RESEARCH ARTICLE

Identifying the combinatorial control of signal-dependent transcription factors

Ning Wang[1,2,3], Diane Lefaudeux[1,2], Anup Mazumder[2¤], Jingyi Jessica Li[1,4], Alexander Hoffmann[1,2]*

1 Institute for Quantitative and Computational Biosciences (QCBio), University of California, Los Angeles, California, United States of America, 2 Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, California, United States of America, 3 Interdepartmental Program in Bioinformatics, University of California, Los Angeles, California, United States of America, 4 Department of Statistics, University of California, Los Angeles, California, United States of America

¤ Current address: National Institute of Biomedical Genomics, Kalyani, India
* ahoffmann@ucla.edu

Abstract

The effectiveness of immune responses depends on the precision of stimulus-responsive gene expression programs. Cells specify which genes to express by activating stimulus-specific combinations of stimulus-induced transcription factors (TFs). Their activities are decoded by a gene regulatory strategy (GRS) associated with each response gene. Here, we examined whether the GRSs of target genes may be inferred from stimulus-response (input-output) datasets, which remains an unresolved model-identifiability challenge. We developed a mechanistic modeling framework and computational workflow to determine the identifiability of all possible combinations of synergistic (AND) or non-synergistic (OR) GRSs involving three transcription factors. Considering different sets of perturbations for stimulus-response studies, we found that two thirds of GRSs are easily distinguishable but that substantially more quantitative data is required to distinguish the remaining third. To enhance the accuracy of the inference with timecourse experimental data, we developed an advanced error model that avoids error overestimates by distinguishing between value and temporal error. Incorporating this error model into a Bayesian framework, we show that GRS models can be identified for individual genes by considering multiple datasets. Our analysis rationalizes the allocation of experimental resources by identifying most informative TF stimulation conditions. Applying this computational workflow to experimental data of immune response genes in macrophages, we found that a much greater fraction of genes are combinatorially controlled than previously reported by considering compensation among transcription factors. Specifically, we revealed that a group of known NFκB target genes may also be regulated by IRF3, which is supported by chromatin immuno-precipitation analysis. Our study provides a computational workflow for designing and interpreting stimulus-response gene expression studies to identify underlying gene regulatory strategies and further a mechanistic understanding.

Author summary

Cells need to sense environmental cues and respond appropriately. One important notion is that different stimuli activate different combinations of transcription factors and that responsive genes are regulated by distinct subsets of these. However, identifying the regulatory strategies by which genes interpret transcription factor activities remains a largely unsolved challenge. In this work we address the question: to what extent are combinatorial transcription factor regulatory strategies identifiable from stimulus-response (input-output) datasets? We present a computational framework to determine the identifiability of gene regulatory strategies, and examine how reliable and quantitative model inference is a function of the quality and quantity of available data. We present an error model that more precisely quantifies uncertainty for perturbation-timecourse data sets by also considering error in the time domain, and achieves an improved performance in identifying and quantifying gene regulatory strategies. With these tools, we generate guidelines for experimental design that optimize limited resources for generating data for model inference. Finally, we apply the workflow to immune response datasets and uncover evidence that many more genes are subject to combinatorial control than previously thought; we offer physical transcription factor binding data to support this finding for one particular group of genes. This demonstrates that the computational workflow may guide studies of the regulatory strategies that govern stimulus-responsive gene expression programs.

Introduction

A primary goal of biology is to understand biological phenomena in terms of the underlying factors, whether these are cells, molecules or genes. These factors form dynamic regulatory networks whose emergent properties are responsible for biological phenomena. Hence, the systems biology approach employs mathematical models that represent or abstract these networks to interpret experimental data.

For studies of how genes are expressed, the advent of experimental assays that are capable of producing genome-wide measurements of mRNA abundance, chromatin-bound factors and modifications has been revolutionary. A variety of computational approaches have been developed to construct correlations networks based on these large datasets [1–6]; these infer correlative connections between regulators that may reflect a direct or indirect causal relationship, a common cause, or mere coincidence. Because correlative approaches often leverage the statistical power of multiple datapoints from similarly expressed genes, they are not well suited in addressing the regulatory precision of individual genes [7]. That is an important limitation, as many pathological conditions can in fact be traced to a single gene culprit, or a handful [8].

In order to leverage the wealth of genome-wide gene expression datasets for developing a mechanistic understanding of gene expression, prior studies have employed mathematical models that represent the functional interactions between the gene and the key transcription factor (TF) [9–13]. In pioneering work, kinetic models with logic gates were applied to time series data to infer genome-wide transcriptional regulatory networks [6], but questions about the goodness of fit, uniqueness of the solution, and model identifiability remained to be addressed. Simple mathematical models do not describe the detailed gene regulatory mechanisms in the nucleus, that may contain hundreds of factors along large stretches of regulatory DNA, but constitute functional abstractions that may be termed the gene regulatory strategies (GRSs). Thus, GRS models describe functional relationships in a mechanistic modeling framework.

Mathematical models of such GRSs involving a single TF have been successfully fit to datasets from individual mammalian genes, when the TF activity may be induced by a stimulus to provide a perturbation with a defined starting timepoint. In this case both TF activity and target gene expression were measured in a timecourse to provide the data for GRS model fitting. Indeed, for immune response genes this approach allowed GRS model parameters to be fit [10,13], or the GRS model topology to be selected from two alternatives [9]. Such experimentally validated GRS models may then guide the development of finer-grained GRN models that more faithfully describe the physical interactions that give rise to the regulatory strategies.

However, many mammalian genes are not regulated by a single TF, but multiple TFs. TFs collaborate either by compensating for or enhancing each other's activities. Characterizing what is termed the "logic" of collaborative TFs regulating genes is critical to understanding how the genome is expressed. To quantitatively capture combinatorial gene regulation by multiple TFs, thermodynamic formulations of Boolean AND- and OR-gate-like relationships may be employed to describe molecular interactions between DNA, TFs, and the polymerase-containing transcription machinery that regulates transcription initiation [14–17]. However, identifying the right model for a specific experimental dataset means that not only the most appropriate model topology (e.g. logic gates) needs to be selected to fit the data, but this should be done by considering a range of parameter values (e.g. TF regulation strengths) that should be optimally fit to the data.

A recent study aimed to identify GRSs with combinatorial TF logics using mammalian immune response datasets [18]. To render the model identifiability challenge more tractable, the complexity of matching combinatorial TF models had to be dramatically reduced by two simplifying measures: first, the model parameters were fixed, reducing the task to identify the best model to match the data; and second, models were matched to the mean expression of co-clustered genes, rather than individual genes, thereby avoiding the confounding effects of technical inaccuracy. This second simplification, however, diminishes the rich diversity of regulatory strategies that occur even within co-clustered genes [19,20]. In sum, we conclude that 1) model selection from a library of GRS topologies and quantitative fitting from a range of parameter values must be done at single gene resolution, and 2) this task poses an as yet unresolved model identifiability challenge.

In this study, we have addressed the GRS model identifiability challenge in three steps. First, we systematically delineated GRS model identifiability, thereby identifying GRS models that are easily distinguished from each other, and others that require substantially more data. Second, we developed a Bayesian computational workflow–including a new error model–to use experimental input-output datasets for GRS model selection and parameter fitting. Third, we applied the computational workflow to newly assembled experimental innate immune datasets. Our results demonstrate the utility of the newly developed computational workflow by applying it to an innate immune gene dataset to reveal potentially combinatorially controlled genes in line with reports in other biological systems [21,22].

Results

Assessing the distinguishability of gene expression patterns produced by distinct GRSs involving three TFs

Upon cellular stimulation, the activities of stimulus-induced TFs are activated and–in a combinatorial manner–enhance transcription initiation of specific genes. To explore the identifiability of distinct GRS models involving TF combinations, we first examined via a systematic model-based analysis how distinguishable the associated gene expression patterns really are when different TF combinations are activated.

We employed an established model formalism in which messenger RNA abundance of an expressed gene is controlled by synthesis and decay [23], in our case we model nascent mRNA such that the decay term corresponds biologically to its processing and release from the chromatin. An ordinary differential equation models RNA abundance dynamics (Fig 1A), where the processing/chromatin release rate (k_{proc}) is first order and RNA synthesis is zeroth order but is modeled by multiplying the RNA synthesis rate constant (k_{syn}) with the fractional promoter activity ($f(t)$). Fractional promoter activity was modeled by an established thermodynamic formulation with a Hill function that captures the TF regulation strength [17,18], as a function of a disassociation constant (K_d). Strong ($K_d = 0.1$) and Weak ($K_d = 10$) regulation strengths were defined as the readily saturated and always linear dose-response relationships, respectively, with Medium regulation strength being in-between (Fig 1B). Here we assume no cooperativity in the activation by a single TF ($n_H = 1$), in line with a previous report [24] (though future work may consider non-linear dose response curves that may have been reported in other systems. Similar thermodynamic formulations [15,17,18] are used to model AND and OR logic gates which consider synergistic or non-synergistic functions between two TFs (Figs 1C and S1, see Methods). We enumerated all possible logics composed by single, dual and triple TFs with AND and OR logic gates (Fig 1D). Single and dual logics can be represented by the same 8 triple logic gates when allowing for null regulation strengths for one or two TFs (see Methods). Thus, only 8 triple logic gates are needed to study all logics formed by the combination of a one, two or three TFs.

To conduct a systematic analysis of GRS identifiability, we enumerated the GRS models (S2 Fig) that cover not only all combinations of AND and OR logics but also 3 TF regulation strengths ($K_d = 0.1, 1, 10$, referred to as Strong, Medium, Weak) and a fixed pre-mRNA chromatin release and processing rate. Of the possible 216 GRSs, 69 GRSs were removed because they cannot be efficiently activated (due to weak TF regulation strengths), and 54 GRSs were removed because they were logically equivalent with other, resulting in a list of 93 unique, potentially identifiable GRSs. To probe these GRSs, we first considered a set of 7 perturbations involving all combinations of high and low activities induced for each of the 3 TF ($2^3 - 1 = 7$, omitting the case of all three TFs being low) (Fig 1E). We simulated gene expression patterns across timepoints for all 93 GRSs (Fig 1F, Methods), and reordered GRSs based on hierarchical clustering (single linkage method) of their gene expression timecourses, using the squared Euclidean distance across all time points and perturbation conditions as the distance metric between them (Fig 1G). This revealed that some distinct logics produce similar gene expression patterns as they were co-clustered. For example, within the TF1-response cluster we find both the TF1 or (TF2 and TF3) GRS and the triple OR gate GRSs, but what all have in common is that TF1 has strong (S) or medium (M) regulation strength. We next examined the tree height for each cluster. It is apparent that the GRSs within the triple OR gate cluster are easiest to distinguish, and the GRSs within the triple AND gate cluster are hardest to distinguish (indistinguishable with the present 7 perturbations). All the results have been further confirmed with different hierarchical clustering methods (average and complete linkages, S3A Fig).

