8-3-2023

Biophysically-informed Models of Heterogeneity in Transcription Factor Activity

Kaiser Loell
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Barak Cohen, Chair
Gary Stormo, Co-Chair
Shankar Mukherji
Mikhail Tikhonov
Tony Tsai

Biophysically-informed Models of Heterogeneity in Transcription Factor Activity

by

Kaiser Loell

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my PhD advisor, Barak Cohen. It is safe to say that he has not disappointed as a mentor, never failing to dig deeper and give constructive feedback at every turn.

I would also like to thank all of the members of the Cohen lab. Everything just said about Barak could equally apply to every other lab member. Of the current and former lab members, I would especially like to thank Max Staller, for getting me started on the yeast project that led to the first half of this thesis, Yawei Wu, for helping out with molecular cloning during the height of the COVID pandemic when it was still difficult to go into lab, Mike White, for being the driving force behind all of the projects related to CRX and the retina in the lab, and Ryan Friedman, for designing and cloning the second CRX library and for helping to devise the toy model that resulted from it.

My thesis committee members, Gary Stormo, Shankar Mukherji, Mikhail Tikhonov, and Tony Tsai, have never ceased to be tremendously insightful in my discussions with them, even from before they were on my committee.

Our collaborators in Joe Corbo's lab have been indispensable for all of our retina-related projects. Connie Myers deserves special thanks for being the one to do all of the hands-on mouse and tissue culture work.

Finally, I'd also like to thank Liang-Bo "Bobo" Wang, for creating the laTex template used to format this thesis.

This work was supported by a T32 training grant from the National Human Genome Research Institute, and ongoing grants from the National Institute of General Medical Sciences (GM140711, GM092910, GM092910).

Kaiser Loell

Washington University in St. Louis

August 2023

ABSTRACT OF THE DISSERTATION

Biophysically-informed Models of Heterogeneity in Transcription Factor Activity

by

Kaiser Loell

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2023

Professor Barak Cohen, Chair

Professor Gary Stormo, Co-Chair

The activity of a transcription factor (TF) can be remarkably heterogeneous, both within and between cells. Between cells, molecular fluctuations drive stochastic differences in the activities of TFs. I will describe a project in which we have quantified how the different components of a TF (its activation domain, affinity for DNA, concentration, and cooperative interactions) influence stochastic fluctuations in TF activity. In addition to stochastic fluctuations between cells, within cells TFs can switch between activating and repressing activities depending on the cis-regulatory content of their target genes. I will describe a second project in which we have modelled how the balance between homo- and heterotypic interactions determines whether a TF will function as an activator or repressor. Understanding and quantitatively modeling the factors underlying TF heterogeneity will be necessary if we are to predict patterns of gene expression from cis-regulatory DNA. Our work suggests that biophysically-informed models will be an important part of capturing the context-specific activities of TFs.

Chapter 1: Introduction

1.1 Transcriptional regulation

The process of transcriptional regulation is fundamental to biology. By enabling cells to express a select subset of the genes encoded in their genomes, it allows a multiplicity of cellular functions and phenotypes to arise from the same genotype. This makes it especially important in multicellular organisms, where it enables a multitude of cell types to be encoded by the same genome. As a consequence, dynamic regulation of gene expression is essential for development. Because of the central nature of gene regulation to development, changes in regulation are the primary drivers of phenotypic change over evolutionary time [1, 2]. Within present-day populations, variants in these regulatory sequences are the major drivers of common traits [3, 4] and variation in gene expression modifies the penetrance of less common coding variants [5, 6]. Understanding the sequence determinants of transcriptional regulation will therefore be essential to understand and predict phenotypic variation and disease risk.

Transcription factors (TFs) are the central players in transcriptional regulation. Transcribing every gene at an appropriate level while avoiding aberrant transcriptional initiation from intra- and intergenic regions requires sequence-specific regulatory mechanisms. TFs provide such a mechanism, with specificity coming from their DNA-binding domains (DBDs), which selectively bind preferred motifs [7, 8]. Sequence variants that alter the degree of match to TF motifs can therefore affect gene expression, leading to striking phenotypes in some cases [9, 10]. TFs also provide cell type and state specificity to regulation, with cell types being defined by combinations of TFs [11–15]. Because many TFs are inducible by signaling pathways [16], TFs are also responsible for gene expression's dynamic, environmentally responsive nature. While chromatin accessibility,

methylation, and histone modifications are frequently invoked as regulatory mechanisms underlying gene expression states, the proteins that deposit, remove, and bind these marks ("epigenetic writers", "erasers", and "readers") lack sequence specificity [17]. Therefore, in order for their effects to be gene-specific, these factors must first be recruited to specific sequences by TFs.

1.2 Transcriptional factor function

TFs recruit chromatin modifiers and other cofactors like Mediator through effector domains. Effector domains can function as activation domains (ADs) or repression domains (RDs) that, as their names indicate, activate or repress transcription, respectively. Because of this, effector domain sequence is a key determinant of whether a TF will activate or repress transcription, and how strongly it will do so. Intriguingly, many TFs contain both activation and repression domains [18]. In fact, many effector domains can act as both activators and repressors depending on the promoter to which they are recruited. Such bifunctionality is frequently observed within the context of endogenous TFs, with many TFs appearing to switch between activation and repression depending on the number and arrangement of their binding sites [19]. This switching can even occur within a single cis-regulatory context, with bifunctional effector domains having activating and repressive effects at different timepoints after recruitment to the same promoter [18]. Furthermore, switching can occur at a single-cell level, with some effector domains being able to split pools of cells into subpopulations with the target gene either active or repressed. Switching is important in signal transduction, with signal-responsive TFs tending to switch from repression in the absence of signal to activation in the presence of signal [16]. These striking observations motivate us to ask what features govern this heterogeneity in the activity of single transcription factors, and what mechanisms cause their transcriptional effects to vary.

1.3 Synthetic reporter assays

Synthetic reporter assays provide an ideal platform for addressing these questions. Synthetic promoter constructs can be built out of known TF binding motifs, allowing us to precisely control the identity of the TFs driving a reporter gene and the spatial configuration in which they bind. Doing so has allowed us to identify a switch between activation and repression in the retinal TF CRX [19]. Furthermore, it is possible to construct synthetic TFs out of effector and DNA-binding domains, and put them under drug-inducible control. By targeting these synthetic TFs to a reporter locus, it is possible to isolate the effects of effector domain sequence on expression from those of DNA binding and TF abundance in the nucleus. This has allowed us and other groups to identify sequence features that contribute to effector domain function [18, 20]. Using these approaches, we can experimentally perturb each of the regulatory features just mentioned and measure their effects on both bulk reporter expression and cell-to-cell variability. However, experimental perturbations alone are not sufficient to understand the mechanisms underlying such heterogeneity in activity: while we have had since 2016 activity measurements of synthetic reporters composed solely of binding sites for CRX and NRL, a cooperating TF [19, 21, 22], we had not been able to fit a predictive model to them prior to the work presented here [Barak Cohen, personal communication].

1.4 Biophysical and machine learning models

Gaining a predictive understanding of the mechanisms underlying regulation requires an appropriate model. The models in use in the field can be broadly classified into biophysical and machine learning models, each of which have unique strengths and weaknesses. Biophysical models, because they are based on specific hypotheses about the mechanisms underlying the phenomena being modeled, are generally highly interpretable, but require prior mechanistic knowledge to be constructed and fit [23, 24]. In contrast, machine learning models, in particular the deep neural nets that have been the focus of much attention recently, can capture complex sequence features in a more unbiased

fashion and frequently achieve high predictive power. However, they are often considered "black boxes" that lack biological interpretability [23, 25, 26]. A related issue afflicting many widespread modeling paradigms is that they do not account for features of the experimental measurement or for experimental or biological noise. This can lead to entanglement between experimental and biological effects and is of particular concern given that it has been shown that models which do not correctly account for experimental noise will generally make incorrect inferences [27]. It is therefore desirable when analyzing biological data to use an approach to modeling that accounts for noise and whose parameters have biophysical interpretations.

1.5 Scope of thesis

In this thesis, I will present two projects combining biophysically informed modeling and synthetic reporter experiments to address questions related to heterogeneity in TF action. In chapters 2 and 3, I use a yeast reporter system driven by synthetic TFs to characterize the effects of different regulatory features on cell-to-cell variability in gene expression. In chapter 4, I use a biophysically-informed machine learning approach to reanalyze accumulated data from high-throughput synthetic reporter assays in the mouse retina, and use the learned parameters to develop a model explaining how CRX switches between activation and repression depending on cis-regulatory context.

Chapter 2: Activation domains can decouple the mean and noise of gene expression

2.1 Introduction

Gene expression is an inherently stochastic process, producing levels of protein and mRNA that fluctuate between genetically identical cells [28–31]. This stochasticity, or noise, in gene expression is unavoidable due to the randomness of molecular motions. Noise in gene expression interferes with information transmission in regulatory networks [32, 33], but also facilitates bistable switching [34–36] and entrainment in systems such as circadian clocks [37–40]. Noise is a fundamental property of gene expression with phenotypic consequences that have been observed across all domains of life and scales of organization [41, 42].

Stochastic noise in gene expression tends to occur in bursts during which many mRNAs are transcribed within a short period, interspersed within longer silent intervals [43–49]. This bursty gene expression has been observed in bacteria [50], yeast [51], and mammalian cells [45]. One major consequence of bursty gene expression noise is its effect on cell fate decisions [52–55], which are frequently determined by stochastic fluctuations in the levels of transcriptional regulators [56–66]. Stochasticity in cell fate decisions represents a "bet hedging" strategy that keeps cellular phenotypes diverse even in the absence of genetic or environmental variation [50, 67–75].

Efforts to understand the molecular causes of expression noise have shown that both cis [66, 76–78] and trans factors [79–82] influence noise in gene expression. Transcription factors (TFs), which constitute the interface between cis-regulatory elements (CREs) and trans-acting cofactors, are a known source of expression noise [83]. TFs are composed of DNA binding domains, which confer specificity for their target CREs, and activation domains (ADs), which recruit trans-acting

cofactors that alter transcription once bound to DNA [84, 85]. Fluctuations in TF binding are a major determinant of gene expression noise [83, 86, 87], but the effects of TF ADs on expression noise have not been determined. Because ADs vary widely in the cofactors they recruit and the contexts in which they are active [88–91], we investigated whether ADs also vary in their effects on noise in gene expression.

Two hypotheses might explain how ADs influence noise in gene expression. One hypothesis is that the noise generated by an AD depends solely on its effect on mean levels of expression. Noise is tightly coupled to mean output levels in most stochastic processes [92]. Thus, increasing a gene's mean expression necessarily increases its expression noise. This hypothesis is consistent with the increase in noise produced by VP64 relative to VP16 [83]. A strong prediction of this hypothesis is that diverse ADs will all generate the same amount of noise for a given mean level of expression. An alternate hypothesis is that an AD's effect on noise depends both on its effect on mean expression and on the specific cofactors that underpin its activity. Different ADs might produce different amounts of noise, even at the same mean output level, because of the distinct biochemical properties of the cofactors they recruit. For example, it has been suggested [87] that recruitment of chromatin remodelers is a key step in generating gene expression variation. If this is the case, TFs with activation domains that recruit chromatin remodelers will have distinctly different effects on noise from those that cannot. This hypothesis is supported by Tan et al's observations of the effects on noise of mutations in the p65 AD that ablate binding to CBP/p300 [93]. However, because only a very limited number of ADs have been tested, it remains unresolved which hypothesis, if either, applies more generally to ADs.

2.2 Results

2.2.1 Comparison of the VP16 and Gcn4 ADs demonstrates that it is possible to decouple expression noise from the mean

We first compared the effects of a strong AD (VP16) and a weak AD (Gcn4) on expression mean and noise by performing dose-response experiments. Varying the levels of induction for each synthetic TF led to Fano factors increasing with mean reporter protein expression. Strikingly, VP16 produced more gene expression noise than Gcn4 for every mean level of reporter expression tested (Figure 2.1B). This result indicates that different ADs can create different amounts of noise in gene expression, even when the mean expression level is the same.

Because the VP16 AD is stronger than the Gcn4 AD, the Gcn4 construct always required higher beta-estradiol concentrations than VP16 to achieve the same mean reporter expression level. This observation led us to a working hypothesis that high expression noise is produced by strong ADs present at low nuclear concentrations, whereas low expression noise is achieved by weak ADs present at high nuclear concentrations.

2.2.2 Comparisons of multiple activation domains reveal that expression noise can be explained by AD strength

Our working hypothesis assumes that the noise generated by an AD is determined solely by its strength and nuclear concentration. Alternatively, different ADs might generate different amounts of noise based on the distinct cofactors they recruit. These two possibilities can be distinguished by measuring the mean and noise induced by a diverse set of ADs at the same beta-estradiol concentration. If the differences in noise were due only to differences in the strengths and nuclear concentrations of ADs, then, for a fixed level of induction, the noise generated by ADs would be predictable from the mean levels of reporter gene expression they produce. Alternatively, if certain ADs had special biochemical properties, then such ADs would deviate from the mean versus noise

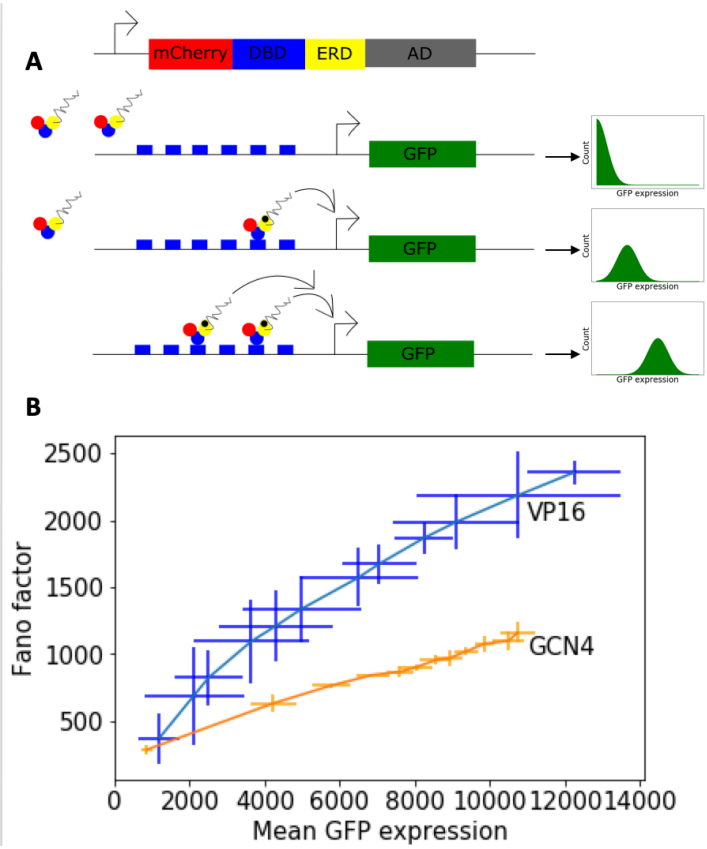

Figure 2.1: Comparison of the VP16 and Gcn4 ADs (A) Reporter system to measure the mean and noise produced by diverse ADs. Synthetic TF constructs are composed of an mCherry tag (red), a fixed zinc-finger DBD (blue), an estrogen response domain (yellow), and a variable AD (grey). Synthetic TFs induce GFP expression (green) from a reporter gene driven by an array of zinc-finger binding sites. Varying levels of beta-estradiol (black) controls the nuclear localization of the synthetic TFs, allowing precise control of nuclear concentration while controlling for AD. (B) Activation domains differ in the amount of noise they induce, even at comparable means. The noise (Fano factor) produced by synthetic TFs carrying either the Gcn4 or VP16 AD is plotted for varying mean levels of reporter gene expression. The trend lines connect the data points and are not model fits to the data. Error bars are standard deviations over three replicates.

trendline.

To test these predictions, we assayed eleven different yeast ADs and a negative control at a fixed beta-estradiol concentration. This experiment revealed a linear relationship between the mean and noise of expression for all ADs tested (Pearson R2=0.90) (Figure 2.2A). We observed similar linear trends at different fixed beta-estradiol concentrations (supplemental Figure A.3j). To confirm that the AD does not affect the TF construct's localization, we imaged the subcellular localization of the TF construct using the mCherry tag. The ratio of nuclear to total mCherry fluorescence was distributed similarly across strains (supplemental Figure A.4), indicating that the AD does not affect the TF's nuclear concentration. In the Saccharomyces Genome Database (https://www.yeastgenome.org/), most of the TFs from which these ADs derive interact with unique cofactors (examples in Figure A.5), suggesting that they activate transcription through different mechanisms. Thus, at a fixed nuclear concentration, diverse ADs follow a predictable trend relating the mean and noise of expression they produce.

We next asked whether mutants that interfere with the mechanism of action of ADs alter the fixed relationship between the mean and noise produced by ADs. We assayed, at a fixed beta-estradiol concentration, the activities of 84 mutants in the Gcn4 AD from Staller et al. This set was chosen to cover a broad range of mutation types and activities. Although the mutants produced a wide range of mean reporter gene activities, we still observed a strong linear relationship between the mean and noise of expression across all mutants (Pearson R2=0.91) (Figure 2.2B).

Taken together, these results demonstrate that it is difficult to uncouple the mean and noise of expression at a fixed nuclear concentration of AD. At a fixed nuclear concentration, diverse ADs and a diverse set of AD mutants both exhibit a tight coupling between mean and noise. These observations support our working hypothesis that noise in gene expression is controlled by the interplay between the strength and nuclear concentration of ADs, and not by the specific mechanism of action of ADs.

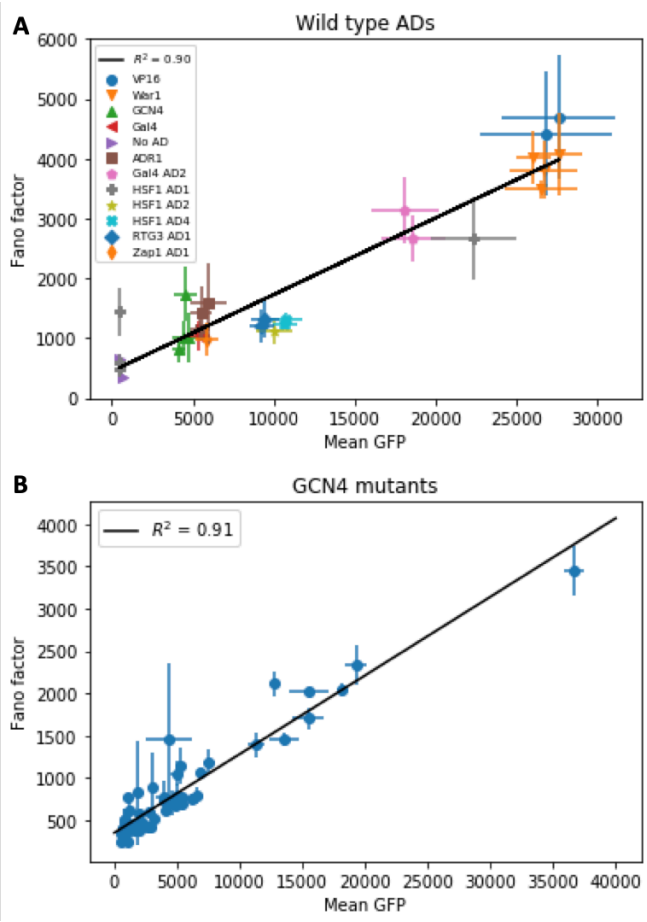

Figure 2.2: Comparisons of multiple activation domains (A) Linear relationship between activation domain strength and noise for diverse ADs. The noise produced by diverse ADs is plotted versus the mean level of reporter gene expression for each AD. Error bars are standard deviations of mean and Fano across three replicates. (B) Linear relationship between activation domain strength and noise for mutants in the Gcn4 AD. The noise produced by each Gcn4 AD mutant is plotted versus its mean reporter gene expression. Error bars are standard deviations over three replicates.

2.2.3 Simulations provide a mechanistic interpretation of the results

We asked what quantitative models of gene expression noise could explain the above observations, and whether we could link perturbations of specific kinetic parameters of those models to the observed effects of perturbing AD strength or nuclear concentration.

Assuming four different models of gene expression kinetics, we performed simulations in which we systematically perturbed kinetic parameters alone or in combination. We then searched for perturbations that could recapitulate the experimental effects of changing either the AD strength or nuclear concentration.

The consensus model of gene expression noise is the random telegraph model [30, 94, 95] (Figure 2.3A), in which it is assumed that DNA transitions between an active and inactive state, and that stochastic bursts of expression initiate only from the active state. This model has been heavily studied and is widely used to model noise in gene expression (36, 37). However, investigators have more recently proposed a number of more complex models incorporating multiple distinct inactive states. Scholes et al [96] describe a model in which a promoter must cycle through two states to produce each mRNA (Figure 2.3B). Another elaboration on the random telegraph model is that of Zoller et al [97], (Fig 3C), in which promoters cycle through multiple refractory states between bursts. Finally, Rodriguez et al. [98] have proposed a multi-state model (Fig 3D.), in which, over long timescales, there is alternation between a state that is permissive for bursting and one that is not. We simulated each of these four models to identify regimes that could reproduce our experimental observations.

To identify perturbations that could reproduce the effects of varying the nuclear concentrations of ADs, we aimed to identify single parameters or linearly related pairs of parameters that, when varied, reproduced our experimental observations. We performed simulations in which we varied one or two parameters in the model through its physiological range while holding the other parameters in the model at fixed values. We performed these simulations many times, each time holding the other parameters in the models at different fixed values. We found no individual parameter that

when varied reproduced our experimental observation that the Fano factor-mean relationship is nonlinear with a decreasing slope. However, linear perturbations to two parameters simultaneously could recapitulate the experimental effects of increasing nuclear concentration (Supplemental Figure A.6). More specifically, we found that, assuming the random telegraph model, increasing Kon (the frequency of transitioning to the active state) while simultaneously increasing Km (the transcription rate while in the active state) (Figure 2.3E) or decreasing Dm (the mRNA degradation rate) (Figure 2.3F) produced trendlines that resembled those associated with increasing nuclear concentration of TF.

All other models we tested either could not explain our observations or provided no additional explanatory value. The mean-noise relationships predicted by the transcription cycle model (Figure 2.3B) sharply contrast with those we observe. Altering either parameter of this model is predicted to cause Fano factor to fall then rise with mean (Figure A.6g-i). Simulations performed across a range of combinations of parameters predicted that noise would be at a minimum when the two rates are equal, with noise increasing with increasing disparity between the parameters. The refractory period model's (Figure 2.3C) predictions also conflict with our observations. This model predicts two qualitatively distinguishable mean-Fano factor relationships, with changes in the rates of steps leading out of the active state causing Fano factor to change linearly with mean (Figure A.6o-p), and changes in the remaining steps having no effect on noise (Figure A.6 j-n).

The effects of varying the AD strength were more straightforward to recreate in models. Again, assuming the random telegraph model, linear increases in Km, with or without concurrent changes in Dm, produced linear mean-Fano relationships matching those we observed when AD strength is varied (Figure 2.3G, 3H). Sherman et al's [94] analytical solution for the moments of the random telegraph model's protein distribution also predicted these mean-Fano relationships (Figure A.6q-s). Varying Km while holding Kon or Koff fixed at different values produced different slopes but otherwise preserved the linear mean-fano relationship produced by ADs of increasing strength, whereas varying Kon or Koff could not reproduce this trend regardless of the value of Km (Supplemental

Figure A.6).

The multi-state model (Figure 2.3D), while able to reproduce our experimental results, did not fit our observations better than the less complex random telegraph model. It predicted mean-variance relationships similar to those of the random telegraph model, with modulating the rates of transition into and out of the long-lived repressed state predicted to have effects resembling those of modulating Koff (Figure A.6t) and Kon (Figure A.6u) in the random telegraph model, respectively, and changes in the rates of the other steps having effects (Figure A.6v-x) similar to those of changes in Km. We conclude that the random telegraph model is the simplest model that captures the effects of both varying the strength and nuclear concentration of ADs. Additionally, we conclude that the nuclear concentration of a TF influences Kon, while AD strength primarily acts through Km. This can further be interpreted in terms of transcriptional burst parameters. Kon in the random telegraph model is effectively equivalent to burst frequency, while burst size is proportional to the ratio of Km to Koff. We therefore conclude that AD strength primarily controls burst size, while the nuclear concentration of an AD affects burst frequency.

