University of Cincinnati

Date: 3/27/2017

I, Jessica Shank, hereby submit this original work as part of the requirements for the degree of Master of Science in Genetic Counseling.

It is entitled:
Topological Domain Variations Among Patients Undergoing Microarray Testing

Student's name: Jessica Shank

This work and its defense approved by:

Committee chair: Kristen Sund, Ph.D.
Committee member: Iouri Chepelev, Ph.D.
Committee member: Lisa Dyer, Ph.D.
Committee member: Lisa Martin, Ph.D.
Committee member: Auvo Reponen, Ph.D.
Committee member: Teresa Smolarek, Ph.D.

UNIVERSITY OF CINCINNATI

22449
Topological domain variations
among patients undergoing microarray testing

A thesis submitted to the
Graduate School
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of
Master of Science

in the Department of Pediatrics
of the College of Medicine
March 2017
by

Jessica Marie Shank

B.S. (Honors) The Ohio State University
May 2013

Committee Chair:
Kristen Sund, LGC, Ph.D.

Committee Members:
Iouri Chepelev, Ph.D.
Lisa Dyer, Ph.D.
Lisa Martin, Ph.D.
Auvo Reponen, Ph.D.
Teresa Smolarek, Ph.D.
ABSTRACT

Chromosomal microarrays have been used for over a decade to improve clinical diagnostic rates by identifying copy number variations (CNVs) across the genome. CNVs in noncoding regions of the genome are difficult to interpret because less is understood about the etiology of disease in these cases. Topologically-associating domains (TADs) describe the higher order chromatin structure that creates a landscape for enhancer-promoter interaction. The human genome has roughly two thousand such regions marked by high levels of chromatin interaction within these regions and separated by genomic boundaries, across which interaction occurs much less frequently. These domains and boundaries may provide the underlying genomic infrastructure for gene regulation and long range activation or repression of gene expression. Disruption of the boundary regions has been associated with human disease in two types of limb abnormalities and in other small disease specific populations. To date only one study has estimated the impact of TAD boundary deletions in an affected population and that study focused on large, known pathogenic deletions that were likely to disrupt multiple boundaries. In this study, a large scale, quantitative, comparative case-control study design was used to examine the presence of deletions across TAD boundaries by comparing patients with clinical indications for microarray testing to population-based controls. 5033 patients with a broad variety of clinical indications for microarray were compared with 312 controls to assess overall deletion burden, enrichment of topologically-associated domain boundary region disruption, and phenotypic trends among patients. A nonparametric t-test was used to assess the difference in the proportion of deletions which span an entire TAD boundary in cases compared to controls. No significant difference was found between cases and controls when both heterozygous and homozygous deletions were compared together. However, cases had significantly more homozygous TADs deletions than controls (p<0.0001). Furthermore, the increase in homozygous TADs in cases could not be explained by the overall number of homozygous deletions. Rare topological domain disruption among the cases was examined to identify candidate TADs deletions associated with phenotypic trends. Chromosome 14q32 was assessed for phenotypic trends and shared genomic properties. These data demonstrate that further analysis is warranted to determine the impact of TAD
disruption on human disease. The approach used here sets a foundation for clinical interpretation of TAD disruption in a clinical setting.

*Keywords*: epigenetics, topological domains, microarray, enhancers, clinical, copy number variations
Table of Contents
INTRODUCTION ............................................................................................................... 1
  Topologically associating domains ............................................................................. 1
  TADs and enhancer-promoter interaction ................................................................. 2
  TADs and impact on human disease ........................................................................... 4
  TADs in the clinical setting ......................................................................................... 5
MATERIALS AND METHODS ......................................................................................... 7
  Study design .................................................................................................................. 7
  Topological domains and boundaries ......................................................................... 8
  Bioinformatics query ..................................................................................................... 9
  Data analysis ................................................................................................................ 10
  Rare TAD analysis ......................................................................................................... 10
RESULTS .......................................................................................................................... 11
  Study population .......................................................................................................... 11
  Case/control comparisons ............................................................................................ 12
  Rare TAD analysis ......................................................................................................... 12
DISCUSSION .................................................................................................................... 14
  Limitations .................................................................................................................... 19
  Conclusion ..................................................................................................................... 19
REFERENCES .................................................................................................................. 21
FIGURES .......................................................................................................................... 25
  Figure 1: Topological domain boundary disruption ..................................................... 25
  Figure 2: Defining clear borders for boundary deletions ............................................. 25
  Figure 3: Homozygous deletions among cases and controls ....................................... 26
  Figure 4: Comparison of deletions among cases and controls .................................... 27
  Figure 5: Filtering of potential TAD boundary regions for rare analysis ................... 28
TABLES ............................................................................................................................. 29
  Table 1: Demographics ............................................................................................... 29
  Table 2: Deletions in hESC and IMR90 cell lines ...................................................... 29
  Table 3: Deletions among cases and controls ......................................................... 30
  Table 4: TAD boundary regions selected for rare analysis ........................................ 30
  Table 5: Phenotypes for patients with 14q32 deletions ............................................ 31
APPENDIX ........................................................................................................................ 32
  Appendix A .................................................................................................................... 32
INTRODUCTION

Molecular genetic testing for clinical disease has become more cost effective and efficient in recent years. Diagnostic rates have improved most for single gene Mendelian conditions caused by mutations in the coding regions of genes (Liu et al., 2015). Microarray testing affords the opportunity to examine the entire genome for copy number variation (CNV), including both deletions and duplications of genetic material in coding and non-coding regions. Non-coding regions of the genome are thought to play a role in gene expression activation or repression but the mechanisms are not yet clearly characterized in clinical populations. Recent studies suggest that the disruption of gene regulation offers an explanation for human disease outside of traditional molecular etiologies like haploinsufficiency due to gene deletion or loss of function mutations. The non-coding genomic landscape remains largely under-investigated and implications for clinical application have yet to be determined.

