UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The Effects of Structural Variation on 3D Chromatin Structure

Permalink

https://escholarship.org/uc/item/2pf3c6r9

Author

Shanta, Omar

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

The Effects of Structural Variation on 3D Chromatin Structure

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Electrical and Computer Engineering

by

Omar Shanta

Committee in charge:

Professor Jonathan Sebat, Chair
Professor Siavash Mirarab, Co-Chair
Professor William Hodgkiss
Professor Lawrence Saul

2019

Copyright
Omar Shanta, 2019
All rights reserved.

The thesis of Omar Shanta is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2019

DEDICATION

To my parents for always supporting my academic endeavors.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures vi

Acknowledgements vii

Abstract of the Thesis viii

Chapter 1 Introduction 1

Chapter 2 Methods 3

Chapter 3 Results 5

Chapter 4 Discussion 12

References 14

LIST OF FIGURES

Figure 3.1: Hi-C bins near a deletion are separated into two categories: Span and Flank Regions 6
Figure 3.2: Linear Regression p-values indicate large deletions have the biggest impact on chromatin structure 7
Figure 3.3: Span region shows significant positive effect for large deletions when the deletion does not intersect a TAD boundary 8
Figure 3.4: Average chromatin structure for large deletions not present at TAD boundaries and present at TAD boundaries 9
Figure 3.5: Hi-C chromatin interaction counts of the 8p23.1 region shows a change in TAD and gene location at an inversion site 11

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Jonathan Sebat, for his guidance and invaluable insights throughout the project as well as the unique opportunities he has provided. I would also like to thank the members of the Sebat Lab and Iakoucheva Lab for all of their advice and feedback. I would like to thank the UCSD ECE program and the San Diego Supercomputer Center for all of the resources available to students. I would also like to thank my thesis committee: Professors Siavash Mirarab, William Hodgkiss, and Lawrence Saul for their time and feedback.

The thesis is currently being prepared for submission for publication of the material. Shanta, Omar; Noor, Amina; Gorkin, David; Qiu, Yunjiang; Ren, Bing; Sebat, Jonathan. The thesis author was the primary investigator of this material. I would like to thank Amina Noor for helping with the initiation of the project. I would also like to thank Bing Ren, David Gorkin and Yunjiang Qiu for their hard work and effort in producing Hi-C data.

Lastly, I would like to thank my family and friends for their continued support throughout my studies. A special thanks to my parents Karim and Angelica Shanta for enabling me to pursue my goals and interests. I would also like to thank my brother Adam Shanta for his moral support throughout the years.

ABSTRACT OF THE THESIS

The Effects of Structural Variation on 3D Chromatin Structure

by

Omar Shanta

Master of Science in Electrical and Computer Engineering

University of California San Diego, 2019

Professor Jonathan Sebat, Chair
Professor Siavash Mirarab, Co-Chair

Three-dimensional spatial organization of chromosomes is defined by highly self-interacting regions 0.1-1 Mb in size termed Topological Associating Domains (TADs). Genetic factors that explain dynamic variation in TAD structure are not understood. We hypothesize that common structural variation (SV) in the human population; including deletions, insertions and inversions of $>$50 bp; can disrupt regulatory sequences and thereby influence TAD formation. To determine the effects of SVs on 3D chromatin organization, we performed chromosome conformation capture sequencing (Hi-C) of lymphoblastoid cell lines from 19 subjects for which SVs had been previously characterized in the 1000 genomes project. Large ($>$10kb) deletions had significant

effects on long-range chromatin interactions. Deletions overall were associated with increased contacts that span the deleted region but had a modest effect on contacts in adjacent regions. Deletions at TAD boundaries were associated with diminished contacts in the adjacent domains, consistent with the structure of TADs being dependent on regulatory sequences at their boundaries. Large inversions in the population had a common signature characterized by a rearrangement of contacts that span its breakpoints. Our results suggest that common SVs in the population impact long range chromatin structure.

