Winter 12-15-2022

Single-Cell Massively Parallel Reporter Assays

Siqi Zhao
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Barak A Cohen, Chair
Jeremy Buhler
Robi Mitra
Zachery Pincus
Gary Stormo

Single-Cell Massively Parallel Reporter Assays

by
Siqi Zhao

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2022
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

Very few can finish a PhD alone. For unfortunate ones like me, research only provides some intrinsic meaning for life. It is the people surrounding me that make it worthwhile. It is through my interactions with them that I become a better person and a better researcher. Many people's lives granted me courage:

Barak showed me it is possible to do science for the sake of science. He encourages me to take risks and shares my failures. He is genuinely interested in understanding. He cares about my well-being and growth.

Zach showed me it is still possible to be a polymath in this day and age.

My college friends, Stella, Emily, Monty, Eric, Charlie, Emma, and Esti, showed me it is possible to pursue an examined life.

My lab mates taught me there are many ways one could excel at research and life. Working with Avi and Clarice has been a wonderful journey.

Emma, Esti, and Charlie have been wonderful friends who are there for all the worst and best moments in my life.

My cats, Oikos and Logos, are wonderful.

My partner, Micah, gives me the courage to face research and life.

And my mother, Zhang Wei, gives me so much support and lives a virtuous life that I can always learn from.

It is also through many people's kindness that I can be who I am. I thank them.

Siqi Zhao

Washington University in St. Louis

August 2022

To my mother.

“Ergo vivida vis animi pervicit et extra
processit longe flammantia moenia mundi
atque omne immensum peragravit mente animoque.”
-Titus Lucretius Carus

ABSTRACT OF THE DISSERTATION

Single-Cell Massively Parallel Reporter Assays

by

Siqi Zhao

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2022

Professor Barak A Cohen, Chair

Our understanding of gene regulation needs to be generalizable as well as specific. A generalizable understanding enables us to transfer our knowledge of one gene to another, and the specificity allows us to understand the precise regulation of gene expression through development. One framework that confers generalizability and specificity is the hierarchical and modular model of gene regulation. To test whether gene regulation is hierarchical and modular requires methods to systematically analyze how different factors collectively control gene expression. This thesis describes the development of two functional genomics methods at single-cell resolution. These methods systematically examined how cellular contexts, chromatin environment, and local regulatory sequence collectively control expression mean and noise. First, scMPRA measures cell-type specific expression of a library of core promoters in K562 and HEK 293 cell lines. scMPRA can also be applied to a complex tissue and performed MPRA *ex vivo*. Both general principles of gene expression and cell-type specific variant effects are found in the newborn mouse retina. scMPRA also measured the cell substate effect on expression mean, I found that cell substate has a large and general effect on expression mean for core promoters. I also deconvolved the extrinsic and intrinsic portion of expression noise, we found that developmental core promoters have larger extrinsic noise. Second, scTRIP measures the chromatin environment effect on expression noise. We found expression noise can be partially explained by expression noise. We also found the expression noise that is independent of expression mean is correlated with specific chromatin marks and transcription factor binding sites. Moreover,

we identified the oscillation between cell substates as a major source of extrinsic noise regardless of the chromatin environment. Using all the information, we trained a logistic regression model with high accuracy. These observations and methods provide a framework to further explore the hierarchical and modular nature of gene regulation.

Chapter 1: Introduction

1.1 Is Gene Expression a Controlled Chaos?

Gene expression is puzzling.

One marvelous event of the living world is how a single cell gives rise to numerous cell types with distinct functions and morphology. Through development, a human zygote would go through 10^{12} divisions and differentiates into more than 200 cell types relying on an identical stretch of genetic material [1]. The functional and morphological diversity for cells in metazoans arises from the spatiotemporal control of gene expression. To control the differentiation process, different genes are tuned to turn on at specific development points. Hence a common analogy of a cell is a precisely tuned machine in which numerous components are orchestrated to perform diverse biological functions.

At the same time, gene expression is incredibly noisy. Due to the thermal noise and the low copy number of regulatory proteins in a cell, the process of gene expression and regulation is incredibly dynamic. It is reported that mRNA levels within a homogeneous population of cells can differ by two orders of magnitude [2]. Similarly, protein counts also span several orders of magnitudes in both prokaryotic and eukaryotic organisms [3].

To understand how gene expression controls development and disease, we must reconcile the seemingly paradoxical phenomena between stochasticity in gene expression and the deterministic order of events in development and normal physiology. We are faced with two questions: (1) What are the mechanisms that control gene expression mean, and (2) How do the mechanisms of gene expression control single-cell variability?

1.2 Transcriptional Regulation: the Interplay between Generalizability and Specificity

The study of gene regulation concerns when, where, and how genes are expressed and silenced. One conceptual challenge in understanding gene regulation is that the genome is one stretch of nucleic acids, and all the spatiotemporal information of gene regulation is essential and stored linearly in the DNA sequence.

A series of experiments conducted by Jacod and Monod [4] first elucidated how the same stretch of genetic material can both provide information for **how** a protein is made, as well as **when** a protein is made. Jacod and Monod proposed that there exist two types of genes. "Operator genes" that are bound by molecules sense the environmental signals, like the change of sugar source; "Structural genes" provide instruction for how a gene carries out its function. The "dual genetic control" mechanism provides a framework for transcriptional regulation. The field of transcriptional regulation later formalized this "dual control" as the coding DNA sequences and non-coding regulatory elements. While through development, all information is read out from the genome; at any given point, trans-regulation comes from regulatory proteins that are produced *prior* to this time point, and cis-regulation is the regulatory ability of regulatory elements.

Later, researchers discovered almost all aspects of gene expression are regulated. Many steps in gene expression, including transcription activation [5], transcription elongation and pausing [6], messenger RNA stability [7], translation rate [8], and post-translational modification [9], contribute to the differences in gene expression in different cell types and developmental stages.

Transcriptional regulation remains central to development. Many classes of transcription factors that interact with non-coding DNA sequences to activate or repress gene expression have been found. One highlight of the past sixty years of genetic research is that the expression of four transcription factors (TF) can induce fibroblast cells to become pluripotent stem cells [10]. Thus, the study of transcriptional regulation has focused on understanding how different non-coding sequences control

spatiotemporal gene expression.

One ultimate goal of studying gene regulation would be to generate knowledge and models that predict spatiotemporal gene expression patterns at the individual cell level. A quantitative and predictive understanding of transcriptional regulation is central to genetics and provides a mechanism that underlies cellular diversity and development.

1.3 The Epistemology of Gene Regulation

There are two specific types of knowledge that underlie our ability to achieve a quantitative and predictive understanding of gene regulation: (1) whether there are generalizable principles and (2) how specificity is achieved. The search for generalizable principles directly determines how predictive our understanding of gene regulation will be. If there are no generalizable principles, i.e., gene regulation is completely shaped by small contingencies in evolution, then we could never predict regulatory activity that has not been measured. However, current evidence suggests that a complete evolutionist view of gene regulation is not true. Generalizable principles have been found. Among the generalizable principles, the central dogma of molecular biology shapes many of the inquiries about gene regulation [11]. The mechanisms for specificity, on the other hand, determine if a quantitative understanding of gene regulation can be achieved. Our understanding of gene regulation must be able to explain how specific genes are active at specific developmental points and in specific tissues. However, the generalizable principles are not directly compatible with our understanding of the specificity of gene regulation. Conceptually speaking, the more specific the mechanism for gene regulation is, the more spatio-temporal control can be achieved. However, the more specific the mechanism for gene expression is, the less generalizable it is. I would argue that our understanding of gene regulation is the interplay between the search for generalizability and specificity.

One of the conceptual frameworks that reconciles the paradox between generalizability and specificity is a hierarchical and combinatorial model of gene regulation. That is, there are hierarchical

structures that are general. If we think transcription factors (TF) are at the higher level of the hierarchical structure, and regulatory elements (RE) are at a lower level, then the rule that governs TF-RE interaction is the same. The specificity comes from the combinatorial nature of the framework. The presence or absence of different sets of TFs can activate or repress different REs. Early studies conducted on viral transcription have shown this framework is powerful in explaining gene regulation [12]. A series of mutagenesis experiments of the SV40 regulatory region was conducted to illustrate that the SV40 enhancer has the following properties: (1) modular with other regulatory elements [13], (2) consists of independent subunits, and [14, 15] (3) those subunits function combinatorially [16].

The hierarchical and combinatorial nature of regulatory elements enabled us to discretize a continuous stretch of the genome into small units. This enabled us to formalize the question of gene regulation to study how those functional units interact with each other. The idea of modularity and discretization is not new to science. In mathematics, Cohen's proof of the continuum hypothesis showed a formal connection between continuous and discrete mathematical quantities [17, 18]. In physics, De Broglie waves provided an explanation for the wave-like (hence continuous) properties of matter [19].

1.4 Testing Modularity among Transcription Factor Binding Sites

Identifying functional units (or modules) and determining their relationship is critical to understanding complex systems like the genome [20]. Transcription factor binding sites (TFBS) are the basic modules for gene regulation. TFBS are short, of six to twelve base pair long stretches of DNA sequences that have biochemical affinity to different TFs [21, 22]. Efforts have been put into understanding how TFBS combinatorially determine the activity of regulatory elements. Firstly, systematic efforts had been put to identify and characterize all functional units in mam-

malian genomes. The Encyclopedia of DNA Elements (ENCODE) used biochemical assays to mark non-coding sequences that interact with regulatory proteins (like transcription factors) and structure protein modifications (like histone modifications) [23, 24]. The main tool for identifying how transcription factors interact with DNA sequence *in vivo* is chromatin immunoprecipitation. To achieve genome-wide coverage, chromatin immunoprecipitation was combined with DNA microarray (ChIP-chip) [25] and later with next-generation sequencing (ChIP-seq) [26–28] to identify DNA sequences that interact with a protein of interest. A newer tool for identifying protein-interacting DNA sequences is Cleavage Under Targets and Release Using Nuclease (Cut&Run) [29]. To date, ENCODE project has generated 406268 datasets to record footprints of different regulatory proteins interacting in different cell types. Identifying TFBS provides a look-up-table-like model of transcriptional regulation.

Secondly, to integrate TFBS properties in regulatory elements, several theoretic frameworks have been developed. Both machine learning-based models and physically-inspired models have been shown to be predictive. For machine-learning-based models, k-mer-based support vector machines (k-mer SVM) were developed to understand the sequence features that distinguish functional and nonfunctional regulatory sequences [30]. Recently, several deep learning models have been shown to predict gene expression using sequence features [31]. Models using thermodynamic principles were also developed to predict gene expression [32, 33]. In those models, TF-TF, TF-TFBS, and TF-Polymerase interactions are parameterized to describe the different biophysical interactions among macromolecules with different sequence combinations. I would argue that both categories of models reply to the fundamental notion of grouping sequences into functional units that interact with transcription factors and cofactors. Thermodynamic models directly parameterize using TFBS. While machine learning models train and test using sequence features, the predictability is often interpreted using TFBS affinity, density, and spacing.

Thirdly, parallel to the theoretical advances, technological advances in measuring regulatory element activity also made a systematic investigation of regulatory code possible. A class of methods

that enable high-throughput testing of regulatory element functions are massively parallel reporter assays (MPRAs). MPRAs combine traditional reporter assays with next-generation sequencing technology. Reporter assays, like luciferase assays, are used to determine if a piece of the non-coding sequence has regulatory activity by measuring the expression of a fluorescent reporter gene [34, 35]. To measure many non-coding sequences' regulatory ability by RNA sequencing, a unique DNA barcode is added to the 3' UTR region of the reporter gene. The relative abundances of DNA barcodes linked to different non-coding sequences measure the regulatory activity of $10^3 - 10^5$ of regulatory sequences at once. MPRAs generate large enough datasets to have the power to examine how different TFBS interact with each other. MPRAs have been applied to screen for functional sequences that have regulatory activities [36], test variant effects on regulatory activity [37], and dissect how TFBS in regulatory elements collectively control expression [38, 39].

Combining the aforementioned strategies, some general rules have been found. For example, the mutation in TATA-Box, the binding site for TBP, reduces regulatory activity[40, 41]. Those general rules often describe the general transcriptional mechanisms. However, for the majority of regulatory elements, the rules underlying their regulatory activity seem to be complex. Studies have shown that TFBS affinity, number, and combination have shown to be important in controlling regulatory activity[42–44]. For example, Gertz et al. used thermodynamic models to predict how to predict the activity of a library of synthetic promoter sequences in *Saccharomyces cerevisiae* [45]. A total of 2807 synthetic promoters with 7 different motifs were tested was tested. The expression of the library of promoters was measured with flow cytometry. The thermodynamic model is predictive. One important result is the complex interactions among TFBS. For example, a single Mig1 binding site is a weak repressor. But a weak Mig1 site adjacent to a strong Mig1 site becomes an activator. These complex interactions among TFBS are more pronounced in mammalian systems. Studies using mammalian cell lines have shown that the orientation, spacing, and adjacent DNA sequence all contribute to the regulatory activity[38, 46]. Taken together, this suggests that while TFBS binding sites are discrete functional units within a regulatory element, the rule governing their interactions

is highly dependent on the context.

1.5 Testing Modularity among Regulatory Elements

In the hierarchical model of gene regulation, TFBS collectively function within longer stretches of non-coding DNA called regulatory elements. While the rules governing how TFBS collectively controls gene expression are complex, the regulatory elements consist of many TFBS that are discrete and have been systematically identified [47]. In metazoans, promoters are often activated or repressed by enhancers to control gene expression. Enhancers are distal regulatory elements that control gene expression in a very tissue-specific manner. Enhancers are necessary for the proper expression of a gene. A single nucleotide change in enhancers could disrupt gene expression [34]. Other regulatory elements, like insulators and boundary elements, are also found important for maintaining a proper expression profile. Testing whether modularity is found on the level of regulatory elements is also critical to understanding transcription regulation.

Though recent studies revealed the complex grammar for how TFBS collectively contributes to the activity of regulatory elements, regulatory elements have been shown to act modularly from the early examples[12, 16]. SV40 enhancer, one of the most well-studied regulatory elements, was shown to have regulatory activity compatible with viral, mammalian, and prokaryotic promoters in diverse cell types [13]. One focus of functional genomics has been to test the hypothesis that general compatibility exists among regulatory elements.

Recent studies using MPRA have proven this to be generally true. Bergman et al. [48] have studied a total of 10^6 enhancer-promoter pairs and measured their regulatory activity in human K562 cells. They showed that enhancers and promoters are in general modular to each other, following a multiplicative relationship. Another study [49] with a similar study design focused on examining the specific cases that do not follow the general trend. They found certain enhancers have larger variability in their regulatory activities. However, this selectivity correlates with the overall strength of the enhancer elements. Taken together, this suggests that the general compatibility exists and is

governed by the regulatory activity.

Beyond the specific test of the modularity among regulatory elements, a necessary step towards understanding transcriptional regulation is to study how regulatory elements behave in the genome. To study the genomic location effect, a study assayed how the expression changes for a promoter integrated into thousands of genomic locations [50]. They found that general chromatin features, like lamina-associated domains, have a uniform effect on expression. This suggests that the genomic environment has a modular, rather than a specific effect on regulatory elements. The general compatibility between promoters and genomic environment is further supported by studies that integrated different promoter elements into known genomic locations and across the genome[51].

This overall modularity among regulatory elements demonstrates generalizability for the rules underlying gene expression. However, if regulatory elements are generally compatible, how is specificity achieved?

1.6 Can Cellular Environment Provide Specificity?

Given the overall modularity among regulation elements, especially the fact that enhancers can activate broad ranges of promoters [48, 51], an important question is how different genes are expressed at different tissues and developmental points. One hypothesis is that the genomic architecture provides such control. The genome is organized into chromatin loops and topologically associated domains (TADs) [52]. Some evidence has suggested that chromatin loops between enhancers and promoters are required for transcription initiation [53]. It is further supported by the existence of cell-type specific chromatin loops [54]. However, this hypothesis that chromatin structure is the cause of specificity, at least the direct cause, is implausible. This is because the mechanism that a chromatin loop forms is very general. The process of chromatin loop formation involves Cohesion and CTCF. Cohesion utilizes binding sites of CTCF as anchors to extrude stretches of DNA to form loops [55, 56]. There is little evidence that this mechanism is cell-type specific or developmental-time-point specific. Moreover, the depletion of chromatin loops does not

have a uniform effect on expression [57].

An alternative hypothesis for specificity is that different combinations of TFs in different cell types provide specificity. While broad compatibility for regulatory elements exists, the different sets of TFs could change the activity of those regulatory elements. Historically, cell-type specific gene expression of the hSV40 enhancer is shown to have different regulatory activity in NIH/3T3, NS-1, MDCK, and HeLa cell lines [58]. If we think development requires the expression of certain genes to reach certain critical thresholds, then this trans-regulation could provide specificity. *Drosophila* development provides direct evidence for this hypothesis. Different enhancers control the gap gene patterning along the *Drosophila* embryo axis [59, 60]. Additional evidence from mouse olfactory neuron development also supports this "enhancer-centric" idea of specificity [61]. During mouse development, more than 1000 types of olfactory neuron receptors are collectively regulated by a small number of enhancers. With the recent advances in single-cell sequencing technologies, putative regulatory elements are found to be cell-type specific [62]. This suggests that the broad regulatory-element-level compatibility may or may not hold in the different cellular contexts. Cell type and its cellular context may serve as a binary selector that alters this broad compatibility.

The specific effect of cellular contexts in different cell types seems to be a general case. Murphy et al. found that the deletion of a transcription factor binding site in the Gnb3 promoter eliminates the activity in bipolar cells, but not in rod photoreceptors [63]. This observation revealed that cellular context might serve as a binary filter for the TFBS in many regulatory elements. Two consequences would follow: (1) By taking the cellular environment into account, the rules governing TFBS might become simpler, and (2) The cellular environment might provide the specificity needed given regulatory elements (enhancer and promoters) seem to have general compatibility.

However, how cellular context influences regulatory element function has not been broadly tested due to the lack of a high throughput method to read out the cell type, TF expression, and the activities of many regulatory elements. One important technical gap in understanding transcriptional regulation is the ability to assay regulatory activity and the cellular environment simultaneously.

1.7 Importance of Understanding Single-cell Level Expression

Development is a result of the collective behaviors of single cells in a multicellular organism. To reconcile the large differences in single-cell expression and the deterministic order of events in development, we also need to understand how non-coding sequences regulate expression at the single-cell level. While organisms may utilize regulatory network structures to buffer expression noise [64], in many cases, differences in single-cell expression in an identical population of cells have shown to have phenotypic significance.

One situation where expression noise is biologically important is symmetry breaking during development [65, 66]. Symmetry breaking is when a group of identical undifferentiated cells commits to different fates. Some symmetry-breaking events during development rely on the single cell expression difference in a group of identical cells. Chang et al. demonstrated the cell-to-cell difference of a single gene could lead to different differentiation outcomes [67]. The authors examined the single-cell expression of a stem cell marker Sca-1 among an isogenic population of a mouse multipotent hematopoietic cell line. They used fluorescence-activated cell sorting to separate two subpopulations of cells with high and low Sca-1 expression and allowed them to differentiate. They found the Sca-1^{high} subpopulation is more likely to take the myeloid lineage whereas the Sca-1^{low} subpopulation differentiates into the erythroid lineage more. This study and other studies in intestinal crypt development[68], olfactory neuron development [61, 69], and retinal photoreceptor patterning Anderson2017-rl all suggested that organisms utilize expression cell-to-cell variability to ensure proper development. Moreover, differences in single-cell expression have been shown to play an important role in cancer drug resistance. Shaffer et al. demonstrated that the rare-cell survival of chemotherapy in melanoma cells is due to non-genetic expression variability [70]. Those cells have sporadic high expression of many cancer resistance markers like JUN, NFGP, EGFR. The results suggest that expression cell-to-cell variability in a population of cells also contributes to disease progression.

One view of single-cell expression is that the expression difference can be completely captured by

the change in mean. In general, the increase in the variance of the single-cell expression distribution can be well explained by the increase in the mean level. In other words, expression variance scales with expression mean [71]. This trend seems to be general. A scaling relationship has been found in *Saccharomyces cerevisiae*, *Escherichia coli*, and human cell lines [72, 73]. This view suggests that understanding the expression mean is enough for the expression variability. However, there are two sets of results that are incompatible with this view: (1) many genes fall off this trend [71], and (2) variants in the non-coding sequences that alter expression variance but not mean have been found [74]. It remains unclear what mechanisms underlie the expression of those genes that fall off this trend. Another view of expression variability is that mechanisms controlling expression variability could be decoupled from the ones that control expression mean. To support this view, Larsson et al. found that promoters with TATA box and Inr element have higher noise and larger burst size [75]. The separate control of single-cell variability is also supported by the finding that TATA-box-containing promoters have higher noise in yeast [76]. Therefore, it is critical to properly study and measure expression variability. Historically, another reason for the emphasis of the average expression level across populations of cells is mainly due to the lack of proper methods. With the recent booming of single-cell methods, notably single-cell RNA sequencing [77, 78] and multiplexed single molecule Fluorescence *in situ* Hybridization [79], universal variability is observed in almost all genes. Hence, an important next step is to elucidate the control mechanisms not only for expression mean, but also for expression variability.

Just as trans- and cis- regulation both control expression mean regulation, there are two sources of information that control expression variability [80]. The intrinsic component of expression variability refers to the variability that arises from the thermodynamic fluctuations of protein-DNA interactions (which could be transcriptional or translational), whereas the extrinsic component of expression noise refers to the global fluctuations that are different among individual cells. A natural hypothesis based on this framework is that regulatory elements control the intrinsic variability, and the cellular context has a large effect on the extrinsic variability. But many aspects of gene

expression could contribute to noise, for example, the differences of protein coding sequences. Thus, to study how regulatory elements and cellular environment control expression variability, it is important to take a synthetic biology, MPRA-like, approach. The ideal method would allow measurements of thousands of regulatory elements' ability to control expression variability, while controlling the protein coding region. However, most of the current studies of cell-to-cell variability rely on individual measurements within a population, a critical need in understanding expression variability regulation is a new high-throughput method for measuring regulatory elements' effect on cell-to-cell variability.

1.8 Towards a Single Cell Level Understanding of Gene Regulation

The conceptual basis of this thesis is that one method can address the technical gaps presented in Section 1.6 and Section 1.7. Understanding how cellular contexts (hence cell types and substates) influence regulatory activity requires measurements of cell type information as well as the activity of regulatory elements. Incidentally, understanding how regulatory sequences control expression noise requires measurements of the activity of regulatory elements at the single-cell level. Both goals require the development of a category of technology that combines single-cell sequencing with functional genomics. This technology ideally can provide two readouts: (1) the regulatory activity of many non-coding sequences at the single-cell level, and (2) the transcriptome associated with each cell. This technology would enable many new inquiries and address many questions in single-cell functional genomics.

First, this method can falsify the hypothesis in Section 1.6 concerning whether cellular context confers specificity for gene regulation. One possible experimental setup is to test if promoters in genes that are only active in one cell type have activities in other cell types. If the hypothesis is true, then the additional cellular context information could help us to improve the specificity of gene

expression models.

Second, this method can study whether variants have cell-type-specific effects. The majority of heritable variation for human diseases maps to the non-coding portions of the genome [81–84]. The interpretation and functional validation of those variants have been extremely challenging. Partially the difficulty arises from the belief that many of those variants have cell-type-specific effects [85, 86]. Testing those non-coding variants in cell lines may not reflect the *in vivo* regulatory effects. Using this category of methods, it is possible to assay the variant effects in the relevant cell types *in vivo*. The results from this line of inquiry will improve our understanding of how variants are associated with human disease.

Thirdly, this method can study how non-coding regulatory elements control expression variability. It is possible to extract sequence features that control not only expression mean but also expression variability if a large number of regulatory elements are measured. Moreover, with the transcriptome information at the single-cell level, we could also examine how global cellular information contributes to expression variability. With that knowledge, we will gain more insights into the paradox of precise regulation of gene expression and its variable nature.

To achieve the aforementioned goals, the critical step is to develop single-cell functional genomics tools.

1.9 Scope of Thesis Work

The central goal of my thesis is to develop high-throughput functional genomics methods at the single-cell level. In this thesis I have developed two single-cell functional genomics methods.

In Chapter 2, I described scMPRA, the method I have developed for multiplexed reporter assay that measures regulatory activity for hundreds of regulatory elements in mixed cell types and at single cell resolution. scMRPA fills a critical technology gap in genomics. It expands MPRAs into directly testing cell-type-specific and cell-state-specific regulatory element activity in multicellular systems. It also enables a high-throughput readout of single-cell expression activity. I measured the

activity of 676 core promoters in a mixture of K562 and HEK293 cells at single cell resolution. I showed the method is reproducible and agrees with orthogonal measurements. On the cell type level, we found that a small number of core promoters are differentially expressed in K562 cells compared to HEK293 cells. Many of those promoters are CpG island promoters. On the cell substate level, we found that the cell cycle phase has a large and consistent effect on promoter activity. Rare cell substates, like the CD38 positive "stemness-like" state, also has a consistent effect on diverse core promoters. To extend scMPRA into a complex tissue, I tested a library of Gnb3 variants in newborn mouse retina. I found the method is sensitive and reproducible in complex tissue. We found about 30% of intentionally designed variants are cell type specific. Lastly, we examined whether single-cell expression difference is controlled by intrinsic factors or extrinsic factors. We found developmental promoters, which are differentially active in different cells, are more affected by global cellular factor fluctuations. Taken together, I developed a novel method that could have wide applications in functional genomics.

In Chapter 3, I described scTRIP, the method I developed in collaboration with Avinash Ramu and Clarice KY Hong in the Cohen Lab. This method measures a single reporter's activity when integrated into hundreds of genomic locations at single cell resolution. We measured the activity of a CMV-BFP reporter at 900 mapped genomic locations in K562 cells. We discovered sequence features around the integration site that explain the observed difference in single-cell variability. We trained a logistic regression model, and found the model has 0.77 accuracy. We also quantified the extrinsic portion of the noise, and discovered that the global cellular fluctuation contributes to roughly 50% of the total expression variability. We identify one of the large factors is that cells in the "stem-like" substate have much higher mean expression and lower variability compared to the "differentiated" substate. Incorporating this information into our logistic regression model, we increased the accuracy to 0.88. Taken together, we developed a novel method that could measure how genomic environment and cellular context control cell-to-cell variability.

Chapter 2: A single-cell massively parallel reporter assay detects cell type specific cis-regulatory activity

Siqi Zhao[1,2], Connie A Myers[3], Clarice Kee Yit Hong[1,2], David M Granas[1,2], Mike A. White[1,2], Joseph C. Corbo[3], and Barak A. Cohen[1,2]

1 The Edison Family Center for Systems Biology and Genome Sciences, 2 Department of Genetics, 3 Department of Pathology and Immunology, Washington University School of Medicine, USA.

This work was done in collaboration with Connie Myers, Clarice Hong, David Granas, Mike White, Joe Corbo, and Barak Cohen. I conceived this project. Barak Cohen, Joe Corbo, Mike White, and I designed the experiments. Connie Mayers carried out the retinal electroporation. Clarice Hong conducted the motif analysis. David Granas performed electroporation in HEK 293 cells. I carried out all other experiments and analyses. Barak and I wrote the paper. This chapter is under revision in the journal *Nature Genetics*.

2.1 Abstract

Massively parallel reporter gene assays are key tools in regulatory genomics, but cannot identify cell-type specific regulatory elements without performing assays serially across different cell types. To address this problem, we developed a single-cell massively parallel reporter assay (scMPRA) to measure the activity of libraries of cis-regulatory sequences (CRSs) across multiple cell-types simultaneously. We assayed a library of core promoters in a mixture of HEK293 and K562 cells and showed that scMPRA is a reproducible, highly parallel, single-cell reporter gene assay. Our results show that housekeeping promoters and CpG island promoters have lower activity in K562 cells relative to HEK293, which reflects developmental differences between the cell types. Within K562 cells, scMPRA identified reporter genes that are specific for the CD34+/CD38- "stem-like" sub-state. We then applied scMPRA to a complex tissue. We tested a library of Gnb3 promoter variants in newborn mouse retina. We found that our method is sensitive, reproducible *ex vivo*. And we found many variants have cell type specific effects. Finally, we deconvolved the intrinsic and extrinsic components of promoter cell-to-cell variability and found that developmental promoters have a higher proportion of extrinsic noise compared to housekeeping promoters, which reflects the responsiveness of developmental promoters to the cellular environment. We anticipate scMPRA will be widely applicable for studying the role of CRSs across diverse cell types.

2.2 Introduction

The majority of heritable variation for human diseases maps to the non-coding portions of the genome[24, 81–83, 87, 88]. This observation has led to the hypothesis that genetic variation in the cis-regulatory sequences (CRSs) that control gene expression underlies a large fraction of disease burden[89–92]. Because many CRSs function only in specific cell types[93], there is intense interest in high-throughput assays that can measure the effects of cell-type-specific CRSs and their genetic variants.

Massively Parallel Reporter Assays (MPRAs) are one family of techniques that allow investigators to assay libraries of CRSs and their non-coding variants *en masse*[36, 37, 94–98]. In an MPRA experiment, every CRS drives a reporter gene carrying a unique DNA barcode in its 3' UTR, which allows investigators to quantify the activity of each CRS by the ratio of its barcode abundances in the output RNA and input DNA. This approach allows investigators to identify new CRSs, assay the effects of non-coding variants, and discover general rules governing the functions of CRSs[34, 36, 39, 99–101]. One limitation of MPRAs is that they are generally performed in monocultures, or as bulk assays across the cell types of a tissue. Performing cell-type specific MPRAs in tissues will require methods to simultaneously readout reporter gene activities and cell type information in heterogeneous pools of cells.

To address this problem, we developed scMPRA, a procedure that combines single-cell RNA sequencing with MPRA. scMPRA simultaneously measures the activities of reporter genes in single cells and the identities of those cells using their single-cell transcriptomes. The key component of scMPRA is a two-level barcoding scheme that allows us to measure the copy number of all reporter genes present in a single cell from mRNA alone. A specific barcode marks each CRS of interest (CRS barcode, "cBC") and a second random barcode (rBC) acts as a proxy for DNA copy number of reporter genes in single cells (**Figure. 2.1 a**). The critical aspect of the rBC is that it is complex enough to ensure that the probability of the same cBC-rBC appearing in the same cell more than once is vanishingly small. In this regime, the number of different cBC-rBC pairs in a single

cell becomes an effective proxy for the copy number of a CRS in that cell. Even if a cell carries reporter genes for multiple different CRS, and each of those reporter genes is at a different copy number, it is still possible to normalize each reporter gene in each individual cell to its plasmid copy number. With this barcoding scheme, we can measure the activity of many CRSs with different input abundances in single cells.

2.3 Results

2.3.1 scMPRA enables single-cell measurement of CRS activity

As a proof of principle, we used scMPRA to test whether different classes of core promoters show different activities in different cell types. Core promoters are the non-coding sequences that surround transcription start sites, where general cofactors interact with RNA polymerase II [40, 51]. Core promoters are divided into different classes by the functions of their host genes (housekeeping vs developmental), as well as by the sequence motifs they contain (TATA-box, downstream promoter element (DPE), and CpG islands). We selected 676 core promoters that we previously tested [51] and cloned them into a double-barcoded MPRA library. In the first stage of library construction each core promoter reporter gene was represented by 10 unique cBCs. We then added rBCs to the library by cloning a 25 nt random oligonucleotide (oligo) directly downstream of the cBCs. The library contains $1.4X10^7$ unique cBC-rBC pairs (**Methods**, **Figure. 2.1 b**). Using this complexity, we calculated that the probability of plasmids with the same cBC-rBC pair occurring in the same cell is less than $2 \cdot 10^{-3}$ with our transfection protocols (**Methods**). Given this low likelihood, the number of rBC per cBC in a cell represents the copy number of a CRS in that cell. Knowing the copy number of CRSs in single cells allows us to normalize reporter gene expression from each CRS to its copy number in individual cells.

We performed a cell mixing experiment to test whether scMPRA could measure cell type specific expression of reporter genes. We transfected K562 and HEK293 cells (**Methods**), and

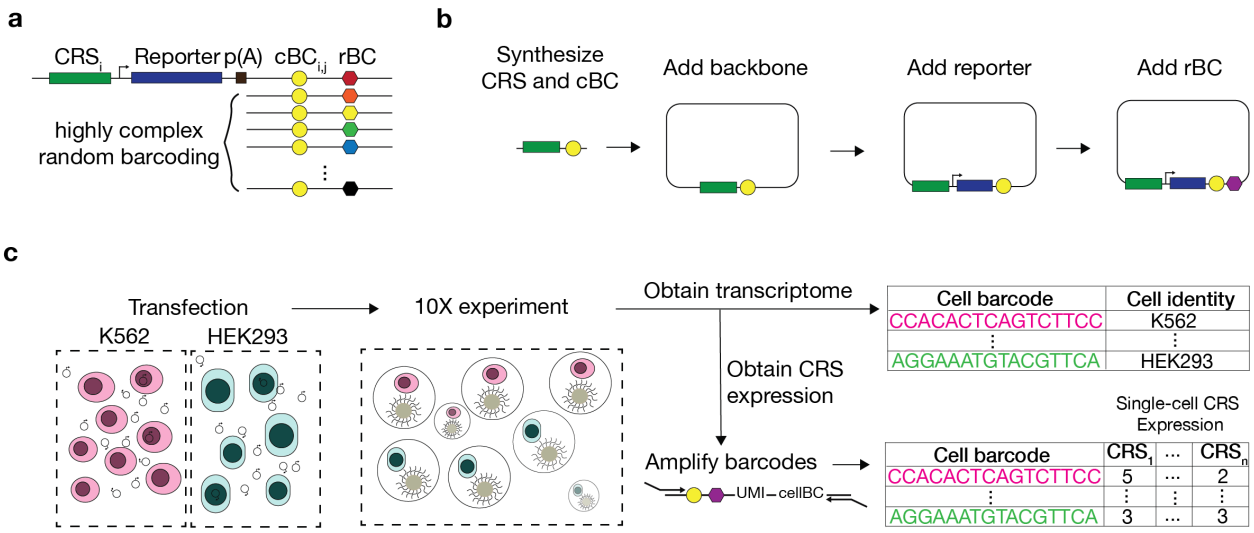

Figure 2.1: **scMPRA measures CRS at single-cell resolution.** (a) Each CRS reporter construct is barcoded with a cBC that encodes the identity of the CRS, as well as a highly complex rBC. The complexity of the cBC-rBC pair ensures that the probability of identical plasmids being introduced into the same cell is extremely low. (b) Cloning strategy for the double barcoded library. CRSs and their corresponding cBCs are synthesized together and cloned into an appropriate backbone. 25 nt rBCs are introduced to the plasmids with Hifi assembly. (c) Experimental overview for scMPRA using mixed cell experiment as an example. K562 cells and HEK293 cells are transfected with the double-barcoded core promoter library. After 24 hours, cells were harvested and mixed for 10X scRNA-seq. Cell identities were obtained through measuring the single transcriptome, and single-cell expression of CRSs was obtained by quantifying the barcodes. The cell identity and CRSs expression were linked by the shared 10X barcodes.

performed scMPRA on a 1:1 mixture of those cell lines (**Figure. 2.1 c**). We harvested cells and prepared them for sequencing using the 10X Chromium™ platform. The mRNA from single cells was captured, converted to cDNA, and pooled together. We then split the samples, with a quarter of the amplified cDNA library used for amplifying the cBC-rBC pairs and three-quarters used to amplify the transcriptome. The resulting reporter barcode abundances and transcriptome of each single cell are linked by their shared 10X cell barcode (**Methods**).

We recovered a total of 3112 cells (1524 in replicate 1 and 1588 in replicate 2) that are unambiguously assigned to one of the two cell types (**Figure. 2.2 a**, **Figure. 2.7 a,b**). We determined the efficiency of our method by calculating the recovery rate of our input promoters. We then calculated

the core promoter expression by taking the average of the cBC expression for the same promoter. We found that scMPRA recovered 99.5% (673 out of 676 core promoters) of the input library for K562 cells and 100% (676 out of 676 core promoters) for HEK293 cells, highlighting the efficiency of our method for recovering input elements.

We next calculated the number of individual cells in which each core promoter is measured.