Topologically associating domains

The physical structure of chromatin is integral to proper gene expression as the landscape influences which areas of the genome are available for transcription. The active and inactive forms of chromatin are known to be structurally different from one another and they play a role in controlled gene expression (Sexton & Cavalli, 2015). New technology has enabled better understanding of the architecture of the mammalian genome including the likelihood of interactions between non-coding regions of the genome involved in long term regulation. Enhancers and promoters, both of which are encoded in DNA, are physically separated from one another but the formation of physical loops through protein interaction facilitates DNA interaction. Transcription factor binding sites are also encoded by the DNA and provide a site to bind transcription factors when expression or suppression occur (Maston, Evans, & Green, 2006; Matharu & Ahituv, 2015; Montavon & Duboule, 2012). This level of genetic regulation which involves controlling transcription of genes beyond the level of the sequence of DNA is termed epigenetics. Epigenetic regulation is controlled by physical spacing of these elements, allowing temporal and spatial coordination of gene activity (Dixon et al., 2012).
Chromosome conformation capture analysis is a technique which analyzes the structure and spatial organization of chromatin. Techniques like 3C, 4C, 5C, and Hi-C are similar approaches which have offered additional insights into the physical interactions between chromosomal regions by looking at the interactions between an increasing number of DNA fragments. These methods quantify interactions between segments of DNA that may be separated by many nucleotides within the genome. In the first step, formaldehyde crosslinking freezes protein-DNA interactions. Then, unbound DNA is digested and removed and remaining DNA is ligated to proximal DNA fragments. Once the DNA molecules are ligated, the crosslink is reversed and protein is removed, leaving only ligated DNA fragments. The library is then amplified and sequenced to quantify those segments that are physically near one another and represent areas of the genome that interact via transcription factors and mediator proteins to regulate gene expression. These methods have been used to delineate TADs and TAD boundary regions throughout the genome (Figure 1). TADs are represented by regions in which interactions are occurring within the same domain. Stretches of DNA across which physical interactions are much less likely are described as TAD boundaries (Lieberman-Aidan, 2009).

Recently, studies examining chromosome conformations have indicated TADs function as an epigenetic layer of compartmentalization within the mammalian genome. The delineation of TADs recognizes the physical separation between genes and their regulatory components. TADs are conserved across species and across cell lines, though a small subset are cell type specific (Dixon et al., 2012). These megabase-sized units are known to span the genome and are thought to contribute to the communication between promoters and enhancers (Dekker & Heard, 2015).

**TADs and enhancer-promoter interaction**

Enhancers and promoters interact by forming enabled loops through protein-DNA and protein-protein interaction. These loops allow physically distant segments of the genome to interact with each other. Interaction between proteins and the chromatin structure regulates which enhancers and promoters are in communication, thereby regulating gene expression. Chromatin interaction techniques identify physical contact within TADs, but contact does not always correlate to functional effect (Vernimmen, De
Gobbi, Sloane-Stanley, Wood, & Higgs, 2007). TAD boundary regions serve as insulators, preventing enhancers from interacting with promoters nearby by physically hindering the ability for this looping to occur (de Laat & Duboule, 2013). Proteins such as CTCF, a transcription factor that has enriched binding sites at TAD boundaries, housekeeping genes and members of the cohesion protein family are commonly found at TAD boundaries. Presence of these regulatory elements suggests an insulation effect of boundary regions (Dixon et al., 2012; Ong & Corces, 2014).

Inappropriate enhancer interaction can lead to dysregulated gene expression through enhancer adoption or enhancer loss of function. Enhancers typically regulate transcription in a spatial or temporal manner. They are typically modular and specific to the promoter with which they usually interact. Enhancers are similar to proximal promoter elements in their function, but are often located much more distally. They can be several hundred kilobase pairs upstream of a promoter, downstream of a promoter in an intron, or even beyond the 3’ end of the gene (Maston et al., 2006).

When enhancers are disrupted, there is potential for regulatory dysregulation. Single nucleotide variants (SNVs) within the enhancer have been shown to disrupt regulation by affecting transcription. SNVs have been shown to alter transcription machinery assembly and disrupt the propagation of transcriptional activating signaling. Studies have shown that most variants result in a 1.3-2 fold difference in target gene expression (Kheradpour et al., 2013; Patwardhan et al., 2012). Point mutations can also disrupt the transcription mechanisms (Haraksingh & Snyder, 2013). Variation in the untranslated transcript can affect microRNAs which can then affect the secondary structure of the RNA and lead to inhibited binding to other transcripts (Cai, Yu, Hu, & Yu, 2009; Kasowski et al., 2010; Stranger et al., 2007). Direct measurement of CNVs and their impact on transcription has been limited. Reports of single nucleotide changes within enhancers have been associated with human disease. Alteration of the long distance enhancer ZRS has been shown to disrupt the sonic hedgehog (SHH) pathway and lead to limb abnormalities. Roughly a dozen single nucleotide variations in this regulatory region have been associated with preaxial polydactyly (Furniss et al., 2008; Lettice et al., 2003; Visel, Rubin, & Pennacchio, 2009). Deletion of enhancers has also been shown to influence disease by altering expression in diseases like
Van Buchem disease and nonsyndromic congenital retinal nonattachment (Ghiasvand et al., 2011; Loots et al., 2005).

**TADs and impact on human disease**

Research is still needed to characterize TAD boundary regions, including the region’s size, genetic makeup and critical binding sites for insulation. TAD boundary regions are thought to serve as protectors, isolating one domain from another and preventing inappropriate communication between enhancers and promoters. When these boundaries are disrupted, rewiring of these communications results in differential gene expression (Lupianez et al., 2015).

Lupianez et. al. were among those who demonstrated human disease as a result of TAD disruption. In their study, unrelated patient families with discrete limb anomalies of unknown etiology were examined for TAD boundary interruption by copy number variation, including deletions, duplications, and inversions. The *EPHA4* gene is housed in its own topologically associating domain, flanked on either side by other TADs which contain genes related to limb development. All located on chromosome 2, the *WNT6* and *IHH* genes are upstream of *EPHA4* while the *PAX3* gene is downstream of *EPHA4*. Patient families with discrete limb anomalies were examined for boundary deletions, duplications, and inversions. Three unrelated families with brachydactyly were found to have heterozygous 1.75-1.9 Mb deletions on chromosome 2q35-36. These deletions include the *EPHA4* gene and extended into the non-coding region, disrupting the predicted domain boundary between the *EPHA4* and *PAX3* genes (Lupianez et al., 2015). There is no evidence to suggest that haploinsufficiency of the *EPHA4* gene alone would result in brachydactyly (Goumy et al., 2014; Li et al., 2015). A patient family with severe polysyndactyly has a duplication near chromosome 2q35. This mutation brings the *EPHA4* promoter closer to the *IHH* gene. The CNVs identified in this work characterize the influence of altered boundary regions and the link between TADs and development of a phenotype. Inappropriate gene expression occurred due to a change in the physical distance between enhancers and promoters. Following the identification of variation among patients, mouse models were used to demonstrate that the disruption of these regions leads to reorganization of the TADs and ectopic gene interactions such as
novel enhancer-gene interactions. The boundaries between domains play a role in facilitating interactions and controlling embryonic development. Disruptions in the chromatin landscape have been shown to alter limb development and likely affect other processes as well (Lupianez et al., 2015).