As 7 perturbations do not enable distinction of all GRSs, we next explored more diverse perturbation conditions such as transient, delayed, gradient, or intermediate amplitude TF activations (Fig 2A). Examining the triple AND gate GRSs that could not be distinguished with the set of 7 perturbations, we found dramatic differences: amplitude and gradient perturbations readily distinguished these 7 GRSs, but transient and delayed TF temporal dynamics did not (Fig 2B). To summarize, we found that the regulation strength of a TF may be identifiable when perturbation data with TF activities of differential amplitudes is available.
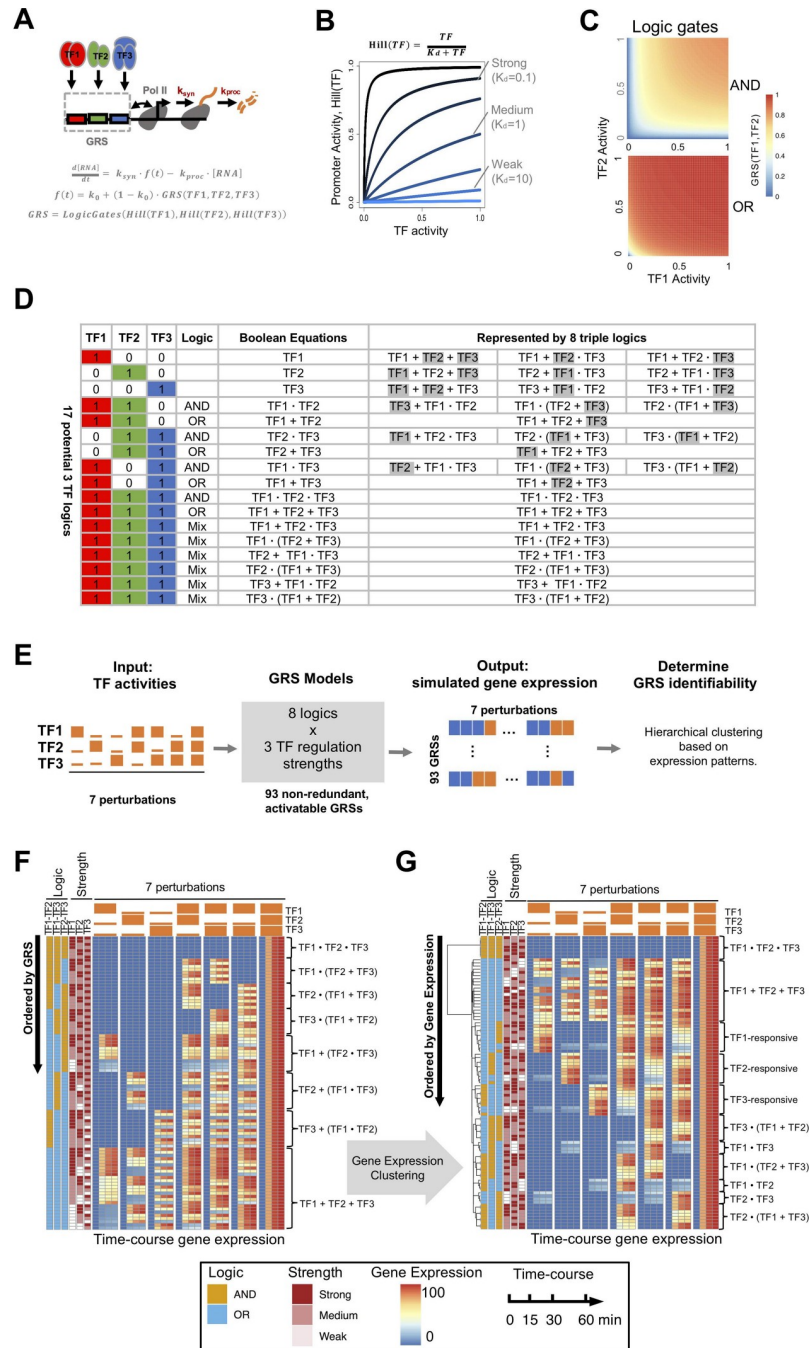
Fig 1. Studying GRSs with stimulus-induced TFs. (A) Schematic of the gene regulation model: stimulus-induced TFs bind to target DNA to induce Pol II-mediated mRNA synthesis, which is followed by its processing and release from the chromatin. Nascent mRNA abundance can be described by a single ordinary differential equation (ODE). Promoter activity is described by thermodynamic models involving Hill functions. (B) Line graphs of promoter activity as a function of a single TF activity depends on the regulation strength (here indicated by $1/K_d$, in black-to-blue scale). We define the regulation strengths as Strong ($K_d = 0.1$), Medium ($K_d = 1.0$), or Weak ($K_d = 10$), as marked in the plot. (C) Heatmaps of promoter activity as a function of logic gates (AND, OR) with varying TF1, TF2 activities as input both with Strong regulation ($K_{d1} = K_{d2} = 0.1$). (D) Enumeration of all 17 possible AND and OR logic combinations by three TFs. These may be represented by 8 triple logics when single and dual TF logics contain null regulation strengths ($K_d \gg 1$) for one or two TFs (marked with grey shading). AND and OR gates are denoted with "·" and "+" in Boolean algebra. (E) Schematic of the analysis workflow using a set of 7 perturbations with high and low TF activities, to probe 93 non-redundant, activatable GRSs (see S2 Fig), by simulating their gene expression patterns and examining those patterns by hierarchical clustering. (F) Heatmap of gene expression at 0, 15, 30, 60 min from all 93

GRSs in response to a set of 7 perturbation conditions involving 3 TFs. Here the heatmap is ordered by GRSs, their regulatory logic and TF regulation strengths (left columns). (G) The data of panel (F) with GRSs are ordered by hierarchical clustering (single linkage approach) of gene expression using the squared Euclidean distance depicted by the tree. This analysis shows that distinct combinatorial logics may give similar gene expression patterns, and that tripe AND gates with distinct regulation strengths cannot be distinguished with 7 perturbations.

When fitting the GRS models to experimental data, we may expect that lower-quality data may affect the identifiability of less well-separated GRSs more than of widely separated GRSs. To explore how the resolution of each GRS and data quality affects GRS identifiability, we cut
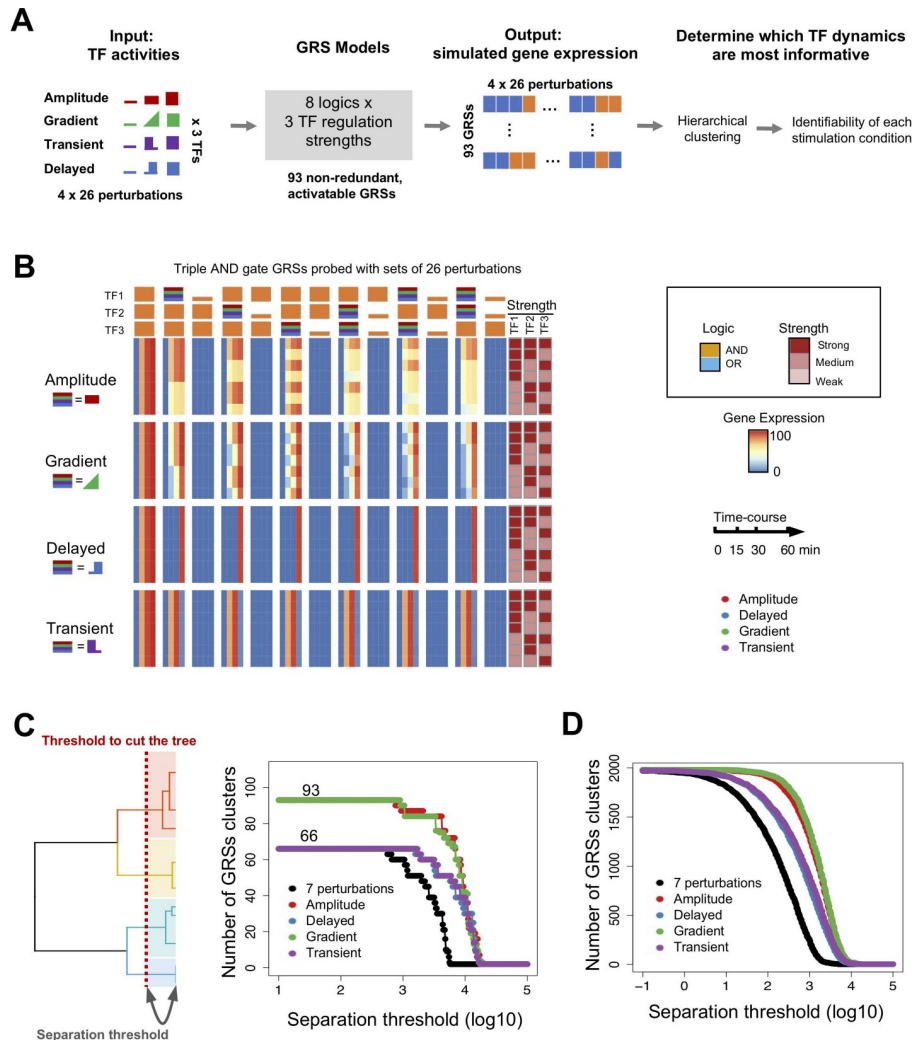
Fig 2. The distinguishability of GRSs depends on the available set of TF perturbations. (A) Schematic of the analysis workflow using 4 sets of 26 perturbations each with amplitude, gradient, transient, and delayed TF dynamics to produce combinatorial TF activities to probe 93 non-redundant, activatable GRSs, by simulating their gene expression patterns and examining those patterns by hierarchical clustering. (B) Heatmap of the triple AND gate GRSs (top 7 rows in (C), probed with one of the 4 sets of 26 perturbations (shown in (B)), containing amplitude modulated TF activities (amplitude, gradient) and temporally modulated TF activities (delayed, transient). The results show that amplitude modulated TF activities best suited to distinguish GRSs are (C) Number of distinguishable GRS clusters as a function of the separation threshold with indicated sets of perturbation combinations. (D) The same plot as panel C but from 1981 inducible GRSs generated by random sampling of 1000 parameter sets for each logic. All 1981 GRS are identified at the lowest separation threshold (-1) for amplitude and gradient perturbations, whereas 1971, 1975, 1976 GRS clusters are identified for high/low, transient, and delayed perturbations, respectively.

the tree at various height thresholds, referred to as separation thresholds, and we determined the number of distinguishable GRS clusters as a function of this threshold with multiple perturbation conditions (Fig 2C). This analysis further confirmed our previous observations: the transient and delayed perturbation conditions only marginally increased GRS identifiability, whereas the amplitude and gradient modulation conditions increased GRS identifiability substantially to distinguishing all 93 GRSs. Moreover, these results were confirmed with alternative hierarchical clustering methods (S3B Fig). The hierarchical tree analysis therefore provides guidance on which GRSs are more likely confused, and indicate which GRSs that are functionally similar to the identified GRS of interest.

The initial study of 93 representative GRSs allowed us to efficiently and systematically reveal the key properties of GRSs with multiple perturbation conditions. We next generalized our results by random sampling of 1000 parameter sets for each logic gate (see Methods). We stringently represented single, dual logics with null regulation strengths of triple logic gates, and sampled all parameters (3 K_d, k_{syn}, k_{proc}, k_0) from the full parameter space. We found that clusters formed (S3C Fig) that were similar to those observed in the analysis of representative GRSs. We next compared multiple perturbation conditions, and confirmed the result: only amplitude and gradient modulation conditions enabled distinguishing all GRSs (Fig 2D).

An error model to quantify data uncertainty in value and time

In practical applications, not only the quantity of data, but also its quality (in terms of associated uncertainty/variability) is expected to affect GRS identifiability. Applying the model to real experimental data requires an error model that reliably quantifies its data uncertainty. Here, we have considered the sources of data uncertainty, resulting from both biological variability and technical uncertainty (Fig 3A). Based on first principles, we simulated biological variability by varying TF abundances and all the parameters in mechanistic models that describe gene responses (3 K_d, k_{syn}, k_{proc}, k_0), as well as technical uncertainty by varying sample preparation timepoints and adding noise to the assay measurements (see Methods). Using the set of 26 amplitude perturbations, we simulated two replicate expression patterns for each GRS and captured four time points from each simulation (Fig 3B).

In timecourse data, higher variance is often observed when the timecourse involves sharp increases or decreases. This higher degree of variance may be caused by uncertainty introduced from the time axis (Fig 3C), for example by samples not being collected at the precise timepoints, or a sample of cells responding more slowly due to culture inconsistencies. In principle, the variance of measurements is composed of both a value error and a temporal error, the latter contributing more to the observed variance when there are rapid changes in the timecourse curve than when changes are slow. To account for the contribution of temporal uncertainty, we developed a time-value error model, in which the temporal error is a function of the derivative of the timecourse. As a result, the time-value error model can avoid the under- or over-estimation of error associated with a conventional error model (Fig 3C). A small number of samples do not allow for stable estimation of variance. Here, we combined global trends from all pooled points with an individual point component to achieve a stable point-specific variance estimation. This approach has been widely used when analyzing RNA-seq data (see Methods). For the value uncertainty, we used polynomial regression to capture the value variance-mean relationship. The temporal uncertainty is modeled by a linear Gaussian model (see Methods).