Chapter 1

Introduction

3D chromatin structure is characterized by Topologically Associated Domains (TADs) and chromatin loops which create physical interactions between genes and distant regulatory sequences [5]. CTCF and the protein complex cohesin are localized to the boundaries of TADs [1, 7, 8], where they serve as barriers to the spread of chromatin. Genetic variation in these sequences has the potential to influence the binding of these factors and contribute to variability in chromatin structure in humans. However, little is known about patterns of topological variation in the population and the underlying genetic mechanisms.

Structural Variants (SVs) are a major source of genetic variability, and SVs have significant functional impact on the genome through the deletion or rearrangement of coding and regulatory sequences. Notably, large SVs that disrupt or re-establish chromatin contacts are associated with two rare monogenic disorders including human limb malformations and female-to-male sex reversal [2, 3, 4, 6]. However, relatively little is known regarding how SVs in humans impact the landscape of chromatin structure genome wide. In this study, we sought to investigate the effect of common SV polymorphism on 3D chromatin structure in a sample of individuals from the 1000 genomes project [9]. Chromatin Conformation Capture (Hi-C) sequencing of lymphoblastoid cell lines (LCLs) of 19 individuals from the 1000 genomes project was conducted. TAD Boundaries

were identified in all 19 individuals. The impact of TAD boundary removal was investigated through linear regression to obtain the relationship between number of chromatin interactions and genotype. We show a significant difference in chromatin structure when deletions and inversions are present.

Chapter 2

Methods

Hi-C data was acquired for 19 subjects from the 1000 Genomes Project. Chromatin contacts were quantified in 40kb bins and TAD boundaries were defined for each of the 19 subjects. Genotypes for 5845 SVs were obtained on the same subjects from the phase 3 SV calls from the 1000 genomes project [9]. The phase 3 SV call set includes 5517 deletions, 101 tandem duplications and 227 inversion/insertion/complex SVs that meet this criteria. Given that deletions vastly outnumber all other classes of variants we focused our primary analysis on these. Only deletion alleles that were present in at least 3/19 subjects were included in our analysis. Deletions were then mapped to overlapping bins within the contact matrix. This process starts by finding the bins of the start/end points for a given deletion and choosing the entire bins as the start/end bins. Additionally, a specific number of bins before and after the deletion (flank bins) were included as a way of testing the effects near and across the deletion. Only 9 of these subjects have both inversion calls and Hi-C data available, so inversions were analyzed individually.

Quantitative effects of deletions on chromatin conformation were tested by Ordinary Least Squares Regression (OLSR). First, bins that overlapped with SVs were masked and specific deletion-flanking and deletion-spanning target regions were defined within 240 kb (6 bins) on either side of the deletion (Figure 3.1A). Regression was performed for each deletion within the

flanking and spanning target regions respectively, controlling for ancestry principal components obtained from SNP genotypes. PLINK software was used to perform the PCA.

Effects were visualized for select loci as heatmaps of regression coefficients. Each heatmap is constructed by applying a regression for all bins separately across a target genomic region. In addition, to visualize "average" effects across multiple loci, matrices were centered on the left and right deletion boundaries, and the median regression coefficient for each bin across multiple loci was displayed as a heatmap (see Figure 4).

To perform the inversion analysis, the Hi-C chromatin interactions for the bins that overlap the inversion and 62 bins on each side of the inversion were extracted. A pearson correlation between number of chromatin interactions and genotype was applied for each bin across the 9 samples that had both Hi-C data and inversion calls available.

Chapter 3

Results

We hypothesize that SVs could influence TAD structure indirectly by disrupting regulatory sequences that control TAD formation. In addition, we anticipate that SVs will have direct effects on the coverage and spacing of paired-end reads similar to the effects that are ordinarily observed for SVs in whole genome sequence data. Therefore, we sought to distinguish each of these types of effects. In the example illustrated in Figure 3.1, a deletion results in loss of all contacts with the deleted region itself as evident by diminished counts (coverage), and as the regions flanking the deletion are brought nearer, there is a simultaneous increase in contacts spanning the deletion (Span). These effects are not expected to directly influence the counts in the flanking regions (Flank).

