Title
Chromosomal Position Effects on Gene Expression Variability and Epigenetic Drug Sensitivity

Author
Zhang, Thanutra

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Chromosomal Position Effects on Gene Expression Variability and Epigenetic Drug Sensitivity

A dissertation submitted in partial satisfaction of the

requirement for the degree Doctor of Philosophy

in Molecular, Cellular and Integrative Physiology

by

Thanutra Zhang

2019

ABSTRACT OF THE DISSERTATION

Chromosomal Position Effects

on Gene Expression Variability and Epigenetic Drug Sensitivity

by

Thanutra Zhang

Doctor of Philosophy in Molecular, Cellular and Integrative Physiology

University of California, Los Angeles, 2019

Professor Roy Wollman, Chair

Chromosomal position effect, also known as position effect variegation, has been extensively studied for almost a century. A systematic approach to study positional effect is to isolate genetic from epigenetic factors specifically to measure the expression of the same gene positioned in different chromatin contexts. Current strategies to target a reporter gene at multiple genomic locations are not capable of increasing both the sensitivity and throughput of data. Here, we developed a new massively parallel method to create and identify isogenic reporter clones. This method allowed us to interrogate the effect of chromatin environment on gene expression variability and epigenetic drug sensitivity as well as identify their underlying mechanisms. In human cells, we found that the protein expression mean and noise significantly are varied by the genomic location of the gene. By mapping our measurements of reporter expression at different genomic loci with epigenetic profiles of the transcription factor enrichment and the distance to chromatin states, we identified the factors that impact gene regulation. Some factors are involved

in mediating both gene expression mean and noise, while other only control one of these features. Moreover, we discovered wide-spread loci-specific sensitivities to epigenetic drugs for three distinct chemical compounds that target histone deacetylase, DNA methylation and bromodomain proteins. By leveraging ENCODE data on chromatin modification, we identified features of chromatin environments that are most likely to be affected by these epigenetic drugs.

The dissertation of Thanutra Zhang is approved.

David Glanzman

Xinshu Xiao

Xia Yang

Sriram Kosuri

Roy Wollman, Committee Chair

University of California, Los Angeles

2019

DEDICATION

To Her Royal Highness Princess Sirindhorn and my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First, I would like to thank Professor Roy Wollman for his suggestions, support and encouragement throughout my graduate studies. In this lab, I really enjoyed doing experiments, meeting with nice people and discussing about science because ‘Science is fun and exciting’. I have learnt a lot from Roy. It is not just about the technical and scientific skills that I gained from this training. It is more about learning to change my mindset to handle the challenges and the inspiration to be a good scientist and mentor.

I am thankful to all my dissertation committee, Professor David Glanzman, Professor Xinshu Xiao, Professor Xia Yang, and Professor Sriram Kosuri for their valuable comments and suggestions for my project.

I would like to express my sincere thanks to Her Royal Highness Princess Sirinhorn for her generosity and financial support since I graduated from high school.

To my family, thank you for always being by my side. My success in every single step of my life came from unconditional love, support and encouragement from my mom, dad and sisters, Jirayapa (Bo) and Paphapit (Bua).

I really appreciate the assistance from every member in Wollman lab. I would like to especially thank Anna Pilko, Robert Foreman and Alok Maity for their help with my project. I also thank everyone else in Wollman lab for all their comments and suggestions.

Thank you to all my friends in San Diego and Los Angeles for their support and encouragement during my past six years. Every meeting, party and board game night with you made my time in this country so memorable.

Chapter Two in full is a work in preparation for submission for publication with the title "Identifying chromatin features that regulate gene expression distribution." Thanutra Zhang, Roy Wollman. The dissertation author is the primary investigator and author of this paper.

Chapter Three in full is a version of the material as it appears in Zhang, T; Pilko,A; Wollman, R. Loci specific epigenetic drug sensitivity. Biorxiv. 2019. The dissertation author was the primary investigator and author of this paper.

Lastly, I am very grateful to know Noravee at the beginning of my journey in the US. His love and support are significant for me to overcome all the challenges in my graduate school.

VITA

2013 Bachelor of Science, Peking University, China

2013-2016 Graduate student researcher, University of California, San Diego

2016-2019 Graduate student researcher, University of California, Los Angeles

2019 Doctor of Philosophy, University of California, Los Angeles

PUBLICATIONS

Zhang T., Pilko A., Wollman R. 2019. Loci specific epigenetic drug sensitivity. Biorxiv. doi: https://doi.org/10.1101/686139

Zhang T., Wollman R. 2019. Identifying chromatin features that regulate gene expression distribution. Manuscript in preparation

He J., **Zhang T.**, Fu X. 2017. Using a novel cellular platform to optimize CRISPR/CAS9 technology for the gene therapy of AIDS. Protein & Cell. doi: 10.1007/s13238-017-0453-z

PRESENTATIONS

Chromosomal Position Effects on Epigenetic Drug Sensitivity
Oral presentation at the 12th Annual International Conference on Systems Biology of Human Diseases, Kaiserin Friedrich-Haus Berlin, Germany, May 27-29, 2018

Massively Parallel Quantitative Analysis of Position Effects on gene expression distribution
Poster presentation at 1st Inaugural Symposium on Multiscale Cell Fate, the NSF-Simons Center for Multiscale Cell Fate Research at UCI, October 1-2, 2018

Massively Parallel Quantitative Analysis of Position Effects on CMV Silencing Kinetics
Poster presentation at the 11th Annual International Conference on Systems Biology of Human Diseases, the California NanoSystems Institute at UCLA, June 4-6, 2018

CHAPTER 1

Introduction

Chromosomal position effect

Eukaryotic transcriptional cpntrol is a complex process with multiple layers of regulation orchestrating a sophisticated network of signals and responses. It involves the coordination of sequence specific transcription factors, chromatin remodelers, epigenetic modifiers and long-range chromatin interactions between promoters and enhancers through chromatin looping factors [1–3]. Moreover, non-coding RNAs, such as enhancer RNAs and long non-coding RNAs [4,5], have also been implicated in regulation of transcription. Our understanding of the interplay between these regulatory players is still far from completed. One powerful tool to explore the effects of regulatory elements and local chromatin context on gene expression is to insert reporter genes into different genomic regions as a sensor. It has been long observed that the behavior of integrated reporters varies greatly depending on their positions in the genome and this phenomenon is called 'position effect' [6–9].

Currently, there are two main strategies to study chromosomal position effect. One is to either target reporter genes to selected genomic loci or randomly insert the reporters into host genome using stable transfection, or transposon- and viral-based delivery [10–15]. The bottleneck of this method comes from the process of validating targeted insertions or establishing clonal cell line and characterizing the integrated location of reporters, making this approach laborious and not scalable. Another strategy is to combine the traditional transgene reporter assay with synthetic

barcoding technology and next generation sequencing to circumvent the need of isolating clonal cell lines. The examples of this advanced approach include Thousands of Reporters Integrated in Parallel (TRIP) [16], Barcoded HIV ensembles (B-HIVE) [17], and parallel targeting of chromosome positions by massively parallel reporter assay (patchMPRA) [18]. The advantage of these methods is the throughput since they can probe the influence of chromatin environment on multifaceted aspects of gene regulation at hundreds to thousands of genomic sites. However, there is a trade off from using this strategy such as the ability to generate single-cell data. As a result, our understanding of chromosomal position effects on several aspects of gene regulation, such as the stochasticity in gene expression, remains elusive.

Next Generation Sequencing and its application

Early DNA sequencing technologies were developed in the late 1970s by Maxam and Gilbert [19,20] and Sanger and colleagues [21] using a method known as fragmentation or wandering-spot analysis and the chain termination, respectively. Since then, Sanger sequencing which is less toxic than Maxam and Gilbert's method has remained the most commonly used DNA sequencing technique to date and revolutionized biology by providing the tools to decipher complete genes and, later, entire genomes [22]. Lately, the Sanger method has been gradually superseded by several “next-generation” sequencing technologies that offer significant increases in cost-effective sequence throughput. The well-known next-generation technologies available nowadays include the 454 pyrosequencing based instrument [23] (Roche Applied Science), Solexa [24,25] (Illumina), SOLiD and Agencourt [26,27] (Applied Biosystems), Heliscope [28,29] (Helicos) and Ion Torrent (founded by Rothberg).

The applications of next generation sequencing (NGS) are beyond simple genome sequencing. Two of the highest impact areas are transcriptome analysis and genome-wide mapping of epigenetic markers and DNA regulatory elements at high resolution. The NGS-based RNA seq has transformed our view of the complex and dynamic nature of the transcriptome. It provides a more detailed and quantitative aspect of gene expression, alternative splicing, and allele-specific expression [30,31]. Deep sequencing when integrated with Chromatin Immunoprecipitation (ChIP) has become a powerful tool to reveal the location of chromatin modifications, transcription factor-bound sites and open region of genome. ChIP–seq has enabled tremendous progress in epigenetic research [32].

In this dissertation

Given that currently available tools to study chromosomal position effect are not capable of measuring single cell data in massive and parallel manner, we were motivated to develop a novel tool to massively establish and characterize clonal lines of reporters by integrating multiple advanced high-throughput technologies, such as synthetic barcode, next generation sequencing and high throughput flow cytometry, into a single pipeline. Furthermore, we aimed to use this tool to examine chromosomal position effects on gene expression variability and epigenetic drug sensitivity and identify their underlying molecular mechanisms.

Reference

1. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159: 1665–1680.

2. Li Q, Barkess G, Qian H. Chromatin looping and the probability of transcription. Trends Genet. 2006;22: 197–202.

3. Dekker J, Misteli T. Long-Range Chromatin Interactions. Cold Spring Harb Perspect Biol. 2015;7: a019356.

4. Lam MTY, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci. 2014;39: 170–182.

5. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol. 2013;14: 699–712.

6. Henikoff S. Position-effect variegation after 60 years. Trends Genet. 1990;6: 422–426.

7. Weiler KS, Wakimoto BT. Heterochromatin and gene expression in Drosophila. Annu Rev Genet. 1995;29: 577–605.

8. Ottaviani A, Gilson E, Magdinier F. Telomeric position effect: from the yeast paradigm to human pathologies? Biochimie. 2008;90: 93–107.

9. Muller HJ. Types of visible variations induced by X-rays inDrosophila [Internet]. Journal of Genetics. 1930. pp. 299–334. doi:10.1007/bf02984195

10. Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, et al. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Dev. 1995;9: 1797–1810.

11. Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, et al. Domain-wide regulation of gene expression in the human genome. Genome Res. 2007;17: 1286–1295.

12. Babenko VN, Makunin IV, Brusentsova IV, Belyaeva ES, Maksimov DA, Belyakin SN, et al. Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome [Internet]. BMC Genomics. 2010. p. 318. doi:10.1186/1471-2164-11-318

13. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, et al. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet. 2011;43: 379–386.

14. Chen M, Licon K, Otsuka R, Pillus L, Ideker T. Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position [Internet]. Cell Reports. 2013. pp. 128–137. doi:10.1016/j.celrep.2012.12.003

15. Chen X, Zhang J. The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. Cell Syst. 2016;2: 347–354.

16. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013;154: 914–927.

17. Chen H-C, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. Nat Struct Mol Biol. 2017;24: 47–54.

18. Maricque BB, Chaudhari HG, Cohen BA. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. Nat Biotechnol. 2018; doi:10.1038/nbt.4285

19. Gilbert W, Maxam A. The Nucleotide Sequence of the lac Operator [Internet]. Proceedings of the National Academy of Sciences. 1973. pp. 3581–3584. doi:10.1073/pnas.70.12.3581

20. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74: 560–564.

21. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

22. Consortium IHGS, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome [Internet]. Nature. 2001. pp. 860–921. doi:10.1038/35057062

23. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380.

24. Bennett ST, Barnes C, Cox A, Davies L, Brown C. Toward the 1,000 dollars human genome. Pharmacogenomics. 2005;6: 373–382.

25. Bentley DR. Whole-genome re-sequencing [Internet]. Current Opinion in Genetics & Development. 2006. pp. 545–552. doi:10.1016/j.gde.2006.10.009

26. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays [Internet]. Nature Biotechnology. 2000. pp. 630–634. doi:10.1038/76469

27. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009;19: 1527–1541.

28. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A. 2003;100: 3960–3964.

29. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008;320: 106–109.

30. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015: 951–969.

31. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63.

32. Park PJ. ChIP–seq: advantages and challenges of a maturing technology [Internet]. Nature Reviews Genetics. 2009. pp. 669–680. doi:10.1038/nrg2641

CHAPTER 2

Identifying chromatin features that regulate gene expression distribution

Abstract

Gene expression variability is ubiquitous across diverse organisms with broad impacts on multiple levels of biological system from affecting the function of transcriptional regulatory networks to originating phenotypic variations. However, the underlying mechanisms that modulate expression noise in higher eukaryotic cells are still poorly understood. Here we addressed this problem through a systematic investigation of molecular contributors to fluctuation in gene expression. Using DNA barcoding and split pool decoding, we created a large library of isogenic reporter clones and identified reporter integration sites in a massive and parallel manner. By mapping our measurements of reporter expression at different genomic loci with multiple epigenetic profiles including the enrichment of transcription factors and the distance to different chromatin states, we identified key factors that impact regulation of gene expression distributions.

Introduction

Fluctuation in gene expression, also termed gene expression noise, is prevalent across multiple organisms ranging from bacteria to mammalian cells [1,2]. Expression noise within genetically identical cells drive phenotypic variation which is important for many biological processes including multicellular development, cell differentiation and lineage decisions, viral decision making as well as bacteria and cancer cell survival during environmental stress [3–9]. Factors contributing to cell-to-cell variability are generally classified into intrinsic and extrinsic noise [10,11]. Lately, our understanding of the molecular regulatory mechanisms underlying the stochasticity in gene expression is increasing. Several lines of evidence support significant roles in expression noise for the presence of TATA-box [12–15], nucleosome occupancy and chromatin remodeling [11,16–18], transcriptional pausing [19,20], chromatin epigenetics [21–25] and concentration of transcription factors [26,27]. Nonetheless, the lack of systematic interrogating the contribution of genomic location to expression variability limits our understanding of the underlying molecular factors.

The influence of local chromatin environment on gene regulation or position-effect variegation has been extensively studied since the classical work in *Drosophila* eyes in 1930s [28–32]. Although, previous studies mostly examined such effects on averaged mRNA or protein productions, positional effects on the heterogeneity of expression is understudied. The pioneering work that addresses this question in a systematic manner was done by Chen and Zhang in *Saccharomyces cerevisiae* and suggests the association between expression noise and three histone modifications, H3K4me1, H3K4me3 and H3K79me3 [33]. Even though yeast and mammals share several features of transcription regulatory mechanisms, there are substantial differences in the

complexity of genomes between these two [34–36]. The yeast genome, which consists of ~12 Mb, is extremely compact while human genome is much bigger, or about 275 times the size of the genome of yeast and contains large amounts of noncoding DNA [37–39]. Considering additional differences in large-scale chromatin dynamics such as long-range chromatin interactions and higher-order chromatin structure, using yeast as a model system to study expression variability may hinder the finding of underlying mechanisms specific to complex genome structure in higher eukaryotes. For example, CTCF, a transcription factor conserved from fly to human but absent from yeast [40,41], is a critical regulator who creates boundaries between topologically associating domains (TAD) and regulates gene expression variability through mediating enhancer-promoter interaction [42–44]. Recently, Dar et al and Dey et al addressed the question of the control of expression noise at different genomic loci in human cell line. However, the genomic positions where reporters integrated were not identified, thus limited the study of the mechanisms behind the observed position effects [18,45]

To specifically investigate positional effects on both expression average and variability, controlling of genetic sequence as well as the measurement at single-cell resolution are required. Two traditional approaches including targeted integration of reporters at selected genomic locations or random insertion of reporters from transposon- or virus-based transfection, followed by reporter mapping can be used [33,46–50]. However, both methods are very laborious and not easy to scale up. Over the last few decades, advances in DNA synthesis and next-generation sequencing technologies (NGS) offer novel and rapid ways to interrogate chromosomal position effects on the scale of thousands of positions [51,52]. Nevertheless, these methods only allow the measurement of populational mRNA average from each location but do not provide information about expression variability, another important feature of gene regulation.

In this study, we utilized a high-throughput method to build and characterize a library of isogenic clones as a platform to study the positional effects on gene expression and heterogeneity at a large number of genomic loci. Significant levels of positional effects on both expression mean and variability were observed across human K562 cells. By leveraging publicly available data of the K562 epigenomic mapping to our reporter measurement, we identified and decomposed the factors that control gene expression mean and noise. Our findings provide a deeper understanding of the mechanisms underlying the stochasticity in gene expression and shed light on novel therapeutic strategy for expression-noise related diseases in humans.

Result

High-throughput generation and identification of isoclonal reporter clones

To obtain genomic scale data on the effect of chromatin environment on gene expression variability, a new high-throughput method was developed to facilitate the creation and identification of isogenic reporter clones in a massively parallel and highly scalable manner. The overview of our method is visualized in Figure 2.1. Briefly, the principle of this method involves tagging individual genomic location with the reporter cassette that contains a unique 16-nucleotide DNA barcode and fluorescent reporter mClover driven by CMV promoter. These barcodes serve as molecular identifiers for mapping the genomic location of the reporter in individual isogenic clones. The barcoded reporters were introduced into K562 cells through lentiviral transduction with low multiplicity of infection (MOI) to ensure single integration of reporter per cell. The founder cells were sorted by fluorescence activated cell sorter (FACS) and expanded for two weeks

and then split into two groups. The first group was used to establish isogenic clones and identify their corresponding barcodes through combinatorial pooled sequencing [53–55]. Particularly, clonal identity is transformed into unique pooling pattern. Each signature when matched with the appearance pattern of a specific barcode, demultiplexed from sequencing reads of pooled samples, will reveal DNA sequence of the barcode belonging to that clone. Once clonal lines were established, the measurements of reporter expression were performed by high-throughput flow cytometry, providing the information of gene expression distribution of each barcoded reporter. The other half of the founder cells was used for parallel mapping of reporter integration sites by applying Thousands of Reporters Integrated in Parallel (TRIP) method [51] which is based on inverse PCR [56] coupled with next generation sequencing. After matching detected location of barcodes with the database of K562 epigenetic profiles, information about local chromatin landscape surrounding each barcode was obtained. We confirmed the accuracy of our method through analysis of randomly picked 5 clones of barcoded cells, extracted genomic DNA, performed targeted PCR and Sanger sequenced. As expected, all revealed barcodes from Sanger sequencing matched with those deconvoluted from combinatorial pooled sequencing. We coupled these two measurements to generate large scale data for investigating the effect of chromatin environment on gene expression variability and interrogating the underlying molecular mechanisms.