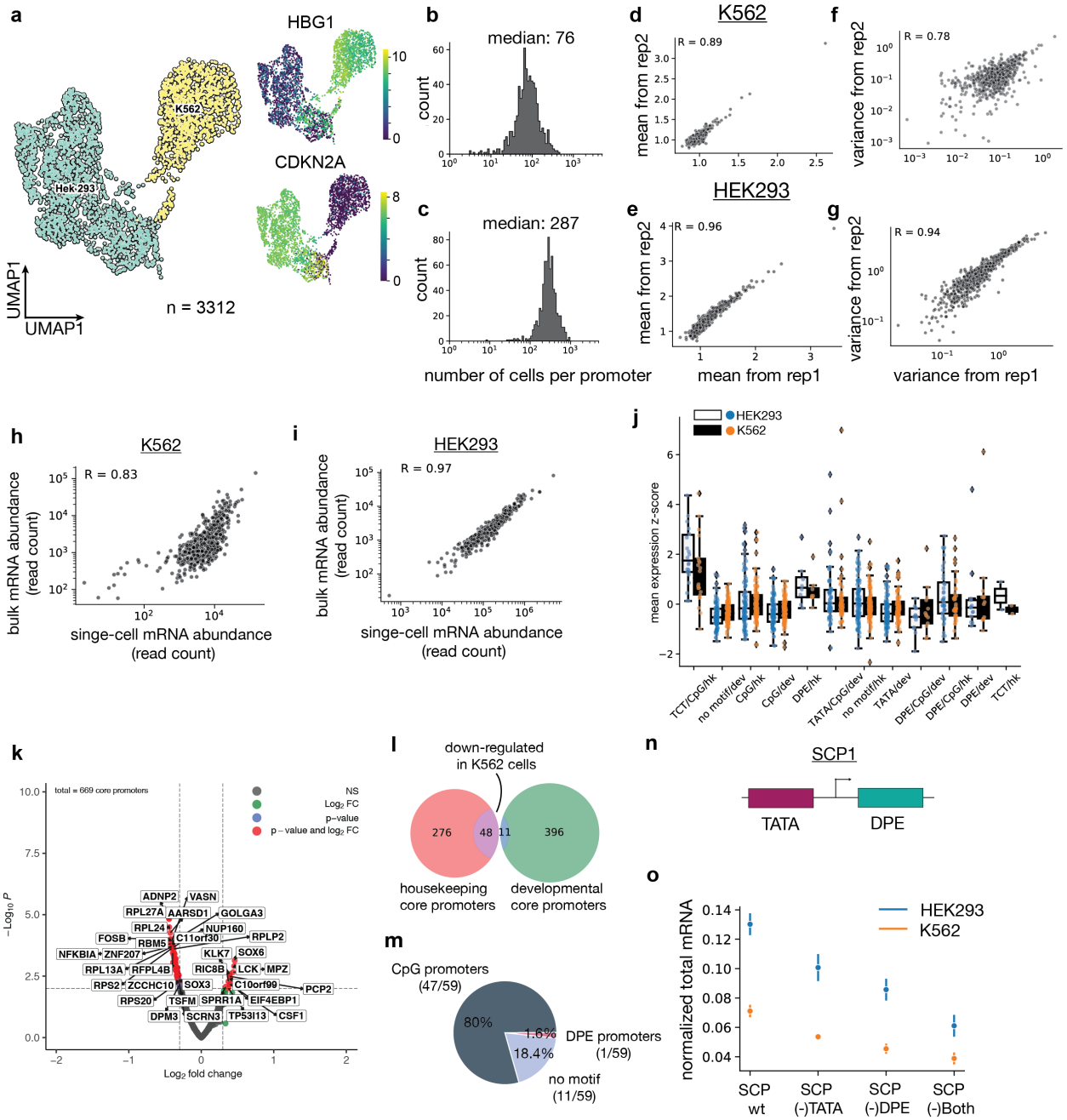

Figure 2.2: **scMPRA detects cell type specific CRS activity.** (a) UMAP of the transcriptome from the mixed cell scMPRA experiment. 3312 out of 3417 cells are assigned to either K562 or HEK293 cells. Cell-type specific genes were used to identify the cell clusters (HBG1 for K562 cells and CDKN2A for HEK293 cells). Cells are labeled by their cell type. (b,c) Histogram of the number of cells per core promoter for HEK293 and K562 cells. (d-g) Reproducibility for expression mean and cell-to-cell variance for both K562 and HEK293 cells (Correlation was calculated using Pearson's R). (h,i) Scatterplot of reproducibility of scMPRA mean expression with bulk MPRA measurement using read count normalization (Correlation was calculated using Pearson's R). (Caption next page.) (j) Boxplot of mean expression from different categories of core promoters in K562 (orange) and HEK293 (blue) cells. Box plot elements: center line: median; limits: 25% quartile (Q1) and 72% quartile (Q3); whiskers, from Q1-1.5 x IQR to Q3 + 1.5 x IQR, where IQR is the interquartile range. (k) Volcano plot for differential expression (DE) of the core promoters in K562 and HEK293 cells (Significant DE reporters have p-value <0.01 and log-2 fold change greater than 0.3). (l) A Venn diagram of the functional characterization (housekeeping vs developmental) of down-regulated reporters in K562 cells. Housekeeping promoters are enriched (p-value = 1.08×10^{-11}, hypergeometric test). (m) Pie chart of the sequence features (CpG, DPE, TATA) of down-regulated reporter genes. CpG promoters are enriched (p=2.18×10^{-6}, hypergeometric test). (n) Schematic of SCP1 binding sites. (o) Expression of wild-type and mutated (TATA-, DPE-, and Both) versions of SCP1 core promoter (error bar: 1 s.d.)

We found that the empirical distribution of the number of cells per core promoter is log normal, with a median of 76 cells per core promoter for K562 cells and 287 cells per core promoter for HEK293 cells (**Figure. 2.2 b,c**). Given that the number of pBC-rBC pair is effectively the number of plasmids per cell, we also calculated the number of plasmid per cell, and found that fewer number of plasmids were incorporated into K562 cells compare to HEK293 cells (median plasmid number in K562 cells: 164, median plasmid number in HEK293 cells: 341. **Supplemental Figure. 2.7 c,d**). The difference in transfection efficiency between these cell types with the same input likely reflects global cellular differences between them, and is representative of the condition when performing scMPRA in different cell types.

We calculated the biological reproducibility and found that scMPRA is highly reproducible in both cell types for measurements of mean expression (K562: Pearson R = 0.89, HEK293: Pearson R = 0.96) and cell-to-cell variance (K562: Pearson R = 0.78, HEK293: Pearson R = 0.94, **Figure.**

2.2 d-g). To validate the measurements, we conducted bulk RNA-seq for the core promoter library in the two cell types separately, and found the bulk measurements correlate well with the aggregated single-cell measurements (**Figure. 2.1 h,i**, **Figure. 2.7 e,f**). This analysis shows that single-cell measurements of library members in as few as 70 individual cells still correlate well with bulk measurements, highlighting the sensitivity of our method.

2.3.2 scMPRA detects cell type specific CRS activity and non-coding variant effect

We asked whether the data allowed us to detect core promoters with differential activity between K562 and HEK293 cells. While different classes of core promoters had similar activities in both cell lines (**Figure. 2.2 j**), our differential analysis using DEseq2[102] identified a small number of promoters (11 out of 669) that are upregulated in K562 cells, and 59 promoters that are downregulated in K562 cells (adjusted $p < 0.01$, log2 fold change > 0.3, **Figure. 2.2 k**). Among the down-regulated promoters, 48 out of 59 core promoters belong to housekeeping genes ($p = 1.08 \times 10^{-11}$, **Figure. 2.2 l**), and 46 out of 59 core promoters are CpG-island-containing core promoters ($p = 2.18 \times 10^{-6}$, **Figure. 2.2 m**). This down-regulation might be explained by the fact that the K562 cell line is a cancer derived cell line, and a hallmark regulatory change in cancer cells is the hypermethylation of CpG promoters[103]. These results demonstrate the ability of scMPRA to detect CRSs with cell-type specific activities.

Another application of scMPRA is to detect cell type specific effects of non-coding variants. To test whether our method can detect the effects of mutations in a given CRS, we included an artificial core promoter SCP1[104] along with mutated versions without a TATA Box or DPE motif in our library (**Figure. 2.2 n**). We first computed the total number of captured reporter gene transcripts, since it is the closest proxy to the bulk expression measurement. We found that deletions of the TATA motif or DPE motif both reduced expression (**Figure. 2.2 o**) and we observed a similar trend in the bulk data (**Figure. 2.7 g**). When we directly calculated the mean of the single-cell expression distribution instead of total number of captured reporter gene transcripts, we found that the deletion of the DPE motif has a stronger effect in K562 cells than in HEK293 cells (40% reduction vs 20% reduction) (Methods, **Figure. 2.7 h,i**). We hypothesized that the differential expression of transcription factors between K562 and HEK293 cells leads to differential sensitivity to the TATA and DPE motifs. We examined the single-cell transcriptome and found that TAF9, which recognizes the DPE motif[105], is more highly expressed in K562 cells compared to HEK293

Cells (Supplementary Fig. 1j, Wilcoxon $p=4.27x10^{-94}$). This observation likely explains why the deletion of the DPE motif has a stronger effect in K562 cells. Our results demonstrate that scMPRA can identify and explain cell-type specific effects of non-coding variants.

2.3.3 scMPRA detects cell sub-state specific CRS activity

Single-cell studies have revealed heterogeneity in cell states even within isogenic cell types[70, 106–108]. Therefore, we asked if scMPRA can identify CRSs with cell-state specific activity. We repeated scMPRA on K562 cells alone and obtained a total of 5141 cells from two biological replicates. Measurements of the mean and variance of each library member were again highly correlated between replicates and agree well with independent bulk measurement (**Figure. 2.8 a-d**).

As the phases of the cell cycle represent distinct cell-states, we asked whether scMPRA could identify reporter genes with differential activity through the cell cycle. We assigned cell cycle phases to each cell using their single cell transcriptome data (**Figure. 2.3 a**) and then calculated the mean expression of each reporter gene in different cell cycle phases. We found that most core promoters in our library are upregulated in the G1 phase of the cell cycle, and some housekeeping promoters are highly expressed through all cell cycle phases (**Figure. 2.3 b**). We also identified core promoters with different expression dynamics through the cell cycle. For example, we found the core promoter for UBA52 remains highly expressed in the S phase, whereas the core promoter for CXCL10 is lowly expressed throughout (**Figure. 2.8 e**). This analysis illustrates the ability of scMPRA to identify CRSs whose expression naturally fluctuates with cellular dynamics.

We then asked whether scMPRA could detect reporter genes with activities that were specific to other cell-states in K562 cells, after normalizing for cell cycle effects. We focused on two specific sub-states that have been reported and experimentally validated for high proliferation rates in K562 cells[109, 110]. The first is the CD34+/CD38- sub-state that has been identified as a leukemia stem-cell subpopulation, and the second is the CD24+ sub-state that is linked to selective activation of proliferation genes by bromodomain transcription factors[106, 107]. To identify these sub-states in our single-cell transcriptome data, we first regressed out the cell cycle effects and confirmed that the single cell transcriptome data no longer clustered by cell cycle phase (Supplementary Fig.2 f). We then identified clusters within K562 cells that have the CD34+/CD38- expression signature, or the CD24+ signature (**Figure. 2.3 c,d**). Although the CD34+/CD38- cells represent only 9.3% of the

cells in a K562 culture, scMPRA revealed two distinct classes of core promoters that are upregulated and downregulated in these cells respectively (**Figure. 2.3 e**). Conversely, the expression patterns of promoters are similar between the CD24+ cluster and cells in the "differentiated" cluster (**Figure. 2.3 e,f**). Motif analysis of the up/down regulated classes of promoters in CD34+/CD38- cells showed that different core promoter motifs are enriched in each class, with the TATA box and Motif 5 being

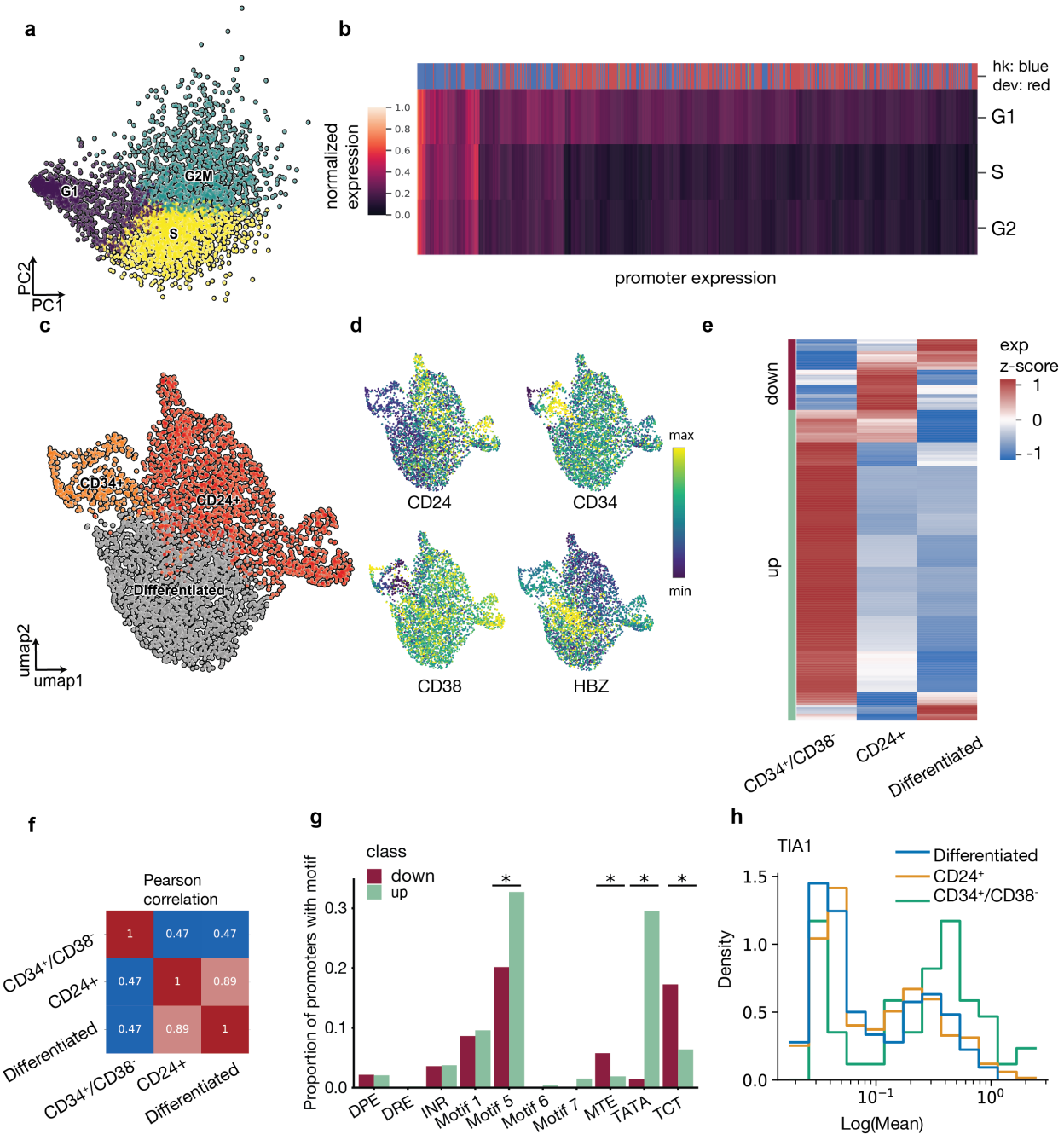

Figure 2.3: **scMPRA detects cell sub-state-specific CRS activity** (a) PCA plot of K562 cells classified based on the cell cycle score. (b) Heatmap of reporter expression in different cell cycle phases (Color bar indicates housekeeping (blue) vs developmental (red) promoters). (c) Representative expression dynamics of reporter genes through cell cycle for UBA52, CSF1, and CXCL10. (d) UMAP embedding of K562 cells with high proliferation sub-states (CD34+/CD38- and CD24+). (e) Marker gene expression signifies different cell sub-states in K562 cells. CD34, CD38 marks the "leukemia stem cell" sub-state; CD24 marks a high proliferation sub-state, and HBZ marks the differentiated leukemia sub-state; left color bar: hierarchical clustering showing 2 clusters based on expression pattern in the three substates. (f) Heatmap showing the correlation matrix of core promoter expression in three substates (CD34+/CD38-, CD24+, and Differentiated). (g) Proportion of promoters in each cluster that contains the indicated core promoter motif. * represents significant enrichment in one cluster over the other ($p < 0.05$, two-sided Fisher's exact test). (h) Histogram of single-cell expression of TIA1 promoter in three substates.

enriched in the upregulated class and MTE and TCT motifs being enriched in downregulated class (**Figure. 2.3 g**, Methods). This result suggests that differences in core promoter usage might be driving the differences between CD34+/CD38- and the other clusters. Because the TATA box is mostly found in developmental core promoters, the CD34+/CD38- subpopulation likely reflects a more "stem-like" cellular environment in these cells. Our analysis highlights the ability of scMPRA to identify CRSs with differential activity in rare cell populations.

With the single-cell expression data, we asked how certain promoters achieve higher expression in the CD34+/CD38- state. We asked whether the single-cell expression distribution for the CD34+/CD38- state is shifted higher than for the other states, or if the range of expression is the same for each sub-state, with only the proportion of cells with high expression changing in each state. To answer this question, we calculated the proportion of cells in each sub-state belonging to the 90th percentile of the total single cell expression distribution. For the majority of promoters, the CD34+/CD38- cluster has a much higher proportion of cells in the 90th percentile (**Figure. 2.9 a**). At the same time, there is no difference in the maximum expression of cells in different sub-states, and this maximum level is mainly set by the promoter identity (**Figure. 2.9 b**). Even for the most differentially expressed promoter in the CD34+/CD38- subpopulation, TIA1, the

expression distributions for cells in the three sub-states cover the same range, but the proportion of cells in the right-tail of the distribution is higher for CD34+/CD38- cells (**Figure. 2.3 h**). This result suggests that the "stem-like" cellular environment of the CD34+/CD38- subpopulation increases the probability of certain promoters having higher expression, without shifting the maximum expression those promoters achieve. Taken together, these analyses highlight how the joint transcriptome and CRS measurements in scMPRA can be used to understand differences in behavior in cellular sub-states.

2.3.4 scMPRA is reproducible and accurate in the mouse retina

To demonstrate that scMPRA is applicable to a complex tissue with multiple cell types, we performed scMPRA in new borne mouse retina explant [37, 39]. We designed a new library consists 114 variants with different variants of the full-length Gnb3 promoter. We chose this library because it had been demonstrated that specific mutations in the Gnb3 promoter have cell-type specific effects [63]. We designed the reporter library with a cBC for each variant. We then added a 25 nt random oligonucleotide downstream of the cBC. To ensure that we could detect the unexpressed variants, we also included a U6 promoter driven cBC that is identical to the cBC that is 3' to the promoter variant. U6 promoter utilizes Pol III transcription system and should not interfere with the Gnb3 promoter expression (**Figure. 2.4 a**). We transfected this library to 3 pairs of dissected newborn mouse retinas per replicate and cultured it for 8 days and performed scMPRA on the dissociated cells (**Figure. 2.4 a**, **Methods**).

We first analyzes the transcriptome data and found that we could recover several cell types from the transfected retina. We obtained a total of 22161 cells from both replicates with a mean of 22161 reads per cell and 1642 genes per cell. We recovered rod photorecepters (87.3%), bipolar cells (3.5%), interneurons(5.2%), and Mueller glia cells (3.9%) (**Figure. 2.4 b**, **Figure. 2.10 a**, **Figure. 2.11 a-c**). Those discovered clusters have very similar transcriptome signatures with previous single-cell RNA-sequencing datasets on the retina (**Figure. 2.10**, **Figure. 2.11 b**). We then process the barcode sequencing data, and found a total of 6232 cells that contain enough reads associated with the barcode library. We look at the transfectability of each cell type. We calculated the number of plasmids per cell per cell type and found that each cell intakes several hundreds of plasmids (**Figure. 2.4 c**). These observations demonstrate that scMPRA can be applied to a complex tissue and study many regulatory elements in parallel.

To calculate the sensitivity of scMPRA in the mouse retina, we subsampled the Rod photoreceptor cells and calculated the reproducibility for expression mean and variance for the different numbers of cells. We found that the minimum number of cells required for obtaining reproducible mean

measurement is around 50 cells, and the minimum number of cells required for obtaining reproducible variance measurement is 250 cells (**Figure. 2.4 d**). All of the cell types we discovered from the retina have more than 50 cells per cell type per replicate. We calculated the reproducibility for each cell type between two biological replicates. We found that the mean expression is very reproducible for all cell types (Rod cells: Spearman R: 0.98, p-value = $3.3 \cdot 10^{-83}$, Bipolar cells: Spearman R: 0.85, p-value = $7.7 \cdot 10^{-34}$, Mueller glia cells: Spearman R: 0.79, p-value = $5.5 \cdot 10^{-26}$, Interneurons: Spearman R: 0.82, p-value = $3.3 \cdot 10^{-29}$, **Figure. 2.4 e**). Agreeing with our sensitivity analysis, we obtain excellent reproducibility for expression variance in Rod cells (Spearman R: 0.94, p-value = $2.2 \cdot 10^{-55}$), but moderate reproducibility for Bipolar cells (Spearman R: 0.69, p-value = $1.6 \cdot 10^{-17}$,**Figure. 2.4 f**). We also found the expression differences for the wild-type Gnb3 promoter that is barcoded twice is 1.2. Given the dynamic range over 200-fold, scMPRA provides excellent resolution.

To examine whether our reporter assay captures the endogenous gene expression, we calculated the Gnb3 gene expression from our transcriptome data, and we found the reporter expression agrees with the endogenous Gnb3 expression (Spearman R: 0.95, p-value = 0.006, **Figure. 2.4 g**). We next examined whether our method recapitulates the previously measured cell-type specific mutation effect. A previous study has demonstrated that the altering of two CRX binding sites into Q50 transcription factor binding sites reduces the expression in Bipolar cells but not in rod photoreceptors, we also observed the same trend (**Figure. 2.4 h,i**). Taken together these observations demonstrate that scMPRA can be applied to complex tissue and our measurements in the newborn mouse retina are sensitive, reproducible, and agree with previous observations.

2.3.5 scMPRA detects many cell-type specific variant effects

We next examined whether the variants we designed have cell-type specific effects (**Figure. 2.4 a**). We applied The χ^2 test and found that 30 out of the 114 variants (30%) are cell type specific (Methods). This result is surprising but is reasonable given we designed those mutations specifically to disrupt key transcription factor binding sites. This suggests that scMPRA is a powerful tool to detect cell-type specific variant effects.

Studies have demonstrated that the disruption of individual CRX sites disrupts the regulatory activity of Gnb3 promoters [63]. In our data, we found our results largely agree with the previous studies in rod and bipolar cells. Disruptions of CRX1, CRX2, and CRX3 sites have small repressive effects on Gnb3 expression, whereas the disruption of CRX4 has a large effect in bipolar and rod cells (**Figure. 2.4 b**). To our surprise, the disruption of CRX4 increases the expression only in Mueller Glia cells. We also included pair-wise mutations of CRX sites in the Gnb3 promoter. Examining the pair-wise mutation data, we found that the mutation of the CRX4 site in combination with another CRX site reduces expression, and suggests that the effect of CRX4 binding site is epistatic to other CRX binding sites. Similarly, when we swapped the CRX binding sites with a Q50 binding site, we observed a general increase of expression, except for the change of the CRX4 site (**Figure. 2.4 c**). A similar trend is also observed in pair-wise swaps of CRX sites into Q50 sites. All pair-wise swaps increase Gnb3 expression, except for any disruption of CRX4 sites. This suggests that the CRX4 site is essential for the proper expression of Gnb3 promoter.

We next examined whether the disruption of an E-box binding site in the Gnb3 promoter has cell-type specific effects. We chose to closely examine the E-box site because E-box motifs are bound by many bHLH transcription factors. Many of those transcription factors, such as NeuroD1 and Math5, are critical for retinal development [111]. We found the mutation of the central base pairs of the E-box into As reduces the expression across cell types, agreeing with the studies demonstrating the 5'-CAAATG-3' motif is a version of the non-binding E-box. We found many of the Ebox mutations have cell-type specific effects (14 out of 33, χ^2 test, adjusted p-value < 0.01,

Figure. 2.4 d). To broadly test whether non-coding regions have cell-type specific effects, we tiled through the conserved region of the gnb3 promoter and shuffled every 5 bps (**Figure. 2.4 e**). We found many cell-type specific variant effects in the first conserved region that is far away from the TSS, whereas no cell-type specific variants within 100 bp around the TSS. This agrees with the results that the core promoter regions (100bp around the TSS) are mainly bound by general transcription factors and are unlikely to be cell type specific [112]. These observations demonstrate that non-coding variants could have very different effects on different cell types, agreeing with many cell-type specific eQTL studies [113].

Besides the cell-type specific variant effects, we also found many variants have uniform effects across cell types. The first such uniform effect is when we changed each CRX binding site into high, medium, and low-affinity sites. We found the change of site affinity has a uniform effect on Gnb3 expression, with the exception of CRX5 sites (**Figure. 2.4 f**). The likely explanation for the specific effect of the CRX5 site is that the disruption of the CRX5 site increases expression, suggesting the site is likely repressive. But the overall trend agrees with the framework that the overall occupancy of a promoter determines its regulatory activity [32]. Another general trend we discovered is that the expression of all variants that are not cell-type specific have lower expression in Mueller glia and interneuron cells. This tracks with the level of endogenous Gnb3 expression. This observation suggests that the correct suite of transcription factors in different cellular contexts set the range of the possible regulatory activity for different regulatory elements, and the variants within the regulatory element modulate the expression level within the range set by the cellular context.

These observations demonstrated that scMPRA could be a powerful tool to functionally test cell-type specific variant effects as well as provide insights into how generalizability and specificity are achieved for gene regulation.

2.3.6 scMPRA enables decomposition of intrinsic and extrinsic noise

Finally, we analyzed the cell-to-cell variability of reporter genes across K562 cells. Cell-to-cell variability, or expression noise, is the phenomenon where gene expression varies among the cells of an isogenic population. Expression noise has important roles in development[67], rare-cell cancer resistance [70, 114], and its origin is a central question in single-cell biology. A common framework for studying expression noise is to decompose it into its intrinsic component, which arises from the thermal fluctuations of macromolecular interactions, and its extrinsic component, which results from fluctuations in the global cellular environment [2, 33, 80, 115, 116]. Intrinsic and extrinsic noise can be decomposed using dual-reporter experiments, where two identical reporter genes are measured across the same single-cells [80]. High covariance of the two reporter genes indicates high extrinsic noise and low intrinsic noise, while independent variation of the two reporters suggests high intrinsic noise and low extrinsic noise. In scMPRA, plasmids with the same CRS but different barcodes are sometimes incorporated into the same cells, effectively serving as a dual-reporter experiment. We extracted pair-wise expression for the same core promoter labeled with different cBCs from our scMPRA data, and computed intrinsic noise and extrinsic noise using a previously developed statistical framework [117] (Methods). We found that different core promoters have distinct intrinsic and extrinsic noise profiles (**Figure. 2.6 a,b**). Globally, we found that intrinsic noise correlates with mean expression levels (Pearson $\rho = 0.455$), while extrinsic noise is not correlated with mean expression (Pearson $\rho = -0.172$, **Figure. 2.6 c,d**). This result agrees with the notion that intrinsic noise arises from the thermodynamics of transcription at different promoters, whereas many sources for extrinsic noise are independent of the specific promoters. We also found that developmental promoters have a higher proportion of noise that is extrinsic, reflecting their role in driving developmental promoters that respond to extrinsic cues during development (**Figure. 2.6 e,f**). This analysis suggests that scMPRA could be a powerful tool to study the mechanistic origin of cell-to-cell variability in a high throughput manner.

2.4 Conclusion

We have presented a method to measure the cell-type and cell-state specific effects of CRSs by devising a barcoding scheme to read out input copy number with mRNA. We demonstrated that scMPRA detects cell-type specific reporter gene activity in a mixed population of cells, and cell-state specific activity in an isogenic population. We also demonstrated that scMPRA can be a powerful tool to study how different CRS control cell-to-cell variability. The assay is reproducible and reports accurate mean levels of reporter gene activity in as few as 35 cells in a complex tissue.

A future direction is to determine the applicability of scMPRA to other complex complex tissues. Many tissues, such as the mouse intestine [118], mouse embryonic kidneys [119], mouse embryonic brain [120], and mouse embryonic spinal cord [121], can be electroporated with high tropism. With the burgeoning of Adeno-associated viral delivery (AVV) systems with distinct tropisms [122–125], it is possible to incorporate a library of reporter genes into mouse brain, heart, liver, and muscle tissues. We anticipate that scMPRA will be widely used to study cis-regulatory effects in a variety of complex tissues.

2.5 Methods

Cell culture

K562 cells were cultured using a medium consisting of Iscove's Modified Dulbecco's Medium (IMDM) + 10% Fetal Bovine Serum (FBS) + 1% non-essential amino acids + 1% pen/strep at 37 C with 5% of CO2. HEK293 cells were cultured using a medium consisting of Eagle's Minimum Essential Medium (EMEM) + 10% Fetal Bovine Serum (FBS) + 1% pen/strep at 37 C with 5% of CO2.

Cloning Strategy

We developed a two-level barcoding technology to enable single-cell normalization for plasmid copy number. We applied this strategy to a promoter library we previously tested in bulk assays24. The original library contains 676 core promoters with a length of 133bp. Each core promoter has 10 promoter barcodes to provide redundancy in the measurements. We then synthesized a single-stranded 90 bp DNA oligonucleotide containing a 25 bp random sequence, a restriction site, and 30 bp homology on each side of the barcode region.

We used Hifi Assembly™ to add the random barcodes to the plasmid library. 4 μg of the plasmid library were split into 4 reactions and digested with 2μl of SalI for 1.5 hours at 37□. The digested products were run at 100V for 2 hours on a 0.7% agarose gel. The correct size band was cut and purified with the Monarch Gel Extraction Kit (New England BioLabs T1020L). The insert single-stranded DNA was diluted in TE to a stock concentration of 100 uM. The insert was then further diluted to 1 uM with ddH2O. Three assembly reactions were pooled together, each reaction containing 100 ng of digested library backbone, 1 uM of insert DNA, 1μl of NEBuffer 2, 10 μl of 2X Hifi assembly mix, and H2O up to 20 ul, the reaction was incubated at 50°C for 1 hour. The assembled product was purified with the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L) and eluted in 12 μl of H2O.

The assembled plasmid was transformed using Gene Pulser Xcell Electroporation Systems by electroporation (BIO-RAD 1652661), 50 μl of ElectroMax DH10B electrocompetent cells

(Invitrogen 18290015) with 1 µl of hifi assembled product at 2 kV, 2000 Ω, 25 nF, with 1 mm gap. 950 µl of SOC medium (Invitrogen 15544034) was added to the cuvette and then transferred to a 15 ml Falcon tube. Two transformations were performed, and each tube was incubated at 37C for 1 hour on a rotator with 300 rpm. The culture was then added to pre-warmed 150 µl LB/Amp medium and grown overnight at 37 C. 1 µl of the culture was also diluted 1:100 and 50 µl of the diluted cultured was plated on a LB agar plate to check the transformation efficiency. For the core promoter library, we obtained more than 4X108 colonies, large enough to cover a complex library.

Estimating Library Complexity

To estimate the library complexity, we sequenced the DNA library using a nested PCR-based Illumina library preparation protocol. Briefly, we first used Q5 polymerase (New England BioLabs M0515) to amplify the region containing the two barcodes with SCARED P17 and SCARED P18 . The total reaction volume is 50µl with 50ng of plasmids with 2.5 µl of 10uM primer each. The annealing temperature is 61°C with an extension time of 10s. 25 cycles of amplification were done. The product was then purified with the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR (SCARED P19, SCARED P20), a 25 µl reaction was set up with 0.25 µl product from the previous step, the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCRDNA Cleanup kit. For the last PCR to add the P5 and P7 Illumina adapters, a reaction with 25 µl of total volume was set up with 2 µl of cleaned product from PCR2, a total of 10 cycles of PCR was done.

The constructed Illumina library was sequenced on an Illumina MiSeq. A total of 1,693,933 reads was generated for this library. A filtering strategy was applied to the raw reads, where reads that do not have matching promoter barcodes and wrong-length random barcodes were filtered out. We obtained a total of 1,359,176 reads (80% of the total reads) that contain the correct promoter barcode and correct length random barcode.

The shallow sequencing of the input plasmid library enabled us to estimate the library complexity

and the probability of two identical copies of the plasmid being transfected into the same cell. We first calculated that each random barcode is attached to 1.9 promoter barcodes on average. For a total of 6760 input promoter barcodes, this suggests that a given random barcode is being reused by 3200 different promoters. The reuse of random barcodes is the effective labeling complexity for the double-barcoding. For the Hifi assembly experiment, we used 300 ng input backbone plasmids containing only the promoter barcode ($4.5X10^9$ total copies and on average 6.65X106 copies of plasmids per promoter barcode). Given the effective labeling complexity, the average copy number of the plasmid containing the same promoter barcode-andom barcode pair is at most $2.08X10^3$. For the transfection experiment done in this study, with 2 μg ($6X10^9$ copy of plasmids) for cell mixing experiment and 10 μg ($3X10^{10}$ copy of plasmids) for K562 along experiment, the estimation of the average copy number for an identical plasmid is $4.4X10^2$ and $2.2X10^3$ respectively.

After obtaining the average copy number for identical plasmids, we estimate the probability of an identical plasmid being transfected into the same cell. We first define this probability as the collision rate. We note that the transfection of the identical copies of different plasmids are independent, so we could only calculate the collision rate for only one of such plasmids. The calculation of the collision rate for a given library member can be formulated as such: given the number of the identical copies of a plasmid, what is the probability of two or more of the copies being transfected into the same cell? We first write the expectation:

$$n^{-m}\sum_{k=0}^{n}\binom{n}{k}\sum_{q=0}^{(n-k)}\binom{n-k}{q}\binom{m}{q}q!m-qn-k-q_{n\geq 2}(n-k-q)!(m-q) \qquad (2.1)$$

where n denotes the total number of cells, m denotes the total number of identical plasmids, k denotes the number of cells with no plasmid, q denotes the cells with exactly 1 plasmid, parentheses denote binomial coefficient, and brackets denote partition function.

The above equation was simplified by substituting with the bivariate generating function, and the expected number is:

$$m(1-(\frac{n-1}{n})^{m-1}) \quad (2.2)$$

For a given transfection experiment, we can estimate the effective percentage of plasmid that is successfully transfected into the cell. Given the estimated copy number for identical plasmids is 4.4X102 and 2.2X103 for mixed cell experiment and K562 alone experiment respectively, the expected number of cells having more than 1 identical plasmid can be calculated with the aforementioned equation, and the probability of two copies of an identical plasmid appearing in the same cell is 0.0004 and 0.002 respectively. On a practical note, researchers have suggested that the effective number of the plasmid that are incorporated into the nucleus is about 0.01 - 0.1 of the input amount [126], hence a library containing around 2.5X105 different members transfected to 1 million cells has a theoretical collision rate around 1

Transfection

K562 cells were transfected using electroporation with the Neon transfection system (Invitrogen MPK5000). 1 million cells were transfected with 2 μg of plasmid DNA (mixed-cell experiment) or 10 μg of plasmid DNA (K562 sub-state experiment), with 3 pulses of 1450 V for 10 ms. The cells were then plated to pre-warmed K562 medium.

HEK293 cells were transfected using the Lipofectamine3000 protocol. 4 μl of p3000 reagent, 4μl of Lipofectamine, and OptiMEM were mixed with 2 μg of plasmid DNA to a volume of 250 μl. The lipofectamine reagents and plasmid were mixed and incubated at room temp for 15 minutes and then added dropwise to the cells.

Bulk RNA extraction and sequencing

We determined the optimal harvest time based on plasmid dilution and protein maturation and found the optimal harvest time is between 22 - 28 hours after transfection. The rationale behind the choice of time is to balance the transcription rate and the plasmid dilution during cell replication.