**TADs in the Clinical Setting**

TADs have been implicated in human disease and progress has been made in beginning to utilize TAD disruption in clinical testing and clinical disease. SNP-based microarray testing evaluates the entire genome for copy number variation and is a standard first tier test, for indications such as developmental delay, offered by clinical genetics laboratories across the nation (Ho et al., 2016; South, Lee, Lamb, Higgins, & Kearney, 2013; Wang & Boyar, 2016). The American College of Medical Genetics guidelines for interpreting CNVs indicate that pathogenicity be determined based on reports in the literature, known genes in the region, clinical databases, and clinical judgement regarding the size of the CNV (South et al., 2013). CNVs in regions with no known function or clinical implication are usually not considered for pathogenicity in mainstream clinical genetics laboratories (Kearney, Thorland, Brown, Quintero-Rivera, & South, 2011).

Genetics laboratories use frequency of the CNV in controls to filter variants of interest. CNVs seen in more than 1 percent of healthy individuals typically are not considered to be pathogenic. Copy number variation exists within those without the occurrence of a clinical phenotype (Jakobsson et al., 2008). In 2006, Redon et al. (2006) reported 1,447 copy number regions identified in the HapMap population which includes individuals who did not report any major medical concerns (Redon et al., 2006). In a separate analysis, controls were analyzed using 1 Mb resolution and 42 copy number variants were found (Qiao et al., 2007). Since phenotypically normal individuals have benign copy number variation and its frequency has varied throughout the literature, it is likely that population-based controls will have TAD disruptions. Common CNVs are found throughout the genome, and likely exist across TAD boundary regions. Standardized methods are in place to assess the clinical impact of CNVs and assessment of TADs at the clinical level is currently being explored.
CNVs are known to have an impact on the level of expression of genes. This can be a result of a gene dosage effect, but it can also be due to long range disruption of gene regulation. Structural changes can affect transcription by rearranging regulatory elements such as promoters and enhancers ultimately disrupting gene expression. Cis regulatory position effects have been reported up to 1.5 Mb from the gene itself (Haraksingh & Snyder, 2013). ChIP-seq and similar methodologies have allowed these regulatory elements to be further examined, but their roles in clinical disease remain largely unknown.

TADs are able to be assessed in the clinical setting and have been shown to offer additional insight into clinical disease. Redin et. al. evaluated balanced chromosomal arrangements (BCAs) for individuals with a range of congenital anomalies. They determined that 7.3% of BCAs identified by a nucleotide resolution cytogenetic approach disrupted TADs encompassing known syndromic loci. BCAs were considered for TAD disruption if they had a dominant loss of function gene nearby. Once considered plausible, heat maps were generated using Juicebox, a software program used to visualize Hi-C data, indicative of intrachromatin interactions (Durand et al., 2016). Breakpoints in 8 participants were predicted to interrupt the long range regulation of the MEF2C gene which is associated with developmental delay, dysmorphic features, and absent speech. TADs disruptions near the FOXG1 and SATB2 genes were also identified by BCA analysis. FOXG1 causes FOXG1 syndrome which includes structural brain abnormalities and impaired brain function while SATB2 is associated with cleft palate and developmental delay (Redin, Brand, & Collins, 2017).

While several studies have identified TADs that appear to explain specific disease, the broader clinical importance and how to evaluate the clinical significance of TADs remains unclear. This study aims to determine whether TADs play a role in unexplained clinical disease by assessing whether there is an enrichment of single TAD boundary deletions in a patient population. In addition, this study aims to suggest an approach to candidate rare TAD boundary deletions that may help explain the etiology of human disease.
MATERIALS AND METHODS

Study design

IRB approval was obtained from Cincinnati Children’s Hospital Medical Center (CCHMC). A quantitative study was performed to compare the number of total deletions and the number of deletions which cross topological domain boundary regions in patients who have undergone microarray testing and population-based controls. Laboratory test requisition and electronic health record data was used in categorical phenotypic analysis to identify overlapping phenotypes in patients with shared TAD boundary disruptions.

A retrospective database review was performed on patients who had clinical testing in the CCHMC Cytogenetics laboratory between October 11, 2010 and July 4, 2016. We excluded data from patients that had somatic microarray testing for an oncology indication and those that were obtained for parental inheritance testing. We also excluded any research participants and patients that were analyzed on an older SNP microarray platform, the Illumina Human-610 Quad Beadchip. Patients were included regardless of whether their final microarray result was reported as normal, abnormal, or as a variant of uncertain clinical significance (VUCS). For abnormal microarray results, we reasoned that TAD boundary deletions may help explain pathogenicity when CNVs in large, gene sparse regions result in a disease phenotype for abnormal cases or cases with CNVs that were classified as VUCS. For normal microarray results, we reasoned that we could identify significant disruptions that were not clinically reported since many are too small in size and do not contain known disease-causing genes. These types of results would be excluded during the clinical interpretation process. Our final patient population included 5033 individuals (patient population, n=5033).

The retrospective study collected patient data from 2 different microarray platforms. 2479 patients were run on the CytoSNP- 850K v1.1 Beadchip (850kv1.1), and 2515 patients were run on the Human-Omni 1 Quad Beadchip (Omni1). CNVs were imputed from SNP genotypes using Illumina CNV Partition v3.1.6 with Genome Studio software, as well as visual inspection of the data.
Population-based controls were obtained from the Cincinnati Genomic Control Cohort (GCC) for this study. Participants in the GCC were recruited between 2007 and 2010 and SNP data was obtained at that time. The GCC consists of 1020 population representative children who had no major medical concerns. The SNP data, which is available for use upon request, were derived from the Illumina Infinium (Omni V) chip. It was then analyzed using Illumina CNV Partition v3.1.6 with Genome Studio software to identify copy number changes.

For the controls, parents reported phenotypic information when they were recruited to the GCC. We excluded participants whose parents reported any phenotypes that could have been due to a genetic syndrome which included: developmental delay, endocrine issues, hearing issues, heart issues, learning/behavioral issues, mental health concerns, neonatal issues, renal/urinary concerns, reproductive problems and skeletal/growth concerns. After exclusion, 908 controls were available for analysis. We randomly selected 312 for this study.

In total, this study had 5345 participants, including both cases and controls. Demographic information for the cases was collected from test requisition information available in the cytogenetics database at CCHMC and from the GCC database for controls.