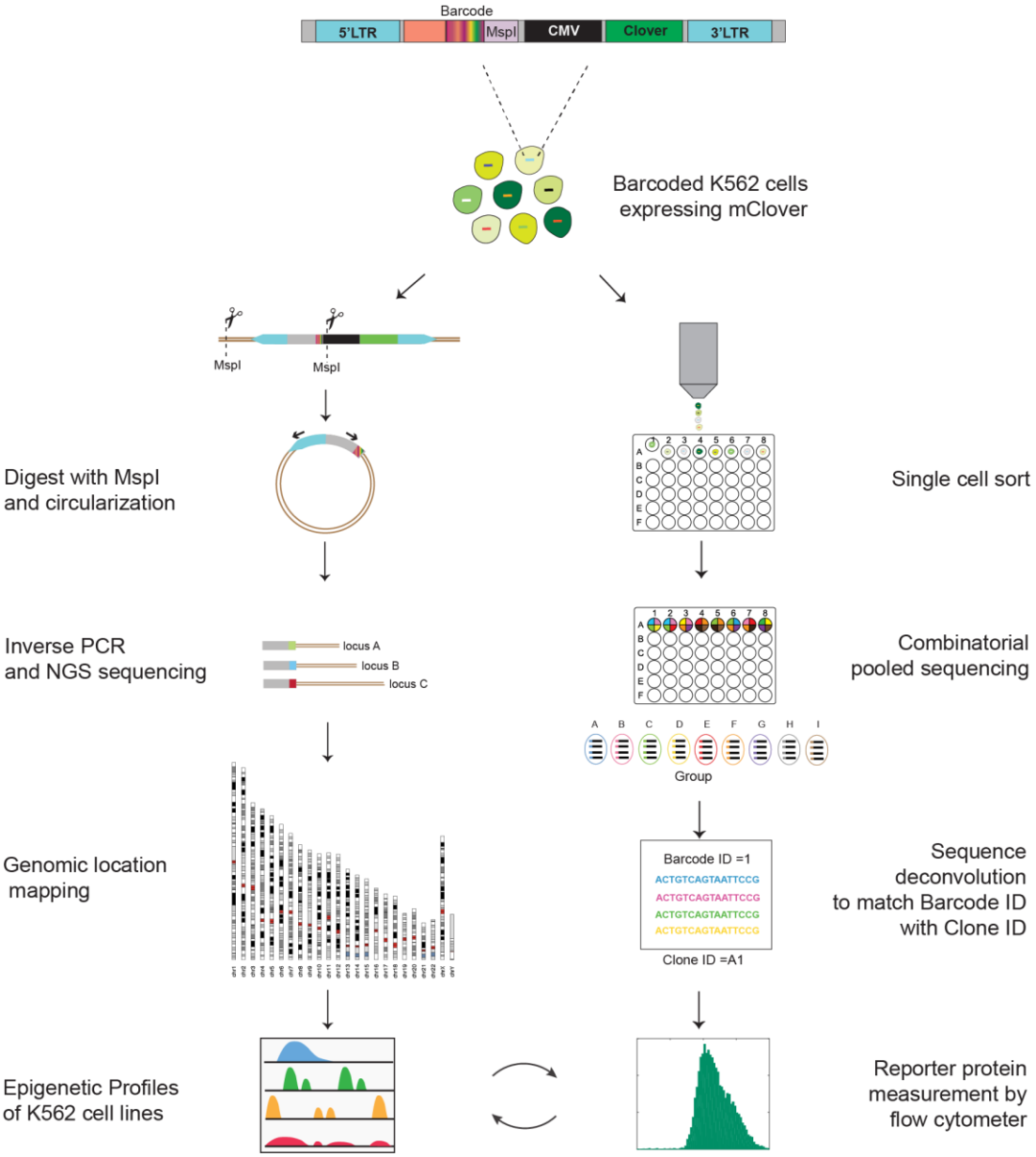


Figure 2.1 Protocol for scalable and parallel measurement of expression distribution and genomic position of reporter genes.

Two high-throughput methods were connected by incorporating random 16-bp DNA barcodes in the reporter cassette containing identical CMV promoter and a gene coding for fluorescent protein Clover. Synthetic reporter was introduced into K562 cells through low m.o.i. lentiviral transduction. Pooled identification of the genomic position of reporter genes based on the unique DNA barcode using TRIP method. Extracted genomic DNA of mixed founder cells was digested with MspI and followed by inverse PCR and deep sequencing. The creation of isogenic cell line, the identification of the DNA barcode sequence unique to each cell line using combinatorial pooled sequencing and the measurement of reporter expression distribution by high-throughput flow cytometry. The coupling of these two measurements allow the generation of large-scale data on the effect of chromatin environment on gene expression variability

Positional effects on multiple features of gene expression

In order to achieve sufficient statistical power to predict the molecular contributors underlying each feature of gene expression, we generated a library of reporter cells on the scale of hundreds of isogenic clones. About thirty percent of cell lines that we established lost reporter expression after two weeks of clonal expansion suggesting their insertions into heterochromatic environments [57]. Since we focus on gene expression variability, these non-expressing clones were excluded from future analysis yielding a total of 90 isogenic clones with confidently mapped reporters.

Established isogenic clones were next profiled for reporter expression at single-cell resolution by high-throughput flow cytometry. Specifically, three measurements were performed on three different experimental dates for each clone. We monitored and controlled batch effects between measurement by co-culturing control cells expressing both Clover and IRFP protein with experimental clones. Forward scatter (FSC), side scatter (SSC), Clover signal and IRFP signal of 50,000 lived cells were collected. To minimize extrinsic noise from differences in cell size, volume and cycle, a very conservative gating on the FSC versus SSC for a subset of live cells was applied. Such conservative gate approach is a validated method previously used in several studies to attenuate extrinsic noise [58–61]. Clover intensity of experimental clones was isolated from internal control cells by IRFP gating and calculated for expression average and noise (Figure2.2A). We chose to quantify protein expression noise by the squared coefficient of variation (CV^2), which is defined as the ratio of variance over the mean squared, as it was widely used in several experimental systems and studies of gene expression noise [58,62–65].

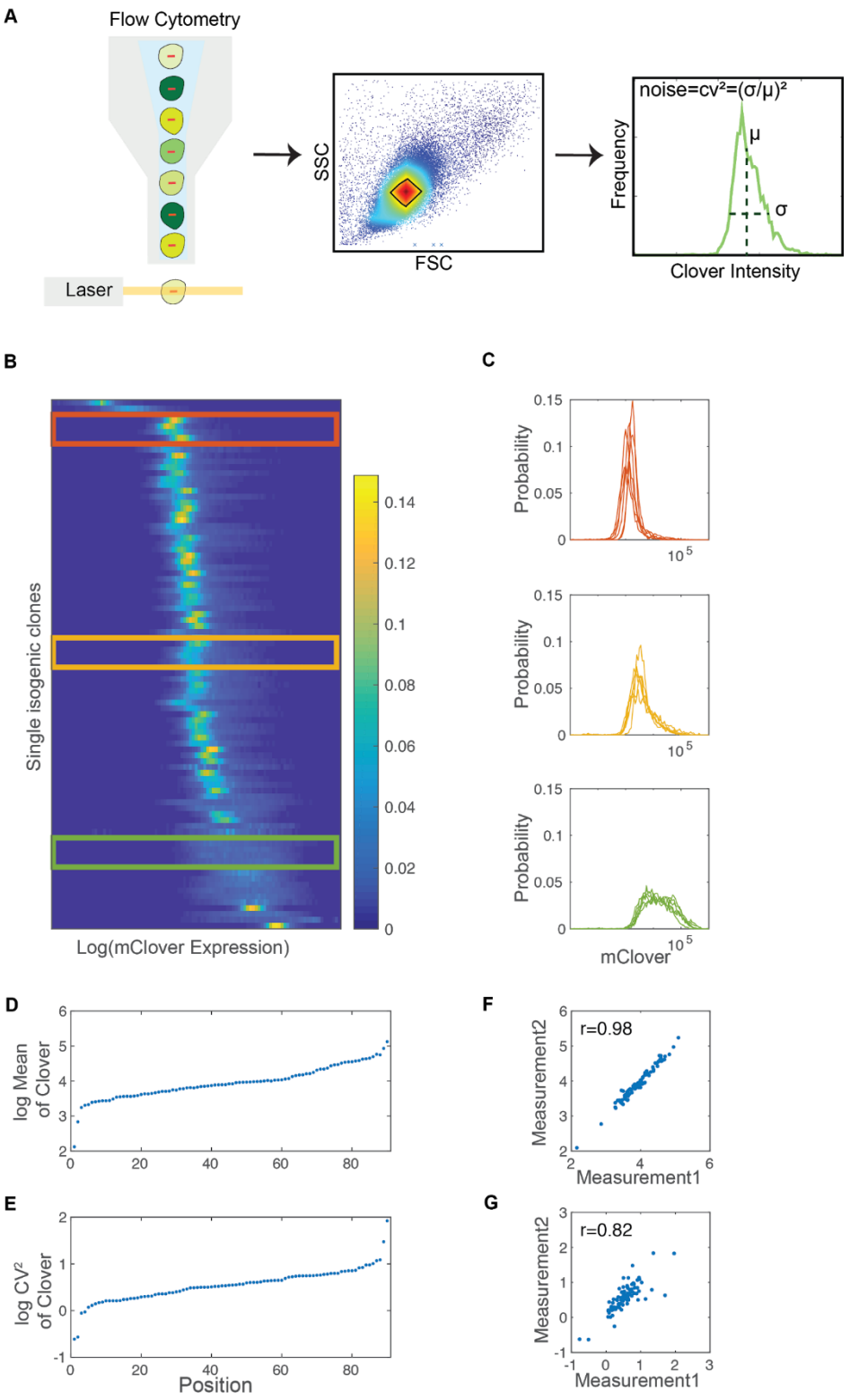


Figure 2.2: Gene expression distribution of a library of isogenic cell lines expressing CMV driven mClover fluorescent reporter shows dramatic gene expression variability.

(Legend on next page)

Figure 2.2: **Gene expression distribution of a library of isogenic cell lines expressing CMV driven mClover fluorescent reporter shows dramatic gene expression variability.**

(A) Reporter expression at the single-cell level was measured by high-throughput flow cytometry for three replicates. A very conservative gate controlling cell size,volume and cycle was applied to 50,000 live cells collected from each isogenic clone to minimize extrinsic noises. Gene expression variation can be quantified by CV^2 which is a measure of noise independent of gene expression levels. (B) Stacked probability density function of log10 expression of mClover in 90 cell lines. Each row in the heatmap represents a single histogram with the probability density function colorcoded. (C) Examples of the histogram of 30 cell lines from the top (red), middle (yellow), and bottom (green) of the stack in B. By comparing these histograms, it clearly shows dramatic difference in expression distribution across different genome positions. (D-E) The mean (D) and noise (E) of Clover levels in 90 isogenic clones. Each point represents averaged fluorescence intensity(D) or the intrinsic noise CV^2 (E) of Clover of a single reporter clone representing the data from one specific genome location. (F-G) The mean(F) and noise(G) of Clover expression of each isogenic clone is plotted for two measurement replicates (Spearman's rank-order correlation coefficient is 0.98 for expression mean (F) and 0.82 for expression noise (G); both axes are in logarithmic scale). Each measurement was performed independently on different experimental setup and dates.

A direct comparison of reporter expression across 90 clones show dramatic variability in the distribution of Clover protein (Figure 2.2B and 2.2C). We observed ~1000-fold difference in mean reporter protein level between the dimmest and the brightest clone (Figure 2.2D). Importantly, our data demonstrates that the stochasticity in gene expression is also position dependent (Figure 2.2E). Reporter expression noise from CMV promoter is altered more than 300 times by their positions. The CV^2 of the most variable clone is ~ 85 while the quietest one is only ~0.25. Expression data are highly correlated between independent measurements (Figure 2.2F and 2.2G) suggesting that extrinsic noise factors cannot explain such positional difference. Moreover, expression distributions of internal control cells are highly consistent across 90 clones discounting the possibility that observed effects arise from technical noises due to culture condition.

Our data show comparable level of the positional variation in expression average and variability when compared with other studies using higher eukaryotes as a model system [45,51]. However, much smaller chromosomal position effects were suggested in other studies carried out

in bacteria or yeast [31,33,50]. This discrepancy may come from their large and complex genome organization. However, we cannot rule out the possibility that come from differences in experimental design and techniques. Overall, our data suggests significant contribution of position effects on multiple features of gene control.

Integrative analysis of transcription factors on differential control of expression mean and noise

To better understand the molecular mechanisms underlying the position effects on expression level and noise, we integrated publicly available high-quality ChIP-seq data from ENCODE [66,67] with our measurement data. Specifically, a window of 50 kilobases surrounding the position of each barcode was calculated for the enrichment of transcription factor (TF) (Figure 2.3A). The Spearman rank correlation coefficient between TF enrichment and expression mean and noise measured by our assay was calculated. From about two hundred tested transcription factors, only a small number of them showed moderately positive or negative correlation with expression mean and noise (Figure 2.3B).

To identify a subset of TFs that play a role in gene expression, we used a multivariate linear regression analysis with stepwise procedure to determine the terms in the model to integratively identify the most likely candidate transcription factors linked to our reporter activities. Model selection approach was essential due to the large number of TFs with high quality ChIP-seq data publicly available. This approach identified a set of transcription factors that statistically explain the reporter expression average and noise, with an estimate of their relative contribution to the predictive power of the model (Figure 2.3C, 2.3D). Interestingly, we found transcription factors

controlling reporter expression average do not always mutually regulate expression noise (Figure 2.3D) and suggests specific factors that orthogonally control two features of gene regulation.

Many of the identified contributing transcription factors have functions involved in chromatin remodeling and transcriptional activation or repression. For example, GATAD2B is one of the factors contributing to the level of reporter expression average. Higher enrichment of this TF is correlated with lower mean of reporter expression. GATAD2B, encoded from the human GATA zinc finger domain containing 2B is beta-subunit of the transcription repressor complex MeCP1-Mi2/nucleosome remodeling and deacetylase complex that involved in chromatin modification and transcription activity [68–70]. TRIM24, on the other hand, has an opposite effect on reporter expression average. Higher enrichment of this TF is associated with increased reporter expression. Previous studies have reported TRIM24 as a transcriptional activator in various signaling pathways [71–73].

Moreover, our results also suggest the role of pioneer transcription factor in gene expression noise. For instance, FOXA1, forkhead box protein A, significantly correlates with the variability of reporter expression. This transcription factor is postulated to have unique properties that allow them to interact with closed nucleosome arrays, initiate epigenetic switch and thereby open condensed chromatin structures [74–77]. Additionally, we found POU5F1, also known as OCT4, is related with high expression noise. POU5F1 possess DNA binding domain that differ from that of FoxA, but also preferentially target silent sites enriched for nucleosomes. As a result, its pioneer activity can initiate cell-fate changes [78,79]. Accordingly, a previous study found genes with super-enhancers that are densely occupied by POU5F1 have unusually high cell-to-cell expression variation [80].

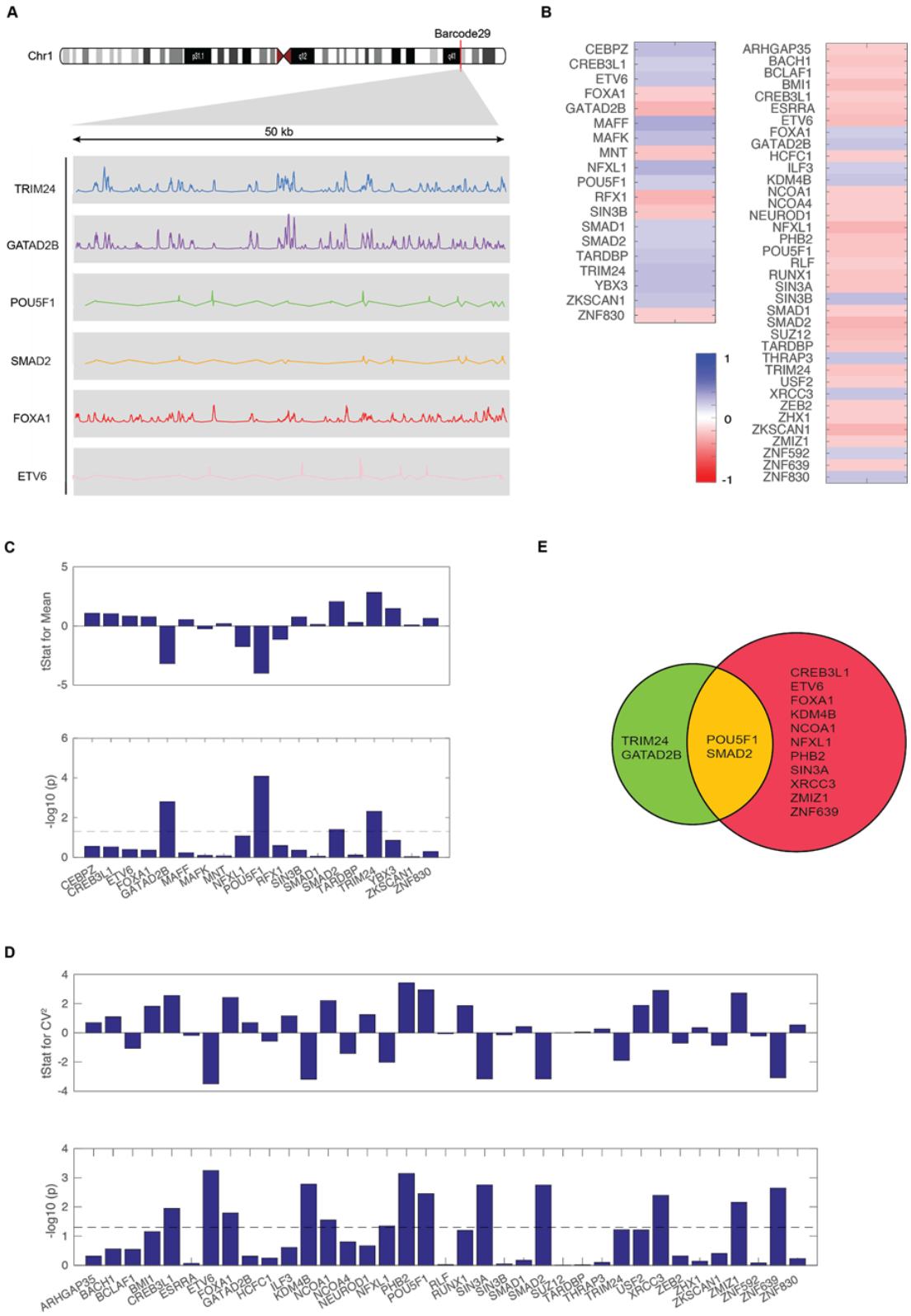


Figure 2.3 Identification of key transcription factors associated with reporter expression levels and noises using multivariate linear regression analysis

(Legend on next page)

Figure2.3: Identification of key transcription factors associated with reporter expression levels and noises using multivariate linear regression analysis

(A) View of the barcode 29 locus on chromosome 1. Genome coordinate of each barcode was obtained from conducting TRIP experiment. Six examples of normalized ChIP-seq signal profile of transcription factor surrounding the integration site of the barcode 29 were visualized for a domain of 50 kb. (B) The heatmap showing the correlation between the enrichment of transcription factors and reporter expression mean and noise. Transcription factors with Spearman's rank-order correlation coefficient more than 0.25 were selected from over 200 tested transcription factors. (C-D) To understand the relationship between the enrichment of TF and the expression levels(C) and noise(D) of reporter in an integrative way, we selected transcription factors showing significant correlation in B to fit multivariate linear regression model. Features with significant level above the threshold or dashed line ($p<0.05$) contributed significantly to the model. (E) A Venn diagram listing transcription factors contributing to reporter expression mean (green) and noise (red).

Distance to chromatin states influences expression mean and CV^2

K562 cell line is a well-established model for studies of chromatin regulation and has the largest number of publicly available datasets generated mainly by the ENCODE project [81]. Although the individual data track of epigenetic mark and regulatory element is informative, the systematic annotations derived from their interrelations contain higher level information and provide deeper insight into the functional element of genome. Therefore, we examined the influence of chromatin states on reporter expression level. Chromatin states of K562 used in our analysis were learned by computationally integrating ENCODE ChIP-seq, DNase-seq, and FAIRE-seq data using a Hidden Markov Model (HMM) [82]. Whole genome of K562 were segment into twenty-five states according to the combination of multiple epigenetic marks and these states were then classified into ten predicted functional elements including active promoter (Tss,TssF), promoter flanking (PromF), inactive promoter (PromP), candidate strong enhancer (Enh, EnhF), candidate strong enhancer or DNase (EnhWF, EnhW, DNaseD, DNaseU, FaireW), distal CTCF or candidate insulator (Ctcf, CtcfO), transcription associated (Gen5', Elon, ElonW,

Gen3', Pol2, H4K20), low activity proximal to active states (Low), polycomb repressed (ReprD, Repr, ReprW) and heterochromatin (Quies, Art) [83].