For both K562 cells and HEK293 cells, we harvested the cells after transfection at 24 hours, and proceeded to extract total mRNA with Qiagen RNeasy kit for K562 cells and Monarch Total RNA

miniprep kit for HEK293 cells. The reverse transcription was done with Superscript IV Reverse Transcriptase (Invitrogen 18090010). The final sequencing library was constructed using a nested PCR strategy. Briefly, we first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the 2 barcodes with SCARED P17 and SCARED P18. The total reaction volume is 50µl with 50ng of backbone with 2.5 µl of 10uM primer each. The annealing temperature is 61°C with an extension time of 10s. 25 cycles of amplification was done. The product was then purified with the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR using primers SCARED P19 and SCARED p20, a 25 µl reaction was set up with 0.25 µl product from the previous step, the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L). For the last PCR to add the P5 and P7 Illumina adapters, a reaction with 25 µl of total volume was set up with 2 µl of cleaned product from PCR2, a total of 10 cycles of PCR was done. The sequencing library was sequenced on an Illumina Mi-seq machine with other samples pooled in the same lane.

10X experiment for scMPRA

We harvested both K562 and HEK293 cells 24 hours after transfection, then followed the cell preparation protocol of 10X genomics. We used the 10X V3.1 chromium kit for our single-cell RNA-seq protocol. All PCRs were performed on an Invitrogen PCR machine. We targeted 2000 cells per replicate for each experiment for the mixed cell experiment. We targeted 2500 cells per replicate for the K562 substrate experiment. We followed the 10X protocol (https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v31-chemistry) with 12 cycles of cDNA amplification. To amplify the Capture-Sequence captured reads. 0.25 µl of 100 uM SCARED P32 was added to the cDNA amplification mix. For step 2.2, we modify the clean-up protocol by saving both the beads and supernatents and for the supernatents, we use a final concentration of 1.2X beads to pull down the DNA fragments. We then take 25% of both the 0.6X and 1.2X pull down products for the next step of PCR. To

construct the illumina sequencing library, we used a 3-step nested PCR strategy. Briefly, we first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the 2 barcodes with SCARED P17 and SCARED P18. We pooled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting. The annealing temperature is 61°C with an extension time of 10s. The product was then purified with the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR using the following 3 primers (SCARED P21, CAS PC2, CAS PP2), we pulled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting , the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCRDNA Cleanup kit (New England BioLabs T1030L). For the last PCR to add the P5 and P7 Illumina adapters (CAS P48, CAS PP4, CAS PC4), we pulled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting, a total of 10 cycles of PCR was done. The transcriptome is generated using the 10X Dual-Index Set TT expression kit (https://support.10xgenomics.com/single-cell-gene-expression/index/doc/technical-note-chromium-next-gem-single-cell-3-v31-dual-index-libraries).

The sequencing was done on an Illumina NextSeq machine. We used 40% of the barcode library, 40% of the balanced scRNA-seq transcriptome, and 20% Phi-X. Sequencing the constructed barcode library with transcriptome and Phi-X is crucial to reduce the sequencing error from the reporter constant sequence. On Read1, only 28 bps contains the 10X cell barcode and UMI wes amplified, to avoid sequencing the constant Poly(A) sequence; On Read2, 105 bps was sequenced. For the mixed experiment, we pool reads from a total of 2 runs of NextSeq High Throughput sequencing runs, and for the K562 cells, we pool 3 runs of NextSeq High Throughput runs.

scRNA-seq data processing

The single-cell RNAseq data were processed using Cellranger 6.0.1 (https://github.com/10XGenomics/cellranger) and Scanpy 1.8.1 [66] (https://github.com/theislab/scanpy) following the standard pipeline. Briefly, different sequencing runs from the same biological replicate were pooled together

and processed with CellRanger 6.1.1; the final output expression matrix was then imported into Scanpy for further normalization. We first removed cells with less than 1000 genes, and genes that were present in less than three cells. We then removed cells with high counts for mitochondrial genes. Next, we normalized the UMI counts to the total cell UMI counts. The normalized expression matrix was used for clustering and visualization with Scanpy. The clustering was done using the Leiden algorithm [127].

scMPRA data processing

The relevant script for processing single-cell MPRA reads can be found on a Github repository (https://github.com/szhao045/scMPRA). The final sequencing product for scMPRA with Read1 contains the cell and molecular information (cellBC and UMI), and Read2 contains the MPRA library information (cBC and rBC). First, we fuzzy-matched the constant sequences before and after both the promoter barcode and random barcode. In this step, we filtered out the reads without correct promoter barcode length, or random barcode length. To increase the speed, we wrote a stand-alone program (https://github.com/szhao045/scMPRA_parsingtools) written with Golang, and can be compiled to work on many operating systems. Second, we filtered out cell barcodes based on the cell barcode list from the CellRanger output barcode list, with error-correction with maximum hamming distance of 1. Third, to mitigate the effect of template-switching during the PCR steps, we plotted the rank read depth for each unique quad of 10X Cell Barcode, UMI, cBC, and rBC. We identified an elbow point with minimum depth of 1 (mixed cell experiment) and 10 (K562 alone experiment), and kept any low-depth unique quad that contains the cBC-rBC pair at most hamming distance of 1 to a high depth pair. Lastly, we remove cells with less than 100 scMPRA-associated UMIs, since the scMPRA reads from those cells were poorly sampled.

Cell cycle analysis

Cell cycle analysis for the scRNA-seq experiment was done with Scanpy 1.8.1 with cell cycle genes [128]. The expression profile of each cell was projected onto a PCA plot based on the list of cell cycle genes using Scanpy.

Motif analysis The core promoters were first clustered according to their expression levels in the different cell sub-state populations by hierarchical clustering. We categorized our data into up/down regulated clusters at the first branching point, aiming to preserve the large structure. We then identified core promoter motifs in each promoter according to the parameters in Zabidi et al[112]. using MAST v4.10.053 and plotted the proportion of promoters containing each motif in each promoter class.

Retinal explant electroporation

Electroporation of plasmid library into newborn mouse retina is performed as described previously [39, 100]. Three pairs of newborn mouse retinas were electroporated per biological replicates. All animal procedures were performed in accordance with a Washington University School of Medicine IACUC-aproved vertebrate animals protocol.

The design of Gnb3 library

We designed 114 different variants of the full length Gnb3 promoter. These mutations can be separated into five categories. First, we mutated each CRX sites as described previous(Drew paper). We mutated CRX sites individually and pair-wise by change the CRX sites into the nonfunctional sequence (5'-CTACTCCC-3'). Second, we changed each CRX sites individually and pair-wise into a Q50 site (5'-TAATTTC-3'). Thirdly, we performed a mutagenesis of mutating every base pair into the three other base pairs for the E-box. We also conducted the pair-wise mutagenesis of the middle two base pairs of the E-box (5'-GCANNTGG-3'). Thirdly, we changed each CRX sites into high affinity (5'-CTAATCCC-3'), medium affinity (5'-CTAAGCCC-3'), and low affinity (5'-CTTATCCC-3') sites. Lastly, we shuffle every five base pairs in the conserved regions in the Gnb3 promoter.

Estimating intrinsic and extrinsic noise

Intrinsic and extrinsic noise were estimated using the statistical framework developed for the dual-reporter experiment. We first extracted the pairwise expression level for cBCs that belong to the same promoter in every single cell. If more than two cBCs are found in the same cell, the

pairwise expressions among them are recorded. We then removed promoters with less than 100 paired single-cell expression measurements (593 out of 676 promoters passed the filtering step). We then applied the statistical framework developed by Fu and Pachter [117]. The derivation is abbreviated and can be found in the original publication. Briefly, let C denote the expression for the first pBC in the cell and let Y denote the expression for the second pBC in the cell. Let ŋint denote the intrinsic noise, and it can be calculated as:

$$\eta_{ext} = \frac{1}{2a\bar{C}\bar{Y}}(\sum_{i=1}^{n}(C_i - Y_i)^2 - n(\bar{C} - \bar{Y})^2), \tag{2.3}$$

where a can be minimized with an MSE, and the extrinsic noise is

$$a = \frac{2n^3 - 7n + 6}{2(n^2 - n)} + \frac{2 - n}{n^2 - n}\frac{\rho}{1 - \rho} + \frac{1}{2(n^2 - n)}(\frac{\rho}{1 - \rho})^2 \tag{2.4}$$

$$\rho = \frac{Cov[C, Y]}{\sqrt{Var[C]}\sqrt{Var[Y]}} \tag{2.5}$$

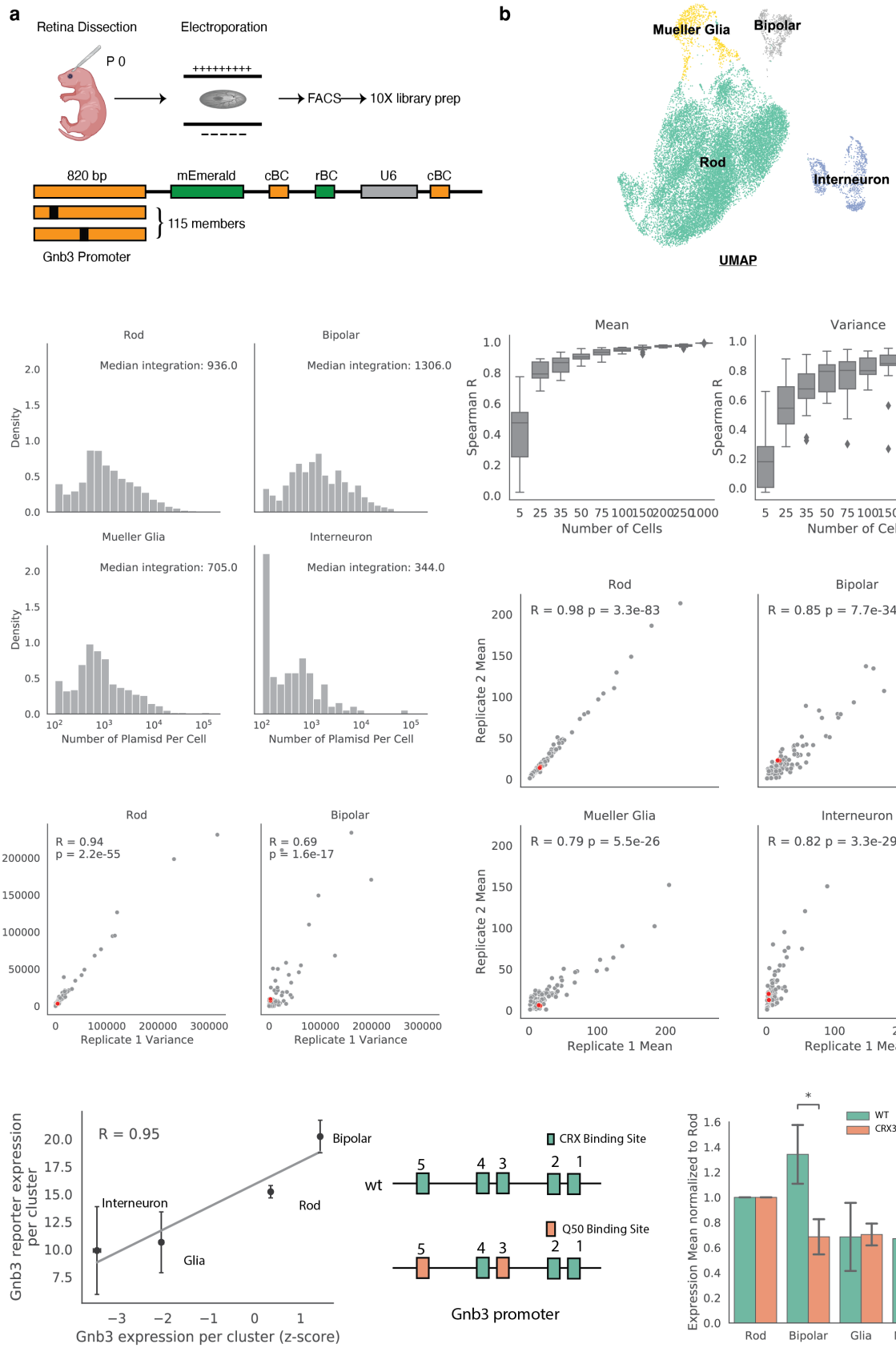
a
Retina Dissection
Electroporation
P 0
FACS
10X library prep
820 bp
mEmerald
cBC
rBC
U6
cBC
115 members
Gnb3 Promoter
b
Mueller Glia
Bipolar
Rod
Interneuron
UMAP
Rod
Median integration: 936.0
Bipolar
Median integration: 1306.0
Mueller Glia
Median integration: 705.0
Interneuron
Median integration: 344.0
Density
Number of Plamisd Per Cell
Mean
Variance
Spearman R
Number of Cells
Rod
R = 0.98 p = 3.3e-83
Bipolar
R = 0.85 p = 7.7e-34
Mueller Glia
R = 0.79 p = 5.5e-26
Interneuron
R = 0.82 p = 3.3e-29
Replicate 2 Mean
Replicate 1 Mean
Rod
R = 0.94
p = 2.2e-55
Bipolar
R = 0.69
p = 1.6e-17
Replicate 2 Variance
Replicate 1 Variance
R = 0.95
Gnb3 reporter expression per cluster
Gnb3 expression per cluster (z-score)
Interneuron
Glia
Rod
Bipolar
CRX Binding Site
Q50 Binding Site
wt
Gnb3 promoter
WT
CRX3Q50 CRX5Q50
Expression Mean normalized to Rod
Rod
Bipolar
Glia
Neuron

Figure 2.4: **scMPRA is reproducible and sensitive in mouse retina** (a) Experimental design for scMPRA in retina. Newborn mouse retinas were dissected and electroporated *ex vivo*. Dissected retinas were cultured for 8 days and harvested for scMPRA experimental protocol. (b) UMAP plot for the transcriptome data associated with the scMPRA experiment. (c) Histograms of the number of plasmids per cell in different clusters. (Upper Left: Rod, Upper Right: Bipolar, Lower Left: Mueller Glia, Lower Right: Interneuron) (d) Scatterplot for the expression mean reproducibility from two biological replicates (Upper Left: Rod, Upper Right: Bipolar, Lower Left: Mueller Glia, Lower Right: Interneuron). Data points in red denote the wild-type Gnb3 promoter activity barcoded differently. (f) Scatterplot for the expression variance reproducibility from two biological replicates (Left panel: Rod, Right panel: Bipolar). (g) The dot plot shows the endogenous Gnb3 expression (10X expression z-score) against the Gnb3 reporter activity (Mean expression). Error bar indicates the 1 std from two biological replicates. (h) Diagram depicting the known cell-type specific variant in Gnb3 promoter (CRX3Q50, CRX5Q50). (i) Expression of the wt and CRX3Q50, CRX5Q50 mutant expression in Rod, Bipolar, Glia, and Neuron. Expression is normalized the Rod expression for each mutant. Error bard indicates 1 std from two biological replicates.

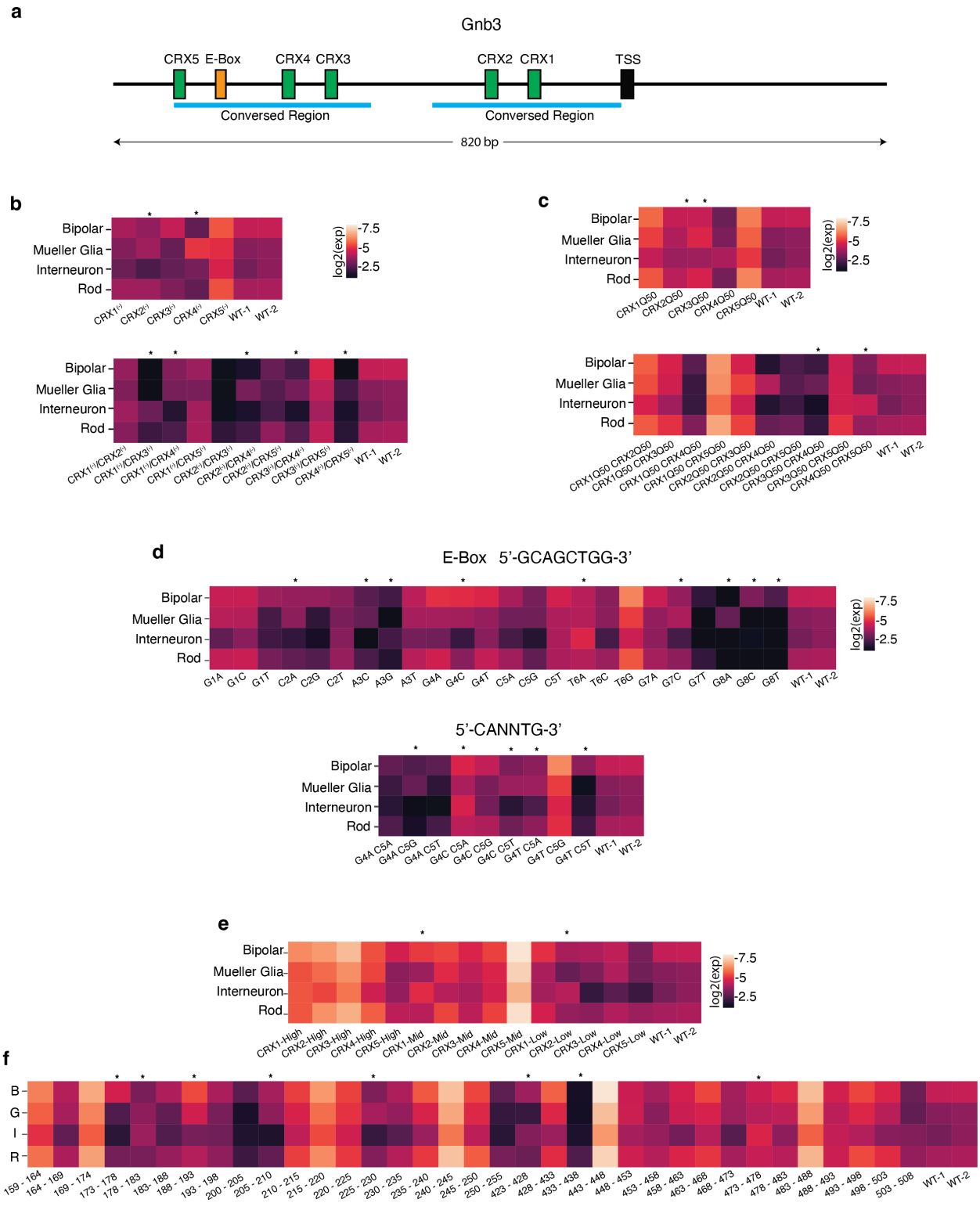
a
Gnb3
CRX5 E-Box
CRX4 CRX3
CRX2 CRX1
TSS
Conversed Region
Conversed Region
820 bp
b
Bipolar
Mueller Glia
Interneuron
Rod
log2(exp)
-7.5
-5
-2.5
WT-1
WT-2
c
CRX1Q50
CRX2Q50
CRX3Q50
CRX4Q50
CRX5Q50
d
E-Box 5'-GCAGCTGG-3'
5'-CANNTG-3'
e
CRX1-High
CRX2-High
CRX3-High
CRX4-High
CRX5-High
CRX1-Mid
CRX2-Mid
CRX3-Mid
CRX4-Mid
CRX5-Mid
CRX1-Low
CRX2-Low
CRX3-Low
CRX4-Low
CRX5-Low
f
B
G
I
R

Figure 2.5: **scMPRA detects many cell-type specific variants in Gnb3** (a) Schematic for the Gnb3 variant library design (b-f) Heatmaps of Log2(express mean) for the Gnb3 variant library in Rod, Bipolar, Mueller Glia, and Interneuron. The colormap for all heatmaps are unified to the minimum and maximum of the whole library. (b) Single and Double mutants of individual CRX sites. (c) Single and Double swap of CRX sites into Q50 sites. (d) Mutagenesis of the E-box motif and double-mutation of the middle base pairs of the E-box motif. (e) Switching the CRX sites into high affinity, medium affinity, and low affinity sites. (f) Shuffling of every 5-bp of the conserved region in Gnb3.

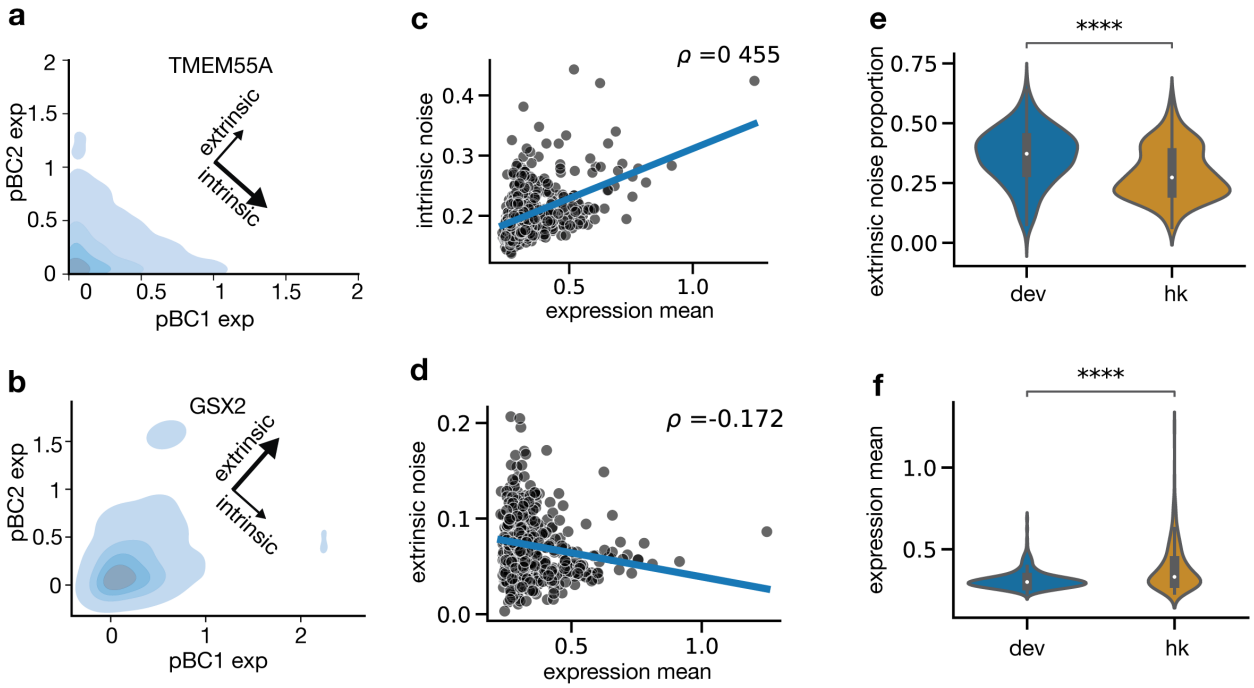

Figure 2.6: **scMPRA deconvolves intrinsic and extrinsic cell-to-cell variability.** a, b) Density plots for single-cell expression of paired cBC expression for the same promoter. TMEM55A has high intrinsic noise, and GSX2 has high extrinsic noise. (c) Scatterplot of expression against intrinsic noise. Blue line shows the linear regression (Pearson ρ = 0.455) (d) Scatterplot of expression against extrinsic noise. Blue line shows the linear regression (Pearson's ρ = -0.172). (e) Violin plot of extrinsic noise proportion for housekeeping and developmental promoters (Mann-Whitney U test. Starts indicate significance: **** : p < 1X10-4) (f) Violin plot of expression mean for housekeeping and developmental promoters (Mann-Whitney U test. Stars indicate significance: **** : p < 1X10-4)

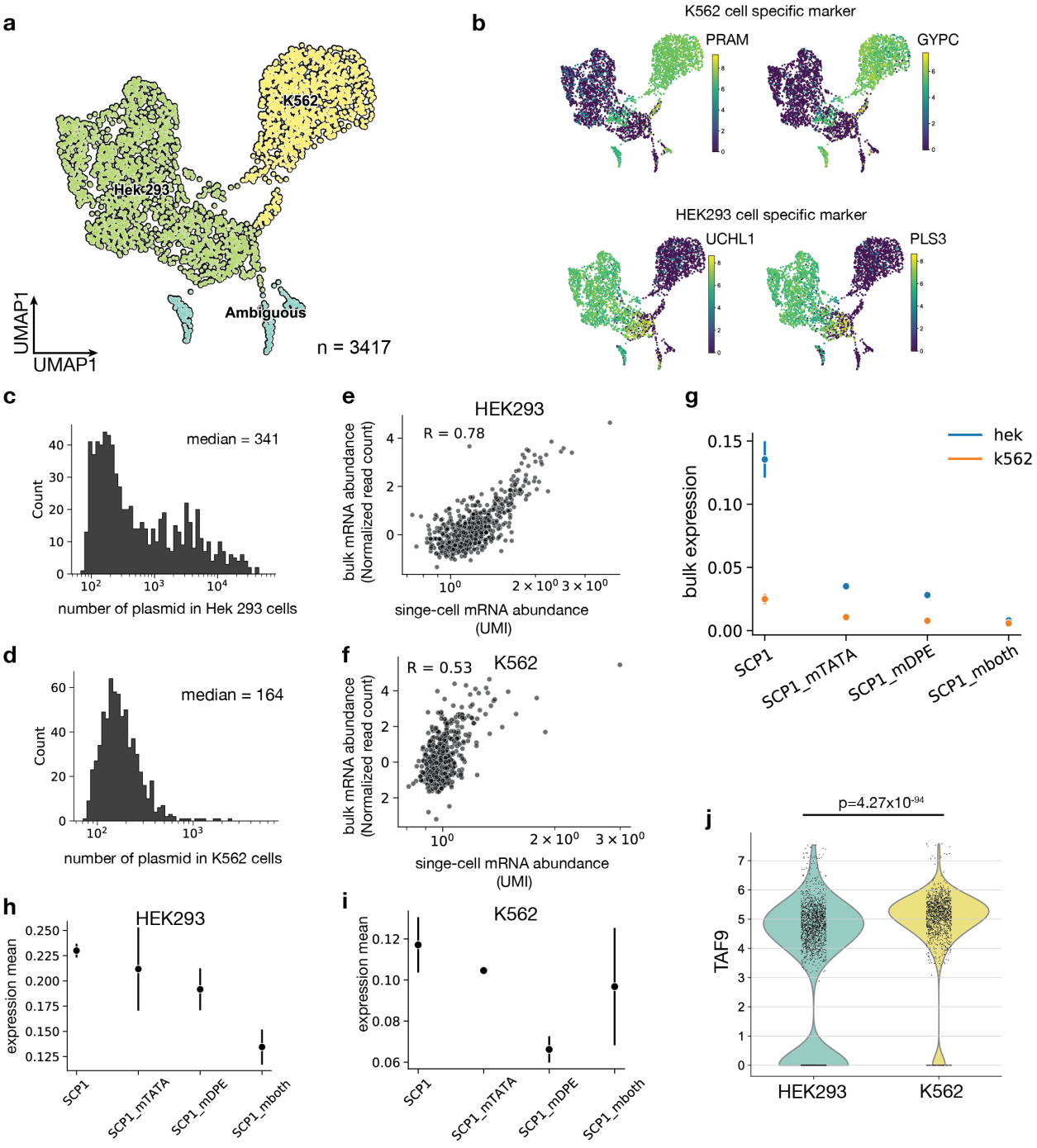
a
K562
Hek 293
Ambiguous
UMAP1
UMAP1
n = 3417
b
K562 cell specific marker
PRAM
GYPC
HEK293 cell specific marker
UCHL1
PLS3
c
median = 341
Count
number of plasmid in Hek 293 cells
d
median = 164
Count
number of plasmid in K562 cells
e
HEK293
R = 0.78
bulk mRNA abundance (Normalized read count)
singe-cell mRNA abundance (UMI)
f
R = 0.53
K562
bulk mRNA abundance (Normalized read count)
singe-cell mRNA abundance (UMI)
g
hek
k562
bulk expression
SCP1
SCP1_mTATA
SCP1_mDPE
SCP1_mboth
h
HEK293
expression mean
SCP1
SCP1_mTATA
SCP1_mDPE
SCP1_mboth
i
K562
expression mean
SCP1
SCP1_mTATA
SCP1_mDPE
SCP1_mboth
j
p=4.27x10-94
TAF9
HEK293
K562

Figure 2.7: **scMPRA measures cell-type specific CRS activity** (a) UMAP of the single-cell transcriptome from the mixed-cell experiment. 105 out of 3417 cells (3%) are labeled by both K562 and HEK293 cell genes. (b) UMAP of the mixed-cell experiment with cells marked by other representative markers for K562 and HEK293 cell expression. (c-d) Histogram of the number of plasmids transfected to K562 cells and HEK293 cells. (e,f) Scatterplot of bulk RNA-seq expression against expression mean from scMPRA (Pearson R for K562 cells: 0.53, Pearson R for HEK293 cells: 0.78). (g) Dot plot of the reporter activity of SCP1 and its mutants from bulk RNA-seq data (error bar: 1 s.d.). (h) Dot plot of the mean reporter activity of SCP1 and its mutants from scMPRA experiment for K562 cells. (i) Dot plot of the mean reporter activity of SCP1 and its mutants from scMPRA experiment for HEK293 cells.(j) Violin plot showing the expression distribution of TAF9 in K562 and HEK293 cells. (Wilcoxon rank sum test, p = 4.27x10-94).

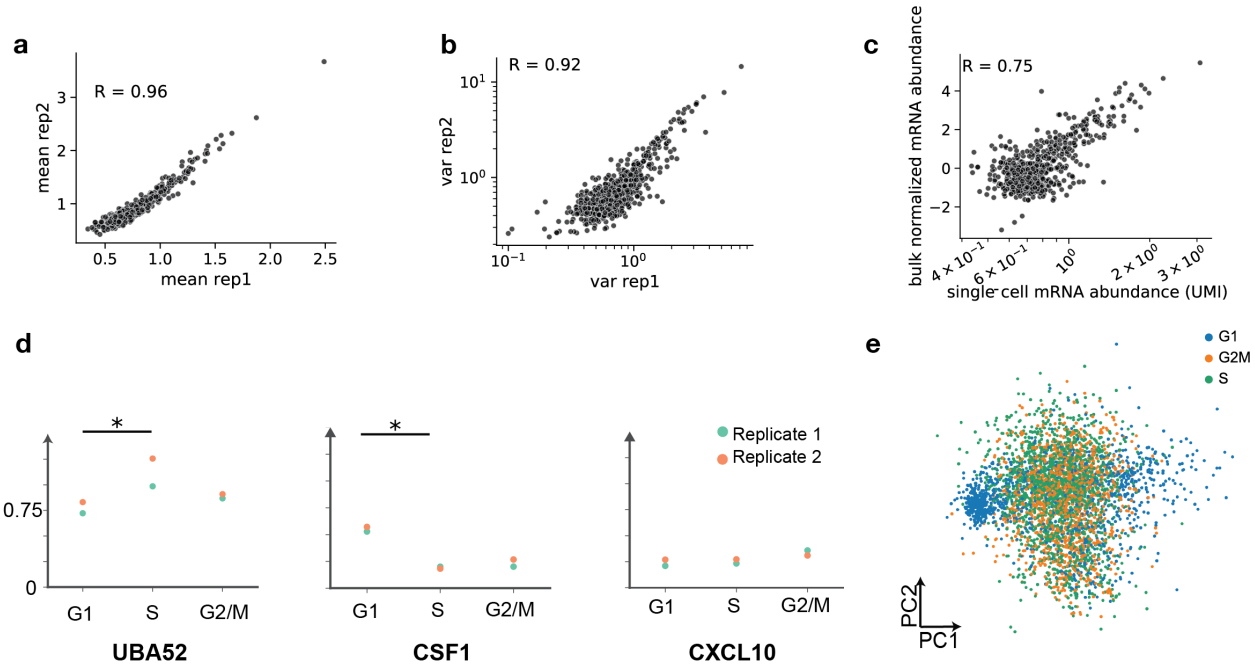

Figure 2.8: **scMPRA measures CRS activity in K562 cell substates.** (a,b) Reproducibility for expression mean and cell-to-cell variance (Pearson Correlation for mean: 0.96, for variance: 0.92). (c) Scatterplot of reproducibility of scMPRA mean expression with bulk MPRA measurement using UMI (Pearson Correlation: 0.75). (d) Different dynamics of expression. For UBA52, the promoter is most highly expressed in S phase, whereas for CSF1, the promoter is most highly expressed in G1 phase. For CXCL10, the promoter is expressed evenly through cell cycle (Stars indicate significance from Wilcoxon rank sum test, *: $p < 0.05$.) (e) Cells no longer cluster together based on cell cycle genes after normalization.

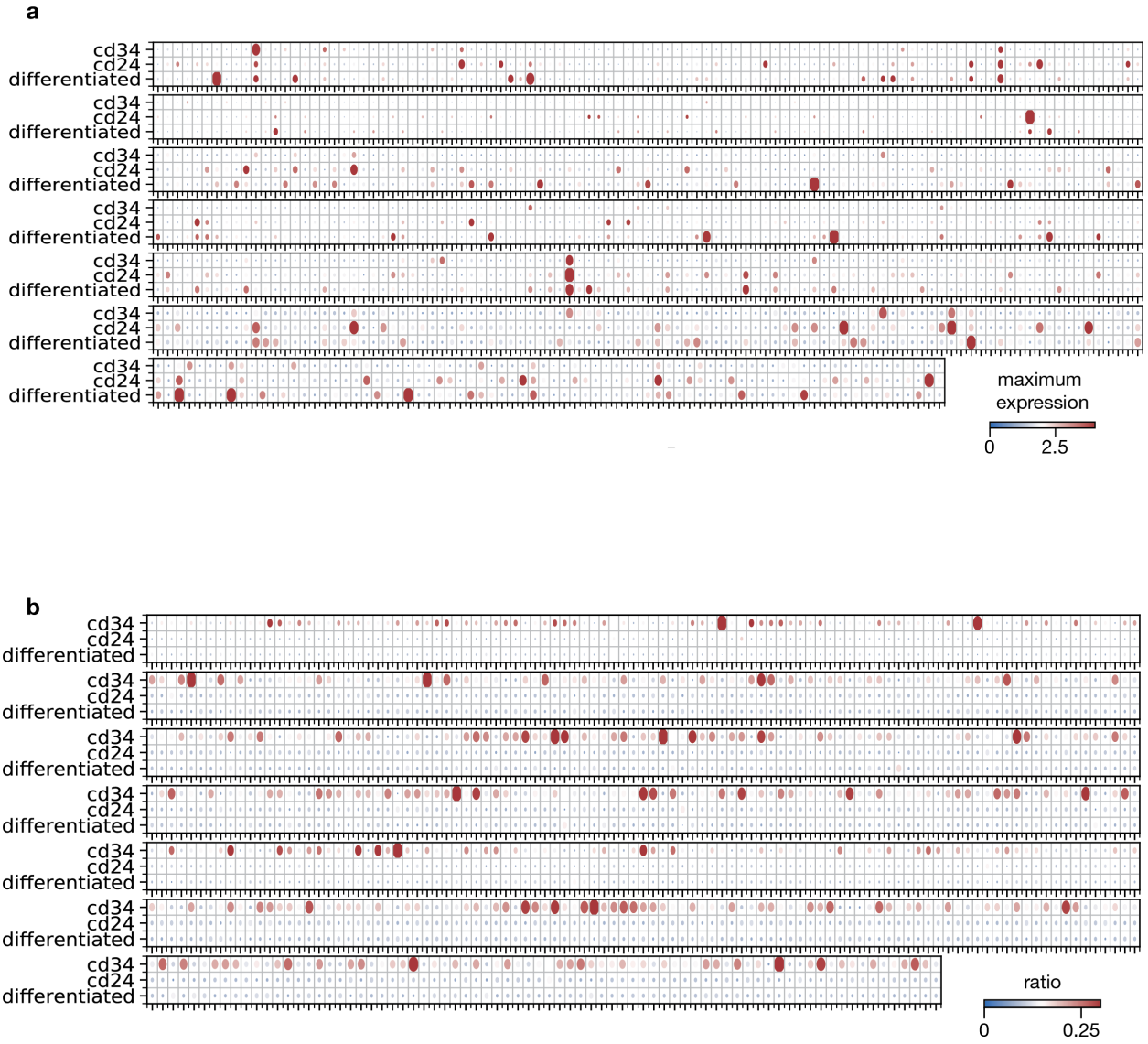

Figure 2.9: **CD34+/CD38- substate changes the probability of cells having higher expression, not the maximum expression level.** (a) Dot plot showing the maximum single-cell expression for the core promoter library in CD34+/CD38-, CD24+, and Differentiated clusters. Color and size both indicate the maximum expression change. (b) Dot plot showing the percentage of cells in CD34+/CD38-, CD24+, and Differentiated clusters that are in the 90th percentile of expression level per promoter. Color and size both indicate the ratio change.