**Topological domains and boundaries**

Topological domain data from the genome-wide higher-order chromatin interaction data in human embryonic stem cells and human lung fibroblast cells were downloaded in human genome build 19 [hg19] coordinates (Dixon et al., 2012). TAD data are available for two human cell lines, hESC and IMR90. Both hESC and IMR90 cell lines were utilized to maximize the possible TAD boundaries of interest and to account for phenotypes that occur in developmentally diverse tissues. While boundary regions are often conserved across cell types, some reported boundary regions are unique to individual cell lines and may represent cell type specific function. Topological domain boundary regions were defined as the region between two adjacent domains. The median boundary size in both the hESC cell type and the IMR90 cell type was 0 kb and the 75th percentile was 40 kb. However, the distributions were
markedly different with the upper quantile range being 749 kb in the hESC cell line and 3.13 Mb in the IMR90 cell line.

Multiple platforms were in use by the Cytogenetics laboratory during the time of patient testing. The 850K v.1.1 chip was the chip with least resolution. The 850K v1.1 chip has 843,390 markers and an average probe spacing of 1.8 kb. The Omni1 chip has 1,134,514 markers and an average probe spacing of 2.63 kb (Illumina, 2016). Based on a CNV call of 10 sequential SNPs, we reasoned that the 850K v1.1 chip minimal resolution was approximately 18 kb. The OmniV platform was used for the GCC. The OmniV chip has 4,301,331 markers and an average probe spacing of 0.68 kb. This platform is much more dense than the two previously mentioned platforms.

**Bioinformatics query**

Two informatics queries were developed. Both queries were validated by randomly cross referencing individuals identified through the query and their data in the cytogenetics database at CCHMC. The purpose of the first query, from here referred to as the genomic query, was to identify and compare deletions among cases and controls within the genome, regardless of location. In this genomic query, a minimum deletion size of 18 kb was chosen to avoid bias based on the platforms’ minimum detection size limits.

The second query, here referred to as the TAD boundary query, was designed to generate a list of individuals who have deletions which entirely cross TAD boundary regions of interest in either the hESC or IMR90 cell line. Previous studies have focused on large, known pathogenic deletions with an average size of 3.1 Mb (Ibn-Salem, 2014). We reasoned that disruption of a single boundary could lead to enhancer adoption and focused the query on individuals who have a deletion across one entire TAD boundary, but does not extend into a second boundary region. TAD boundary regions are identified as a region of increased chromatin-protein interaction and across which there is limited chromatin-chromatin interaction. Because of this, boundary regions vary in size. In order to clearly define the TAD boundary limits, a query was developed based on domain location rather than boundary location. The TAD boundary query identified individuals with heterozygous and homozygous deletions, which began in one
domain and ended in the subsequent domain. Figure 2 outlines the deletions that were detected by the TAD boundary query.

For both the genomic and TAD deletion queries, demographic information was exported from the database including race, ethnicity, age, sex, and date of sample collection. Significance of the clinical microarray testing, platform used, and indications listed on test requisition were included in the output as well. For each deletion, its position, number of SNPs covered, and clinical interpretation were included.

**Data analysis**

The number of deletions across the genome in cases and controls were compared using a non-parametric t-test to identify differences between the groups. A non-parametric t-test was also used to compare the number of deletions that encompass an entire TAD boundary in cases and controls. The proportion of deletions which encompass an entire TAD boundary in cases and controls were then compared using a non-parametric t test. To evaluate whether the presence of TADs were related to the number of deletions in cases and controls, individuals were grouped based on whether they had TADs and non-parametric tests were used in cases and controls separately. To evaluate the relationship between TADs and number of deletions, we used linear regression modeling the number of deletions by TAD, case/control, and TAD*case/control. In the regression analysis, only individuals with 1 or more deletion were included (n=2822). To account for a non-normal distribution, number of deletions was log-transformed.

**Rare TAD analysis**

Case analyses were performed to test the hypothesis that a single TAD boundary deletions cause Mendelian disease in a population referred for microarray testing. Cases were considered for analysis when more than one participant had a deletion crossing a boundary, but the deletion was not commonly seen in the case or control population. We removed TAD deletions that were found in controls or if they were found in greater than 1% of the cases.

Qualitative analysis was performed to further characterize the phenotypes in those with rare TAD boundary disruption. We took a genotype first approach in cases to test the hypothesis that TAD boundary
deletions are enriched in a population referred for microarray testing. TAD boundary disruption was considered rare if a deletion was present in less than 1% of our population which equates to 5 or fewer patients. These rare regions were investigated for phenotypic trends among patients with shared disruption. Phenotypic information was analyzed from both the microarray test requisition form and ICD-10 codes. Overlap was broadly assessed looking for similar indications such as developmental delay or dysmorphic features. Genomic landscape was then assessed using the UCSC genome browser. Tracks were uploaded to allow boundary region depletion to be visualized. Genes and regulatory elements were examined within a region ten times the size of the deletion. Enhancer activity was assessed by H3K27Ac presence. H3K27Ac is a histone marker which is well-established as a distinguisher of active enhancers (Creyghton et al., 2010). Gene content was analyzed using the number of Refseq genes. Refseq genes are those included in Reference Sequence (RefSeq) database which is an open access database consisting of annotated and curated DNA and RNA sequences and their protein products.

A TAD boundary disruption worksheet can be found in appendix A. TAD boundary regions were assessed for pathogenicity by comparing deletions among cases as well as genomic properties. TAD boundary regions were assessed for both gene dosage effect (GDE) and topological domain boundary disruption (TDBD). GDE was gauged based on presence of known Refseq genes in the deletion itself. TDBD was assessed by utilizing the UCSC browser. The patient’s deletion was visualized in relation to TAD boundary regions.

RESULTS

Study population

During the data extraction time period, 5033 patients had SNP microarray analysis. There were 312 controls analyzed from the Cincinnati Genomic Control Cohort (GCC). Demographics are available in Table 1. The cases were more likely to be male (3:2) while the controls had equal representation for number of males and females. The groups also differed slightly in race composition. The cases were mostly white (60.28%), with some individuals of African American descent (9.95%), and 66 multi-racial individuals (1.24%). Information was unknown or unavailable for 1068 (21.22%). The control cohort was
mostly white (83.65%), with some individuals of African American descent (16.00%) and 1 multi-racial individual. The cases had overall ages younger (median age = 3.98 years) than the controls (median age = 10.00 years), but since the diseases of interest onset early in life this difference was not thought to play a role.

We considered boundary data for both the hESC and IMR90 cell lines. hESC and IMR90 datasets were downloaded from the UCSC browser yielding 3062 domains in the hESC cell line and 2291 domains in the IMR90 cell line. From this, we identified 3061 boundary regions and 2290 boundary regions in each cell line respectively by defining the region between domains as a boundary region. Of these, 313 boundary regions were identified in both cell lines while 5038 were unique to their respective cell line. We defined overlapping boundary regions as those which share the start and stop position of the boundary. Both cell lines were utilized to assess rare TAD boundary regions. Table 2 depicts findings in the two cell lines as well as overlap between cell lines.