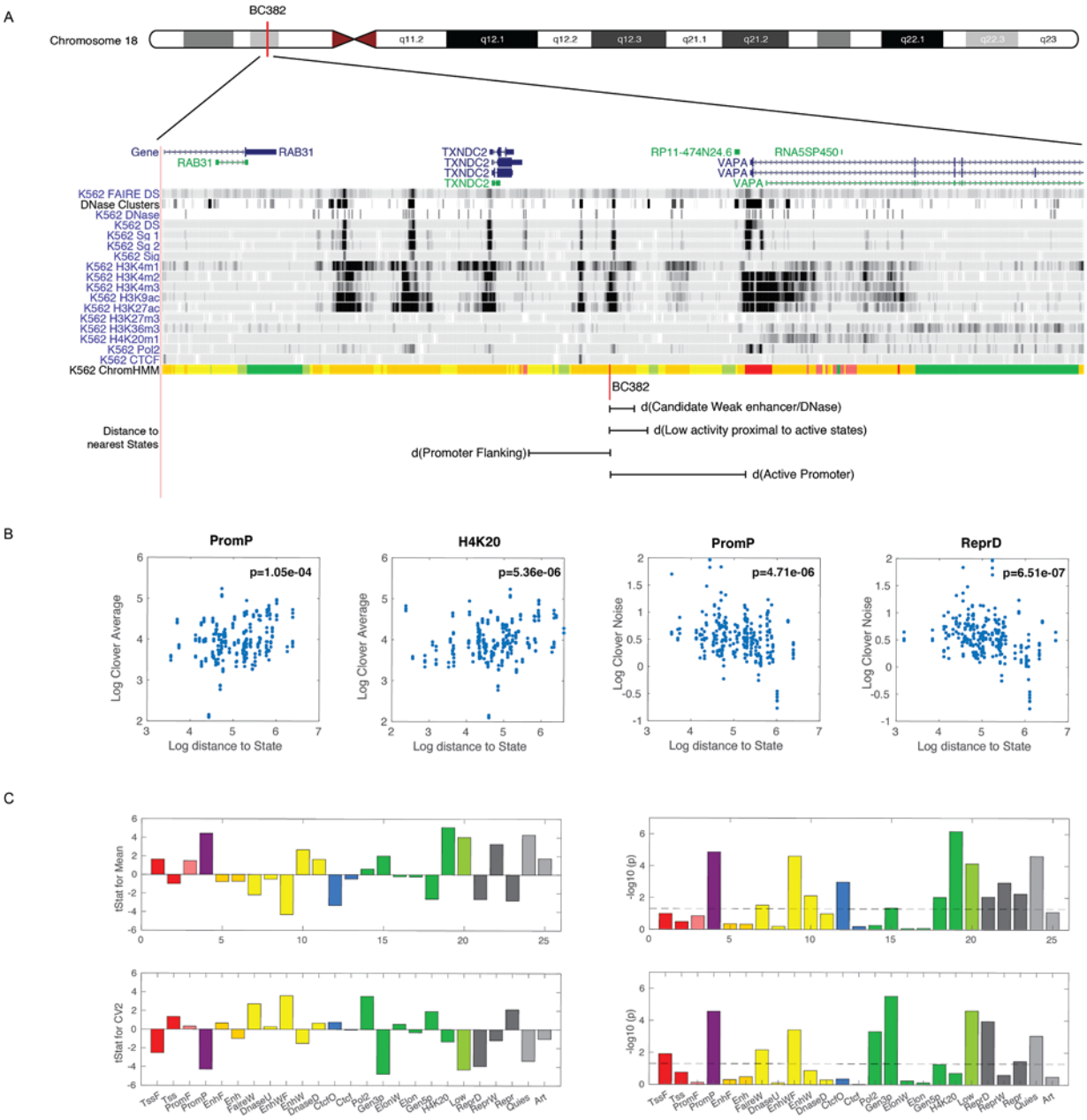


Figure 2.4 Distance to specific chromatin states influences expression mean and variance

(A) Nearest distance from the barcode location to each chromatin state segmented by ChromHMM method was calculated. (B) The correlation plot between distance to specific chromatin state and reporter expression mean or noise. (C) Multivariate linear regression model was used to determine significant chromatin states influencing reporter expression mean and noise.

We calculated the nearest distance between reporter location to each chromatin state (Figure 2.4A) and applied multivariate linear regression to collectively investigate the contribution of the distance of chromatin state to reporter expression (Figure 2.4C). We found reporter expression average was significantly associated with the distance to the following states: H4K20, PromP, Quies, EnhWF, Low, CtcfO, ReprW, Repr, EnhW, ReprD, Gen5p, FaireW, and Gen3p respectively. Only a subset of these chromatin states displayed their connection with reporter expression noise. The significant contributors to expression noise include chromatin state Gen3p, Low, PromP, ReprD, EnhWF, Pol2, Quies, FaireW, TssF and Repr successively. We found the distance to chromatin state H4K20 and CtcfO is highly associated with gene expression mean but not the noise. Reporter genes located closer to CtcfO trend to have higher averaged gene expression. CtcfO is highly enriched in CTCF, PolymeraseII, H3K4me1 and a marker of opened chromatin (DNase). Oppositely, closer distance to H4K20 is likely to decrease expression mean (Figure 2.4B). This relationship is in agreement with several studies previously found the role of H4K20 methylation in transcriptional repression and gene silencing [84–86].

Interestingly, two chromatin states that are highly associated with expression noise detected by our assays (Figure 2.4B) are linked with bivalent chromatin structure. PromP is 'poised promoter' and associated with both the active H3K4me3 mark and the polycomb-repressed H3K27me3 modification. ChromHMM state ReprD has a relatively high frequency of H3K27me3 and DNase sensitivity. Recently, two research groups have reported conflicting chromatin states as one of the determinants of high noise in gene expression [20,80]. Therefore, our results are consistent with previous studies. Moreover, we also found highly significant contribution of the distance to chromatin state Pol2 and Gen3p to gene expression noise exclusively. Although these two states possess similar profiles of high Polymerase II enrichment and relatively open chromatin

structure, the contrasting effects on expression noise were observed. Gen3p state is associated with high expression noise and Pol2 state is vice versa. The key differences of epigenetic marks between these two states are H3K36me3 and H4K20me1 suggesting their roles in enhancing noise in gene expression.

Discussion

Here, we developed a new method to investigate the underlying factors controlling gene expression mean and variability in the human genome (Figure 2.1). We showed that the insertion site affects both features of gene expression (Figure 2.2). Mechanistic insights related to the factors underlying expression mean and variance noise were gleaned by leveraging multiple epigenetic profiles with our measurement data (Figure 2.3 and 2.4). Overall, our results provide new insights into chromatin factors that contribute to regulation of gene expression and highlights the importance of chromosomal context in gene regulation.

Our results illustrate the power of combining two high throughput sequencing based tools. TRIP has high capacity of revealing several thousands of barcodes in a single run of deep sequencing while the use of combinatorial pooled sequencing eliminates the process of individual genomic extraction and PCR amplification per clones which significantly decreases both the cost and time for sequencing preparation. Moreover, our approach is highly scalable. Identification of thousands of clones can be achieved by only tens of genomic extraction and PCR through encoding clone identity in a format of pooling pattern. For instance, 24-choose-4 or 10,626 cell lines can be pooled into 24 flasks in a specific combinatorial pattern such that each cell line is combined into a unique subset of exactly four flasks out of the 24 flasks. Therefore, the only limit of scale is

pipetting time to pool cell lines and measure gene expression. This bottleneck can be overcome by more advanced robotic pipettor [87].

The discovery of DNA methylation and histone post-translational modifications has led to extensive studies on the impact of epigenetic modifications on gene regulation. Although, earlier studies generally considered each modification as a simple code, further exploration revealed complex patterns of their combinations. Recently, the concept of chromatin state, defined by the combinatorial presence and absence of multiple marks, has been introduced into epigenetic field to facilitate the functional interpretations of distinctive genome characteristics. Intuitively, local chromatin environment can be viewed as a unique combination of these basic building blocks. Yet how diversifying the organization of these blocks affect gene expression remains elusive. By directly quantifying the relationship between the distance to chromatin states and gene regulation, we are able to provide functional annotation to chromatin state in a way that is not influenced by the underlying DNA sequence. Our analysis recapitulates the previously reported roles of bivalent chromatin on gene expression variability. Additionally, we found that the distance to chromatin states CtcfO and H4K20 independently control reporter expression mean, while the distance to chromatin states Pol2 and Gen3P solely control reporter expression noise. Further investigation should focus on the mechanistic roles of H3K36me3 and H4K20me1 in gene expression variability. Since previous studies have observed that these two marks could regulate RNAPII-catalyzed transcription elongation by recruiting specific elongation inhibitors and enabling dynamic change in chromatin compaction [88–91], it is possible that modifying these histone marks could control gene expression noise through adjusting transcriptional elongation rate.

Chromatin regulation of gene expression is a highly complex process that is tightly coordinated with a numerous number of transcription factors. It is now widely accepted that

transcription factors modulate gene expression through multiple modes of mechanism that are not restricted to their co-binding at enhancers, suppressors and promoters. They also play architectural roles, activate chromatin remodeling and block nucleosome repositioning [92]. However, there is still limited understanding of the link between transcription factor and stochasticity in gene expression. Our result showing the contribution of various transcription factors on gene expression distribution provide important data to this fundamental question. Interesting, we found several pioneer transcription factors underlie expression noise observed in our study. A better understanding of genetic characteristics and genomic domain that favor the binding of those pioneer factors will shed light on detailed mechanisms of how chromatins connect with transcription factors to modulate gene expression.

We noted that transcription factors found to influence reporter expression mean and noise in our study might be specific to CMV promoter. Future work that generalizes our findings to other promoters is an important next step. The methodological advantages in the creation of scalable assays we describe will be key to addressing these issues. Another key limitation of our results is that they are only based on correlation without direct establishment of causation. It is possible that fluctuations in gene expression cause specific recruitment of histone. Specific manipulation of chromatin state followed by functional assessment of change in reporter gene expression will be needed to address this question. Nonetheless, given the complexity of chromatin states and the technical challenges associated with their manipulation the initial step of establishment of correlation is vital first step in this path.

Overall, our studies provide unprecedented genome-wide information of position effect on gene expression distribution in human genome and uncover the molecular and mechanistic basis of the position effects. New results in this study not only deepen our understanding of the

regulation of gene expression noise in higher eukaryotes genome, but also have potential implications for the development of novel method to tune transcriptional stochasticity.

Acknowledgements

We are thankful to Anna Pilko for her help with a library of barcoded plasmids and Robert Foreman for his help with mapping of reporter integration sites. This work was funded by NIH grants to RW EY024960 and GM111404. T.Z. was also supported by Thailand Her Royal Highness Princess Maha Chakri Sirindhorn fellowship.

Author Contributions

TZ and RW conceptualized the experiments and data analysis.TZ performed the experiments and performed data analysis. TZ and RW wrote and edited the paper.

Declaration of Interests

The authors declare no competing interests.

Lead Contact and Materials Availability

Further information and requests for resources and reagents should be directed to, and will be fulfilled by the corresponding author Roy Wollman (rwollman@ucla.edu).

Material and methods

Cell Lines

K562 cells were maintained in RPMI media (Gibco) supplemented with 10% FBS (Gibco), 1% GlutaMAX(Gibco) and 1% Penicillin Streptomycin. Cells were grown at 37 °C in an incubator with 5% CO2. HEK 293T cells for viral packing were grown under the same conditions in DMEM (Gibco) supplemented with 10% FBS (Gibco), 1% GlutaMAX(Gibco) and 1% Penicillin Streptomycin.

Library construction of barcoded reporter plasmid

The master plasmid excluding barcode was first constructed to contain the following essential elements. Lentiviral production units include HIV-1 truncated 5' LTR, HIV-1 packaging signal, HIV-1 Rev response element (RRE), HIV-1 truncated 3' LTR and Central polypurine tract (cPPT). These components allow proper viral packaging and viral integration into host cells. As a transcription unit, we used cytomegalovirus promoter (CMV) to drive expression of the reporter gene encoding yellow-green fluorescent protein (mClover). Woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) is placed after mClover to enhances mRNA stability and protein yield. Ampicillin resistance gene (β-lactamase) is included for selection of plasmid in bacterial cells.

To generate barcoded plasmid libraries, based lentiviral plasmid was cut upstream of the CMV promoter by ClaI restriction enzyme and purified by ethanol precipitation. The inserted cassette of 127-bp-long oligonucleotide containing a random 16-bp-long barcode sequence (repeats of A,T and G), MspI site, primer priming site and homology arms, were synthesized by Integrated DNA Technology. The assembly reaction of 1:5 vector:insert ratio was carried out for

1 hour at 50 °C using NEBuilder HIFI DNA assembly kit (New England Biolabs, NEB). Assembly products were electroporated into NEB Turbo Competent E.Coli (NEB) and then plated on ampicillin-containing medium. Ampicillin resistant colonies were collected and extracted for plasmids using Maxiprep kit (Invitrogen). Ten sampling clones from the agar plate were analyzed by PCR with forward primer, GATCCTGTAGAACTCTGAACCT, and reverse primer, AGTCGGTGTCTTCTATGGAG, and Sanger sequencing to verify successful cassette insertion and barcode diversity.

Generation of reporter cell lines

The lentiviral library carrying the barcoded reporter cassettes was used to transduce into K562 cells at a multiplicity of infection of 0.01 by culturing cells with barcoded virus in media supplemented with 5 μg/ml polybrene and 20mM HEPES for 2 hours of spinoculation and 24 hours of incubation. Afterwards, cells were collected by gentle centrifugation and the media was replaced with fresh cultured media. Cells were expanded for 72 hours and then subjected to FACS to isolate Clover-positive founder cells. Founder cells were expanded for 14 days and split into two pools. One was used for genomic mapping of barcoded reporter and the other was used for establishing isogenic reporter clones by single cell sorting.

Measurement of reporter protein expression

A million cells were passed through a 35μm mesh filter (Corning 352235) and placed on ice prior to FACS separation. Cells are resuspended in FACS buffer consisting of 96% PBS, 2% fetal bovine serum, 1% 100X Pen/Strep, and 1% 0.5M EDTA (pH 7.4). 50k live cells were collected from each well for FSC-A, SSC-A, the Clover intensity and IRFP intensity using the BD FACSCelesta flow cytometer. Data were analyzed with custom Matlab scripts. A very conservative gating for a live

subset of ~3k cells of similar size, volume, and state, was applied on the FSC versus SSC to reduce extrinsic noise contributions.

Identification of genuine barcodes and genomic integration sites

A library of genuine barcodes in founder cells was first listed. Briefly, barcode region was amplified in first nested PCR from 5 µg of genomic DNA in 50 ul of 20 cycle PCR reaction using Titanium Taq, forward primer: TATGGATCCTGTAGAACTCTG, and reverse primer: GCTCTGCTTATATAGACCTCCCAC. Barcode amplicons were enriched from genomic DNA using SPRI beads (Beckman Coulter) and further amplified in the second nested PCR for 20 cycles using forward primer: TGTAGAACTCTGAACCTAGCT and reverse primer: CGTAAGTTATGTAACGCGGA. Illumina adapter was attached to final amplicon, amplified and sequenced on Illumina HiSeq 3000 platform (1x50bp). Sequencing reads were filtered and analyzed using Matlab Bioinformatics Toolbox. To identify genuine barcode, we used the following algorithm. First, we sorted barcodes according to their counts from most frequent to least frequent. Then, mutant versions of each barcode, defined as barcodes within a Hamming distance of 2, were sequentially removed. We consider remaining sequences as ‘‘genuine’’ barcodes. We recovered 756 genuine barcodes from 3,000 sorted founder cells.

Mapping of reporter integration sites was done by inverse PCR coupled with high-throughput sequencing. Briefly, founder cells were collected and splitted into two replicates. For each replica, 2 µg of genomic DNA was digested with 20 units of MspI (NEB) overnight at 37°C in a volume of 100 µl. Subsequently, three sets of ligation reactions were set up by incubating 600 ng of purified digested DNA with 2 µl of high-concentration T4 DNA ligase (NEB, M0202T) overnight at 4°C in a volume of 400 µl. The ligation reactions were purified by phenol-chloroform

isoamyl alcohol extraction and ethanol precipitation. DNA pellets were dissolved in 30 µl of water. Two rounds of PCR were performed to amplify and enrich fragments containing both the barcodes and flanking genomic DNA regions. For the first round of nested PCR, five sets of 25-cycle reaction in a volume of 50 µl were performed using 5 µl of ligated products as templates, forward primer: TATGGATCCTGTAGAACTCTG, reverse primer: GCTTCAGCAAGCCGAGTCCTGCGTCGAG and Phusion Hot Start Flex 2X Master Mix (NEB). Amplicon was pooled together, cleaned by DNA Clean & Concentrator kit (Zymo), and diluted in 50 ul of water. For the second round of nested PCR, four sets of 15-cycle reaction in a volume of 50 ul were done with 5 ul of cleaned amplicon from first PCR, forward primer:TGTAGAACTCTGAACCTAGCT, reverse primer:GCTTTCAGGTCCCTGTTCGG. Purified sample was further ligated with Illumina adapter, amplified and sequenced on Illumina HiSeq 3000 platform (2x150bp). Sequencing reads were filtered and analyzed using Matlab Bioinformatics Toolbox. The genomic regions associated with genuine barcodes were extracted from mapping reads and aligned against the human genome (hg38) using STAR [93]. Detected integration sites from each replicate were compared and assigned to each genuine barcode only if top candidate site from both replicates are identical. Genome coordinate of reporter integration site was converted to human reference genome(hg19) using UCSC liftOver tool [94] for comparison to ChIP-Seq data.

Combinatorial pool sequencing

To simultaneously reveal the identity of reporter cell lines linked by DNA barcodes in a single run, combinatorial pooled sequencing was used. Specifically, clonal numbers were encoded in a form of pooling pattern and DNA barcodes were decoded from such known pattern. To increase decoding accuracy, we designed pooling signature to be unique four selected pools out of

total eighteen pools. Cells from each clone were split into four selected pools according to the design. Sequentially, genomic DNA from individual pool of mixed clones was extracted and used as templates for PCR to amplify barcode using same procedure described in the method of identification of genuine barcode list. Forward primers of second nested PCR contain 6-bp index DNA to label PCR products from each pool, which allow high-throughput multiplex sequencing. Sequences were filtered and demultiplexed using Matlab Bioinformatics Toolbox. Genuine barcodes from all pools were first listed. For each detected barcode, normalized counts per pools were calculated and pools showing high reads above the threshold were identified. Barcodes with four detected pools were first assigned to the clone showing matched pooling design. Some barcodes were found in more than four pools when sister cells, expanded from one founder cell, were sorted into multiple wells during single-cell sort. A list of merged pooling signature of two unassigned clones was matched with barcodes showing complexed readout. Clones with two inserted barcodes (~2% of the population) were excluded from the library of reporter cell lines.

Validation of Combinatorial pool sequencing

For the validation of combinatorial pool sequencing, 5 clones were randomly chosen and extracted for genomic DNA using PureLink Genomic DNA Mini Kit (Invitrogen). 200ng of purified genomic DNA was used as a template for PCR amplification with a set of validation primers, forward:TAGTGAACGGATCTCGACG, reverse:GCTCTGCTTATATAGACCTCCCAC. PCR products were cleaned by DNA Clean & Concentrator kit (Zymo) and Sanger sequenced to verify the barcode sequences.

Sequencing, quality control and demultiplexing

Libraries were sequenced on Illumina Hiseq3000 sequencing systems by UCLA technology center for genomics & bioinformatics. Low quality sequences were filtered out by mismatched read length and low-quality scores (>25% of base with score <20) using Matlab Seqfilter. Sample indexes were trimmed with Matlab Seqtrim and reads from different samples were demultiplexed by Matlab Seqsplit.