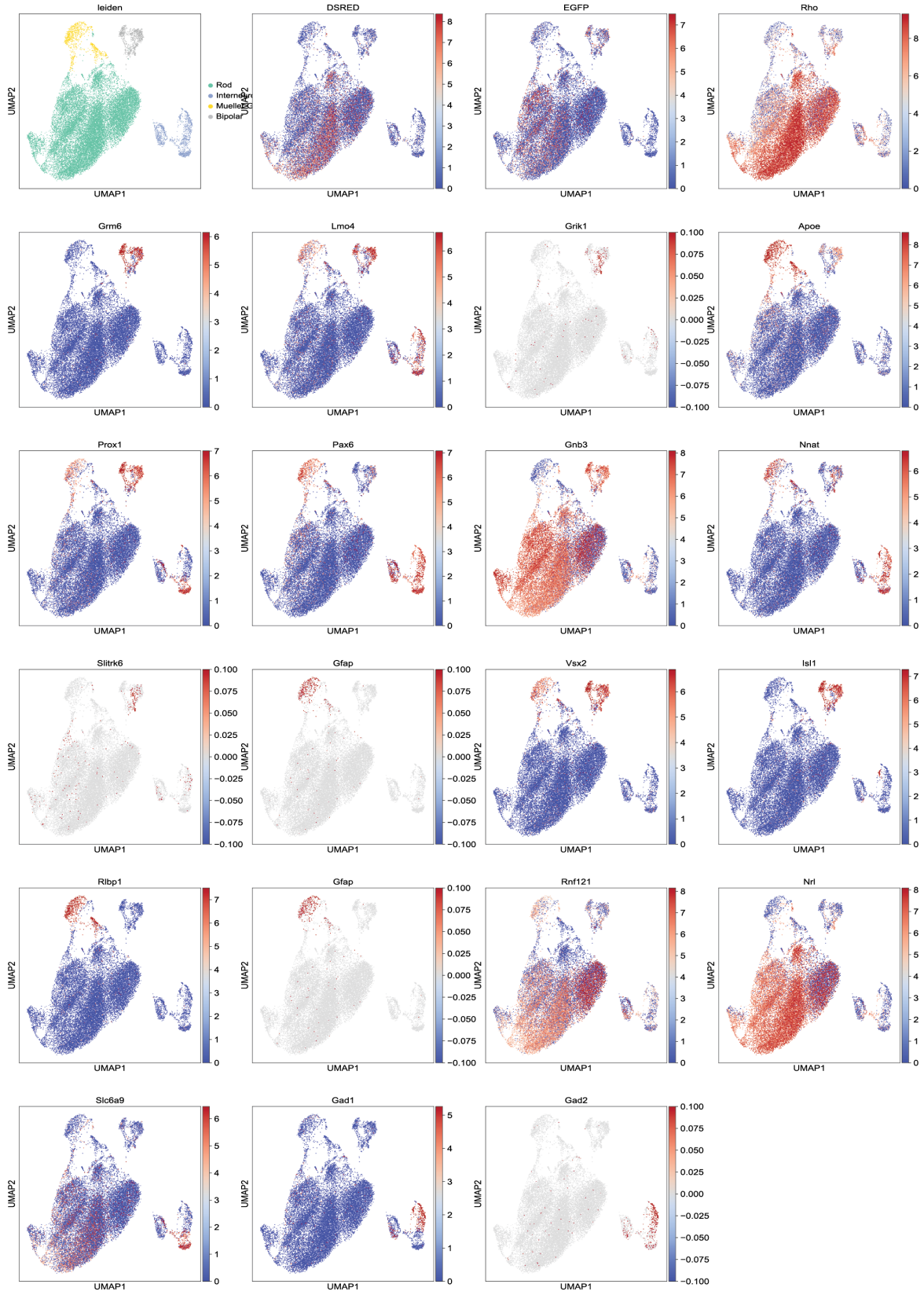

Figure 2.10: **Several cell types are recovered from the scMPRA experiment in retina**

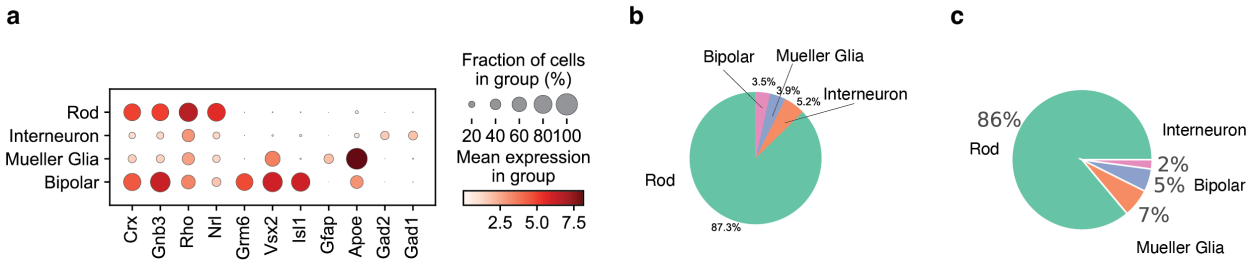

Figure 2.11: **Cell types recovered from scMPRA retina experiment have distinct cellular signatures** (a) Heatmap showing the marker gene expression in the 4 cell types measured with scMPRA in retina. (b) Pie chare showing the percentage of cells recovered with scMPRA in retina after filtering. (c) Pie chart showing the percentage of cells with scMPRA in retina after filtering.

Chapter 3: Quantifying the Impact of Cellular and Genomic Environments on Single-Cell Variability

Siqi Zhao[1], Clarice KY Hong[1], Avinash Ramu[1] and Barak A. Cohen

The Edison Family Center for Systems Biology and Genome Sciences and the Department of Genetics, Washington University School of Medicine, USA.

1 These authors contributed equally.

This work was done in collaboration with Clarice Hong, Avinash Ramu, and Barak Cohen. Clarice Hone, Avinash Ramu, and I conceived this project. Clarice Hone, Avinash Ramu, Barak Cohen, and I designed the experiments. Clarice Hone, Avinash Ramu, and I carried out all experiments and analyses. Clarice Hone, Avinash Ramu, Barak Cohen, and I wrote the paper.

3.1 Abstract

Individual cells from an isogenic population of cells can display high levels of variability in gene expression. In mammalian cells, large-scale maps of histone modifications and TF binding have allowed us to understand the effects of genomic environments on controlling mean expression levels. However, we have not been able to leverage such maps for single-cell variability, largely due to the lack of methods to assay single-cell variability across the genome in a high-throughput manner. Here we present single-cell TRIP (scTRIP), a method to measure single-cell variability of reporter genes integrated throughout the genome in parallel. We show that different chromatin and TF features regulate gene expression mean and variability, and build a model to classify genomic locations that have high vs low expression variability. We further show that scTRIP data can be used to quantify intrinsic and extrinsic contributions to expression variability. Additionally, another advantage of scTRIP is that the transcriptome of each cell can be simultaneously readout with reporter gene activity. Using the transcriptome information, we find that part of the extrinsic noise can be explained by the cellular substates of individual cells, and these substates can be used to improve our prediction model. Taken together, we show that scTRIP is a powerful tool to measure gene expression mean and single-cell variability across the genome, and that the data generated by scTRIP is a useful complementary resource for understanding how genomic environments control gene expression.

3.2 Introduction

Individual cells from an isogenic population in the same cellular environment can display high levels of variability in gene expression [2]. This can lead to variable outcomes in processes such as cell differentiation [67, 129–131], response to environmental stimuli [132, 133], viral latency [134]and chemotherapeutic drug resistance [70, 114, 135]. One of the major goals of biology is to understand the molecular mechanisms controlling gene expression mean and single-cell variability and to develop tools for precise control of gene expression at the individual cell level.

Single-cell variability in mRNA expression is influenced by two major components: intrinsic features such as nearby cis-regulatory sequences and genomic environments and extrinsic features such as variation in cellular states [3, 136–139]. Classic experiments of position effect variegation in yeast have shown that chromatin states can influence gene expression noise [140]. Other features of genomic environments such as enhancers, histone modifications and transcription factors (TFs) have been shown to regulate single-cell variability [141–146]. However, these studies were either performed with scRNA-seq, which does not control for the effects of different promoter elements or rely on low-throughput imaging methods. The only study with large numbers of genomic integrations was not able to assign locations to the measured reporter genes [144]. Thus, we still lack a systematic way of quantifying the impact of individual components of genomic environments on single-cell variability.

Gene expression noise also hinders the production of pure cell populations from reprogramming and influences gene therapy outcomes because of the variability in expression of integrated transgenes. While advances in genome engineering technologies now allow researchers to integrate transgenes at most desired genomic locations, the selection of appropriate sites for transgene overexpression remains non-trivial, with no location in human cells validated as a safe harbor locus [147, 148]. This is mainly due to the lack of methods to systematically screen for loci that have high expression, low variability and do not impact cellular function. Here we developed a high-throughput method to measure the mean and variance of a common reporter gene at different

genomic locations in parallel. We showed that different chromatin features are associated with high and low mean independent noise at different genomic locations. We used a multivariate logistic regression model to identify other features in genomic environments that impact the transcriptional noise. We also quantified the extrinsic noise and showed that the differences in cellular substates explain part of the extrinsic noise. We also showed the performance of our model improved using cellular substate information as additional features. Lastly, we showed that scTRIP can measure the effect of integrations on endogenous gene expression in a high throughput manner. Taken together, we show that scTRIP is a powerful tool to study how genomic environment and cellular context control single-cell variability.

3.3 Results

3.3.1 High-throughput method to measure expression mean and variance across the genome

We developed a high-throughput method to measure the mean and variance of the same reporter gene across the genome based on the Thousands of Reporters Integrated in Parallel (scTRIP) method [50]. First, we generated a library of reporter genes that are identical except for a short random barcode in the 3'UTR of the gene (**Figure. 3.1 A**). The reporter gene consists of a CMV promoter driving the expression of a fluorescent protein and contains a capture sequence from the 10x Genomics feature barcoding kit, which allows us to capture the mRNAs using orthogonal capture sequences on the 10x Genomics gel beads (see rationale below). The library of reporter genes are then integrated into cells using the piggyBac transposase system. We can then map the integration location of each barcode to assign barcodes to specific genomic locations (Methods).

To capture the expression mean and variance of each reporter gene, we used the 10x Chromium Single Cell 3' Kit (v3.1) with Feature Barcoding Technology. The gel beads from the Feature Barcoding Kit contain capture sequences that can be used to capture our reporter gene which contains corresponding capture sequences (**Figure. 3.1 A**). Even though the reporter gene is polyadenylated and can be captured with oligo-dT sequences, we reasoned that the reporter mRNA would be competing with the transcriptome to bind to oligo-dT, while the capture sequence would be free to bind only to our reporter gene. This would potentially increase the efficiency of capturing the reporter mRNA transcripts and reduce the number of 'dropout' transcripts that is pervasive in scRNA sequencing. We then amplified only our reporter genes from the 10x pool for sequencing and mRNA barcode counting (Methods).

3.3.2 scTRIP measurements are highly accurate and reproducible

We performed scTRIP in K562 cells because of the abundance of public epigenetic data available. We reasoned that 100 single cells for each IR location would be required to obtain a good estimate of mean and variance. Each scTRIP experiment contains many single cell clone expansions, all the cells from the same clone share the same genomic integrations. Since we targeted approximately 20,000 cells per 10x run, the upper limit of the numbers of clones we can test in one experiment is 200. Because 10x also has a high dropout rate we targeted 100 clones per experiment in order to ensure that we obtained high quality data. Each clone has an average of 5 integrations, this theoretically allows us to assay 500 Integrated Reporter (IR) locations in one experiment.

To compute the technical reproducibility of scTRIP we used two pools of cells that were grown out of the same initial clones and hence had identical insertions of the reporter genome. We ran each pool separately through the scTRIP workflow and computed the mean and variance at each IR location from each replicate. We observed 600 identical IR locations detected in both the replicates, indicating that scTRIP is able to capture most of the locations with IRs. Across both replicates, after QC, we had data from 7680 cells and a total of 2,940,912 UMIs. The replicates were well correlated in terms of the mean and variance measured at each IR location (**Figure 3.2 A,B**, mean correlation (Pearson's r) = 0.76, variance correlation (Pearson's r) = 0.72) indicating that the measurements obtained with scTRIP are highly reproducible.

In order to determine how well scTRIP estimates the mean levels of expression we measured the mean expression at each IR location using bulk RNA sequencing of the barcodes. We then measured the mean level in the same pool of cells with the same integrations as scTRIP. We observed a good correlation between the mean levels measured using the bulk assay and using scTRIP (**Figure 3.7 A**, correlation (Pearson's r) = 0.48). Since the bulk mean levels are measured by integrating over all the RNA molecules from all the cells and is a different measurement from scTRIP where we count the number of UMIs at each IR location and average over the observed number of cells, we did not expect the two measurements to be perfectly correlated. To further validate our measurement of

single-cell variability, we performed single-molecule Fluorescence in situ Hybridization (smFISH) on two known locations (Methods). We found our results agree with the smFISH measurements (**Figure 3.7 B,C**). These analyses demonstrated that our method is accurate and reproducible in measuring the expression mean and noise.

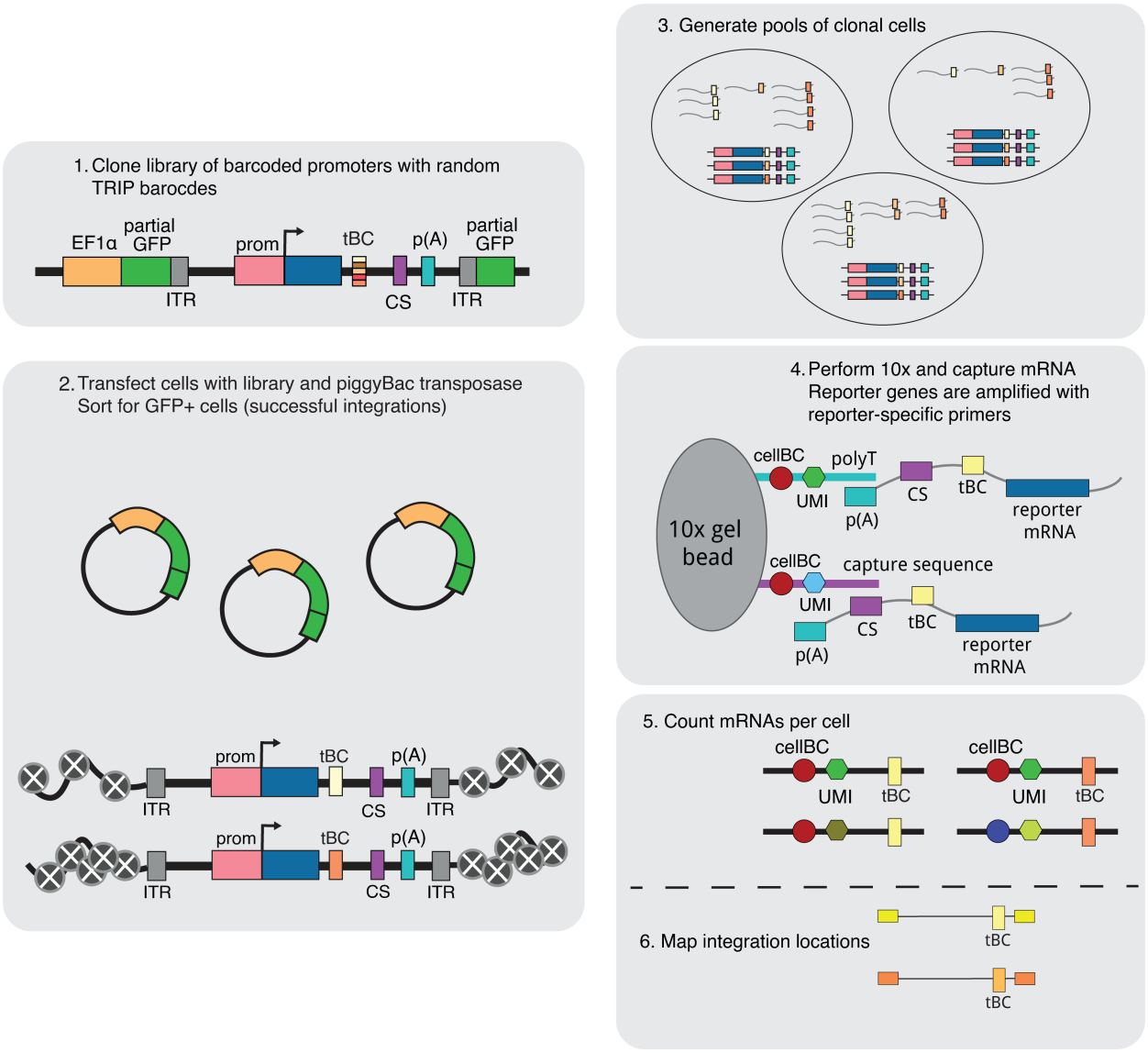

Figure 3.1: Overview of the scTRIP workflow.

3.3.3 Measurements of mean-independent noise across different chromosomal environments

In total, we performed four experiments and generated mean and variance measurements for 934 integrations. The integrations were spread across the genome and found in different chromHMM annotations [149] (**Figure 3.8 A,B**), allowing us to study the effects of diverse chromosomal environments on expression variance.

Previous studies of noise have shown that the expression mean is highly correlated with variance [71, 72]. Here we also find that expression mean is also highly correlated with expression variance, indicating that a large proportion of a gene's variance is explained by its mean (**Figure 3.2 C**). To identify chromosome features that control expression variance independently of mean levels we attempted to use common noise metrics such as the fano factor and CV2. These metrics normalize expression variance to different powers of expression mean (1 for fano factor and 2 for CV2). However, even when we normalize expression variance by CV2, we still observed some correlation with expression mean (**Figure 3.8 C,D**), likely because the expression from non-endogenous locations in the genome is highly variable (Akhtar et al. 2013). This led us to a metric we refer to as mean-independent noise (MIN) that regresses out the mean effect (Methods). As expected, MIN levels were completely uninfluenced by expression mean levels (**Figure 3.2 D**). Thus, we used MIN as a measure of expression noise for all the analyses below.

3.3.4 Expression mean and noise are associated with different chromosomal features

We first sought to find chromosomal features that might explain differences in MIN levels between different genomic locations. Large-scale studies of chromatin features by ENCODE and others

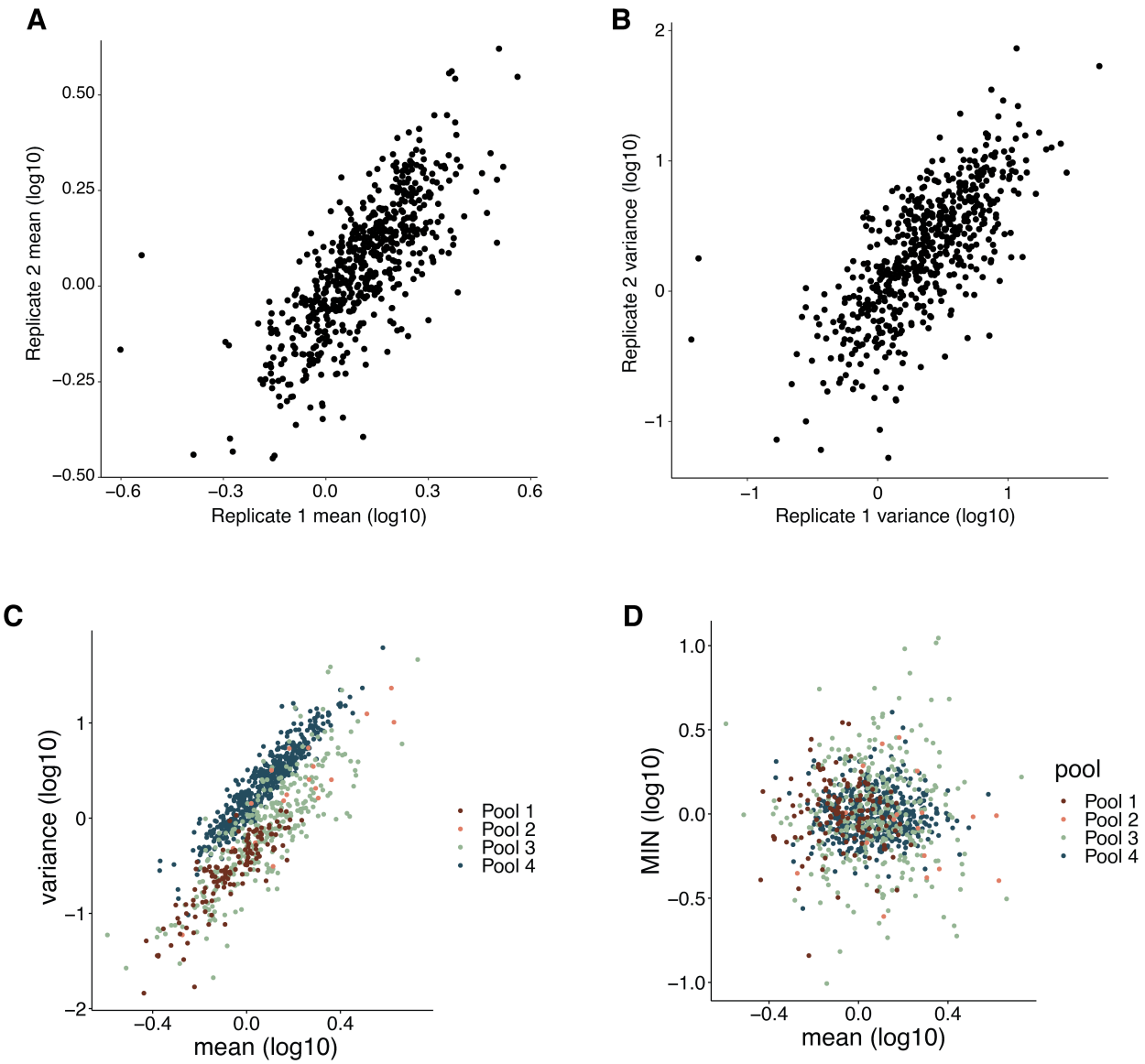

Figure 3.2: scTRIP measurements are accurate and reproducible. (A) Correlation of mean levels between replicates. (B) Correlation of variance levels between replicates. (C) Mean and variance are correlated within each experiment. (D) Mean-independent noise corrects for mean effects on variance.

have shown that the mean expression of a gene is correlated with its surrounding chromatin features [50, 150]. As expected, when we look at the mean expression levels of IRs present in different chromHMM annotations [149], we find that active chromatin regions (enhancers, promoters and transcribed regions) tend to have higher mean levels of expression compared to silent chromatin environments (**Figure 3.3 A**). However, we observed no correlation between chromHMM annotations and MIN levels (**Figure 3.3 B**), suggesting that such broad chromatin categories are not predictive of gene expression noise.

To look more closely into the individual features that might control mean and noise, we first split the IRs into bins of high/low mean levels and high/low MIN levels. We then looked at chromatin features that might correlate with mean and/or MIN levels. As expected, IRs with higher means had higher levels of active chromatin modifications such as H3K27ac, H3K4 methylation, H3K79me2 and H3K9ac (**Figure 3.3 C**). Conversely, IRs with high vs low MIN did not exhibit significant differences between H3K27ac or H3K4me1 levels, and low MIN locations showed slightly elevated levels of H3K4me2/3, H3K79me2 and H3K9ac (**Figure 3.3 D**). These results suggest that different chromatin modifications are associated with regulating expression mean and noise, and that more active genomic locations might be reducing the MIN of our reporter genes. This is also consistent with previous analyses of scRNA-seq data showing that the lack of active histone modifications is associated with high MIN [141].

Another factor that can impact expression noise levels is the binding of TFs around the IR. To identify TFs that might affect noise we looked for TFs that are enriched around either high or low MIN IRs. We find that sequences at low MIN IRs are enriched for transcriptional activators such as MEF2, TBP and SP1, while sequences at high MIN IRs are enriched for CREM and AP-1 family motifs (**Figure 3.3 E**), suggesting that different TFs can modulate noise independently of mean expression levels.

Next we wanted to systematically identify the features that are associated with low MIN and high MIN IR locations. We fit a logistic regression model using the chromatin modifications and

sequence features to classify high and low MIN locations. We were able to train a model with 76% accuracy on the training data. When applied to holdout data from a different pool, the model was able to achieve an accuracy of 67%. The most important sequence features that helped with the predictions are the H3K4me3 mark, TF motis (RARG, FOXO4, HIF1A, TFAP4, CREM, ATF1, NFIC, CLOCK and NFIA) and whether the IR location was inside a gene or outside (**Figure 3.3 F**). Being inside a gene reduced the probability of being a high noise lR location, this could be due to the local regulatory elements that control a gene's expression. Similar to our results above, lower H3K4me3 increased the probability of being a high noise IR location. H3K4me3 is associated with active chromatin and agrees with the idea that higher activity reduces IR MIN. This is consistent with a previous study showing that H3K4me3 reduces noise at endogenous genes [141]. However, we also find that the direction of effect for some of the other histone modifications from our study is different from the effects on endogenous genes, highlighting the importance of controlling for gene sequences. Presence of some TF motifs increase the probability of being a high IR location (NFIC, CREM, TFAP4, CLOCK), whereas other TFs reduce the probability of being a high noise location (RARG, NFIA, ATF1, FOXO4, HIF1A). We used a similar logistic regression framework to identify features that can separate IR locations with high mean levels from IR locations with low mean levels and achieved an accuracy of 78%. The sequence features that increase the probability of being a high mean IR location are higher levels of H2AZ, lower levels of H3K27me3, lower levels of H3K4me2 and a higher number of ATAC-seq peaks in the IR location. The other significant features that increased the probability of being a high mean IR location are higher numbers of motifs of the ALX1, ETV1 transcription factors and fewer motifs of the ELF2 transcription factor.

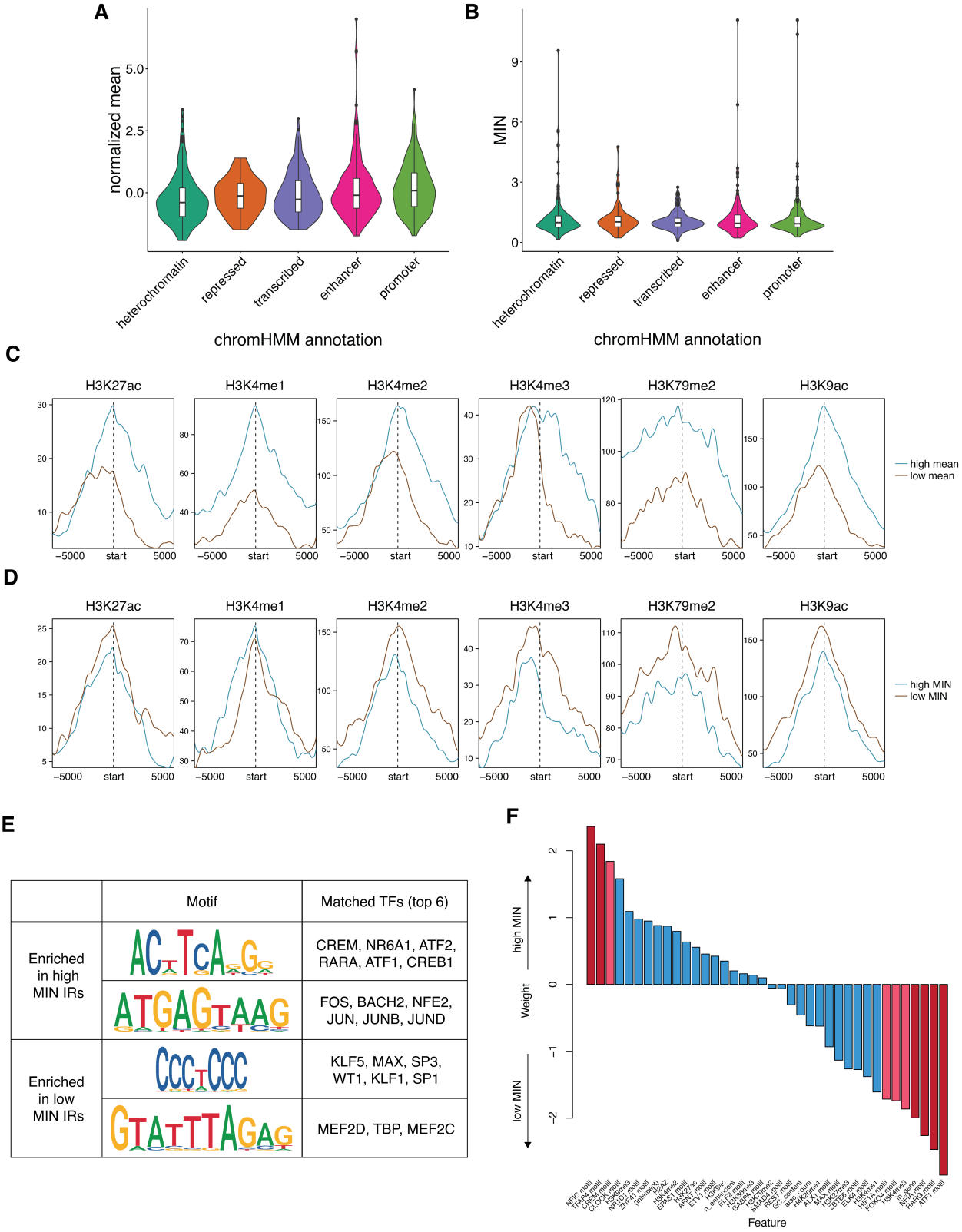
A
normalized mean
5.0
2.5
0.0
heterochromatin
repressed
transcribed
enhancer
promoter
chromHMM annotation
B
MIN
9
6
3
0
heterochromatin
repressed
transcribed
enhancer
promoter
chromHMM annotation
C
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K79me2
H3K9ac
−5000
start
5000
high mean
low mean
D
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K79me2
H3K9ac
−5000
start
5000
high MIN
low MIN
E
Motif
Matched TFs (top 6)
Enriched in high MIN IRs
CREM, NR6A1, ATF2, RARA, ATF1, CREB1
FOS, BACH2, NFE2, JUN, JUNB, JUND
Enriched in low MIN IRs
KLF5, MAX, SP3, WT1, KLF1, SP1
MEF2D, TBP, MEF2C
F
high MIN
Weight
low MIN
2
1
0
−1
−2
Feature

Figure 3.3: **Expression mean and noise are associated with different chromosomal features.** (A) Mean levels of expression are correlated with chromHMM annotations. (B) chromHMM annotations do not impact expression MIN. (C) Active histone modifications associated with high mean IRs. Start indicates the location of the IR. (D) MIN levels are impacted by different histone modifications. (E) Motifs enriched in high or low MIN IRs respectively, and potential TFs that match these discovered motifs. (F) Logistic regression weights of various intrinsic features associated with high or low MIN IRs. Red bars: p-value < 0.05; Pink bars: 0.05 < p-value < 0.1 from the logistic regression model.

3.3.5 Estimating the extrinsic portion of the noise

Our work and many other studies have demonstrated that the global cellular fluctuation, i.e, extrinsic factors, are a source of single-cell variability [3, 33, 64, 115, 116, 151, 152]. We sought to ask if the global, extrinsic factors could explain some portion of the MIN. Historically, two-reporter experiments can be used to quantify extrinsic noise. The covariance of two different reporter activities measures the extrinsic noise [3]. Our system allows us to measure the identical reporter that is integrated in different genomic locations, effectively serving as a type of two-report experiment. We wanted to examine how different IR locations co-fluctuate with each other. This should give us a good estimate of the extrinsic noise.

Each scTRIP experiment contains many single cell clone expansions that share the same integrations (**Figure 3.4 A**). We first used the sharing of the IR location barcodes to identify those single-cell clonal expansions (Methods). We identified 192 original clones, and the mean number of integration per clone is 3 (**Figure 3.9 A**). 45 out of 192 those clones contain more than 1 integration (**Figure 3.4 B**). To validate whether those clones are correctly identified, we individually mapped 16 of those clones, and found most of those barcodes are correctly assigned (**Figure 3.4 B**). We first asked how different the expression is from different IR locations from the same clone. To quantify this difference, we calculated the standard deviation of IR expression from each cell in each clone (Methods). If all noise is extrinsic, we would expect no variability among copies of the same reporter in different IR locations. if all noise is intrinsic, we would expect a high standard

deviation. To estimate the standard deviations of IR locations with completely intrinsic noise, we shuffled the cell labels and computed the standard deviation for each clone. This serves as an upper limit for the standard deviation per clone. We found that for all clones, the standard deviations lie between the total extrinsic and the total intrinsic noise regimes (**Figure 3.4 C**). This suggests that some portions of the expression noise could be explained by some extrinsic factors.

To quantify the amount of extrinsic and intrinsic noise, we computed the pair-wise IR expressions for all clones that contain more than one IR. We employed a previously established statistical framework [117] to quantify the intrinsic and the extrinsic component of noise (Methods). We found that intrinsic noise composes 54% of the total noise (**Figure 3.4 D, E**). This result agrees with previous studies of single-cell variability that found a considerable portion of expression noise is extrinsic. These results show that scTRIP measurements allow us to quantify both intrinsic and extrinsic contributions to single-cell variability.