Quantitative effects of 2180 common deletions on Hi-C data were tested by linear regression. A null distribution was also obtained by randomly permuting the genotype in each sample 100 times according to a uniform distribution. The genomic inflation factor (λ) is used to determine bias in the model based on the median of the chi-squared statistic for the observed p-values divided by the expected median of the chi-squared distribution with one degree of freedom. A λ farther from 1 denotes a model with a larger amount of bias. A Quantile-Quantile plot of observed regression p-values relative to a null distribution shows a substantial enrichment

of effects for large (>10 kb) common deletions for deletion spanning (Figure 3.2, $\lambda = 2.4$) and deletion-flanking regions ($\lambda = 1.24$) within 240 kb on either side of the deletion, whereas there was comparatively little inflation of the test statistic for deletions smaller than 10kb ($\lambda = 1.12$ and 1.07 for span and flank respectively). These results are consist with the observed effects being attributable to large-scale effects of deletions on nearby chromatin contacts.

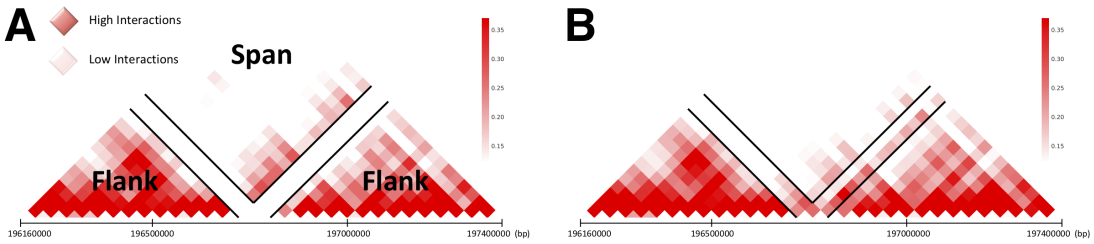

Figure 3.1: **Hi-C bins near a deletion are separated into two categories: Span and Flank Regions.** An 80kb deletion located at chr1:196728877-196808865 is depicted by averaging the homozygous DEL subjects in panel A and homozygous REF subjects in panel B. Hi-C bins that overlap with a deletion are removed and 6 bins on both sides of the deletion are extracted for analysis. Red indicates a high number of chromatin interactions between bins while white indicates a lower number of interactions. The matrix of interactions was separated into regions that interact across the deletion (span) and regions that do not cross the deletion (flank) as they can exhibit different behavior with the removal of the deletion bins.

The number of flank bins was chosen by a sweep to see the effects of short-range versus long-range interactions. The effects decreased with an increase in bin size, so 6 bins was used for the rest of the analysis. The QQ plots helped choose the covariates that mitigate the effects of bias and to identify the correct model (Equation 3.1). In this model, Y is the genotype with range [0,1,2] where 0 is reference, 1 is heterozygous for the deletion and 2 is homozygous for the deletion. X (Flank/Span) is the median chromatin interactions of the 42 bins in the flank category or 36 bins in the span category. α, β and γ are regression coefficients that are produced by the model and we are interested in the magnitude and direction of α. Ancestry and intra-chromosomal PCs are potential covariates that need to be accounted for. C is a constant that helps fit the line. Using a model with only ancestry PCs as covariates produced the results with the least amount of

bias.

$$Y = \alpha X_{Flank/Span} + \beta(ancestryPCs) + \gamma(chromPCs) + C \tag{3.1}$$