Multivariate linear regression analysis

Fluorescence data of each reporter cell line was collected in triplicate from different experimental setup. The mean and CV2 of Clover intensity were calculated from a population of more than 3,000 cells with stringent control of size, volume and state. To understand the relationship between TF enrichment and reporter expression, Spearman's rank-order correlation coefficient between Clover mean or CV2 and averaged transcription factor enrichment within a window of 50 kb was first assessed. We used K562 ChIP-Seq datasets of transcription factor in the format of fold change over control, uniformly processed from two replicates, from the ENCODE database (Michael Snyder,Stanford) at https://www.encodeproject.org. Transcription factors with correlation coefficient of more than 0.2 were selected for multivariate linear regression analysis using Matlab. The analysis of nearest distance to chromatin state and Clover expression was processed in the same procedure but without the preselection step based on the Spearman rank-order correlation coefficient. Chromatin state data was obtained from Genome Segmentations from ENCODE (ChromHMM Segmentations) at UCSC (Data version: ENCODE Jan 2011 Freeze)

References

1. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010;467: 167–173. doi:10.1038/nature09326

2. Balázsi G, van Oudenaarden A, Collins JJ. Cellular decision making and biological noise: from microbes to mammals. Cell. 2011;144: 910–925. doi:10.1016/j.cell.2011.01.030

3. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135: 216–226. doi:10.1016/j.cell.2008.09.050

4. Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. Nature. 2002;420: 231–237. doi:10.1038/nature01258

5. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453: 544–547. doi:10.1038/nature06965

6. Samoilov M, Plyasunov S, Arkin AP. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci U S A. 2005;102: 2310–2315. doi:10.1073/pnas.0406841102

7. Boettiger AN, Levine M. Synchronous and Stochastic Patterns of Gene Activation in theDrosophilaEmbryo [Internet]. Science. 2009. pp. 471–473. doi:10.1126/science.1173976

8. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV. Stochastic Gene Expression in a Lentiviral Positive-Feedback Loop: HIV-1 Tat Fluctuations Drive Phenotypic Diversity [Internet]. Cell. 2005. pp. 169–182. doi:10.1016/j.cell.2005.06.006

9. Süel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB. Tunability and noise dependence in differentiation dynamics. Science. 2007;315: 1716–1719. doi:10.1126/science.1137455

10. Elowitz MB. Stochastic Gene Expression in a Single Cell [Internet]. Science. 2002. pp. 1183–1186. doi:10.1126/science.1070919

11. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science. 2004;304: 1811–1814. doi:10.1126/science.1098641

12. Choi JK, Kim Y-J. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat Genet. 2009;41: 498–503. doi:10.1038/ng.319

13. Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, et al. Noise-mean relationship in mutated promoters [Internet]. Genome Research. 2012. pp. 2409–2417. doi:10.1101/gr.139378.112

14. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. Genetic properties influencing the evolvability of gene expression. Science. 2007;317: 118–121. doi:10.1126/science.1140247

15. Zoller B, Nicolas D, Molina N, Naef F. Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol Syst Biol. 2015;11: 823. doi:10.15252/msb.20156257

16. Boeger H, Griesenbeck J, Kornberg RD. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. Cell. 2008;133: 716–726. doi:10.1016/j.cell.2008.02.051

17. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, et al. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol. 2008;4: e1000216. doi:10.1371/journal.pcbi.1000216

18. Dey SS, Foley JE, Limsirichai P, Schaffer DV, Arkin AP. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. Mol Syst Biol. 2015;11: 806. doi:10.15252/msb.20145704

19. Ribeiro AS, Häkkinen A, Healy S, Yli-Harja O. Dynamical effects of transcriptional pause-prone sites. Comput Biol Chem. 2010;34: 143–148. doi:10.1016/j.compbiolchem.2010.04.003

20. Kar G, Kim JK, Kolodziejczyk AA, Natarajan KN, Torlai Triglia E, Mifsud B, et al. Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. Nat Commun. 2017;8: 36. doi:10.1038/s41467-017-00052-2

21. Lim HN, van Oudenaarden A. A multistep epigenetic switch enables the stable inheritance of DNA methylation states. Nat Genet. 2007;39: 269–275. doi:10.1038/ng1956

22. Miller-Jensen K, Dey SS, Schaffer DV, Arkin AP. Varying virulence: epigenetic control of expression noise and disease processes. Trends Biotechnol. 2011;29: 517–525. doi:10.1016/j.tibtech.2011.05.004

23. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332: 472–474. doi:10.1126/science.1198817

24. Weinberger L, Voichek Y, Tirosh I, Hornung G, Amit I, Barkai N. Expression noise and acetylation profiles distinguish HDAC functions. Mol Cell. 2012;47: 193–202. doi:10.1016/j.molcel.2012.05.008

25. Nicolas D, Zoller B, Suter DM, Naef F. Modulation of transcriptional burst frequency by histone acetylation. Proc Natl Acad Sci U S A. 2018;115: 7153–7158. doi:10.1073/pnas.1722330115

26. Octavio LM, Gedeon K, Maheshri N. Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression. PLoS Genet. 2009;5: e1000673. doi:10.1371/journal.pgen.1000673

27. Senecal A, Munsky B, Proux F, Ly N, Braye FE, Zimmer C, et al. Transcription factors modulate c-Fos transcriptional bursts. Cell Rep. 2014;8: 75–83. doi:10.1016/j.celrep.2014.05.053

28. Henikoff S. Position-effect variegation after 60 years. Trends Genet. 1990;6: 422–426. Available: https://www.ncbi.nlm.nih.gov/pubmed/2087785

29. Weiler KS, Wakimoto BT. Heterochromatin and gene expression in Drosophila. Annu Rev Genet. 1995;29: 577–605. doi:10.1146/annurev.ge.29.120195.003045

30. Ottaviani A, Gilson E, Magdinier F. Telomeric position effect: from the yeast paradigm to human pathologies? Biochimie. 2008;90: 93–107. doi:10.1016/j.biochi.2007.07.022

31. Bryant JA, Sellars LE, Busby SJW, Lee DJ. Chromosome position effects on gene expression in Escherichia coli K-12. Nucleic Acids Res. 2014;42: 11383–11392. doi:10.1093/nar/gku828

32. Muller HJ. Types of visible variations induced by X-rays inDrosophila [Internet]. Journal of Genetics. 1930. pp. 299–334. doi:10.1007/bf02984195

33. Chen X, Zhang J. The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. Cell Syst. 2016;2: 347–354. doi:10.1016/j.cels.2016.03.009

34. Galagan JE, Henn MR, Ma L-J, Cuomo CA, Birren B. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 2005;15: 1620–1631. doi:10.1101/gr.3767105

35. Rando OJ, Chang HY. Genome-wide views of chromatin structure. Annu Rev Biochem. 2009;78: 245–271. doi:10.1146/annurev.biochem.78.071107.134639

36. Court F, Miro J, Braem C, Lelay-Taha M-N, Brisebarre A, Atger F, et al. Modulated contact frequencies at gene-rich loci support a statistical helix model for mammalian chromatin organization. Genome Biol. 2011;12: R42. doi:10.1186/gb-2011-12-5-r42

37. Mohanta TK, Bae H. The diversity of fungal genome [Internet]. Biological Procedures Online. 2015. doi:10.1186/s12575-015-0020-z

38. Cooper GM. The Cell: A Molecular Approach [Internet]. Sinauer Associates; 2000. Available: https://books.google.com/books/about/The_Cell.html?hl=&id=DCdyQgAACAAJ

39. Morton NE. Parameters of the human genome. Proc Natl Acad Sci U S A. 1991;88: 7474–7476. doi:10.1073/pnas.88.17.7474

40. Cai M, Davis RW. Yeast centromere binding protein CBF1, of the helix-loop-helix protein family, is required for chromosome stability and methionine prototrophy [Internet]. Cell. 1990. pp. 437–446. doi:10.1016/0092-8674(90)90525-j

41. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci U S A. 2012;109: 17507–17512. doi:10.1073/pnas.1111941109

42. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function [Internet]. Nature Reviews Genetics. 2014. pp. 234–246. doi:10.1038/nrg3663

43. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485: 376–380. doi:10.1038/nature11082

44. Ren G, Jin W, Cui K, Rodrigez J, Hu G, Zhang Z, et al. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression [Internet]. Molecular Cell. 2017. pp. 1049–1058.e6. doi:10.1016/j.molcel.2017.08.026

45. Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Natl Acad Sci U S A. 2012;109: 17454–17459. doi:10.1073/pnas.1213530109

46. Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, et al. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Dev. 1995;9: 1797–1810. doi:10.1101/gad.9.14.1797

47. Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, et al. Domain-wide regulation of gene expression in the human genome. Genome Res. 2007;17: 1286–1295. doi:10.1101/gr.6276007

48. Babenko VN, Makunin IV, Brusentsova IV, Belyaeva ES, Maksimov DA, Belyakin SN, et al. Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome [Internet]. BMC Genomics. 2010. p. 318. doi:10.1186/1471-2164-11-318

49. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, et al. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet. 2011;43: 379–386. doi:10.1038/ng.790

50. Chen M, Licon K, Otsuka R, Pillus L, Ideker T. Decoupling epigenetic and genetic effects through systematic analysis of gene position. Cell Rep. 2013;3: 128–137. doi:10.1016/j.celrep.2012.12.003

51. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013;154: 914–927. doi:10.1016/j.cell.2013.07.018

52. Chen H-C, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. Nat Struct Mol Biol. 2017;24: 47–54. doi:10.1038/nsmb.3328

53. Patterson N, Gabriel S. Combinatorics and next-generation sequencing. Nat Biotechnol. 2009;27: 826–827. doi:10.1038/nbt0909-826

54. Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, et al. DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 2009;19: 1243–1253. doi:10.1101/gr.092957.109

55. Cao C-C, Sun X. Combinatorial pooled sequencing: experiment design and decoding. Quantitative Biology. 2016;4: 36–46. doi:10.1007/s40484-016-0064-3

56. Ochman H, Gerber AS, Hartl DL. Genetic applications of an inverse polymerase chain reaction. Genetics. 1988;120: 621–623. Available: https://www.ncbi.nlm.nih.gov/pubmed/2852134

57. Elgin SCR, Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol. 2013;5: a017780. doi:10.1101/cshperspect.a017780

58. Singh A, Razooky B, Cox CD, Simpson ML, Weinberger LS. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. Biophys J. 2010;98: L32–4. doi:10.1016/j.bpj.2010.03.001

59. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature. 2006;441: 840–846. doi:10.1038/nature04785

60. Skupsky R, Burnett JC, Foley JE, Schaffer DV, Arkin AP. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. PLoS Comput Biol. 2010;6. doi:10.1371/journal.pcbi.1000952

61. Dar RD, Hosmane NN, Arkin MR, Siliciano RF, Weinberger LS. Screening for noise in gene expression identifies drug synergies. Science. 2014;344: 1392–1396. doi:10.1126/science.1250220

62. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. Nat Genet. 2006;38: 636–643. doi:10.1038/ng1807

63. Keren L, Hausser J, Lotan-Pompan M, Vainberg Slutskin I, Alisar H, Kaminski S, et al. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. Cell. 2016;166: 1282–1294.e18. doi:10.1016/j.cell.2016.07.024

64. Zoller B, Little SC, Gregor T. Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. Cell. 2018;175: 835–847.e25. doi:10.1016/j.cell.2018.09.056

65. Keren L, van Dijk D, Weingarten-Gabbay S, Davidi D, Jona G, Weinberger A, et al. Noise in gene expression is coupled to growth rate. Genome Res. 2015;25: 1893–1902. doi:10.1101/gr.191635.115

66. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489: 91–100. doi:10.1038/nature11245

67. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

68. Brackertz M, Boeke J, Zhang R, Renkawitz R. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. J Biol Chem. 2002;277: 40958–40966. doi:10.1074/jbc.M207467200

69. Brackertz M, Gong Z, Leers J, Renkawitz R. p66alpha and p66beta of the Mi-2/NuRD complex mediate MBD2 and histone interaction. Nucleic Acids Res. 2006;34: 397–406. doi:10.1093/nar/gkj437

70. Feng Q, Zhang Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. Genes Dev. 2001;15: 827–832. doi:10.1101/gad.876201

71. Groner AC, Cato L, de Tribolet-Hardy J, Bernasocchi T, Janouskova H, Melchers D, et al. TRIM24 Is an Oncogenic Transcriptional Activator in Prostate Cancer. Cancer Cell. 2016;29: 846–858. doi:10.1016/j.ccell.2016.04.012

72. Tsai W-W, Wang Z, Yiu TT, Akdemir KC, Xia W, Winter S, et al. TRIM24 links a non-canonical histone signature to breast cancer [Internet]. Nature. 2010. pp. 927–932. doi:10.1038/nature09542

73. Lv D, Li Y, Zhang W, Alvarez AA, Song L, Tang J, et al. TRIM24 is an oncogenic transcriptional co-activator of STAT3 in glioblastoma. Nat Commun. 2017;8: 1454. doi:10.1038/s41467-017-01731-w

74. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011;25: 2227–2241. doi:10.1101/gad.176826.111

75. Sérandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C, et al. Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome Res. 2011;21: 555–565. doi:10.1101/gr.111534.110

76. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell. 2002;9: 279–289. Available: https://www.ncbi.nlm.nih.gov/pubmed/11864602

77. Belikov S, Astrand C, Wrange O. FoxA1 binding directs chromatin structure and the functional response of a glucocorticoid receptor-regulated promoter. Mol Cell Biol. 2009;29: 5413–5425. doi:10.1128/MCB.00368-09

78. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell. 2015;161: 555–568. doi:10.1016/j.cell.2015.03.017

79. Soufi A, Donahue G, Zaret KS. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell. 2012;151: 994–1004. doi:10.1016/j.cell.2012.09.045

80. Faure AJ, Schmiedel JM, Lehner B. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. Cell Syst. 2017;5: 471–484.e4. doi:10.1016/j.cels.2017.10.003

81. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016;44: D726–32. doi:10.1093/nar/gkv1160

82. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9: 215–216. doi:10.1038/nmeth.1906

83. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41: 827–841. doi:10.1093/nar/gks1284

84. Bierhoff H, Dammert MA, Brocks D, Dambacher S, Schotta G, Grummt I. Quiescence-induced LncRNAs trigger H4K20 trimethylation and transcriptional silencing. Mol Cell. 2014;54: 675–682. doi:10.1016/j.molcel.2014.03.032

85. Stender JD, Pascual G, Liu W, Kaikkonen MU, Do K, Spann NJ, et al. Control of proinflammatory gene programs by regulated trimethylation and demethylation of histone H4K20. Mol Cell. 2012;48: 28–38. doi:10.1016/j.molcel.2012.07.020

86. Evertts AG, Manning AL, Wang X, Dyson NJ, Garcia BA, Coller HA. H4K20 methylation regulates quiescence and chromatin compaction. Mol Biol Cell. 2013;24: 3025–3037. doi:10.1091/mbc.E12-07-0529

87. de Groot R, Lüthi J, Lindsay H, Holtackers R, Pelkmans L. Large-scale image-based profiling of single-cell phenotypes in arrayed CRISPR-Cas9 gene perturbation screens. Mol Syst Biol. 2018;14: e8064. doi:10.15252/msb.20178064

88. Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, et al. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell. 2005;123: 581–592. doi:10.1016/j.cell.2005.10.023

89. Joshi AA, Struhl K. Eaf3 Chromodomain Interaction with Methylated H3-K36 Links Histone Deacetylation to Pol II Elongation [Internet]. Molecular Cell. 2005. pp. 971–978. doi:10.1016/j.molcel.2005.11.021

90. Keogh M-C, Kurdistani SK, Morris SA, Ahn SH, Podolny V, Collins SR, et al. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell. 2005;123: 593–605. doi:10.1016/j.cell.2005.10.025

91. Bell O, Wirbelauer C, Hild M, Scharf AND, Schwaiger M, MacAlpine DM, et al. Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. EMBO J. 2007;26: 4974–4984. doi:10.1038/sj.emboj.7601926

92. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13: 613–626. doi:10.1038/nrg3207

93. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

94. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14: 144–161. doi:10.1093/bib/bbs038

CHAPTER 3

Loci specific epigenetic drug sensitivity

Abstract

Therapeutic targeting of epigenetic modulators offers a novel approach to the treatment of multiple diseases. The cellular consequences of chemical compounds that target epigenetic regulators (epi-drugs) are complex. Epi-drugs affect global cellular phenotypes and cause local changes to gene expression due to alteration of a gene chromatin environment. Despite increasing use in the clinic, the mechanisms responsible for cellular changes are unclear. Specifically, to what degree the effects are a result of cell-wide changes or disease related locus specific effects is unknown. Here we developed a platform to systematically and simultaneously investigate the sensitivity of epi-drugs at hundreds of genomic locations by combining DNA barcoding, unique split-pool encoding and single cell expression measurements. Internal controls are used to isolate locus specific effects separately from any global consequences these drugs have. Using this platform, we discovered wide-spread loci specific sensitivities to epi-drugs for three distinct epi-drugs that target histone deacetylase, DNA methylation and bromodomain proteins. By leveraging ENCODE data on chromatin modification, we identified features of chromatin environments that are most likely to be affected by epi-drugs. The measurements of loci specific epi-drugs sensitivities will pave the way to the development of targeted therapy for personalized medicine.

Introduction

The location of a gene on the chromosome is known to affect its expression. Position effect was first observed in Drosophila by Muller in 1930 [1,2] and intensively investigated afterward [3–5]. Many years after the original work in Drosophila, it is now well documented that gene expression levels are influenced by chromatin environment [2,6–10]. Chromatins play a key role in the regulation of gene expression and are responsible for cell maintenance and differentiation. Chromatin regulation is complex and is an area of active research. The three-dimensional structure of the chromatins plays an important role in gene regulation by controlling accessibility of transcriptional machinery and the spatial proximity of a gene from cis regulatory elements such as enhancers. In addition, the specific three-dimensional folding will change the spatial distribution of transcription factors and other regulatory molecules such as lncRNAs. The spatial proximity of these regulatory molecules then plays a key role in controlling gene expression patterns. The three-dimensional structure itself is highly correlated with specific histone and DNA modification patterns. Overall, the complex multi-layered regulation of chromatin on gene expression pattern causes each gene to exist in a unique chromatin environment that plays an important role in determining gene expression distribution, i.e. both its average level as well as population variability [6,9,11].

The proper regulation of gene expression is vital for health and dysregulation of gene expression is associated with a large number of pathologies. Advances in DNA sequencing allow the classification of the specific disease based on the underlying changes of gene expression, the basis of large parts of precision medicine approaches. Given the large knowledge that is accumulating on what changes in gene expression are associated with disease conditions, it is only

natural to attempt to correct these pathologies by modification of underlying gene expression patterns. This quest has long history with initial attempts related to antisense oligos [12]. Similarly, the discovery of RNA interference (RNAi) sparked many attempts to develop therapies based on RNAi with the goal of manipulating gene expression [13]. However, despite the conceptual simplicity, translating these concepts into therapy was challenging [14–16].

Given the influence of local chromatin environment on gene expression, strategies that target epigenetic regulators are being investigated. Two main strategies are the pharmacological use of epi-drugs to influence gene expression and targeted approaches for epigenetic editing. Pharmacological approach uses inhibitors to the readers/writers/erasers of epigenetic marks. The pharmacological approach that is being developed to address a wide range of diseases is continuously expanding [17–23]. Multiple targeting strategies for epi-drugs are being explored including specific loss and gain of function [24–31], synthetic lethality [32–35], and to overcome drug resistance [35–37]. A common theme across these strategies is the use of epi-drugs to manipulate gene expression patterns e.g. suppress oncogenes or activate tumor suppressor genes [38]. However, the precision of epi-drugs induced gene expression targeting, i.e. the fraction of overall changes to gene expression that are desired for therapy, is currently very low. This low precision limits the usability of epi-drugs [39–41]. The alternative strategy is based on targeted recruitment of epigenetic modulators into specific sites. CRISPR mediated sequence specific targeting of epigenetic regulators is used to cause changes in gene expression pattern of specific loci. [42–44] The key advantage of epigenetic engineering is their precision. However, many challenges have to be addressed before these approaches can be translated into the clinics.