To test the utility of the time-value error model, we applied it to our simulated data and compared results to when the conventional model (which does not consider temporal uncertainty) is applied. By comparing estimated data uncertainty with empirical uncertainty
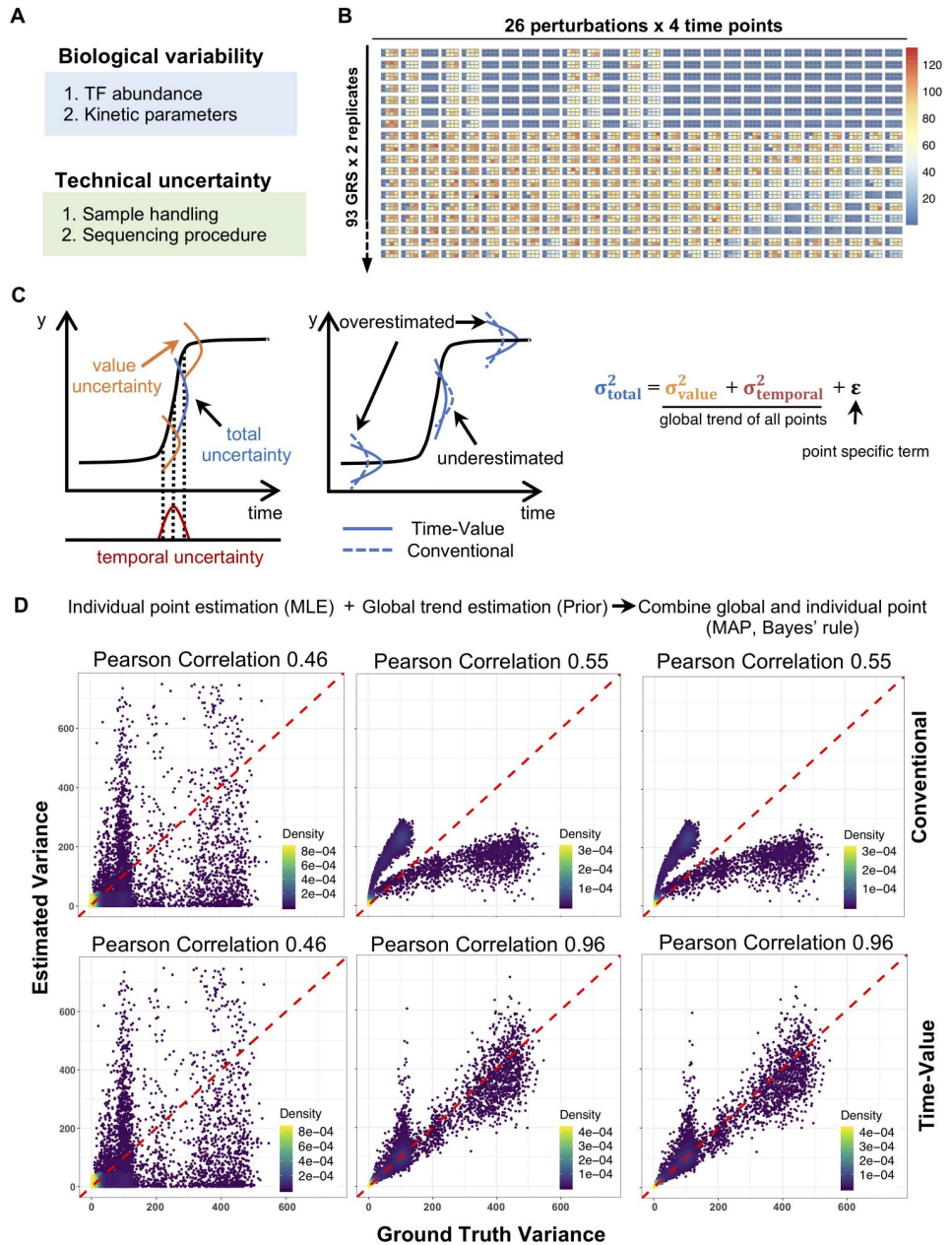
A

Biological variability

1. TF abundance
2. Kinetic parameters

Technical uncertainty

1. Sample handling
2. Sequencing procedure

B

26 perturbations x 4 time points

93 GRS x 2 replicates

C

$$\sigma^2_{total} = \underbrace{\sigma^2_{value} + \sigma^2_{temporal}}_{\text{global trend of all points}} + \underbrace{\varepsilon}_{\text{point specific term}}$$

D   Individual point estimation (MLE) + Global trend estimation (Prior) ➔ Combine global and individual point (MAP, Bayes' rule)

Pearson Correlation 0.46   Pearson Correlation 0.55   Pearson Correlation 0.55

Pearson Correlation 0.46   Pearson Correlation 0.96   Pearson Correlation 0.96

Fig 3. An error model for stimulus-response data that leverages timecourse information to avoid under and over-estimates of measurement uncertainty. (A) Uncertainty sources in data simulation. We considered both biological variability and technical uncertainty. (B) Simulated gene expression data with uncertainty (see Methods). Replicate datasets for 18 of the 93 GRSs are shown here. (C) Diagram of the time-value error model. Left panel shows how observed uncertainty can be decomposed into value and temporal uncertainties. Middle panel shows how the conventional model will under- and over-estimate data uncertainty. Right panel shows that error is decomposed into value uncertainty, temporal uncertainty, and point-specific uncertainty. (D) Uncertainty estimation with the temporal-value and conventional error model. For the left column, we estimated data uncertainty from each point using Maximum Likelihood Estimation (MLE) (98% points are shown). For the middle column, we estimated the global trend as prior. By applying Bayes' rule, we obtained posterior estimates, right column. Pearson correlations are calculated between estimated variance and ground truth variance. Top row corresponds to the conventional model, and bottom row to our error model.

https://doi.org/10.1371/journal.pcbi.1009095.g003

(calculated from 1000 samples), pooling points together can largely improve the estimation accuracy over the directly calculated raw variance, with a Pearson correlation from 0.46 to 0.96 for the time-value model, and from 0.46 to 0.55 for the conventional model (Figs 3D and S4). Interestingly, we clearly see two modes of estimation in the conventional model: one is an over-estimate and the other is an under-estimate, which is consistent with our expectation (Fig 3C and 3D). It also shows that two replicates alone are not able to improve estimation accuracy by combining point-specific information, and more replicates are required to improve the point-specific estimation.

A workflow to fit GRS models to data

Identifying GRSs from experimental data requires not only an error model to characterize data uncertainty, but also an optimization pipeline that enables fitting the model to multiple pairs of input-output datasets of different qualities (Fig 4A). Here, we used a Bayesian framework to derive a likelihood function that serves as the target function and allows fitting the model to datasets of different qualities, while searching the full space for all the parameters (3 K_d, k_{syn}, k_{proc}, k_0). We applied this optimization workflow to the simulated data, and obtained a fitted model which recapitulates the ground truth data (Fig 4B). Compared to using raw variance, the time-value error model correctly identified all the combinatorial logic gates irrespectively of the associated regulation strength (Fig 4C). Compared to the conventional error model and raw variance, the time-value error model better constrained the estimated regulation strengths within two-fold of the true values (Fig 4C). Further, we found that this improved performance of the time-value error model in GRS parameter estimation over the conventional model and raw variance is irrespective of the degree of uncertainty or noise level (S5A Fig). Interestingly, we found that when we consider both combinatorial logic gate and regulation strength, a total of 15 GRSs are mis-classified when using either the conventional model or the raw variance. These 15 GRSs are associated with clusters whose GRSs we previously found are hard to distinguish in our theoretical analysis (S5B Fig), like triple AND gate, and single TF regulated clusters. In addition, the confused GRSs are mis-classified mostly with their neighbors characterized by the tree structure (S5B Fig).

Guidelines for designing stimulation/perturbation studies

To identify GRSs from experimental data, not only the quality but also the quantity of available data and the type of perturbation conditions matter. Specifically, we showed that GRSs can produce very different gene expression patterns under different perturbation conditions (Figs 1F and 2B), and thereby leads to different identifiability under those conditions. Therefore, we wondered whether we could determine which sets of perturbation conditions are most informative to distinguish GRSs (Fig 5A). In practice, researchers, given a fixed budget to generate for example 52 datasets, need to decide whether to generate a large number of replicates to reduce experimental error prioritizing more perturbations that would provide additional perturbation-response information (Fig 5A).

Examining the identifiability (defined as the $\log_2$ likelihood ratio between the ground truth and the most similar alternative GRS) using simulated, imperfect data (due to biological variability and measurement uncertainty) from all possible perturbation combinations, we first compared GRS identifiability for different sets of single or double knockout or knockdown conditions (HLL, HHL, HMM, HHM), which correspond to setting TF perturbation or activity levels to fully activated (H) as in wild-type, low (L) for knockouts, or medium (M) for knockdowns. We found that the single knockout (HHL) is the most informative of these perturbations (Fig 5B), but combining the three single and three double knockout conditions
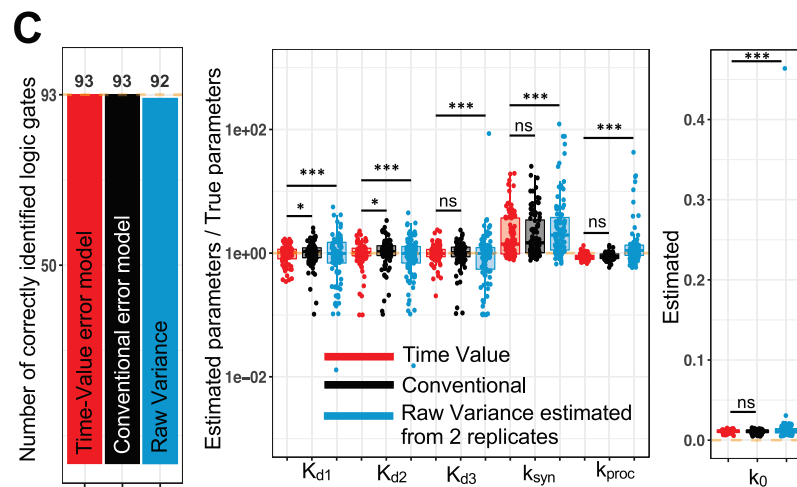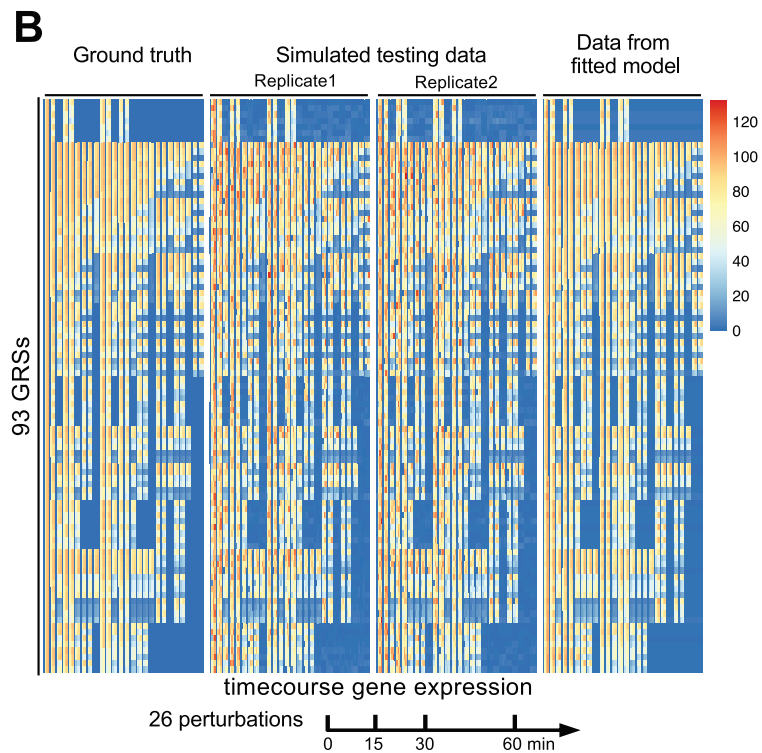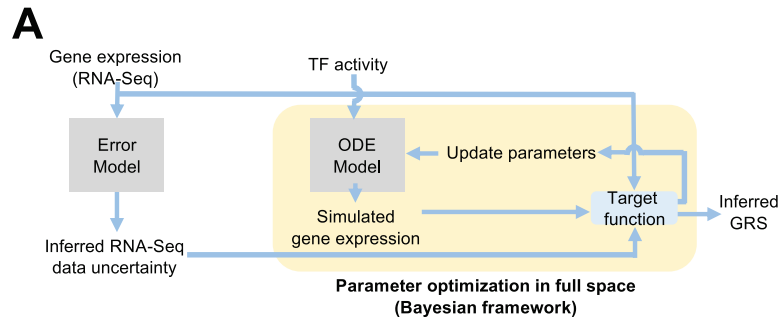
Fig 4. Bayesian framework to parameterize model with data uncertainty. (A) Diagram of the computational pipeline composed of data quality assessment and model parameterization. (B) Gene expression from simulated testing data and fitted model generated data. Rows correspond to the 93 GRSs in the same order as in Fig 2C. Columns corresponds to the 26 amplitude perturbation conditions at 0, 15, 30, 60 min. (C) Bar plot presenting the number of correctly identified logic gate and boxplot of estimated parameters for 93 GRSs fitted models. Percentage of estimated parameters within 2-fold of the true value are for Time-Value model, Conventional model, raw variance respectively: K_{d1} 95%, 91%, 70%; K_{d2} 91%, 84%, 70%; K_{d3} 91%, 90%, 71%. The mean absolute percent deviations are for Time-Value model, Conventional model, raw variance respectively: K_{d1} 24%, 28%, 58%; K_{d2} 26%, 36%, 48%; K_{d3} 23%, 28%, 138%.

https://doi.org/10.1371/journal.pcbi.1009095.g004

(HLL, HHL) provides a substantial boost to distinguishing the 93 GRSs, more so than when knockdowns (HHM, HMM) alone were used. However, combining knockdowns (HHM, HMM) can help distinguish the weakly identifiable GRSs by distinguishing the regulation strength features, with the comprehensive set of 26 perturbations (including mixed knockout and knockdown) providing a further boost in identifiability.

Next, we considered how to optimally allocate resources that would allow for 52 datasets. If only 4 perturbation conditions (of WT and three single knockout conditions) are employed, providing them in 13 replicates substantially increases identifiability (median GRS identifiability metric of 427 vs. 54). However, spreading resources to more perturbations (either 13 × quadruplets, or better 26 duplicates), will further improve identifiability (median GRS identifiability metric 498, 536 vs. 427) (Fig 5C). Further, we found generating more perturbations is still more informative even with the small number of datasets (S6 Fig). These studies suggest that simulations may be helpful in prioritizing limited resources for designing experimental studies to elucidate GRSs based on input-output data.