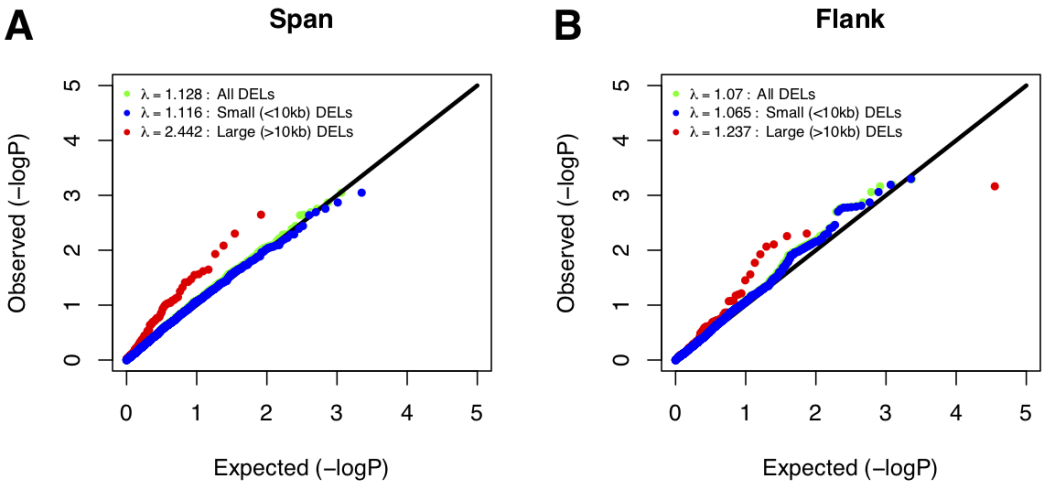

Figure 3.2: **Linear Regression p-values indicate large deletions have the biggest impact on chromatin structure.** Linear Regression was performed with a few different models to determine how many covariates were necessary to explain away group variability. Adding only ancestry principal components (PCs) 1 and 2 as covariates had the lowest inflation factor for both flank and span regions and was used as the optimal model. The deletions were then stratified by size at a 10kb threshold to determine the source of the effect. A) Large deletions in the span region have the strongest effect while the contribution from small deletions is non-existent. B) Large deletions also show an effect to a lesser extent in the flank region.

We examined whether there was a strong directionality of effects for 112 large common deletions by a wilocoxon rank sum test of the regression coefficients. In addition, we examined whether effects differed based on the position of the deletions relative to TADs. Large deletions at TAD boundaries (TB, N = 16 deletions) were associated with a decrease in chromatin interactions in the flanking TADs within 240 kb on either side (Figure 3.3), while a consistent directionality of effect was not observed for deletions overall or for deletions that did not intersect with a TAD boundary (non-TB, N = 96 deletions). Deletions overall were associated with an increase in long-range contacts that span the deletion (Figure 3.3), as would be expected when the distance

between flanking regions is shortened. This effect was also significant for non-TB deletions. A strong positive effect was not observed for chromatin interactions spanning the TAD boundary deletions. Thus, we do not find evidence that common TB deletions typically result in "TAD fusion". However, a larger number of TB deletions must be assessed to draw general conclusions about their effects.