Despite the popularity of the use of epi-drugs to cause changes in gene expression patterns, there are many unknowns resulting from gaps in existing measurement capabilities of the effect of

epi-drugs on gene expression. Epi-drugs change gene expression due to direct, locus-dependent changes, and indirect or nonspecific effects [38]. Existing approaches to identify direct effects of epi-drugs rely on a combination of RNAseq and multiple ChIPseq to show that gene expression changes are coupled to changes in local histone modification [45]. However, the reliance on ChIPseq makes this approach limited as these measurements are only semi-quantitative [46], it is often hard to interpret their functional effects on gene expression [47], and they are challenging to scale to large numbers of samples [48]. Therefore, it is currently impossible to rigorously identify changes to gene expression that are due to specific modifications to the chromatin environment and not a result of non-specific or indirect effects. Therefore, there is a gap in current capabilities of mapping the specific and local impact of drug manipulating chromatin modifiers on gene expression.

Here we developed new measurement technology for the Massive And Parallel Measurement of Epigenetic Drug Sensitivity (MAPMEDS). MAPMEDS is based on comparison of drug effect at specific locus compared to the drug effect of hundreds of other loci. Statistical comparison of the drug effects on expression of a reporter fluorescent protein allows the deconvolution of global effects and locus specific sensitivities. MAPMEDS utilizes DNA barcodes to pool together the time-consuming step of genomic position identification of the reporters. Split-pool approach enables mapping of DNA barcodes to individual reporter cell lines. Using MAPMEDS we demonstrate the widespread existence of loci specific sensitivities for three epi-drugs that target histones acetylation, DNA methylation and proteins with bromodomains. By leveraging ENCODE data on the chromatin environment in each location, we show what types of environments are more susceptible to the different epi-drugs. Overall, these results shed light on

how epi-drugs cause changes in gene expression, the information that can be used for development of more precise targeting strategies.

Result

Overview of MAPMEDS

MAPMEDS is based on two key innovations: 1) A split-pool strategy for the creation of cell lines that uses DNA barcodes to identify what DNA labeling each cell line has and what is the genomic integration position of that barcode (Fig. 3.1a-d). Each expression reporter incorporates unique DNA barcode into reporter cassette that contain identical promoter and fluorescent reporter. These barcodes serve as unique identifiers for mapping the genomic locations of reporters and revealing clonal identities without the need of individual genomic extraction. The use of split-pool encoding allows the mapping between isoclonal line expression and the DNA barcode and thereby connecting genomic information with expression measurements. 2) A new strategy that uses in-well controls and statistical tools enables the separation of gene expression changes into locus-specific and global non-specific (Fig. 3.1ef). The use of in-well controls minimizes batch effects and is a key to identify locus specific sensitivities. Collectively, these two steps allow the parallel creation of large number of reporter cell lines and their use to identify loci specific epigenetic drug sensitivities.

MAPMEDS utilizes DNA barcoded expression reporters integrated as single copy per cell. In order to generate founder cells with one copy of barcoded reporter, lentiviral transduction at low multiplicity of infection (MOI) was used. In short, we first created a library of lentiviral plasmids containing barcode, MspI restriction site, cytomegalovirus (CMV) promoter and GFP

variant, mClover [49] (Fig. 3.1a). The barcodes are 16-bp sequences of random Adenosine, Thymine and Guanine. Cytosine is excluded to keep barcode intact for future application to examine DNA methylation through bisulfite conversion. Barcoded lentiviruses were packed and transduced at MOI of ~0.01 into K562 cells, leukemia cell line with abundant available epigenetic profiles [50] (Fig. 3.1b). Reporter K562 cells were selected at 72 hours post-transduction by fluorescent activated cell sorter (FACS) to establish a pool of 3,000 founder cells. Founder cells were expanded for 2 weeks and split into two pools (Fig. 3.1c). One half was used to identify the genomic location using inverse PCR [6]. The other half was used to establish individual isoclonal lines through single cell sorting. Once the cell lines were established, we utilized a combinatorial pooling approach to combine all the cell lines into a small number of pooled samples that were used to identify the barcode identity, and hence the genomic integration site, of the ORFs for all cell lines (Fig. 3.1d). The result of this procedure is a library of clonal cell lines that can be used as a powerful resource to examine drug sensitivity at diverse epigenetic environments.

In order to isolate locus-specific changes from any global changes in gene expression patterns, we implemented a new strategy using in-well controls and statistical tools. A reference "non-specific" cell population was created by using a polyclonal population of multiple integration cells so that the overall population has thousands of integration sites of same reporter cassettes. Far-red fluorescent protein iRFP670 [57] was used to mark this reference population. We split these control cells and co-cultured with each target cell line in multi-well plates.

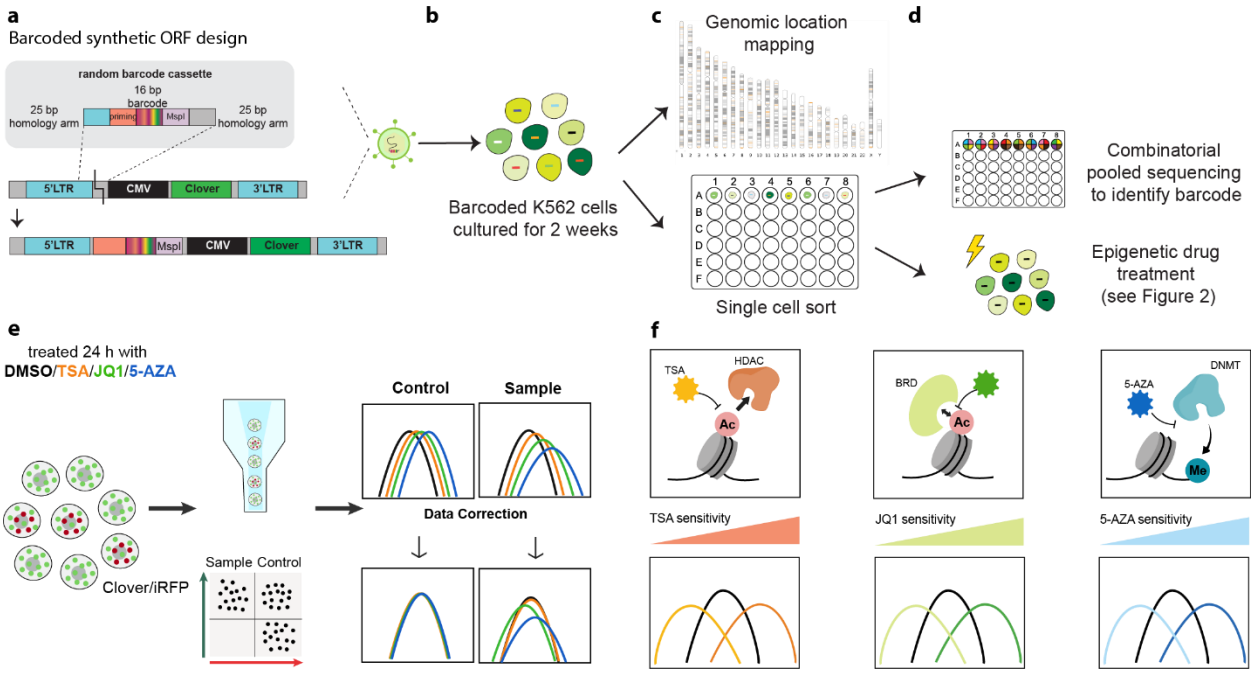


Figure 3.1 An overview of MAPMEDS

(a) Schematic structure of barcoded lentiviral constructs. The library vector contains short barcode, CMV promoter and mClover fluorescent protein as a reporter. The barcode is random 16-bp-long DNA with repeats of A,G and T. MspI restriction site is integrated upstream CMV promoter for genomic location mapping. **(b)** Barcoded lentivirus was packed and transduced into K562 cells at low MOI to create founder cells with singly integrated reporter. **(c-d)** Barcoded founder cells, selected by flow cytometry, were expanded for two weeks and split into two pools. Cells in the first pool were collected for locating reporter integration site. Founder cells in the second pool were sorted into 96-well plates to establish clonal cell lines. Barcode of each clone was simultaneously identified by split-pool encoding and deep sequencing. Library of characterized reporter clones is a useful resource to examine loci specific epigenetic drug sensitivity. **(e)** Loci specific effects were decoupled from global effects through mixing individual barcoded clones of interests with control cells expressing mClover and IRFP from multiple integration sites. Co-cultured cells were treated with TSA, JQ1 and 5'AZA for 24 hours. Expression of reporter proteins were measured by flow cytometers. Distribution of mClover expression in control cells was used to remove global effects of drugs. **(f)** A cartoon illustrating known mechanisms of actions of Trichostatin A, JQ1 and 5-Azacytidine.

Integration landscapes of reporters

To map the integration sites of reporters, we split half of founder cells into three sub-pools and further expanded the population. The first pool was used to reveal a list of genuine barcodes. We detected 756 candidate genuine barcodes after two weeks of culture. Two other sub-pools are technical replicates for locating reporter integration sites by an inverse PCR method coupled to paired-end high-throughput sequencing (Fig. 3.2a). In short, genomic DNA from each pool was isolated, digested with MspI enzyme and self-ligated. Barcodes and adjacent genomic DNA were amplified and deep sequenced. After barcode demultiplexing, genomic sequence was mapped to human genome assembly GRCh38. Genomic coordinate was assigned to corresponding barcode when mapped results from two replicates are matched. We observed reporter integration throughput the genome (Fig. 3.2b, Fig. 3S1a) with enriched pattern similar to previous study (Fig. 3S1b) [8]. Reporters were integrated at various genomic environments serving as diversified resources to study epigenetic drug sensitivity.

Scalability and Robustness of MAPMEDS

Pooled-sample sequencing is a cost-effective and practical strategy for many studies, especially the ones related to the discovery of rare mutation and single nucleotide polymorphism associated diseases [51–53]. Combinatorial pooled sequencing is an extension of standard pooled sampling where each sample exists in a few pools creating a many to many mappings between experimental conditions and samples. Combinatorial pooling improves the sensitivity and robustness of pooled sequencing since it includes built in error corrections. It reduces the cost and time for library preparation exponentially [54–56].

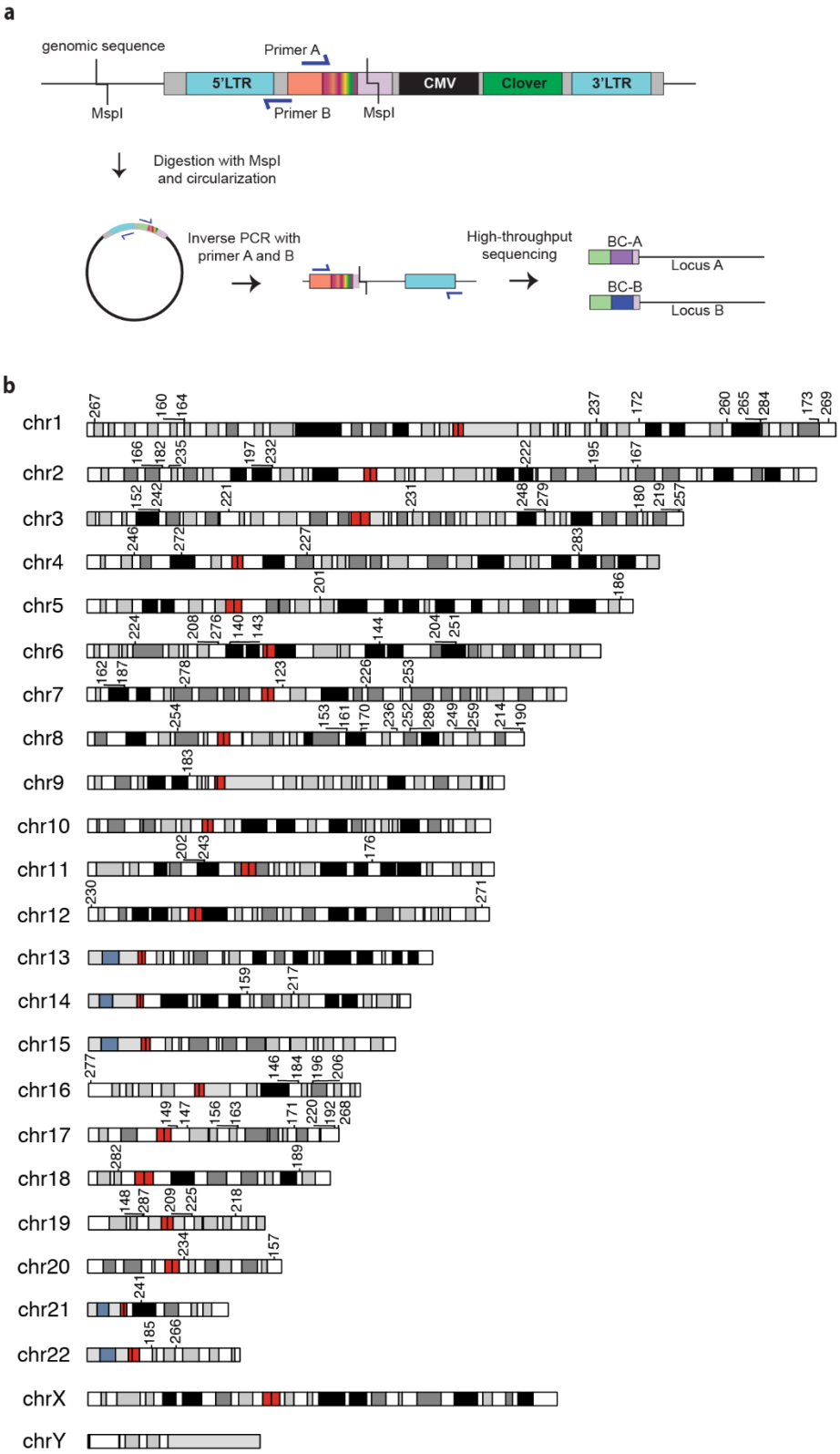


Figure 3.2: Diverse insertion landscapes of barcoded reporter

(Legend on next page)

Figure 3.2: Diverse insertion landscapes of barcoded reporter

(a) Reporter mapping by inverse PCR. Genomic DNA of founder cells was extracted, digested with restriction enzyme MspI and self-ligated to stitch barcode with its neighboring genome. Ligated product was amplified and followed by next generation sequencing. **(b)** Ideogram plot displaying reporter integration sites of individual clones in the library. Centromere position is indicated in red and stalk is marked in light blue. Heterochromatic region, which tend to be rich with adenine and thymine and relatively gene-poor, is represented by black and variation of grey. R-band in white on the ideogram is less condensed chromatin that is transcriptionally more active.

Conceptually, in combinatorial pooled sequencing, the identity of each sample is encoded in the composition of pools and this pooling pattern serves as a reference for decoding sequences belonging to corresponding sample (Fig. 3.3a). Clonal lines are mixed into few pools according to a predefined design. Each individual pool is genomic DNA extracted, barcode amplified, and index tagged using nested PCR. Amplicons from all pools are mixed together and sequenced. Subsequently, sequencing data is sorted by barcodes and their appearance patterns determined by the well indexes. The measured patterns are compared with the design pooling signatures to match barcodes with clone numbers.

The use of combinatorial pooling circumvents the need for individual genomic extraction and PCR per clones which significantly reduces the cost of reagents and hands-on time for sequencing preparation. In our design, each clone is mapped into 4 selected pools from total of 18 pools and this set-up allows up to 18-choose-4 or 3060 sample identification. This approach is highly scalable because sample size can be easily increased by adjusting number of pool and bit of code. We checked the accuracy of barcode identification using combinatorial pool sequencing by targeted PCR and Sanger sequencing. For 5 randomly chosen clones, the barcode deconvolution was all corrected which confirm the robustness of our method.

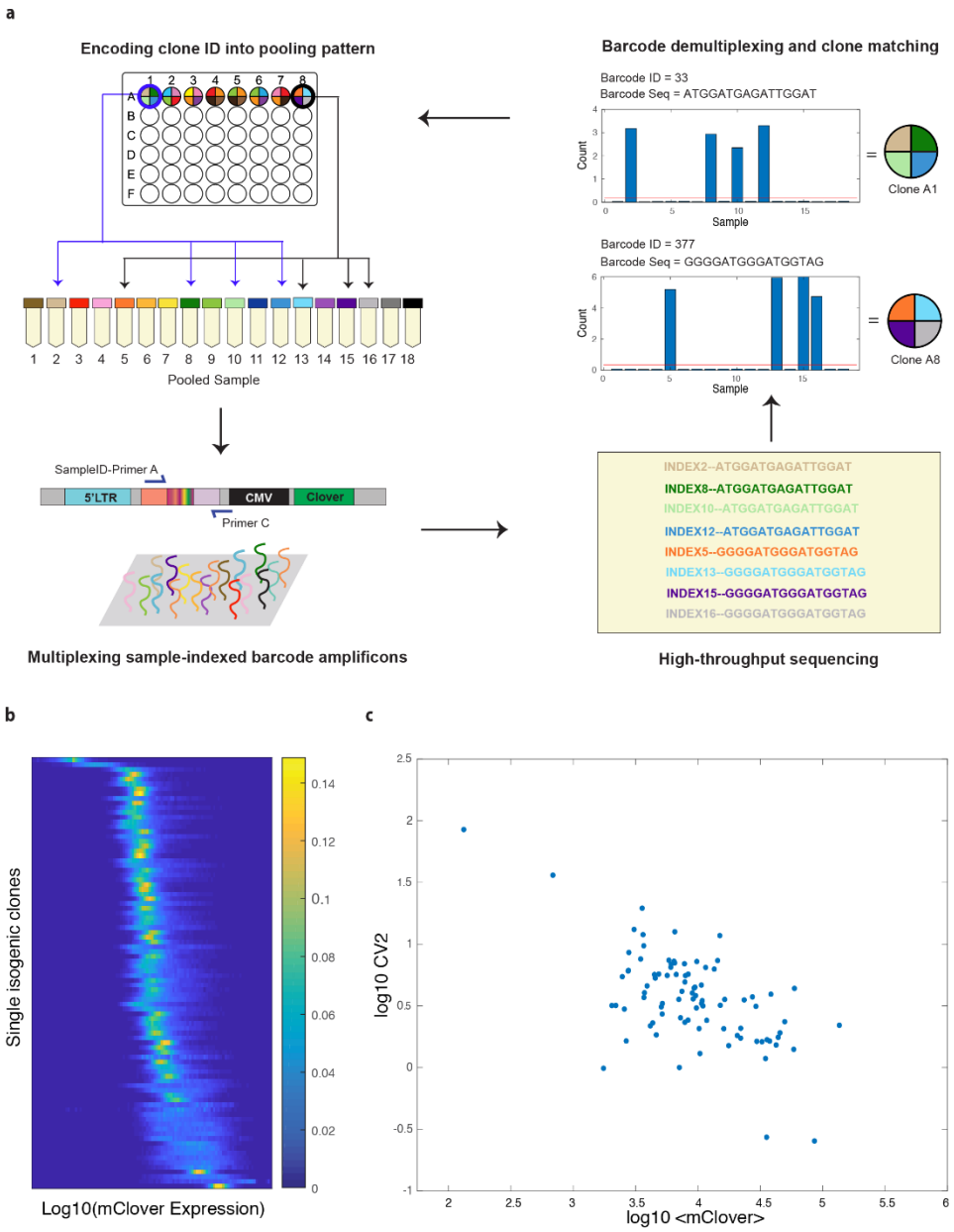


Figure 3.3: Combinatorial pooling massively and parallel identify barcode of individual clones.

(a) The combinatorial pooled sequencing involves encoding and decoding steps. Each individual clone was split into 4 out of 18 pooled samples according to the designs. Genomic DNA from each pooled sample was extracted and barcodes were amplified and labeled with 6-bp sample index. Amplicons from all samples were mixed together and prepared for NGS. Detected barcodes were deconvoluted to match barcode ID with clone ID. **(b)** The expression distribution of mClover protein was measured by high-throughput flow cytometer and displayed by stacked probability density function. Each row in the heatmap represents a single histogram from a single clonal line with the probability density function color-coded. **(c)** A scatter plot of reporter expression noise, measured as log-transformed squared coefficient of variation (CV^2), and mClover mean across all examined positions affirms unique chromatin environments across the genome.