GRSs of immune response genes

We applied the newly developed computational workflow to identify GRSs for immune response genes in bone marrow derived macrophages (BMDMs). NFκB, IRF, and MAPK are three signaling pathways governing immune response gene expression. We obtained nascent RNA expression from chromatin associated RNA-seq (caRNA-seq) by processing recently deposited data [19]. We measured NFκB and IRF3 activities, and inferred MAPK-regulated transcription factor activity (given the absence of a direct TF assay) from the expression patterns of well-established immediate early response genes (Egr1, Fos, and Dusp4) that are neither NFκB, nor IRF targets [18,19] (S7A Fig). Here, we focused on the primary immune response genes (see Methods for selecting induced genes) within the first hour after stimulation, to avoid potential secondary effects from signaling cascades or feedbacks.

We first quantified data uncertainty of the time-course caRNA-seq data with the newly developed error model (Fig 3), and then estimated parameters from models with 8 logics with the described Bayesian framework (Fig 4). We mapped the 8 triple logics back to the original 17 logics (Fig 1D) by identifying dual or single TF logics *via* regulation strengths (Fig 6A). Given the available data, most genes were matched to multiple GRSs (Figs 6B and S7B). However, interestingly, our analysis suggested that only 26% of genes were matched to GRSs governed by a single TF, whereas 74% of genes could only be accounted for by logics involving 2 or 3 TFs (Fig 6C). This suggests that combinatorial gene regulation is much more common than previously reported [18,19].

Next, we focused on genes that are potentially synergistically regulated by NFκB and IRFs, i.e. genes to which any of the 5 logics may apply: NFκB AND IRFs, NFκB AND IRFs AND MAPK, NFκB AND (IRFs OR MAPK), IRFs AND (NFκB OR MAPK), MAPK OR NFκB AND IRFs. We identified not only all 5 previously reported NFκB-regulated genes that are IRF-dependent (Fig 6D) [19], but also a bigger set of additional genes. Many of these showed
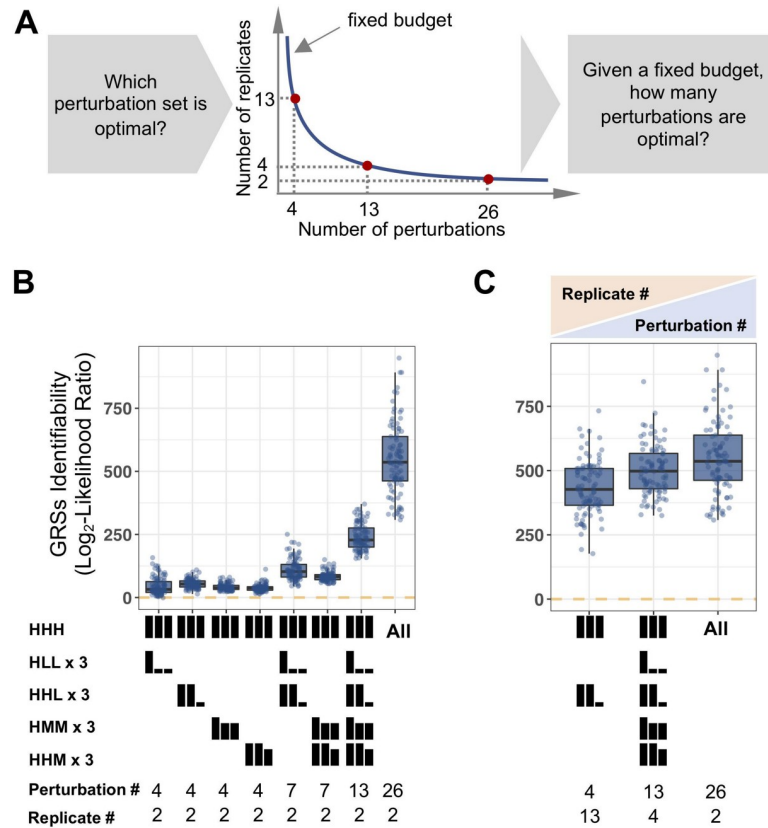
Fig 5. Principles of designing the experimental perturbation studies. (A) Schematic of the relationship between the number of replicates and perturbation conditions given a fixed budget to generate for example 52 datasets. (B) Effect of combining different perturbation sets on the identifiability (defined as the log2 likelihood ratio between the ground truth and the most similar alternative GRS) of the 93 GRSs. Testing 2 replicates for each condition, the number of perturbation sets determines GRS identifiability. Mathematically, we defined identifiability of each GRS as the lowest squared Euclidian distance between the gene expression responses of the ground truth GRS and any other GRS. (C) Comparison of the trade-off between the number of replicates and number of perturbation conditions for the identifiability of the 93 GRSs. When 52 datasets can be generated, employing more perturbation datasets is preferable to having more replicates for fewer perturbation sets.

potential involvement of MAPK that may compensate for loss of IRFs, and thus show little gene reduction in IRF3$^{-/-}$ cells compared to wild type (Fig 6D). This illustrates the limitation of using knockout data (IRF3$^{-/-}$) as a defining feature of whether a gene is regulated the TF, or, in this case, IRF-NFκB AND gate genes. Ppp1r15a and Srgn are examples of such genes that we identified as being potentially synergistically regulated by NFκB and IRFs with little reduction in expression in IRF3$^{-/-}$ cells (Figs 6D and 6E and S7C). Further analysis of ChIP-Seq data revealed induced binding of both RelA and IRF3 in the promoter and gene body of both genes (Fig 6F), confirming the potential regulation by both RelA and IRF3. To elucidate the regulation of these genes more precisely and identify unique GRS models requires additional perturbations, such as combinatorial knockout or inhibition of MAPK and IRF3.

Discussion

Over the past two decades a wealth of cellular transcriptomic and epigenomic data have been generated but there has only been limited progress in understanding the multi-TF gene regulation networks (GRNs) of mammalian cells [1,25]. Here we have presented a theoretical
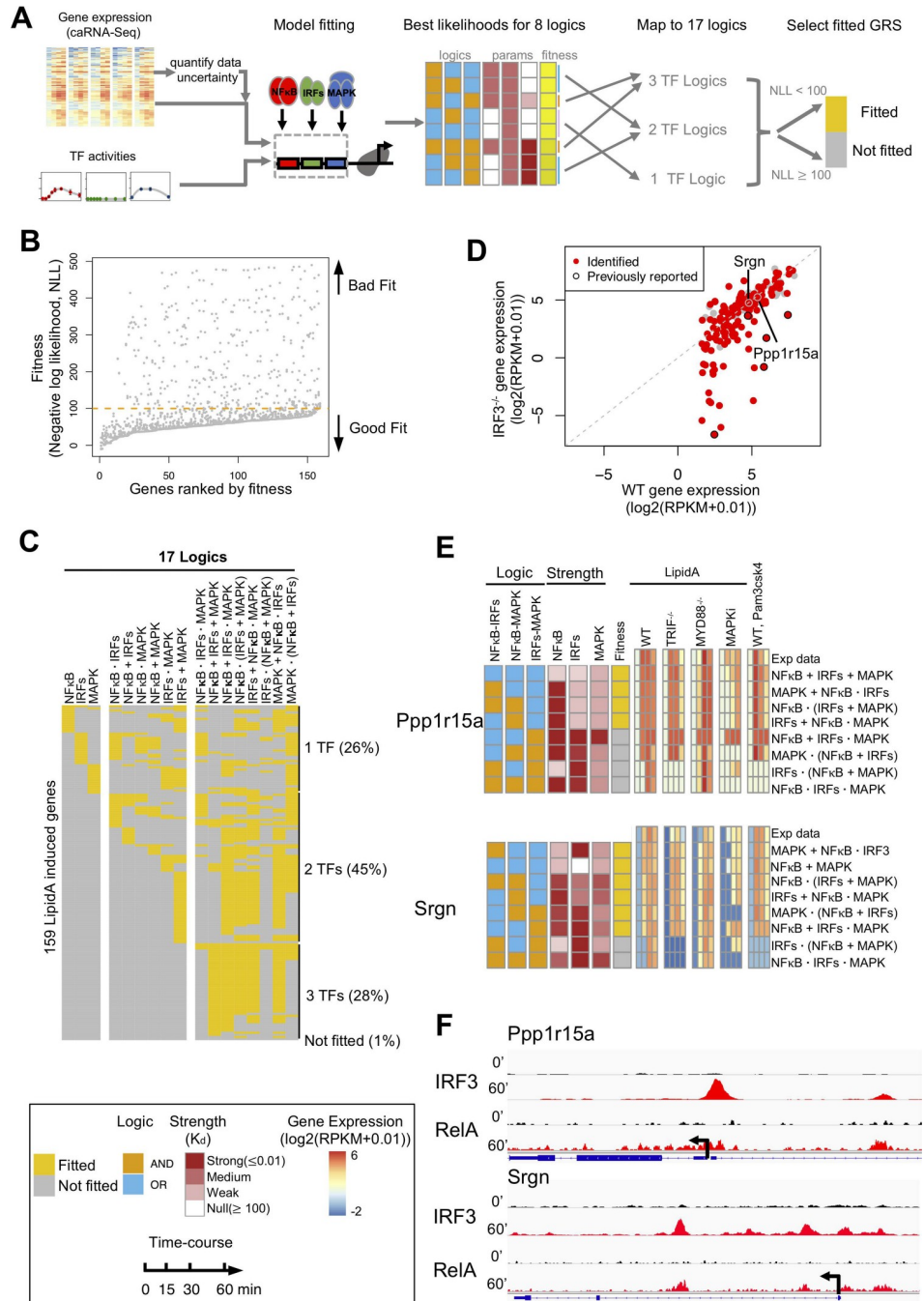
Fig 6. Mapping GRSs involving NFκB, IRFs, MAPK to endotoxin-induced immune genes. (A) Scheme of fitting models to experimental data. Gene expression data and TF activities for different stimuli were used to fit 8 possible model topologies and best likelihoods are determined; this was followed by mapping to 17 possible logics (Fig 1D), which were then determined to whether they fit the data, or not. (B) The plot of fitness (negative log likelihood) of 8 triple TF logic gates for all lipid A-induced genes identified by [19]. An arbitrary threshold to select fitted logic gates is marked as an orange dashed line. (C) The plot of fitted GRS logic gates for lipid A-induced genes identified by [19]. Only 26% of these genes fit a GRS governed by a single TF; most require GRS models that are controlled by at least two. (D) Peak expression of IRF3$^{-/-}$ is assessed against WT peak expression for all the lipid A-induced genes. Genes that are identified as potentially synergistically regulated by NFκB and IRFs are marked in red. The 5 previously reported IRF3-NFκB-regulated genes [19] are marked with a black circle. (E) For two example genes, heatmaps of experimentally measured caRNA-seq time-course data and simulation data by the best-fit GRS models, ordered by fit quality from top (best) to bottom. (F) IGV genome browser tracks of IRF3 and RelA ChIP-seq data in resting and 60 min lipid A-stimulated macrophages of the two genes.

analysis and a computational framework to examine the feasibility to infer abstracted mechanistic models of GRSs as stepping stones to develop the more detailed mechanistic understanding of GRNs. We show that GRS model identifiability is a function of the quantity and quality of data, and present a GRS model fitting computational workflow that includes an improved error model. We then apply this workflow to experimental data of innate immune gene expression in macrophages, and demonstrate its utility by revealing that the data supports alternative hypotheses that were previously missed by conventional analysis approaches. This work has been inspired by multiple previous studies to obtain mechanistic insights from genome wide genomic data using mathematical models involving logic gates [6]. Our work in particular examined the identifiability of different mechanistic models by measuring TF activity directly, and by considering different data uncertainty levels. Therefore, our work serves as a stepping stone to gain finer mechanistic insights of gene regulation.

Our computational analysis revealed a non-intuitive mapping relationship between model/ mechanism and gene expression responses. We showed that not all GRSs are equally identifiable, and nor are all GRS regulatory features equally identifiable. For example, GRSs formed by differential regulatory strengths are much harder to distinguish within the triple AND gate than within OR gates. But the triple AND gate logic feature is readily identifiable from all other GRSs with less available data. By displaying the model/mechanism-gene expression mapping relationship as a hierarchical tree, we found core regulation motifs that are hallmarks of specific GRS clusters and therefore play dominant roles in determining the input-output relationship. This finding suggested that some important insights can be gained reliably even when available datasets do not allow for the identification of a single GRS. For example, the triple AND gate and the double AND gate as core regulation motifs can be easily identified even when distinct GRSs that differ in the regulation strengths within each motif cannot be determined easily. Similarly, the emergent dynamical systems properties (e.g. perfect adaptation) may be mediated by diverse networks that nevertheless share common network motifs, as revealed by a function-topology map [26]. It is interesting that in our mechanism-gene expression map, we similarly found conserved core regulation motifs enriched in GRSs with similar input-output relationships. As such core regulation motifs determine their function, they are readily identifiable even from limited datasets.