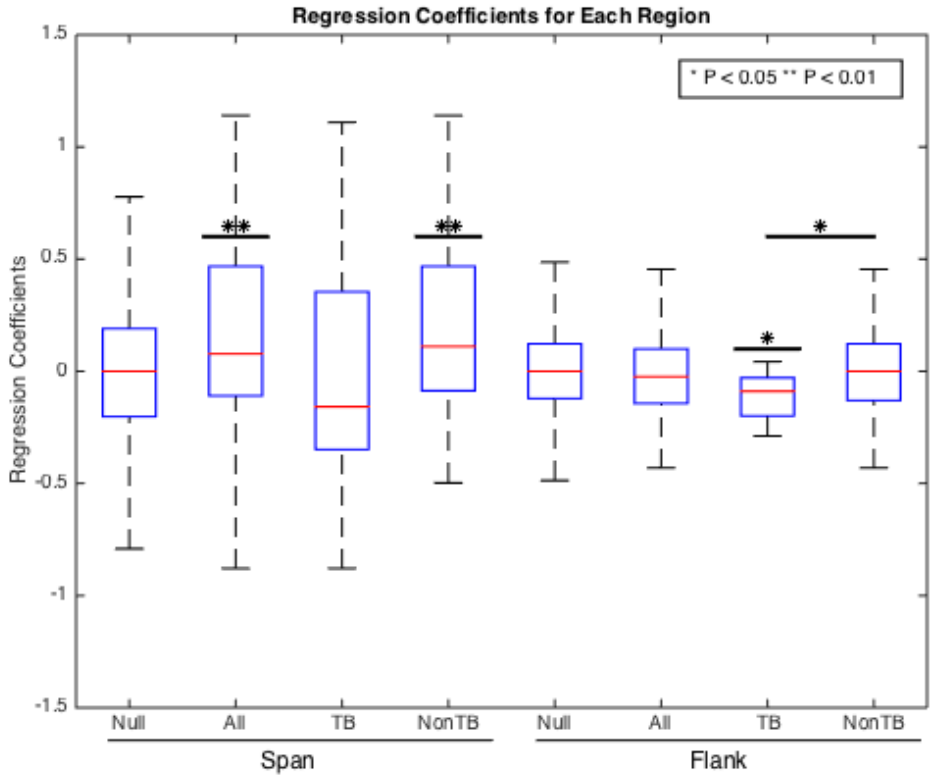

Figure 3.3: **Span region shows significant positive effect for large deletions when the deletion does not intersect a TAD boundary.** Wilcoxon test was performed for each category of median regression coefficient distributions against a zero-median distribution after being stratified by overlap with TAD boundaries. TAD boundary overlap with deletions consistently shows negative values in the flank region while an absence of TAD boundary overlap with deletions consistently shows positive values in the span region.

When a deletion affects a piece of the chromosome that is not a TAD boundary, there is a significant positive effect between the regions that span the deletion (Figure 3.4A). This is consistent with the results from Figure 3.3 and since the deletions that do intersect TAD

boundaries do not have this trend, the promotion of interactions is driven by the deletions that are not at TAD boundaries. A reason for the noisy coefficients in the TAD boundary overlap deletions could be due to sample size. Of the 112 large deletions, only 16 of them overlap with TAD boundaries and 96 of them do not. If more loci were available, there could be more support to identify the presence of TAD fusion. Currently, the results do not support the existence of TAD fusion with the deletion of TAD boundaries. What can be seen from the TAD boundary overlap is the amount of negative effects in the flank region (Figure 3.4B). Even though there are only 16 loci, the flank region is mostly negative and there is higher confidence that TAD disruption occurs at the edges of the TAD close to the deletion at the boundary.

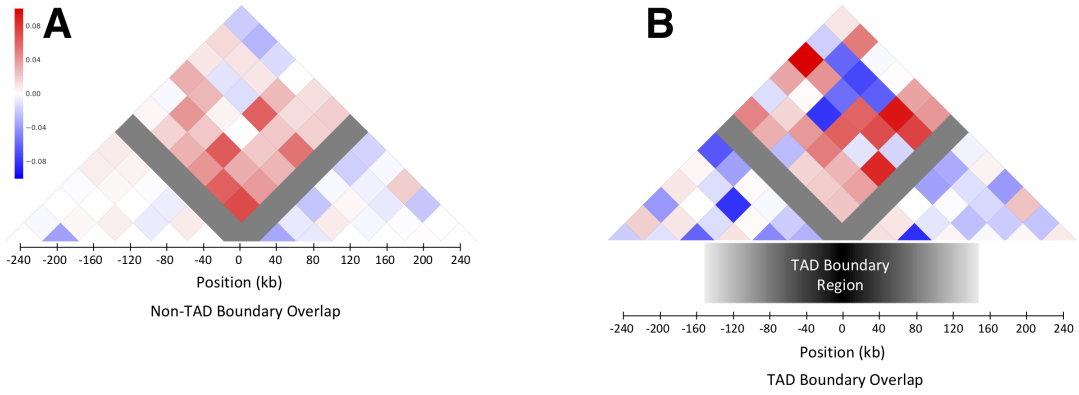

Figure 3.4: **Average chromatin structure for large deletions not present at TAD boundaries and present at TAD boundaries.** Linear regression was performed for each bin and the coefficients were averaged across large deletions using a median. The amount and location of overlap between deletions and TAD boundaries varied, but most are within the boundary. A) Deletions not present at TAD boundaries have positive values in the span region with neutral values in the flank region. B) Deletions that intersect with TAD boundaries have negative values in the flank region and seemingly noisy values in the span region.