2.3 Discussion

Our results demonstrate that the mean and noise of gene expression can be separated by varying the strength and nuclear concentration of ADs. By varying the induction levels of different ADs, we observed noise levels spanning a roughly 2-fold range at comparable means. These results raise the possibility of synthetically manipulating expression mean and noise independently by using ADs of varying strength while tuning TF occupancy on DNA to compensate. Doing so would have many applications in synthetic biology, allowing perturbation experiments aimed at determining the effects of noise on gene expression networks, development, and cellular physiology, and opening the way for engineering stochastic fate decisions during cellular reprogramming. Likewise, natural selection could tune the noisiness of a gene's expression, independent of its mean levels, by operating on variation that affects the strength and nuclear concentrations of ADs.

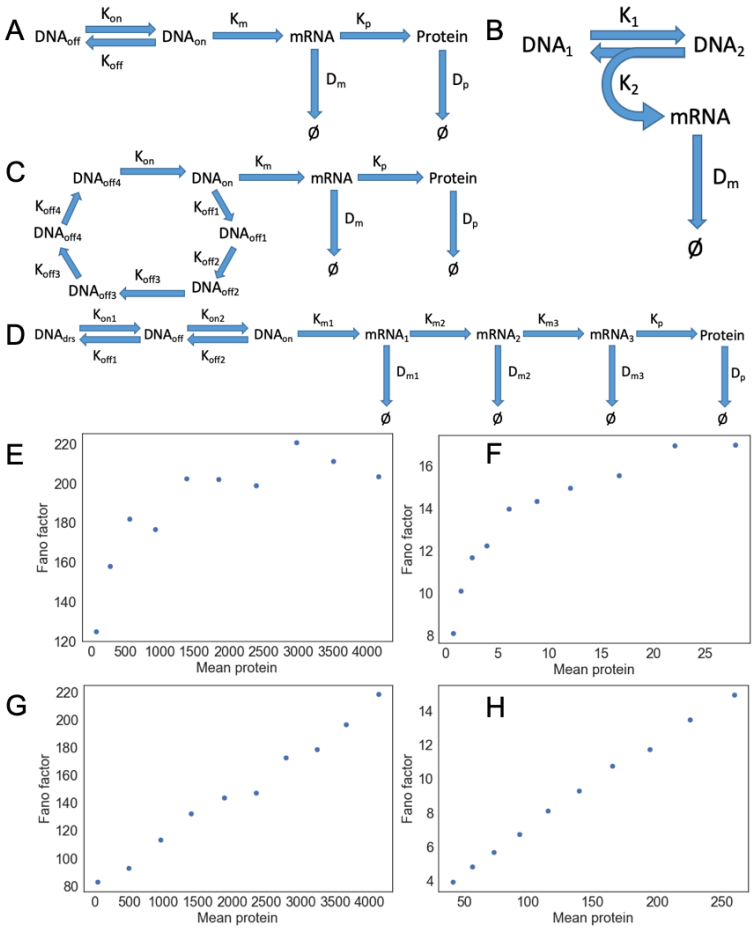

Figure 2.3: Stochastic models of gene expression. Cartoon diagrams of (A) the Random Telegraph Model, (B) the Transcription Cycle model of Scholes et al, (C) The Refractory Model of Zoller et al, and (D) the Multi-State Model of Rodriguez et al. The mean (X axis) and Fano factor (Y axis) of protein expression predicted by the Random Telegraph Model are plotted for simulations in which (E) Kon and Km are varied simultaneously, (F) Kon and Dm are varied simultaneously, (G) Km is varied, or (H) Km and Dm are varied simultaneously.

Although manipulating ADs allowed us to decouple the mean and noise of expression, we did not find evidence that ADs with different mechanisms of action produced different amounts of noise. At any fixed nuclear concentration of AD, the noise produced by an AD was largely predictable from its effects on mean expression. Manipulating the class of AD did not alter the relationship between mean and noise.

Our results can be explained in terms of the effects of AD strength and TF concentration in the nucleus. TF binding to DNA occurs for only brief periods [99–102] leading to intermittent transcriptional activation. At low TF concentrations there is low TF occupancy on the DNA. Coupled with a strong AD, this regime leads to large but infrequent bursts of expression, which produces high noise. Conversely, at high TF concentrations there is much higher promoter occupancy, leading to more frequent bursts approaching a continuous rate of mRNA production, and therefore relatively low noise. Maintaining the same mean level of expression in this regime requires lowering the strength of the AD to compensate for more frequent bursting. This model is consistent with Pelet et al's [87] observation that transient nuclear localization of the TF Hog1 leads to bimodal expression and greater noise than is induced by sustained induction.

Our working hypothesis that both nuclear concentration and AD strength determine noise can be formulated quantitatively in terms of the random telegraph model of gene expression kinetics. The components of the random telegraph model are both necessary and sufficient to explain our results: models such as Zoller et al's and Scholes et al's that cannot be reduced to the random telegraph model are unable to produce the mean-fano relationships we observe experimentally, and those such as Rodriguez et al's that add additional steps to the random telegraph model do not provide additional explanatory power. Our results thus favor models without cycles or refractory states as the more parsimonious explanation. This result is consistent with Zoller et al's finding that synthetic promoters have only one inactive state.

Based on the simulation and experimental results, we posit that AD strength primarily affects Km, the rate of transcription from the promoter in its active state, while the abundance of TF in the

nucleus determines Kon, the rate at which the promoter switches into the transcriptionally active state. Under these assumptions, the random telegraph model predicts mean-Fano factor relationships resembling those we observed experimentally. This is largely consistent with Senecal et al's results and their model suggesting that altering AD strength primarily affects Km, while modulating TF level primarily affects Kon ("burst frequency") and TF lifetime on the promoter affects Koff ("burst duration"). However, our results require an additional effect of TF level on Km or Dm that was not seen by Senecal et al. The next chapter will investigate the role of DBD affinity for sequence motifs, cooperative TF binding, and other cis-factors in TF regulation of transcriptional noise.

2.3.1 Limitations

Synthetic TFs were used throughout this study. Their use is necessary to control for DNA binding and isolate the effects of AD sequences and nuclear concentration, but could in theory alter the properties of the ADs from their physiological context. Additionally, we performed all experiments in yeast because it is a highly tractable model system. Because of this, the generalizability of our results to mammalian contexts depends on the conservation of transcriptional regulatory mechanisms across eukaryotes.

2.4 Experimental model and subject details

2.4.1 A reporter system that measures the effects of different ADs on expression noise

To compare the effects of diverse ADs on expression noise, we set up a reporter system that allowed us to measure the mean and noise of expression generated by TFs that differ only in the AD they carry. We repurposed the reporter system described in Staller et al. [20] (Figure 2.1A), which uses synthetic S. cerevisiae TFs consisting of an AD of interest, a fluorescent tag (mCherry), an estrogen response domain (ERD) and a fixed synthetic zinc-finger DNA binding domain (DBD)

[103]. Using a synthetic DBD avoids interference with endogenous yeast TFs, and the ERD allows us to control the nuclear localization of the synthetic TFs with beta-estradiol. Swapping the ADs on these synthetic TFs allowed us to directly compare the effects of different ADs while keeping the rest of the TF constant and avoiding competition with endogenous TFs. The synthetic TF is integrated at a single allele of the URA3 locus.

We measured the activities of these synthetic TFs by reading out the fluorescence of a GFP reporter construct via flow cytometry. The reporter construct contains a tandem array of zinc-finger binding sites and is integrated at the YBR032w locus. To probe the impact of varying TF stimulation on target expression, we performed dose-response experiments with increasing amounts of beta-estradiol and measured the resulting distribution of GFP by flow cytometry. We did not observe any effect on cell growth from the beta-estradiol, either in the side and forward scatter measurements (Figure A.1a), or in the growth rates (Figure A.2). We then computed the mean (x) and Fano factor (σ^2/x) of expression for each TF at each beta-estradiol concentration. We used the Fano factor as the measure of noise because it normalizes for different mean levels of expression [104–107].

A gene's expression noise can be divided into an intrinsic and extrinsic component [108]. We focused on the effect of ADs on the intrinsic component because current models of stochastic gene expression best capture intrinsic noise. We therefore attempted to exclude extrinsic noise from our measurements.

The presence of significant extrinsic noise in our data was indicated by a strong correlation between the forward scatter, side scatter, GFP, and mCherry fluorescence in the raw data. However, after controlling for cell size by gating on forward scatter, there was little correlation between the expression of the TF construct as measured by mCherry signal and GFP expression (Figure A.1b). We speculate that there is no residual correlation between mCherry and GFP because of the time lag between the production of mCherry-tagged TFs and the expression and maturation of the GFP reporter, or because the mCherry signal reflects the TF concentration throughout the cell and is not specific for the nucleus. The observation that gating on forward scatter removes the correlation

between TF and reporter gene fluctuations indicates that fluctuations in cell size are the major source of extrinsic noise in our data.

To screen out this source of extrinsic noise, we filtered our data by gating on cells expressing similar levels of mCherry, which removed 56% of the total variance in GFP expression. We chose mCherry as the marker to filter by to exclude a small (< 1%) outlier subpopulation of events with similar scatter to most cells but very high mCherry and low GFP. While filtering on mCherry does not exclude all extrinsic noise related to forward scatter, additional filtering on forward scatter leaves few cells, which lowered reproducibility without affecting the overall shape of the mean-fano relationships. Using this system we compared the effects of diverse ADs on both the mean and noise of expression.

2.4.2 Yeast strains

We repurposed the yeast strains created by Staller et al [20]. These strains (MATa/MATalpha, synthetic TF-Kanr::ura3/URA3 GFP reporter-natR::YBR032w/YBR032w) are generated from crosses between derivatives of FY4 (MATa, synthetic TF-Kanr::ura3) and FY5 (MATalpha, GFP reporter-natR::YBR032w). Each synthetic TF construct consists of an mCherry tag, a murine Zif268 DNA binding domain (DBD), a human estrogen response domain (ERD), and an activation domain (AD), driven by the yeast ACT1 promoter. The activation domains were either sampled from the library of GCN4 mutants described in Staller et al by picking clones at random, or derived from wild type ADs (Table A2). Sequence files of the all wild-type synthetic TF-AD fusions are available in Supplementary folder 1. The GFP reporter is the same as reported in Staller et al. and consists of a fast maturing GFP variant, with expression driven by six upstream Zif268 binding sites and the P3 promoter.

2.4.3 Culture conditions

All strains were obtained from frozen glycerol stocks. For the experiments involving wild-type activation domains, cultures were seeded by pipetting 2 uL of each strain onto a YPD agar plate, streaking out, and growing for 2 days in a 30 degree incubator. For each experiment, individual colonies were then picked from the plates and grown out overnight in tubes containing 3 mL of SC dextrose medium. These tubes were constantly rotated on a wheel in a 30 degree incubator throughout the growth process. The optical densities of 1:10 dilutions of the overnight cultures were then measured, and all cultures were diluted to an OD of 0.0225. For each combination of strain and level of induction, 200 uL of diluted culture was added to a well of a deep 96 well plate. 50 uL of diluted beta-estradiol was then added to the well (final concentrations are listed in Table A1). The diluted cultures were incubated with the beta-estradiol in a shaker incubator set to 300 rpm for 5 hours before measuring.

For the experiments involving mutant activation domains, the initial cultures were seeded by pinning from the 96 well plate containing the glycerol stocks onto a YPD agar plate. Liquid cultures were then seeded by pinning from the agar cultures into a deep well plate with 300 uL of SC in each well. The deep well plate was agitated in a shaker incubator overnight.

2.5 Method details

2.5.1 Molecular cloning and transformations

To generate the remaining yeast strains not described in Staller et al (MATa/MATalpha, synthetic TF-Kanr::ura3/URA3 GFP reporter-natR::YBR032w/YBR032w), eleven AD sequences (Table A2) were ordered as DNA fragments from SynBio Technologies (Monmouth Junction, NJ). Those AD fragments have homology arms for cloning into pMV219 (Addgene), a plasmid vector that contains the pACT1-mCherry-DBD-ERD cassette described above. More specifically, pMV219 was digested with Nhe1-HF (NEB, R3131S) and Asc1-HF (NEB,R0558S), AD sequences were

cloned downstream of the ERD using HIFI assembly (NEB, E2621S). The entire pACT1-mCherry-DBD-ERD-AD region was then PCR amplified using primers YP16 and YP17 (Supplement) with homology targeting the URA3 locus. The PCR product was then transformed into FY4 S288c yeast by incubating with a mixture of 30-33% PEG, 100 uM lithium acetate, and 0.3 mg/mL boiled salmon sperm DNA for 30 minutes at 30 C followed by 1 hour of heat shock at 42 C. The transformed yeast were spun down, resuspended in YPD, and plated on nonselective medium, followed by replica plating onto YPD+G418 plates (200 mg/ml). The KanR positive yeast were then struck out onto SC-URA and 5-FOA (1 mg/ml) plates to test for loss of the URA3 locus.

Integrations were confirmed by colony PCR targeting the upstream breakpoint. To extract genomic DNA, URA- yeast strains were grown overnight in YPD, spun down and resuspended in 500 ul each of lysis buffer containing 100 mM Tris, 50 mM EDTA, and 1% SDS. They were then vortexed with silica beads for 2 minutes each. The liquid was then pipetted off the beads, mixed with 275 uL of 7M ammonium acetate pH7, and incubated for 5 minutes 65 C then 5 minutes on ice. 500 uL of chloroform was added, and the mixture was vortexed and spun for 2 mins. The supernatant was then added to 1 mL isopropanol, incubated for 5 minutes at room temperature, and spun down for 5 minutes. The resulting pellet was then washed with 70% ethanol, air dried, and dissolved. For PCR, the genomic DNA was first digested with Nhe1. 2.5 uL of digest was then mixed with 2.5 uL of each primer (put sequences in supplement), 2.5 uL water, and 10 uL of NEB OneTaq and run for 34 cycles of 30 seconds at 94 C, 30 seconds at 55 C, and 1 minute at 72 C. The resulting PCR product was loaded directly onto an agarose gel, which was then run and imaged.

2.5.2 Beta-estradiol induction and flow cytometry

All single-cell fluorescence measurements were collected using a Beckman Coulter cytoflex S flow cytometer. In the experiments with wild-type activation domains, the optical densities of 1:10 dilutions of the overnight cultures were first measured, and all cultures were diluted to an OD of 0.0225. For each combination of strain and level of induction, 200 uL of diluted culture was added

to a well of a deep 96 well plate. 50 uL of diluted beta-estradiol was then added to the well (final concentrations are listed in Table A1). The diluted cultures were incubated with the beta-estradiol in a shaker incubator set to 300 rpm for 5 hours before measuring. Finally, 100000 such single-cell fluorescences were collected for each combination of activation domain and induction level.

In the experiments with mutant activation domains, 5 uL of overnight culture from each well was first diluted into 200 uL of SC in the corresponding well of a new deep well plate. 50 uL of 1 mM beta-estradiol diluted 1:1000 was added to each well of the new plate, and the plate was again incubated for 5 hours in a shaker set to 300 rpm before measuring. Strains for which we suspected contamination or other experimental error were then subjected to verification experiments. These involved growing them up as liquid cultures in tubes and then streaking them out on agar plates. Single colonies were then picked and used to start liquid cultures which were then induced and measured using the same protocols used for the wild-type activation domains, except that, for each mutant, we collected measurements for a fixed 45 seconds.

2.5.3 Imaging

Yeast strains MY447 (expressing a VP16 fusion), YM23.31 (expressing a War1 fusion), MY450 (expressing a GCN4 fusion), and MY460 (expressing a Gal4 fusion) were grown overnight from streakouts in 1 mL of SC. They were then incubated for 4 hours in 250 uL of SC with or without 200 uM beta-estradiol. The yeast were immobilized on agarose pads and imaged on a Zeiss LSM 880 Confocal.

2.6 Quantification and statistical analysis

2.6.1 Simulations

Sets of simulations were run to predict the effects on mean and noise of varying one or two parameters in a given stochastic model of gene expression, while keeping all others constant. The models

simulated were the random telegraph model [30, 94, 95], the multi-state model of Rodriguez el at [98], the refractory period model of Zoller et al [97], and the transcription cycle model of Scholes et al [96]. These simulations were run using the implementation of Gillespie's stochastic simulation algorithm [109] in BioNetGen [110]. Model files are available in Supplementary folder 2, and sets of input parameters used for each simulation are available in Supplemental Table A3. For each set of parameters, 1000 individual simulations were run and a mean and a Fano factor was computed for the distribution of protein expression at the ends of the individual runs. The relationships between mean and Fano factor over multiple sets of parameters within the same model were plotted using the Matplotlib and Seaborn packages in Python.

2.6.2 Flow data analysis

Outliers in forward and side scatter area and height were removed at the collection stage through gating. Summary statistics of GFP fluorescence (mean and Fano factor) were then computed over the cells whose mCherry fluorescence fell between 300 and 400 units, and the results were plotted, again using Matplotlib and Seaborn.

2.6.3 Image analysis

The images were segmented in CellProfiler using the GFP signal, and total mCherry signal was quantified for the nuclear and cytoplasmic segments of each cell.

Chapter 3: The effects of cis-regulatory sequence on noise in gene expression

3.1 Introduction

Using our yeast reporter system, we have dissected the effects of trans regulatory factors on gene expression noise. However, cis-regulatory factors are also major influences on noise in gene expression. Multiple studies have found effects on noise from cis-regulatory features such as TATA boxes in promoters [111], chromatin architecture [111–114], and the presence of enhancers [66, 78, 115]. However, recent modeling and experiments [86, 114] suggests that transcription factor binding is the primary driver of gene expression noise. Notably, while many studies have made large-scale perturbations to higher-order features like the ones mentioned, few have specifically measured the effect on noise of perturbing TF binding sites.

There are multiple features of TF binding sites (TFBS) that can be perturbed and are potentially drivers of noise. The most basic is the affinity of the cis-regulatory element for its cognate TF(s), which can be modulated by introducing mismatches between the TFBS sequence and the TF's consensus binding motif. Furthermore, Parab et al [86] suggest that the number of distinct TFs binding to a promoter, and the cooperative and competitive interactions between them, is the primary determinant of expression noise levels in yeast. These parameters can also be straightforwardly manipulated, the former simply by adding binding sites for additional TFs, and the latter by altering the spacing between the binding sites. In general, binding sites spaced such that TFs bind close together and on the same side of the double helix will favor interactions between TFs, while spacings that place TFs further apart and on opposite sides of the double helix will disfavor such interactions [116]. Finally, while most studies in the field have focused on activation, repression domains are also

widespread within TFs, and analysis of developmental single-cell transcriptomic data suggests that repression, rather than activation, is the primary contributor to noise in expression [117]. Due to its highly modular and controllable nature, our yeast reporter system allows us to make cis-regulatory changes to modify any of these features. Taking advantage of this fact, we have made perturbations systematically targeting each of these features and measured their effects on noise.

3.2 Results

3.2.1 Cooperativity increases mean reporter expression, with variable effects on noise

To specifically probe the effects of binding site spacing and TF cooperativity on noise, we designed two reporters (Figure 3.1A and B) to perturb these parameters in opposing fashions. One of these reporters is designed to maximize interactions between TF molecules bound at the promoter. We achieved this by spacing all binding sites for the synthetic activator 21 base pairs, that is, two helical turns apart. In contrast to the original reporter, in which the activator binding sites are unevenly spaced, this places all sites on the same side of the helix. We also designed a reporter to achieve the opposite aim. This reporter has binding sites spaced 16 base pairs or 1.5 helical turns apart. By placing adjacent binding sites on opposite sides of the double helix, this arrangement is intended to minimize interactions between TF molecules bound at neighboring positions.

The new reporters were integrated into the HIS locus in MATalpha haploids and crossed with the MATa haploids expressing the synthetic TF constructs generated in Chapter 1. When we measured the mean and noise of these crosses' reporter expression distributions, we observed clear differences between the new reporters. In the context of activation domain swaps (Figure 3.2A), consistent helical phasing, and thus presumably cooperativity, appears to increase mean expression. However, the effects on noise are difficult to discern in this context, due to massive clonal variability between lines with independent integrations, even though all of the crosses have the reporter integrated

at the same locus. A potential explanation for this clonal variability is variable spreading of the chromatin state in regions surrounding the HIS locus leading to variable silencing, aka, position effect variegation [118]. The crosses expressing the reporter designed to maximize cooperativity appeared to form two clusters, with the mean and noise of one overlapping with that of the reporter designed to minimize cooperativity, and the other exhibiting high mean and noise. Within the high mean, high noise group, noise appeared to decrease with mean. However, because most of the variability within this group is clonal and not related to genotype, it is difficult to analyze its behavior further within the scope of this study.

Examining the behavior of the new reporters in the context of a single cross and variable nuclear concentration of the activator, patterns start to become clearer (Figure 3.2B). We took crosses of each reporter with a single MATa clone expressing an HSF1 activation domain fusion and measured their expression distributions across a range of beta-estradiol levels, as in chapter 1. In this context, cooperativity appears to modify the mean-noise relationship, with the reporters designed to maximize cooperativity have higher mean expression but similar noise levels. This leads to a decoupling between mean and noise opposite that observed in chapter 1, with the more active constructs in terms of mean having expression that is less noisy when levels of induction are adjusted to equalize mean expression, rather than more noisy as in the case of different activation domain fusions.

3.2.2 Repression by REB1 suppresses noise

We then asked what effect repression has on noise. More specifically, we asked whether the effect on noise of reducing reporter expression by adding binding sites for a repressor is distinguishable from that of reducing reporter expression by reducing activator binding. To answer this question, we designed two reporters (Figure 3.1C and D). One of these reporters has binding sites for the TF REB1 inserted between the activator binding sites of the original reporter. (REB1 acts as a repressor in this context, leading to mean expression levels lower than observed from the original reporter.) To design the other, we made point mutations in the original activator binding sites to its binding

affinity for the synthetic TF construct. These reporters were again integrated at the HIS locus in MATalpha haploids and crossed with the MATa haploids generated in Chapter 1.

When we measured the mean and noise of the new reporters' expression across all crosses under uniform culture conditions (Figure 3.2A), we observed a striking difference between the reporters. The reporter with weakened activator binding sites exhibited strong clonal variability and varied widely in both mean and noise. The reporter with added REB1 sites, on the other hand, exhibited almost uniformly low noise despite occupying the same range of mean expression levels as the reporter with weakened activator binding. While noise did increase with mean among the lowest-expression crosses, noise otherwise varied little with mean and was consistently lower than among crosses expressing the reporter with weakened activator binding.

We also observed this pattern in the context of fixed genotypes and varying induction levels (Figure 3.2B). When we induced differing levels of activator nuclear localization by varying the level of beta-estradiol in the growth medium, crosses expressing an HSF1 AD fusion TF and the reporter with added REB1 binding sites again exhibited consistently low noise. This again contrasts with the reporter with weakened activator binding sites, whose expression was noisier and much less responsive to the level of TF stimulation. Overall, it appears that REB1 greatly suppresses noise in its targets' expression.

3.2.3 Distinct mechanisms of repression have distinct effects on noise

The striking difference in noise between the two reporters led us to ask what molecular mechanisms might underlie the effect of repression on noise. To link mechanisms of repression to specific noise behaviors, we constructed a simple model of repressor action amenable to stochastic simulation via the Gillespie algorithm. In this model, we assume that there are two TFs that can bind to the promoter, an activator and a repressor, which bind at different sites. We further assume that activator binding corresponds to a transcriptionally "on" state, with the overall kinetics of the transcription following the random telegraph model, with the addition of a distinction between repressor bound

and unbound states. Under these assumptions, there are two basic mechanisms by which repressor binding can reduce transcription: it can either reduce activator binding (Figure 3.3A), or reduce the rate of transcription while the activator is bound (Figure 3.3C). We simulated the effects of progressively adding repressor to the system in these two scenarios.