In considering input-output datasets that involve perturbations at a given timepoint, we found it useful to precisely decompose value and temporal uncertainty for timecourse data. Our analysis of the "time-value" error model with synthetic data indicates that accurately accounting for data uncertainty can improve model identifiability by correctly penalizing less reliable data, but avoiding under- and over-estimates of the error associated with value. The time-value error model may be useful for timecourse data involving sharp changes, observed in data sets from highly dynamic biological processes like signal transduction or gene activation in response to stresses, immune threats or growth factors, where temporal error can greatly contribute to the data uncertainty.

After developing a practical computational workflow, we used simulated datasets to provide some guiding principles for experimental design for generating the most informative stimulus-response datasets. Model-guided experimental design is addressed in a rich literature [27–29], and some of our insights are intuitive. First, it is clear that given a limited budget, a greater number of distinct perturbation conditions should be prioritized over more replicates that would diminish error uncertainty. As an error model can be parameterized from replicates of a single perturbation condition, duplicates of other conditions primarily serve the purpose of ensuring the experiment did not in some way fail. Second, although perturbations that involve complete knockout may better distinguish GRSs in general, perturbations with partial activities are particularly helpful for determining regulation strengths. These involve dose response studies or siRNA-mediated knockdown rather than complete knockout. Not surprisingly, we

found that the optimal experimental perturbation condition set depends to some extent on the GRSs to be identified. This suggests that the optimal strategy is an iterative approach. To begin with, a set of perturbations that activate TFs singly or doubly can first segregate GRSs into multiple clusters. In the following round, based on these segregated clusters of GRSs and their shared features, additional perturbations may be selected to further dissect these GRSs and find other associated regulatory features.

Our work also provides guiding principles for mapping other types of observed data to the molecular network. The general approach outlined here—generating a model-gene expression map by comparing all possible hypotheses and fitting the model to integrated data sets—can be extended beyond the question of TF combinatorial regulation of transcription initiation. For example, epigenomic datasets that describe chromatin remodeling (ATAC-Seq), TF-DNA interactions (ChIP-Seq), chromatin conformation (Hi-ChIP, PLAC-seq), and protein-RNA interactions (CLIP-Seq), may be used to distinguish alternate models of multiple or sequential steps of gene regulation. Furthermore, the abstracted combinatorial models described here may be iteratively refined for example by addressing the TFs regulating different and sequential kinetic steps leading to transcriptional initiation [30,31], and may be used to identify additional regulatory mechanisms.

Applying the computational workflow to publicly available and newly generated input-output data of macrophages responding to immune stimuli illustrates the power of this quantitatively rigorous analysis workflow. Previous purely experimental studies relied heavily on knockout data to infer the involvement of a transcription factor. However, biological systems are characterized by overlapping functions of its components, a feature that confers functional robustness [32]. Hence, the regulatory role of a transcription factor on a particular gene is not reliably reflected in its functional requirement. In the case of pathogen-response gene expression, the potential for functional overlap manifests itself in many more genes being potentially combinatorially regulated than previously reported [18,19]. Indeed, physical binding data of the transcription factor IRF3 supports the notion that this transcription factor is more widely contributing to gene regulation than previously reported [19,33]. Further, the described computational workflow is extensible for keeping rigorous account of which GRSs may be excluded for each immune response gene as additional datasets become available.

In sum, as we lack a quantitative understanding of the molecular mechanisms governing many important biological regulatory systems, which would enable detailed regulatory network models, we describe here an approach for distinguishing quantitatively between alternate regulatory strategies using input-output datasets. Such an approach connects biological knowledge to data-driven approaches, and may prompt and guide subsequent mechanistic experimental and mathematical modeling studies. Considering TF combinatorial regulation alone, there are numerous examples that may benefit from the described workflow. In macrophages, the response to pathogens and cytokines is mediated by a handful of inducible TFs that are thought to engage in AND and OR gates [19,20]. Similarly, in hemogenic reprogramming or in induced pluripotent stem cell reprogramming initiated by three Yamanaka factors, combinatorial TF regulation triggers a cascade of successive molecular events that mediate the biological process [34,35]. Thus, the described computational framework may serve as a manual to gain quantitative understanding of gene regulation, and provide guidance for optimal experimental designs.

Methods

Model formulation

In this work, we quantitatively studied combinatorial GRSs formed by three TFs. Hence, we modeled nascent RNA expression from activated gene by considering RNA synthesis and first

order RNA processing. RNA abundance dynamics can be modeled using a single ODE:

$$\frac{dRNA}{dt} = k_{syn} \cdot f(t) - k_{proc} \cdot RNA \tag{1}$$

Here k_{proc} represents the nascent RNA processing rate, k_{syn} a constant synthesis rate that is modulated by the fractional promoter activity $f(t)$, which includes both combinatorial regulation from the GRS and some basal activity.

$$f(t) = (1 - k_0) \cdot GRS(TFs(t)) + k_0 \tag{2}$$

$f(t)$ describes how TFs regulate promoter activity, which is composed by a logic gate function GRS that depends on the TFs activities and some basal promoter activity denoted k_0.

Formulation of logic gates

We considered all possible synergistic (AND) and non-synergistic (OR) gene regulation strategies involving three TFs. As it has been shown in Fig 1D, there are 3 single logics (one for each TF), and 6 dual logics by considering two options "·" and "+" two options for the 3 pairs of the TFs ($2 \cdot 3 = 6$). The triple logics come from the 6 dual logics with an additional "·" and "+" for the third and by removing 4 redundant logics (2 of TF1+TF2+TF3, and 2 of TF1·TF2·TF3), leading to the final 8 triple logics ($2 \cdot 6-4 = 8$). The single and dual logics can be represented by a triple logic with null regulation strength of one or two TFs, see Fig 1D. This can also be derived by Boolean algebra, for example, TF1+TF2+TF3 = TF1+TF2 with TF3 = 0. Therefore, considering the 8 triple TFs logics is enough to represent all single, dual and triple logics. As synergistic (AND) and non-synergistic (OR) gene regulations can be modeled with thermodynamic models [15,17,18], the eight the derived logic gates represents four main structures that can be modeled by these formulas:

$$\text{TF}_1 \text{ AND TF}_2 \text{ AND TF}_3: \; G = \frac{[TF_1(t)]}{K_{d1} + [TF_1(t)]} \cdot \frac{[TF_2(t)]}{K_{d2} + [TF_2(t)]} \cdot \frac{[TF_3(t)]}{K_{d3} + [TF_3(t)]} \tag{3}$$

$$\text{TF}_1 \text{ AND } (\text{TF}_2 \text{ OR TF}_3): \; G = \frac{[TF_1(t)]}{K_{d1} + [TF_1(t)]} \cdot \left(1 - \frac{K_{d2}}{K_{d2} + [TF_2(t)]} \cdot \frac{K_{d3}}{K_{d3} + [TF_3(t)]}\right) \tag{4}$$

and similarly by rotation for TF$_2$ AND (TF$_1$ OR TF$_3$) and for TF$_3$ AND (TF$_1$ OR TF$_2$):

$$\text{TF}_2 \text{ AND } (\text{TF}_1 \text{ OR TF}_3): \; G = \frac{[TF_2(t)]}{K_{d2} + [TF_2(t)]} \cdot \left(1 - \frac{K_{d1}}{K_{d1} + [TF_1(t)]} \cdot \frac{K_{d3}}{K_{d3} + [TF_3(t)]}\right) \tag{5}$$

$$\text{TF}_3 \text{ AND } (\text{TF}_1 \text{ OR TF}_2): \; G = \frac{[TF_3(t)]}{K_{d3} + [TF_3(t)]} \cdot \left(1 - \frac{K_{d1}}{K_{d1} + [TF_1(t)]} \cdot \frac{K_{d2}}{K_{d2} + [TF_2(t)]}\right) \tag{6}$$

$$\text{TF}_1 \text{ OR } (\text{TF}_2 \text{ AND TF}_3): \; G = 1 - \frac{K_{d1}}{K_{d1} + [TF_1(t)]} \cdot \left(1 - \frac{[TF_2(t)]}{K_{d2} + [TF_2(t)]} \cdot \frac{[TF_3(t)]}{K_{d3} + [TF_3(t)]}\right) \tag{7}$$

and similarly by rotation for TF$_2$ OR (TF$_1$ AND TF$_3$) and for TF$_3$ OR (TF$_1$ AND TF$_2$):

$$\text{TF}_2 \text{ OR } (\text{TF}_1 \text{ AND TF}_3): \; G = 1 - \frac{K_{d2}}{K_{d2} + [TF_2(t)]} \cdot \left(1 - \frac{[TF_1(t)]}{K_{d1} + [TF_1(t)]} \cdot \frac{[TF_3(t)]}{K_{d3} + [TF_3(t)]}\right) \tag{8}$$

$$\text{TF}_3 \text{ OR } (\text{TF}_1 \text{ AND TF}_2): \; G = 1 - \frac{K_{d3}}{K_{d3} + [TF_3(t)]} \cdot \left(1 - \frac{[TF_1(t)]}{K_{d1} + [TF_1(t)]} \cdot \frac{[TF_2(t)]}{K_{d2} + [TF_2(t)]}\right) \tag{9}$$

$$\text{TF}_1 \text{ OR TF}_2 \text{ OR TF}_3 : \ G = 1 - \frac{K_{d1}}{K_{d1} + [TF_1(t)]} \cdot \frac{K_{d2}}{K_{d2} + [TF_2(t)]} \cdot \frac{K_{d3}}{K_{d3} + [TF_3(t)]} \tag{10}$$

Enumeration of representative GRSs. Our focus here is the synergistic and/or non-synergistic combinatorial regulation by 3 TFs to produce a specific GRS. Hence for each regulatory logic, we considered that each TF may have one of three regulation strengths (strong (S), medium (M), and weak (W)). Thus, we considered a total of 27 regulation strength combinations for 3 TFs and 8 logic gates, yielding a total 216 possible GRSs. Of these 216 GRSs, several may in fact not be efficiently activated. For example, AND gate configurations in which one component has a "Weak" regulatory strength cannot be activated well within the present scheme. In this way, we identified 69 poorly activated GRSs and removed them from further consideration. In addition, several GRSs were found to be logically equivalent. For example, the TF1 is logically equivalent to either a triple OR gate in which TF2 and TF3 have "Weak" regulatory strength, or a TF1 or (TF2 and TF3) gate in which TF2 or TF3 have "Weak" regulatory strength. In this way we removed 54 redundant GRSs. These first principle considerations led us to a list of 93 potentially identifiable GRSs (see S2 Fig).

Perturbations

We defined perturbations as following functions.

$$\text{Simple perturbation}: \ f(t) = \begin{cases} 1, high \\ 0, low \end{cases} \tag{11}$$

In addition to the simple perturbation, we have:

$$\text{Amplitude}: \ f(t) = 0.5, 0 \le t \le 60 \tag{12}$$

$$\text{Gradient}: \ f(t) = \frac{t}{60}, 0 \le t \le 60 \tag{13}$$

$$\text{Delayed}: \ f(t) = \begin{cases} 0, 0 < t \le 30 \\ 1, 30 < t \le 60 \end{cases} \tag{14}$$

$$\text{Transient}: \ f(t) = \begin{cases} 1, 0 < t \le 30 \\ 0, 30 < t \le 60 \end{cases} \tag{15}$$

GRS simulation

In our work, we defined GRS as the combination of logic gate and regulation strength parameters, which together determine gene expression levels. For testing purpose, we assume TFs activities range from 0 to 1 (i.e. TFs are normalized to avoid assay specific dynamic span) and define the high level as 1, the medium level as 0.5, and the low level as 0. Based on Hill function, we define regulation strength ranging from weak to strong with strong corresponding to $K_d = 0.1$, medium to $K_d = 1$, and weak to $K_d = 10$, given that K_d has minimal effect on promoter activity outside this range. We consider gene with a short RNA processing rate (7min, i.e. k_{proc} = 0.1 min^{-1}). Here, we simulate a timecourse gene expression at 0, 15, 30, 60 min, but our approach can be generalized into any time points served for the interests of experiments. To

Table 1. Model Parameters.

Parameters	Description	Range	Value
[TFs] (Inputs)	TF amplitude	High	1
		Medium	0.5
		Low	0
K_d	TF regulation strength	Strong	0.1
		Medium	1
		Weak	10
k_0	Basal gate value		0
k_{syn}	Nascent mRNA synthesis rate		1 RPKM/min
k_{proc}	Nascent mRNA processing rate		0.1 min^{-1}

compare all the GRSs, we simulated resulting gene expression with k_{syn} = 1 RPKM/min (where RPKM means reads per kilobase per million of mapped reads), and normalized the gene expression such that its maximal value corresponds to 100 across all the perturbation conditions. This is equivalent to adjusting k_{syn}. In Table 1, you can find a summary of the different parameters used for the simulations:

Computational analysis

We define the distance between two gene expression profiles as the squared Euclidean distance for each timepoints across all perturbation conditions:

$$d(gene_i, gene_j) = \sum_{p \in perturbations} \sum_{t \in timepoints} (ge_{i,p}(t) - ge_{j,p}(t))^2 \tag{16}$$

where $ge_{i,p}(t)$ means gene expression of gene i for perturbation p at time t. We used a hierarchical clustering with single linkage approach to cluster GRSs based on their gene expression profile. With single linkage approach, the distance between two clusters of GRSs is defined as the minimum distance of inter-pairs, as we considered two groups of GRSs is not identifiable if one inter-pair distance is below a certain tree height threshold, called in the rest of the manuscript separation threshold.