Our results suggest that large SVs have detectable effects on chromatin conformation. Since the above analysis focused on deletions, it did not assess the largest common SVs known to exist in the population, which include large inversions of 8p23.1 (3.87 Mb) and 7q11.1 (2.45 Mb). To characterize the effects of large inversions on chromatin conformation, inversion genotypes were obtained from 50X single molecule (PacBio) sequencing for 9 subjects in the 1000

genomes project), and the correlation of chromatin contacts across the region was visualized. The availability of a full assembly of the 8p23.1 inversion haplotype, enabled us to map TAD structure of the inversion haplotype directly by mapping HiC data of subjects that were homozogous for the 8p23.1 inversion to a "patched" version of the hg19 reference genome that contains the assembled inversion haplotype. TAD structures of the reference and inversion haplotypes were similar, and the same 5 TADs were defined (Figure 3.5A). As expected, the most dramatic effects of the inversion involve contacts that span the inversion breakpoints (Figure 3.5B). The black triangle illustrates the inversion region and it can be seen that these effects span distances $>$2 megabase from the breakpoint.

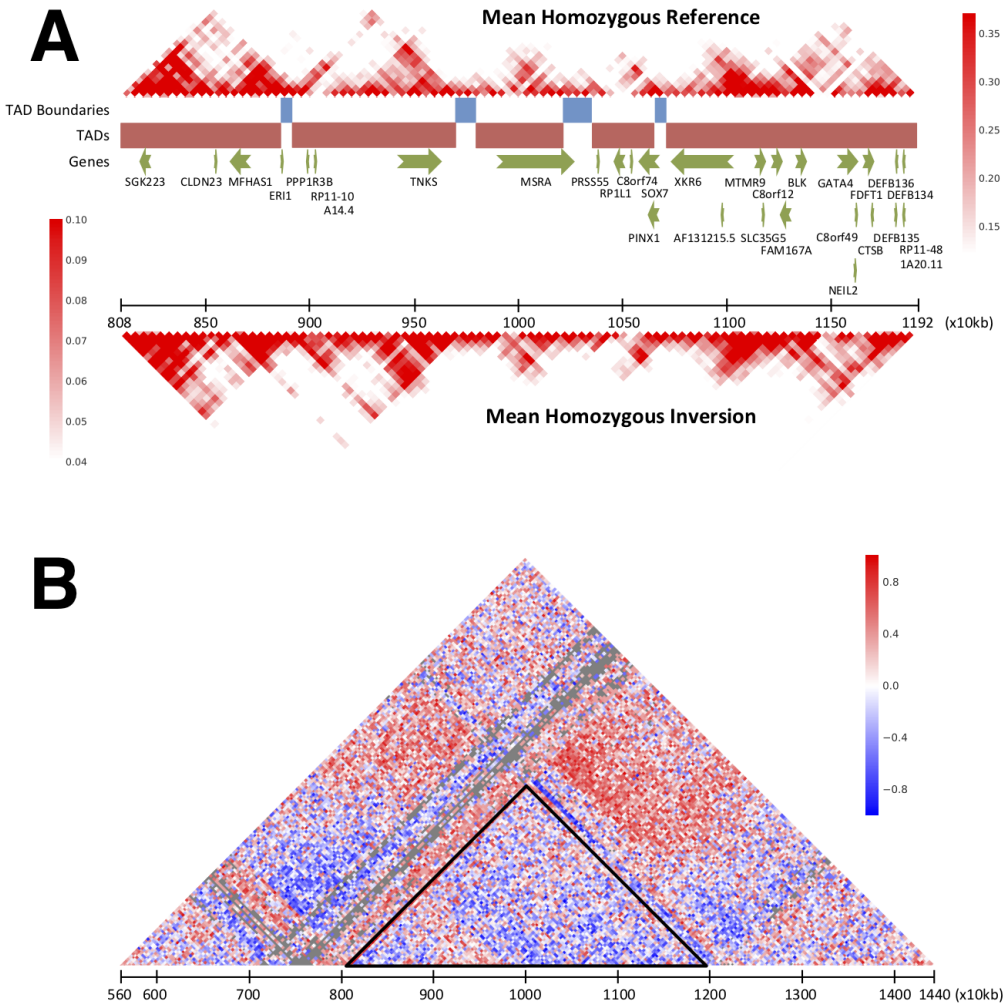

Figure 3.5: **Hi-C chromatin interaction counts of the 8p23.1 region shows a change in TAD and gene location at an inversion site.** A) Hi-C matrices for samples that were homozygous for the absence of an inversion and homozygous for the inversion at 8p23.1 were averaged separately and annotated. The TAD structure is preserved in a mirrored fashion along with their associated genes. Chromatin interactions for the inversion were mirrored to aid visual comparison with the reference. B) A correlation heatmap shows an increase in chromatin interactions between regions that are closer together after the inversion. The inversion region is depicted by the black triangle.

Chapter 4

Discussion

HiC has enabled discoveries related to understanding the structural and functional basis of the genome. Variation in chromatin structure within the population has not been studied. We show that large common deletions have significant effects on patterns of chromatin conformation, with effects that are sufficiently large to be detectable in our small samples of 19 subjects.

Large common deletions have a distinctive signature characterized by strong positive effects on contacts that span the deletion. In addition, we find evidence that deletions, particularly those that occur at the boundaries of TADs, negatively impact the conformation of the flanking TAD regions. Effects of deletions on TAD structure has potential significance for gene regulation. However, larger sample sizes with power to detect the effects of small ($<$10 kb) or low-frequency (MAF 1-5%) SVs are needed to get a broader picture of how large deletions at TBs affect TAD structure. More power also needed to investigate specific regulatory sequences and how these determine effects on gene regulation.