Simulations suggest that qualitatively distinct mechanisms of repression will have clearly distinct effects on noise. In the scenario in which repressor binding reduces the time spent in the "on" state by inhibiting activator binding, noise goes up as the level of the repressor increases and the mean level of expression goes down (Figure 3.3B). Conversely, the repressor that reduces the rate of transcription in the "on" state without inhibiting repressor binding is predicted to simultaneously suppress mean expression and noise as its level increases (Figure 3.3D). These qualitative predictions hold across a broad range of assumed parameter values. The consistently low noise from the REB1-binding reporter is clearly consistent with the latter scenario in which REB1 reduces the rate of transcription in the active, activator-bound state without affecting the time spent in this state.

3.2.4 The effect of MIG1 on noise resembles that of decreased activator binding

We then asked whether other repressors exhibit the same noise behavior as REB1. The original reporter construct contains a binding site for the repressor MIG1. MIG1 is a key regulator of yeast's response to glucose, and is induced by the presence of glucose in the growth medium. We thus assayed the effect of removing MIG1 activity by growing the yeast on raffinose as an alternative carbon source, and asked whether it would mirror that of manipulating REB1 binding.

Looking first at the effect of carbon source on the reporters with varying activator binding site spacing in the context of varying beta-estradiol levels (Figure 3.4A), we observed that removing MIG1 appeared to phenocopy increases in beta-estradiol induction. Yeast expressing these reporters followed the same trendlines in mean-noise space when grown on either carbon source, with those grown in raffinose exhibiting higher mean and noise. Turning to the effect of carbon source on

reporters with reduced expression (Figure 3.4B), we observed a striking pattern from the reporter with weakened activator binding sites, which actually produced lower noise at high levels of beta-estradiol and mean expression when the yeast were grown on raffinose and thus lacking MIG1 activity. This is consistent with the mechanism of repression in which inhibition of activator binding reduces time spent in the "on" state, and contrasts sharply with the effects of MIG1 binding. Finally, when yeast expressing the REB1-binding reporter were grown on raffinose, the effect of REB1 overrode the effect of removing MIG1, with the reporters producing low noise throughout the full range of beta-estradiol concentrations.

3.3 Discussion

Here we have assayed the effects of a diverse array of cis-regulatory perturbations on noise, across a wide range of trans-regulatory contexts. In contrast to the effects of trans-regulatory perturbations, which generally alter mean and noise in the same direction and differ primarily quantitatively, we find that cis-regulatory perturbations have much more complex and heterogeneous effects on noise which are often opposite to their effects on mean. This implicates cis-regulatory architecture as key to specifying noise levels, and suggests that interactions between multiple TFs bound to the same regulatory element are central to the mechanisms generating expression noise.

When we manipulated the spacing between TF binding sites, we found that spacings that facilitate intermolecular interactions reduce noise relative to mean. This stands in stark contrast to the prediction in [86] that cooperative binding would lead to higher noise than independent binding. A possible explanation for this discrepancy is that the model in [86] assumes that cooperative binding is switchlike and all-or-nothing in nature. This represents an extreme case of cooperativity that might not be reflective of TFs' behavior in the cell. In the case of our reporter system, the estrogen response domain is known to dimerize, while the other components are not known to specifically self interact. Because there are more than two activator binding sites in our reporters, dimerization will increase the effective affinity of the TF constructs for conducively spaced binding sites, while

more distant sites can still be bound independently of one another. The effect of cooperativity in our system might therefore be to increase promoter occupancy without making binding and activation significantly more switch-like.

An especially striking finding from our experiments is that different repressors can have opposing effects on noise that are consistent with distinct mechanisms of repression. This contrasts with the noise effects of activators, which, as we observed in Chapter 2, form a single allelic series. These observations are consistent with recent high throughput screens for effector domain function [18], which have found that activation domains tend to be unstructured and homogenous in their sequence features, while repressor domains are more structurally diverse and can be grouped into distinct classes. This also resembles the results of targeting experiments, which show that repressive chromatin modifiers that act on different marks have clearly distinct kinetic profiles [119]. However, previous studies of repression kinetics have focused on the rates of silencing and reactivation, while our simulations suggest that the primary factor differentiating repressors with distinct noise profiles is whether repression excludes activator binding. Such a difference could be explained by repressive cofactors differing in their effects on accessibility. Under this hypothesis, repressors that reduce noise along with mean would recruit cofactors that do not impact accessibility, while repressors that increase noise would recruit cofactors that promote a heterochromatic, inaccessible state that prevents other TFs from binding. Analyses of coaccessibility and gene expression via single-cell sequencing provide support for this model, suggesting that there are indeed different classes of repressors that act through mechanisms with opposing effects on accessibility [120]. Overall, our results implicate repressors as prime candidates for regulating noise independently of mean, as has been observed by Antolovic et al [117].

3.4 Methods

Molecular cloning, transformations, yeast culture, and flow cytometry measurements followed the same protocols as in chapter 1, except for the following modifications. Reporter sequences were

designed in Snapgene and ordered from synbio. The reporters were synthesized and cloned into vector pMVS008P_TDH3VenusNATMX6 by synbio. To add homology arms, we performed PCR directly on resuspended lyophilized plasmid DNA, using the primers HIS_UP_homology2 and HIS_dwn_homology2 targeting the HIS locus. The annealing temperature was progressively raised over the course of PCR, being 50 C for the first 2 cycles, 65 C for the next 2, and 72 C for the remainder. The inserts were transformed into FY5 (BAC45) MATalpha yeast as in chapter 1, with NAT used to select for transformants instead of G418. To confirm integration at the HIS locus, transformants were struck out on HIS- medium and screened for lack of growth. Additionally, we performed PCR on DNA extracted by the protocol in chapter 1, using primers HIS_up_in_1 and HIS_up_out_1 targeting the 5' integration breakpoint, and HIS_dnw_in_1 and HIS_dwn_out_1, targeting the 3' integration breakpoint. Transformants with confirmed integrations at HIS were crossed with the MATa transformants generated in chapter 1 by mixing YPD liquid cultures in a deep-well plate and leaving at room temperature overnight. The crosses were then pinned onto medium containing both selectable markers (G418 and Nat) to select for diploids. Reporter expression was then induced and measured as in chapter 2, and the resulting data was analyzed using the same computational pipeline.

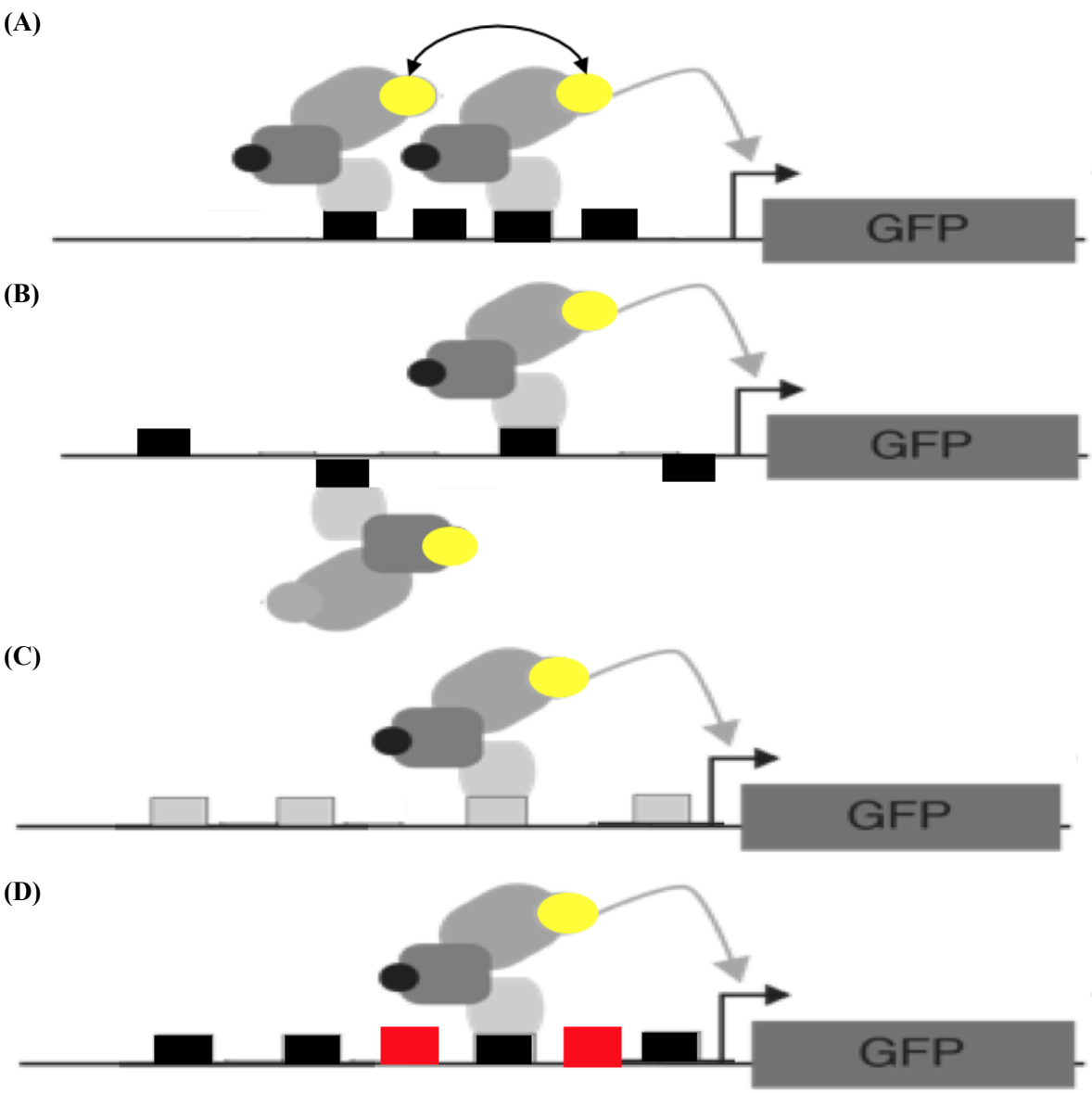

Figure 3.1: Promoter architectures assayed in this chapter. (A) Reporter with binding sites for synthetic activator spaced 21 bp (2 full helical turns) apart to maximize cooperativity (B) Reporter with binding sites for synthetic activator spaced 16 bp (1.5 full helical turns) apart to minimize cooperativity (C) Reporter with point mutations in binding sites for synthetic activator to reduce binding affinity (D) Reporter with added binding sites for REB1 to introduce repression

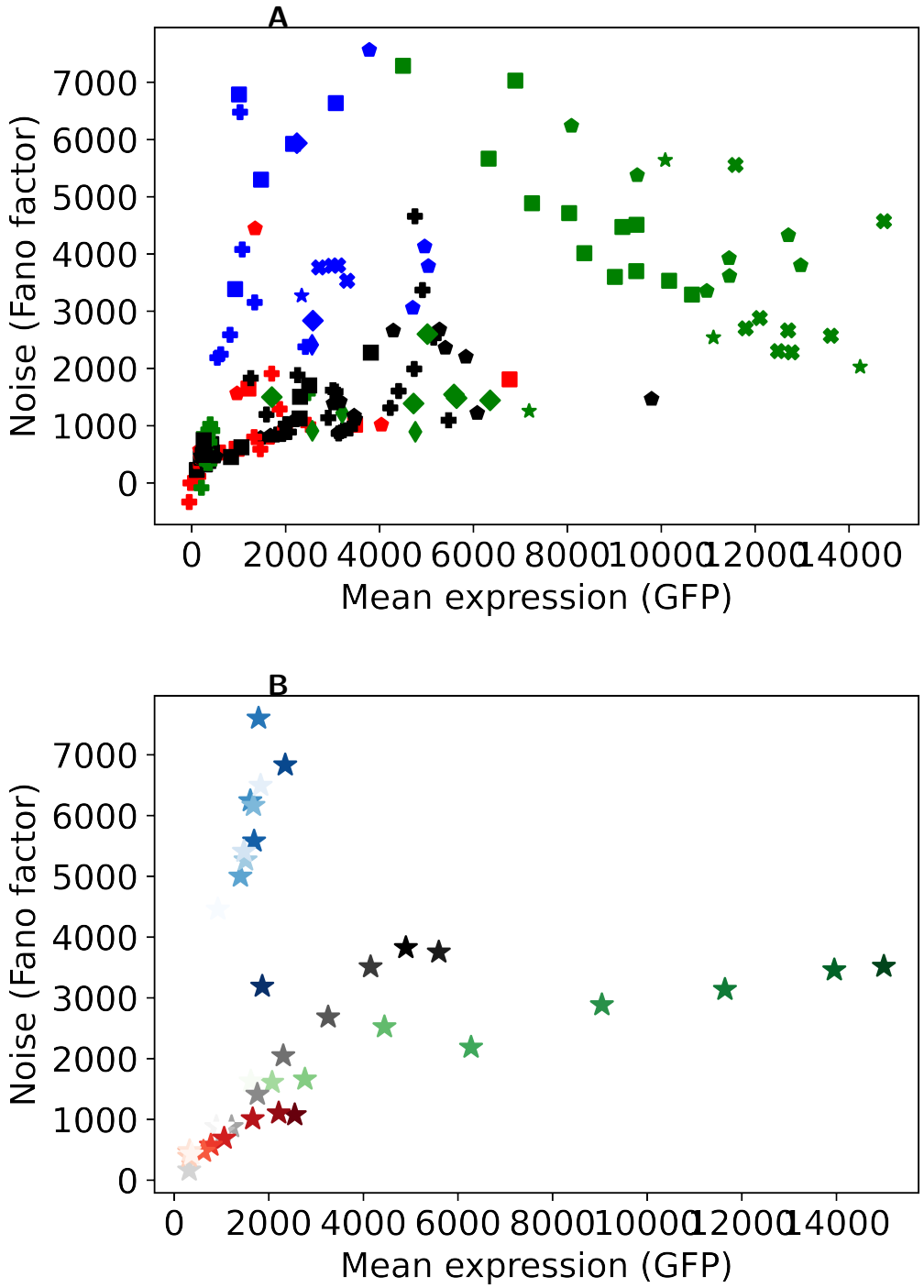

Figure 3.2: Effects of cis-regulatory perturbations on mean and noise. (A) Noise levels quantified using Fano Factor (on Y axis) vs. Means (on X axis) of GFP expression across pairings of CRE (indicated by hue) and AD (indicated by shape) (B) Noise levels quantified using Fano factor (on Y axis) vs. Means (on X axis) of GFP expression across pairings of CRE (indicated by hue) and beta-estradiol level (indicated by shade)

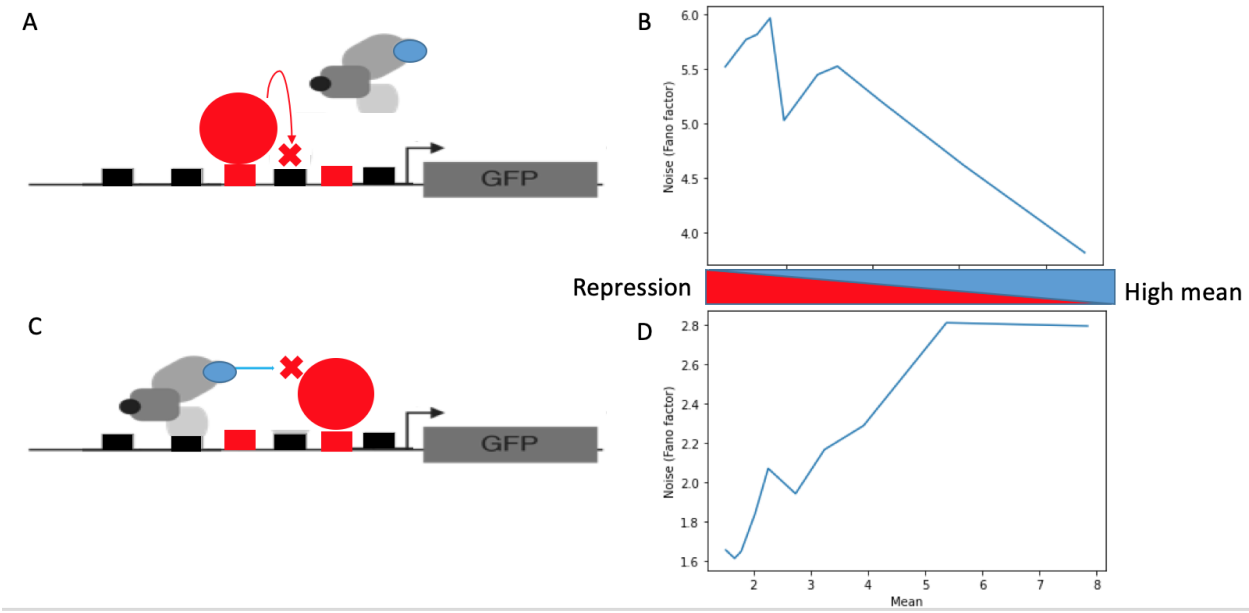

Figure 3.3: Predicted effects of repression on noise. Repressors which act through a mechanism that prevents activator binding (A) are predicted to increase noise as they decrease mean (B). Repressors that act through a mechanism that allows activator binding but prevents transcription from the activator bound state (C) reduce both noise and mean (D).

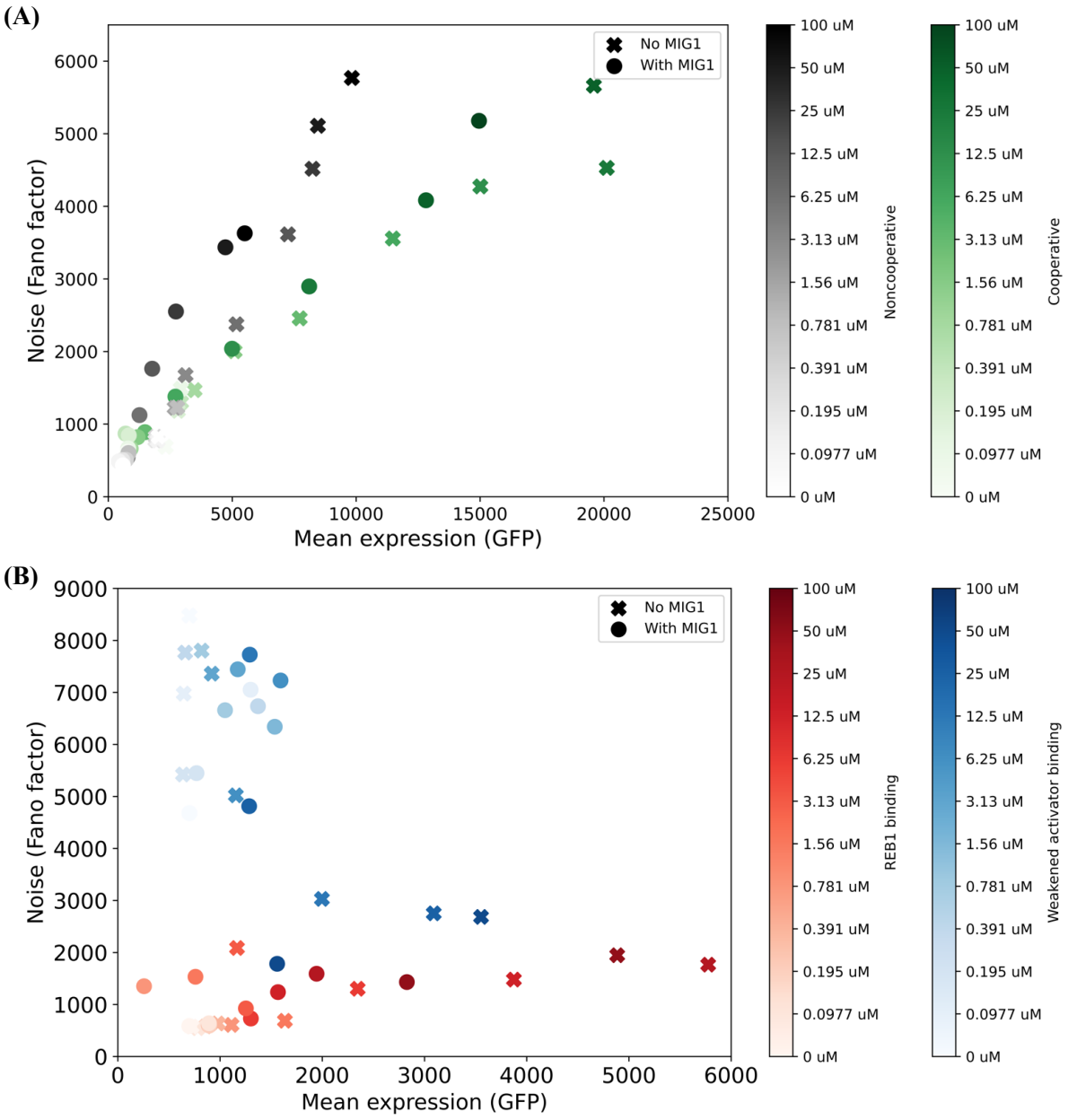

Figure 3.4: Effects of MIG1 perturbations on mean and noise. (A) (Top) Noise levels quantified using Fano Factor (on Y axis) vs. Means (on X axis) of GFP expression. Yeast grown in dextrose medium, and therefore with MIG1 activity, are indicated by circles, while yeast grown in raffinose medium, and thus without MIG1 activity, are indicated by crosses. The reporter with activator binding sites spaced 2 full helical turns apart to promote cooperativity is in green, and the reporter with activator binding sites spaced 1.5 helical turns apart to suppress cooperativity is in black. Beta-estradiol concentration is indicated by shading. (B) (Bottom) Noise levels quantified using Fano Factor (on Y axis) vs. Means (on X axis) of GFP expression. Yeast grown in dextrose medium, and therefore with MIG1 activity, are indicated by circles, while yeast grown in raffinose medium, and thus without MIG1 activity, are indicated by crosses. The reporter with point mutations in the activator binding sites to reduce affinity is in blue, and the reporter with added repressor binding sites is in red. Beta-estradiol concentration is indicated by shading.

Chapter 4: Transcription factor interactions explain the context-dependent activity of CRX binding sites

4.1 Introduction

A typical mammalian transcription factor (TF) binds hundreds or thousands of cis-regulatory sequences (CRSs) in the genome [121–123]. CRSs that are bound by the same TF vary widely in their activity, and can include strong enhancers, transcriptional silencers, or sequences with weak or no cis-regulatory activity [19, 91, 123–138]. Such dramatic functional differences among CRSs with similar TF binding sites (TFBSs) show that local sequence context modulates the contribution of a TFBS to cis-regulatory activity, yet how this occurs is not well understood. Proposed models of cis-regulatory grammar that may account for context-dependence vary in their emphasis on the importance of interactions between TFs, and they suggest different degrees of flexibility in the possible functional arrangements of TFBSs [139–142]. The enhanceosome model proposes that strict geometrical constraints determine whether CRS-bound TFs can activate transcription, suggesting that context-dependent effects of TFBSs are strongly influenced by highly specific interactions between them [139–141, 143]. The contrasting billboard model proposes that active CRSs are defined by the presence of a sufficient number of bound TFs, with no strong constraints governing their arrangement [142]. The billboard model implies that the context of a TFBS is determined primarily by additive effects of the surrounding TFBSs, with few specific interactions between sites. Other models of cis-regulatory grammar propose that individual TFBSs are weak on their own and depend on strong cooperative interactions [16], that particular TFBSs recruit specific, required transcriptional cofactors [91], or that the balance between sites for transcriptional activators and sites for repressors determines whether a CRS is an enhancer or silencer [144–146]. The degree

to which these proposed features of cis-regulatory grammar modulate the local context within a CRS is not well understood. As a result, accurately predicting the activity of CRSs or the effects of genetic variants in TFBSs remains an unsolved problem.