Chromosomal Position Effects on Protein Expression Distribution

To support the notion that each integration site exists in a different chromatin environment, we examined the positional effect of our expression reporters. Across nearly a hundred positions, we observed variable expression distribution of mClover protein (Fig. 3.3b). Some clones are completely silent while many are highly expressed with expression average of approximately 1000-fold higher than the lowest-expressing cells (Fig. 3.3c). This variation range is comparable to a study that measured averaged mRNA expression across mouse genome [6], but higher than other studies carried out in bacteria or yeast [7,9,10]. Broader magnitude of positional effect in expression level detected in mammalian genomes may come from their large and complex genome organization and the discrepancy in experimental design and techniques.

Our observation on variable expression distribution confirms that the location where reporter gene inserted affects its expression and differences in genomic landscapes and epigenetic profiles are suggested to explain such differential expression in several studies. For example, lamina-associated domain and chromatin compaction significantly attenuates transcriptional activity [6] and certain histone marks, including H3K36me3, are correlated with expression level of the reporters [9,10]. Intuitively, epigenetic drugs that target chromatin regulator should also be impacted by distinct genomic environments. However, such loci specific sensitivity has not been previously measured and this motivates us to systematically measure positional effects on the sensitivity of epigenetic drug using our library of isogenic clones established and characterized by MAPMED.

Epigenetic drugs show position-dependent sensitivity

As a proof of concept, we chose three epigenetic drugs representing three mechanisms of inhibition. Trichostatin A (TSA) is histone deacetylases (HDACs) inhibitor [58,59] and effective in the treatment of several types of cancer including promyelocytic leukemia [60], lung cancer [61], breast cancer [62] and also enhancing the response of chemotherapy of multiple cancers [63,64]. JQ1 is a small-molecule inhibitor of BRD2 and BRD4, members of the bromodomain and extra-terminal domain (BET) protein family. JQ1 competitively blocks the binding of bromodomain proteins and acetylated chromatin which results in transcriptional attenuation [65,66]. JQ1 has been reported as a promising cancer therapeutic strategy in several cancers [67–69]. Azacytidine is a cytidine analogue that inhibits DNA methylation through loss of DNA methyltransferase (DNMT) activity [70]. Azacitidine was approved by the U.S. Food and Drug Administration (FDA) for the treatment of all subtypes of myelodysplastic syndrome (MDS) since 2004 [71]. These three drugs were added to co-cultured cells and incubated for 24 hours before sample collection and fluorescent quantification.

The expression level of mClover and iRFP was measured using high throughput sampler (HTS) flow cytometer. Cells from each sample were separated into reference and target cells based on iRFP gating. To quantify the site-specific effect, we normalized the expression levels of the sample cells by the average change between the drug and DMSO in the control population. After correction, the site-specific effect of a drug that are not captured by reference population were quantified by the Kolmogorov–Smirnov (KS) statistics that compares distribution similarity between sample cells under drug and sample cells with DMSO (Fig 3.4a). Any effects that are only site specific are effectively measured as changes beyond those also occur in the reference polyclonal population. Differential magnitudes of drug sensitivity were observed across examined

locations. Some loci, such as clone 224 and clone 260, show indifferent distribution of mClover in all conditions, but many positions are either hyposensitive or hypersensitive to at least one of epigenetic drugs (Fig.4b). These non-uniform responses indicate unique chromatin environment at each genomic location.

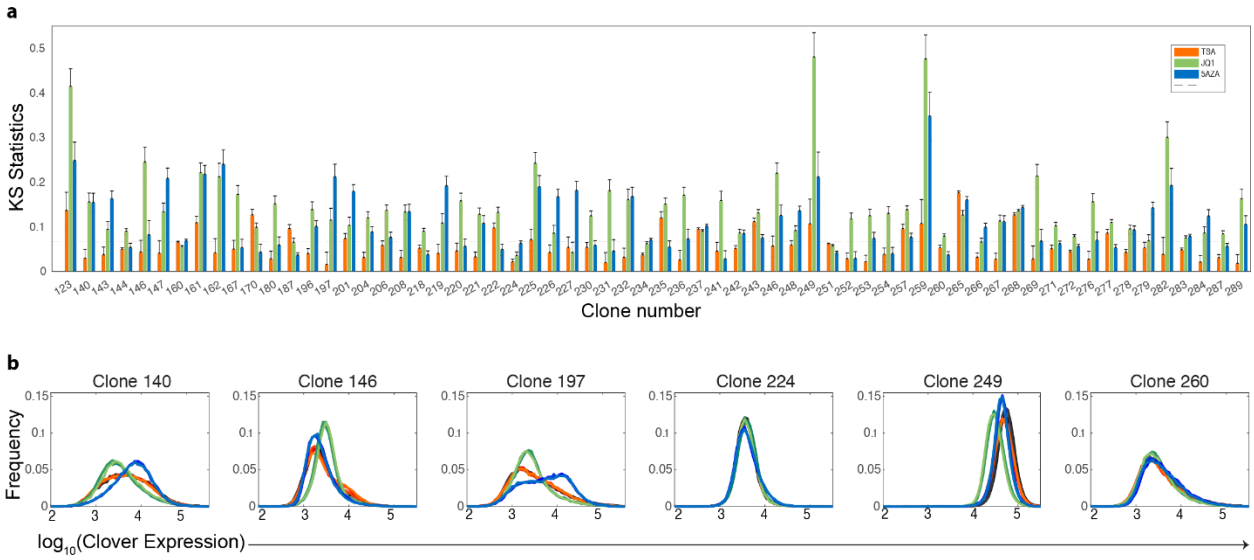


Figure 3.4 Chromosomal position effects influence magnitudes of epigenetic drug effects.

(a) A bar graph of calculated the test statistic of two-sample Kolmogorov–Smirnov test between epigenetic drug treatment and DMSO. (b) Histogram plot showing mClover distributions of selected clones after 24-hour treatment of DMSO (Black), TSA (Orange), JQ1 (Green) and 5'Azacytidine (Blue).

The distributions of KS statistics in control cells fall within the three standard deviation limits. (Fig. 3S2). Therefore, we used this criterion to identify the number of 'hits' per drug as it provides a non-parametric estimate of our false discovery rate to be <0.015. Our drug screening shows ~16% and ~13% of the positions that have down-regulated and up-regulated reporter expression after TSA treatment. The number of hits in TSA screen is lower than those in JQ1 and 5-Azacytidine. The percentage of down-regulated and up-regulated loci are 64% and 30% in JQ1 and 40.6% and 29.7% in 5-Azacytidine respectively. Smaller fraction of TSA-sensitive sites

observed in our data agrees with other studies findings that HDAC inhibitors only target less than 10% of the genome [72–74]. Overall, our data demonstrates a chromosomal position effects on epigenetic drug sensitivity.

Analysis of histone modification profiles identifies chromosomal environments susceptible to JQ1 sensitivity

We compared our maps of epigenetic drug sensitivity to a collection of available epigenomic map from K562 cells focusing on histone modifications. ChIP-seq signals, expressed by fold change over control, in a window of 10 kb around barcodes of interest were considered. We note that histone modification profiles were mapped in K562 cells without any genetic engineering. Insertion of reporters potentially change the pre-existing epigenetic landscape or cause sequence-specific or protein-specific interactions between regional chromatin and synthetic ORF. However, previous studies suggest that integrated reporters generally do not perturb the chromatin landscape but adopt the local chromatin state [10,75]. Complex sequence- or protein-specific interactions between local chromatin and synthetic reporters were not observed in previous study as well [11].

After selecting top JQ1-sensitive clones with distinct downregulation and upregulation of reporter expression (Fig. 3.5a), we compared their epigenetic profiles. Interestingly, we found JQ1 sensitivity is associated with certain histone modifications (Fig. 3.5b). Structural study reveals the binding of JQ1 to the acetyl-lysine binding pocket of BET bromodomains [65]. Such competitive binding disrupts bromodomain/acetyl histone interaction and therefore transcriptional activation. Genomic regions enriched in acetylated histones were hypothesized to display higher magnitude of reporter downregulation. Indeed, we observed that clones with lower expression after JQ1 treatment show higher enrichment of H3K9ac and H3K27ac. Moreover, we also found that these

clones have significantly higher enrichment of H2A.Z. This data supports previous studies demonstrating that BRD2 interacts with H2A.Z to mediate transcription initiation [76,77] and thus suggests that our assay recapitulated well-established results.

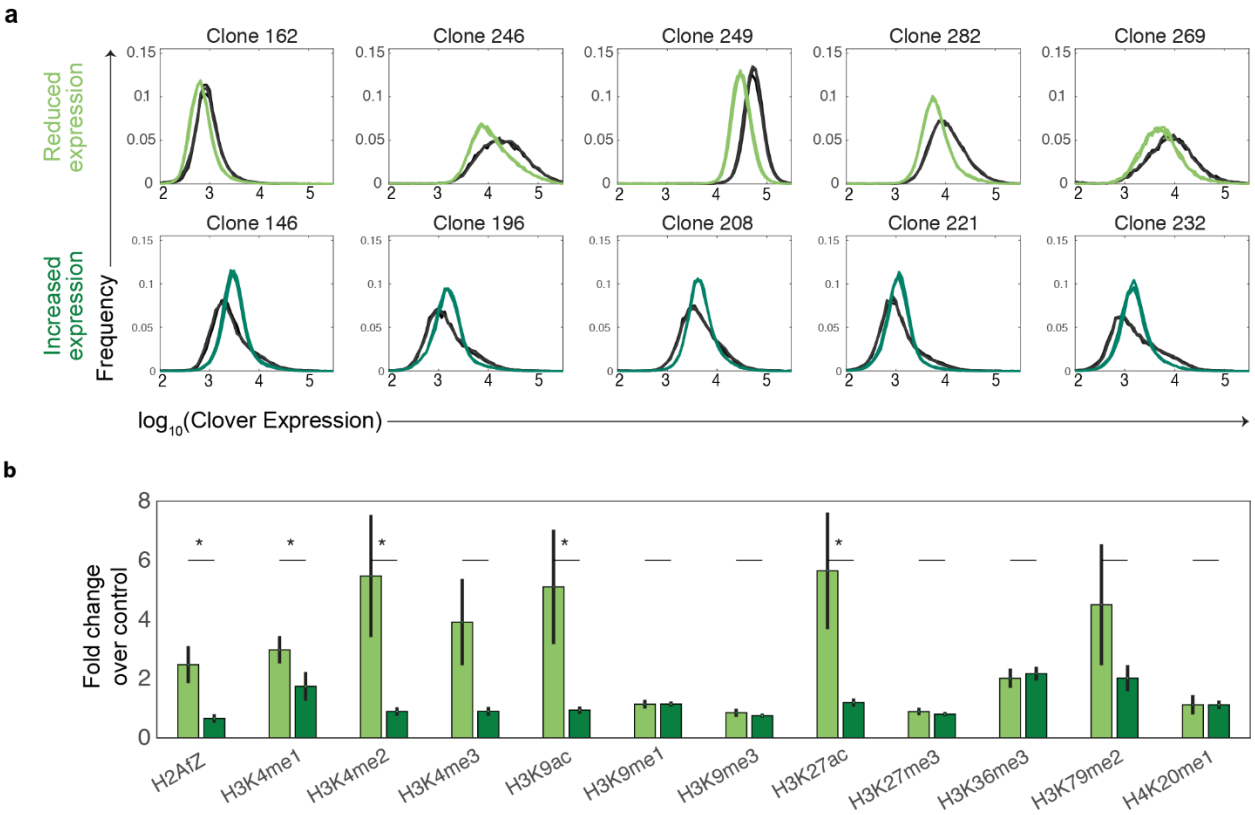


Figure 3.5: H2A.Z influences sensitivity of bromodomain inhibitor to expression alteration

(a) Examples of mClover distribution from clones showing significant reduced (top row) and increased (bottom row) mClover expression to JQ1 drug. **(b)** Bar graph shows histone enrichments for comparison of positions displaying reporter down-regulation (light green) versus those exhibiting up-regulation (dark green) after JQ1 treatment. Fold change over control of each histone mark was averaged within a window of 10 kb. The p-values were determined by two-sample t-test.

Differential sensitivity to 5-AZA treatment happened through DNA methylation-independent mechanism

Considering Azacytidine as a well-known inhibitor of DNA methylation, it is likely that genomic regions with hypermethylated promoter will response to such epigenetic drug the most. Moreover, CMV promoter used in our reporter is highly enriched in CpG sites which should be susceptible to DNA methylation. Therefore, we hypothesized that reporters inserted at diverse genomic environments will have different DNA methylation level and result in differential drug sensitivity. To test this hypothesis, we examined DNA methylation at CMV promoter, which contain 30 CG sites, using target bisulfite sequencing (Fig. 3.6a).

Surprisingly, even in cases where we observed distinct reporter downregulation and upregulation in two groups of top sensitive clones (Fig. 3.6b), their CMV promoters are mostly unmethylated and indistinguishable (Fig 3.6c). Our data suggests that differential sensitivity to 5-Azacytidine treatment may not result from the direct inhibition of promoter methylation but instead reflect changes in cell regulatory machinery that has stronger effects on these loci than average. This observation agrees with previous studies that report 50-60% of induced genes by 5-Aza-CdR did not have CpG islands within their 5'region [78,79]

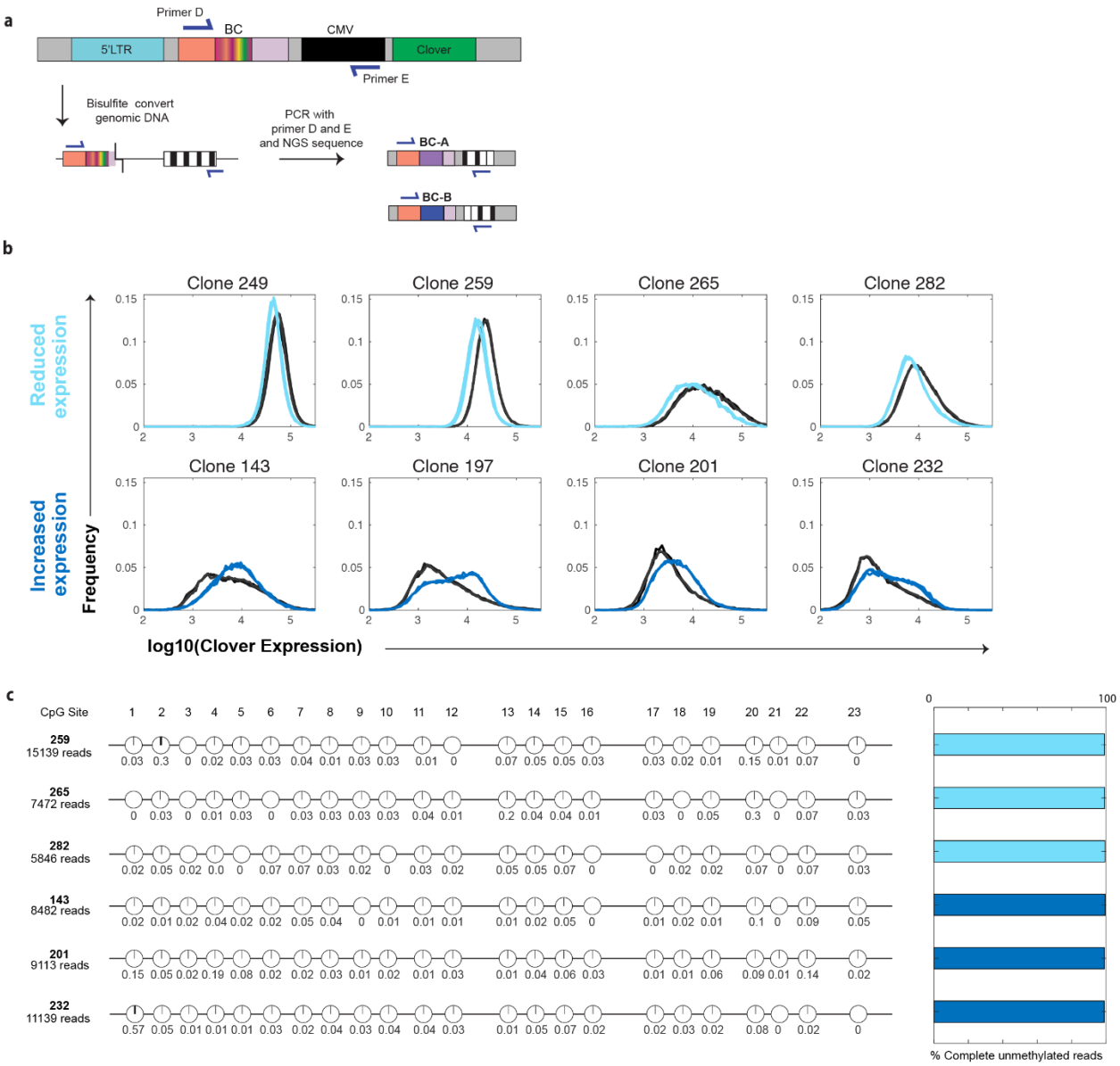


Figure 3.6: Chromosomal position effects 5 Azacytidine sensitivity through DNA methylation-independent mechanism.

(a) DNA methylation profiles of selected clones were simultaneously revealed by bisulfite conversion, targeted PCR and deep sequencing. **(b)** Examples of mClover distribution from clones with reduced expression (top row) and induced (bottom row) to 5 Azacytidine. **(c)** Locus-specific bisulfite sequencing of the CMV promoters of representative clones exhibiting hyposensitivity and hypersensitivity to 5 Azacytidine.

Discussion

Here we developed a new method MAPMEDS that can identify loci specific drug sensitivity in a robust and scalable manner. The method uses DNA barcoding to generate cell lines with known integration position of an expression reporter at genome scale and includes a statistical procedure to quantify the locus specific effect that a drug has on gene expression. We used MAPMEDS to evaluate the position specific effects of three common epi-drugs. We found that up to 60% of positions change in a manner that is different than average including both hyper and hypo sensitivity. Through analysis of the chromatin features that are enriched in sites with hyper/hypo drug sensitivities we were able to characterize what aspects of chromatin environment make it more (or less) sensitive to a specific drug. The development of MAPMEDS have both translational and basic science implications.

Locus-specific sensitivity measurements will support the development of new treatment strategies that use existing drugs and the development of new, more precise drugs. Identification of correct dosage of epi-drugs is challenging [80–82]. Comparison of the dose-response curve of the overall change in gene expression to the dose-response changes in gene expression that are due to locus-specific modification will help identifying drug concentrations that maximally impact target genes while limiting non-specific effects. Similarly, it will be possible to improve the precision of drug combinations [19,83]. New drugs and lead compounds could be identified based on predefined desired locus-specific changes in expression patterns. For example, in breast cancer that develops resistance to PI3K therapy, it was shown that co-treatment with a bromodomain inhibitor JQ-1 helps mitigate drug resistance since it silences the compensatory upregulation of RTKs [84]. The use of JQ-1 to achieve such desired effect is limited by the fact that JQ-1 has very

broad effects on expression changes across the genome. The tools we propose to develop will allow screening for new compounds and JQ-1 derivatives that maximizes the effects on the desired genes such as EGFR and INSR while limiting other undesirable changes.

Precision medicine is based on stratification of patients and assigning specific therapies based on the molecular information often associated with the key aberrant pathways. Therefore, in many cases, the nature of needed changes in gene expression are known. Treatment is limited due to the lack of therapies that can cause such desired changes in gene expression. Current manipulations of gene expression patterns using epigenetic targeting drugs such as JQ-1 are imprecise [39,40]. The new measurement technology developed here will support the future development of more precise epigenetic drug-mediated gene expression targeting. The reduction in side effects and the ability to screen for locus-specific changes in gene expression will lead to a large array of new therapies.