GRS analysis with randomly sampled parameters

We followed the same procedures described in GRS simulation to examine GRS with randomly sampled parameters. Specifically, we generated 1000 sets of parameters of each logic gate by sampling uniformly from parameter space (described in Table 2) in logarithm scale. We only selected the activatable GRS for the downstream analysis. The activatable GRS are defined as GRS that can produce an output higher than the maximum value of previously identified 69 poorly activated GRS (here the threshold is 1.66 for GRS output) when the 3 input TFs are fully active (i.e. activity of 1). As sampling effects of k_{syn} will be counteracted by the normalization of gene expression, the random sampling of k_{syn} was replaced by constant number (1 RPKM/min).

Table 2. Parameter ranges.

Parameters	Description	Sampling Range
K_d	TF regulation strength	0.01–100
k_0	Basal gate value	0–1
k_{proc}	Nascent mRNA processing rate	0.01 min^{-1}–10 min^{-1}

Error model

Reliable point specific estimation.   To achieve reliable point specific estimation from small sample size, we leveraged global trend from all the points to stabilize the estimation. For this purpose, we used Bayes' rule to combine global trend (prior distribution) with point specific information using maximum likelihood estimation (MLE) to generate a stabilized estimate (posterior distribution). This empirical Bayes approach have been broadly used in handling small sample problem in bulk and single cell RNA-seq data, including by DESeq2 [36], voom [37], and BASiCS [38].

Specifically, we first estimate the error parameters ξ_{prior}, which are defined based on the error model described later, from global trend by pooling all the points (i.e. all genes and time points) together. Next, we combine the global trend with data of each time point i to estimate the posterior distribution $\xi_{i,posterior}$ of each time point, which is given by:

$$\xi_{i,posterior} = \underset{\xi_i}{\text{argmax}} \ \mathcal{L}(\xi_i)P(\xi_i) \tag{17}$$

Where $\mathcal{L}(\xi_i)$ is the likelihood function that represents information from each time point i, and $P(\xi_i)$ corresponds to the prior knowledge on the distribution of ξ_i which we assume to be normal with mean $\hat{\xi}_{prior}$ estimated from global trend, and $\sigma^2_{\xi_{prior}}$ as variance. Here, $\sigma^2_{\xi_{prior}}$ allows scaling the importance of the global trend over the point specific information. We empirically chose the prior variance based on our confidence of how point-specific information contributes to the estimation. Specifically, we first start with a very small prior variance 0.00001 so that the posterior would almost be the same. Then we gradually increase the prior variance, thereby putting more weight on point-specific information, and examine whether the performance (correlation between ground truth and estimation) is improving. By doing so, we may choose the optimal prior variance given the number of available replicates. If there are few replicates (less than 10), then we tend to put high weight i.e. confidence (small $\sigma^2_{\xi_{prior}}$) on the global trend so that the final posterior estimation is similar to the global trend. If a large number of replicates are available (more than 10), we may increase the weight of the point specific information (larger $\sigma^2_{\xi_{prior}}$).

Uncertainty estimation.   Uncertainty originates from various sources. Here, we have considered uncertainties coming from biological variabilities including TF abundance variations, gene response time variations, chemical parameters variations, and technical uncertainty including sample preparation timepoints variation and assay measurement variation. We separately modeled them into value uncertainty (TF abundance variations, chemical parameters variations, and assay measurement variation), and temporal uncertainty (gene response time variations and sample preparation timepoints variation), as those two sources of uncertainties are orthogonal (value and temporal), and have two different mathematical forms (temporal uncertainty is curve shape dependent, and value uncertainty is not).

Temporal uncertainty estimation.   To model the temporal uncertainty of a point (t, y_t) due to potential uncertainty in the exact time the measurement was done, we assume that it follows a Gaussian distribution with mean t and standard deviation σ_t. The resulting uncertainty in y due to the uncertainty in t is denoted by $\sigma_{temporal}$.

For a small neighborhood around t, the curve, which passes the point in such a small region, can be approximated by a line (linearization) with the slope as tangent of the curve. Therefore with the assumption that the uncertainty of t follows a Gaussian distribution, then y will also follow a Gaussian distribution, where variance is given by [39]:

$$\sigma_{temporal}^2 = Slope^2 \cdot \sigma_t^2 \tag{18}$$

This can be considered as an error propagation function, which is:

$$(\Delta y)^2 = (Jacobian \cdot \Delta t)^2 \tag{19}$$

Value uncertainty estimation. As we know multiple uncertainty sources can cause observed value uncertainty, we assume the overall observed value uncertainty follows a Gaussian distribution.

To obtain robust value uncertainty estimation, we modeled mean-variance relationship by a polynomial regression, where σ^2_{value} is representing value variance that can be predicted by the mean of point μ. This helps stabilize the estimation of variance by leveraging its dependency relationship with robust estimator, mean. Similar approach to capture mean-variance relationship was introduced by DESeq2 [36], where it uses fixed relationship, and BASiCS [38], where it uses more flexible semi-parametric approach with Gaussian radial basis function (GRBF) kernels. We use polynomial regression as it can capture the constant, linear, and quadratic variance components. It is given by:

$$\sigma^2_{value} = f(\mu) = \alpha_0 + \alpha_1 \cdot \mu + \alpha_2 \cdot \mu^2 \tag{20}$$

Here, we only include the first two orders, as this already captures well the mean-value variance dependency. Higher order terms can be added when dealing with more complex datasets.

Distribution of gene expression. We separately modeled background and induced gene expression distribution. These two parts are usually caused by different sources, and exhibit different level of uncertainty. In addition, the distribution for background uncertainty is desired to have $\mathbb{R}_{\geq 0}$ as its support, as all measurements are positive. Therefore, the distribution of gene expression is modeled by gamma and normal distributions, depending on the expression level, which is given below:

$$gene_i \sim \begin{cases} \Gamma(k_{bg,i}, \theta_{bg,i}), if\ \mu_i \leq \eta \\ \mathcal{N}(\mu_i, \sigma^2_{induc,i}), if\ \mu_i > \eta \end{cases} \tag{21}$$

Here, we connect data mean and variance to the gamma distribution's mean and variance to parameterize k, θ:

$$(k_{bg,i} - 1) \cdot \theta_{bg,i} = \mu_i$$

$$k_{bg,i} \cdot \theta_{bg,i}^2 = \sigma^2_{bg} \tag{22}$$

where,

$$\sigma^2_{induc,i} = \alpha_{induc,0,i} + \alpha_{induc,1,i} \cdot \mu_i + \alpha_{induc,2,i} \cdot \mu_i^2 + Slope_i^2 \cdot \sigma^2_{induc,t,i} \tag{23}$$

In addition, we set the threshold η to be 3 RPKM.

Implementation of error model

Simulation of data uncertainty. We followed the first principle to generate data uncertainty from both biological variability and technical uncertainty. We generated biological variability by varying TF amplitude, gene response time and parameters of the model by sampling from the distribution specified in Table 3. We generated technical uncertainty by varying sample preparation timepoints, and adding noise to the assay measurement, as specified in Table 3.

Table 3. Simulated uncertainty.

Uncertainty Source	Form	Distribution
TF amplitude	$TF \cdot (1+\varepsilon)$	$\varepsilon \sim \mathcal{N}(0, 0.0004)$
Gene response time	$t+\varepsilon$	$\varepsilon \sim \mathcal{N}(0, 25)$
Parameters (K_d, k_{syn}, k_{proc})	$Parameters \cdot e^{\varepsilon}$	$\varepsilon \sim \mathcal{N}(0, 0.0004)$
Parameters (k_0)	$k_0 + \varepsilon$	$\varepsilon \sim U(0, 0.02)$
Sample preparing timepoints	$t+\varepsilon$	$\varepsilon \sim \mathcal{N}(0, 25)$
Assay measurement	$y+\varepsilon$	$\varepsilon \sim \mathcal{N}(0, 0.01 \cdot y^2)$

Slope estimation. To reliably estimate the slope of the curve, we first interpolated the curve with a piecewise cubic hermite function (pchip function in the R signal package), and then estimated the slope from the interpolated curve. This allowed us to leverage information from all points and the shape of the curve. Specifically, the slope is calculated by the central difference:

$$Slope = \frac{\Delta y_{fit}}{\Delta t} = \frac{y_{fit, t+\frac{\Delta t}{2}} - y_{fit, t-\frac{\Delta t}{2}}}{\left(t + \frac{\Delta t}{2}\right) - \left(t - \frac{\Delta t}{2}\right)} \tag{24}$$

The choice of Δt depends on temporal uncertainty level, as one can imagine we should take bigger Δt when temporal uncertainty is high. Here, we set $\Delta t = 2 \cdot \hat{\sigma}_t$ ($\hat{\sigma}_t$ is the unbiased estimated temporal uncertainty). We set the slope to be 0 for sample at time 0, as it comes from unstimulated cells and should correspond to steady state. We interpolate the time series to range from 0 to 80 min to allow for a better estimate of the slope for the last time point.

Error inference

Likelihood function. We first derived the likelihood function for the inference with parameters $\xi = (\sigma_{bg}^2, \alpha_0, \alpha_1, \alpha_2, \sigma_t^2)$. The probability distribution for measured y_i at time point i is given by:

$$y_i | \xi_i, Slope_i^2, \mu_i \sim \begin{cases} \Gamma\left(y_i \middle| \frac{\mu_i^2}{\sigma_{i,bg}^2}, \frac{\sigma_{i,bg}^2}{\mu_i}\right), if \ \mu_i \leq 3 \\ \mathcal{N}(y_i | \mu_i, \sigma_{total,i}^2), if \ \mu_i > 3 \end{cases} \tag{25}$$

Where $\mu_i = \bar{y}_i$. Therefore, the point likelihood $\mathcal{L}_{point}$ is given by:

$$\mathcal{L}_{point}(\xi_i) = I(\mu_i \leq 3)\Gamma\left(y_i \middle| \frac{\mu_i^2}{\sigma_{i,bg}^2}, \frac{\sigma_{i,bg}^2}{\mu_i}\right) + I(\mu_i > 3)\mathcal{N}\left(y_i | \mu_i, \sigma_{total,i}^2\right) \tag{26}$$

Here, Γ and $\mathcal{N}$ are used to denote the probability density function (PDF) of gamma and gaussian distributions respectively. We will use this notation to describe both the distribution or its PDF based on the context in the rest of the manuscript.

Next, we estimated ξ_{prior} from global trend of points using global likelihood $\mathcal{L}_{global}$, which is given by the multiplication of the point likelihood from all the perturbations p, genes g and time points t. $\mathcal{L}_{global}$ is given by:

$$\mathcal{L}_{global}(\xi_{prior}) = \prod_p \prod_g \prod_t \mathcal{L}_{point}(\xi_{prior}) \tag{27}$$

Finally, we use Bayes' rule to estimate $\xi_{i,posterior}$ by combining global trend with point specific information. We assume ξ_{prior} follows a normal distribution, with mean as $\hat{\xi}_{prior}$ and variance as $\sigma^2_{\hat{\xi}_{prior}}$. Based on the confidence of prior information, we can adjust $\sigma^2_{\hat{\xi}_{prior}}$ to change the weights of point information and global trend. It is given by:

$$\xi_{i,posterior} = \underset{\xi_i}{\text{argmax}}\ \mathcal{L}(\xi_i)P(\xi_i) = \underset{\xi_i}{\text{argmax}}\ P(\boldsymbol{y}_i|\xi_i, Slope_i^2, \mu_i)\mathcal{N}(\xi_i|\hat{\xi}_{prior}, \sigma^2_{\hat{\xi}_{prior}}) \tag{28}$$

For the computational efficiency during optimization, we convert all the target function as negative log, and use Limited-memory BFGS (L-BFGS) method which is implemented in the optim function in R for optimization. More details can be found in the parameter estimation and model selection part.

As MLE of variance is a biased estimator, we have corrected it to the unbiased estimator by:

$$\hat{\sigma}^2_{total} = \frac{N}{N-1}\hat{\sigma}^2_{total,MLE} \tag{29}$$

where N is the number of replicates (here N = 2). All the estimated parameters for variance have been corrected by:

$$\hat{\xi} = \frac{N}{N-1}\hat{\xi}_{MLE} \text{ with } \hat{\xi} = \left(\hat{\sigma}^2_{bg,MLE}, \hat{\alpha}_{0,MLE}, \hat{\alpha}_{1,MLE}, \hat{\alpha}_{2,MLE}, \hat{\sigma}^2_{t,MLE}\right) \tag{30}$$

Conventional model. In the conventional model, we simply set σ^2_t to be zero, and keep the rest to be the same for comparison.