Local sequence context has strong effects on the function of binding sites for the retinal TF Cone-rod homeobox (CRX) [147–149]. CRX is a paired-type K50 homeodomain TF and a critical regulator of transcription in multiple retinal cell types, where it contributes to both activation and repression of cell type-specific genes [128, 147, 149–158]. Using massively parallel reporter assays (MPRAs), we previously found that genomic CRX-bound sequences include strong enhancers and silencers [19, 136, 159, 160]. The activities of these CRSs, whether activating or repressing, depend on both CRX binding sites and CRX protein, which demonstrates that the effects of CRX sites are modulated by context [19, 159]. Yet how local sequence context determines whether a CRX-bound region functions as an enhancer or a silencer is poorly understood. We previously found that synthetic CRSs with sites for CRX and the rod photoreceptor-specific leucine zipper TF NRL were often strongly activating [19], and CRX cooperatively interacts with NRL at some promoters [22, 128, 147, 153, 155, 156, 161]. We also found that CRX-bound silencers tend to contain more copies of the CRX motif than CRX-bound enhancers [19, 136, 159], while CRX-bound enhancers are enriched in sites for other TFs relative to silencers [136]. However, it is unclear why CRX binding sites have an activating effect at some elements and a repressing effect at others. We hypothesized that interactions between CRX and other co-bound TFs determine whether a sequence functions as an enhancer or silencer, and we sought to capture those interactions in a quantitative model.

A key advantage of synthetic CRSs is that their binding site composition can be systematically varied to generate informative training data for models of cis-regulatory grammars. We used our MPRA data from synthetic CRSs to train neural network-based models using MAVE-NN, a modeling framework designed to efficiently model data from massively parallel functional assays [24]. We find that the effects of CRX sites are explained by a model that includes positive, additive

contributions of individual TFBSs, negative homotypic interactions between sites for the same TFs, and positive heterotypic interactions between sites for different TFs. The model explains the observations that CRX sites produce context-dependent activation and repression, and that the addition of an NRL site converts silencers to enhancers. The model also accounts for our finding that CRX-bound enhancers have sites for diverse TFBS, while CRX-bound silencers lack this diversity. More generally, our results suggest that context-dependent activity of binding sites for transcriptional activators can be explained by the balance between the negative effects of interactions between sites for the same TF, the positive effects of individual TFBSs, and heterotypic cooperativity between sites.

4.2 Results

4.2.1 Positive heterotypic and negative homotypic interactions explain the effects of CRX and NRL sites on expression

We previously reported that both genomic and synthetic CRSs with many binding sites for CRX tend to act as silencers, while CRSs with fewer CRX sites tend to act as enhancers [19, 136]. Our prior results from a reporter library of 1,299 synthetic CRSs showed that sequences composed of only CRX and NRL binding sites exhibit activity that ranges from strong activation to repression [19]. These CRSs were tested by MPRA in mouse retinal explants, which preserve all retinal cell types and cell type-specific TFs that comprise the native context in which CRX is active. We observed that sequential addition of CRX sites upstream of a basal promoter led first to increased activation and then to repression below basal levels when three or four CRX sites were present (Figure 4.1A and B). Repressive CRSs with four CRX sites could be converted to strongly activating sequences by replacing one CRX site with a site for NRL. Synthetic sequences composed of multiple CRX sites and one NRL site were more active than equal length CRSs composed of only CRX sites (Figure 4.1B) or only NRL sites (Figure A.7). We found that genomic CRX-bound sequences followed a

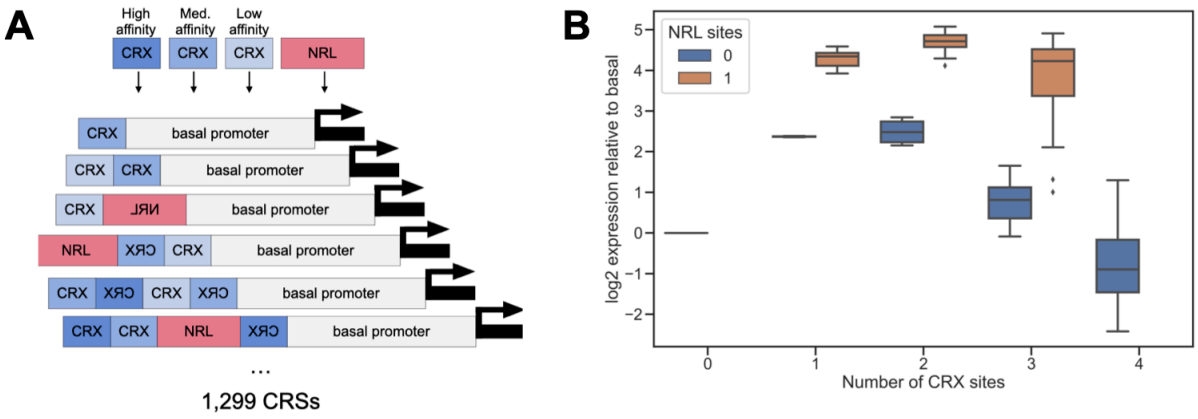

Figure 4.1: Synthetic CRSs sites reveal context-dependent effects of CRX and NRL sites. (A) Design of synthetic CRS MPRA library reported in [19]. Combinations of CRX and NRL sites (up to four TFBSs) were cloned adjacent to either a Rho or a Hsp68 basal promoter. TFBSs could be in either forward or reverse orientation. (B) MPRA activity (y-axis) of CRSs composed only of high affinity CRX sites (blue) is consistently lower than that of CRSs with high affinity CRX sites and one NRL site (orange), relative to the Rho basal promoter. Sequential addition of high affinity CRX first activates, then represses the Rho basal promoter. Plot shows a subset of the data reported in [19].

similar pattern [19]. Thus, our previous experiments with systematically varied synthetic CRSs show that a sequence context composed of only two types of TFBSs strongly modulates the effects of CRX binding sites. However, it is unclear what kinds of interactions among CRX and NRL sites could account for such context-dependent activity.

To discover interactions among CRX and NRL sites that might explain context-dependent activity, we trained a model using MAVE-NN, a recently published neural network framework that is designed specifically to model data obtained from massively parallel functional assays [24]. A key strength of MAVE-NN is that it deconvolves sequence-function relationships from the confounding effects of experimental non-linearities and noise. Using MAVE-NN, we modeled our synthetic CRS data. As noted above, the synthetic CRSs were composed of systematically varied combinations of CRX and NRL sites [19]. Sequences included up to four sites in either the forward or reverse orientation. TFBSs included high, medium, and low affinity versions of CRX sites and the consensus site for NRL (Figure 4.1A). CRSs were cloned upstream of either the murine Rho or Hsp68 basal promoter. The library included all 584 possible combinations of one, two, and three TFBSs, and 715 sequences randomly sampled from all possible combinations of four sites. We trained MAVE-NN models with different architectures to predict MPRA activities from sequence alone. We reasoned that due to the small number of TFBSs included in the synthetic CRSs and the uniform spacing between them, additive models with or without interaction terms would capture most of the effects of CRX and NRL sites on reporter activity. Each model infers a relationship between a reporter gene sequence and its latent phenotype, which represents the intrinsic activity of the CRS that is indirectly read out by the MPRA. MAVE-NN simultaneously models (1) the relationship between DNA sequence and the latent phenotype, and (2) the nonlinear relationship between the latent phenotype and the noisy MPRA measurement. MAVE-NN quantifies the performance of the models using an information theoretic measure called predictive information [24, 27]. Predictive information is the mutual information between the inferred latent phenotype and the MPRA measurement, and it represents how well the model captures the relationship between a reporter gene's inferred intrinsic activity and its MPRA output. We used predictive information to compare the performance of four different model architectures: (1) an additive model lacking interactions between TFBSs, (2) a nearest-neighbor model that only allows interactions between neighboring TFBSs, (3) a pairwise interaction model allowing interactions between all pairs of

TFBSs regardless of spacing, and (4) a "black box" multilayer perceptron model that makes no prior assumptions about the interactions between TFBSs. To train the models, the data was randomly split among training (80%), validation (10%), and test (10%) sets. All performance metrics were computed from the test set. Model parameters for analysis were taken from the best performing model out of multiple random initializations.

Of the three model architectures that included additive and interaction terms, we found that the pairwise interaction model achieved the best overall performance (Figure 4.2A and B). This model provided 1.8 bits of predictive information, roughly equivalent to an accurate three-way classification of CRSs by activity. The predictive information of the pairwise model (1.82 bits) is approximately half that of the multilayer perceptron "black box" model (3.00 bits). The disparity in predictive information between the pairwise and black box models suggests that additional higher order interactions between TFBSs likely account for much of the unexplained activity of the synthetic CRSs. However, this unexplained activity likely consists of small discrepancies between sequences with similar activities, because the pairwise model captured a substantial fraction of the variation in reporter activity (Figure 2B and C, R2 = 0.889). To understand how additive and interaction effects of TFBSs might explain the context-dependent activity of CRX sites, we examined the parameters of the pairwise model. We found that the additive contributions of all TFBSs, averaged over all four possible positions in the synthetic CRSs, were positive (Figure 4.2D). This is consistent with the roles of CRX and NRL as transcriptional activators. The average additive contribution of CRX sites increases with site affinity, particularly for sites in the forward orientation. High affinity CRX sites have a stronger positive, additive effect than NRL sites, suggesting CRX is a stronger activator (Figure 4.2D). The additive terms of the model reflect the expected effects of simple transcriptional activators whose probability of binding to a CRS is determined by the number and affinity of binding sites. However, these positive, additive terms do not account for the context-dependent effects of CRX binding sites observed in the data shown in Figures 4.1B and A.7.

Examining the interaction terms of the model, we observed a pattern of positive, heterotypic

cooperativity between CRX and NRL sites, and negative homotypic interactions between binding sites for the same TF (Figure 4.2D). Negative homotypic interactions are strongest between NRL sites and between high affinity CRX sites, and they decrease with decreasing CRX site affinity. CRX sites also show a strong, affinity-dependent negative interaction with the Rho basal promoter, which contains three CRX sites [147, 162]. Negative homotypic interactions were especially strong between adjacent sites but can be observed at all distances in the synthetic CRSs (Figure A.8). Positive interactions between CRX and NRL sites also occur at all distances and depend on binding site affinity. The modeling results suggest that activating and inhibiting interactions between CRX and NRL are the primary determinants of the activity of CRSs with binding sites for these two TFs.

Taken together, the parameters of the pairwise interaction model reveal a cis-regulatory grammar that accounts for the observed context-dependent activity of CRX and NRL sites. Consistent with the known roles of CRX and NRL as transcriptional activators, sites for these TFs consistently make positive, independent contributions to activity. However, negative homotypic interactions reduce activation or lead to repression when multiple sites for the same TF are placed together. The repressive effect of negative homotypic interactions can be overcome by the strong positive cooperativity between CRX and NRL. An important feature of this cis-regulatory grammar is that additive effects and interactions scale differently with the number of TFBSs. The independent contributions of TFBSs increases linearly with the number of TFBSs, while the interaction effects increase with the square of the number of TFBSs. These differences in scaling have a strong impact on CRSs with multiple sites and explain why the replacement of a single binding site can convert a silencer to an enhancer (Figure 4.1B).

4.2.2 Positive heterotypic interactions require CRX protein

We previously reported that many genomic and synthetic CRSs with CRX binding sites either retain or gain activity in Crx-/- retina, despite the loss of CRX protein [13]. Activity in Crx-/- retina still requires intact CRX sites, indicating that another TF, possibly the CRX ortholog OTX2, acts at these

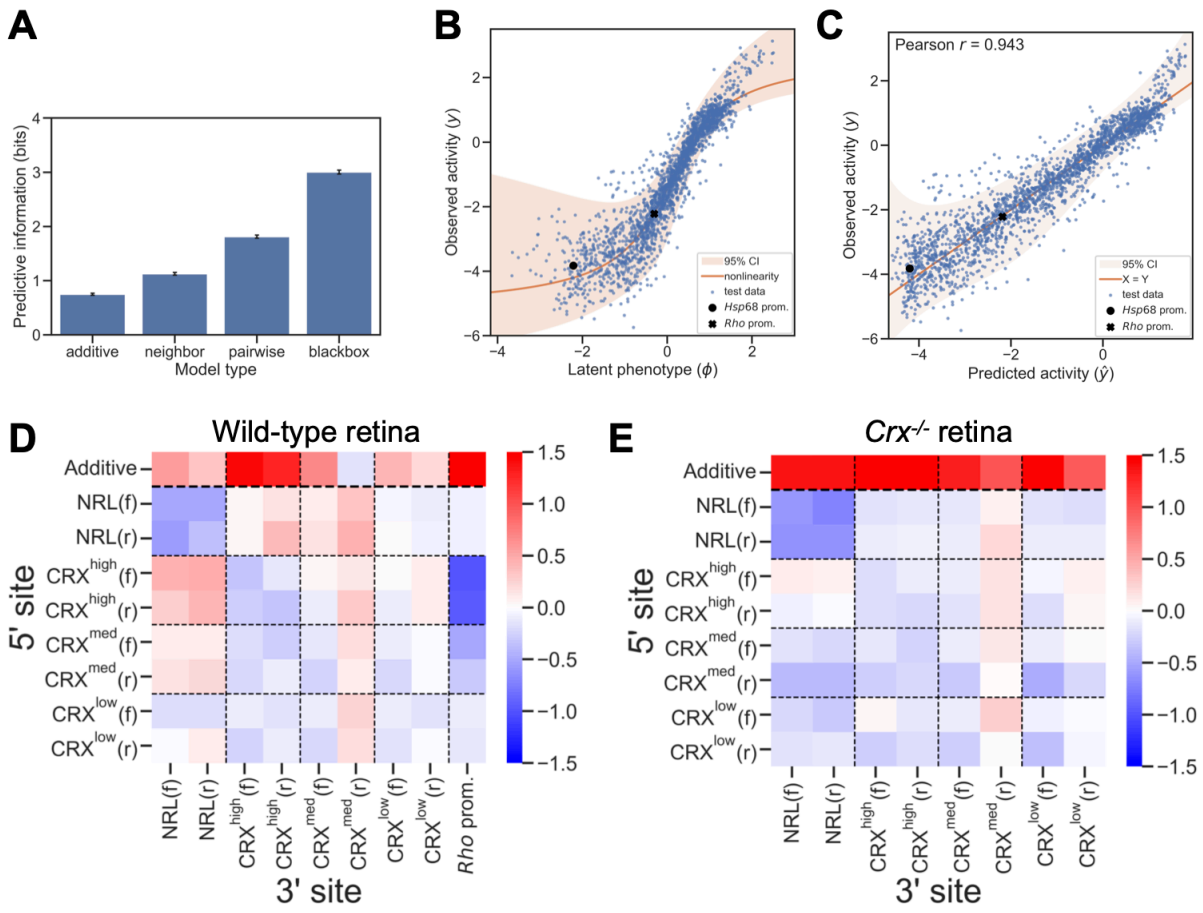

Figure 4.2: A model of CRX and NRL-driven cis-regulatory activity in wild-type retina. (A) The performance of different model architectures (measured as predictive information) fit to MPRA measurements of the CRX-NRL library in wild-type retina. Error bars indicate standard error. (B) The observed activity (y-axis) of test set sequences in wild-type retina compared to the latent phenotype (x-axis) inferred by the pairwise model. (C) The observed activity (y-axis) of test set sequences in wild-type retina compared to the activity predicted by the pairwise model (x-axis). (D) Model parameters for additive and pairwise contributions of CRX and NRL sites and the Rho promoter to activity in wild-type retina, averaged across the four positions in synthetic CRSs. For pairwise interactions rows indicate the 5' site and columns indicate the 3' site. Forward and reverse orientation of the TFBS is indicated by (f) and (r). (E) Model parameters for additive and pairwise contributions of CRX and NRL sites to activity in Crx-/- retina, averaged across positions and spacings.

sites when CRX is absent. To examine how additive and pairwise interactions among TFBSs change in the absence of CRX, we trained a pairwise interaction model on prior data from the synthetic CRS library tested in Crx-/- retina, with the Rho basal promoter. The Crx-/- model performed similarly to that trained on data from wild-type retina (2.21 bits of predictive information, Figure A.8 A-C, R2 = 0.900 for predicted versus observed activity). In this model, additive effects of all TFBSs remained positive (Figure 4.2E), indicating that these CRSs continue to be bound by transcriptional activators in Crx-/- retina. Unlike the model for wild-type retina, the additive contributions of CRX sites did not show a strong dependence on affinity. Negative homotypic interactions remain in Crx-/- retina, though they are attenuated for CRX sites. Notably, the positive interaction between CRX and NRL sites was absent, indicating that the interaction between these two sites depends specifically on CRX and NRL, and that other TFs that bind these sites in Crx-/- retina do not interact. Despite the loss of positive cooperativity between CRX and NRL sites, the model suggests that synthetic CRSs in Crx-/- retina maintain or increase their activity due to stronger additive contributions of lower affinity binding sites and a modest attenuation of negative homotypic interactions between CRX sites. The negative homotypic interactions with the Rho basal promoter are likely attenuated as well, though we were unable to explicitly model the effects of the Rho promoter because there was no data from a library without the Rho promoter. Taken as a whole, the model suggests that cooperative interactions depend on the specific identities of the TFs involved, while the positive additive and negative homotypic effects hold more generally among TFs, though with varying effect sizes.

4.2.3 Additional retinal TFs contribute independently to CRS activity

CRX and NRL are critical for establishing rod photoreceptor identity, and together they drive high expression of many key rod photoreceptor genes [22, 151, 155, 161, 163]. However, a cooperative interaction between CRX and NRL is not sufficient to explain the context-dependent effects of CRX sites in enhancers, because most CRX-bound enhancers do not contain a copy of the NRL motif [136, 159]. To investigate how other TFBSs contribute to the activity of CRSs that contain CRX sites, we

designed a new library of 6,600 synthetic CRSs. The library included TFBSs for CRX, NRL, and three additional TFs expressed in photoreceptors, NEUROD1, RORB, and MAZ [158, 164, 165]. We previously found that motifs for these TFs were enriched in CRX-bound enhancers, but not silencers [136]. We sought to discover whether these additional TFBSs interact cooperatively with CRX sites, or whether they contributed independently to enhancer activity. We designed synthetic CRSs composed of five sites and we systematically varied the TFBSs composition across the library. Each CRS included either two or three CRX sites and one or two sites for each of two additional TFs (Figure 4.3A). Synthetic CRSs were cloned upstream of the Rho minimal promoter and tested by MPRA with three replicate transfections in explanted retinas (mean R2 between replicates = 0.950, Figure A.9A). We used the data to train different models and again found that the pairwise interaction model performed better than the additive or nearest-neighbor models (predictive information = 2.65). No additional performance was gained from the black box model (Figure A.9B). The pairwise model captured most of the variance in CRS activity (Figure 4.3B and C, R2 between predicted and observed expression = 0.970).

Examining the average additive effects of TFBSs in the pairwise interaction model, we found that higher affinity sites for NRL, NEUROD1, RORB, and MAZ contributed positively to activity, while lower affinity sites had weaker positive effects or negative effects on activity (Figure 4.3D). High affinity sites for NEUROD1 and RORB in particular made strong additive contributions to activity. In contrast to the models above, the additive contributions of CRX sites in this model were negative. This is likely due to the design of the library, which only includes CRSs with two or three CRX sites, making it difficult to deconvolve additive effects of individual CRX sites from the effects of negative homotypic interactions between CRX sites. The interaction terms of the model exhibit a mixture of moderate positive and negative effects that depend on the affinity and order of the TFBSs (Figs. 3D and S3C). We looked specifically for interactions between other TFs and CRX by examining the interaction terms between high affinity CRX sites and each type of TFBS (Figure 4.3E). As with the prior model, there were strong, affinity-dependent negative homotypic

interactions between CRX sites. High affinity NRL sites interacted positively with CRX sites only in some positions. Sites for RORB showed consistent positive interactions with CRX sites across positions, while sites for NEUROD1 and MAZ generally did not interact with CRX sites. The results of this model suggest that, while CRX does cooperatively interact with some TFs, such interactions are not necessary to overcome the repressive effects of negative homotypic interactions. Instead, the additive contributions from diverse TFBSs can shift the balance towards activation.

4.2.4 Balance between positive and negative interactions can explain context-dependent effects of binding sites for transcriptional activators

The MAVE-NN models suggest that the context dependency of sites for transcriptional activators like CRX and NRL can be explained by the balance between the negative effects of homotypic interactions and the positive effects of individual TFBSs and heterotypic interactions between them. These effects create context dependence without the need for repressor TFBSs or dual-function TFs with distinct activation and repression domains. Instead, the results presented above suggest that some transcriptional activators self-inhibit when present at higher occupancy on a CRS. The negative effects of self-inhibition can be overcome by positive cooperativity with a different TF, or by the non-cooperative action of a diverse collection of TFs that, as a collective, engage in fewer negative homotypic interactions. Under this model, the TFBS composition at enhancers and silencers shifts the balance between these effects in favor of either activation or repression. At enhancers, positive cooperativity and the independent contributions of diverse activator TFBSs outweigh the effects of negative homotypic interactions, while at silencers negative homotypic interactions predominate.

To demonstrate how context dependence is achieved under such cis-regulatory grammar of balanced effects, we implemented a simplified model that expresses CRS activity as the sum between positive and negative contributions of activator TFBSs. This model recapitulates the non-monotonic relationship we observed between the number CRX sites and CRS activity (Figure 4.1B). In the

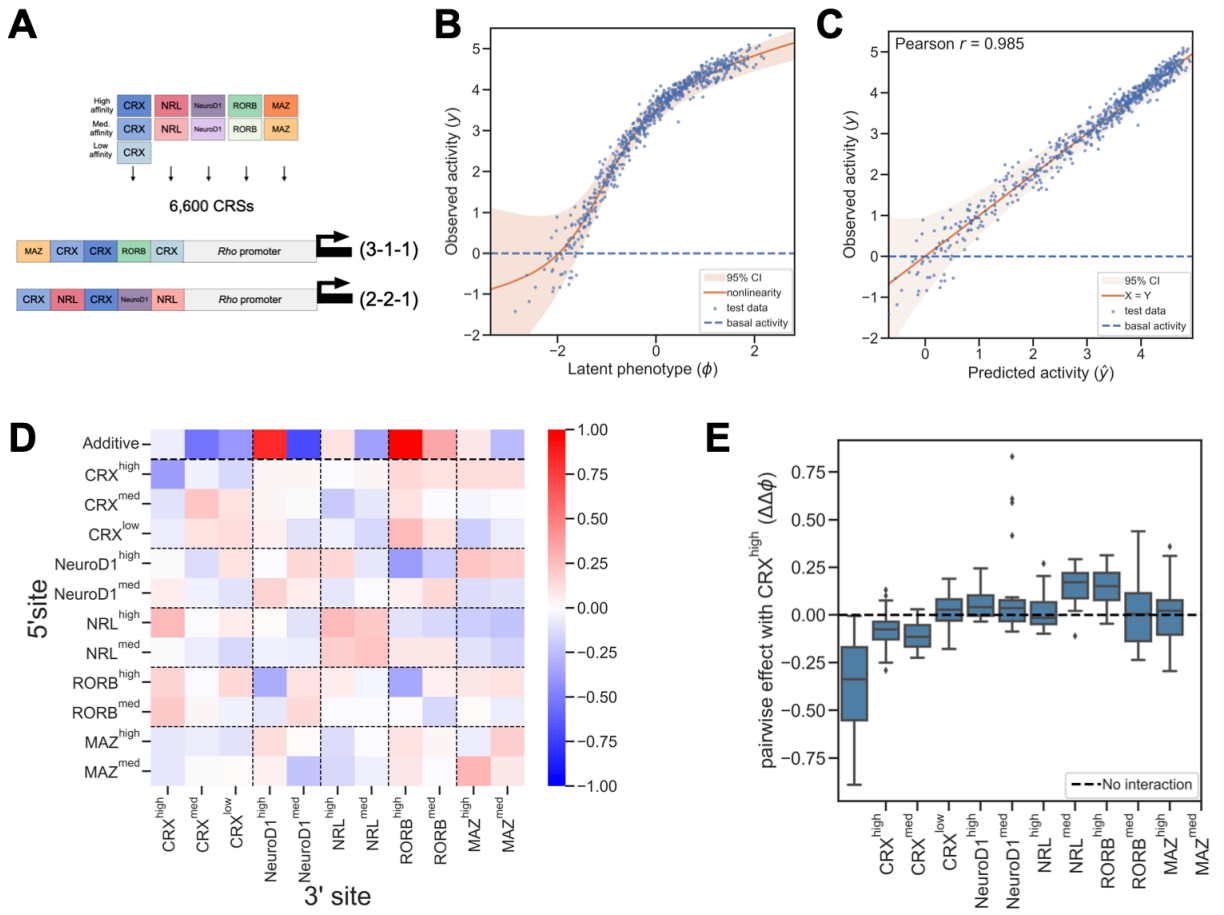

Figure 4.3: A model of cis-regulatory activity driven by diverse TFBSs in wild-type mouse retina. (A) Design of MPRA library of synthetic CRSs with additional lineage-specific TFBSs. CRSs contained five sites placed adjacent to the Rho basal promoter. Each CRS contained either three CRX sites and two sites for other TFs (3-1-1) or two CRX sites, two sites for another TF, and one site for a third TF (2-2-1). (B) Observed activity (y-axis) of test set sequences compared to the latent phenotype (x-axis) predicted by the pairwise model. (C) Observed activity (y-axis) of test set sequences compared to the activity predicted by the pairwise model (x-axis). (D) Model parameters representing additive and pairwise contributions of TFBSs averaged across positions. (E) Distributions of interactions with high-affinity CRX binding sites across all positions, broken down by partner TF.

model, we assume that CRS activity is the sum of (1) positive, additive contributions from sites for transcriptional activators, (2) positive cooperativity between sites for different TFs, and (3) negative interactions between sites for the same TF. For CRSs composed only of sites for two different TFs, as in Figure 4.1, this sum is

$$A = \alpha_x x + \alpha_y y + \beta_{xy} xy - \frac{\gamma_x x(x-1) + \gamma_y y(y-1)}{2} \quad (1)$$

where A is activity of a CRS, x is the number of sites for the first TF, y is the number of sites for the second TF, and α, β, γ are weights reflecting the relative strength of each contribution to activity. The first two terms represent the additive contribution of each TFBS, the third term represents positive cooperativity between all pairs of sites for different TFBSs, and the final term represents negative interactions between all pairs of sites for the same TF.