From basic science perspective, our understanding of chromatin regulation of gene expression is far from complete. MAPMEDs has the capacity to generate large datasets that look at changes across many positions and epigenetic drugs. Such large dataset will provide key insights into how changes in the local chromatin environment can affect gene expression in a manner isolated from any global changes. The use of DNA barcodes as part of MAPMEDS enables the pooled measurement of changes to local chromatin environment at scale. These data will provide invaluable insight into chromatin regulation.

MAPMEDS have two unique aspects that set it apart from other measurement approaches such as TRIP and BHIV [6,8] that aimed at measuring positional effects at scale. Unlike other approaches, MAPMEDS is based on the library creation of cell lines. The use of library of cell lines provides single cell data on changes in expression variability. Indeed, many of the locus

specific drug effects we saw did not simply shifted the population but changes the shape of the distribution. These effects would have been missed with bulk population measurements. Additionally, once the initial work in creating the cell line library was invested, the measurement of the drug effects is straightforward and can be scaled to large drug libraries. Other approaches such as TRIP and BHIV will require full RNA sequencing for each drug tested. The downside of the cell line library approach is that it is hard to scale it to more than a few hundred sites. However, recent advances in in-cell barcodes make it possible to generate the cell line library in pooled format, opening the way to a few orders of magnitude increase in the number of positions that can be measured. Future development of MAPMEDS to include these in-cell barcodes will further increase its utility as a platform for the discovery of more precise and useful epi-drugs.

Acknowledgements

We are thankful to Robert Foreman for his help with mapping of reporter integration sites. This work was funded by NIH grants to RW EY024960 and GM111404. T.Z. was also supported by Thailand Her Royal Highness Princess Maha Chakri Sirindhorn fellowship.

Author Contributions

TZ and RW conceptualized the experiments and data analysis.TZ performed the experiments and performed data analysis. TZ and RW wrote and edited the paper.

Declaration of Interests

The authors declare no competing interests.

Methods

Cell Lines and cell culture

The human K562 cells (Sigma-Aldrich) were grown at 37 °C in RPMI 1640 medium (Gibco) supplemented with 10% FBS (Gibco), 1% penicillin-streptomycin (Gibco) and 1% GlutaMAX (100x) (Gibco) under a 95% air and 5% CO2 atmosphere.

Construction of library reporter plasmid

The based plasmid without barcode was first constructed to contain the following elements. Lentiviral production units include HIV-1 truncated 5' LTR, HIV-1 packaging signal, HIV-1 Rev response element (RRE), HIV-1 truncated 3' LTR and Central polypurine tract (cPPT). These components allow proper viral packaging and viral integration into host cells. As a transcription unit, we used cytomegalovirus promoter (CMV) to drive expression of the reporter gene encoding yellow-green fluorescent protein (mClover). Woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) is placed after mClover to enhances mRNA stability and protein yield. Ampicillin resistance gene (β-lactamase) is included for selection of plasmid in bacterial cells.

To generate barcoded plasmid libraries, based lentiviral plasmid was cut upstream of the CMV promoter by ClaI restriction enzyme and purified by ethanol precipitation. The inserted cassette of 127-bp-long oligonucleotide containing a random 16-bp-long barcode sequence (repeats of A,T and G), MspI site, primer priming site and homology arms, were synthesized by Integrated DNA Technology. The assembly reaction of 1:5 vector:insert ratio was carried out for 1 hour at 50C using NEBuilder HIFI DNA assembly kit (New England Biolabs, NEB). Assembly products were electroporated into NEB Turbo Competent E.Coli (NEB) and then plated on ampicillin-containing medium. Ampicillin resistant colonies were collected and extracted for

plasmids using Maxiprep kit (Invitrogen). Ten sampling clones from the agar plate were analyzed by PCR and Sanger sequencing to verify successful cassette insertion and barcode diversity (Primer details in Table S1).

Generation of founder cell library and cell lines

Barcoded reporter and third generation lentiviral packaging plasmids were transfected into HEK 293T cells to generate a library of barcoded lentivirus. Viral supernatant was collected and concentrated by Lenti-X-concentrator (Takara) at 48-hour post transfection. K562 cells were transduced with barcoded virus in cultured media supplement with 5 μg/ml polybrene and 20mM HEPES for 2 hours of spinoculation and 24 hours of incubation. An m.o.i of approximately 0.01, corresponding to 1% infectivity estimated by flow cytometer, was used to ensure that the majority of cells were labeled with single barcode per cell. Founder cells were selected by fluorescence-activated cell sorting (FACS) at 72 hours post transduction. Founder cells were expanded for two weeks and split into two pools. In the first pool, cells were subject to mapping the genomic location of barcoded reporter. In the second pool, cells were single-cell sort to establish cell lines of unique barcode.

Identification of genuine barcode list

A library of genuine barcodes in founder cells was first listed. Briefly, barcode region was amplified in first nested PCR from 5 μg of genomic DNA in 50 ul of 20 cycle PCR reaction using Titanium Taq. Barcode amplicons were enriched from genomic DNA using SPRI beads (Beckman Coulter) and further amplified in the second nested PCR for 20 cycles. Illumina adapter was attached to final amplicon, amplified and sequenced on Illumina HiSeq 3000 platform (1x50bp). Sequencing reads were filtered and analyzed using Matlab Bioinformatics Toolbox. To identify

genuine barcode, we used the following algorithm. First, we sorted barcodes according to their counts from most frequent to least frequent. Then, mutant versions of each barcode, defined as barcodes within a Hamming distance of 2, were sequentially removed. We consider remaining sequences as ‘‘genuine’’ barcodes. We recovered 756 genuine barcodes from 3,000 sorted founder cells.

Mapping of reporter integration sites

Mapping of reporter integration sites was done by inverse PCR coupled with high-throughput sequencing. Briefly, founder cells were collected and splitted into two replicates. For each replica, 2 µg of genomic DNA was digested with 20 units of MspI (NEB) overnight at 37C in a volume of 100 µl. Subsequently, three sets of ligation reactions were set up by incubating 600 ng of purified digested DNA with 2 µl of high-concentration T4 DNA ligase (NEB, M0202T) overnight at 4C in a volume of 400 µl. The ligation reactions were purified by phenol-chloroform isoamyl alcohol extraction and ethanol precipitation. DNA pellets were dissolved in 30 µl of water. Two rounds of PCR were performed to amplify and enrich fragments containing both the barcodes and flanking genomic DNA regions (Primer details in Table S1). For the first round of nested PCR, five sets of 25-cycle reaction in a volume of 50 µl were performed using Phusion Hot Start Flex 2X Master Mix (NEB) and 5 µl of ligated products as templates. Amplicon was pooled together, cleaned by DNA Clean & Concentrator kit (Zymo), and diluted in 50 ul of water. For the second round of nested PCR, four sets of 15-cycle reaction in a volume of 50 ul were done with 5 ul of cleaned amplicon from first PCR. Purified sample was further ligated with Illumina adapter, amplified and sequenced on Illumina HiSeq 3000 platform (2x150bp). Sequencing reads were filtered and analyzed using Matlab Bioinformatics Toolbox. The genomic regions associated with genuine barcodes were extracted from mapping reads and aligned against the human genome (hg38)

using STAR [85]. Detected integration sites from each replicate were compared and assigned to each genuine barcode only if top candidate site from both replicates are identical. Mapping of reporter integration sites were plotted on the ideogram (Fig.2b) using R and karyoploteR package [86]. Genome coordinate of reporter integration site was converted to human reference genome(hg19) using UCSC liftOver tool (Kuhn et al., 2013) for comparison to ChIP-Seq data.

Combinatorial pool sequencing

Identity of reporter cell lines, linked by DNA barcodes, were simultaneously revealed in a single run using combinatorial pooled sequencing. Clonal numbers were encoded in a form of pooling pattern. To increase decoding accuracy, we designed pooling signature to be unique four selected pools out of total eighteen pools. Cells from each clone were split into four pools according to the design. Sequentially, genomic DNA from individual pool of mixed clones was extracted and used as templates for PCR to amplify barcode using same procedure described in the method of identification of genuine barcode list. Forward primers of second nested PCR contain 6-bp index DNA to label PCR products from each pool, which allow high-throughput multiplex sequencing (Primer details in Table S1). Sequences were filtered and demultiplexed using Matlab Bioinformatics Toolbox. Genuine barcodes from all pools were first listed. For each detected barcode, normalized counts per pools were calculated and pools showing high reads above the threshold were identified. Barcodes with four detected pools were first assigned to the clone showing matched pooling design. Some barcodes were found in more than four pools when sister cells, expanded from one founder cell, were sorted into multiple wells during single-cell sort. A list of merged pooling signature of two unassigned clones was matched with barcodes showing complexed readout. Clones with two inserted barcodes (~2% of the population) were excluded from the library of reporter cell lines.

Epigenetic drug treatment

We first created control cells expressing both mClover and IRFP670 from multiple integration sites. Briefly, K562 cells were transduced with CMV-IRFP670 lentivirus at high m.o.i. and sorted for IRFP positive cells. Lentiviral transduction of CMV-mClover was followed and dual reporter cells were selected by FACS. Control cells were co-cultured with individual reporter clones. 0.5 million cells of mixed samples were separately treated with DMSO, 400 nM of TSA, 5 μM of 5' Azacytidine and 1 μM of JQ1 for 24 hours. Afterward, cells were collected and measured for expression distribution of mClover and IRFP using BD FACSCelesta flow cytometer. Experiments were done in three replicates per drug treatment per clone. IRFP expression was used to separate control cells from sample cells. To eliminate non loci-specific effects, log10-transformed mclover expression of control cells with epigenetic drug treatment was calibrated to match corresponding expression in DMSO condition. Same adjustment was applied to reporter cells in the same well. the Kolmogorov–Smirnov test was performed by Matlab software to compare histogram similarity of mClover distribution under different conditions.

Validation of Combinatorial pool sequencing

For the validation of combinatorial pool sequencing, 5 clones were randomly chosen and two of them are sister clones, sharing same barcode. Genomic DNA (200ng) was used as a template for amplification with a set of validation primers (Primer details in Table S1). PCR products were cleaned by DNA Clean & Concentrator kit (Zymo) and Sanger sequenced to verify the barcode sequences.

Analysis of ChIP-Seq data

Following ChIP-seq data sets were downloaded from ENCODE: H2A.Z (ENCFF191EXE), H3K4me1 (ENCFF526QTS), H3K4me2 (ENCFF118MMT), H3K4me3 (ENCFF715DGL), H3K9ac (ENCFF602QRW), H3K9me1 (ENCFF526UWC), H3K9me3 (ENCFF834YLI), H3K27ac (ENCFF010PHG), H3K27me3 (ENCFF445UCR), H3K36me3 (ENCFF678IWR), H3K79me2 (ENCFF003CLZ) and H4K20me1 (ENCFF143CUR). Fold change over control signals were averaged within a window of 10 kb centered around integration sites using R and Bioconductor packages.

Methylation analysis of CMV promoter

To assess the levels of DNA methylation of CMV promoter, targeted bisulfite conversion was performed. Genomic DNA from clones of interest was extracted and bisulfite converted with Qiagen's EpiTect bisulfite kit according to manufacturer's instructions. Bisulfite converted genomic DNA was used as a template for two rounds of nested PCR using Invitrogen's Phusion U polymerase (Primer details in Table S1). Amplicons from all samples were pooled together, purified by DNA Clean & Concentrator kit (Zymo) and deep sequenced. Sequencing reads were filtered and demultiplexed by barcode using Matlab Bioinformatics Toolbox. Changes in cytosine base at CpG sites were converted into a matrix of methylation status.

References

1. Muller HJ. Types of visible variations induced by X-rays inDrosophila [Internet]. Journal of Genetics. 1930. pp. 299–334. doi:10.1007/bf02984195

2. Henikoff S. Position-effect variegation after 60 years. Trends Genet. 1990;6: 422–426.

3. Weiler KS, Wakimoto BT. Heterochromatin and Gene Expression in Drosophila [Internet]. Annual Review of Genetics. 1995. pp. 577–605. doi:10.1146/annurev.ge.29.120195.003045

4. Ottaviani A, Gilson E, Magdinier F. Telomeric position effect: from the yeast paradigm to human pathologies? Biochimie. 2008;90: 93–107.

5. Elgin SCR, Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol. 2013;5: a017780.

6. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013;154: 914–927.

7. Bryant JA, Sellars LE, Busby SJW, Lee DJ. Chromosome position effects on gene expression in Escherichia coli K-12. Nucleic Acids Res. 2014;42: 11383–11392.

8. Chen H-C, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. Nat Struct Mol Biol. 2017;24: 47–54.

9. Chen X, Zhang J. The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. Cell Syst. 2016;2: 347–354.

10. Chen M, Licon K, Otsuka R, Pillus L, Ideker T. Decoupling epigenetic and genetic effects through systematic analysis of gene position. Cell Rep. 2013;3: 128–137.

11. Maricque BB, Chaudhari HG, Cohen BA. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. Nat Biotechnol. 2018; doi:10.1038/nbt.4285

12. Stephenson ML, Zamecnik PC. Inhibition of Rous sarcoma viral RNA translation by a specific oligodeoxyribonucleotide. Proc Natl Acad Sci U S A. 1978;75: 285–288.

13. Wood M, Yin H, McClorey G. Modulating the expression of disease genes with RNA-based therapy. PLoS Genet. 2007;3: e109.

14. Paroo Z, Corey DR. Challenges for RNAi in vivo [Internet]. Trends in Biotechnology. 2004. pp. 390–394. doi:10.1016/j.tibtech.2004.06.004

15. Dutta T. Challenges in siRNA/ shRNA delivery and development of RNAi therapeutics for cancer [Internet]. Journal of Bioequivalence & Bioavailability. 2010. doi:10.4172/0975-0851.1000065

16. Boudreau RL, Rodríguez-Lebrón E, Davidson BL. RNAi medicine for the brain: progresses and challenges. Hum Mol Genet. 2011;20: R21–7.

17. Schnekenburger M, Florean C, Dicato M, Diederich M. Epigenetic alterations as a universal feature of cancer hallmarks and a promising target for personalized treatments. Curr Top Med Chem. 2016;16: 745–776.

18. Pfister SX, Ashworth A. Marked for death: targeting epigenetic changes in cancer. Nat Rev Drug Discov. 2017;16: 241–263.

19. Raynal NJ-M, Da Costa EM, Lee JT, Gharibyan V, Ahmed S, Zhang H, et al. Repositioning FDA-Approved Drugs in Combination with Epigenetic Drugs to Reprogram Colon Cancer Epigenome. Mol Cancer Ther. 2017;16: 397–407.

20. Jones PA, Issa J-PJ, Baylin S. Targeting the cancer epigenome for therapy. Nat Rev Genet. 2016;17: 630–641.

21. Bertino EM, Otterson GA. Romidepsin: a novel histone deacetylase inhibitor for cancer. Expert Opin Investig Drugs. Taylor & Francis; 2011;20: 1151–1158.

22. Lee H-Z, Kwitkowski VE, Del Valle PL, Ricci MS, Saber H, Habtemariam BA, et al. FDA Approval: Belinostat for the Treatment of Patients with Relapsed or Refractory Peripheral T-cell Lymphoma. Clin Cancer Res. AACR; 2015;21: 2666–2670.

23. Mann BS, Johnson JR, Cohen MH, Justice R, Pazdur R. FDA approval summary: vorinostat for treatment of advanced primary cutaneous T-cell lymphoma. Oncologist. AlphaMed Press; 2007;12: 1247–1252.

24. Raynal NJ-M, Lee JT, Wang Y, Beaudry A, Madireddi P, Garriga J, et al. Targeting Calcium Signaling Induces Epigenetic Reactivation of Tumor Suppressor Genes in Cancer. Cancer Res. 2016;76: 1494–1505.

25. Rathert P, Roth M, Neumann T, Muerdter F, Roe J-S, Muhar M, et al. Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. Nature. nature.com; 2015;525: 543–547.

26. Zhu H, Bengsch F, Svoronos N, Rutkowski MR, Bitler BG, Allegrezza MJ, et al. BET Bromodomain Inhibition Promotes Anti-tumor Immunity by Suppressing PD-L1 Expression. Cell Rep. Elsevier; 2016;16: 2829–2837.

27. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell. Elsevier; 2016;164: 1073.

28. Karpf AR, Peterson PW, Rawlins JT, Dalley BK, Yang Q, Albertsen H, et al. Inhibition of DNA methyltransferase stimulates the expression of signal transducer and activator of transcription 1, 2, and 3 genes in colon tumor cells. Proc Natl Acad Sci U S A. National Acad Sciences; 1999;96: 14007–14012.

29. Puissant A, Frumm SM, Alexe G, Bassil CF, Qi J, Chanthery YH, et al. Targeting MYCN in neuroblastoma by BET bromodomain inhibition. Cancer Discov. AACR; 2013;3: 308–323.

30. Rocchi P, Tonelli R, Camerin C, Purgato S, Fronza R, Bianucci F, et al. p21Waf1/Cip1 is a common target induced by short-chain fatty acid HDAC inhibitors (valproic acid, tributyrin and sodium butyrate) in neuroblastoma cells. Oncol Rep. spandidos-publications.com; 2005;13: 1139–1144.

31. Brueckner B, Garcia Boy R, Siedlecki P, Musch T, Kliem HC, Zielenkiewicz P, et al. Epigenetic reactivation of tumor suppressor genes by a novel small-molecule inhibitor of human DNA methyltransferases. Cancer Res. AACR; 2005;65: 6305–6311.

32. Kadoch C. Lifting Up the HAT: Synthetic Lethal Screening Reveals a Novel Vulnerability at the CBP–p300 Axis. Cancer Discov. American Association for Cancer Research; 2016;6: 350–352.

33. Pfister SX, Markkanen E, Jiang Y, Sarkar S, Woodcock M, Orlando G, et al. Inhibiting WEE1 Selectively Kills Histone H3K36me3-Deficient Cancers by dNTP Starvation. Cancer Cell. Elsevier; 2015;28: 557–568.

34. Oike T, Ogiwara H, Tominaga Y, Ito K, Ando O, Tsuta K, et al. A Synthetic Lethality–Based Strategy to Treat Cancers Harboring a Genetic Deficiency in the Chromatin Remodeling Factor BRG1. Cancer Res. American Association for Cancer Research; 2013;73: 5508–5518.

35. Bitler BG, Aird KM, Garipov A, Li H, Amatangelo M, Kossenkov AV, et al. Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. Nat Med. nature.com; 2015;21: 231–238.

36. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell. 2010;141: 69–80.

37. Vinogradova M, Gehling VS, Gustafson A, Arora S, Tindell CA, Wilson C, et al. An inhibitor of KDM5 demethylases reduces survival of drug-tolerant cancer cells. Nat Chem Biol. nature.com; 2016;12: 531–538.

38. de Groote ML, Verschure PJ, Rots MG. Epigenetic Editing: targeted rewriting of epigenetic marks to modulate expression of selected target genes. Nucleic Acids Res. 2012;40: 10596–10613.

39. Azad N, Zahnow CA, Rudin CM, Baylin SB. The future of epigenetic therapy in solid tumours--lessons from the past. Nat Rev Clin Oncol. nature.com; 2013;10: 256–266.

40. Altucci L, Rots MG. Epigenetic drugs: from chemistry via biology to medicine and back. Clin Epigenetics. 2016;8: 56.

41. Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. Nat Biotechnol. 2010;28: 1069–1078.

42. Chen Z, Li S, Subramaniam S, Shyy JY-J, Chien S. Epigenetic Regulation: A New Frontier for Biomedical Engineers. Annu Rev Biomed Eng. 2017;19: 195–219.

43. Falahi F, Sgro A, Blancafort P. Epigenome engineering in cancer: fairytale or a realistic path to the clinic? Front Oncol. 2015;5: 22.

44. Mlambo T, Nitsch S, Hildenbeutel M, Romito M, Müller M, Bossen C, et al. Designer epigenome modifiers enable robust and sustained gene silencing in clinically relevant human cells. Nucleic Acids Res. 2018;46: 4456–4468.

45. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546: 431–435.

46. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9: e1003326.

47. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010;107: 2926–2931.