**Case/Control Comparisons**

Overall deletions among cases and controls were distributed similarly and no significant difference was found between them when using a minimum size of 18 kb. As TADs could be a function of the overall number of deletions, we then tested the relationship between homozygous TADs and homozygous deletions. The number of heterozygous TAD deletions was greater among controls (23.08%) than in cases (18.02%). However, homozygous TAD deletions was higher in cases compared to controls (figure 3, p<0.0001). Of cases, 291 (5.78%) had homozygous TADs deletions while only 3 (0.96%) of controls had homozygous deletions (table 3). We found having 1 or more homozygous TAD deletions was associated with more overall deletions in both cases and controls, but this relationship was stronger in controls (p=0.0004, figure 4).

**Rare TAD Analysis**

TAD boundary deletions were described as rare if they were present in less than 1% of the patient population but affected more than one case. For the rare TAD analysis, boundary region deletions were
investigated further if they were not present in the control cohort. The filtering process used to determine which TAD boundary regions were assessed is depicted in Figure 5.

Forty-five rare TAD boundary regions were identified: 23 unique to the hESC cell line, 26 unique to the IMR90 cell line, and 2 identified in both. Regions were determined to be of interest if phenotypic overlap was present, enhancer activity was in a relatively appropriate location, and no genes appeared to be potentially causative. Of the rare regions evaluated, 5 had phenotypic overlap among the patients warranting review of the regions (table 4). These regions include 7q11.23, 13q31.3, 14q32.33, 18p11.21, and 20q12. The regions on chromosome 7, 18, and 20 included common variants curated by DGV. The region on chromosome 13 had the exact same breakpoints in each of the five patients. The region on chromosome 14 has common variants seen in population-representative samples, but none overlap entirely with our patients’ deletions. Furthermore, most deletions of 14q are terminal and our patients’ were not. Variation within 14q is common and likely represents a hot spot within the genome, but gold standard variants within DGV and segmental duplications are terminal. For these reasons, we explored the region on chromosome 14 further.

Five patients had deletions near 14q32 (table 5). Three of five patients’ deletions (60%) were not considered clinically significant. One patient’s deletion was reported as a VU. The fifth patient’s deletion was called benign although it was significantly smaller than the other four deletions. Three of the four patients had developmental delay and four of the five patients had dysmorphic features. Genes in the largest deletion include INF2 and AKT1. INF2 is involved in Charcot-Marie Tooth and glomerulosclerosis. Genes in the overlapping regions include CRIP1 and CRIP2 which are transport protein genes. The deletions in two of the patients could result in an enhancer being moved closer to the PACS2 gene. PACS2 (Phosphofurin Acidic Cluster Sorting Protein 2) codes for a multifunctional sorting protein which controls communication between the endoplasmic reticulum and mitochondria possibly through ion channel trafficking. It is a homolog of PACS1 which is a trans-Golgi-membrane traffic regulator that is upregulated during embryonic brain development in humans and has low expression after birth (Simmen et al., 2005).
We assessed our study population for several known chromosome loci associated with TAD disruption. The regions we examined were those mentioned by Redin et. al. (Redin et al., 2017) as previously well-documented with human disease. We examined PAX6-WT1, TWIST1, SOX9, SHH, and POU3F4. Of these 6 distinct regions, none appeared in our cohort. We would expect that of 5345 individuals, some would have disruptions in these regions. The topological domains defined by these regions vary in size from 40 kb to 3.04 Mb. Our approach could have omitted individuals with these regions disrupted because their deletions may have been too large to meet our analytic parameters.

DISCUSSION

Previous studies have used animal models and patient populations to describe the clinical impact of TADs on human disease (Ibn-Salem et al., 2014; Lupianez et al., 2015; Ordulu et al., 2016; Redin et al., 2017). Those studies considered patient populations with large known deleterious deletions of multiple TAD boundaries (Ibn-Salem et al., 2014) and patients with balanced chromosomal rearrangements (Redin et al., 2017). In this study, we created an algorithm to identify smaller deletions with the logic that disruption of a single TAD boundary could be enough to result in human disease. The major finding of this study was that there is an enrichment of homozygous TAD boundary disruption in our patient population as compared to a control population. Although future studies are needed, this finding supports the importance of epigenetic regulation through TADs in human health and disease. Clinical laboratories should consider the incorporation of TAD analytic strategies into their workflow to increase the yield of diagnostics for patients with rare disease.

The work presented here focused on small deletions which disrupted exactly one TAD boundary and compared these deletions between cases and population-based controls. Previous studies have utilized public databases such as DECIPHER to assess CNVs (Ibn-Salem et al., 2014). In the Ibn-Salem paper, TAD boundary deletions of interest were limited to the single largest CNV in a patient determined to be known deleterious, likely due to its size. The average size of deletions in this study was 3.68 Mb for cases taken from DECIPHER. Topological domains are roughly one Mb in size and thus most deletions of this size would disrupt at least one boundary region. On average, the deletions investigated in their study...
encompassed 3.3 boundaries. Ibn-Salem et. al. accounted for this by defining a topological domain boundary disruption as a deletion which completely interrupted a boundary region but also had a tissue-specific enhancer upstream of a tissue-associated gene. The authors indicated that a shortfall of the study was the inability to see smaller deletions that could interrupt a TAD boundary, causing disease (Ibn-Salem et al., 2014).

Ibn-Salem et. al. examined two hypotheses for disease, gene dosage effect (GDE) and topological boundary domain disruption (TDBD). GDE pathogenicity of a deletion is size-dependent. A larger deletion is likely to encompass more genes and is statistically more likely to cause a gene dosage problem. TDBD pathogenicity is likely size independent as it depends more on the location of the deletion. Ibn-Salem found that 6.72% of the DECIPHER deletions may contribute to phenotypic dysregulation due to TDBD. Relatedly, they found that small deletions overlapping only one boundary showed rates of 10% TDBD, a rate higher than that when all deletions were considered.

The current study sought to examine those smaller deletions, reasoning that disruption of a single boundary could be enough to cause disease through enhancer adoption. Dixon et. al. defined boundary regions in hESC and IMR90 cell lines through chromatin conformation capture techniques. When analyzing the genome, data was binned using 40kb due to resolution limitations of the computer algorithm. Our study defined boundaries as the region between two domains, as localized by Dixon et. al., rather than a genomic coordinate. The TAD boundary query identified patients who had a deletion which encompassed an entire boundary region but did not extend into a subsequent boundary region. The size of boundary regions differed between cell lines. Due to the average boundary size being larger in the IMR90 cell type, patients needed to have much larger deletions to be identified through the IMR90 cell type when compared to the hESC cell type.