Large common inversions have corresponding effects on chromatin interactions that span the inversion breakpoints, and these effects extend for distances of up to 2Mb. TAD structures within the large inverted segments of two common inversions appears to be well preserved on the inversion haplotypes, suggesting that the sequences within these regions contains the necessary

regulatory information to determine their 3D structures.

Our analysis has shown that gene regulation is directly affected by structural variation through the alteration of chromatin structure. We hope that this finding will inspire ideas that can implicate variants by using 3D structural information with larger sample sizes.

The thesis is currently being prepared for submission for publication of the material. Shanta, Omar; Noor, Amina; Gorkin, David; Qiu, Yunjiang; Ren, Bing; Sebat, Jonathan. The thesis author was the primary investigator of this material. I would like to thank Amina Noor for helping with the initiation of the project. I would also like to thank Bing Ren, David Gorkin and Yunjiang Qiu for their hard work and effort in producing Hi-C data.

References

[1] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. In *Nature*, 2012.

[2] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. In *Nature*, 2016.

[3] F. R. Goodman. Limb malformations and the human hox genes. In *American Journal of Medical Genetics*, 2002.

[4] N. Jansz, A. Keniry, M. Trussart, H. Bildsoe, T. Beck, I. D. Tonks, A. W. Mould, P. Hickey, K. Breslin, M. Iminitoff, M. E. Ritchie, E. McGlinn, G. F. Kay, J. M. Murphy, and M. E. Blewitt. Smchd1 regulates long-range chromatin interactions on the inactive x chromosome and at hox clusters. In *Nature Structural and Molecular Biology*, 2018.

[5] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. In *Science*, 2009.

[6] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. In *Cell*, 2015.

[7] R. P. McCord. How to build a cohesive genome in 3d. In *Nature*, 2017.

[8] M. Merkenschlager and E. P. Nora. Ctcf and cohesin in genome folding and transcriptional gene regulation. In *Annual Review of Genomics and Human Genetics*, 2016.

[9] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter,

S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, T. . G. P. Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel. An integrated map of structural variation in 2,504 human genomes. In *Nature*, 2015.