We calculated the expected activities of all possible CRSs with up to four sites (Figure 4.4A), making the simplifying assumption that the relative strengths of the different terms in eq. 1 are similar and setting all weights equal to 1. The simulated activities recapitulate patterns of expression observed in the library of synthetic CRSs with combinations of CRX and NRL sites. Starting with the basal promoter alone (zero TFBSs), increasing the number of sites for a single TF leads first to an increase and then a decrease in activity (leftmost column or top row in Figure 4.4A, compare with Figs. 1B and S1). The highest activities are obtained from CRSs with combinations of sites from both TFs. In the model, a CRS with four CRX sites is repressive. Replacing one of those sites with an NRL site converts the CRS from a silencer to an enhancer, an effect also observed in the data (Figs. 4A and 1B.) While our model relies on simplifying assumptions that are unlikely to fully hold in vivo, it successfully recapitulates the major trends observed in our data.

We also modeled the effects of TFBS diversity in the absence of cooperative interactions. The MAVE-NN model of synthetic CRSs with five TFBSs suggests that a diversity of TFBSs can shift the balance in favor of activation, even without cooperative interactions (Figure 4.3D). For a CRS composed of a given number of sites, greater TFBS diversity reduces negative homotypic interactions. As a result, the independent positive effects of each TFBS predominate. To model these

effects, we assumed that (1) bound TF activators always make positive, independent contributions to activity and (2) all TFs engage in negative homotypic interactions. We calculated the sums of positive and negative effects for CRSs with five total TFBSs, but different numbers of CRX sites (and therefore varying amounts of negative homotypic interactions). In these simulated CRSs, the total additive contribution is constant and equal to the total number of TFBSs (Figure 4.4B, blue circles). As the diversity of the TFBSs increases, the number of negative homotypic interactions is reduced (Figure 4.4B, orange crosses). A CRS with five CRX sites is therefore highly repressive, while replacing some CRX sites with different TFBSs increases the activity of the CRS (Figure 4.4B, green squares). This simplified model demonstrates how strong activity can be achieved by the independent effects of diverse TFBSs, even in the absence of cooperative interactions. This model suggests an explanation for our prior observation that CRX-bound strong enhancers have more diverse TFBSs than CRX-bound silencers [136].

4.3 Discussion

Because the effects of a TFBS often strongly depend on local sequence, the activity of cis-regulatory DNA is not a simple function of TFBS composition. To accurately predict the activities of cis-regulatory sequences and the effects of genetic variants that occur within them, we need models of cis-regulatory grammar that accurately account for the influence of sequence context. Different occurrences of a TF binding motif can differ in their effects due to post-translational modifications of TFs [166, 167], the presence of other co-bound factors [91, 125, 126, 129, 131, 137, 145, 168–171], and binding by different TFs with similar sequence specificities [130, 172, 173]. Our results suggest that context dependence can also be determined by the overall balance between independent and interaction effects of individual sites for transcriptional activators. At the core of this model is a distinction between additive, independent effects of individual TF molecules and effects of interactions between molecules. In the case of CRX, the independent and interaction effects influence cis-regulatory activity in opposite directions, with CRX molecules independently

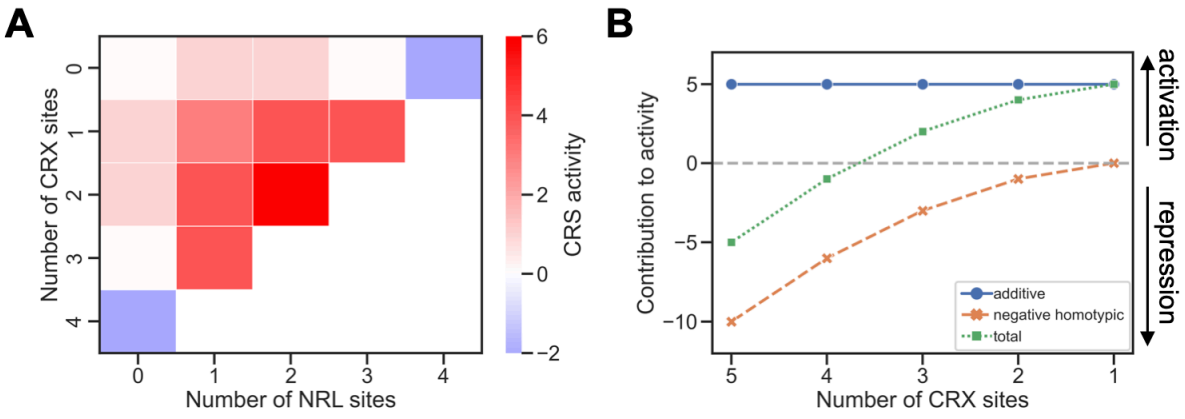

Figure 4.4: Simplified balance model of context-dependent effects of binding sites for transcriptional activators. (A) Simulated CRS activities calculated by eq. 1 for sequences with up to four TFBSs for CRX or NRL. Stepwise addition of sites for a single TF first increase then decrease activity. The first two columns show predicted expression of CRSs with only CRX sites or with CRX sites plus one NRL site. Compare with the measured values in Figure 4.1B. (B) Model of CRSs with five TFBSs shows how TF diversity reduces negative homotypic interactions and increases CRS activity. As CRX sites are replaced with sites for different TFs, TF diversity increases (x-axis) and the number of negative homotypic interactions decreases (orange crosses) and the overall CRS activity increases (blue squares). The total additive contribution of TFBSs (green circles) is equal to the total number of TFBSs and remains constant.

contributing to activation while engaging in repressive homotypic interactions with one another. The independent, activating effects scale linearly with the number of binding sites, while the number of repressive homotypic interactions scales with the square of binding site number. As the number of binding sites in a CRS increases, negative homotypic interactions grow faster than the activating effects of individual binding sites. As a result, sequences with many CRX sites are likely to act as silencers, a pattern that we observe with both synthetic CRSs and genomic CRX-bound sequences [19, 136].

In this model, other TFs can influence whether a CRX-bound CRS will activate or repress transcription in two ways. First, a TF like NRL may form positive cooperative interactions with CRX. Positive cooperative interactions also scale with the square of the numbers of binding sites, thus the addition of relatively few sites for a cooperating TF can shift the balance towards activation. This explains our observation that, for a repressive CRS with four CRX sites, converting one position to an NRL site changes activity of the CRS from repression to activation (Figure 4.1B). This effect in synthetic CRSs is consistent with our previous finding that genes near CRX-bound loci that are co-bound by NRL are more highly expressed than genes near regions bound by CRX alone [19]. In the second way, the presence of other bound TFs at a CRS can also overcome negative homotypic effects by making strong additive contributions, without cooperativity, as in the case of NEUROD1. At a CRS, having fewer sites for a single TF and more sites for a variety of TFs minimizes negative homotypic interactions and allows the additive effects of TFBSs to predominate. The repressive effects of homotypic interactions can thus be overcome by positive cooperativity, or by the presence of diverse TFBSs.

This model suggests an explanation for our prior observation that CRX-bound enhancers often have more diverse TFBSs than CRX-bound silencers [136]. Genomic sequences that act as silencers tend to contain clusters of multiple CRX sites and few sites for other TFs, while genomic enhancers tend to contain sites for a variety of photoreceptor TFs [19, 136, 159]. These patterns are recapitulated in synthetic CRSs whose TFBS composition is systematically varied. We previously quantified

the role of TFBS diversity in distinguishing silencers from enhancers using a phenomenological metric that we called information content, which considers the number and diversity of TFBSs in a CRS [136]. High information content sequences have a greater number and diversity of TFBSs and are more likely to be enhancers, while silencers are often low information content because they lack TFBS diversity. Our MAVE-NN models suggest that high information content sequences are enhancers because binding site diversity allows the additive contributions of TFBSs to outweigh negative homotypic interactions. This is in contrast to low-diversity silencers, which our model suggests are dominated by negative homotypic interactions. TFBS diversity is a feature of enhancers in other cell types, where a similar balance between positive additive effects and negative homotypic interactions may occur [174].

In our models, negative homotypic interactions strongly influence the context-dependent effects of binding sites for several photoreceptor TFs, particularly CRX and NRL. The existence of such interactions is supported by data from both synthetic CRSs and genomic sequences. Sequential addition of CRX or NRL binding sites upstream of the Rho basal promoter first increases, then decreases transcription, sometimes below basal levels (Figure 4.1B and A7) [19]. Genomic CRX-bound sequences that act as silencers when measured by MPRA have more copies of the CRX motif than sequences that act as enhancers [19, 136, 159]. We have shown that this silencing activity depends on both CRX motifs and CRX protein [19]. Similar negative homotypic effects have been reported for several TFs, including liver specific factors [175], yeast Gcn4 [176], pluripotency TFs [134, 177], and Sp3 [178]. These findings suggest some TFs may self-inhibit or recruit repressors when they highly occupy certain CRSs [138, 179, 180]. The homeodomain TF WUSCHEL activates transcription as a monomer at low concentration, but forms repressive dimers at higher concentration [181, 182]. A recent study found that homeodomain TFs in particular are enriched in activation domains that also exhibit the ability to repress [18]. Our model of context dependency suggests that the balance between positive effects and negative homotypic interactions can account for the dual activities of some TFs, without the need for dedicated repressors.

4.4 Methods

4.4.1 Model fitting

For the CRX + NRL library, binding site arrangements and MPRA activities were extracted from Database S3 of [19]. For the library with CRX, NEUROD1, NRL, RORB, and MAZ sites (CDNRM library), the MPRA experiment was performed as described below. Data files are described below under Data availability. To encode arrangements of TFBSs as input sequences for MAVE-NN, we used single letters to represent each type of binding site. To create input sequences of uniform length for the CRX + NRL library, dummy binding sites labeled "O" were prepended to each arrangement to render all CRSs four sites long, and an additional letter indicating the basal promoter (Rho or Hsp68) was then appended to the end. Models were trained using mave-nn package version 1.01 until convergence on the processed data, using hyperparameters given in Table 1. The models were specified using the Skewed-T GE noise model with a heteroskedasticity order of 2. We used the consensus gauge with basal Hsp68 as the consensus sequence to obtain parameters from the models trained on CRX-NRL data in wild-type retina, and the uniform gauge for the remaining models. To ensure consistent training outcomes, we trained each model from a large number of random initializations (25 for the CRX + NRL library in wild-type retina, 20 for the library in Crx-/- retina, and 50 for the CDNRM library), with the numbers chosen to achieve maximum performance and reproducibility. We picked the best-performing model of each type for further evaluation. Model performance was evaluated by cross-validation with an 80-10-10 percent training-validation-test set split. The measurements were split randomly between sets and the same split was used for all random initializations.

4.4.2 CDNRM library design

We designed a library of 6600 synthetic CRSs composed of combinations of binding sites for CRX, NRL, NEUROD1, RORB, and MAZ. The library was designed to vary TFBS diversity

around CRX sites. It contained all possible arrangements of either 3 sites for CRX and 1 site each for two other TFs (3-1-1 sequences); or 2 sites for CRX, 2 sites for a second TF, and 1 site for a third TF (2-2-1 sequences). CRX sites in a CRS were either all high affinity or a mixture of affinities. TFBS orientation was held constant. High, medium and low affinity CRX sites were those used in the CRX + NRL library [19]. The NRL, NEUROD1, RORB, and MAZ motifs were randomly selected from sites identified in genomic strong enhancers that lose activity when the corresponding motif is deleted [136]. Note that NRL sites in the CDNRM library vary from the site in the CRX + NRL library, but all sites for the synthetic CRSs have been shown to be active in genomic sequences. Binding site sequences were padded to make all motifs 12 bp, then a constant buffer sequence was added (AGCTAC<motif>GT) to create a 20 bp "building block" that maintains helical spacing when sites were combined, similar to our procedure for prior libraries of synthetic CRSs [13,16,65]. The 12 bp motifs used with the core motif underlined, are: high affinity CRX, TGCTAATCCCAC; medium affinity CRX, TGCTAAGCCAAC; low affinity CRX, TGCTGATTCAAC; high affinity NRL: AATTTGCTGACC; medium affinity NRL, GGCCTGCTGACC; high affinity NEUROD1, CAACAGATGGTA; medium affinity NeuroD1, CGGCAGGTGGTA; high affinity RORB, AATTAGGTCACT; medium affinity RORB, ATCTGGGTCAGT; high affinity MAZ, GGGGGAGGGGGG; medium affinity MAZ, GCGGGCGGGGGG.

4.4.3 MPRA library cloning

Synthetic CRSs were each represented in the library with 3 unique barcodes. As standards, the library included 20 genomic sequences taken from [18] that span the dynamic range of the MPRA and 150 scrambled sequences as negative controls. The Rho basal promoter was tagged with 90 barcodes to ensure precise measurement of basal levels. Barcoded CRSs were synthesized as two sub-libraries on a single chip using custom oligonucleotide synthesis from Agilent Technologies. The oligonucleotide libraries were cloned as previously described [18]. Briefly, we amplified oligos using either primer pairs MO563 (GTAGCGTCTGTCCGTGAATT) and MO564 (CTGTAGTAGTAGTTGGCGGC)

or RZFP3 (TCTAGACTGCGGCTCGAATT) and RZFP4 (AGATCTAATGCATACGCGGC), and cloned them into the vector pJK03 (AddGene #173,490). The rod-specific Rho promoter, the DsRed reporter gene, and a multiplexing barcode (mBC) was cloned between the synthetic sequence and the cBC. One sub-library was assigned mBC TAGTAACGG, the other was assigned CCTACTAGT. The final plasmid libraries were mixed together at equimolar concentrations.

Retinal explant electroporation: Animal procedures were performed in accordance with a Washington University in St Louis Institutional Animal Care and Use Committee-approved vertebrate animals protocol. Electroporations into retinal explants into P0 CD-1 mice and RNA extractions were performed as described previously [13,18,42,49]. We performed three replicate electroporations. cDNA and the input plasmid pool was sequenced on the Illumina NextSeq platform. We obtained an average sequencing depth of >675 reads per barcode.

4.4.4 MPRA data processing

Sequencing reads were filtered to retain only perfect barcode matches. After filtering we retained 95% of sequencing reads. Barcodes with fewer than 50 reads in the plasmid pool were considered missing and removed. Barcode read counts were normalized by total sample reads. MPRA activity scores for each replicate were calculated by dividing RNA by DNA values, averaging across barcodes for each CRE, then normalizing to the activity of the basal promoter [18]. Replicates were averaged and the log2 transformed values were used for model training.

Chapter 5: Conclusions

5.1 Repression is heterogeneous and involves complex interactions, while activation is homogeneous and additive

In the studies I have described in the previous chapters, we have uncovered a great degree of complexity and context-dependency in the activities of transcription factors. One of the most striking aspects of our findings is the asymmetry between activation and repression, with each manifesting a different form of complexity. More specifically, activators are homogeneous in terms of noise, but act upon multiple rate constants. Furthermore, activation can occur through specific cooperative interactions. On the other hand, in our retinal MPRA data, repression primarily occurred through homotypic interactions between large numbers of adjacent binding sites. However, in terms of noise, we observed clear differences in activity and thus mechanism between different repressors, quite unlike what we saw with activators. In the following section, I will propose a theoretical framework to explain these seemingly disparate observations, and conclude by arguing that an understanding of noise will be necessary to model gene regulation even if modeling noise is not the immediate objective.

5.2 Potential mechanisms for repression by an excess of activator

One of our most counterintuitive observations is that transcription factors that act as activators when bound individually frequently act as repressors at high occupancy on the DNA. We suspect that this effect is not specific to a particular transcription factor, given that we observed this phenomenon in

both TFs (CRX and NRL) for which we were powered to detect homotypic interactions. Additionally, multiple studies [181, 183] have observed repressive homotypic interactions, with Mukund et al [183] even observing it generally among the strongest activation domains. There are a number of imaginable explanations for this effect. One possible explanation for the MPRA results is that the "CRX" binding sites in fact bind multiple transcription factors with differing degrees of homotypic cooperativity. The effects of more cooperative transcription factors will thus increasingly predominate as the number of sites increases, leading to the observed switch if the more cooperative TFs are repressors and the less cooperative TFs are activators. However, the switch between activation and repression would be expected to be confined to a specific set of TF binding sites under this model, contrary to the observations of this phenomenon in multiple settings. Additionally, Mukund et al [183] observed this switch in a targeting system using combinations of just effector domains, a setting where differential DNA binding has been taken out of the picture. Switching between activation and repression has also been observed [18] in the context of targeting single effector domains to different promoters, again ruling out mechanisms involving DNA binding.

If the switch between activation and repression does not depend on differential DNA binding, it must be occurring after DNA binding and mediated by transcriptional effector domains. One can easily imagine a similar differential binding mechanism occurring after DNA binding. To elaborate, it is easily to imagine differential cooperativity in recruitment of different cofactors by effector domains, such that repressive cofactors are primarily recruited by multiple effector domains acting cooperatively, while activating cofactors can be recruited by effector domains acting independently. This would thus promote activation at low transcription factor occupancy and repression at high occupancy. This model can be further extended to encompass phase separation. Under this version of the model, individual effector domains would again recruit activating cofactors, while a sufficient number of effector domains in close proximity would nucleate a condensate, with corepressors preferentially partitioning into such liquid drops, thus leading to repression.

Another possible explanation is rooted in kinetics. Galburt [184] has shown that factors with no

DNA binding specificity can have site-specific effects on transcription based on the kinetic landscape of a promoter. This is because many mechanisms of action, such as catalyzing DNA opening or stabilizing the open complex, increase the rates of both forward and reverse reactions within the transcriptional process. Thus, depending on the rates in the absence of the factor, introducing the factor can either increase or decrease the overall rate of transcription depending on which step(s) are rate-limiting. A concrete example of this can be derived from the standard recruitment model of TF activity [185] in which an activation domain interacts (with cofactors which in turn interact) with polymerase, thus bringing it to the transcription start site. While this might increase transcription if the activator is present at low occupancy, excessive binding of polymerase to the activation domains could prevent it from escaping the promoter, thus reducing transcription when occupancy of the transcription factor is high. Importantly, Galburt shows [184] that the transcription factor must impact multiple rates for this switching to occur. This is consistent with our results from the yeast reporter system, with modulation of activator occupancy through nuclear concentration consistently inducing mean-variance relationships suggestive of changes in multiple kinetic parameters, regardless of the specific activation domain used.

5.3 Potential explanations for the greater complexity of repression

Another striking finding from our experiments is the marked asymmetry between activators and repressors in isolation, with activators having qualitatively homogeneous effects on noise consistent with a complex shared mechanism and repressors producing sharply distinct effects on noise that can be straightforwardly modeled. A partial explanation for this difference lies in the inherent nature of activation and repression. Because functional transcription initiation is confined to a set of defined sites that is small compared to the genome, activators can act autonomously, while repression is meaningful only in the context of activating sequence features. This inherent disparity points to

a potential evolutionary explanation for the differences between activation and repression. Since activation can occur independently of any other factor binding, there might be selective pressure for activators to preserve autonomous activity, constraining the space of mechanisms available to them. Repressors, on the other hand, because the sites where they are functionally significant always have activators bound as well, are free to evolve context-dependent mechanisms of action. This helps to explain the results from our yeast reporters. While autonomously acting activators must act directly upon transcription initiation, repressors can act by excluding activator binding, opening up a mechanism with strongly distinct effects on noise.

This framework can also help explain how repressive homotypic interactions can become widespread. Multifunctional transcription factors are particularly free to evolve complex interactions. Because the different domains and activities contained within the same protein are always present together, mutations in one can be compensated by changes in the other components. This allows the different components to coevolve, and thus develop highly interdependent mechanisms. The interactions that evolve between domains and functions could then occur between adjacent molecules of the same factor as well, leading to complex homotypic interactions.

Besides the neutral mechanism just discussed, another factor potentially driving the evolution of homotypic repressive interactions is selection for signal responsive elements to be repressed under basal conditions. In the absence of signal, signaling pathways tend to exert a repressive effect upon CRSs bound by their associated signal-responsive TFs [16]. This has the function of ensuring that associated constitutive TFs that provide cell type specificity do not inappropriately activate transcription in the absence of signal. In the presence of signal, cooperative interactions between signal-responsive and cell type-specific TFs create highly specific patterns of gene expression. Selection for such regulatory logic will naturally favor the development of positive heterotypic interactions. Negative homotypic interactions might also be selected for because they help to prevent inappropriate activation by a single TF in the absence of additional cooperating TFs.

5.4 Models of noise and nonlinearity are necessary to gain a predictive understanding of gene expression

Theory and practical experience both indicate that modeling noise and characteristics of the experimental measurement process is a vital part of any attempt to gain a predictive understanding of gene expression. As related by Atwal and Kinney [27], maximum likelihood inference is likely to give incorrect results if an incorrect noise model is used. To achieve accurate inference, the noise model must either be correctly specified ahead of time or learned alongside the model of deterministic effects using mutual information as the objective. This comports with practical experience. It is widely accepted in the field [186, 187] that correcting for batch effects or experimental variability is an essential part of any bioinformatic analysis. In our own experience [Barak Cohen, personal communication], all models prior to MAVE-NN had failed to fit the building block MPRA data used here. This includes both biophysical models and standard machine learning architectures, suggesting that this is not due to an inability to capture deterministic biological effects. Rather, what sets MAVE-NN apart from previous approaches to modeling is that it explicitly models experimental noise and nonlinearity and learns it alongside the genotype-phenotype mapping, thus implementing the mutual information objective Atwal and Kinney [27] describe. The necessity of modeling these features of the data has multiple implications for future analyses as new MPRA methods are developed to probe the effects of ever more complex sequence features on ever more complex phenotypes.