Most of the boundary regions utilized in the study presented here were unique to either the hESC or IMR90 cell type. However, previous studies have indicated that there is significant overlap between cell lines and that TADs are consistent across cell lines (Dixon et al., 2012; Ibn-Salem et al., 2014). In the Dixon study, domain coordinates were defined in both the hESC and IMR90 cell types. Of the regions
identified, 1289/2471 (52.17%) overlapped in both cell types. Our findings did not demonstrate significant overlap in boundaries defined across cell lines, identifying only 5.85% of boundary regions in both cell types. Further characterization of the findings in each cell line is available in table 2. In order to be considered overlapping here, the start coordinates of one domain and next subsequent domain would have needed to match exactly. In previous work, the binning process may have offered more generous definitions of overlap. Dixon et. al. performed Hi-C experiments and assessed chromatin interaction frequencies. We used previously reported positions of topological domains to define boundaries and assess patients for disruption. These two methodologies led to considerably different findings and highlight the importance of defining metrics such as boundary location in further studies of TADs.

The role of TADs in both epigenetic regulation and clinical disease are being explored, but studies investigating the clinical impact are sparse. Strides are being made in identifying specific effects of disrupted TADs and developing approaches to assess clinical relevance of genomic variation. Numerous considerations must be made when assessing TADs and their implications in disease. TADs are currently being assessed for clinical disease by examining chromosomal breakpoints and examining intrachromatin interactions. Redin et. al. examined TAD disruption within balanced chromosomal rearrangements (BCAs) detected by karyotyping in individuals with congenital and/or developmental anomalies. The majority of their participants had normal chromosomal microarray testing (62.6%). Authors used Hi-C contact data to localize TAD disruption and assess its pathogenicity. In their study, the authors only considered genes known to have a loss of function effect (Redin, et al., 2017). Ordulu et. al. looked at 10 prenatal participants each of whom had BCAs identified on karyotype and normal microarray testing. In these prenatal samples, genes were considered if they were known to be likely haploinsufficient (haploinsufficiency indices <10%), hemizygous or imprinted genes, disrupted genes due to the breakpoint, or genes associated with the phenotype of the participant (Ordulu, et. al., 2016). Both of these studies identified increased yield when considering TADs as a pathogenic mechanism for disease, with the first identifying 7.3% of BCAs to be associated with regulation disruption (Redin, et al., 2017). Additionally, Ibn-Salem proposed a mechanism for bioinformatics signatures to identify TDBD. TADs
could be identified by looking for TAD disruption and enhancers near the disruption. GDE, as an alternative mechanism for disease, could be excluded by ruling out those with one or more genes in the CNV and plausible enhancer activity absent (Ibn-Salem et. al, 2014). These previous studies as well as proposed mechanisms for analysis offer a general workflow for assessing CNVs for pathogenicity. A similar workflow for TAD CNVs, a subset of genomic CNVs, can be utilized to develop an assessment of TADs for pathogenicity.

While laboratories differ in their approach to the assessment of CNVs, the ACMG guidelines suggest a framework for assessment (South et al., 2013). The currently workflow for CNV assessment takes into consideration the size of the CNV, the gene content, and the frequency of the CNV in disease and healthy populations to determine implications for human disease. In addition, laboratories may wish to create additional work flows for the evaluation of TAD CNVs.

Because CNVs which interrupt TADs are a subset of CNVs as a whole, factors to consider in assessing pathogenicity of TAD disruption are similar to that of CNVs. The two main aspects to consider are the frequency of the TAD disruption in case and control populations and the likelihood of pathogenicity. In order to assess pathogenicity, the genomic properties of the landscape surrounding the TAD and the implications of the alteration should be considered. Regions of interest for TAD pathogenicity include those where recurrent disruption of a boundary is occurring at a locus in an affected population with a consistent phenotype. The phenotype of the loci with regulatory interruption is similar to that of a phenotype associated with haploinsufficiency of a nearby gene, and ultimately a demonstrated gene expression effect as a result of the genomic interruption. Disrupted regulation could result in phenotypic variability through two mechanisms either through deletion of an enhancer or novel enhancer interactions, a mechanism known as enhancer adoption. Deletion of an enhancer could lead to gene suppression due to absence of the enhancer to facilitate transcription. Enhancer adoption could lead to a gene being expressed at a time or place when it typically is not activated.

Next generation sequencing at chromosomal breakpoints could offer insight into TAD variation. In addition, many metrics need to be considered when evaluating the clinical implications of TADs such as
cell type, patient phenotype, and genomic analysis techniques like laboratory methodology and technology utilized. There are potential analytic workflows to incorporate in a clinical setting in order to integrate knowledge of TADs and their roles. We propose a process for exploratory TAD pathogenicity analysis.

A TAD boundary disruption worksheet can be found in appendix A. TAD boundaries of interest were identified by presence in more than one individual, each of whom had no clinically significant SNP microarray result. Deletions seen in our control population were not considered for analysis. The patients’ phenotypes were compared and assessed for overlap based on either similar symptoms or similar systems known to be developmentally related such as ectoderm cell derivatives. If there was a phenotypic overlap which warranted further exploration, the genomic landscape was then examined.