48. van Galen P, Viny AD, Ram O, Ryan RJH, Cotton MJ, Donohue L, et al. A Multiplexed System for Quantitative Comparisons of Chromatin Landscapes. Mol Cell. 2016;61: 170–180.

49. Lam AJ, St-Pierre F, Gong Y, Marshall JD, Cranfill PJ, Baird MA, et al. Improving FRET dynamic range with bright green and red fluorescent proteins. Nat Methods. 2012;9: 1005–1012.

50. Qu H, Fang X. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. Genomics Proteomics Bioinformatics. 2013;11: 135–141.

51. Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, et al. Quantification of rare allelic variants from pooled genomic DNA. Nat Methods. 2009;6: 263–265.

52. Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, et al. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet. 2010;42: 851–858.

53. Ezquerra-Inchausti M, Anasagasti A, Barandika O, Garai-Aramburu G, Galdós M, López de Munain A, et al. A new approach based on targeted pooled DNA sequencing identifies novel mutations in patients with Inherited Retinal Dystrophies. Sci Rep. 2018;8: 15457.

54. Patterson N, Gabriel S. Combinatorics and next-generation sequencing. Nat Biotechnol. 2009;27: 826–827.

55. Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, et al. DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 2009;19: 1243–1253.

56. Cao C-C, Sun X. Combinatorial pooled sequencing: experiment design and decoding. Quantitative Biology. 2016;4: 36–46.

57. Shcherbakova DM, Verkhusha VV. Near-infrared fluorescent proteins for multicolor in vivo imaging. Nat Methods. 2013;10: 751–754.

58. Xu WS, Parmigiani RB, Marks PA. Histone deacetylase inhibitors: molecular mechanisms of action [Internet]. Oncogene. 2007. pp. 5541–5552. doi:10.1038/sj.onc.1210620

59. Yoshida M, Horinouchi S, Beppu T. Trichostatin A and trapoxin: novel chemical probes for the role of histone acetylation in chromatin structure and function. Bioessays. 1995;17: 423–430.

60. Fenrick R, Hiebert SW. Role of histone deacetylases in acute leukemia [Internet]. Journal of Cellular Biochemistry. 1998. pp. 194–202. doi:3.0.co;2-h">10.1002/(sici)1097-4644(1998)72:30/31+<194::aid-jcb24>3.0.co;2-h

61. Platta CS, Greenblatt DY, Kunnimalaiyaan M, Chen H. The HDAC inhibitor trichostatin A inhibits growth of small cell lung cancer cells. J Surg Res. 2007;142: 219–226.

62. Vigushin DM, Ali S, Pace PE, Mirsaidi N, Ito K, Adcock I, et al. Trichostatin A is a histone deacetylase inhibitor with potent antitumor activity against breast cancer in vivo. Clin Cancer Res. 2001;7: 971–976.

63. Piacentini P, Donadelli M, Costanzo C, Moore PS, Palmieri M, Scarpa A. Trichostatin A enhances the response of chemotherapeutic agents in inhibiting pancreatic cancer cell proliferation. Virchows Arch. 2006;448: 797–804.

64. Zhang X, Yashiro M, Ren J, Hirakawa K. Histone deacetylase inhibitor, trichostatin A, increases the chemosensitivity of anticancer drugs in gastric cancer cell lines. Oncol Rep. 2006;16: 563–568.

65. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, et al. Selective inhibition of BET bromodomains. Nature. 2010;468: 1067–1073.

66. Shi J, Vakoc CR. The mechanisms behind the therapeutic activity of BET bromodomain inhibition. Mol Cell. 2014;54: 728–736.

67. Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J, Jacobs HM, et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Cell. 2011;146: 904–917.

68. Bid HK, Phelps DA, Xaio L, Guttridge DC, Lin J, London C, et al. The Bromodomain BET Inhibitor JQ1 Suppresses Tumor Angiogenesis in Models of Childhood Sarcoma. Mol Cancer Ther. 2016;15: 1018–1028.

69. Costa DD, Da Costa D, Agathanggelou A, Perry T, Weston V, Petermann E, et al. BET inhibition as a single or combined therapeutic approach in primary paediatric B-precursor acute lymphoblastic leukaemia [Internet]. Blood Cancer Journal. 2013. pp. e126–e126. doi:10.1038/bcj.2013.24

70. Christman JK. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy [Internet]. Oncogene. 2002. pp. 5483–5495. doi:10.1038/sj.onc.1205699

71. Kaminskas E. FDA Drug Approval Summary: Azacitidine (5-azacytidine, VidazaTM) for Injectable Suspension [Internet]. The Oncologist. 2005. pp. 176–182. doi:10.1634/theoncologist.10-3-176

72. Richon VM, Sandhoff TW, Rifkind RA, Marks PA. Histone deacetylase inhibitor selectively induces p21WAF1 expression and gene-associated histone acetylation [Internet].

Proceedings of the National Academy of Sciences. 2000. pp. 10014–10019. doi:10.1073/pnas.180316197

73. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, et al. Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions [Internet]. Science. 2009. pp. 834–840. doi:10.1126/science.1175371

74. Van Lint C Emiliani S Verdin E. The expression of a small fraction of cellular genes is changed in response to histone hyperacetylation. Gene Expr. 1996;5: 245–253.

75. Corrales M, Rosado A, Cortini R, van Arensbergen J, van Steensel B, Filion GJ. Clustering of Drosophila housekeeping promoters facilitates their expression [Internet]. Genome Research. 2017. pp. 1153–1161. doi:10.1101/gr.211433.116

76. Draker R, Ng MK, Sarcinella E, Ignatchenko V, Kislinger T, Cheung P. A Combination of H2A.Z and H4 Acetylation Recruits Brd2 to Chromatin during Transcriptional Activation [Internet]. PLoS Genetics. 2012. p. e1003047. doi:10.1371/journal.pgen.1003047

77. Vardabasso C, Gaspar-Maia A, Hasson D, Pünzeler S, Valle-Garcia D, Straub T, et al. Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma [Internet]. Molecular Cell. 2015. pp. 75–88. doi:10.1016/j.molcel.2015.05.009

78. Liang G, Gonzales FA, Jones PA, Orntoft TF, Thykjaer T. Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. Cancer Res. 2002;62: 961–966.

79. Schmelz K, Sattler N, Wagner M, Lübbert M, Dörken B, Tamm I. Induction of gene expression by 5-Aza-2′-deoxycytidine in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) but not epithelial cells by DNA-methylation-dependent and -independent mechanisms [Internet]. Leukemia. 2005. pp. 103–111. doi:10.1038/sj.leu.2403552

80. Shen H, Laird PW. In epigenetic therapy, less is more. Cell Stem Cell. 2012;10: 353–354.

81. Miousse IR, Murphy LA, Lin H, Schisler MR, Sun J, Chalbot M-CG, et al. Dose-response analysis of epigenetic, metabolic, and apical endpoints after short-term exposure to experimental hepatotoxicants. Food Chem Toxicol. 2017;109: 690–702.

82. Fardi M, Solali S, Farshdousti Hagh M. Epigenetic mechanisms as a new approach in cancer treatment: An updated review. Genes Dis. 2018;5: 304–311.

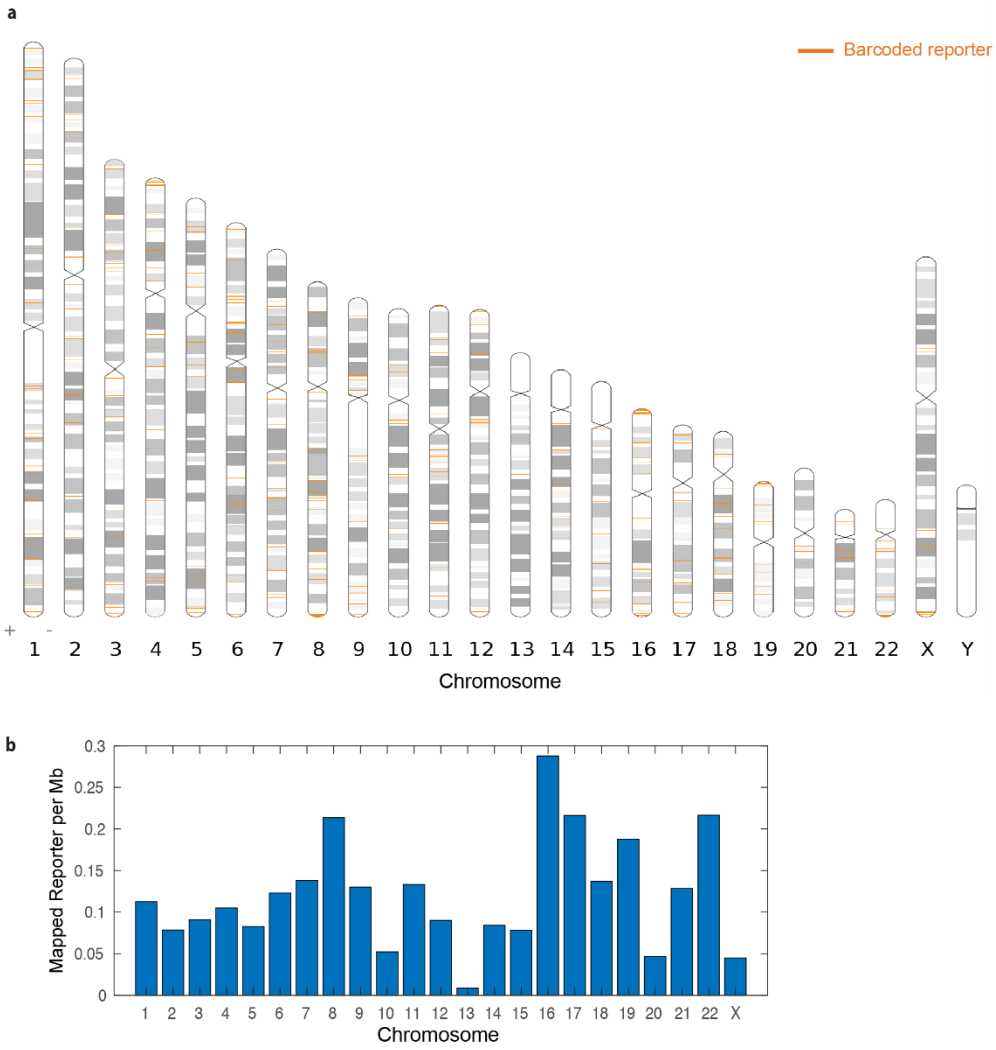
83. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. Elsevier; 2016;166: 740–754.

84. Stratikopoulos EE, Dendy M, Szabolcs M, Khaykin AJ, Lefebvre C, Zhou M-M, et al. Kinase and BET Inhibitors Together Clamp Inhibition of PI3K Signaling and Overcome Resistance to Therapy. Cancer Cell. 2015;27: 837–851.

85. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

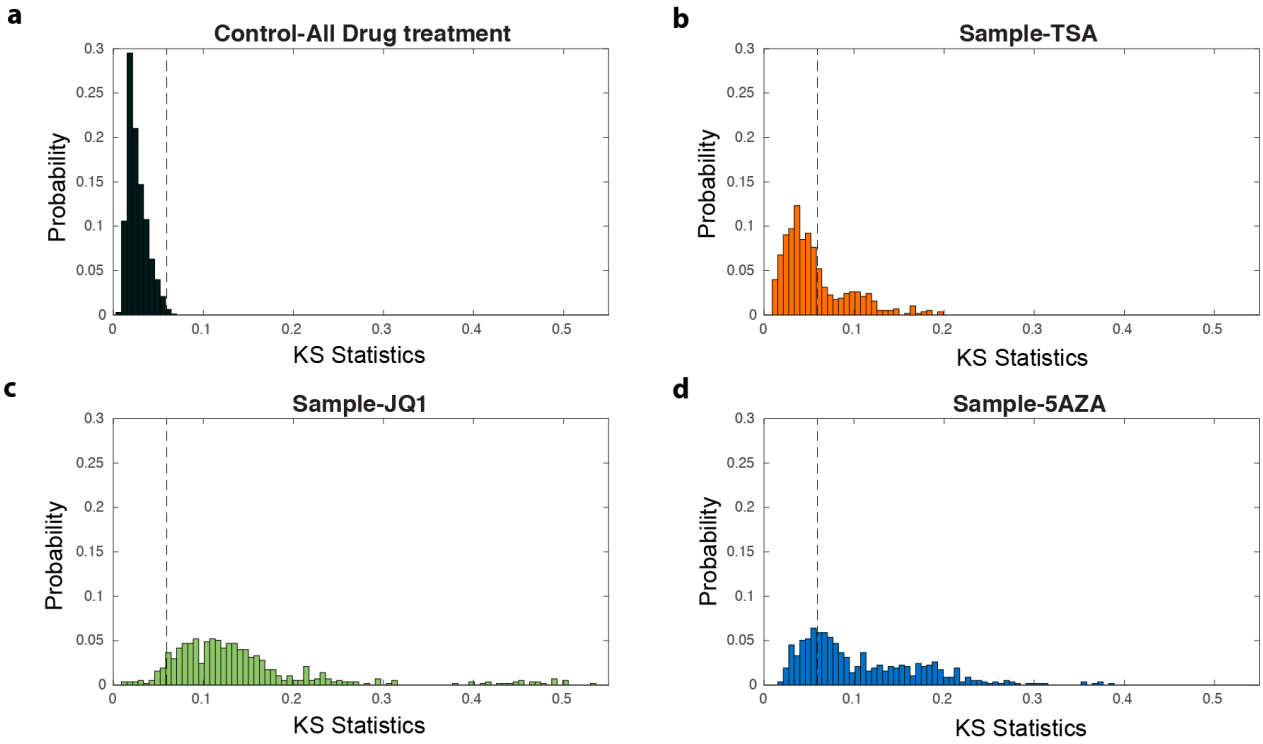
86. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data [Internet]. Bioinformatics. 2017. pp. 3088–3090. doi:10.1093/bioinformatics/btx346

Supplemental Figures



Supplementary Figure 3.1: Genome-wide map of barcoded reporter insertion

(a) Positions of mapped barcoded-CMV-mClover reporters along all chromosome. Each synthetic reporter is represented as an orange tick on the ideogram. We detected 756 candidate genuine barcodes from three thousand original founder cells after two weeks of expansion. **(b)** A bar graph representing the frequency of mapped reporter per megabase.



Supplementary Figure 3.2: Drug screening shows different number of hits for different epi-drug treatment

(a) The distributions of KS statistics in control cells fall within the three standard deviation limits. **(b-d)** The number of hits in the drug screen for TSA(b), JQ1(c) and 5-AZA(d) is determined by Kolmogorov-Smirnov statistics passing three standard deviation limits of control groups.

Supplemental Table

Name	Sequence	FW/RV	Description
DBC_Amp_FWD	GATCCTGTAGAACTCTGAACCT	F	Verify barcode insertion in plasmid library
DBC_Amp_REV	AGTCGGTGTCTTCTATGGAG	R	Verify barcode insertion in plasmid library
What_Nest1_FWD	TATGGATCCTGTAGAACTCTG	F	Identify geniune barcode sequence
What_Nest1_REV	GCTCTGCTTATATAGACCTCCCAC	R	Identify geniune barcode sequence
What_Nest2_FWD	TGTAGAACTCTGAACCTAGCT	F	Identify geniune barcode sequence
What_Nest2_REV	CGTAAGTTATGTAACGCGGA	R	Identify geniune barcode sequence
Where_Nested1_FWD	TATGGATCCTGTAGAACTCTG	F	Mapping reporter location. First Nest PCR.
Where_Nested1_REV	GCTTCAGCAAGCCGAGTCCTGCGTCGAG	R	Mapping reporter location. First Nest PCR.
Where_Nested2_FWD_A	AGTCATGTAGAACTCTGAACCTAGCT	F	Mapping reporter location, Second Nest PCR, Replication1
Where_Nested2_FWD_B	CTAGTTGTAGAACTCTGAACCTAGCT	F	Mapping reporter location, Second Nest PCR, Replication2
Where_Nested2_REV	GCTTTCAGGTCCCTGTTCGG	R	Mapping reporter location, Second Nest PCR, Replication1&2
CombiWhat_Nest2_FWD_A	ATCACGTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_B	CGATGTTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_C	TTAGGCTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_D	TGACCATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR

CombiWhat_Nest2_FWD_E	ACAGTGTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_F	GCCAATTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_G	CAGATCTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_H	ACTTGATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_I	GATCAGTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_J	TAGCTTTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_K	GGCTACTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_L	CTTGTATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_M	AGTCAATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_N	AGTTCCTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_O	ATGTCATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_P	CCGTCCTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_Q	GTCCGCTGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Nest2_FWD_R	GTGAAATGTAGAACTCTGAACCTAGCT	F	Combinatorial pooled sequencing, Second Next PCR
CombiWhat_Validate_F	TAGTGAACGGATCTCGACG	F	Validate combinatorial pooled sequencing
CombiWhat_Validate_R	GCTCTGCTTATATAGACCTCCCAC	R	Validate combinatorial pooled sequencing

Table 1 A list of primers used in this study

CHAPTER 4

Conclusion and future directions

Here we developed a novel method that combine two high throughput techniques together. One is Thousands of Reporters Integrated in Parallel (TRIP) which has high capacity of revealing the level of gene expression from thousands of genomic locations in a single run of deep sequencing. However, the drawback of this method is the sensitivity of data since it only provides populational average per specific location labeled by the unique barcode. To get single cell data, clonal establishment and characterization is required. Nevertheless, this process is very laborious and not scalable. We circumvented this problem by integrating combinatorial pooled sequencing with TRIP. Combinatorial pooled sequencing eliminates the process of individual genomic extraction and PCR amplification per clones which significantly decreases both the cost and time for sequencing preparation. We showed that our approach is highly robust, massive, parallel and scalable.

Our novel method allows massive and parallel establishment and characterization of clonal line of reporters which has numerous applications for addressing important questions yet still underexplored in biology. One interesting question is the effect of chromosomal position effects on gene expression variability in higher eukaryote genome. We used newly developed method to investigate the underlying factors controlling gene expression mean and variability in the human genome. We showed that the insertion site affects both features of gene expression. Mechanistic insights related to the factors underlying expression mean and variance noise were gleaned by

leveraging multiple epigenetic profiles with our measurement data. Overall, our results provide new insights into chromatin factors that contribute to regulation of gene expression and highlights the importance of chromosomal context in gene regulation.

Moreover, we also investigated the contribution of heterogenous chromatin environment on the sensitivity of epigenetic drugs using newly developed method. The position specific effects of three common epi-drugs targeting histone deacetylase (TSA), DNA methylation (5'Azacytidine) and bromodomain proteins (JQ1) were evaluated. We found that up to 60% of genomic positions change in a manner that is different than average including both hyper and hypo sensitivity. Through analysis of the chromatin features that are enriched in sites with hyper/hypo drug sensitivities, we were able to characterize what aspects of chromatin environment make it more, or less, sensitive to a specific chemical compound.

Future work for this dissertation should focus on applying other approaches to validate new finding. Our studies took advantage of big repository of well characterized and publicly available K562 epigenetic data. However, these profiles might not be perfectly identical to our K562 cells used in this study. Therefore, it will be beneficial to verify our novel finding using other approaches such as performing ChIP-qPCR experiments on reporter promoter and gene with epigenetic marks representing chromatin states or transcription factors of interest. Moreover, investigation of chromosomal position effects on more epigenetic drugs, promoters and cell lines will be useful to validate the generalization of our finding.

Overall, we demonstrated that our method has both translational and basic science implications. Here, we used gene expression noise and the sensitivity of epigenetic drug as proof-of-concept studies. However, our approach is highly flexible and can be easily adapted to probe the influence of chromatin environment on any biological processes that involve the use of genetic

and epigenetic information and their readouts need single cell data. For example, our method is a great tool to study the contribution of chromatin context on the efficiency of genome editing tool CRISPR. Deeper understanding of the factors that affect this process will lead to improvement on their efficacy and safety and further propel this technology toward therapeutic applications.