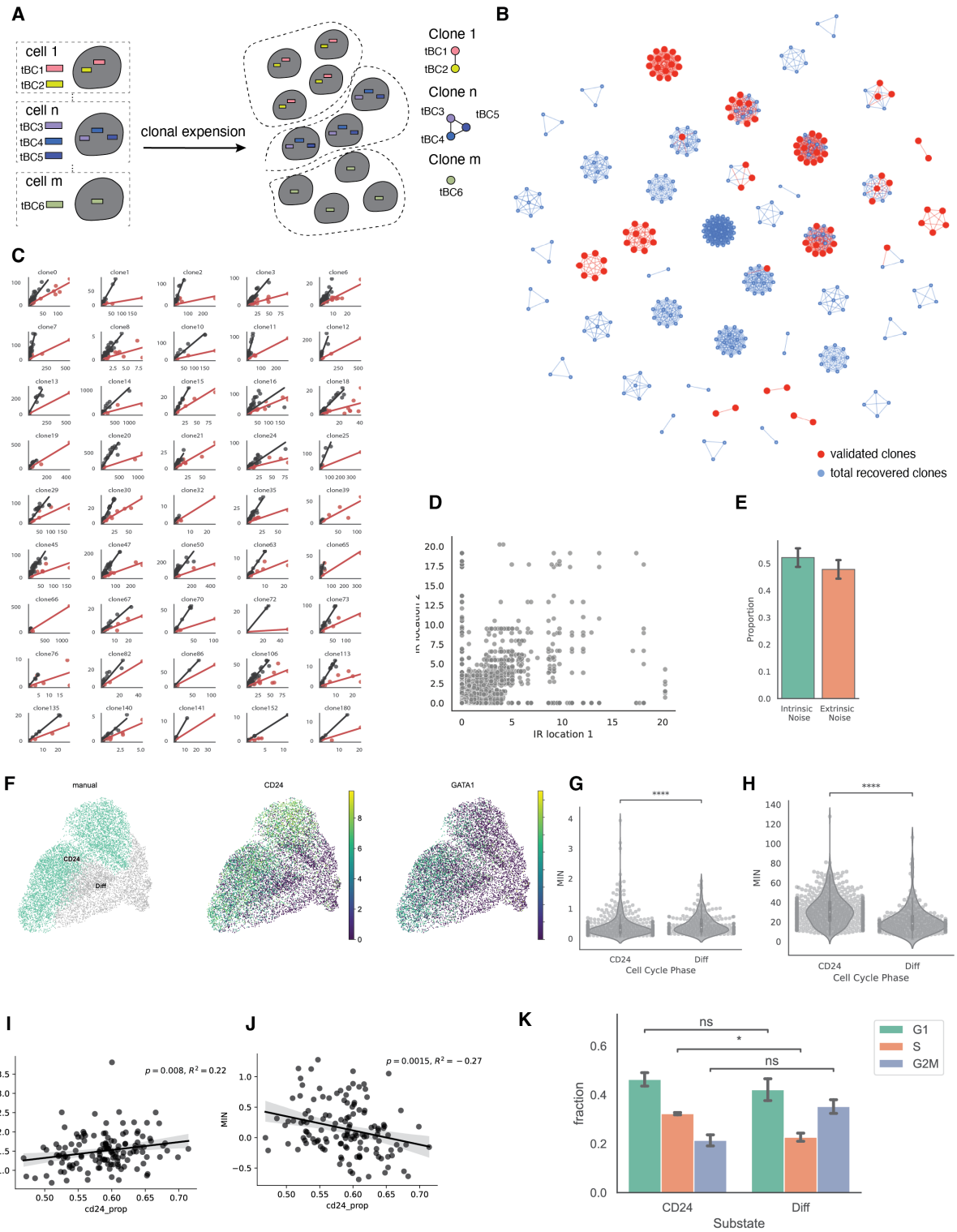
A
cell 1
tBC1
tBC2
cell n
tBC3
tBC4
tBC5
cell m
tBC6
clonal expension
Clone 1
tBC1
tBC2
Clone n
tBC3
tBC5
tBC4
Clone m
tBC6
B
validated clones
total recovered clones
C
D
IR location 1
E
Proportion
Intrinsic Noise
Extrinsic Noise
F
manual
CD24
Diff
CD24
GATA1
G
MIN
CD24
Diff
Cell Cycle Phase
****
H
MIN
CD24
Diff
Cell Cycle Phase
****
I
p = 0.008, R² = 0.22
mean
cd24_prop
J
p = 0.0015, R² = −0.27
MIN
cd24_prop
K
ns
*
ns
G1
S
G2M
fraction
CD24
Diff
Substate

Figure 3.4: **scTRIP quantifies extrinsic portion of expression noise.**(A) Schematic for identifying different original clones (B) A network representation of the different clones identified, red nodes indicate those IR locations are individually validated and appear only in the identified clones. (C) Scatterplot for the mean and standard deviation of expressions from individual cells from the same clone with more than 1 IR integrations. Red: the actually observed data. Black: IR expression with the cell barcode label is randomly shuffled. Fitted lines: linear fit of the mean and standard deviation of expression from individual cells. (D) Pair-wise expression for any two IR locations observed in the same cell. The trend along the diagonal suggests the existence of extrinsic noise. (E) Quantification of intrinsic and extrinsic proportion of noise. Error bar: two technical replicates. (F) UMAP plot of the single-cell transcriptome data. "Stem-like" subcluster has high CD24 and high GATA1 expression. (G,H) violin plots of expression mean and MIN at two substates (student t-test, ****: $p < 0.0001$). (I,J) scatterplots of proportion of cells in the "stem-like" substate against mean and MIN. Line: linear fit with 95% CI. Spearman correlation between mean and proportion of cells in the "stem-like" substate: 0.22, p-value = 0.08. (K) Barplot of the fraction of cells in different cell cycle phases for cells in the "stem-like" substate and the "differentiated" substate (student t-test, *: $p<0.05$).

3.3.6 Cell substates have global effects on expression noise independent of the genomic environment

We sought to ask what cellular mechanisms control extrinsic sources of the mean independent noise. One of the advantages of scTRIP is the simultaneous readout of the reporter expression and the single-cell transcriptome. The transcriptome information provides a good measurement of the cell substate information for a population of cells. To quantify the global, extrinsic factors on the reporter expression at different genomic locations, we sequenced the single-cell transcriptome associated with 356 out of the 934 genomic locations (Methods). The single-cell transcriptome provides a global profile of the cellular information associated with each genomic location at the single cell level.

We hypothesized that the cell substates within an isogenic population could be an important source of extrinsic noise. To test whether this hypothesis is true, we focused on the oscillation between the "stem-like" cell sub-state and the "more differentiated" state in K562 cells [106, 153]. The "stem-like" substate is marked by high CD24 expression, and proliferates at a higher rate.

We processed the transcriptome data that are associated with the IR locations, and identified the clusters with high CD24 expression, and confirmed that this cluster has the other high-proliferation signatures such as GATA1, GATA2 (**Figure 3.4 F**). We then calculated the expression mean and MIN for each IR location in those substates. We found that despite the difference in genomic environment, IR locations in the "stem-like" substate have higher mean and lower MIN (**Figure 3.4 G,H**). This suggests that the cell-substate differences could be a global source affecting MIN.

Given this cell-substate difference of expression mean and MIN, we hypothesized that the proportion of cells in the "stem-like" substates for each IR location could partly explain the MIN. We quantified the percentage of cells for each IR location is in "stem-like" substate, and found that higher proportion of cells in the "stem-like" substate leads to a slightly higher mean expression (Spearman R=0.22, p-value = 0.008), and a lower MIN (Spearman R= -0.27, p-value = 0.0015) (**Figure 3.4 I,J**). To further support this high-proliferation hypothesis, we examined the proportion of cells in different cell cycle phase in the "stem-like" substate and the "differentiated" substate. We found that there is more cells in the S phase in the "stem-like" substate compare to the more "differentiated" state (**Figure 3.4 K**). We examined the differences of mean and MIN in different cell cycle phases, and found that expression mean is higher and MIN is lower in the S phase compare to other phases (**Figure 3.9 B,C**). These results suggest that the difference in proliferation could be a major source of extrinsic noise, and scTRIP could be a powerful tool to study how extrinsic sources affect expression noise.

3.3.7 Cellular information improves classification of low vs high MIN IR locations

Since extrinsic cellular information appears to play an important role in regulating expression noise, we trained a different logistic regression model using cellular information features. Using just these extrinsic features (cell cycle proportions (S, G2) and CD24 proportion) we were able to train a model with 75% accuracy, which is comparable to the model with only intrinsic features. This result implies that the cellular features explain a lot of the variance in MIN between high and low IR locations. The proportion of cells in G2 and the proportion of cells in the CD24 state were significant predictors in this model (Supplementary Data1). Being in G2 increases the probability of a high IR location whereas having a higher proportion of CD24 cells reduced the probability of being in a high IR location (**Figure 3.5 A**). When we combined the significant intrinsic features from the previous model with the extrinsic features, the model accuracy moved up to 84% (**Figure 3.5 B**). In the combined model, the extrinsic features have higher weights than the intrinsic genomic environment features (**Figure 3.5 C**), suggesting that the cellular information may play a larger role in regulating MIN compared to genomic environments. We observed a similar role for extrinsic features in classifying IR locations with high mean levels from IR locations with low mean levels. The model accuracy for the combined model with both intrinsic and extrinsic features is 88%. In the combined model, the proportion of cells in the CD24 cell-state is the most highly weighted feature along with the number of motifs of the ELF2 transcription factor. The proportion of cells in the CD24 state increases the probability of being a high mean IR location but reduces the probability of being a higher MIN IR location consistent with Figures 4 I, J.

3.3.8 The integration of a trans gene has complex effect on endogenous gene expression

Finally, another application of scTRIP is its ability to screen the effect of integrating a trans-gene at a high throughput level. To achieve safe and effective gene transfer therapy, we must have practical knowledge of the genomic locations that enable stable expression of the trans-gene of interest and reduce the disruption of the endogenous gene expression. Historically, trans-genes are often integrated into several "safe harbor" loci[154]. Those loci are mainly located in the introns of stably expressed genes to prevent silencing. Recent efforts have attempted to screen for loci that are far away from known genes, yet the long term stable expression needs more valuation.

We wanted to examine how those trans gene integrations altered the endogenous gene expression. We focused on the IR locations (100 out of 952) that are integrated into gene bodies. We used decoupler(Badia-i-Mompel et al. 2022) to calculate pseudo-bulk expression for clones that contain

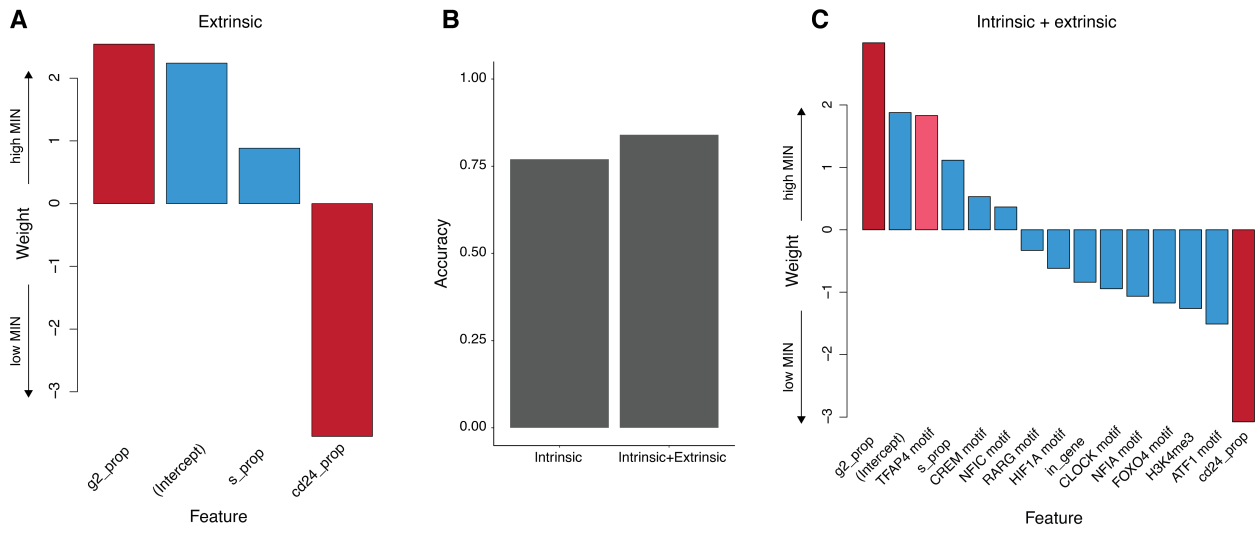

Figure 3.5: **Cellular information improves classification of low vs high MIN IR locations.**(A) Weights of logistic regression model using extrinsic (cellular) features alone. (B) Addition of extrinsic features helps to improve the accuracy of the model. (C) Weights of logistic regression model using both intrinsic and extrinsic features. The most significant features are still the proportion of cells in the G2 phase and CD24+ phase. Red bars: p-value < 0.05; Pink bars: 0.05 < p-value < 0.1 from the logistic regression model.

the integration and the pseudo bulk of the remaining cells that do not have the IR integration serve as the control expression (**Figure 3.6 A**). We found that most integrations do not alter the endogenous gene expression (**Figure 3.6 B**, **Figure 3.10 A**). Among the locations with significantly differentially expression, 3 out of 4 IR integrations increases gene expression (**Figure 3.6 C**). We also randomly shuffled the gene labels to compute the background differential expression, and found that there were no significantly differentially expressed genes once the labels were shuffled (Supplementary Figure 5A). The discovery of upregulation of some endogenous genes agrees with the studies of safe harbor locus that the integration of a trans gene often increases the endogenous gene expression [148]. Taken together, scTRIP can detect the effect of the integration of an exogenous gene. This result illustrates that scTRIP could be a powerful tool to screen for genomic locations with specific expression profiles and minimize the effect on endogenous gene expression.

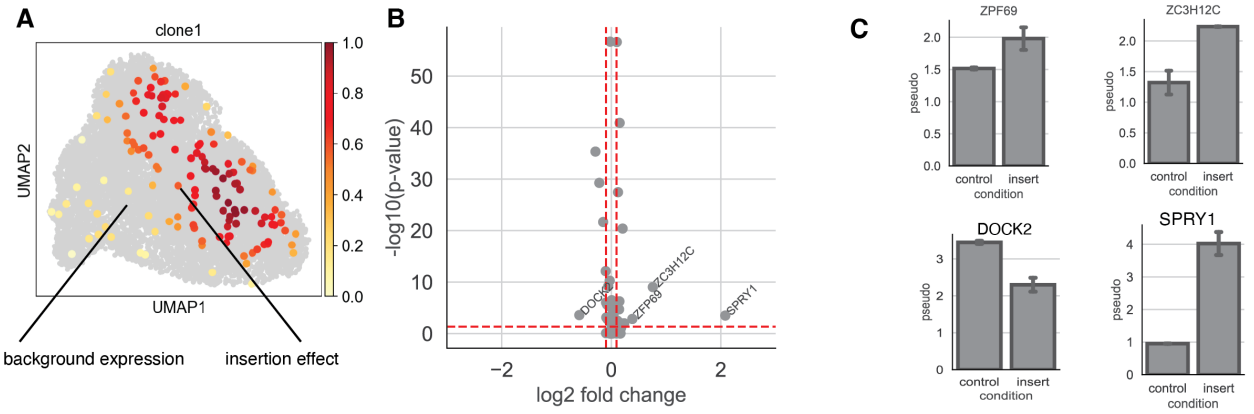

Figure 3.6: **scTRIP measures the insertion effect of a trans gene.**(A) Schematic for expression change detection in the transcriptome data. (B) Volcano plot of log2 fold change and -log10(p-value) from a Fisher's Exact Test. Red dotted line: cut off for fold change (0.5), cut off for p-value: 0.05. (C) Barplots of difference of expression between genes without IRs (control) and genes with IRs (insert). Error bar is derived from two technical replicates.

3.4 Discussion

Since the first single-cell experiment revealed the variability of gene expression in an isogenic population [3], many individual chromatin and sequence features are identified to modulate single-cell variability [2, 74, 129, 155]. However, there has yet to be a systematic study of the impact of different genomic features on large numbers of identical genes.

In this paper we developed scTRIP, a high throughput method to measure the expression mean and noise at different genomic locations in parallel. One key advantage of scTRIP is that the reporter gene used in all locations is exactly identical, allowing us to isolate the effects of the genomic environments without being confounded by the identity of the promoters. We identified different chromatin marks that are associated with high or low MIN, and used a logistic regression model to identify features of the genomic environments that might control MIN. Our observations indicate that the features that control expression noise are independent of the features controlling expression mean. Many studies have been developing tools for the orthogonal control of mean and single-cell variability. To this end, our results expand the potential set of tools that we can use for independent modulation of expression mean and single-cell variability.

We also quantified the extrinsic portion of expression noise and identified that the oscillation between a “stem-like” substate and a “differentiated” substate is an important source of extrinsic noise. Our data suggests that extrinsic noise might be more important in regulating MIN than genomic environments.

A logical next step is to apply scTRIP to more diverse chromatin environments in the genome. A larger goal will be to construct a detailed map of the MIN landscape across the genome, much like the maps of mean expression levels generated by ENCODE. While we used the 10x Genomics platform in this study, it would be possible to also use other single-cell RNA methods such as sciRNA-seq[78, 156] to assay many more locations at once.

3.5 Methods

scTRIP library cloning To clone the reporter gene for scTRIP, we first cloned a CMV-BFP reporter gene containing the 10x capture sequence 1 (CS1) into a TRIP vector containing two parts of a split-GFP reporter gene [157]. When the reporter gene construct is integrated into the genome, the split-GFP combines to produce functional GFP, allowing us to sort for cells that have successful reporter gene integrations. We next added a library of random barcodes to the plasmid by digesting the plasmid with XbaI followed by HiFi assembly (New England Biolabs) with a single-stranded oligo containing 16 random N's (TRIP barcodes; tBC) and homology arms to the plasmid (CAS P57).

Generation of cell lines for scTRIP

K562 cells were maintained in Iscove's Modified Dulbecco′s Medium (IMDM) + 10% FBS + 1% non-essential amino acids + 1% penicillin/streptomycin. We selected two K562 cell lines previously used in our lab that each contain a 'landing pad' at a unique location with a pair of asymmetric Lox sites for recombination (loc1 - chr8:146022171, loc2 - chr11:16,258,750). For each cell line, we replaced the original landing pad cassette with the same BFP-CS1 reporter in the scTRIP library so that we can capture the reporters from the landing pad and the scTRIP library using the same primers. Pool 1 was derived from the loc2 cell line, while Pools 2, 3 and 4 were derived from the loc1 cell line.

The scTRIP library and piggyBac transposase were co-transfected into K562 (LP cell lines) cells at a 3:1 ratio using the Neon Transfection System (Life Technologies). For each experiment, we transfected 2.4 million cells with 9μg of scTRIP library and 3μg of transposase. The cells were sorted after 24 hours for GFP-positive cells to enrich for cells that have integrated scTRIP reporters. For Pools 1 and 2, cells were sorted into pools of 100 cells each and allowed to grow until there were sufficient cells for RNA/DNA extraction and scTRIP experiments. Pool 3 contained the same cells as Pool 2, except that single cells were allowed to grow individually and pooled by hand just before the scTRIP experiments. This allowed for a more even representation of each individual

clone (which contains unique integrations) in the final pool. For Pool 4, transfected cells were first sorted into 96-well plates with 2 cells/well and allowed to grow individually and 100 wells were pooled by hand for scTRIP experiments.

Bulk expression measurements

We harvested DNA and RNA from scTRIP pools using the TRIzol reagent (Life Technologies). The RNA was treated with two rounds of DNase using the Rigorous DNase treatment protocol (Ambion) and the cDNA was synthesized with oligo-dT primers using the SuperScript IV First Strand Synthesis System (Invitrogen). The barcodes were then amplified from cDNA and genomic DNA (gDNA) using the Q5 High Fidelity 2X Master Mix (New England Biolabs) with primers specific to our reporter gene (CAS P58-59). We performed two technical replicates per cDNA/gDNA sample per pool. For each technical replicate, we performed 4 PCRs on each sample respectively and pooled the PCRs for purification. 4ng from each replicate was then further amplified with 2 rounds of PCR to add Illumina sequencing adapters (CAS P60-63). The barcodes were sequenced on the Illumina NextSeq platform.

Sequencing reads were processed using a custom script. Reads that contained tripBCs in the correct sequence context were included in downstream analysis. We further filtered tripBCs such that they are at least 3 hamming distance apart from every other barcode to account for mutations that occurred during PCR and sequencing. We added a pseudocount to the RNA counts to include barcode pairs that had DNA but no RNA reads, and calculated the expression of each barcode pair as log2(RNA/DNA). We then calculated the average expression of each integration from 2 technical replicates.

scTRIP integration mapping

To map the locations of scTRIP integrations, we digested pooled gDNA with a combination of AvrII, NheI, SpeI and XbaI for 16 hours. The digestions were purified and self-ligated at 16°C for another 16 hours. After purifying the ligations, we performed inverse PCR to amplify the barcodes with the associated genomic DNA region (CAS P59 and P64). For each pool, we performed 2

technical replicates with 8 PCRs per replicate and pooled the PCRs of each replicate for purification. We then used 8ng of each replicate for further amplification with 2 rounds of PCR to add Illumina sequencing adapters (CAS P55 and P65). The sequencing library was sequenced on the Illumina NextSeq platform.

The barcodes of each read were matched with the sequence of its integration site. The integration site sequences were then aligned to hg38 using bwa with default parameters. Only barcodes that mapped to a unique location were kept for downstream analyses.

smFISH

Single-molecule FISH was performed on the two 'landing pad' locations that were in the original cell lines used for scTRIP (see Generation of cell lines for scTRIP above). ClampFISH probes for the reporter genes were designed using the Raj Lab Probe Design Tool (rajlab.seas.penn.edu). Each probe was broken into three arms to be synthesized by IDT. The 5' of the left arm is labeled by a hexynyl group, and the 3' of the right arm is labeled by NHS-azide. The right arm fragment was purified by HPLC. All three components were resuspended in nuclease-free H2O to a concentration of 400 uM. The three arms were ligated by T7 ligase (NEB, Cat M0318L), at 25 C overnight. then purified using the Monarch PCR and DNA cleanup Kit (NEB, Cat T1030S) and eluted with 40 ul of nuclease-free water. After the ligation, each probe is stored at -20 C. The list of oligos used in this paper can be found in Supplementary Table S2. ClampFISH was performed according to the suspension cell line protocol of clampFISH (Rouhanifard et al. 2018). 0.7 million cells were collected and fixed in 2 mL of fixing buffer containing 4% formaldehyde for 10 min, then permeabilized in 70% EtOH at 4 C for 24 hours. The primary ClampFISH probes were then hybridized for 4 hours at 37 C in the hybridization buffer (10% Dextran Sulfate, 10% Formamide, 2X SSC, 0.25% Triton X). After hybridization, cells were spun down gently at 1000 rcf for 2 min. Cells were washed twice with the washing buffer (20% formamide, 2X SSC, 0.25% Triton X) for 30 min at 37 C. The secondary probes were then hybridized to cells at 37 C for 2 hours and the cells were then washed twice with washing buffer for 30 min at 37 C. The primary and secondary probes

are “clamped” in place through a click reaction (CuSO4 75 uM, BTTAA 150 uM, Sodium Ascorbate 2.5 mM in 2X SSC) for 20 min at 37 C. The cells were then washed twice in the washing buffer at 37C for 30 min each wash. Then, the cells were hybridized with the hybridization buffer with tertiary probes for 2 hours at 37C. We complete 6 cycles of hybridization for all our experiments. After the final washes, cells were incubated at 37 C with 100mM DAPI for 20 min, washed twice with PBS, resuspended in the anti-fade buffer, and spun onto a 1.5 coverslip (part number) using a cytospin cytocentrifuge (Thermo Scientific), mounted onto a glass slide, sealed with a sealant, and stored at 4C.

scTRIP library using the 10X Genomics platform

We used the Chromium Single Cell 3’ kit (v3.1) from 10X Genomics for scTRIP sequencing.

Cell preparation

We followed manufacturers instructions for preparing single-cell suspensions as described in the kit guidelines for 10X kit v3.1. We used a cell counter to measure the number of cells and viability, and used cell preparations with greater than 95% cell viability. We targeted 20,000 cells per replicate by loading 51,260 cells in the instrument at a concentration of 1,100 cells per ul.

Cell barcoding and reverse transcription

10X Genomics Chromium sequencing library preparation was done following the 10X Genomics user guidelines for 3’ reagent kit v3.1 with the following modifications: no 10X template switching oligo (PN 3000228) was added to the Master Mix (Step 1.1). To correct for the missing volume, 2.4 ul of Molecular Grade H2O was added to the master mix per reaction. For the cDNA amplification (Step 2.2), no 10X provided reagents were used. Instead, a custom set of primers (CAS P20) were used with 14 cycles of amplification with the provided 10X protocol (Step 2.2 d).

Barcode PCR and library preparation

Nested PCR was used for sqeuencing library construction. PCR library construction was split into capture sequence captured and poly(A) sequence captured PCRs. Both PCR reactions were done with 2 ul purified cDNA from 10X GEM-RT, 2.5 ul 10 uM reporter-specific forward primer

(CAS P44, CAS P45), 2.5 ul 10 uM capture-sequence, and poly(A) sequence-specific primers (CAS P20 and CAS P32), and 25 ul Q5 Mastermix (NEB) in 50 ul total volume. 10 cycles of amplification were done with Tm = 61 for the capture sequence and Tm = 68 for the poly(A) sequence, with an extension of 30s at 72 C. After PCRs were finished, clean up was done with NEB PCR clean up kit.

A final round of PCR was done to add standard Illumina P5 and P7 adapters. PCR products from poly(A) and capture sequence reactions are combined with equal concentration. The final PCR reaction contains 2 ul of mixed PCR product, 1.25 ul 10 uM forward primer for capture sequence (CAS P49), 1.25 ul 10 uM reverse primer for capture sequence (CAS PC4), 1.25 ul 10 uM reverse primer for poly(A) sequence (CAS P48), 1.25 ul 10 uM forward primer for poly(A) sequence (CAS PP4), 25 ul of Q5 Mastermix (NEB), and 18 ul of H2O. 10 cycles of amplification was done with T_m = 59 C, with the extension time of 30s at 72C. A final extension with 5 minutes at 7C was done. After PCR reactions, an SPRI Select Beads(Beckman Coulter) size selection was done with 0.65X concentration.

scTRIP data processing

Parsing of reads

We first identified the reads that match the constant sequence in our reporter gene. We used two versions of constant sequence to match against, depending on if the read was captured using the polyA sequence on the mRNA or the capture sequence specific to the 10X beads. We used a fuzzy match algorithm to capture reads that have a mismatch at these positions due to sequencing error. From each read, we parsed out the cell barcode, 10X UMI and TRIP barcode. We then collapsed reads with identical cell barcodes, UMI and TRIP barcodes into one “quad” and kept track of the number of reads supporting each quad. For downstream analysis, we filtered out quads with low numbers of supporting reads since these are likely to be enriched for PCR artifacts. We next processed the quads to error correct the cell barcodes and trip barcodes before estimating the mean and variance using a model based approach. These steps are described below.

Cell barcode error correction

To correct for PCR artifact and sequencing errors, a custom script was used to error-correct for 10x cell barcodes(https://github.com/barakcohenlab/castools/). Briefly, we first acquired the empirical distribution of the hamming distances among observed 10X cell Barcodes, we found that more than 99% of 10X cell Barcodes are more than 6 hamming distances away from each other, making error correction a feasible approach to denoise the data. We employed an approach similar to the original trip paper27. We first identify cell barcodes that match perfectly to the 10X cell barcode white list, when we order them based on their abundance. We then compare the cell barcodes that are not in the white list against the ordered whitelisted cell barcodes, if the Hamming distance between the unwhitelisted cell barcodes is within 2 hamming distances away from the whitelisted cell barcodes, we correct the unwhitelisted cell barcode. With cell barcode correction, we recovered 12% of the previously unusable reads.

TRIP barcode error correction

Due to the random synthesis of the TRIP barcode, a different approach is taken for error correction for the TRIP barcodes. Briefly, all the TRIP barcodes are ranked based on abundance. We then find any reads containing TRIP barcodes that are within 4 hamming distances to the most abundant TRIP barcode and correct them to the most abundant TRIP barcode. We next remove the most abundant TRIP barcodes and any corrected barcodes. We repeat this process until we have iterated through all TRIP barcodes. The relevant script can be found n the GitHub repository associated with this manuscript (https://github.com/barakcohenlab/castools/)

Transcriptome analyses associated with scTRIP

Processing the single-cell transcriptome data

The single-cell RNAseq data was processed with CellRanger 6.0.1 and Scanpy 1.9.1. Briefly, the raw reads were processed with the standard single-cell expression cell line pipeline line. The resulting expression matrix was then input into Scanpy for further visualization and clustering.

Computing cell-cycle specific expression mean and noise at different genomic locations

To compute the cell-cycle specific expression mean and noise at different genomic locations,

individual cells were assigned a cell cycle phase of G1, S, or G2/M using a previously reported set of cell-cycle specific marker genes with Scanpy 1.9.1 45. To ensure an accurate measurement of expression mean and noise, genomic locations with measurement less than 15 in any phase were excluded from the cell cycle analysis. Based on this filtering criterion, 345 out of 942 genomic locations were used for this analysis. The relevant scripts can be found in the GitHub repository associated with this manuscript (https://github.com/barakcohenlab/castools/).

Identifying single cell clones

We identified the individual clones that for the second batch of scTRIP experiment based on the fact that the second batch of scTRIP cells grew out of 100 two-cell clones. Since most of the clones will have unique integrations into unique genomic locations, the cells expanded from the same clone will have unique sets of trip barcodes. Given the biological expression noise at single-cell level, as well as the systematic noise with single-cell RNAseq, not all barcodes will present in all cells, nor will the cell barcodes will be uniquely linked to correct sets of trip barcodes. To identify correct sets of barcodes, we only used tripBC that are confidently linked to a given cell by filtering with an automatic inflection point detection algorithm on the barcode abundance histogram. The relevant scripts can be found in the GitHub repository associated with this manuscript (https://github.com/barakcohenlab/castools/). We then aggregate the per cell trip barcode combinations to link any trip barcodes that occurred in any cells. With this method, we identified 196 clones with a median integration number of 3.

Validation of individual clones

We extracted gDNA from 16 clones that were grown out from Pool 4. We then amplified the barcodes from each clone using Q5 High Fidelity 2X Master Mix (M0492, New England Biolabs) with primers specific to our reporter gene (CAS P58-59). For each clone, we performed 4 PCRs and pooled the PCRs for purification. 4ng from each clone was then further amplified with 2 rounds of PCR to add Illumina sequencing adapters (CAS P60-63). The barcodes were sequenced on the Illumina NextSeq platform.

K562 Hi-C

We performed Hi-C on wild-type K562 cells with the Arima Hi-C kit (A510008) according to the manufacturer's protocols (3 replicates, 870 million reads total). The reads were then processed with the Juicer pipeline46 to generate HiC contact files for each replicate. We then used the peakHiC tool47 to call loops from each IR with the following parameters: window size = 80, alphaFDR = 0.5, minimum distance = 10kb, qWr = 1. Using these parameters each IR was looped to a median of 3 regions (range 0-7).

Logistic Regression model for Intrinsic and Extrinsic features associated with MIN

We used histone ChIP-seq and ATAC-seq datasets from ENCODE(ENCODE Project Consortium 2012) and overlapped their signals with each IR using used bedtools v2.27.1(Quinlan and Hall 2010). For all features we considered the 20kb upstream and downstream of each IR respectively. For each histone modification, we computed the mean ChIP signal around the IRs. For ATAC-seq, we calculated the total number of peaks with the bedtools map count option. To look for TF motifs we counted the numbers of each motif for TFs expressed in K562s (see above) in each surrounding IR sequence using FIMO(Grant et al. 2011) (MEME suite 5.0.4). Because this resulted in a long list of TFs we further filtered the TFs to include only those with a significant correlation with MIN levels in the regression model. To determine the numbers of enhancers interacting with each IR we annotated the loops called from peakHiC above with chromHMM enhancer annotations using the GenomicInteractions R package(Harmston et al. 2015) and counted the number of enhancers.

For the extrinsic features, we calculated the proportion of cells in the "stem-like" substate and "differentiated" substate and different cell cycle phases based on the barcodes that appeared in those substates. We removed IR locations that have less than 30 cells in any of the substates.

We used the glm function in R (version 3.6.3) to fit logistic regression models. We separated the IR locations into top 20% MIN and bottom 20% MIN and used logistic regression to classify locations. We first fit a model with just local sequence features (chromatin modifications, number of TF motifs, number of loops, whether the IR location is in a gene, GC content and the number

of ATAC-seq peaks). We used data from one experiment for training the model and used data from another experiment as a holdout set of data to estimate the performance of the classifier. We next fit a model with cellular information for each IR location: proportion of cells with data for the IR location in S phase of the cell cycle, in G2 phase and the proportion of cells that are in the "stem-like" substate of K562 cells(Litzenburger et al. 2017). Lastly, we fit a model that incorporated the extrinsic features and the significant predictors from the intrinsic features model.

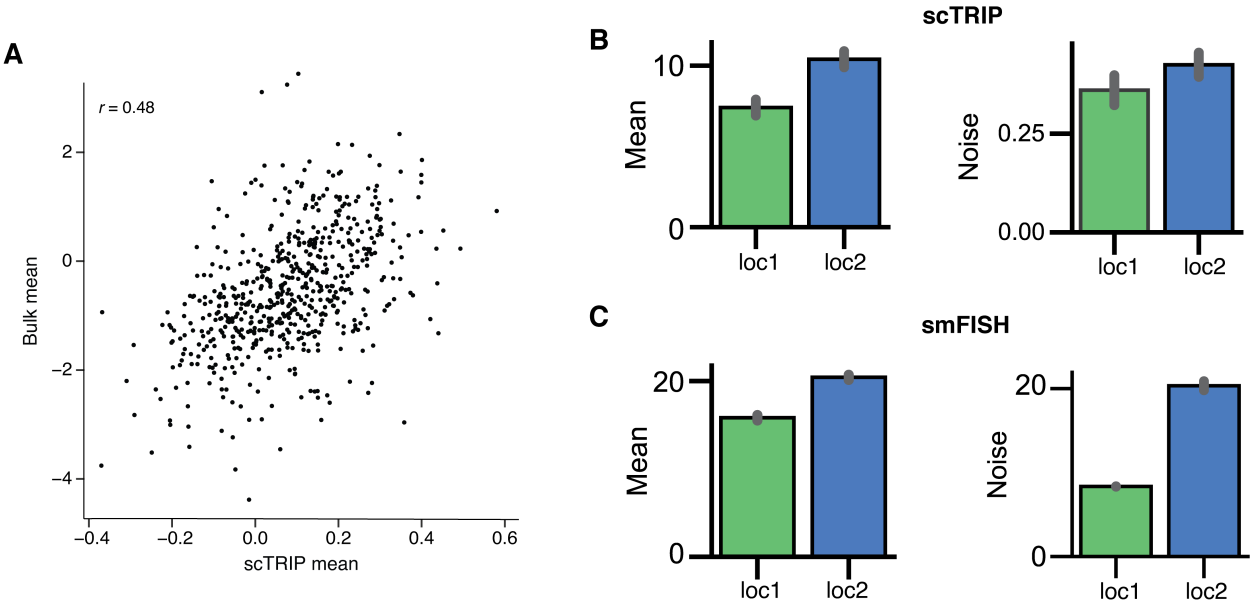

Figure 3.7: **scTRIP is reproducible.**(A) Correlation between means expression measured by scTRIP vs measured in bulk. (B, C) Mean and noise levels of two landing pad locations (loc1 and loc2) measured by scTRIP (B) or smFISH (C). Error bars represent 1 std from two biological replicates. .

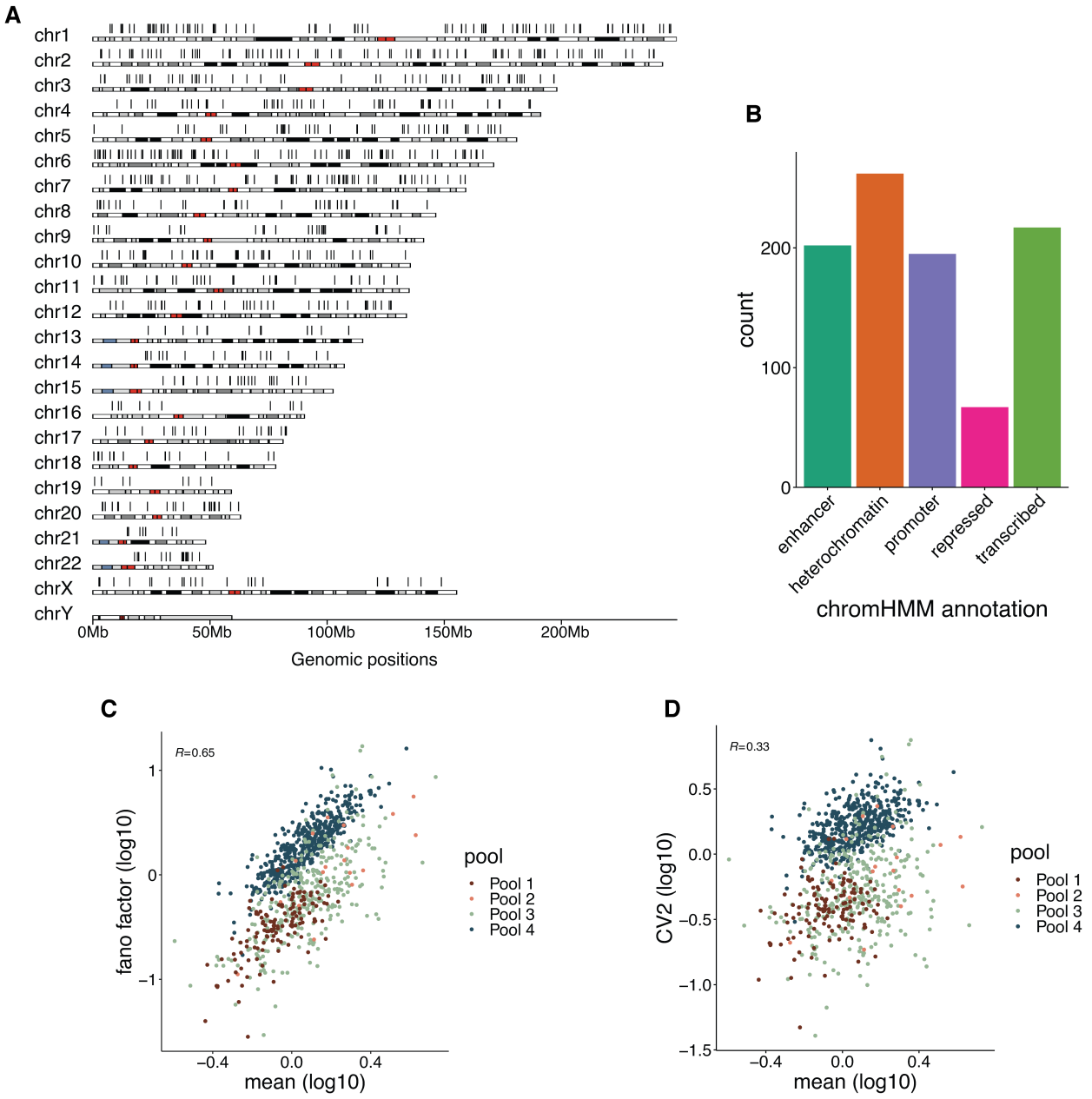

Figure 3.8: **Measurements of mean-independent noise across different chromosomal environments.**(A) IR locations are distributed all throughout the genome. Each black bar above the ideogram represents a separate integration. (B) IR locations are found distributed across different chromatin types. (C, D) Expression mean is well correlated with fano factor (C) and CV2 (D).