Nonlinear activation functions are a widespread motif in biology. Both saturating and sigmoid curves are common in biochemistry, with Michaelis-Menten and Hill functions arising from simple physical chemical models. Both of these curves are commonly fit to biochemical data to recover biophysical parameters. While there have been some recent publications [188, 189] that fit binding curves to genome-wide data from ChIP or Cut and Tag, and the Garcia lab has done extensive work using input-output functions as the organizing framework [190–192], there has been relatively

little exploration of global nonlinearities in cis-regulatory activity. However, there is evidence that such nonlinearities might be a useful conceptual tool for understanding a broad range of regulatory phenomena. For example, massively parallel assays of enhancer-promoter compatibility have consistently found that the relationship between intrinsic enhancer and promoter activity and expression is predominantly multiplicative, with a relatively small residual component [193, 194]. This residual component has been explained differently by different authors. Bergman et al [194] describe the residual as a specific interaction between two classes of enhancers and two classes of promoters. Hong et al, on the other hand, propose that the deviations from multiplicative scaling arise from detection limits and saturation imparting a sigmoidal character to reporter activity. A fine-grained examination of the results from Bergman et al [194] also lends support to this model. Close examination of the features distinguishing the proposed classes of enhancers and promoter reveals a preponderance of features, like accessibility, H3K27 acetylation, and activity in MPRAs and CRISPR screens, that can be interpreted as proxies for overall level of activity. This suggests that the deviations from the multiplicative relationship in Bergman et al's [194] data might also arise from a global nonlinearity superimposed upon an simple underlying principle. Approaches to modeling like MAVE-NN that explicitly account for global nonlinearity hold promise for untangling specific interactions from global epistasis due to nonlinearity in experimental measurements.

Noise is another ubiquitous biological phenomenon that must be accounted for in order to obtain accurate models. Until recently, noise in MPRA data has been largely an experimental artifact because the assays have been performed in bulk. However, single cell MPRAs (scMPRAs) have recently been developed to probe cis-regulatory effects on cell type- and state-specific gene expression and cell-to-cell variability [114, 195, 196]. By the nature of these assays, their output exhibits both technical and biological noise. This implies that any attempt to model single-cell gene regulation will have to account for the intrinsic noise in gene expression, even if the objective is to understand cell-type specific regulation and not specifically noise. Because, as we have seen here, it is possible to decouple mean and noise, studies that only probe effects on mean expression will not

be sufficient to derive an adequate noise model for single-cell data. Therefore, understanding gene regulation in its cellular context will require a dialog between theory and experiment that treats noise and nonlinear input-output relationships not as epiphenomena, but inherent features of biology. Approaches to modeling designed with such a dialog in mind will be vital tools for understanding gene regulation.

References

1. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* (2006) (cited on p. 1).
2. Davidson, E. H. The regulatory genome: gene regulatory networks in development and evolution. *Elsevier* (2010) (cited on p. 1).
3. Gazal, S., Loh, P. R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S. & Price, A. L. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature genetics* (2018) (cited on p. 1).
4. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* (2021) (cited on p. 1).
5. Raj, A., Rifkin, S. A., Andersen, E. & Van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* (2010) (cited on p. 1).
6. Castel, S. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature genetics* (2018) (cited on p. 1).
7. Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics* (2000) (cited on p. 1).
8. Stormo, G. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics* (2010) (cited on p. 1).
9. Lim, F., Ryan, G., Le, S., Solvason, J., Steffen, P. & Farley, E. Affinity-optimizing variants within the ZRS enhancer disrupt limb development. *bioRxiv* (2022) (cited on p. 1).
10. Jindal, G. *et al.* Affinity-optimizing variants within cardiac enhancers disrupt heart development and contribute to cardiac traits. *bioRxiv* (2022) (cited on p. 1).
11. Serrano-Saiz, E., Poole, R. J., Felton, T., Zhang, F., De La Cruz, E. D. & Hobert, O. Modular control of glutamatergic neuronal identity in C. elegans by distinct homeodomain proteins. *Cell* (2013) (cited on p. 1).
12. Hobert, O. Terminal selectors of neuronal identity. *Current topics in developmental biology* (2016) (cited on p. 1).
13. Hobert, O. & Kratsios, P. Neuronal identity control by terminal selectors in worms, flies, and chordates. *Current opinion in neurobiology* (2019) (cited on p. 1).

14. Hobert, O. Homeobox genes and the specification of neuronal identity. *Nature Reviews Neuroscience* (2021) (cited on p. 1).

15. Leyva-Diaz, E. & Hobert, O. Robust regulatory architecture of pan-neuronal gene expression. *Current Biology* (2022) (cited on p. 1).

16. Barolo, S. & Posakony, J. W. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes and development* (2002) (cited on pp. 1, 2, 35, 58).

17. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics* (2017) (cited on p. 2).

18. DelRosso, N. *et al.* Large-scale mapping and systematic mutagenesis of human transcriptional effector domains. *bioRxiv* (2022) (cited on pp. 2, 3, 29, 51, 56).

19. White, M. A., Kwasnieski, J., Myers, C., Shen, S., Corbo, J. & Cohen, B. A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Reports* (2016) (cited on pp. 2, 3, 35–39, 50–53, 84).

20. Staller, M. V., Holehouse, A. S., Swain-Lenz, D., Das, R. K., Pappu, R. V. & Cohen, B. A. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell systems* (2018) (cited on pp. 3, 16, 18).

21. Corbo, J. C., Myers, C. A., Lawrence, K. A., Jadhav, A. P. & Cepko, C. L. A typology of photoreceptor gene expression patterns in the mouse. *Proceedings of the National Academy of Sciences* (2007) (cited on p. 3).

22. Hao, H. *et al.* Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis. *PLoS Genetics* (2012) (cited on pp. 3, 36, 43).

23. Tareen, A. & Kinney, J. Biophysical models of cis-regulation as interpretable neural networks. *arXiv* (2019) (cited on pp. 3, 4).

24. Tareen, A., Kooshkbaghi, M., Posfai, A., Ireland, W., McCandlish, D. & Kinney, J. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology* (2022) (cited on pp. 3, 36, 39).

25. Koo, P. & Eddy, S. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS computational biology* (2019) (cited on p. 4).

26. Koo, P. & Ploenzke, M. Improving convolutional network interpretability with exponential activations. *bioRxiv* (2019) (cited on p. 4).

27. Atwal, G. & Kinney, J. Learning Quantitative Sequence–Function Relationships from Massively Parallel Experiments. *Journal of Statistical Physics* (2016) (cited on pp. 4, 39, 59).

28. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* (2002) (cited on p. 5).

29. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS biology* (2006) (cited on p. 5).

30. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* (2008) (cited on pp. 5, 11, 22).

31. Kepler, T. B. & Elston, T. C. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal* (2001) (cited on p. 5).

32. Levchenko, A. & Nemenman, I. Cellular noise and information transmission. *Current opinion in biotechnology* (2014) (cited on p. 5).

33. Rieckh, G. & Tkačik, G. Noise and information transmission in promoters with multiple internal states. *Biophysical journal* (2014) (cited on p. 5).

34. Miller, C. A. & Beard, D. A. The effects of reversibility and noise on stochastic phosphorylation cycles and cascades. *Biophysical journal* (2008) (cited on p. 5).

35. Bishop, L. M. & Qian, H. Stochastic bistability and bifurcation in a mesoscopic signaling system with autocatalytic kinase. *Biophysical journal* (2010) (cited on p. 5).

36. Hsu, C., Jaquet, V., Maleki, F. & Becskei, A. Contribution of bistability and noise to cell fate transitions determined by feedback opening. *Journal of molecular biology* (2016) (cited on p. 5).

37. Kellogg, R. A. & Tay, S. Noise facilitates transcriptional control under dynamic inputs. *Cell* (2015) (cited on p. 5).

38. Gupta, A., Hepp, B. & Khammash, M. Noise induces the population-level entrainment of incoherent, uncoupled intracellular oscillators. *Cell systems* (2016) (cited on p. 5).

39. Heltberg, M., Kellogg, R. A., Krishna, S., Tay, S. & Jensen, M. H. Noise induces hopping between NF-κB entrainment modes. *Cell systems* (2016) (cited on p. 5).

40. Meng, J. H. & Riecke, H. Synchronization by uncorrelated noise: interacting rhythms in interconnected oscillator networks. *Scientific reports* (2018) (cited on p. 5).

41. Raj, A. & Van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* (2008) (cited on p. 5).

42. Gandrillon, O., Kolesnik-Antoine, D., Kupiec, J. J. & Beslon, G. Chance at the heart of the cell. *Progress in biophysics and molecular biology* (2012) (cited on p. 5).

43. Paré, A., Lemons, D., Kosman, D., Beaver, W., Freund, Y. & McGinnis, W. Visualization of individual Scr mRNAs during Drosophila embryogenesis yields evidence for transcriptional bursting. *Current Biology* (2009) (cited on p. 5).

44. Singh, A., Razooky, B., Cox, C. D., Simpson, M. L. & Weinberger, L. S. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical journal* (2010) (cited on p. 5).

45. Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A. & Itzkovitz, S. Bursty gene expression in the intact mammalian liver. *Molecular cell* (2015) (cited on p. 5).

46. Chubb, J. R. & Liverpool, T. B. Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Current opinion in genetics and development* (2010) (cited on p. 5).

47. Dar, R., Razooky, B., Singh, A., Trimeloni, T., McCollum, J., Cox, C., Simpson, M. & Weinberger, L. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences* (2012) (cited on p. 5).

48. Dar, R. D., Shaffer, S. M., Singh, A., Razooky, B. S., Simpson, M. L., Raj, A. & Weinberger, L. S. Transcriptional bursting explains the noise–versus–mean relationship in mRNA and protein levels. *PLoS one* (2016) (cited on p. 5).

49. Tantale, K. *et al.* A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nature communications* (2016) (cited on p. 5).

50. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* (2005) (cited on p. 5).

51. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural and molecular biology* (2008) (cited on p. 5).

52. Symmons, O. & Raj, A. What's luck got to do with it: single cells, multiple fates, and biological nondeterminism. *Molecular Cell* (2016) (cited on p. 5).

53. Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* (2011) (cited on p. 5).

54. Losick, R. & Desplan, C. Stochasticity and cell fate. *Science* (2008) (cited on p. 5).

55. Bell, M. L., Earl, J. B. & Britt, S. G. Two types of Drosophila R7 photoreceptor cells are arranged randomly: A model for stochastic cell□fate determination. *Journal of Comparative Neurology* (2007) (cited on p. 5).

56. Miller, A. C., Seymour, H., King, C. & Herman, T. G. Loss of seven-up from Drosophila R1/R6 photoreceptors reveals a stochastic fate choice that is normally biased by Notch. *Development* (2008) (cited on p. 5).

57. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* (2008) (cited on p. 5).

58. Pina, C., Fugazza, C., Tipping, A., Brown, J., Soneji, S., Teles, J., Peterson, C. & Enver, T. Inferring rules of lineage commitment in haematopoiesis. *Nature cell biology* (2012) (cited on p. 5).

59. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* (2009) (cited on p. 5).

60. Enver, T., Pera, M., Peterson, C. & Andrews, P. W. Stem cell states, fates, and the rules of attraction. *Cell stem cell* (2009) (cited on p. 5).

61. Cross, M. A. & Enver, T. The lineage commitment of haemopoietic progenitor cells. *Current opinion in genetics and development* (1997) (cited on p. 5).

62. Wolff, S. *et al.* Inheritance of OCT 4 predetermines fate choice in human embryonic stem cells. *Molecular Systems Biology* (2018) (cited on p. 5).

63. Van Roon, M. A., Aten, J. A., Van Oven, C. H., Charles, R. & Lamers, W. H. The initiation of hepatocyte-specific gene expression within embryonic hepatocytes is a stochastic event. *Developmental biology* (1989) (cited on p. 5).

64. Fiering, S., Northrop, J. P., Nolan, G. P., Mattila, P. S., Crabtree, G. R. & Herzenberg, L. A. Single cell assay of a transcription factor reveals a threshold in transcription activated by signals emanating from the T-cell antigen receptor. *Genes and development* (1990) (cited on p. 5).

65. Dingemanse, M. A., de Boer, P. A., Moorman, A. F., Charles, R. & Lamers, W. H. The expression of liver-specific genes within rat embryonic hepatocytes is a discontinuous process. *Differentiation* (1994) (cited on p. 5).

66. Walters, M. C., Fiering, S., Eidemiller, J., Magis, W., Groudine, M. & Martin, D. I. Enhancers increase the probability but not the level of gene expression. *Proceedings of the National Academy of Sciences* (1995) (cited on pp. 5, 23).

67. Rouzine, I. M., Weinberger, A. D. & Weinberger, L. S. An evolutionary role for HIV latency in enhancing viral transmission. *Cell* (2015) (cited on p. 5).

68. Singh, A. & Weinberger, L. S. Stochastic gene expression as a molecular switch for viral latency. *Current opinion in microbiology* (2009) (cited on p. 5).

69. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* (2005) (cited on p. 5).

70. Weinberger, L. S., Dar, R. D. & Simpson, M. L. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature genetics* (2008) (cited on p. 5).

71. Schwall, C., Loman, T., Martins, B., Cortijo, S., Villava, C., Kusmartsev, V., Livesey, T., Saez, T. & Locke, J. Tunable phenotypic variability through an autoregulatory alternative sigma factor circuit. *Molecular Systems Biology* (2021) (cited on p. 5).

72. St-Pierre, F. & Endy, D. Determination of cell fate selection during phage lambda infection. *Proceedings of the National Academy of Sciences* (2008) (cited on p. 5).

73. Zeng, L., Skinner, S. O., Zong, C., Sippy, J., Feiss, M. & Golding, I. Decision making at a subcellular level determines the outcome of bacteriophage infection. *Cell* (2010) (cited on p. 5).

74. Dar, R. D., Hosmane, N. N., Arkin, M. R., Siliciano, R. F. & Weinberger, L. S. Screening for noise in gene expression identifies drug synergies. *Science* (2014) (cited on p. 5).

75. Razooky, B. S., Pai, A., Aull, K., Rouzine, I. M. & Weinberger, L. S. A hardwired HIV latency program. *Cell* (2015) (cited on p. 5).

76. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* (2004) (cited on p. 5).

77. Anderson, C. *et al.* Natural variation in stochastic photoreceptor specification and color preference in Drosophila. *eLife* (2017) (cited on p. 5).

78. Waymack, R., Fletcher, A., Enciso, G. & Wunderlich, Z. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. *eLife* (2020) (cited on pp. 5, 23).

79. Ahmad, K. & Henikoff, S. Modulation of a transcription factor counteracts heterochromatic gene silencing in Drosophila. *Cell* (2001) (cited on p. 5).

80. Hensel, Z., Feng, H., Han, B., Hatem, C., Wang, J. & Xiao, J. Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. *Nature structural and molecular biology* (2012) (cited on p. 5).

81. Kalo, A., Kanter, I., Shraga, A., Sheinberger, J., Tzemach, H., Kinor, N., Singer, R., Lionnet, T. & Shav-Tal, Y. Cellular levels of signaling factors are sensed by β-actin alleles to modulate transcriptional pulse intensity. *Cell Reports* (2015) (cited on p. 5).

82. Kafri, P., Hasenson, S. E., Kanter, I., Sheinberger, J., Kinor, N., Yunger, S. & Shav-Tal, Y. Quantifying β-catenin subcellular dynamics and cyclin D1 mRNA transcription during Wnt signaling in single living cells. *eLife* (2016) (cited on p. 5).

83. Senecal, A., Munsky, B., Proux, F., Ly, N., Braye, F., Zimmer, C., Mueller, F. & Darzacq, X. Transcription factors modulate c-Fos transcriptional bursts. *Cell Reports* (2014) (cited on pp. 5, 6).

84. Näär, A. M., Lemon, B. D. & Tjian, R. Transcriptional coactivator complexes. *Annual review of biochemistry* (2001) (cited on p. 6).

85. Govind, C. K., Yoon, S., Qiu, H., Govind, S. & Hinnebusch, A. G. Simultaneous recruitment of coactivators by Gcn4p stimulates multiple steps of transcription in vivo. *Molecular and cellular biology* (2005) (cited on p. 6).

86. Parab, L., Pal, S. & Dhar, R. Transcription factor binding process is the primary driver of noise in gene expression. *PLoS Genetics* (2022) (cited on pp. 6, 23, 28).

87. Pelet, S., Rudolf, F., Nadal-Ribelles, M., de Nadal, E., Posas, F. & Peter, M. Transient activation of the HOG MAPK pathway regulates bimodal gene expression. *Science* (2011) (cited on pp. 6, 15).

88. Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J. & Bentley, D. Three functional classes of transcriptional activation domain. *Molecular and cellular biology* (1996) (cited on p. 6).

89. Brown, S. A., Weirich, C. S., Newton, E. M. & Kingston, R. E. Transcriptional activation domains stimulate initiation and elongation at different times and via different residues. *The EMBO journal* (1998) (cited on p. 6).

90. Duarte, F. M., Fuda, N. J., Mahat, D. B., Core, L. J., Guertin, M. J. & Lis, J. T. Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. *Genes and development* (2016) (cited on p. 6).

91. Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F. & Stark, A. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* (2015) (cited on pp. 6, 35, 48).

92. Vallania, F. L. M., Sherman, M., Goodwin, Z., Mogno, I., Cohen, B. A. & Mitra, R. D. Origin and consequences of the relationship between protein mean and variance. *PLoS one* (2014) (cited on p. 6).

93. Tan, D. *et al.* Quantitative control of noise in mammalian gene expression by dynamic histone regulation. *eLife* (2021) (cited on p. 6).

94. Sherman, M. S. & Cohen, B. A. A computational framework for analyzing stochasticity in gene expression. *PLoS computational biology* (2014) (cited on pp. 11, 12, 22).

95. Tiberi, S., Walsh, M., Cavallaro, M., Hebenstreit, D. & Finkenstädt, B. Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics* (2018) (cited on pp. 11, 22).

96. Scholes, C., DePace, A. H. & Sánchez, Á. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell system* (2017) (cited on pp. 11, 22).

97. Zoller, B., Nicolas, D., Molina, N. & Naef, F. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology* (2015) (cited on pp. 11, 22).

98. Rodriguez, J., Ren, G., Day, C. R., Zhao, K., Chow, C. C. & Larson, D. R. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell* (2019) (cited on pp. 11, 22).

99. Normanno, D. *et al.* Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher. *Nature communications* (2015) (cited on p. 15).

100. Zhang, Z. *et al.* Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis. *Genes and development* (2016) (cited on p. 15).

101. Liu, Z. & Tjian, R. Visualizing transcription factor dynamics in living cells. *Journal of Cell Biology* (2018) (cited on p. 15).

102. Hansen, A. S., Amitai, A., Cattoglio, C., Tjian, R. & Darzacq, X. Guided nuclear exploration increases CTCF target search efficiency. *Nature chemical biology* (2020) (cited on p. 15).

103. McIsaac, R. S., Oakes, B. L., Wang, X., Dummit, K. A., Botstein, D. & Noyes, M. B. Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic acids research* (2013) (cited on p. 17).

104. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* (2003) (cited on p. 17).

105. Munsky, B., Neuert, G. & Van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* (2012) (cited on p. 17).

106. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature genetics* (2002) (cited on p. 17).

107. Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. *Science* (2013) (cited on p. 17).

108. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences* (2002) (cited on p. 17).

109. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics* (1976) (cited on p. 22).

110. Harris, L. *et al.* BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics* (2016) (cited on p. 22).

111. Faure, A. J., Schmiedel, J. M. & Lehner, B. Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell systems* (2017) (cited on p. 23).

112. Chen, X. & Zhang, J. The genomic landscape of position effects on protein expression level and noise in yeast. *Cell systems* (2016) (cited on p. 23).

113. Berry, S., Dean, C. & Howard, M. Slow chromatin dynamics allow polycomb target genes to filter fluctuations in transcription factor activity. *Cell systems* (2017) (cited on p. 23).

114. Hong, C., Ramu, A., Zhao, S. & Cohen, B. Effect of genomic and cellular environments on gene expression noise. *bioRxiv* (2022) (cited on pp. 23, 60).

115. Bartman, C., Hsu, S., Hsiung, C., Raj, A. & Blobel, G. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Molecular cell* (2016) (cited on p. 23).

116. Davis, J., Insigne, K., Jones, E., Hastings, Q., Boldridge, W. & Kosuri, S. Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. *Cell systems* (2020) (cited on p. 23).

117. Antolović, V., Miermont, A., Corrigan, A. & Chubb, J. Generation of single-cell transcript variability by repression. *Current Biology* (2017) (cited on pp. 24, 29).

118. Elgin, S. & Reuter, G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harbor perspectives in biology* (2013) (cited on p. 25).

119. Bintu, L., Yong, J., Antebi, Y., McCue, K., Kazuki, Y., Uno, N., Oshimura, M. & Elowitz, M. Dynamics of epigenetic regulation at the single-cell level. *Science* (2016) (cited on p. 29).

120. Janssens, J. *et al.* Decoding gene regulation in the fly brain. *Nature* (2022) (cited on p. 29).

121. Partridge, E. *et al.* Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* (2020) (cited on p. 35).

122. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* (2020) (cited on p. 35).

123. Consortium, E. P., Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B. & Shoresh N., e. a. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* (2020) (cited on p. 35).

124. Jiang, J., Cai, H., Zhou, Q. & M., L. Conversion of a dorsal-dependent silencer into an enhancer: evidence for dorsal corepressors. *EMBO journal* (1993) (cited on p. 35).

125. Alexandre, C. & Vincent, J.-P. Requirements for transcriptional repression and activation by Engrailed in Drosophila embryos. *Development* (2003) (cited on pp. 35, 48).

126. Grass, J., Boyer, M., Pal, S., Wu, J., Weiss, M. & Bresnick, E. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proceedings of the National Academy of Sciences* (2003) (cited on pp. 35, 48).

127. Iype, T., Taylor, D., Ziesmann, S., Garmey, J., Watada, H. & Mirmira, R. The transcriptional repressor Nkx6.1 also functions as a deoxyribonucleic acid context-dependent transcriptional activator during pancreatic beta-cell differentiation: evidence for feedback activation of the nkx6.1 gene by Nkx6.1. *Molecular endocrinology* (2004) (cited on p. 35).

128. Peng, G.-H., Ahmad, O., Ahmad, F., Liu, J. & Chen, S. The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. *Human Molecular Genetics* (2005) (cited on pp. 35, 36).

129. Martínez-Montañés, F., Rienzo, A., Poveda-Huertes, D., Pascual-Ahuir, A. & Proft, M. Activator and repressor functions of the Mot3 transcription factor in the osmostress response of Saccharomyces cerevisiae. *Eukaryotic Cell* (2013) (cited on pp. 35, 48).

130. Rister, J., Razzaq, A., Boodram, P., Desai, N., Tsanis, C., Chen, H., Jukam, D. & Desplan, C. Single-base pair differences in a shared motif determine differential Rhodopsin expression. *Science* (2015) (cited on pp. 35, 48).

131. Rachmin, I., Amsalem, E., Golomb, E., Beeri, R., Gilon, D., Fang P., . & Tshori, S. FHL2 switches MITF from activator to repressor of Erbin expression during cardiac hypertrophy. *International Journal of Cardiology* (2015) (cited on pp. 35, 48).

132. Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P. & Lander, E. S. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences* (2017) (cited on p. 35).

133. Carleton, J., Berrett, K. & Gertz, J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α-Bound Enhancers. *Cell systems* (2017) (cited on p. 35).

134. King, D., Hong, C., Shepherdson, J., Granas, D., Maricque, B. & Cohen, B. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife* (2020) (cited on pp. 35, 51).

135. Huang, Z., Liang, N., Goñi, S., Damdimopoulos, A., Wang, C., Ballaire, R. & Treuter, E. The corepressors GPS2 and SMRT control enhancer and silencer remodeling via eRNA transcription during inflammatory activation of macrophages. *Molecular cell* (2021) (cited on p. 35).

136. Friedman, R., Granas, D., Myers, C., Corbo, J., Cohen, B. & White, M. Information content differentiates enhancers from silencers in mouse photoreceptors. *eLife* (2021) (cited on pp. 35–37, 43, 44, 48, 50, 51, 53).

137. Loker, R., Sanner, J. & Mann, R. Cell-type-specific Hox regulatory strategies orchestrate tissue identity. *Current Biology* (2021) (cited on pp. 35, 48).

138. Tokuhiro, S. & Satou, Y. Cis-regulatory code for determining the action of Foxd as both an activator and a repressor in ascidian embryos. *Developmental biology* (2021) (cited on pp. 35, 51).

139. Jindal, G. & Farley, E. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell* (2021) (cited on p. 35).

140. Long, H., Prescott, S. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* (2016) (cited on p. 35).

141. Spitz, F. & Furlong, E. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* (2012) (cited on p. 35).

142. Arnosti, D. & Kulkarni, M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry* (2005) (cited on p. 35).

143. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* (1995) (cited on p. 35).