As the genomic landscape was analyzed, two pathogenicity hypotheses were explored. The gene dosage effect was explored by assessing the function of genes in a deletion. Topological Domain Boundary Disruption (TDBD) was assessed by examining enhancer presence near the boundary region. Enhancers were assessed by presence or absence of the H3K27Ac marker. H3K27Ac is a modification to histone H3. It is associated with higher levels of transcription and thus indicates presence of an active enhancer. H3K4Me3 is another modification of histone H3 and it is known to be associated with promoters. We looked for the presence of H3K27AC to identify active enhancers as well as the absence of H3K4Me3 to rule out the likelihood that the region was a promoter. We looked at the presence of these markers across cell types including blood (GM12878), stem cell (H1-hESC), muscle (HSMM), endothelial (HUVEC), leukemia (K562), skin (NHEK) and lung fibroblasts (NHLF). We considered whether a deletion may move an enhancer into proximity of a gene that has potential to be involved in disease pathogenesis as well as whether the deletion of an enhancer may inhibit gene expression. We assessed whether the deletion seen in other control populations such as the Database of Genomic Variants, or other case populations. This process offers an approach to TAD CNV assessment to be included in a clinical workflow through the addition of relevant tracks in analytic software and through education of analytic staff.
This study has several limitations. Due to the exploratory nature of this study, a broad approach was utilized. Phenotypes of patients included in the study are diverse and may have homogenized any trends that exist within specific subtypes. The phenotypic information that was available was from test requisition forms and ICD-10 codes, two sources which are limited by provider report and are not collected in a standardized manner. A second limitation of our study was that certain structural rearrangements were not examined due to the limitations of microarray testing. Balanced inversions and balanced translocations cannot be appropriately assessed through microarray, though other studies cite these as sources of TAD disruption (Lupianez et al., 2015; Redin et al., 2017). Furthermore, the effect of duplications cannot be assessed by microarray alone since it is impossible to tell where the duplicated material has been inserted in the genome. Another limitation of this study was that the number of cases heavily outweighed the number of controls. Future studies may consider using larger control datasets that are publically available. Future studies would also benefit from utilization of the same microarray platform. The 850K, Omni1, and OmniV chips have unique markers and vastly different spacing which creates noise that is difficult to assess across chips. Failure to account for differences between the cases and controls based on both sample size and microarray platform could have limited our ability to find an effect. A prospective study in which TAD disruption is examined in patients undergoing microarray could uncover trends that were not able to be examined in the present study. A prospective study would also aid in collecting more thorough phenotypic information for assessment of similarities. Though this study did not confirm any TAD boundary deletions as causative for clinical disease, some of the findings in the rare disease population supported by a body of literature suggest that these types of disruptions are responsible for clinical phenotypes. Based on this, we suggest that systematic clinical microarray review of TAD boundaries could uncover genomic variation disease pathways that have previously been unreported and improve diagnostics.

The findings of this study support the role of TADs in clinical disease and demonstrate the complexity of assessing the implications of TAD variation within the human genome. Reports of TAD disruption have been limited to specific clinical indications such as limb abnormalities and comparison of
TAD disruption to individuals without clinical disease has been limited. This study is the first, to our knowledge, to demonstrate an increase in the number of homozygous TAD boundary disruption in a clinical population of patients undergoing microarray for a variety of indications. Our findings showed no significant difference between cases and controls in heterozygous deletions which disrupt TAD boundary regions. This demonstrates a need to further investigate TADs, but more importantly highlights the need to develop a process in assessing TADs and heavily consider what factors may or may not be playing a role in TAD disruption. Study metrics such as TAD boundary location, CNV size, cytogenetic technology, and cell type are all important components to consider when developing future studies.

There are additional studies from which our study population could benefit. One approach would be assessing individuals in our study population who have large deletions that include minimal genes. This subset of our population would likely not have an identified mechanism for disease and TAD disruption could potentially clarify the mechanism of disease. Our study population has a noted enrichment of homozygous deletions which warrant further investigation as well. After assessing gene sparse regions and individuals with homozygous deletions, data collected in this study should be revisited. There is likely something in our approach that is limiting the ability to observe the influence of TAD disruption on rare disease. We would expect to find that in some cases a single deletion or mutation within a TAD boundary could result in disease.

The clinical applications of TAD boundary disruptions are becoming better understood as technology becomes more accessible and data becomes more widely available. Analyzing TAD disruption posits a new lens through which to view human genome variation. Microarray review of clinically significant TADs could uncover genomic variation disease pathways that have previously been unreported. The assessment of TADs and their impact on disease is complex and requires understanding of cytogenetic technology, epigenetic factors across the genome, and clinical manifestations of diseases of interest. Future evaluation of additional cohorts and use of diverse research design will aid in increasing the knowledge of TADs and their role in human disease.
REFERENCES


FIGURES

Figure 1. Topological Domain Boundary Disruption. TAD boundary disruption through alteration of physical spacing within the genome could lead to inappropriate regulation between long range regulatory units and genes.

Figure 2. Defining clear borders for TAD boundary deletions. TAD boundary deletions met criteria if they encompassed an entire boundary domain (light gray) but did not extend beyond the two adjacent TADs (dark gray). Solid lines represent deletions which would be picked up by the TADs query. Dotted lines represent deletions which would not have met the TADs query specifications.
Figure 3. *Homozygous and heterozygous deletions among cases and controls. The number of TAD deletions is greater in cases when heterozygous deletions are considered and the number of TAD deletions is greater in controls when homozygous deletions are considered.*

![Bar chart showing proportion with TAD deletions for cases and controls](chart.png)
Figure 4. Comparison of deletions among cases and controls. Number of homozygous TADs based on presence of homozygous TADs and case/control status. Logistic regression showed that both presence of homozygous TAD deletions and case/control status are predictors of the number of homozygous TAD deletions.
**Figure 5.** Filtering of potential TAD boundary regions of interest for rare analysis.

**hESC Rare Analysis Filtering Process**

- Shared boundary region deletions in more than one patient
  - N=62

- Patients with TAD boundary region disruption (CNV 0 or 1)
  - N=903

- Shared boundary regions disrupted in less than 1% of patients
  - N=38

- Patients with boundaries disrupted in less than 1% of patients
  - N=119

- Boundary regions disrupted in patients with normal CNV calls
  - N=29

- Patients with normal calls for CNV
  - N=99

- TAD boundaries not disrupted in controls
  - N=23

- Patients with TAD deletions not disrupted in controls
  - N=74

**IMR90 Rare Analysis Filtering Process**

- Shared boundary region deletions in more than one patient
  - N=50

- Patients with TAD boundary region disruption (CNV 0 or 1)
  - N=318

- Shared boundary regions disrupted in less than 1% of patients
  - N=39

- Patients with boundaries disrupted in less than 1% of patients
  - N=118

- Boundary regions disrupted in patients with normal CNV calls
  - N=32

- Patients with normal calls for CNV
  - N=95

- TAD boundaries not disrupted in controls
  - N=26

- Patients with TAD deletions not disrupted in controls
  - N=73
### Table 1. Demographics of cases and controls.