Model parameterization with Bayesian framework

We considered gene expression data from multiple perturbation conditions $p = 1,\ldots,N$, with multiple time points $t = 0, 15, 30, 60$ min of each perturbation. This yields the probability to observe experimental data $\boldsymbol{y}^{(obs)}$ as:

$$P(\boldsymbol{y}^{(obs)}) = \prod_p\prod_i P(\boldsymbol{y}^{(obs)}_{p,t_i}) \tag{31}$$

The statistical model of observing the experimental data $\boldsymbol{y}^{(obs)}$ given its true value $\boldsymbol{y}_{p,t_i}$ can be expanded as:

$$P(\boldsymbol{y}^{(obs)}) = \prod_p\prod_i I(\boldsymbol{y}_{p,t_i} \leq 3)\Gamma(\boldsymbol{y}^{(obs)}_{p,t_i}|\boldsymbol{y}_{p,t_i}, \sigma^2_{bg}) + I(\boldsymbol{y}_{p,t_i} > 3)\mathcal{N}(\boldsymbol{y}^{(obs)}_{p,t_i}|\boldsymbol{y}_{p,t_i}, \sigma^2_{p,t_i}) \tag{32}$$

Likelihood function. In our work, the experimental data $\boldsymbol{y}^{(obs)}$ is explained by the combination of two parts, 1) model simulated $\boldsymbol{y}_{p,t_i}$, which is determined by the GRS and kinetic parameters described by:

$$\boldsymbol{\theta} = (\underbrace{\text{logic gate}, K_{d1}, K_{d2}, K_{d3}}_{\text{GRS}}, \underbrace{k_{syn}, k_{proc}, k_0}_{\text{kinetic parameters}}) \tag{33}$$

2) data uncertainty, which is estimated from global trend $\hat{\xi}_{prior} = (\hat{\sigma}^2_{bg}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_t)$ in time-value error model. Therefore, the likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_p\prod_i P(\boldsymbol{y}^{(obs)}_{p,t_i}|\boldsymbol{y}_{p,t_i}(\boldsymbol{\theta}), \hat{\xi}_{prior}) = \prod_p\prod_i(I(\boldsymbol{y}_{p,t_i} \leq 3) \cdot \Gamma(\boldsymbol{y}^{(obs)}_{p,t_i}|\boldsymbol{y}_{p,t_i}(\boldsymbol{\theta}), \hat{\sigma}^2_{bg}) + I(\boldsymbol{y}_{p,t_i}$$
$$> 3) \cdot \mathcal{N}(\boldsymbol{y}^{(obs)}_{p,t_i}|\boldsymbol{y}_{p,t_i}(\boldsymbol{\theta}), \hat{\alpha}_0 + \hat{\alpha}_1 \cdot \boldsymbol{y}_{p,t_i}(\boldsymbol{\theta}) + \hat{\alpha}_2 \cdot \boldsymbol{y}_{p,t_i}(\boldsymbol{\theta})^2 + Slope^2_{p,t_i} \cdot \hat{\sigma}^2_t)) \tag{34}$$

We took the negative log likelihood for the optimization.

Computational analysis with data uncertainty. We define the distance between gene expression i and gene expression j as the average of $NLL(i,j)$ and $NLL(j,i)$, where $NLL(i,j)$ is the negative log likelihood to observe gene expression i, given gene expression j as ground truth. We then performed hierarchical clustering with single linkage approach and the defined distance matrix.

Computational workflow implementation. We first applied time-value error model to estimate data uncertainty level. We took average of two replicates of TF activities as input. We then incorporate averaged TF activities as input, simulated gene expression data as output and their estimated data uncertainty into the model parameterization part.

Parameters estimation and model selection. To estimate parameters and select logic gates, we first optimized parameters for each logic gate, and picked the best fitted logic gate and its estimated parameters among all eight logic gates. In addition, we constrained the range of K_d from 10 fold weaker and stronger than the weak and strong regulation strength in the process of optimization (i.e. from 0.01 to 100), as K_d has minimal effect on the GRS output beyond this range. All the parameters are optimized in logarithmic scale, as it enables the algorithm to quickly search large space. We used multi-start local optimization approach, as it has been shown outperformed or at least perform as well as some global optimization methods [40]. Specifically, we employed the optim function with method Limited-memory BFGS (L-BFGS) in R for optimization, because it provides constrains for parameters, and it is computationally efficient. We used 300 randomly sampled initial sets of parameters for optimization for each logic gate. To ensure confidence that we sample the parameter space sufficiently, we examined the identifiability of each GRS as a function of the number of random initializations, and ensured it converges with 300 initial sets. The best fitted parameter set of 300 initial sets is chosen for the parameters of each logic gate. As all the logic gates have the same model complexity i.e. the same number of parameters, we simply select the logic gate based on their likelihood instead of using Akaike information criterion (AIC) or Bayesian information criterion (BIC). While our workflow allows people to select multiple good fitted models based their own criteria, we simply selected the logic gate with highest likelihood as the best fitted model.

Varying data uncertainty level. We tested our error model and pipeline with different level of data uncertainty. As each uncertainty sources don't equally contribute to the final data uncertainty, we adjusted the altered level of each uncertainty source, so that the final error contribution from each of uncertainty sources would be approximately the same. After we simulated the data, we examined the overall uncertainty (mean of empirical estimated variance of all the points) caused by all the uncertainty sources (Table 4).

Macrophage cell culture

Primary Bone Marrow Derived Macrophages (BMDMs) were prepared by culturing bone marrow cells from femurs of female 8–12 weeks old WT mice or different knock-out mice in

Table 4. Simulated uncertainty in multiple uncertainty levels.

Uncertainty Source	Form	Distribution	θ				
TF amplitude	$TF \cdot (1+\varepsilon)$	$\varepsilon \sim \mathcal{N}(0,\theta)$	0.25x	0.5x	4.0×10^{-4}	1.1x	1.45x
Gene response time	$t+\varepsilon$	$\varepsilon \sim \mathcal{N}(0,\theta)$	0.25x	0.56x	25	2x	4x
Parameters (K_d, k_{syn} k_{proc})	$Parameters \cdot e^{\varepsilon}$	$\varepsilon \sim \mathcal{N}(0,\theta)$	0.25x	0.5x	4.0×10^{-4}	1.1x	1.45x
Parameters (k_0)	$k_0+\varepsilon$	$\varepsilon \sim U(0,\theta)$	0.5x	0.75x	2.0×10^{-2}	1.05x	1.2x
Sample handing time	$t+\varepsilon$	$\varepsilon \sim \mathcal{N}(0,\theta)$	0.25x	0.56x	25	2x	4x
Sequencing error	$y+\varepsilon$	$\varepsilon \sim \mathcal{N}(0,\theta \cdot y^2)$	0.25x	0.56x	1.0×10^{-2}	1.1x	1.4x
Overall Uncertainty			0.21x	0.53x	1x	1.64x	2.66x

L929-conditioned medium by standard methods [18]. BMDMs were grown for 7 days and stimulated with different agonists on day 8. BMDMs were stimulated with 100 ng/ml LipidA, 100 ng/mL LPS, as well as with a TLR2/1 agonist, the synthetic triacylated lipoprotein Pam3CSK4 (PAM) (3 μg/mL).

TF activity quantification and normalization

Western blotting analysis and EMSAs were conducted with standard methods as described previously [18]. Briefly, nuclear extracts were prepared by hypotonic cell lysis and high salt extraction of nuclear proteins. The band intensities for Western blots or EMSA gels were measured in Image-Quant software. The samples within a timecourse that had peak intensities for different perturbations were run on the same gel, quantified and used for normalization of respective band intensities of different agonists.

To quantify the TF activity upon stimulation, we first linearly scaled the value of band intensities so that for each perturbation to the range that the basal and peak band intensity to be 1% and 100% by the formula:

$$TF_{scaled} = 0.99 \frac{TF_{raw} - TF_{basal}}{TF_{peak} - TF_{basal}} + 0.01 \tag{35}$$

where TF_{scaled} and TF_{raw} are the values of TF band intensity before and after scaling. TF_{peak} and TF_{basal} are the peak and basal band intensity before scaling.

To compare TF activities in different perturbation conditions, we have normalized them by setting peak of wild type LipidA stimulation to be 1. We then normalized all the other perturbation conditions by multiplying each of the other perturbation conditions p with factor p_{norm}:

$$p_{norm} = \frac{TF_p^{(peak)}}{TF_{WTLipidA}^{(peak)}} \tag{36}$$

where $TF_p^{(peak)}$ and $TF_{WTLipidA}^{(peak)}$ are the peak intensities of perturbation p and perturbation for wild type stimulated with LipidA that has been measured from the same gel.

Inference of MAPK-regulated transcription factor activity. We first defined the MAPK targeted genes: Egr1, Fos, and Dusp4, that are neither NFκB nor IRF targets. For those target genes, we estimate TF activities from their gene expression in RNA-seq. Specifically, for individual target genes, we linearly convert their gene expression (RPKM) so that in the WT stimulated with LipidA condition, the basal and peak value are 1% and 100%, by using the same formula described in the last section. Finally, we take the averaged gene expression from all the converted target genes as the inferred MAPK-regulated transcription factor activity.

RNA-seq data processing

We have collected all the chromatin associated RNA-seq data from previously published work [19]. BAM files of chromatin associated RNA-seq data have been downloaded from GEO with the series accession number GSE67357. In the original BAM files, reads has been aligned to the mouse genome (NCBI37/mm9) with TopHat v1.3.3 by allowing up to two mismatches per read in one alignment. We followed the same standard in the paper to calculate RPKM for the chromatin associated RNA-seq by dividing all mapped reads within the transcription unit (both intron and exon) by the length of the entire locus.

Induced genes have been selected by the same criteria described in the paper [19]: the peak RPKM of induced genes is larger than 3 at any given time point, and the gene was induced more than 10 fold at any time point comparing to the time point 0min, with statistical

significance p < 0.01 defined by the edgeR package in R Bioconductor [41]. In addition, a gene was also considered induced if it had more than a 5 fold gene induction at 15min.

For MAPK inhibited perturbation, we have collected both MAPK inhibited LipidA stimulation, and control condition (wild type in solvent with LipidA stimulation). To make the MAPK inhibited condition comparable to the other conditions, we normalized each time points by multiplying by the factor $\frac{RNA_{WT}(t_i)}{RNA_{control}(t_i)}$, where $RNA_{WT}(t_i)$ and $RNA_{control}(t_i)$ are the gene expression (RPKM) at time point i in wild type stimulated with LipidA and wild type with solvent stimulated with LipidA.

ChIP-Seq data processing

We have collected all the RelA and IRF3 ChIP-Seq data from previously published work [19], and downloaded the bigWig files from GEO with the series accession number GSE67357. In the bigWig files, the reads had been aligned to the mouse genome (NCBI37/mm9) with Bowtie2. We examined and visualized the tracks of ChIP-seq data with Integrative Genomics Viewer (iGV) [42].

Identify GRSs for immune response genes

Error quantification for chromatin associated RNA-seq data.   We applied the developed error model to the collected chromatin associated RNA-seq data to estimate data uncertainty. For RNA-seq data with only single replicate (TRIF$^{-/-}$ with LipidA stimulation, wild type with Pam3csk4 stimulation), we used the values estimated from wild type with LipidA stimulation.

Final model for studying immune response genes.   We applied the same developed mechanistic model to identify GRSs for immune response genes. Specifically, to better capture basal gene expression, we set $k_0 = 0$ for the model, and used experimental basal gene expression (gene expression in unstimulated condition, 0min) as gene basal expression for fitting, with formula:

$$y_{sim,p,i} = y_{sim\,raw,p,i} + y_{basal,p} \tag{37}$$

where $y_{sim,p,i}$ and $y_{sim\,raw,p,i}$ are the final and raw model (with $k_0 = 0$) simulated gene expression at time point i and perturbation condition p. $y_{basal,p}$ is the experimental basal gene expression at perturbation p.

Fitting for experimental data.   We applied the same Bayesian framework and likelihood function to identify GRS for immune response genes. After parameter optimization for all the 8 logic gates, we mapped them back to the 17 logics gates by assigning the 8 triple TF logics gates to single or dual logic gates if one or two TFs has null regulation strength (estimated $K_d > 100$). Then, we only selected the potential GRSs for downstream analysis by setting a threshold on the fitting score (NLL < 100).