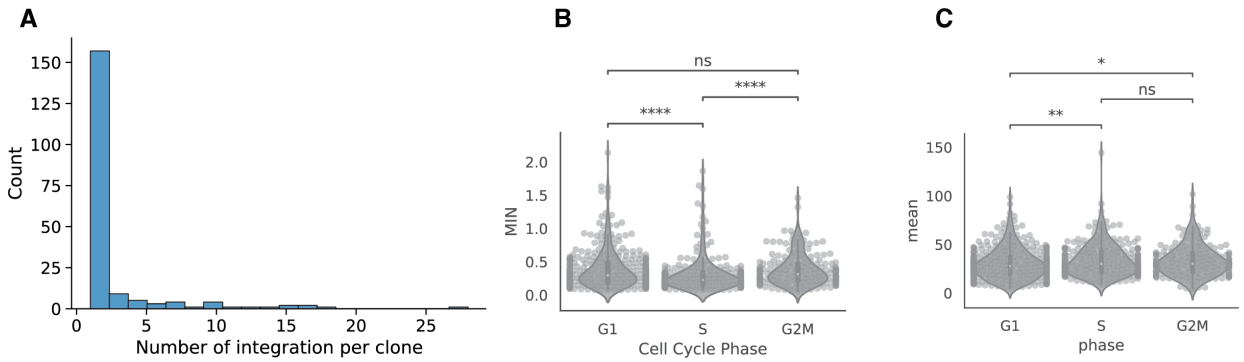

Figure 3.9: **Quantifying the extrinsic source of expression noise.**(A) Number of integrations per clone The mean number of integration is 3. (B,C) Scatterplot of expression mean and MIN in different cell cycle phases. (Student t-test: *: p-value < 0.05, **: p-value < 0.01, ****: p-value < 0.00001).

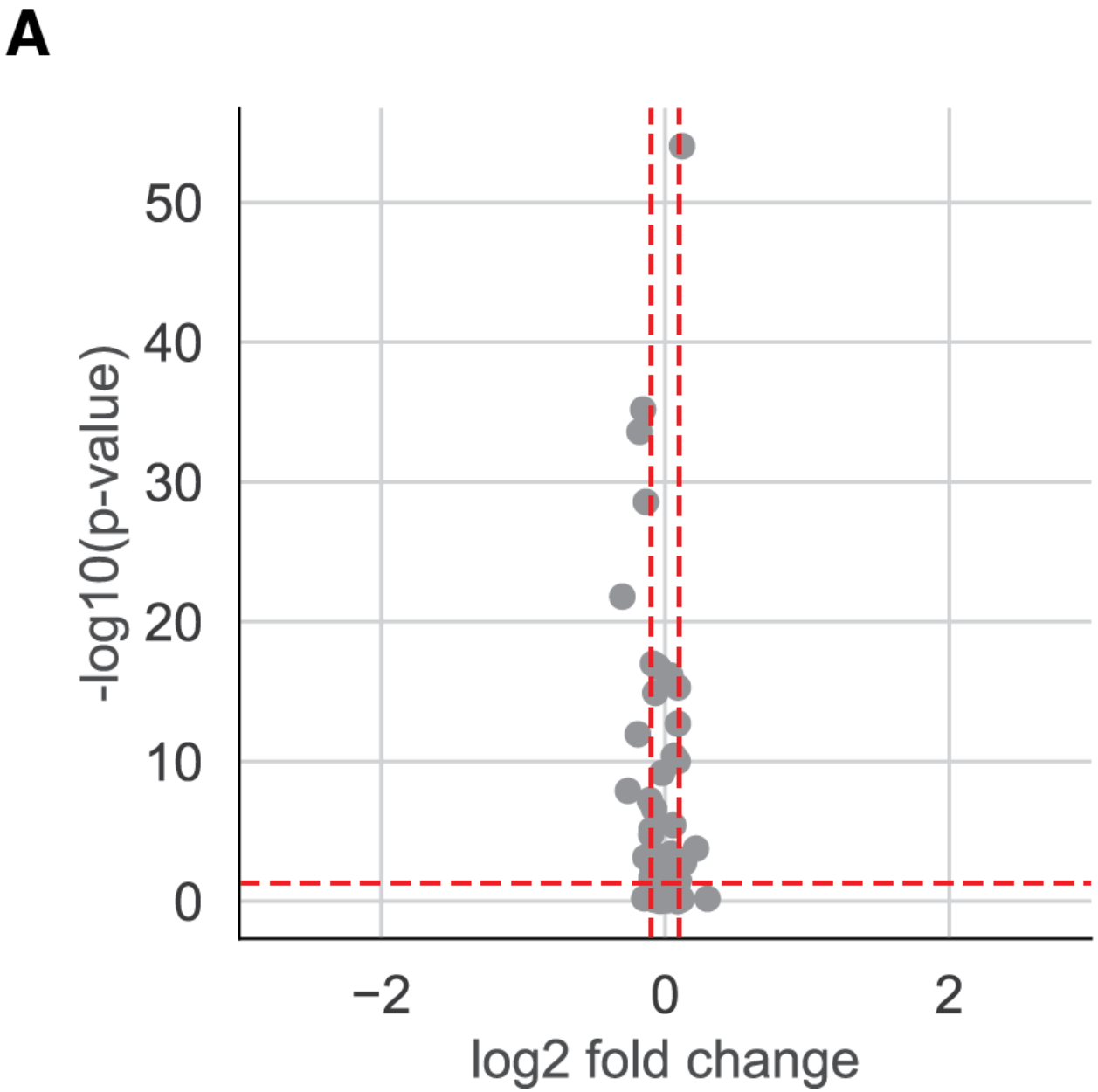

Figure 3.10: **Shuffling of IR location label and the gene labels removes differential expression effect.**

Chapter 4: Discussion

This thesis describes the development of two functional genomics methods at single-cell resolution. These methods systematically examined how cellular contexts, chromatin environment, and local regulatory sequences collectively control expression mean and noise. scMPRA can systematically examine how different local regulatory sequences and cellular contexts (cell type and cell substates) control expression mean and noise. scTRIP can systematically examine how the chromatin environment and the cellular contexts control expression mean and noise. The overall observations suggest the hierarchical and modular model of gene regulation is like true. We found that (1) the expression level of the upstream regulatory directly tracks with the overall expression level of a library of variants in the downstream gene's promoter, (2) cell cycle phase and stable cell substates have a largely general effect on a library of core promoters, (3) the cell substates' effect on expression mean and noise is independent of the genomic context. The specificity seems to arise from the quantitative difference in activity of local regulatory sequences, rather than the qualitative differences such as categories. We did observe the cell substates effect on developmental core promoters is different from the effect on housekeeping core promoters, but it can be explained by the fact that housekeeping core promoters are overall stronger. At the same time, we observed many cell-type specific variants in Gnb3 promoter, but it can be explained by the creation of a novel binding site for other TFs. These observations also provide answers to some standing questions about gene regulation.

Are there cell-type specific non-coding variants?

scMPRA enabled us to directly test if non-coding variants are cell-type specific. We designed a library of mutants that disrupt the conserved regions and TFBS. We found more than half of them have large effects on the regulatory activity. Moreover, we found about one-third of those designed

variants have cell-type specific effects. This is likely an inflated number, given we intentionally altered the non-coding region that has regulatory functions. Nevertheless, some of those cell-type specific variants can have more than 2-fold differences across cell types. This suggests that cell-type specific differences in regulatory activity could be a mechanism to confer specificity.

Are there generalizable principles cross cell types?

Because the known TF of the regulatory element is present at different concentrations in all the cell types in our system, we also found that the wild-type expression of the promoter we tested tracks the endogenous expression of the gene in different cell types. Moreover, the variants that are not cell-type specific seem to follow this general rank order, where if the wild-type expression is high in one cell type, the variant is also higher in this cell type compared to others. This observation suggests that many variants have the same effect in different cell types. This supports the idea that the overall affinity of a regulatory element is likely modular to the level of the TFs in different cell types if the relevant TFs are present.

What controls expression noise?

scMPRA and scTRIP enabled us to systematically examine how different local sequences, different genomic environments, and different cellular contexts collectively control noise. Not surprisingly, but anticlimactically, we found that expression mean explains more than 60% of expression noise. Expression mean remains the most predictive factor of expression noise. It is not surprising because the same biochemical mechanism produces expression mean and noise. We found that the rest expression noise can be explained by the sequence features and cellular context. Chromatin marks and TFBS are associated with high and low expression noise. Cell cycle and cell substates have a very general effect on expression noise regardless of sequence features. Here we also found some modular effects between sequence features and cellular context, suggesting some modularity of gene regulation.

How is the hierarchical and modular model of gene regulation confer specificity?

The existence of cell-type specific variant effects, and some scaling effect of the non-cell-type

specific variants likely points us to a threshold model of gene expression. The level of transcription factors set the possible range for a regulatory element, and variants in regulatory elements control the difference within this range. Combining the variant effect and the concentrations of transcription factors, many levels of expression can be achieved for different proteins downstream. In this model, to break symmetry and to turn on different developmental programs from an isogenic population (i.e. undifferentiated stem cells), the high and low expression of some transcription factors in different portions of the cells can lead to very different levels of activity for regulatory sequences with a different configuration, hence initiating diverse pathways that lead to differentiation. This threshold model suggests the existence of modularity of gene regulation.

These methods also have wide applicability to (1) examine how cellular context, and chromatin environment affect expression mean and noise, (2) detect cell-type specific variant effects *in vivo*, and (3) design synthetic sequences to control expression mean and noise.

One immediate next step for scMPRA is to determine its applicability in other tissues. Several tissues can be electroporated *ex vivo*, such as mouse small intestine, mouse embryonic kidney, and mouse embryonic spinal cord [118–121]. Other tissues can be transduced with the AAV virus, such as the mouse brain and the mouse heart [122–125]. Those tissues would be excellent candidates for testing developmental programs and cell-type specific disease variants.

At the same time, a large-scale screening of different genomic locations using scTRIP could also define the genomic landscape for gene transfer therapy. Identifying genomic locations that allow high expression mean and low expression noise, while having a minimal effect on endogenous gene expression.

The development of these single-cell functional genomics methods was inspired by the observations from the MPRA experiments. For example, the complex rules governing how TFBS within regulatory elements and the simple rules underlying how regulatory elements collective control expression lead to the motivation of measuring the cellular context in which those regulatory elements function. This leads to the development of single-cell functional genomics methods. And

these new methods, scMPRA and scTRIP, could potentially lead to new observations and paradoxes that require new methods to answer. This cycle seems to be a part of the scientific process.

Appendix A: Modular effects of gene promoters and chromatin environments on noise in gene expression

Siqi Zhao[1,2], Zachary Pincus[2,3], Barak A Cohen[1,2]

1 The Edison Family Center for Genome Sciences, 2 Department of Genetics, 3 Department of Developmental Biology, Washington University School of Medicine in St. Louis, USA

This work was done in collaboration with Zach Pincus, and Barak Cohen. I conceived this project. Barak Cohen and I designed the experiments. Zach Pincuse helped with image acquisition. I carried out all other experiments and analyses. Barak and I wrote the paper.

A.1 Abstract

Genetically identical cells growing in the same environment can have large differences in gene expression. Both locally acting cis-regulatory sequences (CRS) and the regional properties of chromosomal environments influence the noisiness of a gene's expression. Whether or not local CRS and regional chromosomal environments act independently on noise, or whether they interact in complex ways is unknown. To address this question, we measured the expression mean and noise of reporter genes driven by different CRS at multiple chromosomal locations. While a strong power law relationship between mean expression and noise explains 60% of noise for diverse promoters across chromosomal locations, modeling the residual mean-independent noise suggests that chromosomal environments have strong effects on expression noise by influencing how quickly genes transition from their inactive states to their active states and that the effects of local CRS and regional chromatin on noise are largely independent. Our results support a modular genome in which regional chromatin modifies the inherent relationship between the mean and noise of expression regardless of the identity of the promoter sequence.

A.2 Introduction

Genetically identical cells growing in the same environment can display large differences in the mRNA levels of expressed genes [2, 3, 158]. This cell-to-cell variance in mRNA abundance among isogenic cells is called noise in gene expression and drives important aspects of biology. Cell fate decisions for olfactory sensory neurons [69], retinal pigment cells [159], intestinal crypt cells [160], neural precursors in the spinal cord [161, 162], and hematopoiesis [67] are governed by the noisy expression of lineage-determining transcription factors. Random fluctuations in gene expression also affect how individual melanoma cells respond to chemotherapy [70]. Single-cell transcriptome profiles from scRNA-seq suggest that expression noise is universal for many cell types [163–165] and likely accounts for different cell states within the same cell type [166]. Noise in gene expression is thus an important component of both development and disease.

Noise in gene expression results from random fluctuations in the molecular processes that control transcription. Single-cell studies show that transcription is not continuous, but instead occurs in discrete bursts of activity [167, 168]. A gene's promoter sequences [96, 169], its epigenetic modifications [170], or its patterns of DNA looping [171] can all influence the burst size and frequency of genes. For different genes, the burst frequency can range from just a few minutes to tens of hours [167, 172]. Consequently, cells in the midst of bursting will have higher mRNA counts than cells waiting for their next burst of expression. The bursty dynamics of mammalian transcription is an important source of expression noise.

Two types of regulatory information can influence expression noise: locally acting CRS, such as promoters and proximal enhancers, and more distally acting factors, such as long-range enhancers and sequences that control the regional chromatin state. In this study, we define the regional "chromosomal environment" as the combination of a gene's distally acting enhancers and its broader chromatin landscape. Classic studies of Position Effect Variegation (PEV) show that different chromosomal environments can have large effects on noise [173]. More recent work has shown how features associated with local cis-regulatory sequences (CRS), such as the TATA-box [174],

the number of transcription factor binding sites [175], the local nucleosome occupancy [176], and local epigenetic modifications [170] can all affect expression noise. Two studies used randomly integrated reporter genes to show that different chromatin environments can affect expression noise [144, 176]. However, how the effects of regional chromosomal environments combine with the effects of local regulatory features to control noise remains an open question.

Local and regional cis-regulatory information may act independently to control noise in expression, or there may be complex interactions between local CRS and their regional chromosomal environments. If these two sources of cis-regulatory information act independently on noise, then the effects of chromosomal environments should be the same for different classes of local CRS. Alternatively, a gene's noise may depend on complex interactions between local CRS and their chromosomal environments: certain locally acting cis-elements may only increase noise in certain chromosomal environments. We previously showed that local and regional cis-regulatory information act independently to control mean levels of expression [35]. In this study we sought to determine whether the modularity between local and regional cis-regulatory information extends to the control of noise in expression.

To address this question, we constructed a series of reporter genes driven by different promoters and integrated them into several defined genomic locations in the K562 cell line. We measured both the mean and noise of each reporter gene at each genomic location with single molecule Fluorescence in situ Hybridization (smFISH). Our results show that across all reporters at all locations, 60% of expression noise depends only on the mean level of the reporter gene and is captured by a power law. The residual mean-independent noise is mainly associated with the epigenetic signature of the regional chromosomal environment. Given the same mean level of expression, more active chromatin environments generally result in less expression noise than more repressive environments. By comparing diverse reporter genes integrated at the same chromosomal locations we found that the effects of different chromosomal environments on expression noise were independent of the specific promoter. Taken together our results support the notion of a modular genome in which local

and regional cis-regulatory information act independently to control both mean levels of expression and noise in expression.

A.3 Results

A.3.1 An experimental system to quantify noise at diverse chromosomal locations

We constructed a system to separate the contributions of locally acting CRS and regional chromosomal environments to expression noise. We used a set of previously generated landing pad cell lines that contain reporter genes integrated in defined chromosomal locations across the K562 genome [35]. Each landing pad cell line has an identical reporter gene cassette in a different mapped genomic location (**Figure A.1 A,B**). In this study, we used landing pads integrated at twenty-two different genomic locations, with different epigenetic marks (**Figure A.6**).

Each landing pad carries a common reporter gene consisting of a locally acting cis-regulatory sequence (CRS) driving the expression of Enhanced Green Fluorescent Protein (eGFP) and flanked by a pair of asymmetric Lox sites allowing us to easily exchange the reporter gene cassette. (**Figure A.1 A**). For all twenty-two locations, we integrated an eGFP reporter driven by the cytomegalovirus (CMV) promoter. To explore the dependence of noise on local CRS, we integrated three other reporter cassettes with tdTomato reporter gene driven by PSMP2, HBZ, and CHMP2A promoters at four locations (**Figure A.1 B**). To simplify the system, we intentionally chose not to include an intron in the reporter cassette to avoid potential confounding effects of different genomic locations on the efficiency of splicing [177]. With this system, we can systematically probe how different chromosomal environments affect the activities of different local CRS.

One of the challenges of studying the effect of transcriptional regulation is to exclude the confounding effects of post-transcriptional regulation. Hence, for each reporter gene at each chromosomal location we performed clampFISH, a smFISH method, to measure the absolute

mRNA counts in individual cells [178]. To accurately quantify both expression mean and noise, we quantified mRNA molecules from more than 8x105 cells and obtained more than 1000 single-cell mRNA measurements per cell line per replicate. The experimental measurements are highly reproducible for both mean and noise at each genomic location (R^2 = 0.947 for mean, R^2 = 0.88 for noise, **Figure A.1 D,E**). The absolute measurements of mRNA molecules using smFISH allowed us to use computational modeling to infer transcriptional dynamics.

A.3.2 The mean and noise of gene expression are linked by a power-law relationship

We observe dramatic changes in the mean and noise of expression that depend on genomic locations. We observed a 6-fold difference in mRNA mean and a 7-fold difference in noise among the genomic locations we sampled (**Figure A.1 C**). This dynamic range of mean and noise is consistent with the range observed in a previous study where the investigators measured mRNA in cell lines with random integrations in a mammalian cell line [176]. Thus, our data is consistent with both classic work on PEV and more recent single-cell studies, which show that changing the genomic location of a gene has large effects on its expression mean and noise.

We first established the general relationship between the mean and noise of expression in our data. In all stochastic processes there is a strong dependency of noise on the mean output levels. Because changing the genomic location of a gene affects its mean expression, its noise will also necessarily change due to the dependency of the noise on the mean. Therefore, it is important to establish the general relationship between expression mean and noise in our experiments so that we can decompose the effects of genomic location on noise into effects that are due solely to changes in mean expression levels and those that are independent of mean level changes.

To characterize the general relationship between the mean and noise in our data we performed a log-log regression of expression noise on mean expression levels across all CRS at all genomic locations. Consistent with previous findings, this analysis revealed a power law relating expression

noise to mean expression levels. Differing from studies using fluorescence-based single-cell quantification of genome-wide protein abundance, which report a degree of 1.96, we observed a smaller degree of 1.15 (noise = mean$^{1.15}$, R^2 = 0.6) (**Figure A.1 F**) [71, 176, 179]. To determine the significance of the fitted degree for the power law relation between expression mean and noise, we performed 105 bootstrap simulations and obtained a p-value of 9*10-5 for the fit to the actual data (**Figure A.1 G**). We interpret the higher degree for the power law relationship for the protein distribution to represent additional noise introduced by post-transcriptional steps in gene expression [180]. In our experiments, 60% of the noise in mRNA levels is explained solely by the effects of different chromosomal locations on mean levels. The same power law relation holds regardless of the local CRS identity: there are no promoter specific mean-noise relationships in our data. When we plot the mean-noise relationship for different promoters alone, the power law relation remains largely the same (**Figure A.7 A-D**). This result suggests that a general aspect of the transcriptional mechanism accounts for a significant fraction of observed noise, regardless of the specific sequence context.

The power law's degree of 1.15 reflects the bursty nature of mammalian gene expression. If the reporter genes were constitutively producing mRNA, then we would expect Poisson dynamics where the degree for scaling is 1.0 [72, 180]. The non-Poisson relationship in our data suggests the existence of "ON" and "OFF" states for transcription activation [2], which agrees with the bursty dynamics observed in mammalian cell systems and Drosophila melanogaster [168, 171]. Consistent with the existence of an OFF state, for each reporter gene at each location there was always a subpopulation of cells with no mRNA (**Figure A.7 D**). This suggests that in some cells, transcripts are degraded completely before the reporter gene transitions to the ON state. To confirm that this observation is not due to false negative labeling of mRNAs, we performed smFISH on the introns of the constitutively expressed gene Actb, which encodes the beta actin protein. For Actb, less than 1% of the cells have no labeled transcript (**Figure A.7 E**), which suggests that the larger fraction of cells with no reporter gene mRNAs are not false negatives resulting from experimental

artifacts such as poor fixation or permeabilization. Both the observation of a non-Poisson dynamics and the existence of cells with no mRNA present suggests that there are regulated, slow steps in transcription that generate noise in gene expression. We sought to explore the mechanism behind the difference in transcriptional dynamics at different genomic locations.

A.3.3 Mean Independent Noise (MIN) describes expression noise without the mean effect.

The power law relation shows that mean expression levels explain 60% of expression noise. We define the residual 40% of expression noise that is not explained by mean levels as Mean Independent Noise (MIN). In all subsequent analyses we use MIN as the metric of noise because it shows no dependence on mean levels (**Figure A.2 A**), while other commonly used metrics of noise, such as the Fano Factor (σ^2/μ) or CV2 (σ^2/μ^2), still show some dependence on mean levels (**Figure A.8 A,B**). Because MIN is based on the power law fit to our data (MIN $\sigma^2/\mu^{1.15}$ it more naturally expresses the relationship between mean expression levels and noise. In all subsequent analyses we use MIN as a measure of the effects of chromosomal environments on expression noise.

A.3.4 Chromatin states explain Mean Independent Noise at different chromosomal locations

What mechanisms control MIN? We hypothesized that regional chromatin environments have direct effects on MIN. Before analyzing whether regional chromatin environments affect MIN, we addressed whether the reporter gene causes gross alterations in the chromatin environment. We performed 4C [181] on two cell lines containing reporter genes and found that the contact maps agree with the contact frequencies derived from the Hi-C data on cells without reporter genes (**Figure A.8 E,F**) [55, 150]. We then asked whether the different epigenetic properties of genomic locations explain the differences in expression mean and MIN at different genomic locations [23]. We first examined individual epigenetic marks, and found some active chromatin marks are enriched in

active regions 5kb around the insertion site (**Figure A.10 A-F**). However, due to the sparseness of the peaks of histone modifications within 500 bp around our insertion sites, we decided to look at the aggregated epigenetic information (**Figure A.10 G**). We then plotted the MIN from cell lines with CMV promoters against the genomic annotations of the 100 bp flanking the reporter genes. Genomic locations labeled as Transcribed by the chromHMM and Segway [149, 182] combined annotations have lower MIN than regions labeled as Repressed (**Figure A.2 B,C**). We observed similar results using either the Fano Factor or CV^2 as the noise metric (**Figure A.8 C,D**). We hence classified our genomic locations as active (Transcribed, Weak Enhancer, and CTCF site based on ChromHMM segmentations) and repressed [149, 182, 183], which combine diverse epigenetic regulatory information into annotations of chromosomal locations.

Interestingly, we did not find strong differences in mean expression levels between active and repressed locations (**Figure A.2 D**). Previous studies of the genomic location effect on expression mean have shown that reporters integrated into the active locations have statistically higher mean expression [50]. However, those statistically higher expression mean levels are often driven by a small number of genomic locations. To illustrate this fact, we used a previously published dataset containing the mean measurement of a reporter gene integrated in thousands of genomic locations. We then subsampled the data and compared the z-score of the expression mean level at active and initiative locations. We found that there are no significant differences in mean expression levels when we randomly sample 22 locations from the whole dataset (**Figure A.2 E**). Taken together, our results suggest that the epigenetic differences between chromosomal locations have larger effects on the noise of gene expression than on mean levels of expression.

A.3.5 Computational model reveals the dynamics that changes cell-to-cell variability at different genomic locations

To explore the underlying mechanisms that explain the observed difference in MIN at different genomic locations, we studied how different genomic locations control the bursting dynamics

of expression. Recent studies suggest that the control of transcriptional bursts directly affects expression noise [172]. Theoretical and experimental works show that the bursty dynamics of single-cell gene expression can be described by the two-state ON/OFF model (Fig. 3A)[2, 72, 180]. At the same time, this framework has successfully been used to connect single-cell variability with transcriptional dynamics experimentally [2, 170, 176, 179, 184]. The ON/OFF model abstracts gene expression into four macroscopic processes, each with a corresponding rate constant. Kon and Koff describe the transition of chromatin between the ON and OFF states, Km describes the rate of mRNA production (which only occurs when the chromatin is in the ON state), and Kd describes the rate of RNA degradation. These rate constants set the burst size (Km/Koff) and burst frequency (Kon/Koff) of a gene [144, 185].

We employed an exhaustive fitting strategy for the ON/OFF model to identify parameter sets that might explain the differences in MIN between different genomic locations (Methods). We found a median of 23 sets of parameters that fit each of our experimental distributions and those sets of parameters cluster distinctively based on different genomic locations (**Figure A.3 B**). Compared to the thousands of sets of parameters obtained from fitting those parameters with protein distribution data, our fitted result is 100-fold less degenerative, this allows us to infer the differences in dynamics at different genomic locations.

We first examined the general trend for how different parameters in the ON/OFF model correlate with mean expression levels. We found that only the transition to the ON state is positively correlated with the increase of expression mean (**Figure A.3 C**); there is no change for the rate for transitioning to the OFF state (**Figure A.3 D**). This suggests that the genomic location effect on mean is mainly associated with the opening of the chromatin, and there is a constant process controlling the transition to the OFF state. This result agrees with the recent study showing that the inactivation of chromatin is a constant, and active process [186]. We also found a weak correlation between expression mean and Km, the rate of transcription (**Figure A.3 E**). Overall, the computational modeling agrees with experimental observations of how mammalian chromatin environments affect the mean of gene

expression.

We next asked how the different parameters of the ON/OFF model correlate with MIN. Surprisingly, we found that the transition to the ON state is negatively correlated with MIN (**Figure A.2 F**). This suggests that the faster the transition to the ON state, the lower the MIN. Thus, Kon appears to have opposing effects on mean and MIN, which suggests that expression mean and noise can be orthogonally controlled. We found that Km is also positively correlated with MIN (**Figure A.2 H**), suggesting that the rate of transcription increases both the mean and MIN of expression.

Overall, the ON/OFF model revealed that chromosomal environments mainly affect Km, the macroscopic parameter describing the rate of transcription, and Kon, the parameter describing the rate of activation. However, we discovered that MIN decreases as Kon increases, suggesting faster activation is a potential mechanism for suppressing expression noise. Moreover, the orthogonal effect of Kon on expression mean and MIN suggests that mean and MIN can be decoupled by tuning different macroscopic steps of gene expression.

A.3.6 Active chromosomal environment produces less expression noise by creating frequent but small transcriptional bursts

We then asked if the ON/OFF model explains the difference in MIN at active and repressed genomic locations. We found that reporter genes driven by the CMV promoter tend to have faster Kon when integrated in active locations compared to repressed locations, while Koff is not distinguishable for the transcribed and repressed locations (**Figure A.3 I,J**). This suggests that at active locations, reporter genes transition faster to the ON state than at repressed locations. However, repressed locations have slightly higher transcription rates (Km) compared to active locations (**Figure A.3 K**). To achieve the same mean level at a repressed location with slower ON rate, the transcription rate must be higher to compensate for the lower activity of the chromatin transition.

We then estimated the transcriptional burst size and frequency from the fitted parameters (**Figure A.3 C**). Intuitively, we found positive correlations between burst size and frequency with expression

mean (**Figure A.4 A,B**), increasing either burst frequency or burst size can lead to the increase of mean expression, and different chromatin environments can modulate both burst size and burst frequency. Interestingly, while burst size still is positively correlated with MIN, burst frequency is negatively correlated with MIN (**Figure A.4 D,E**). We also found that active locations have higher burst frequencies and lower burst sizes, while repressed locations have lower burst frequencies and higher burst sizes (**Figure A.4 F,G**). This suggests that faster burst dynamics reduce MIN without reducing expression mean level. This finding provides a potential mechanism for reducing expression noise without lowering the mean expression through controlling the transcriptional dynamics.

Overall, we found that expression mean and MIN are decoupled through differential control of different steps of transcription. We found that active genomic locations have MIN through faster burst dynamics but slightly smaller burst sizes compared to repressed locations. Moreover, the faster dynamics is due to faster transitions to the ON state, and not to slower transitions to the OFF state. These results suggest that expression mean and MIN can be decoupled through increasing the ON transition, but not by increasing the rate of transcription.

Independence of the effects of genome location and local CRS on noise in gene expression

To determine whether the effect of chromatin environments on MIN depends on the specific local CRSs, we analyzed the expression distribution of reporter genes with three other promoters (CHMP2A, PSMB2, and HBZ) at the same two active and two repressed genomic locations. Interestingly, the resulting mean-noise relationship still falls on the same power-law relationship (**Figure A.7 A**). While the promoter identity determines the mean level of expression all four locations (**Figure A.5 A**), we found that chromatin environments have a strong effect on MIN, regardless of the specific promoter identity (**Figure A.5 B**). To quantify the contribution of local CRSs and chromatin environment on MIN, we performed a two-way ANOVA, and found that genomic location explains 52% of the variance of the MIN ($p = 0.039$), while promoter identity is not a significant predictor of the observed variance of the MIN. In contrast, promoter identity

explains 65% of the variance of the mean ($p = 0.0012$), and the genomic location explains 20% of variance of the mean ($p = 0.046$) (**Figure A.5 E**). The ANOVA analysis affirms the observation that the regional chromatin environment of a gene has a strong effect on the MIN.

Agreeing with the data from CMV promoters, reporter genes with the other promoters integrated in the active locations also have smaller burst sizes and higher burst frequencies compared to those integrated in the repressed locations (**Figure A.5 C,D**). The same trends also hold for the individual macroscopic parameters for the ON/OFF model; we observe a higher Kon but similar Koff at active locations regardless of the promoter identity (**Figure ?? A,B**).

In summary, we found that the regional chromatin environment and the local promoter identity independently control expression. While promoter identity has a large effect on expression mean, the activeness of the chromatin environment has a large effect on MIN.

A.4 Discussion

Taken together, our results suggest that there is a strong inherent relationship between mean mRNA levels and expression noise, but that this relationship can be modified by the epigenetic properties of different chromosomal environments. About sixty percent of all noise in mRNA expression is set solely by a gene's mean level of mRNA production through a power law relationship, regardless of the identity or genomic location of the local CRSs. This observation suggests strong mechanistic constraints on noise that originate from fundamental properties of the transcriptional machinery. However, our data also show that the different properties of chromosomal environments can, in part, uncouple expression noise from mean levels of expression. We found that changing the dynamics of the transitions between ON/OFF states has an orthogonal effect on mean and mean-independent noise. This mechanism could allow natural selection to select for noisy expression without changing a gene's mean expression level. Our data also suggest that local and regional sources of cis-regulatory information act independently on a gene's expression noise. Reporter genes have lower noise when integrated at active regions of the genome than at repressed regions, and this

is true regardless of the identities of their promoters. Local and regional cis-regulatory information also control different aspects of transcriptional dynamics. The local cis-regulatory information controls the basic rates of transcription, regardless of its chromosomal environment, whereas the chromosomal environment has a large effect on the activation and inactivation dynamics of the gene, and consequently has a larger effect on mean-independent expression noise. We speculate that more active chromosomal environments have higher local concentrations of the components of the transcriptional machinery available, resulting in less heterogeneous transcription among individual cells. Meanwhile at more repressed genomic regions, the binding of the transcriptional machinery is a rarer event, increasing time intervals between transcriptional bursts and promoting heterogeneous expression in the population. This model suggests the possibility for predicting the noise in gene expression based on genomic location.

A.5 Methods

Landing Pad Design and Cell Line Maintenance. K562 cell lines carrying landing pads constructions were taken from our previously published work [35]. There are several important features of the landing pad cassette: First, a pair of asymmetric Lox sites is in the cassette, allowing integration of different DNA sequences into the same location. Second, a 12-bp DNA sequence barcode was cloned downstream of one of the Lox sites, allowing the mapping of the genomic location of each landing pad. Those locations were mapped individually to the Human hg19 genome. Twenty-two different landing pads were chosen for the experiments presented in this work. Landing pad K562 cells were cultured using a medium consisting of Iscove's Modified Dulbecco's Medium (IMDM) + 10% Fetal Bovine Serum (FBS) + 1% non-essential amino acids + 1% pen/strep.

Plasmid Design and Construction. The transfer vectors containing promoter cassette were subcloned with the following steps: First, the sequence of each promoter (PSMB2, CHMP2A, and HBZ) was extracted from the Ensembl genome browser [187]. The 700 bp sequence before the TSS for each promoter was synthesized by Twist Bioscience. Second, the synthesized fragment

was cloned into a pGL4.23 transfer vector backbone driving a tdTomato reporter using NEB HIFI assembly.

clampFISH Probe Design and Generation. ClampFISH probes for the reporter genes were designed using the Raj Lab Probe Design Tool (rajlab.seas.penn.edu). Each probe was broken into three arms to be synthesized by IDT. The 5' of the left arm is labeled by a hexynyl group, and the 3' of the right arm is labeled by NHS-azide. The right arm fragment was purified by HPLC. All three components were resuspended in nuclease-free H2O to a concentration of 400 uM. The three arms were ligated by T7 ligase (NEB, Cat M0318L), at 25 C overnight. then purified using the Monarch PCR and DNA cleanup Kit (NEB, Cat T1030S) and eluted with 40 ul of nuclease-free water. After the ligation, each probe is stored at -20 C.

clampFISH Experimental Procedure. ClampFISH was performed according to the suspension cell line protocol of clampFISH [178]. 0.7 million cells were collected and fixed in 2 mL of fixing buffer containing 4% formaldehyde for 10 min, then permeabilized in 70% EtOH at 4 C for 24 hours. The primary ClampFISH probes were then hybridized for 4 hours at 37 C in the hybridization buffer (10% Dextran Sulfate, 10% Formamide, 2X SSC, 0.25% Triton X). After hybridization, cells were spun down gently at 1000 rcf for 2 min. Cells were washed twice with the washing buffer (20% formamide, 2X SSC, 0.25% Triton X) for 30 min at 37 C. The secondary probes were then hybridized to cells at 37 C for 2 hours and the cells were then washed twice with washing buffer for 30 min at 37 C. The primary and secondary probes are "clamped" in place through a click reaction (CuSO4 75 uM, BTTAA 150 uM, Sodium Ascorbate 2.5 mM in 2X SSC) for 20 min at 37 C. The cells were then washed twice in the washing buffer at 37C for 30 min each wash. Then, the cells were hybridized with the hybridization buffer with tertiary probes for 2 hours at 37C. We complete 6 cycles of hybridization for all our experiments. After the final washes, cells were incubated at 37 C with 100mM DAPI for 20 min, washed twice with PBS, resuspended in the anti-fade buffer, and spun onto a 1.5 coverslip using a cytospin cytocentrifuge (Thermo Scientific), mounted onto a glass slide, sealed with a sealant, and stored at 4C.