<table>
<thead>
<tr>
<th></th>
<th>Case population</th>
<th>Control population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total participants</strong></td>
<td>5033 (94%)</td>
<td>312 (6%)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>480 (10%)</td>
<td>50 (16%)</td>
</tr>
<tr>
<td>American Indian/Alaskan Native</td>
<td>6 (0.12%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Asian</td>
<td>74 (1.47%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>5 (0.10%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>White</td>
<td>3034 (60.28%)</td>
<td>261 (83.65%)</td>
</tr>
<tr>
<td>Other</td>
<td>1434 (28.49%)</td>
<td>1 (0.32%)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>171 (3.40%)</td>
<td>3 (0.99%)</td>
</tr>
<tr>
<td>Non-hispanic</td>
<td>3793 (75.36%)</td>
<td>309 (99.04%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1069 (21.24%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2009 (39.91%)</td>
<td>156 (50%)</td>
</tr>
<tr>
<td>Male</td>
<td>3007 (59.74%)</td>
<td>156 (50%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>22 (0.44%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>3.98 years</td>
<td>10.00 years</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>1.05-10.48 years</td>
<td>7-13 years</td>
</tr>
</tbody>
</table>

### Table 2. Deletions in hESC and IMR90 cell lines

<table>
<thead>
<tr>
<th></th>
<th>hESC</th>
<th>IMR90</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Boundary Regions</strong></td>
<td>3060 (57.20%)</td>
<td>2289 (42.79%)</td>
<td>313 (5.85%)</td>
</tr>
<tr>
<td><strong>Boundary regions with TAD deletions present</strong></td>
<td>813 (26.57%)</td>
<td>587 (25.64%)</td>
<td>164 (52.39%)</td>
</tr>
<tr>
<td><strong>Deletions across TADs</strong></td>
<td>4145 (66.15%)</td>
<td>2121 (33.84%)</td>
<td>501 (7.99%)</td>
</tr>
<tr>
<td><strong>Cases with TAD deletions</strong></td>
<td>2200 (43.71%)</td>
<td>1331 (26.44%)</td>
<td>355 (7.05%)</td>
</tr>
<tr>
<td><strong>Controls with TAD deletions</strong></td>
<td>209 (66.99%)</td>
<td>174 (55.77%)</td>
<td>30 (9.62%)</td>
</tr>
</tbody>
</table>
Table 3. Deletions among cases and controls.

<table>
<thead>
<tr>
<th></th>
<th>Heterozygous Deletions</th>
<th>Homozygous Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>907/5033 (18.02%)</td>
<td>291/5033 (5.78%)</td>
</tr>
<tr>
<td>Controls</td>
<td>72/312 (23.08%)</td>
<td>3/312 (0.96%)</td>
</tr>
</tbody>
</table>

Table 4. TAD boundary regions selected for rare analysis

<table>
<thead>
<tr>
<th>TAD region</th>
<th>Patients with deletion</th>
<th>Shared phenotype</th>
<th>Reason for rule out</th>
</tr>
</thead>
<tbody>
<tr>
<td>7q11.23</td>
<td>5</td>
<td>Dysmorphic features</td>
<td>Seen in DGV</td>
</tr>
<tr>
<td>13q3.13</td>
<td>5</td>
<td>Fetal demise</td>
<td>Same exact breakpoints</td>
</tr>
<tr>
<td>14q32.33</td>
<td>5</td>
<td>Developmental delay, dysmorphic features</td>
<td></td>
</tr>
<tr>
<td>18p11.21</td>
<td>3</td>
<td>Fetal anomalies</td>
<td>Common variant curated by DGV</td>
</tr>
<tr>
<td>20q12</td>
<td>3</td>
<td>Mixed receptive-language disorder, lack of coordination</td>
<td>Common variant curated by DGV</td>
</tr>
<tr>
<td>Breakpoints</td>
<td>Size (bp)</td>
<td>Test Requisition</td>
<td>EHR Information</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Patient 1</strong></td>
<td>105080065-106067618</td>
<td>987553 Characterize known chromosome abnormality; dysmorphic features</td>
<td>No EHR information available</td>
</tr>
<tr>
<td><strong>Patient 2</strong></td>
<td>105771428-106188230</td>
<td>416802 Dysmorphic features; failure to thrive</td>
<td>No EHR information available</td>
</tr>
<tr>
<td><strong>Patient 3</strong></td>
<td>105781428-106016434</td>
<td>235006 Developmental delay; dysmorphic features; hypotonia; other</td>
<td>Developmental delay, mixed receptive-expressive language disorder, Speech/language delay, other speech disturbance, myopia, other forms of nystagmus, ocular torticollis, optic nerve atrophy, ptosis left eyelid, other encephalopathy, localization-related epilepsy, wilms' tumor, regurgitation, feeding difficulties, gastrocutaneous fistula, elevated blood pressure, nephrocalcinosis, acquired scoliosis, chemotherapy induced cardiomyopathy</td>
</tr>
<tr>
<td><strong>Patient 4</strong></td>
<td>105900546-105970176</td>
<td>69630 Atrial septal defect; respiratory failure; other</td>
<td>Mixed receptive-expressive language disorder, speech disturbance, dysphagia, global developmental delay, gross motor delay, regular astigmatism, strabismic amblyopia, monocular esotropia, lack of coordination, gait abnormality, tightness of heel cord, muscle weakness, gastrostomy status, feeding by g-tube, attention deficit hyperactivity disorder, adjustment disorder with mixed disturbance of emotions and conduct type, dysplasia of kidney, periventricular leukomalacia, atrial septal defect, ostium secundum, s/p fontan procedure, prematurity, dysphagia, oropharyngeal phase, S/P Fontan, acute kidney injury, Ebstein anomaly</td>
</tr>
<tr>
<td><strong>Patient 5</strong></td>
<td>105943679-106190145</td>
<td>246466 Developmental Delay; Dysmorphic craniofacial features; Macrocephaly</td>
<td>Global developmental delay, fine motor delay, decreased movement of arm, muscle weakness, sensory processing difficulty, strabismic amblyopia, hyperopia, accommodative esotropia, lack of coordination, abnormal posture, hyperreflexia, macrocephaly, arachnoid cyst, other symbolic dysfunctions</td>
</tr>
</tbody>
</table>
Appendix A. Rare TAD analysis worksheet.

Patient Phenotype

1. Do the patients have a normal microarray result?
2. Are there any other known genetic explanations for the patient’s clinical indications?
3. Do the patients have an overlapping phenotype? If not, does and EHR review help rule in/rule out a phenotype?
   a. Clinical Indications
   b. EHR Extract
4. Is there a particular cell type that might be associated with the patient phenotype?

PRIORITIZED or NOT PRIORITIZED

Gene Dosage Effect (GDE)

1. What genes are in the deletion? What do those genes do?

Topological domain boundary disruption (TDBD)

1. Are there markers of enhancers (H3K27Ac) near the boundaries of the deletion? (If the deletion is associated with a particular phenotype, consider cell line to use for H3K27Ac mark.)
2. What genes are close by (within 10X)?
3. Would the deletion place an enhancer in close proximity to a gene it does not typically regulate?
4. Would the deletion remove an enhancer that might be important for a nearby gene?

Both

1. Have overlapping deletions previously been reported?
   a. In control populations?
   b. In case populations?

What are the next steps?