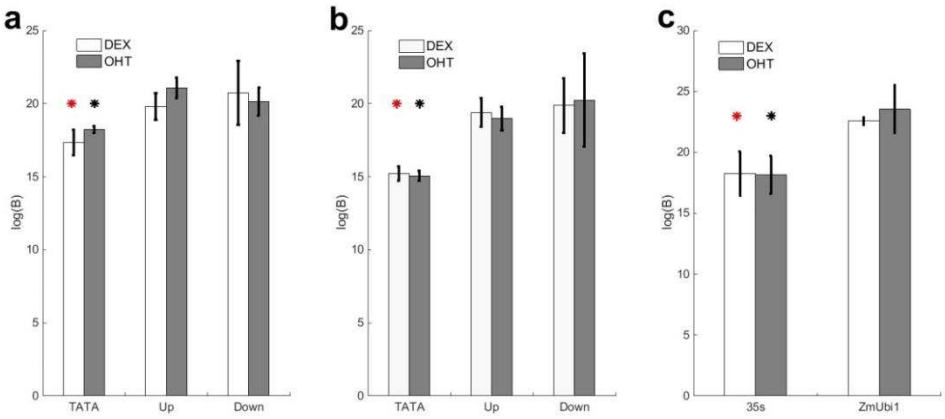

Figure B.9 ANOVA and HSD Tukey analysis. (**a**) Arabidopsis log(*B*) values for all good performing gene circuits categorized first by inducer type (*DEX* or *OHT*), and then by binding site position. Red star indicates a significant difference of DEX-inducible promoters when binding sites are near the TATA-box compared with binding sites either upstream or downstream of the constitutive element ($P < 0.01$). Black star indicates the same for OHT promoters ($P = 0.04$ for upstream, and $P = 0.01$ for downstream). (**b**) Sorghum log(*B*) values for all good performing gene circuits categorized first by inducer and then by binding site position. Red and black stars have the same meaning as in **a**. (*P*-values for significance are $P < 0.01$ for both upstream and downstream for DEX, and $P = 0.02$ for upstream and $P = 0.06$ for downstream, respectively, for OHT.) (**c**) Sorghum log(*B*) values for all good performing gene circuits categorized first by inducer and then by constitutive element (*CaMV 35S* or *ZmUbi1*). Red star indicates a significant difference in DEX promoters with these two elements ($P < 0.01$). Black star indicates the same for OHT promoters ($P < 0.01$). *TATA*, promoters with binding sites just upstream of the TATA-box; *Up*, binding sites placed upstream of the constitutive scaffold; *Down*, binding sites placed downstream of the constitutive scaffold; *35S* or *ZmUbi1*, refers to the constitutive promoter scaffold used. l*og(B)*, logarithm of the bulk promoter strength, *B*. Bars are standard deviations.

REFERENCES

[1] B. Peirce, "Criterion for the rejection of doubtful observations," vol. 2, pp. 161-163, 1852.

[2] S. M. Ross, "Peirce's criterion for the elimination of suspect experimental data," *J Eng Technol,* vol. 20, pp. 38-41, 2003.

[3] J. Fox, Applied Regression Analysis and Generalized Linear Models, Thousand Oaks, CA,: SAGE Publications, 2008.

[4] Y. Tadesse, L. Sagi, R. Swennen and M. Jacobs, "Optimisation of transformation conditions and production of transgenic sorghum (Sorghum bicolor) via microparticle bombardment," *Plant Cell Tiss Org,* vol. 75, pp. 1-18, 2003.

[5] J. Odell, F. Nagy and N. Chua, "Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter," *Nature,* vol. 313, pp. 810-812, 1985.

[6] M. Sanger, S. Daubert and R. M. Goodman, "Characteristics of a strong promoter from figwort mosaic virus: comparison with the analogous 35S promoter from cauliflower mosaic virus and the regulated mannopine synthase promoter," *Plant Mol Biol ,* vol. 14, pp. 433-443, 1990.

[7] C. H. Shaw, G. H. Carter and M. D. Watson, "A functional map of the nopaline synthase promoter," *Nucleic Acids Res,* vol. 12, pp. 7831-7846, 1984.

[8] J. M. Jeoung, S. Krishnaveni, S. Muthukrishnan, H. N. Trick and G. H. Liang, "Optimization of sorghum transformation parameters using genes for green fluorescent protein and beta-glucuronidase as visual markers," *Hereditas,* vol. 137, pp. 20-28, 2002.

[9] K. A. Schaumberg, M. S. Antunes, T. K. Kassaw, W. Xu, C. S. Zalewski, J. I. Medford and A. Prasad, "Quantitative characterization of genetic parts and circuits for plant synthetic biology," *Nature Methods,* vol. 13, pp. 94-100, 2016.