Imaging. All images were taken within 72 hours of mounting of the slides. All images were captured by a 63X oil-immersion inverted wide field scope (Leica DMi8) with customized stage, camera (Andor Zyla 5.5) and filter sets (Chroma VCGR-SPX-P01). Automated image acquisition script was achieved through a custom imaging system developed in the previous publication [188]. We acquired z-stacks (1.2 μm between stacks) of stained cells.

Image Processing. Once the images were collected, we perform a maximum z-projection to reduce the z-stacks into 2D images. We then used CellProfiler 3.0 for segmenting cells [189, 190]. Briefly, trans images were first preprocessed to enhance the edges of the cells using the Prewitt algorithm. Then nuclei were identified with global minimum cross-entropy thresholding from DAPI images. Relying on the size and location of the nucleus, cell boundaries were segmented with a Watershed Algorithm. Once the cells were segmented, we then used the rajlabimagetools to quantify RNA FISH spots (https://github.com/arjunrajlaboratory/rajlabimagetools/). At this step we manually inspected each segmented cell and removed poorly segmented cells and adjusted the threshold for signal detection.

Gillespie Simulation for the ON/OFF Model. To investigate possible explanations for the observed differences in mean-independent noise associated between different chromosomal environments, we employed a two-state stochastic model for extrapolating the transcriptional dynamics. A parameter sweep of a stochastic model is computationally expensive, but makes the least assumptions of the relationships among the kinetic parameters. To set up the simulation using the previously established framework, we first constructed the Chemical Master Equation (CME) for the ON/OFF model [2, 33, 180, 191], and by convention, we separate the Chemical Master Equation into two equation:

$$\frac{dP_0(m)}{dt} = K_{off}P_1(m) + K_d(m+1)P_0(m+1) - (mK_d + K_{on})P_0(m) \quad \text{(A.1)}$$

$$\frac{dP_1(m)}{dt} = K_{on}P_0(m) + K_d(m+1)P_1(m) + K_mP_1(m-1) - (K_m + mK_d + K_{off})P_1(m) \quad \text{(A.2)}$$

After establishing the CME for the ON/OFF model, we used the Gillespie Algorithm to simulate 8.8 million sets of parameters (Gillespie, 1976). We chose the estimated physiological range of the rates

of activation, transcription, and mRNA degradation (Kon: $0.00001 - 50s^{-1}$, Koff: $0.00001 - 50s^{-1}$, Km: $0.002 - 1s^{-1}$, $0.0002 - 0.003s^{-1}$)[172, 192, 193]. Kon and Koff were first broken down into intervals ($0.00001 - 0.001s^{-1}$ and $0.001 - 50s^{-1}$). For the first interval, 37 values for each parameter were chosen logarithmically to ensure finer sampling for the smaller values. For the second interval, 50 values for each parameter were chosen logarithmically. Each set of parameters was simulated for 1X103 individual trajectories till reaching stationary distribution, and the final distribution was recorded. The simulation was done on the Washington University High Throughput Computing Facility at Center for Genome Sciences (https://htcf.wustl.edu) with 200 cores with 2 Gb of RAM per core.

Fitting Experimental Data to Simulated Results. To fit our experimentally measured mRNA distribution with the ON/OFF model, Kolmogorov-Smirnov test was employed to determine the closest simulated distributions to the experimentally measured distribution. All sets of parameters that accept the null hypothesis for the K-S test ($p > 1x10^{-2}$) are reported for subsequent analyses.

Epigenome Data Analysis. For the epigenetic data analyses at different integration sites, we considered the genomic region of interest as the 2500 bps flanking the integrated site. We then downloaded various K562 epigenome datasets. For H3K9ac, H3K27ac, H3K9me3, H3K4me3, H3K27me3, Dnase-seq we used pyranges [194] module to overlay the signal with the integration sites. The ChromHMM and Segway combined segmentations for K562 cells were downloaded from UCSC genome browser and overlaid with the integrated sites.

TRIP Data Analysis. We downloaded TRIP data for mouse embryonic stem cells (mES cells) from Akhtar et al. [50] and the ChromHMM segmentation for mES cells from Pintachuda et al. [195]. We lifted the TRIP data from mm9 to mm10 using the UCSC Liftover Tool [196]. We then overlay the TRIP expression data with the ChromHMM segmentation using pyranges [194].

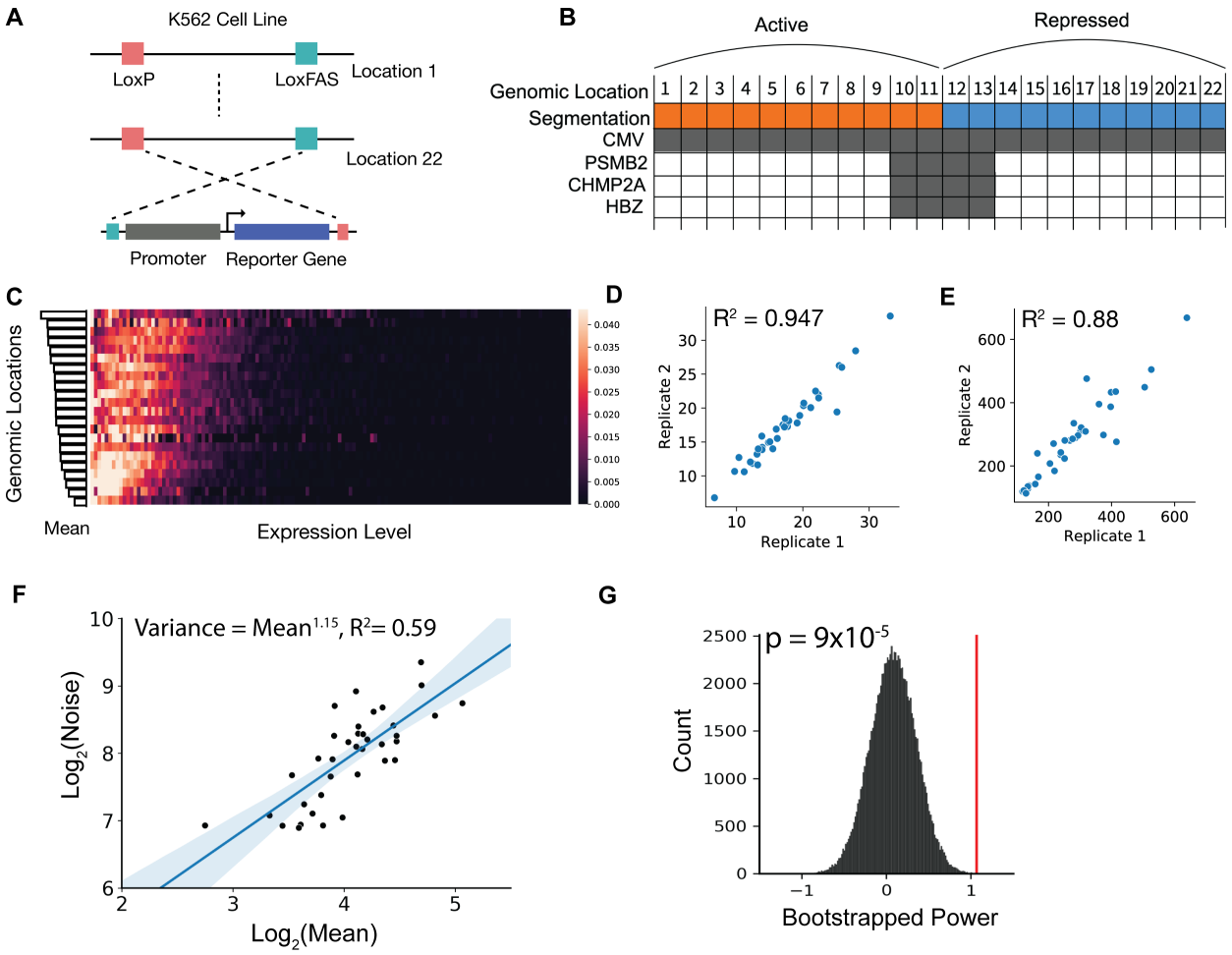

Figure A.1: **Experimental design for studying the cell-to-cell variability associated with different genomic locations**(A) Strategy for integrating the same reporter cassette into mapped genomic locations. In each K562 cell line, a construct containing a pair of LoxP and LoxFAS sites is integrated and mapped. The sequence between the pair of lox sites contains a reporter cassette of eGFP driven by a CMV promoter. Three other reporter cassettes containing PSMB2, HBZ, and CHMP2A promoters replace the CMV promoter through cre recombination. (B) Schematics for selecting K562 cell lines. 8 active locations and 14 repressed locations were selected based on the ChromHMM+Segway combine segmentation (see Methods). (C) Heatmap of single-cell mRNA expression of 22 cell lines with CMV promoter, ordered from highest mean expression level to lowest. Each position in the matrix represents the mRNA counts of a single cell, and the color represents the percentage of cells. (D, E) Scatter plots showing the reproducibility derived from two replicates for mean (D) and noise (E) of the measured single-cell expression distributions. (F) Linear regression of mean against noise in the log scale revealed a power-law dependence. Mean expression level explains about 59% of the observed noise, with the power of 1.15. (G) Histogram of the bootstrap of the power of the power-law relationship between mean and noise. The red line shows the fitted power from experimental data.

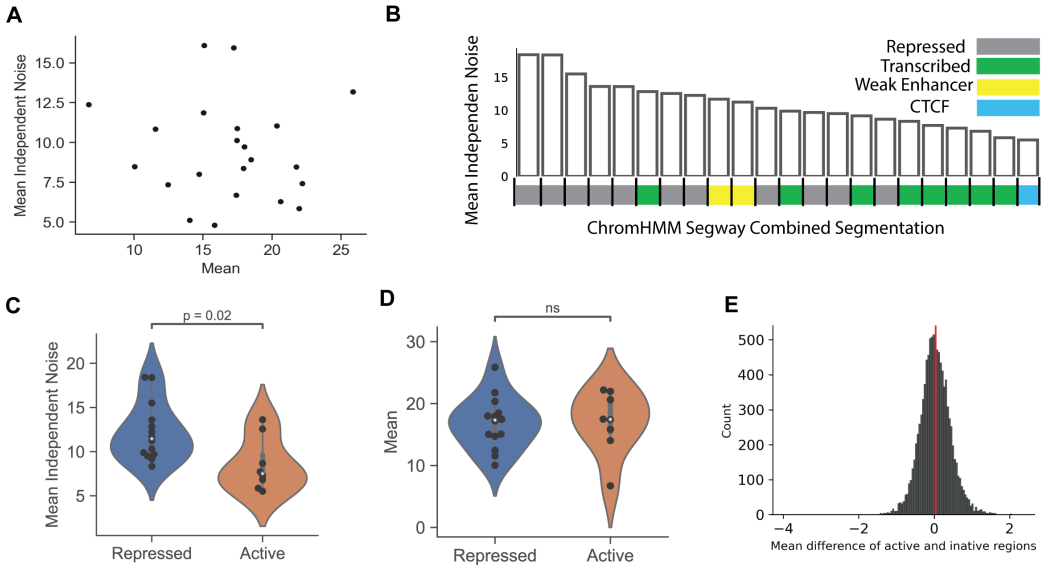

Figure A.2: **Active genomic locations have lower expression noise.** (A) Expression mean and MIN are plotted for each genomic location expressing eGFP driven by the CMV promoter. (B) Residual noises from the power-law fitting are plotted with the order from highest residual to the lowest. Each bar corresponds with the combined ChromHMM+Segway segmentation annotations. (C) Violin plot for MIN at active and repressed locations. (D) Violin plot for mean at repressed and active locations. (E) Histogram of trip mean expression differences from Ahktar et al 2013 for 105 times using the number of observations from our experimental data (n=22), red vertical line indicates the actual mean difference observed from our data.

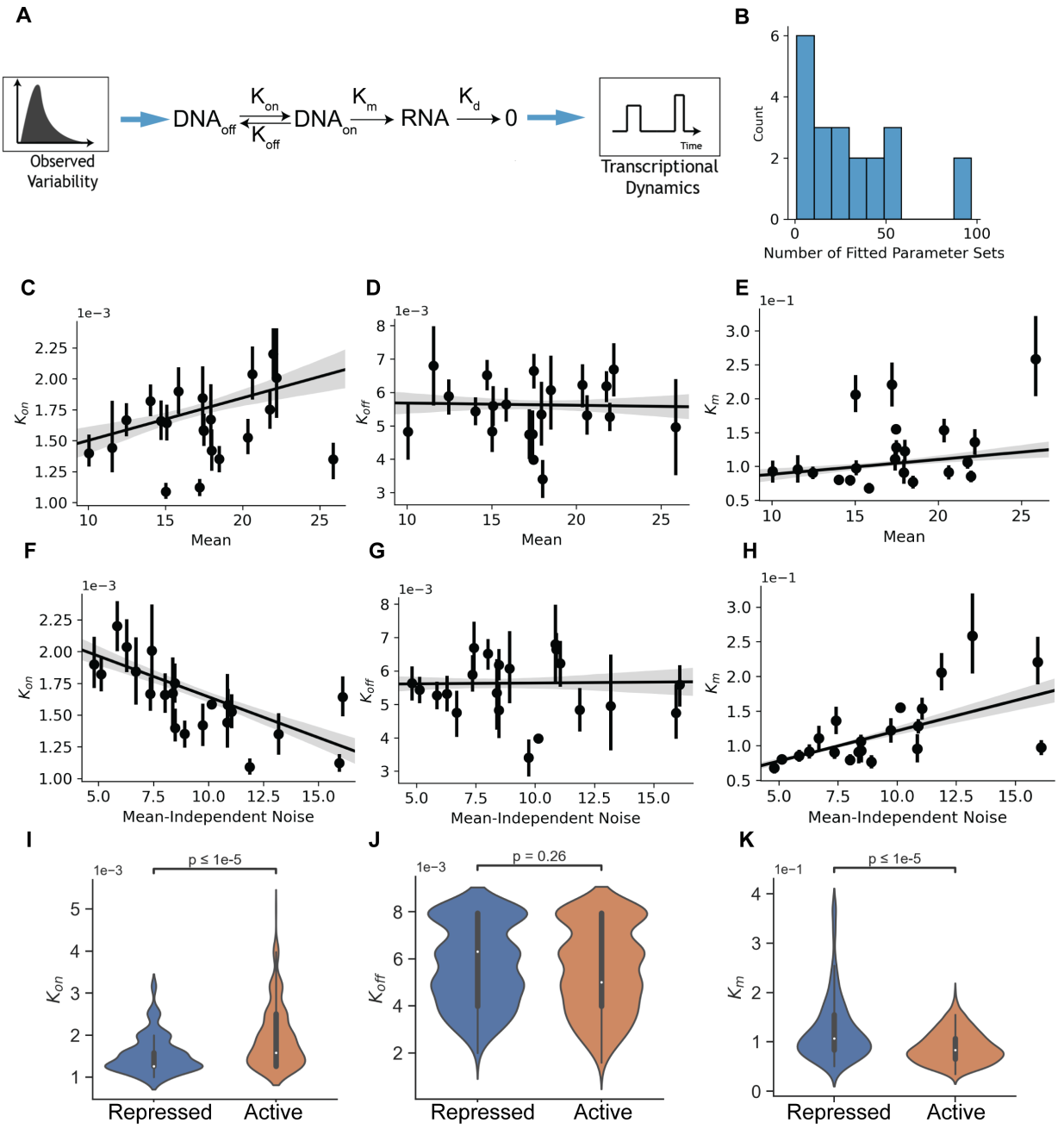

Figure A.3: **The ON/OFF model reveals the dynamics that drive different MIN at active and repressed locations.** (A) Schematics for stochastic simulation for the ON/OFF model. (B) Number of fitted sets of parameters for each experimentally measured distribution (n = 34, median = 23). (C–E) Scatter plots of mean against Kon (C) (R2 = 0.03, p = $1.7\text{x}10^{-6}$), Koff (D) (R2 = 0.0001, p = 0.73), and Km (E) (R2 = 0.02, p = 0.00015),). Error bar indicates all fitted values from the stochastic simulation. Shadow shows the 95% C.I.of the linear fit. (F–H) Scatter plots of MIN against Kon (F) (R2 = 0.1, p = $5.2\text{x}10^{-17}$), Koff (G) (R2 < $1\text{x}10^{-5}$, p = 0.8), and Km (H) (R2 = 0.3, p = $3.38\text{x}10^{-52}$). Error bar indicates all fitted values from the stochastic simulation. Shadow shows the 95% C.I. of the linear fit. (I-K) Violin plots of all fitted Kon (I), Koff (J), and Km (K) at active and repressed locations.

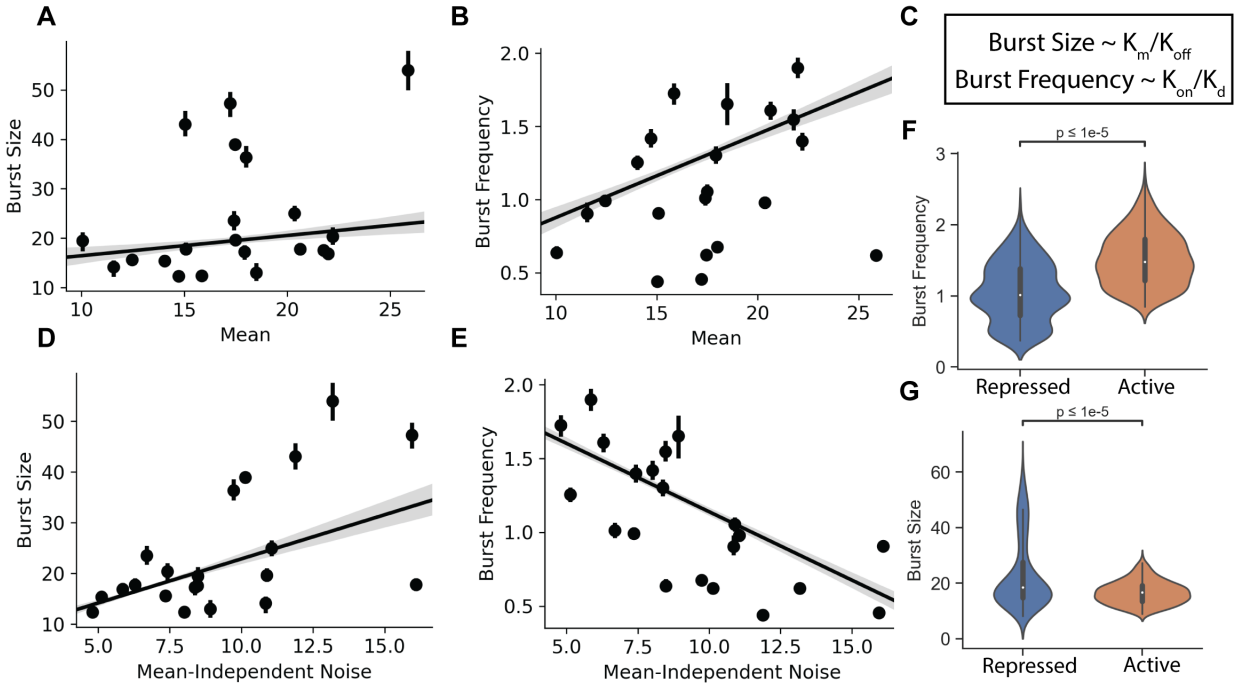

Figure A.4: **The ON/OFF model reveals the dynamics that drive different MIN at active and repressed locations.** (A) Schematics for stochastic simulation for the two-state ON/OFF model. (B) Number of fitted sets of parameters for each experimentally measured distribution (n = 34, median = 23). (C–E) Scatter plots of mean against Kon (C) (R2 = 0.03, p = $1.7\text{x}10^{-6}$), Koff (D) (R2 = 0.0001, p = 0.73), and Km (E) (R2 = 0.02, p = 0.00015),). Error bar indicates all fitted values from the stochastic simulation. Shadow shows the 95% C.I. of the linear fit. (F–H) Scatter plots of MIN against Kon (F) (R2 = 0.1, p = $5.2\text{x}10^{-17}$), Koff (G) (R2 < $1\text{x}10^{-5}$, p = 0.8), and Km (H) (R2 = 0.3, p = $3.38\text{x}10^{-52}$). Error bar indicates all fitted values from the stochastic simulation. Shadow shows the 95% C.I. of the linear fit. (I-K) Violin plots of all fitted Kon (I), Koff (J), and Km (K) at active and repressed locations.

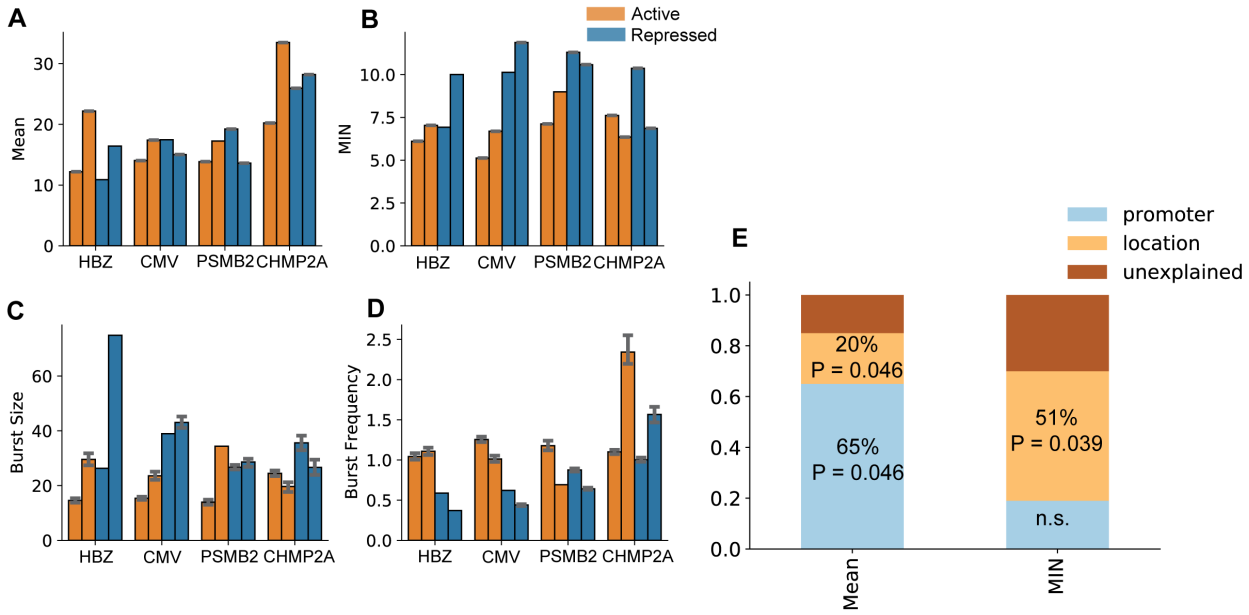

Figure A.5: **Active transcriptional locations have higher burst frequency but lower burst size.** (A,B) Bar plots showing the mean and MIN of different promoters at two active and two initiative genomic locations. (C,D) Bar plots showing the burst size (C) and burst frequency (D) of different promoters at two active and two initiative genomic locations.(E) Two-way ANOVA analysis for promoter identity and genomic location contribution to expression mean and MIN.

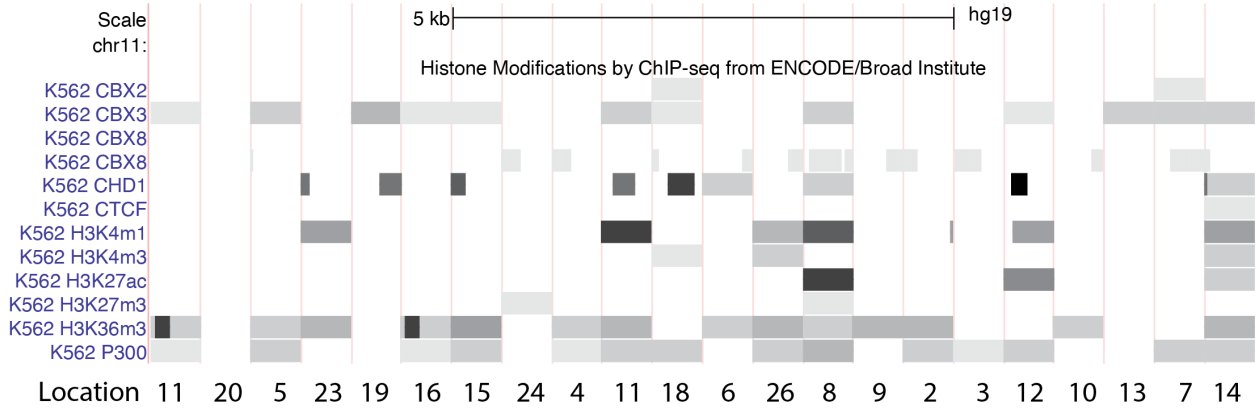

Figure A.6: **Genomic locations of landing pads have different epigenetic signatures.**

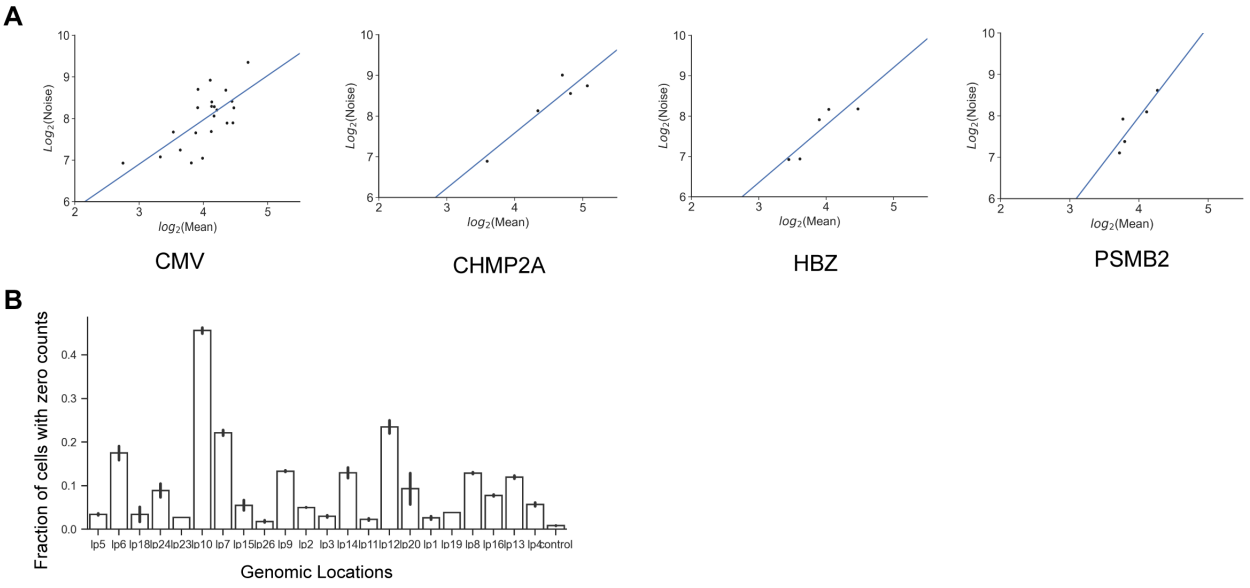

Figure A.7: **The Power-law relationship for different promoters at different genomic locations are largely the same and suggests a slow dynamics.**(A) Power-law relations for expression mean and noise for different promoters at different genomic locations in K562 cells. (B) The percentage of cells with no mRNA molecule labeled, compared to ACTB intron control.

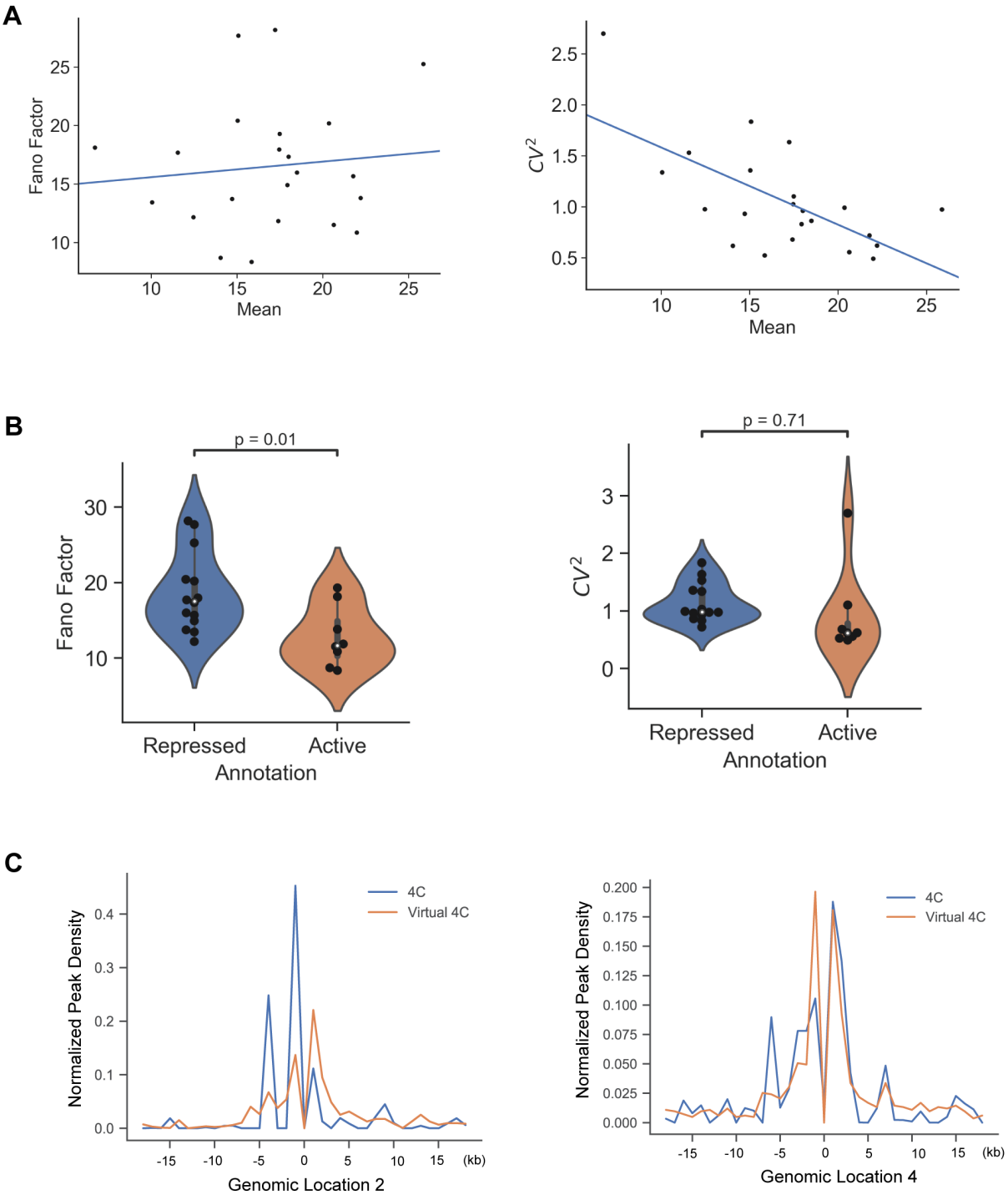

Figure A.8: **Specific noise metric does not change our observation.** (A) Power-law relations for expression mean and Fano factor and CV2 for CMV promoters at different genomic locations in K562 cells. (B) Box plot showing Fano Factor and CV2 as noise metric for active and repressed locations. (C) Comparison of peak density between 4C (blue line) and Hi-C derived virtual 4C (orange line) for Location 2 and Location 4. Each peak represents the normalized read density centered on the insertion site.

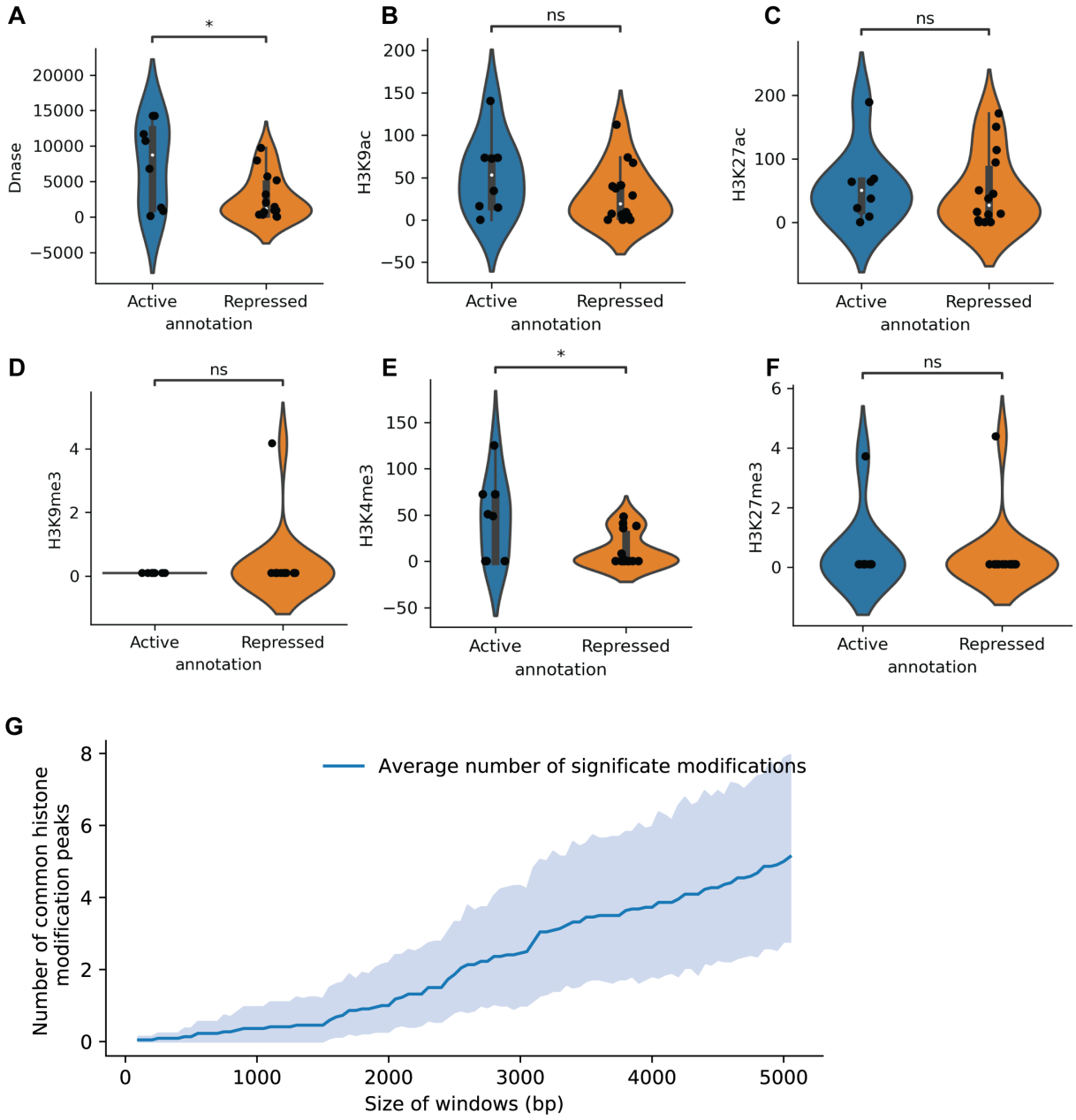

Figure A.9: **Epigenetic Analysis for different genomic locations.** ChIP-seq datasets of histone modification are downloaded from the ENCODE and are correlated with active and repressed regions. (A-C). Violin plot showing active histone modifications peaks 5kb surrounding the integration site at active and repressed locations. (D-F). Violin plot showing repressive histone modifications peaks 5kb surrounding the integration site at active and repressed locations. (G) Number of histone ChIP-seq peaks (H3K9ac, H3K27ac, H3K9me3, H3K4me3, H3K27me3) and Dnase-seq peaks within a given window centered on the integration site for 22 genomic locations in K562 cells Solid blue line: average number of peaks, blue shade: standard error).