144. Pang, B. & Snyder, M. Systematic identification of silencers in human cells. *Nature genetics* (2020) (cited on p. 35).

145. Pang, B., van Weerd, J., Hamoen, F. & Snyder, M. Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology* (2022) (cited on pp. 35, 48).

146. Segert, J., Gisselbrecht, S. & Bulyk, M. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends in Genetics* (2021) (cited on p. 35).

147. Corbo, J. C., Lawrence, K. A., Karlstetter, M., Myers, C. A., Abdelaziz, M., Dirkes, W. & Langmann, T. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome research* (2010) (cited on pp. 36, 41).

148. Hughes, A., Enright, J., Myers, C., Shen, S. & Corbo, J. Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. *Scientific reports* (2017) (cited on p. 36).

149. Murphy, D., Hughes, A., Lawrence, K., Myers, C. & Corbo, J. Cis-regulatory basis of sister cell type divergence in the vertebrate retina. *eLife* (2019) (cited on p. 36).

150. Furukawa, T., Morrow, E. & Cepko, C. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* (1997) (cited on p. 36).

151. Chen, S., Wang, Q., Nie, Z., Sun, H., Lennon, G., Copeland, N., Gilbert, D., Jenkins, N. & Zack, D. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron* (1997) (cited on pp. 36, 43).

152. Freund, C. L., Gregory-Evans, C. Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L. & McInnes, R. R. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* (1997) (cited on p. 36).

153. Cheng, H., Khanna, H., Oh, E., Hicks, D., Mitton, K. & Swaroop, A. Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. *Human Molecular Genetics* (2004) (cited on p. 36).

154. Srinivas, M., Ng, L., Liu, H., Jia, L. & Forrest, D. Activation of the blue opsin gene in cone photoreceptor development by retinoid-related orphan receptor beta. *Molecular endocrinology* (2006) (cited on p. 36).

155. Hsiau, T.-C., Diaconu, C., Myers, C., Lee, J., Cepko, C. & Corbo, J. The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One* (2007) (cited on pp. 36, 43).

156. Hennig, A., Peng, G.-H. & Chen, S. Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Research* (2008) (cited on p. 36).

157. Montana, C., Lawrence, K., Williams, N., Tran, N., Peng, G., Chen, S. & Corbo, J. Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant. *Journal of Biological Chemistry* (2011) (cited on p. 36).

158. Cherry, T. *et al.* Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease. *Proceedings of the National Academy of Sciences* (2020) (cited on pp. 36, 44).

159. White, M., Myers, C., Corbo, J. & Cohen, B. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences* (2013) (cited on pp. 36, 43, 50, 51).

160. Hughes, A., Myers, C. & Corbo, J. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome research* (2018) (cited on p. 36).

161. Mitton, K., Swain, P., Chen, S., Xu, S., Zack, D. & Swaroop, A. The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. *Journal of Biological Chemistry* (2000) (cited on pp. 36, 43).

162. Kwasnieski, J., Mogno, I., Myers, C., Corbo, J. & Cohen, B. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences* (2012) (cited on p. 41).

163. Mears, A., Kondo, M., Swain, P., Takada, Y., Bush, R., Saunders, T., Sieving, P. & Swaroop, A. Nrl is required for rod photoreceptor development. *Nature genetics* (2001) (cited on p. 43).

164. Morrow, E., Furukawa, T., Lee, J. & Cepko, C. NeuroD regulates multiple functions in the developing neural retina in rodent. *Development* (1999) (cited on p. 44).

165. Lerner, L., Peng, G.-H., Gribanova, Y., Chen, S. & Farber, D. Sp4 is expressed in retinal neurons, activates transcription of photoreceptor-specific genes, and synergizes with Crx. *Journal of Biological Chemistry* (2005) (cited on p. 44).

166. Méthot, N. & Basler, K. Hedgehog controls limb development by regulating the activities of distinct transcriptional activator and repressor forms of Cubitus interruptus. *Cell* (1999) (cited on p. 48).

167. Parker, D., White, M., Ramos, A., Cohen, B. & Barolo, S. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Science Signaling* (2011) (cited on p. 48).

168. Tsai, A., Muthusamy, A., Alves, M., Lavis, L., Singer, R., Stern, D. & Crocker, J. Nuclear microenvironments modulate transcription from low-affinity enhancers. *eLife* (2017) (cited on p. 48).

169. Peng, Y. & Jahroudi, N. The NFY transcription factor functions as a repressor and activator of the von Willebrand factor promoter. *Blood* (2002) (cited on p. 48).

170. Pompeani, A., Irgon, J., Berger, M., Bulyk, M., Wingreen, N. & Bassler, B. The Vibrio harveyi master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Molecular Microbiology* (2008) (cited on p. 48).

171. Sánchez-Tilló, E., de Barrios, O., Valls, E., Darling, D., Castells, A. & Postigo, A. ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene* (2015) (cited on p. 48).

172. Samuel, A., Housset, M., Fant, B. & Lamonerie, T. Otx2 ChIP-seq reveals unique and redundant functions in the mature mouse retina. *PLoS One* (2014) (cited on p. 48).

173. Bulajić, M., Srivastava, D., Dasen, J., Wichterle, H., Mahony, S. & Mazzoni, E. Differential abilities to engage inaccessible chromatin diversify vertebrate Hox binding patterns. *Development* (2020) (cited on p. 48).

174. Singh, G., Mullany, S., Moorthy, S., Zhang, R., Mehdi, T., Tian, R., Duncan, A., Moses, A. & Mitchell, J. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome research* (2021) (cited on p. 51).

175. Smith, R., Taher, L., Patwardhan, R., Kim, M., Inoue, F., Shendure, J., Ovcharenko, I. & Ahituv, N. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature genetics* (2013) (cited on p. 51).

176. Van Dijk, D., Sharon, E., Lotan-Pompan, M., Weinberger, A., Segal, E. & Carey, L. Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators. *Genome research* (2017) (cited on p. 51).

177. Fiore, C. & Cohen, B. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome research* (2016) (cited on p. 51).

178. Majello, B., De Luca, P. & Lania, L. Sp3 is a bifunctional transcription regulator with modular independent activation and repression domains. *Journal of Biological Chemistry* (1997) (cited on p. 51).

179. Sauer, F. & Jäckle, H. Dimerization and the control of transcription by Krüppel. *Nature* (1993) (cited on p. 51).

180. Zuo, P., Stanojević, D., Colgan, J., Han, K., Levine, M. & Manley, J. Activation and repression of transcription by the gap proteins hunchback and Krüppel in cultured Drosophila cells. *Genes and development* (1991) (cited on p. 51).

181. Rodriguez, K., Do, A., Senay-Aras, B., Perales, M., Alber, M., Chen, W. & Reddy, G. Concentration-dependent transcriptional switching through a collective action of cis-elements. *Science advances* (2022) (cited on pp. 51, 56).

182. Sloan, J., Hakenjos, J., Gebert, M., Ermakova, O., Gumiero, A., Stier, G., Wild, K., Sinning, I. & Lohmann, J. Structural basis for the complex DNA binding behavior of the plant stem cell regulator WUSCHEL. *Nature communications* (2020) (cited on p. 51).

183. Mukund, A., Tycko, J., Allen, S., Robinson, S., Andrews, C., Ludwig, C., Spees, K., Bassik, M. & Bintu, L. High-throughput functional characterization of combinations of transcriptional activators and repressors. *bioRxiv* (2022) (cited on p. 56).

184. Galburt, E. The calculation of transcript flux ratios reveals single regulatory mechanisms capable of activation and repression. *Proceedings of the National Academy of Sciences* (2018) (cited on pp. 56, 57).

185. Phillips, R., Kondev, J., Theriot, J. & Garcia, H. *Physical biology of the cell.* (2012) (cited on p. 57).

186. Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K. & Irizarry, R. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* (2010) (cited on p. 59).

187. Tran, H., Ang, K., Chevrier, M., Zhang, X., Lee, N., Goh, M. & Chen, J. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* (2020) (cited on p. 59).

188. Hansen, J. & Cohen, B. A quantitative metric of pioneer activity reveals that HNF4A has stronger in vivo pioneer activity than FOXA1. *Genome Biology* (2022) (cited on p. 59).

189. Neikes, H. *et al.* Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq. *Nature Biotechnology* (2023) (cited on p. 59).

190. Alamos, S., Reimer, A., Westrum, C., Turner, M., Talledo, P., Zhao, J., Luu, E. & Garcia, H. Minimal synthetic enhancers reveal control of the probability of transcriptional engagement and its timing by a morphogen. *Cell systems* (2023) (cited on p. 59).

191. Lammers, N., Flamholz, A. & Garcia, H. Competing constraints shape the non-equilibrium limits of cellular decision making. *Proceedings of the National Academy of Sciences* (2023) (cited on p. 59).

192. Kim, Y., Rhee, K., Liu, J., Jeammet, S., Turner, M., Small, S. & Garcia, H. Predictive modeling reveals that higher-order cooperativity drives transcriptional repression in a synthetic developmental enhancer. *eLife* (2022) (cited on p. 59).

193. Hong, C. & Cohen, B. Genomic environments scale the activities of diverse core promoters. *Genome research* (2022) (cited on p. 60).

194. Bergman, D. *et al.* Compatibility rules of human enhancer and promoter sequences. *Nature* (2022) (cited on p. 60).

195. Zhao, S., Hong, C., Myers, C., Granas, D., White, M., Corbo, J. & Cohen, B. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nature genetics* (2023) (cited on p. 60).

196. Lalanne, J. *et al.* Multiplex profiling of developmental enhancers with quantitative, single-cell expression reporters. *bioRxiv* (2022) (cited on p. 60).

Appendix A: Supplement

A.1 Supplementary Figures

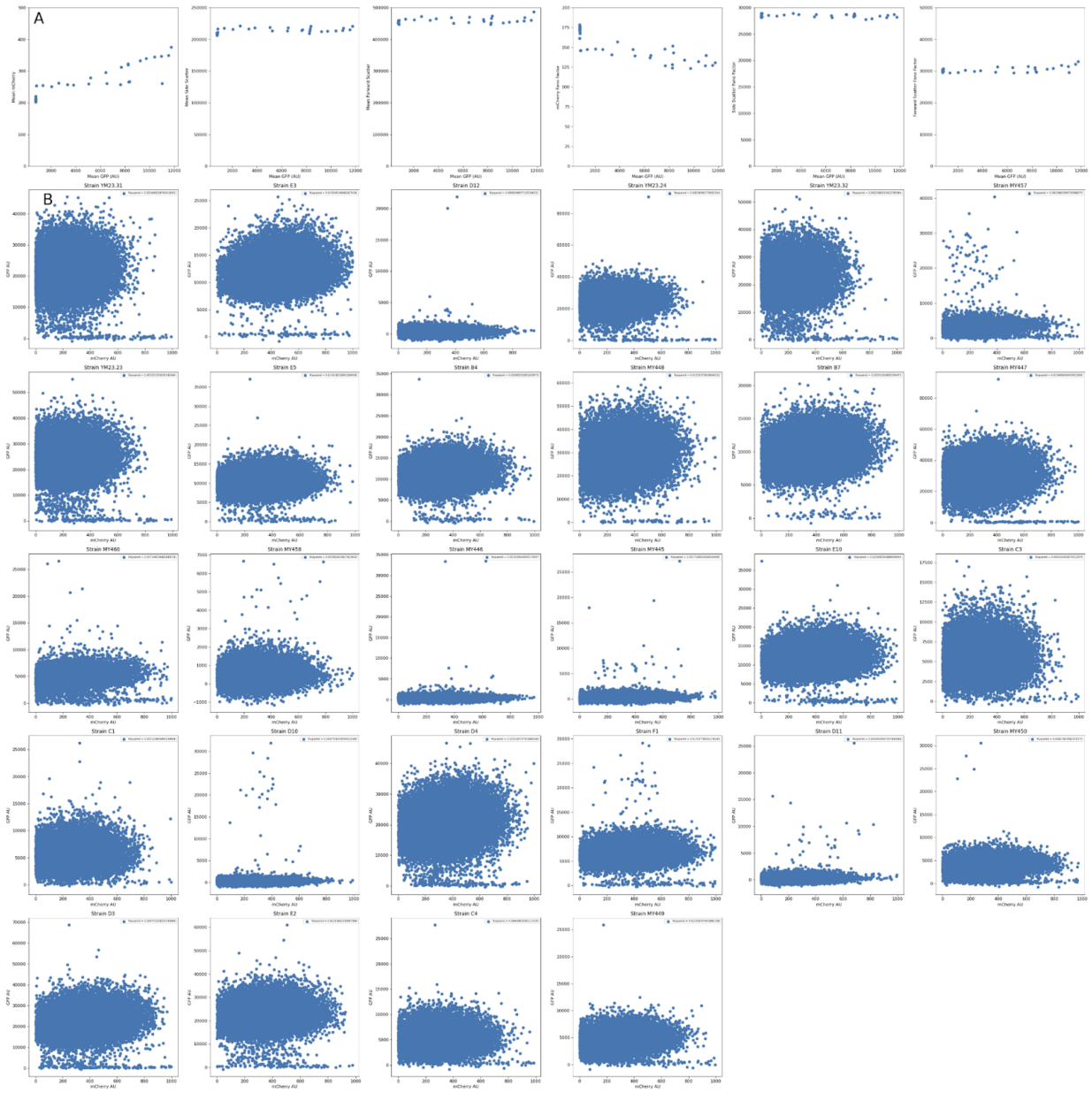

Figure A.1: mCherry expression does not correlate with beta-estradiol or GFP expression after controlling for cell size, Related to STAR methods A: mCherry expression and scatter measurements by GFP expression in a beta-estradiol dose-response experiment B: mCherry expression vs. GFP expression across all strains expressing WT activation domains, controlling for forward scatter

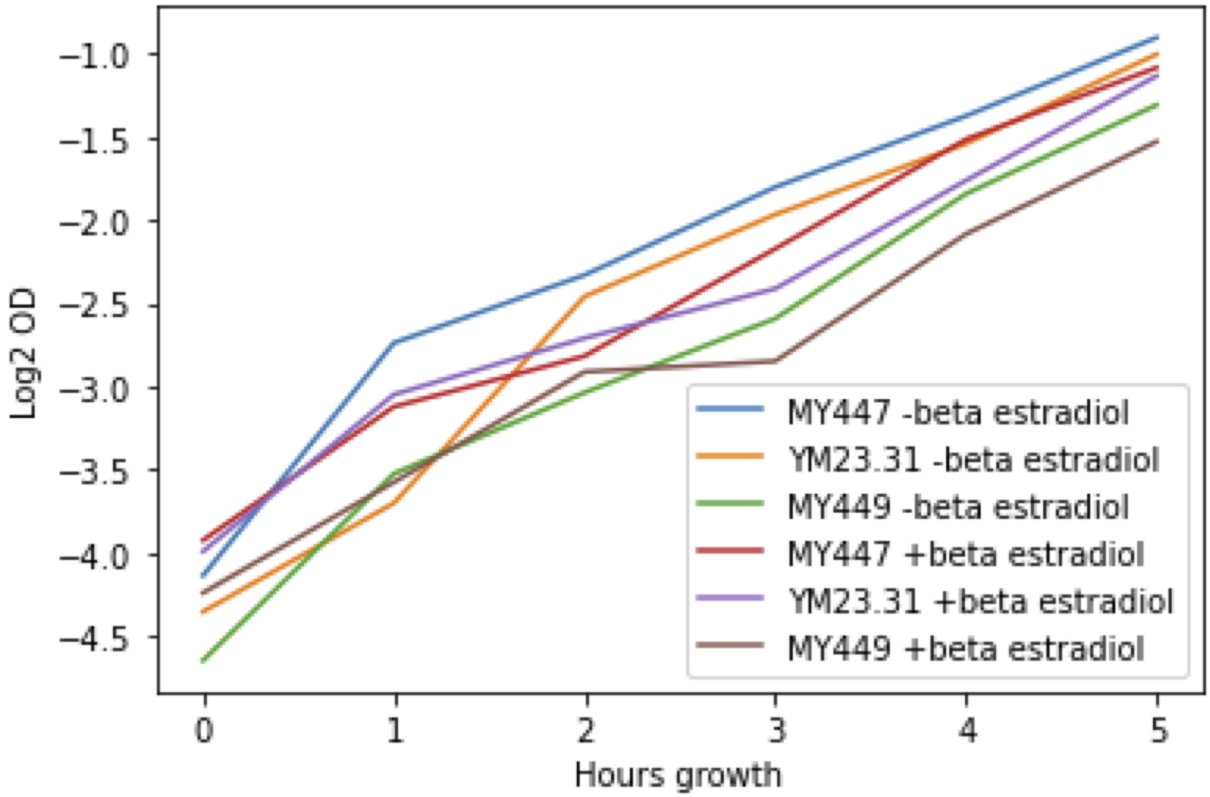

Figure A.2: Yeast growth curves in the presence and absence of beta-estradiol, Related to STAR methods

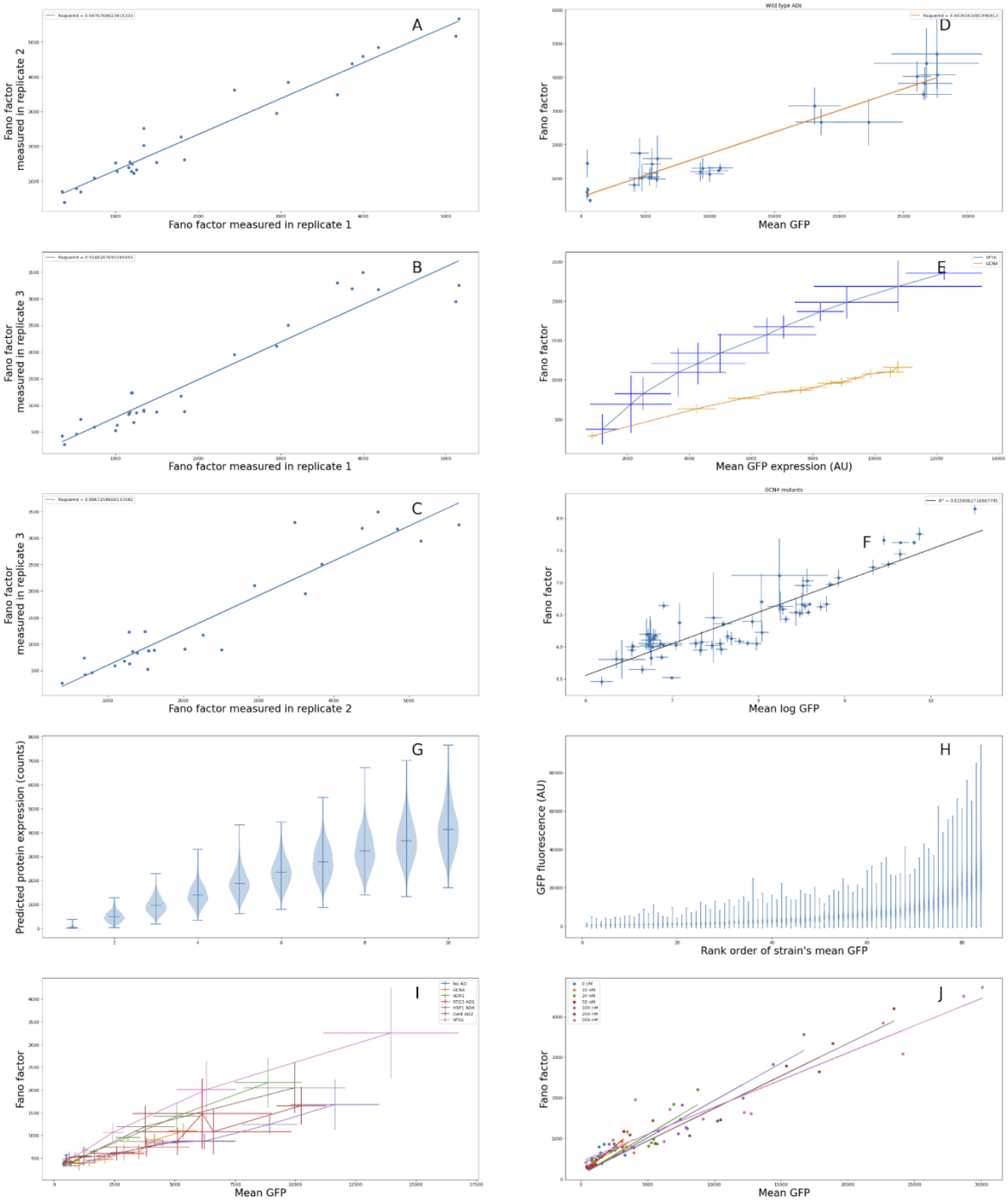

Figure A.3: Additional plots of mean-Fano factor relationships, related to Figure 2 A-C: Reproducibility of Fano factor estimates with cells filtered by forward scatter in addition to mCherry D: Means and Fano factors induced by yeast activation domains with cells filtered by forward scatter in addition to mCherry E: Means and Fano factors induced by VP16 and GCN4 at varying levels of beta- estradiol with cells filtered by forward scatter in addition to mCherry F: Means and Fano Factors of GCN4 mutants plotted on log scale G: Violin plots of predicted GFP expression distributions across ADs H: Violin plots of observed GFP expression distributions across ADs (ordered by mean GFP) I: Mean-Fano curves created by varying induction of 7 different activation domains Fig. S3j: Mean-Fano trendlines across activation domains at different beta-estradiol concentrations (in box)

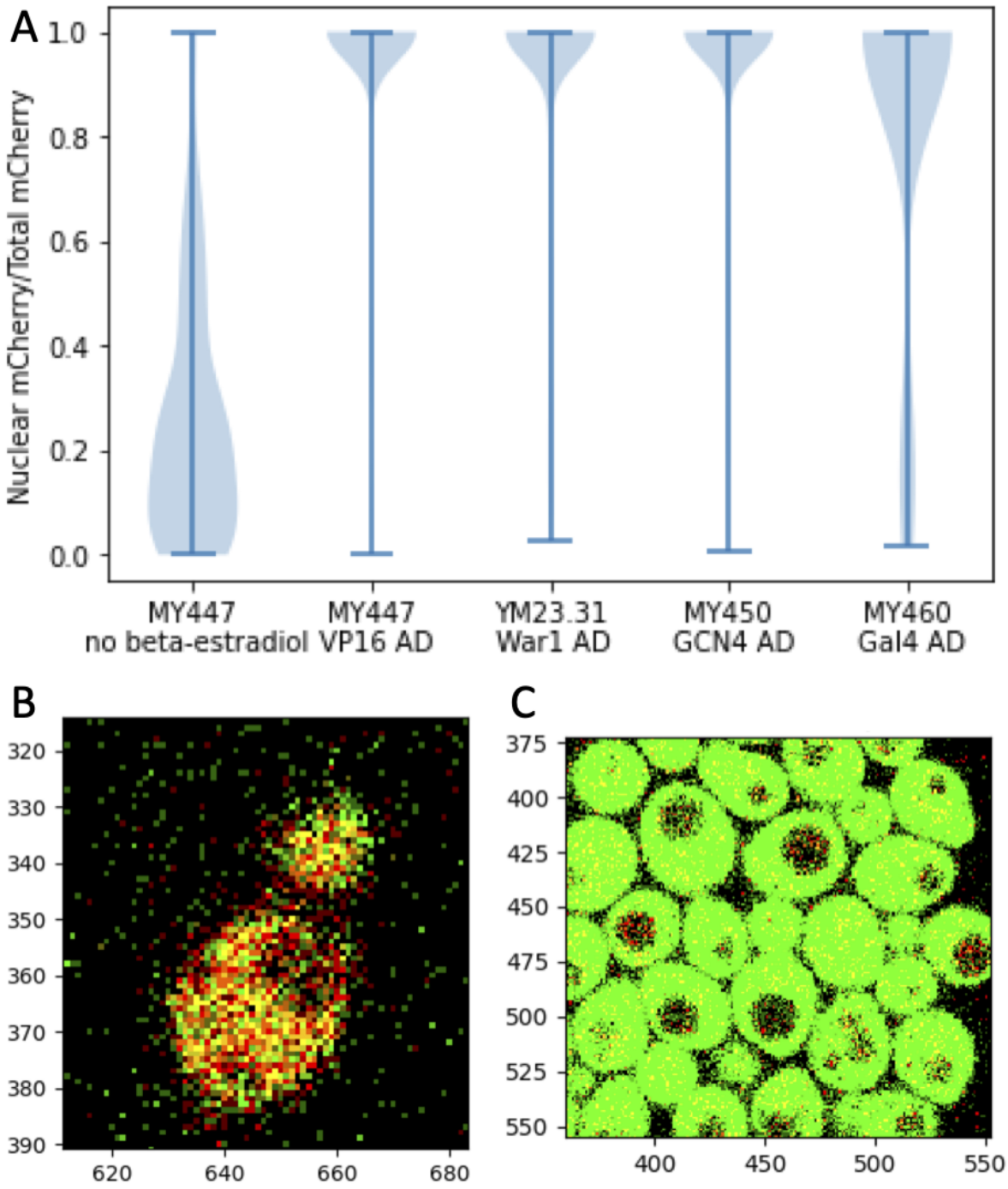

Figure A.4: Imaging shows nuclear localization of synthetic TF depends mostly on beta-estradiol concentration, Related to STAR methods A: Distributions of ratios of nuclear to total mCherry signal across conditions and strains. Strains MY447, YM23.31, MY450, and MY460 were induced with 200 uM beta-estradiol and the mCherry signal in the nucleus was quantified with fluorescence microscopy. The GFP fluorescence from the reporter gene was used to define cytoplasmic fluorescence. B: Example image of mCherry and GFP fluorescence in the absence of beta-estradiol C: Example image of mCherry and GFP fluorescence in the presence of beta-estradiol