Supporting information

S1 Fig. Exploring how logic gates behave as a function of variable TF regulation strengths. Related to Fig 1. Heatmaps of promoter activity as a function of logic gates (AND, OR) and varying TF1, TF2 regulation strengths.
(TIFF)

S2 Fig. Extended details of GRS analysis. Related to Fig 2. Diagram explaining how the 216 possible GRSs were reduced to 93 potentially identifiable GRSs. First 69 poorly activated GRSs were removed (left side), as well as 54 redundant GRSs given that they are logically equivalent to a single TF (top right) or to two TFs with an AND gate (bottom right). This results in 93

potentially identifiable GRSs.
(TIFF)

S3 Fig. Confirmation of conclusions with multiple alternative approaches. Related to Fig 2. (A). The heatmap as Fig 1G but reordered with different hierarchical clustering methods (complete and average linkage) in comparison to the used single linkage method used in Fig 1G. (B). The same plot as Fig 2C but only using average or complete linkage methods for hierarchical clustering. (C). Heatmap of gene expression in response to the 7 perturbation conditions from 1981 activatable GRSs that are generated by random sampling of all the parameters. (TIFF)

S4 Fig. Variance of each time points. Related to Fig 3. Comparison between directly estimated raw variance of two replicates and ground truth.
(TIFF)

S5 Fig. Comparison between time-value error model and conventional model. Related to Fig 4. (A). Comparison of the estimated regulation strength between time-value error model, conventional error model, and raw variance with different levels of noise. (B). Confusion matrix of all the 93 estimated GRSs from time-value, conventional error models and raw variance.
(TIFF)

S6 Fig. Extended study of perturbations and replicates trade-off. Related to Fig 5 Comparison of the number of replicates and number of perturbation conditions for the identifiability of the 93 GRSs with a total of 8 datasets.
(TIFF)

S7 Fig. Experimental data and identified GRS for lipid A induced genes. Related to Fig 6. (A). Graphs of measured and inferred TF activities in the perturbation conditions used by the caRNA-seq study. (B). Heatmap of measured nascent mRNA expression data together with fitness for all the candidate logic gates. The left side shows fitness (negative log likelihood) of all the possible logics. The white boxes are the logics that do not account for the data when mapping 8 logics back to the 17 logics. Genes are ordered by their expression with hieratical clustering approach. (C) Heatmaps of experimentally measured caRNA-seq time-course data and simulation data by the best-fit GRS models using the Time-Value error model or raw variance, as indicated. (D) Comparison of the negative log likelihood (NLL) of the best GRS model that contains any synergistic component to the negative log likelihood (NLL) of the best GRS model that does not contain any synergistic component.
(TIFF)

Author Contributions

Conceptualization: Ning Wang, Alexander Hoffmann.

Data curation: Anup Mazumder.

Formal analysis: Ning Wang, Diane Lefaudeux.

References

1. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004 Apr 16; 117(2):185–98. https://doi.org/10.1016/s0092-8674(04)00304-6 PMID: 15084257.

2. Salleh FH, Arif SM, Zainudin S, Firdaus-Raih M. Reconstructing gene regulatory networks from knock-out data using Gaussian Noise Model and Pearson Correlation Coefficient. Computational biology and chemistry. 2015 Dec 1; 59:3–14. Epub 2015 Jun 17. https://doi.org/10.1016/j.compbiolchem.2015.04.012 PMID: 26278974.

3. Zhang X, Zhao XM, He K, Lu L, Cao Y, Liu J, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics. 2012 Jan 1; 28(1):98–104. Epub 2011 Nov 15. https://doi.org/10.1093/bioinformatics/btr626 PMID: 22088843.

4. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics 2006 Mar (Vol. 7, No. 1, pp. 1–15). BioMed Central. https://doi.org/10.1186/1471-2105-7-S1-S7 PMID: 16723010; PMCID: PMC1810318.

5. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biol. 2007 Jan 9; 5(1):e8. https://doi.org/10.1371/journal.pbio.0050008 PMID: 17214507; PMCID: PMC1764438.

6. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome biology. 2006 May; 7(5):1–6. Epub 2006 May 10. https://doi.org/10.1186/gb-2006-7-5-r36 PMID: 16686963; PMCID: PMC1779511.

7. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional regulatory circuits: predicting numbers from alphabets. Science. 2009 Jul 24; 325(5939):429–32. https://doi.org/10.1126/science.1171347 PMID: 19628860; PMCID: PMC2745280.

8. Vonsattel JP, DiFiglia M. Huntington disease. Journal of neuropathology and experimental neurology. 1998 May 1; 57(5):369. https://doi.org/10.1097/00005072-199805000-00001 PMID: 9596408.

9. Sen S, Cheng Z, Sheu KM, Chen YH, Hoffmann A. Gene regulatory strategies that decode the duration of nfκb dynamics contribute to LPS-versus TNF-Specific gene expression. Cell systems. 2020 Feb 26; 10(2):169–82. Epub 2020 Jan 22. https://doi.org/10.1016/j.cels.2019.12.004 PMID: 31972132; PMCID: PMC7047529.

10. Ngo KA, Kishimoto K, Davis-Turak J, Pimplaskar A, Cheng Z, Spreafico R, et al. Dissecting the regulatory strategies of NF-κB RelA target genes in the inflammatory response reveals differential transactivation logics. Cell reports. 2020 Feb 25; 30(8):2758–75. https://doi.org/10.1016/j.celrep.2020.01.108 PMID: 32101750; PMCID: PMC7061728.

11. Hao N, O'shea EK. Signal-dependent dynamics of transcription factor translocation controls gene expression. Nature structural & molecular biology. 2012 Jan; 19(1):31. https://doi.org/10.1038/nsmb.2192 PMID: 22179789; PMCID: PMC3936303.

12. Hansen AS, O'shea EK. Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression. Molecular systems biology. 2013; 9(1):704. https://doi.org/10.1038/msb.2013.56 PMID: 24189399; PMCID: PMC4039373.

13. Mukherjee SP, Behar M, Birnbaum HA, Hoffmann A, Wright PE, Ghosh G. Analysis of the RelA: CBP/p300 interaction reveals its involvement in NF-κB-driven transcription. PLoS Biol. 2013 Sep 3; 11(9):e1001647. Epub 2013 Sep 3. https://doi.org/10.1371/journal.pbio.1001647 PMID: 24019758; PMCID: PMC3760798.

14. Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. Proceedings of the National Academy of Sciences. 2003 Apr 29; 100(9):5136–41. Epub 2003 Apr 17. https://doi.org/10.1073/pnas.0930314100 PMID: 12702751; PMCID: PMC404558.

15. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: applications. Current opinion in genetics & development. 2005 Apr 1; 15(2):125–35. https://doi.org/10.1016/j.gde.2005.02.006 PMID: 15797195; PMCID: PMC3462814.

16. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: models. Current opinion in genetics & development. 2005 Apr 1; 15(2):116–24. https://doi.org/10.1016/j.gde.2005.02.007 PMID: 15797194; PMCID: PMC3482385.

17. Li Z, Bianco S, Zhang Z, Tang C. Generic properties of random gene regulatory networks. Quantitative Biology. 2013 Dec; 1(4):253–60. https://doi.org/10.1007/s40484-014-0026-6 PMID: 25328770; PMCID: PMC4198180.

18. Cheng CS, Behar MS, Suryawanshi GW, Feldman KE, Spreafico R, Hoffmann A. Iterative modeling reveals evidence of sequential transcriptional control mechanisms. Cell systems. 2017 Mar 22; 4 (3):330–43. Epub 2017 Feb 22. https://doi.org/10.1016/j.cels.2017.01.012 PMID: 28237795; PMCID: PMC5434763.

19. Tong AJ, Liu X, Thomas BJ, Lissner MM, Baker MR, Senagolage MD, et al. A stringent systems approach uncovers gene-specific mechanisms regulating inflammation. Cell. 2016 Mar 24; 165(1):165–79. Epub 2016 Feb 25. https://doi.org/10.1016/j.cell.2016.01.020 PMID: 26924576; PMCID: PMC4808443.

20. Cheng Q, Behzadi F, Sen S, Ohta S, Spreafico R, Teles R, et al. Sequential conditioning-stimulation reveals distinct gene-and stimulus-specific effects of Type I and II IFN on human macrophage functions. Scientific reports. 2019 Mar 27; 9(1):1–4. https://doi.org/10.1038/s41598-018-37186-2 PMID: 30626917; PMCID: PMC6437173.

21. Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, et al. A validated regulatory network for Th17 cell specification. Cell. 2012 Oct 12; 151(2):289–303. Epub 2012 Sep 25. https://doi.org/10.1016/j.cell.2012.09.016 PMID: 23021777; PMCID: PMC3503487.

22. Johnson JS, De Veaux N, Rives AW, Lahaye X, Lucas SY, Perot BP, et al. A comprehensive map of the monocyte-derived dendritic cell transcriptional network engaged upon innate sensing of HIV. Cell reports. 2020 Jan 21; 30(3):914–31. https://doi.org/10.1016/j.celrep.2019.12.054 PMID: 31968263; PMCID: PMC7039998.

23. Alon U. An introduction to systems biology: design principles of biological circuits. CRC press; 2019 Jul 12. https://doi.org/10.1201/9780429283321

24. Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al. Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. Molecular cell. 2010 Feb 12; 37(3):418–28. https://doi.org/10.1016/j.molcel.2010.01.016 PMID: 20159560.

25. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nature biotechnology. 2020 Jan; 38(1):56–65. Epub 2019 Dec 2. https://doi.org/10.1038/s41587-019-0315-8 PMID: 31792407; PMCID: PMC6954276

26. Ma W, Trusina A, El-Samad H, Lim WA, Tang C. Defining network topologies that can achieve biochemical adaptation. Cell. 2009 Aug 21; 138(4):760–73. https://doi.org/10.1016/j.cell.2009.06.013 PMID: 19703401; PMCID: PMC3068210.

27. Apgar JF, Toettcher JE, Endy D, White FM, Tidor B. Stimulus design for model selection and validation in cell signaling. PLoS Comput Biol. 2008 Feb 15; 4(2):e30. https://doi.org/10.1371/journal.pcbi.0040030 PMID: 18282085; PMCID: PMC2323406.

28. Molinelli EJ, Korkut A, Wang W, Miller ML, Gauthier NP, Jing X, et al. Perturbation biology: inferring signaling networks in cellular systems. PLoS Comput Biol. 2013 Dec 19; 9(12):e1003290. Epub 2013 Dec 19. https://doi.org/10.1371/journal.pcbi.1003290 PMID: 24367245; PMCID: PMC3868523.

29. Gross T, Blüthgen N. Identifiability and experimental design in perturbation studies. Bioinformatics. 2020 Jul 1; 36(Supplement_1):i482–9. https://doi.org/10.1093/bioinformatics/btaa404 PMID: 32657359; PMCID: PMC7355299.

30. Letsou W, Cai L. Noncommutative biology: sequential regulation of complex networks. PLoS Comput Biol. 2016 Aug 25; 12(8):e1005089. https://doi.org/10.1371/journal.pcbi.1005089 PMID: 27560383; PMCID: PMC4999240.

31. Scholes C, DePace AH, Sánchez Á. Combinatorial gene regulation through kinetic control of the transcription cycle. Cell systems. 2017 Jan 25; 4(1):97–108. Epub 2016 Dec 29. https://doi.org/10.1016/j.cels.2016.11.012 PMID: 28041762; PMCID: PMC5469051.

32. Stelling J, Sauer U, Szallasi Z, Doyle FJ III, Doyle J. Robustness of cellular functions. Cell. 2004 Sep 17; 118(6):675–85. https://doi.org/10.1016/j.cell.2004.09.008 PMID: 15369668

33. Ourthiague DR, Birnbaum H, Ortenlöf N, Vargas JD, Wollman R, Hoffmann A. Limited specificity of IRF3 and ISGF3 in the transcriptional innate-immune response to double-stranded RNA. Journal of leukocyte biology. 2015 Jul; 98(1):119–28. Epub 2015 Apr 20. https://doi.org/10.1189/jlb.4A1014-483RR PMID: 25896227; PMCID: PMC4467169.

34. Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, et al. Cooperative binding of transcription factors orchestrates reprogramming. Cell. 2017 Jan 26; 168(3):442–59. Epub 2017 Jan 19. https://doi.org/10.1016/j.cell.2016.12.016 PMID: 28111071; PMCID: PMC5302508.

35. Gomes AM, Kurochkin I, Chang B, Daniel M, Law K, Satija N, et al. Cooperative transcription factor induction mediates hemogenic reprogramming. Cell reports. 2018 Dec 4; 25(10):2821–35. https://doi.org/10.1016/j.celrep.2018.11.032 PMID: 30517869; PMCID: PMC6571141.

36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014 Dec; 15(12):1–21. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281; PMCID: PMC4302049.

37. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology. 2014 Feb; 15(2):1–7. https://doi.org/10.1186/gb-2014-15-2-r29 PMID: 24485249; PMCID: PMC4053721.

38. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. Cell systems. 2019 Oct 23; 9(4):401. Epub 2018 Aug 29. https://doi.org/10.1016/j.cels.2019.08.003 PMID: 31647917; PMCID: PMC6167088.

39. Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012 Sep 7.

40. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PloS one. 2013 Sep 30; 8(9):e74335. https://doi.org/10.1371/journal.pone.0074335 PMID: 24098642; PMCID: PMC3787051.

41. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1; 26(1):139–40. Epub 2009 Nov 11. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308; PMCID: PMC2796818.

42. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011 Jan; 29(1):24–6. https://doi.org/10.1038/nbt.1754 PMID: 21221095; PMCID: PMC3346182.