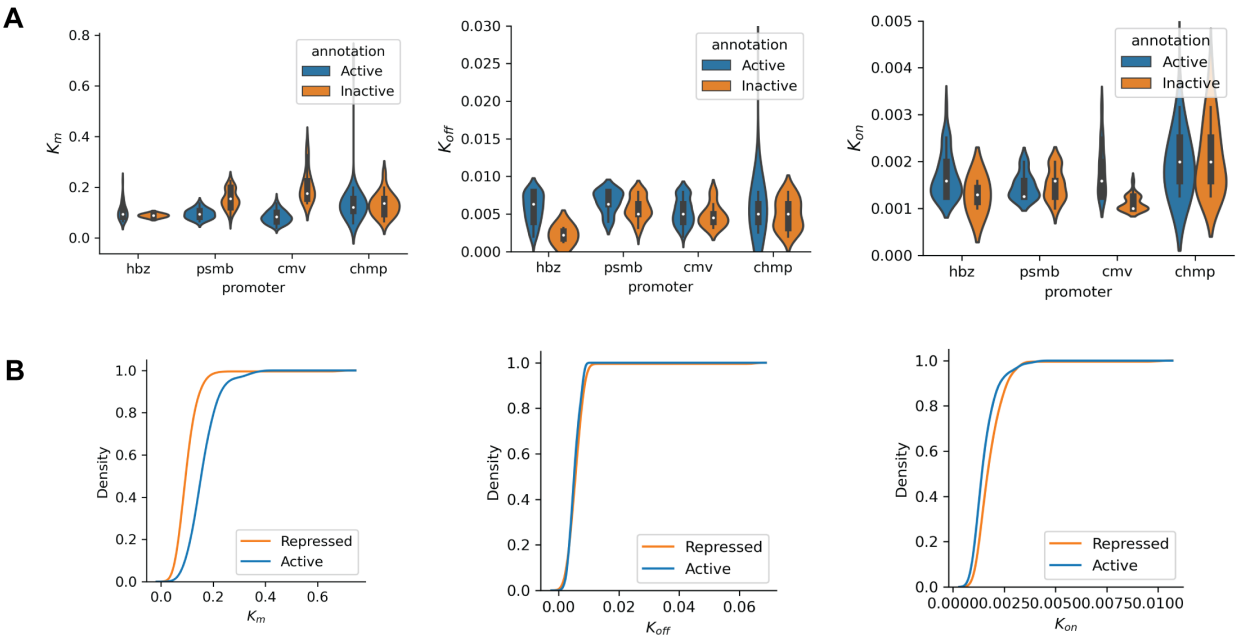

Figure A.10: **Fitted rates for different promoters and different genomic locations.** (A) Different rates (Km, Koff, and Kon) of the ON/OFF model for different promoters at active and repressed locations. (B) Cumulative distribution function for different rates at active and repressed locations

References

1. Bianconi, E. *et al.* An estimation of the number of cells in the human body. en. *Ann. Hum. Biol.* **40,** 463–471 (2013) (cited on p. 1).
2. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. en. *PLoS Biol.* **4,** e309 (2006) (cited on pp. 1, 33, 55, 73, 92, 96, 99, 105).
3. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. en. *Science* **297,** 1183–1186 (2002) (cited on pp. 1, 55, 65, 73, 92).
4. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. en. *J. Mol. Biol.* **3,** 318–356 (1961) (cited on p. 2).
5. Roeder, R. G. & Rutter, W. J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. en. *Nature* **224,** 234–237 (1969) (cited on p. 2).
6. Noe Gonzalez, M., Blears, D. & Svejstrup, J. Q. Causes and consequences of RNA polymerase II stalling during transcript elongation. en. *Nat. Rev. Mol. Cell Biol.* **22,** 3–21 (2021) (cited on p. 2).
7. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. en. *Cell* **169,** 1187–1200 (2017) (cited on p. 2).
8. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. en. *Cell* **136,** 731–745 (2009) (cited on p. 2).
9. Walsh, C. T., Garneau-Tsodikova, S. & Gatto Jr, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. en. *Angew. Chem. Int. Ed Engl.* **44,** 7342–7372 (2005) (cited on p. 2).
10. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. en. *Cell* **126,** 663–676 (2006) (cited on p. 2).
11. Crick, F. H. On protein synthesis. en. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958) (cited on p. 3).
12. Herr, W. The SV40 enhancer: Transcriptional regulation through a hierarchy of combinatorial interactions. *Semin. Virol.* **4,** 3–13 (1993) (cited on pp. 4, 7).
13. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. en. *Cell* **27,** 299–308 (1981) (cited on pp. 4, 7).

14. Zenke, M., Grundström, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A. & Chambon, P. Multiple sequence motifs are involved in SV40 enhancer function. en. *EMBO J.* **5,** 387–397 (1986) (cited on p. 4).

15. Ondek, B., Gloss, L. & Herr, W. The SV40 enhancer contains two distinct levels of organization. en. *Nature* **333,** 40–45 (1988) (cited on p. 4).

16. Herr, W. & Clarke, J. The SV40 enhancer is composed of multiple functional elements that can compensate for one another. en. *Cell* **45,** 461–470 (1986) (cited on pp. 4, 7).

17. Cantor, G. Uber eine elementare Frage der Mannigfaltigkeitslehre. *Jahresber. Dtsch. Math.-Ver.* **1,** 75–78 (1891) (cited on p. 4).

18. Cohen, P. J. The independence of the continuum hypothesis. en. *Proc. Natl. Acad. Sci. U. S. A.* **50,** 1143–1148 (1963) (cited on p. 4).

19. De Broglie, L. Recherches sur la théorie des Quanta. *Ann. Phys.* **10,** 22–128 (1925) (cited on p. 4).

20. Holland, J. H. Complex Adaptive Systems. *Daedalus* **121,** 17–30 (1992) (cited on p. 4).

21. Stormo, G. D. DNA binding sites: representation and discovery. en. *Bioinformatics* **16,** 16–23 (2000) (cited on p. 4).

22. Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. en. *Nat. Genet.* **38,** 1289–1297 (2006) (cited on p. 4).

23. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. en. *Nature* **489,** 57–74 (2012) (cited on pp. 5, 97).

24. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. en. *Nature* **447,** 799–816 (2007) (cited on pp. 5, 17).

25. Ren, B. *et al. Genome-Wide Location and Function of DNA Binding Proteins* 2000 (cited on p. 5).

26. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. en. *Science* **316,** 1497–1502 (2007) (cited on p. 5).

27. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. High-resolution profiling of histone methylations in the human genome. en. *Cell* **129,** 823–837 (2007) (cited on p. 5).

28. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. en. *Nat. Methods* **4,** 651–657 (2007) (cited on p. 5).

29. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. en. *Elife* **6** (2017) (cited on p. 5).

30. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. en. *Nucleic Acids Res.* **41,** W544–56 (2013) (cited on p. 5).

31. De Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of enhancers. *bioRxiv* (2021) (cited on p. 5).

32. Fiore, C. & Cohen, B. A. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. en. *Genome Res.* **26,** 778–786 (2016) (cited on pp. 5, 32).

33. Sherman, M. S., Lorenz, K., Lanier, M. H. & Cohen, B. A. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. en. *Cell Syst* **1,** 315–325 (2015) (cited on pp. 5, 33, 65, 105).

34. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. en. *Genome Res.* **24,** 1595–1602 (2014) (cited on pp. 6, 7, 17).

35. Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. en. *Nat. Biotechnol.* (2018) (cited on pp. 6, 93, 94, 103).

36. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. en. *Science* **339,** 1074–1077 (2013) (cited on pp. 6, 17).

37. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19498–19503 (2012) (cited on pp. 6, 17, 29).

38. Friedman, R. Z., Granas, D. M., Myers, C. A., Corbo, J. C., Cohen, B. A. & White, M. A. Information content differentiates enhancers from silencers in mouse photoreceptors. en. *Elife* **10** (2021) (cited on p. 6).

39. White, M. A., Kwasnieski, J. C., Myers, C. A., Shen, S. Q., Corbo, J. C. & Cohen, B. A. A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. en. *Cell Rep.* **17,** 1247–1254 (2016) (cited on pp. 6, 17, 29, 42).

40. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. en. *Nat. Rev. Mol. Cell Biol.* **19,** 621–637 (2018) (cited on pp. 6, 18).

41. De Pater, B. S., de Kam, R. J., Hoge, J. H. & Schilperoort, R. A. Effects of mutations in the TATA box region of the Agrobacterium T-cyt gene on its transcription in plant tissues. en. *Nucleic Acids Res.* **15,** 8283–8292 (1987) (cited on p. 6).

42. Grassi, E., Zapparoli, E., Molineris, I. & Provero, P. Total Binding Affinity Profiles of Regulatory Regions Predict Transcription Factor Binding and Gene Expression in Human Cells. en. *PLoS One* **10,** e0143627 (2015) (cited on p. 6).

43. Liu, J., Shively, C. A. & Mitra, R. D. Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. en. *Nucleic Acids Res.* **48,** e50 (2020) (cited on p. 6).

44. Chen, C.-H. *et al.* Determinants of transcription factor regulatory range. en. *Nat. Commun.* **11,** 2472 (2020) (cited on p. 6).

45. Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. en. *Nature* **457,** 215–218 (2009) (cited on p. 6).

46. King, D. M., Hong, C. K. Y., Shepherdson, J. L., Granas, D. M., Maricque, B. B. & Cohen, B. A. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. en. *Elife* **9** (2020) (cited on p. 6).

47. Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M. & Weng, Z. Functional analysis of transcription factor binding sites in human promoters. en. *Genome Biol.* **13,** R50 (2012) (cited on p. 7).

48. Bergman, D. T. *et al. Compatibility logic of human enhancer and promoter sequences* en. 2021 (cited on pp. 7, 8).

49. Martinez-Ara, M., Comoglio, F., van Arensbergen, J. & van Steensel, B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol. Cell* (2022) (cited on p. 7).

50. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. en. *Cell* **154,** 914–927 (2013) (cited on pp. 8, 57, 62, 98, 106).

51. Hong, C. K. Y. & Cohen, B. A. *Genomic environments scale the activities of diverse core promoters* en. 2021 (cited on pp. 8, 18).

52. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. en. *Science* **361,** 1341–1345 (2018) (cited on p. 8).

53. Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A. & Blobel, G. A. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. en. *Mol. Cell* **62,** 237–247 (2016) (cited on p. 8).

54. Grubert, F. *et al.* Landscape of cohesin-mediated chromatin loops in the human genome. en. *Nature* **583,** 737–743 (2020) (cited on p. 8).

55. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. en. *Cell* **159,** 1665–1680 (2014) (cited on pp. 8, 97).

56. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E6456–65 (2015) (cited on p. 8).

57. Choudhary, M. N., Friedman, R. Z., Wang, J. T., Jang, H. S., Zhuo, X. & Wang, T. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. en. *Genome Biol.* **21,** 16 (2020) (cited on p. 9).

58. Schirm, S., Jiricny, J. & Schaffner, W. *The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity* 1987 (cited on p. 9).

59. Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T. & Bialek, W. Positional information, in bits. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 16301–16308 (2013) (cited on p. 9).

60. Tkacik, G., Callan Jr, C. G. & Bialek, W. Information flow and optimization in transcriptional regulation. en. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 12265–12270 (2008) (cited on p. 9).

61. Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E. S. & Lomvardas, S. Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. en. *Elife* **6** (2017) (cited on pp. 9, 10).

62. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. en. *Nature* **598,** 129–136 (2021) (cited on p. 9).

63. Murphy, D. P., Hughes, A. E., Lawrence, K. A., Myers, C. A. & Corbo, J. C. Cis-regulatory basis of sister cell type divergence in the vertebrate retina. en. *Elife* **8** (2019) (cited on pp. 9, 29, 31).

64. Peng, W., Song, R. & Acar, M. Noise reduction facilitated by dosage compensation in gene networks. en. *Nat. Commun.* **7,** 12959 (2016) (cited on pp. 10, 65).

65. Grün, D. Revealing dynamics of gene expression variability in cell state space. en. *Nat. Methods* **17,** 45–49 (2020) (cited on p. 10).

66. Wolf, L., Silander, O. K. & van Nimwegen, E. Expression noise facilitates the evolution of gene regulation. en. *Elife* **4** (2015) (cited on pp. 10, 40).

67. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. en. *Nature* **453,** 544–547 (2008) (cited on pp. 10, 33, 55, 92).

68. Itzkovitz, S., Blat, I. C., Jacks, T., Clevers, H. & van Oudenaarden, A. *Optimality in the Development of Intestinal Crypts* 2012 (cited on p. 10).

69. Chess, A., Simon, I., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. en. *Cell* **78,** 823–834 (1994) (cited on pp. 10, 92).

70. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. en. *Nature* **546,** 431–435 (2017) (cited on pp. 10, 25, 33, 55, 92).

71. 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. en. *PLoS One* **9,** e102202 (2014) (cited on pp. 11, 60, 96).

72. Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. & Barkai, N. Noise in protein expression scales with natural protein abundance. en. *Nat. Genet.* **38,** 636–643 (2006) (cited on pp. 11, 60, 96, 99).

73. So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R. & Golding, I. General properties of transcriptional time series in Escherichia coli. en. *Nat. Genet.* **43,** 554–560 (2011) (cited on p. 11).

74. Bonny, A. R., Fonseca, J. P., Park, J. E. & El-Samad, H. Orthogonal control of mean and variability of endogenous genes in a human cell line. en. *Nat. Commun.* **12,** 292 (2021) (cited on pp. 11, 73).

75. Larsson, A. J. M. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* (2019) (cited on p. 11).

76. Carey, L. B., van Dijk, D., Sloot, P. M. A., Kaandorp, J. A. & Segal, E. Promoter sequence determines the relationship between expression level and noise. en. *PLoS Biol.* **11,** e1001528 (2013) (cited on p. 11).

77. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. en. *Cell* **161,** 1202–1214 (2015) (cited on p. 11).

78. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. en. *Nature* **566,** 496–502 (2019) (cited on pp. 11, 73).

79. Moffitt, J. R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C. & Zhuang, X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 14456–14461 (2016) (cited on p. 11).

80. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. en. *Science* **297,** 1183–1186 (2002) (cited on pp. 11, 33).

81. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. en. *Genome Res.* **22,** 1748–1759 (2012) (cited on pp. 13, 17).

82. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. en. *Science* **337,** 1190–1195 (2012) (cited on pp. 13, 17).

83. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. en. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (2009) (cited on pp. 13, 17).

84. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. en. *Nat. Genet.* **42,** 565–569 (2010) (cited on p. 13).

85. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. en. *Cell* **169,** 1177–1186 (2017) (cited on p. 13).

86. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. en. *Science* **373,** 1464–1468 (2021) (cited on p. 13).

87. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. en. *PLoS Genet.* **8,** e1002637 (2012) (cited on p. 17).

88. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. en. *Am. J. Hum. Genet.* **99,** 139–153 (2016) (cited on p. 17).

89. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. en. *Hum. Mol. Genet.* **25,** R157–R165 (2016) (cited on p. 17).

90. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. en. *Hum. Mol. Genet.* **24,** R102–10 (2015) (cited on p. 17).

91. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. en. *Science* (2019) (cited on p. 17).

92. Aygün, N. *et al.* Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. en. *Am. J. Hum. Genet.* **108,** 1647–1668 (2021) (cited on p. 17).

93. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. en. *Nat. Rev. Genet.* **12,** 283–293 (2011) (cited on p. 17).

94. Ireland, W. T. *et al.* Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. en. *Elife* **9** (2020) (cited on p. 17).

95. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. en. *Nat. Biotechnol.* **30,** 265–270 (2012) (cited on p. 17).

96. Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z. & Segal, E. *Probing the effect of promoters on noise in gene expression using thousands of designed sequences* 2014 (cited on pp. 17, 92).

97. Kinney, J. B., Murugan, A., Callan Jr, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. en. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9158–9163 (2010) (cited on p. 17).

98. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. en. *Nat. Biotechnol.* **30,** 271–277 (2012) (cited on p. 17).

99. Chaudhari, H. G. & Cohen, B. A. Local sequence features that influence AP-1 cis-regulatory activity. en. *Genome Res.* **28,** 171–181 (2018) (cited on p. 17).

100. Hughes, A. E. O., Myers, C. A. & Corbo, J. C. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. en. *Genome Res.* **28,** 1520–1531 (2018) (cited on pp. 17, 42).

101. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. en. *Cell* **165,** 1519–1529 (2016) (cited on p. 17).

102. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biol.* **15,** 550 (2014) (cited on p. 23).

103. Amabile, G. *et al.* Dissecting the role of aberrant DNA methylation in human leukaemia. en. *Nat. Commun.* **6,** 7091 (2015) (cited on p. 23).

104. Juven-Gershon, T., Cheng, S. & Kadonaga, J. T. Rational design of a super core promoter that enhances gene expression. en. *Nat. Methods* **3,** 917–922 (2006) (cited on p. 23).

105. Shao, H., Revach, M., Moshonov, S., Tzuman, Y., Gazit, K., Albeck, S., Unger, T. & Dikstein, R. Core promoter binding by histone-like TAF complexes. en. *Mol. Cell. Biol.* **25,** 206–219 (2005) (cited on p. 23).

106. Moudgil, A. *et al.* Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. en. *Cell* **182,** 992–1008.e21 (2020) (cited on pp. 25, 68).

107. Litzenburger, U. M., Buenrostro, J. D., Wu, B., Shen, Y., Sheffield, N. C., Kathiria, A., Greenleaf, W. J. & Chang, H. Y. Single-cell epigenomic variability reveals functional cancer heterogeneity. en. *Genome Biol.* **18,** 15 (2017) (cited on p. 25).

108. Min, M. & Spencer, S. L. Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. en. *PLoS Biol.* **17,** e3000178 (2019) (cited on p. 25).

109. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. en. *Nat. Med.* **3,** 730–737 (1997) (cited on p. 25).

110. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. en. *Nat. Biotechnol.* **25,** 1315–1321 (2007) (cited on p. 25).

111. Zagozewski, J. L., Zhang, Q., Pinto, V. I., Wigle, J. T. & Eisenstat, D. D. The role of homeobox genes in retinal development and disease. en. *Dev. Biol.* **393,** 195–208 (2014) (cited on p. 31).

112. Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O. & Stark, A. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. en. *Nature* **518,** 556–559 (2014) (cited on pp. 32, 42).

113. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. en. *Science* **369** (2020) (cited on p. 32).

114. Emert, B. L., Cote, C. J., Torre, E. A., Dardani, I. P., Jiang, C. L., Jain, N., Shaffer, S. M. & Raj, A. Variability within rare cell states enables multiple paths toward drug resistance. *Nat. Biotechnol.* (2021) (cited on pp. 33, 55).

115. Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by underlying cell state. en. *Mol. Syst. Biol.* **16,** e9146 (2020) (cited on pp. 33, 65).

116. Hilfinger, A. & Paulsson, J. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. en. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12167–12172 (2011) (cited on pp. 33, 65).

117. Fu, A. Q. & Pachter, L. Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. en. *Stat. Appl. Genet. Mol. Biol.* **15,** 447–471 (2016) (cited on pp. 33, 43, 66).

118. Abud, H. E., Lock, P. & Heath, J. K. Efficient gene transfer into the epithelial cell layer of embryonic mouse intestine using low-voltage electroporation. *Gastroenterology* **126,** 1779–1787 (2004) (cited on pp. 34, 88).

119. Gupta, I. R., Lapointe, M. & Yu, O. H. Morphogenesis during mouse embryonic kidney explant culture. en. *Kidney Int.* **63,** 365–376 (2003) (cited on pp. 34, 88).

120. Dal Maschio, M. *et al.* High-performance and site-directed in utero electroporation by a triple-electrode probe. en. *Nat. Commun.* **3,** 960 (2012) (cited on pp. 34, 88).

121. Kinoshita-Kawada, M. *et al.* Explant Culture of the Embryonic Mouse Spinal Cord and Gene Transfer by ex vivo Electroporation. en. *Bio Protoc* **9,** e3373 (2019) (cited on pp. 34, 88).

122. Byrne, L. C., Day, T. P., Visel, M., Strazzeri, J. A., Fortuny, C., Dalkara, D., Merigan, W. H., Schaffer, D. V. & Flannery, J. G. In vivo-directed evolution of adeno-associated virus in the primate retina. en. *JCI Insight* **5** (2020) (cited on pp. 34, 88).

123. Chan, Y. K. *et al.* Engineering adeno-associated viral vectors to evade innate immune and inflammatory responses. en. *Sci. Transl. Med.* **13** (2021) (cited on pp. 34, 88).

124. Wang, D., Tai, P. W. L. & Gao, G. Adeno-associated virus vector as a platform for gene therapy delivery. en. *Nat. Rev. Drug Discov.* **18,** 358–378 (2019) (cited on pp. 34, 88).

125. Bryant, D. H., Bashir, A., Sinai, S., Jain, N. K., Ogden, P. J., Riley, P. F., Church, G. M., Colwell, L. J. & Kelsic, E. D. Deep diversification of an AAV capsid protein by machine learning. en. *Nat. Biotechnol.* **39,** 691–696 (2021) (cited on pp. 34, 88).

126. Cohen, R. N., van der Aa, M. A. E. M., Macaraeg, N., Lee, A. P. & Szoka Jr, F. C. Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. en. *J. Control. Release* **135,** 166–174 (2009) (cited on p. 38).

127. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. en. *Sci. Rep.* **9,** 5233 (2019) (cited on p. 41).

128. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. en. *Science* **352,** 189–196 (2016) (cited on p. 41).

129. Desai, R. V. *et al.* A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. en. *Science* **373** (2021) (cited on pp. 55, 73).

130. Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J. & Martinez Arias, A. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. en. *PLoS Biol.* **7,** e1000149 (2009) (cited on p. 55).

131. Abranches, E., Guedes, A. M. V., Moravec, M., Maamar, H., Svoboda, P., Raj, A. & Henrique, D. Stochastic NANOG fluctuations allow mouse embryonic stem cells to explore pluripotency. en. *Development* **141,** 2770–2779 (2014) (cited on p. 55).

132. Topolewski, P., Zakrzewska, K. E., Walczak, J., Nienałtowski, K., Müller-Newen, G., Singh, A. & Komorowski, M. Phenotypic variability, not noise, accounts for most of the cell-to-cell heterogeneity in IFN-γ and oncostatin M signaling responses. en. *Sci. Signal.* **15,** eabd9303 (2022) (cited on p. 55).

133. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. en. *Nature* **459,** 428–432 (2009) (cited on p. 55).

134. 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. en. *Cell* **122,** 169–182 (2005) (cited on p. 55).

135. Yang, C., Tian, C., Hoffman, T. E., Jacobsen, N. K. & Spencer, S. L. Melanoma subpopulations that rapidly escape MAPK pathway inhibition incur DNA damage and rely on stress signalling. en. *Nat. Commun.* **12,** 1747 (2021) (cited on p. 55).

136. Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. en. *Science* **342,** 1188–1193 (2013) (cited on p. 55).

137. Das Neves, R. P., Jones, N. S., Andreu, L., Gupta, R., Enver, T. & Iborra, F. J. Connecting variability in global transcription rate to mitochondrial variability. en. *PLoS Biol.* **8,** e1000560 (2010) (cited on p. 55).

138. Stewart-Ornstein, J., Weissman, J. S. & El-Samad, H. Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. en. *Mol. Cell* **45,** 483–493 (2012) (cited on p. 55).

139. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. en. *Nat. Genet.* **31,** 69–73 (2002) (cited on p. 55).

140. Elgin, S. C. R. & Reuter, G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. en. *Cold Spring Harb. Perspect. Biol.* **5,** a017780 (2013) (cited on p. 55).

141. Faure, A. J., Schmiedel, J. M. & Lehner, B. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Systems* **5,** 471–484.e4 (2017) (cited on pp. 55, 62, 63).

142. Wu, S., Li, K., Li, Y., Zhao, T., Li, T., Yang, Y.-F. & Qian, W. Independent regulation of gene expression level and noise by histone modifications. en. *PLoS Comput. Biol.* **13,** e1005585 (2017) (cited on p. 55).

143. Weinberger, L., Voichek, Y., Tirosh, I., Hornung, G., Amit, I. & Barkai, N. Expression noise and acetylation profiles distinguish HDAC functions. en. *Mol. Cell* **47,** 193–202 (2012) (cited on p. 55).

144. Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L. & Weinberger, L. S. Transcriptional burst frequency and burst size are equally modulated across the human genome. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17454–17459 (2012) (cited on pp. 55, 93, 99).

145. Larson, D. R., Fritzsch, C., Sun, L., Meng, X., Lawrence, D. S. & Singer, R. H. Direct observation of frequency modulated transcription in single cells using light activation. *Elife* **2,** e00750 (2013) (cited on p. 55).

146. Senecal, A., Munsky, B., Proux, F., Ly, N., Braye, F. E., Zimmer, C., Mueller, F. & Darzacq, X. Transcription factors modulate c-Fos transcriptional bursts. en. *Cell Rep.* **8,** 75–83 (2014) (cited on p. 55).

147. Pavani, G. & Amendola, M. Targeted Gene Delivery: Where to Land. en. *Front Genome Ed* **2,** 609650 (2020) (cited on p. 55).

148. Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. en. *Mol. Ther.* **24,** 678–684 (2016) (cited on pp. 55, 72).

149. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. en. *Nucleic Acids Res.* **41,** 827–841 (2013) (cited on pp. 60, 62, 98).

150. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. en. *Nature* **489,** 57–74 (2012) (cited on pp. 62, 97).

151. Ham, L., Brackston, R. D. & Stumpf, M. P. H. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Phys. Rev. Lett.* **124,** 108101 (2020) (cited on p. 65).

152. Ham, L., Jackson, M. & Stumpf, M. P. Pathway dynamics can delineate the sources of transcriptional noise in gene expression. en. *Elife* **10** (2021) (cited on p. 65).

153. Litzenburger, U. M., Buenrostro, J. D., Wu, B., Shen, Y., Sheffield, N. C., Kathiria, A., Greenleaf, W. J. & Chang, H. Y. Single-cell epigenomic variability reveals functional cancer heterogeneity. en. *Genome Biol.* **18,** 15 (2017) (cited on p. 68).

154. Aznauryan, E., Yermanos, A., Kinzina, E., Devaux, A., Kapetanovic, E., Milanova, D., Church, G. M. & Reddy, S. T. *Discovery and validation of human genomic safe harbor sites for gene and cell therapies* 2022 (cited on p. 71).

155. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. en. *Cell* **135,** 216–226 (2008) (cited on p. 73).

156. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357,** 661–667 (2017) (cited on p. 73).

157. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. en. *Bioinformatics* **26,** 841–842 (2010) (cited on p. 74).

158. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. en. *Nat. Struct. Mol. Biol.* **15,** 1263–1271 (2008) (cited on p. 92).

159. Wernet, M. F., Mazzoni, E. O., Çelik, A., Duncan, D. M., Duncan, I. & Desplan, C. Stochastic spineless expression creates the retinal mosaic for colour vision. en. *Nature* **440,** 174–180 (2006) (cited on p. 92).

160. Tóth, B., Ben-Moshe, S., Gavish, A., Barkai, N. & Itzkovitz, S. Early commitment and robust differentiation in colonic crypts. en. *Mol. Syst. Biol.* **13,** 902 (2017) (cited on p. 92).

161. Dasen, J. S., Liu, J.-P. & Jessell, T. M. Motor neuron columnar fate imposed by sequential phases of Hox-c activity. en. *Nature* **425,** 926–933 (2003) (cited on p. 92).

162. Dasen, J. S., Tice, B. C., Brenner-Morton, S. & Jessell, T. M. A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. en. *Cell* **123,** 477–491 (2005) (cited on p. 92).

163. Cembrowski, M. S. & Menon, V. Continuous Variation within Cell Types of the Nervous System. en. *Trends Neurosci.* **41,** 337–348 (2018) (cited on p. 92).

164. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. en. *Nature* **525,** 251–255 (2015) (cited on p. 92).

165. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. en. *Science* **356** (2017) (cited on p. 92).

166. Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A. & Alon, U. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. *Cell Systems* **8,** 43–52.e5 (2019) (cited on p. 92).

167. Lenstra, T. L., Rodriguez, J., Chen, H. & Larson, D. R. Transcription Dynamics in Living Cells. en. *Annu. Rev. Biophys.* **45,** 25–47 (2016) (cited on p. 92).

168. Zoller, B., Little, S. C. & Gregor, T. Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. en. *Cell* **175,** 835–847.e25 (2018) (cited on pp. 92, 96).

169. Tunnacliffe, E., Corrigan, A. M. & Chubb, J. R. Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication. en. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 8364–8369 (2018) (cited on p. 92).

170. Nicolas, D., Zoller, B., Suter, D. M. & Naef, F. Modulation of transcriptional burst frequency by histone acetylation. en. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 7153–7158 (2018) (cited on pp. 92, 93, 99).

171. Fukaya, T., Lim, B. & Levine, M. Enhancer Control of Transcriptional Bursting. en. *Cell* **166,** 358–368 (2016) (cited on pp. 92, 96).

172. 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* (2018) (cited on pp. 92, 99, 106).

173. Wallrath, L. L. & Elgin, S. C. Position effect variegation in Drosophila is associated with an altered chromatin structure. en. *Genes Dev.* **9,** 1263–1277 (1995) (cited on p. 92).

174. Hornung, G., Bar-Ziv, R., Rosin, D., Tokuriki, N., Tawfik, D. S., Oren, M. & Barkai, N. Noise-mean relationship in mutated promoters. en. *Genome Res.* **22,** 2409–2417 (2012) (cited on p. 92).

175. Soltani, M., Bokes, P., Fox, Z. & Singh, A. Nonspecific transcription factor binding can reduce noise in the expression of downstream proteins. en. *Phys. Biol.* **12,** 055002 (2015) (cited on p. 93).

176. Dey, S. S., Foley, J. E., Limsirichai, P., Schaffer, D. V. & Arkin, A. P. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. en. *Mol. Syst. Biol.* **11,** 806 (2015) (cited on pp. 93, 95, 96, 99).

177. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. en. *Nat. Commun.* **10,** 4572 (2019) (cited on p. 94).

178. Rouhanifard, S. H. *et al.* ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. en. *Nat. Biotechnol.* (2018) (cited on pp. 95, 104).

179. Das, D., Dey, S., Brewster, R. C. & Choubey, S. Effect of transcription factor resource sharing on gene expression noise. en. *PLoS Comput. Biol.* **13,** e1005491 (2017) (cited on pp. 96, 99).

180. Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2,** 157–175 (2005) (cited on pp. 96, 99, 105).

181. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. en. *Nat. Genet.* **38,** 1341–1347 (2006) (cited on p. 97).

182. Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A. & Noble, W. S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. en. *Nat. Methods* **9,** 473–476 (2012) (cited on p. 98).

183. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. en. *Nat. Methods* **9,** 215–216 (2012) (cited on p. 98).

184. Lammers, N. C., Galstyan, V., Reimer, A., Medin, S. A., Wiggins, C. H. & Garcia, H. G. Multimodal transcriptional control of pattern formation in embryonic development. en. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 836–847 (2020) (cited on p. 99).

185. 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. en. *PLoS One* **11,** e0158298 (2016) (cited on p. 99).

186. Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and conventional nuclei. en. *Nature* **570,** 395–399 (2019) (cited on p. 99).

187. Yates, A. D. *et al.* Ensembl 2020. en. *Nucleic Acids Res.* **48,** D682–D688 (2020) (cited on p. 103).

188. Zhang, W. B., Sinha, D. B., Pittman, W. E., Hvatum, E., Stroustrup, N. & Pincus, Z. Extended Twilight among Isogenic C. elegans Causes a Disproportionate Scaling between Lifespan and Health. en. *Cell Syst* **3,** 333–345.e4 (2016) (cited on p. 105).

189. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. en. *PLoS Biol.* **16,** e2005970 (2018) (cited on p. 105).

190. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. en. *Genome Biol.* **7,** R100 (2006) (cited on p. 105).

191. Sherman, M. S. & Cohen, B. A. A computational framework for analyzing stochasticity in gene expression. en. *PLoS Comput. Biol.* **10,** e1003596 (2014) (cited on p. 105).

192. Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U. & Naef, F. Mammalian genes are transcribed with widely different bursting kinetics. en. *Science* **332,** 472–474 (2011) (cited on p. 106).

193. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional pulsing of a developmental gene. en. *Curr. Biol.* **16,** 1018–1025 (2006) (cited on p. 106).

194. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. en. *Bioinformatics* **36,** 918–919 (2020) (cited on p. 106).

195. Pintacuda, G. *et al.* hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. en. *Mol. Cell* **68,** 955–969.e10 (2017) (cited on p. 106).

196. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. en. *Genome Res.* **12,** 996–1006 (2002) (cited on p. 106).

197. Turabian, K. L. *A Manual for Writers of Term Papers, Theses, and Dissertations* 5th (University of Chicago Press, Chicago, 1987).

198. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821. doi:10.1126/science.1225829 (2012).

199. Chen, Q., Shi, J., Tao, Y. & Zernicka-Goetz, M. Tracing the origin of heterogeneity and symmetry breaking in the early mammalian embryo. en. *Nat. Commun.* **9,** 1819 (2018).

200. Hatleberg, W. L. & Hinman, V. F. Modularity and hierarchy in biological systems: Using gene regulatory networks to understand evolutionary change. en. *Curr. Top. Dev. Biol.* **141,** 39–73 (2021).

201. Dominguez, R. & Holmes, K. C. Actin structure and function. en. *Annu. Rev. Biophys.* **40,** 169–186 (2011).

202. Telser, A. *Molecular Biology of the Cell, 4th Edition* 2002.

203. Anderson, C. *et al.* Natural variation in stochastic photoreceptor specification and color preference in Drosophila. en. *Elife* **6** (2017).

204. Lorenz, K. & Cohen, B. A. Causal variation in yeast sporulation tends to reside in a pathway bottleneck. en. *PLoS Genet.* **10,** e1004634 (2014).

205. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. en. *Cell* **154,** 914–927 (2013).

206. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. en. *Cell* **135,** 216–226 (2008).

207. Lombardo, A. *et al.* Site-specific integration and tailoring of cassette design for sustainable gene transfer. en. *Nat. Methods* **8,** 861–869 (2011).

208. Ko, M. S., Nakauchi, H. & Takahashi, N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. en. *EMBO J.* **9,** 2835–2842 (1990).

209. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. en. *Nature* **453,** 544–547 (2008).

210. Desai, R. V. *et al.* A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. en. *Science* **373** (2021).

211. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. en. *Nature* **546,** 431–435 (2017).

212. Emert, B. L., Cote, C. J., Torre, E. A., Dardani, I. P., Jiang, C. L., Jain, N., Shaffer, S. M. & Raj, A. Variability within rare cell states enables multiple paths toward drug resistance. en. *Nat. Biotechnol.* **39,** 865–876 (2021).

213. Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. en. *Mol. Ther.* **24,** 678–684 (2016).

214. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. en. *Bioinformatics* **27,** 1017–1018 (2011).

215. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S. & Aiden, E. L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. en. *cels* **3,** 95–98 (2016).

216. Bianchi, V. *et al. Detailed Regulatory Interaction Map of the Human Heart Facilitates Gene Discovery for Cardiovascular Disease* en. 2019.

217. Harmston, N., Ing-Simmons, E., Perry, M., Barešić, A. & Lenhard, B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. en. *BMC Genomics* **16,** 963 (2015).

218. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. en. *Bioinformatics Advances* **2,** vbac016 (2022).

219. Transcriptional burst frequency and burst size are equally. *https://www.pnas.org › contenthttps://www.pnas.org › content.*

220. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. en. *Nature* **518,** 317–330 (2015).

221. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. en. *PLoS Comput. Biol.* **9,** e1003118 (2013).

222. Bailey, T. L. STREME: Accurate and versatile sequence motif discovery. en. *Bioinformatics* (2021).

223. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. en. *Genome Biol.* **8,** R24 (2007).

224. Corrales, M., Rosado, A., Cortini, R., van Arensbergen, J., van Steensel, B. & Filion, G. J. Clustering of Drosophila housekeeping promoters facilitates their expression. en. *Genome Res.* **27,** 1153–1161 (2017).

225. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22,** 403–434 (1976).

226. Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L. & Weinberger, L. S. Transcriptional burst frequency and burst size are equally modulated across the human genome. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17454–17459 (2012).

227. Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. & Weissman, J. S. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. en. *Nature* **441,** 840–846 (2006).

228. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. & de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). en. *Nat. Genet.* **38,** 1348–1354 (2006).