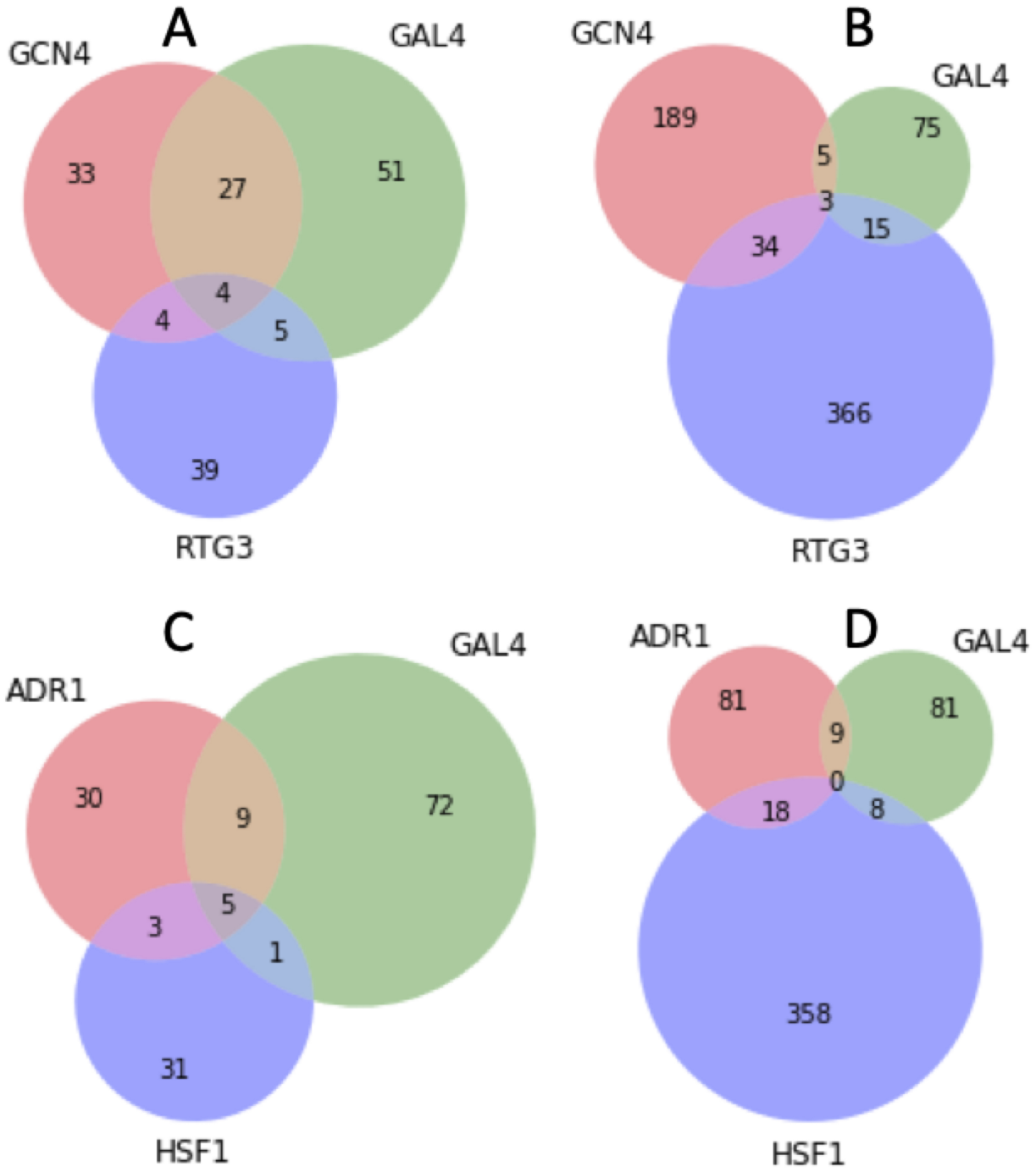

Figure A.5: Overlap between AD interaction partners, Related to STAR methods A: Shared and unique physical interaction partners of GCN4, GAL4, and RTG3 B: Shared and unique genetic interaction partners of GCN4, GAL4, and RTG3 C: Shared and unique physical interaction partners of ADR1, GAL4, and HSF1 D: Shared and unique genetic interaction partners of ADR1, GAL4, and HSF1

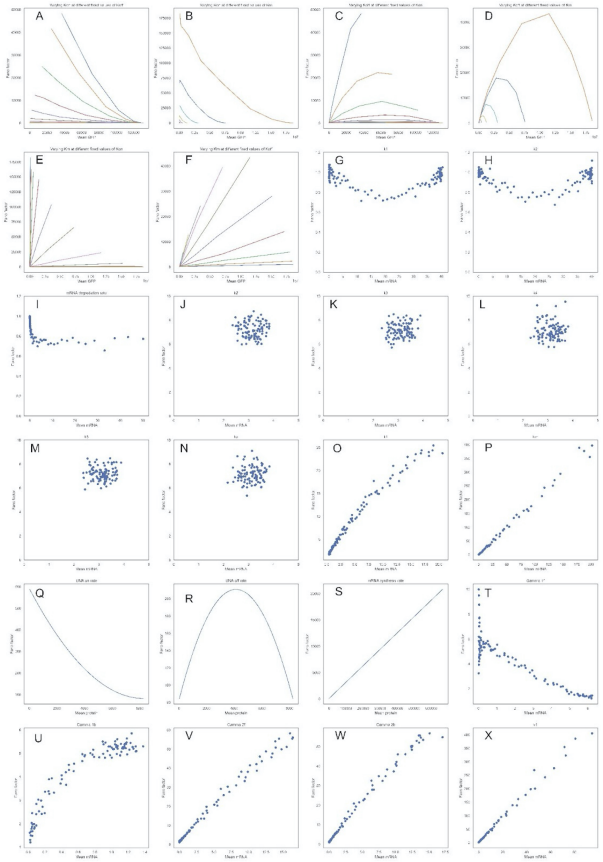

Figure A.6: Additional simulation results, related to Figure 3 A: Predicted effect of varying only Kon at different fixed values of Koff B: Predicted effect of varying only Kon at different fixed values of Km C: Predicted effect of varying only Koff at different fixed values of Kon D: Predicted effect of varying only Koff at different fixed values of Km E: Predicted effect of varying only Km at different fixed values of Kon F: Predicted effect of varying only Km at different fixed values of Koff G: Predicted effect of varying K1 in Scholes et al's transcription cycle model H: Predicted effect of varying K2 in Scholes et al's transcription cycle model I: Predicted effect of varying degradation rate in Scholes et al's transcription cycle model J: Predicted effect of varying Koff2 in Zoller et al's refractory period model K: Predicted effect of varying Koff3 in Zoller et al's refractory period model L: Predicted effect of varying Koff4 in Zoller et al's refractory period model M: Predicted effect of varying Koff5 in Zoller et al's refractory period model N: Predicted effect of varying Kon in Zoller et al's refractory period model O: Predicted effect of varying Koff in Zoller et al's refractory period model P: Predicted effect of varying Km in Zoller et al's refractory period model G: Effect of varying Kon predicted by Sherman et al's analytical solution R: Effect of varying only Koff predicted by Sherman et al's analytical solution S: Effect of varying only Km predicted by Sherman et al's analytical solution T: Predicted effect of varying Kon1 in Rodriguez et al's model U: Predicted effect of varying Koff1 in Rodriguez et al's model V: Predicted effect of varying Kon2 in Rodriguez et al's model W: Predicted effect of varying Km1 in Rodriguez et al's model

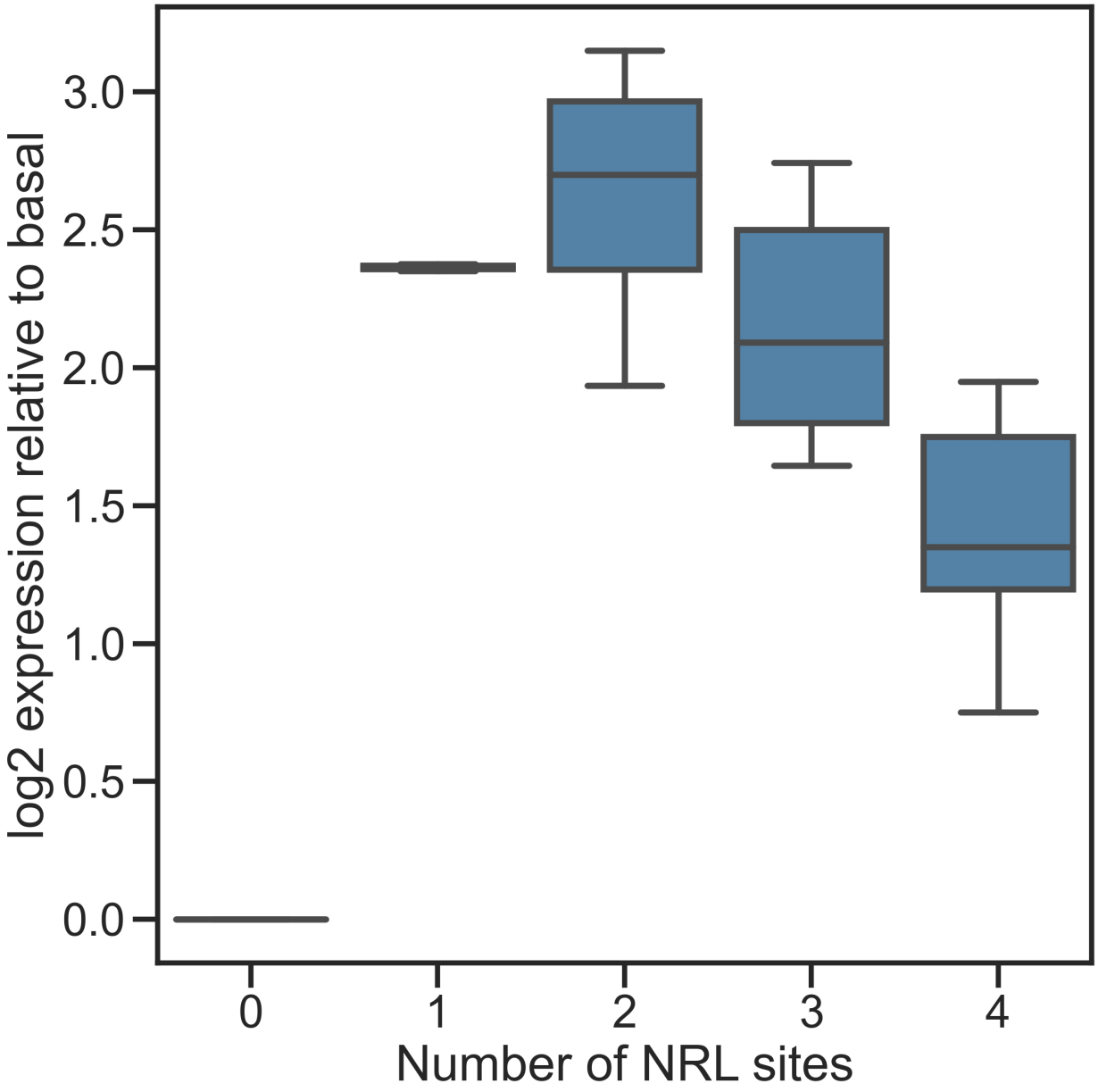

Figure A.7: Increasing NRL sites reduces MPRA activity in synthetic CRSs. Synthetic CRSs composed only of NRL sites show an increase, then a decrease in activity relative to the Rho basal promoter as the number of sites is increased. Plot shows a subset of the data reported in [19].

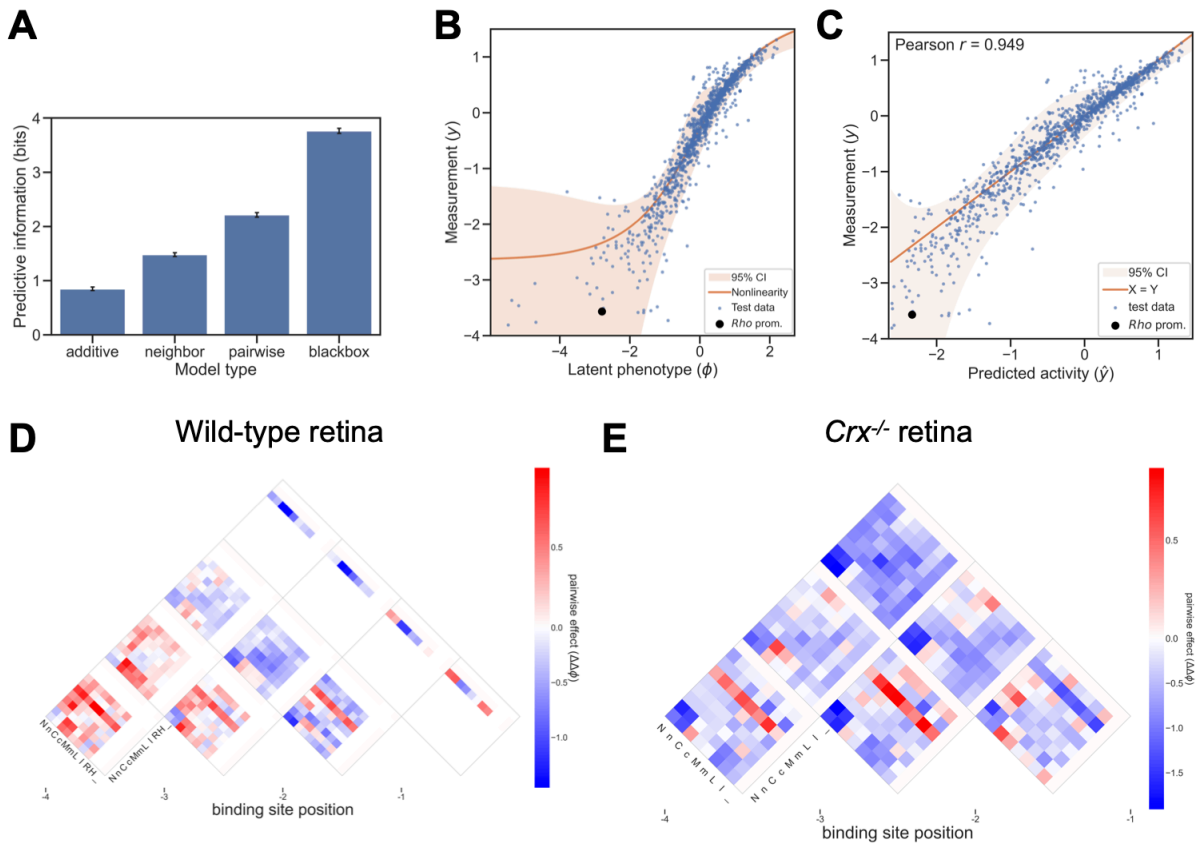

Figure A.8: A model of CRX and NRL-driven cis-regulatory activity in Crx-/- retina. (A) The performance of different model architectures (predictive information) fit to MPRA measurements of synthetic CRSs in Crx-/- retina. (B) Observed activity (y-axis) of test set sequences in Crx-/- retina compared to the latent phenotype (x-axis) predicted by the pairwise model. (C) Observed activity (y-axis) of test set sequences in Crx-/- retina compared to the activity predicted by the pairwise model (x-axis). (D) Model parameters for position-specific pairwise contributions of CRX and NRL sites in wild-type retina. Forward and reverse orientation of binding sites is indicated by capital or lower case letter. CRX sites are either high (C or c), medium (M or m), or low (L or l) affinity. NRL sites are labeled N or n and the Rho promoter is labeled R. There are no model parameters for Hsp68 (H, used as the overall basal sequence) or the placeholder site _ used to equalize the lengths of input sequences. See methods for details. (E) Position-specific pairwise contributions of CRX and NRL sites to activity in Crx-/- retina.

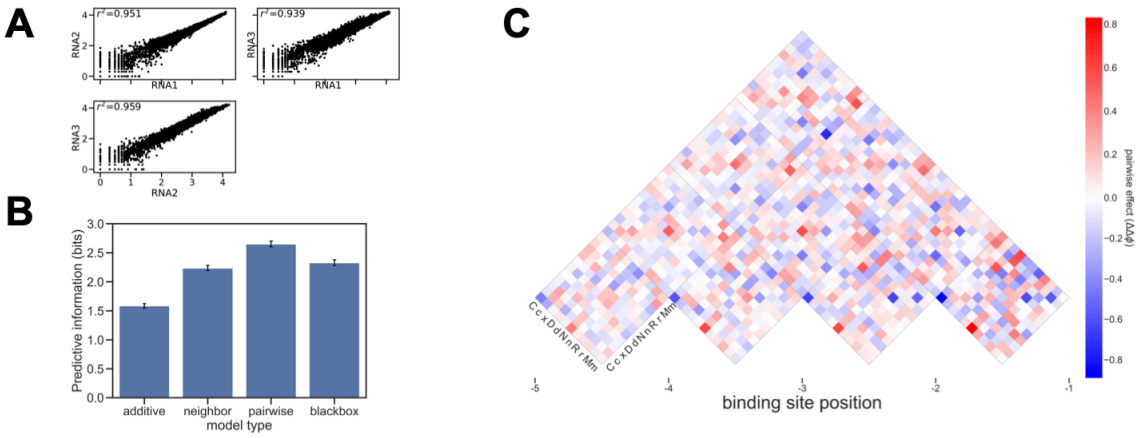

Figure A.9: A model of cis-regulatory activity driven by diverse TF binding sites in wild-type retina. (A) Reproducibility of MPRA measurements across three replicates. (B) Performance of modes fit to measurements of the MPRA library of CRSs composed of five TFBSs, expressed in terms of predictive information. (C) Position-specific pairwise contributions of diverse TF binding sites. Model parameters shown are the average of multiple training runs from different random starts as described in the methods. Capital and lowercase letters represent high and medium affinity sites for CRX (C), NEUROD1 (D), NRL (N), RORB (R), and MAZ (M). Low affinity CRX sites are represented by x.

Table A.1: Beta-estradiol concentrations used in creating Figure 1. 200 nM was used for all conditions in Figure 2. Related to Figure 1.

GCN4	16 uM	8 uM	4 uM	2 uM	1 uM	500 nM
250 nM	125 nM	64 nM	32 nM	16 nM	0 nM	
VP16	1.2 nM	1.1 nM	1 nM	0.9 nM	0.8 nM	0.7 nM
0.5 nM	0.4 nM	0.3 nM	0.2 nM	0.1 nM	0 nM	

A.2 Supplementary tables

Table A.2: AD sequences and names of strains

AD name	Sequence	Strain names
VP16	ELHLDGEDVAMAHADALDDFDLDMLGDGDSPGPGFTPHD-SAPYGALDMADFEFEQMFTDALGIDEYGG	MY447, MY448
War1	TNIKPSPSSSVDNLNDYLTDINSLAWGVNSLNDEFWTDLFM-NDI	YM23.23, YM23.24, YM23.31, YM23.32
GCN4	IPELDDAVVESFFSSSTDSTPMFEYENLEDNSKEWTSLFD-NDIPVTTDDVSLADK	MY449, MY450, MY457, MY461
Gal4	SAAHHDNSTIPLDFMPRDALHGFDWSEEDDMSDGLPFLKT-DPNNNGF	MY459, MY460
ADR1 AD4	DFVDFQELLDNDTLGNDLLETTAVLKEFELLHDDSVSATATS	C1, C3, C4
Gal4 AD2	NFNQSGNIADSSLSFTFTNSSNGPNLITTQTNSQALSQPI-ASSNVHDNFMNNEITASKIDDGNNSKPLSPGWTDQTAY-NAFGITTGMFNTTTMDDVYNYLFDDEDTPPNPKKE	D3, D4
HSF1 AD1	IFTTDRTDASTTSSTAIEDIINPSLDPQSAASPVPSSSFFHD-SRKPSTSTHLVRRGTPLGIYQTNLYGHNSRENTNPN-STLLSSKLLAHPPVPYGQNPDLLQHAVYR	B2, D10, D11, D12
HSF1 AD2	FTSRDPNNQTSENTFDPNRFTMLSDDDLKKDSHT-NDNKHNESDLFWDNVHRNIDEQDARLQNLENMVHILSP-GYPNKSFNNKTSSTNTNSNMESAVNVNSPGFNLQDYLT-GESNSPNSVHSVPSNGSGSTPLPMPNDNDTEHASTSVNQ-GENGSGLTPFLTVDDHTLNDNNTSEGSTRVSPDIKFSATEN-TKVSDNLPS	E3
HSF1 AD4	DFVDFQELLDNDTLGNDLLETTAVLKEFELLHDDSVSATATS	B4, E5
RTG3 AD1	MNNNESEAENQRLLDELMNQTKVLQETLDFSLVTPTPHHND-DYKIHGSAYPGGETPAQ	E10

Table A.3: Parameter values used in simulations in Figure 3 in units of S-1 (Ranges given for parameters that were varied).

	Kon	Koff	Km	Dm	Kp	Dp
Fig. 3e	0.022-0.22	0.22	0.14-1.4	0.044	5.2	0.02
Fig. 3f	0.0022-0.034375	0.022	1.4	4.4-1.76	5.2	0.02
Fig. 3g	0.22	0.22	0.014-1.4	0.044	5.2	0.02
Fig. 3h	0.22	0.22	1.4-3.5	4.4-1.76	5.2	0.02

A.3 Sequences

A.3.1 Primer sequences in Chapter 2

YP16: ACTGCACAGAACAAAAACCTGCAGGAAACGAAGATAAATCTTCGTACGCTGCAGGTC-GAC YP17: GTGAGTTTAGTATACATGCATTTACTTATAATACAGTTTTTCGATGAATTC-GAGCTCGTT URA3_UP_Kai_1: GAAGGTTAATGTGGCTGTGG URA3_UP_Kai_2: GATTCG-GTAATCTCCGAGCAGAAG URA3_UP_Kai_rc: CCTACCAATAGCTTCAACACTGTTGAG

A.3.2 Primer sequences in Chapter 3

HIS_UP_homology2: GGCAAGATAAACGAAGGCAAAGATGACCTGCAGGTCGACGGATCC HIS_dwn_homology2: GCGAGGTGGCTTCTCTTATGGCAACCGCCGATGAATTCGAGCTCG HIS_up_in_1: CTTCTTTGCGTCCATCC HIS_up_out_1: CCACCTAGCGGATGACTC HIS_dwn_in_1: CCATGGTATGGATGAATTGTAC HIS_dwn_out_1: CCCGTTCCTCCATCTC