The Pennsylvania State University

The Graduate School

EXPLORING SEQUENCE ARCHITECTURES AT FLANKING REGIONS OF COPY NUMBER VARIANTS

A Thesis in

Bioinformatics and Genomics

by

Hossain Mohammad Nayeim Khan

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2019

The thesis of Hossain Mohammad Nayeim Khan was reviewed and approved* by the following:

Santhosh Girirajan
Associate Professor of Genomics
Thesis Adviser

Anton Nekrutenko
Professor of Biochemistry and Molecular Biology

Yifei Huang
Assistant Professor of Biology

George (PJ) Perry
Associate Professor of Anthropology and Biology
Chair of Bioinformatics and Genomics Graduate Program

*Signatures are on file in the Graduate School

ABSTRACT

Copy number variations (CNVs) represent a subtype of structural variations where a portion of the genome is deleted or duplicated resulting in copy number changes of genes within the region. These copy number changes happen due to aberrant recombination between specific repeat architectures facilitated by a degree of sequence homology present within these repeats. In this study, a comprehensive and an exhaustive catalog of 150,802 copy number variants have been leveraged to explore the flanking regions of copy number variants and enumerate the abundances of different subtypes of repeat architectures. Thus, looking at the flanking regions of copy number variants may indicate which repeat architecture is relatively more abundant and may contribute more to the formation of larger fraction of copy number variants. Alu elements (~67.12 percent) were found to be the most abundant repeat architecture followed by segmental duplications (~22.59 percent) and L1 LINE elements (~21.71 percent) in both upstream and downstream flanking regions. Furthermore, AluY, AluSx and AluSx1 were found to be relatively more abundant among other 34 different Alu Subtypes at the flanking regions. Interestingly, the newer AluY subfamilies were the least abundant in the flanking regions. In the case of L1 elements, 114 different L1 LINE subtypes were present and exhibited a heterogenous pattern in these flanking regions. Among those 114 subtypes, 4 specific L1 subtypes L1MB3, L1MB5, L1MB7, L1MB8 were relatively more abundant than the rest. Strikingly, all these 4 subtypes are relatively old L1s and have existed even before the evolution of primates. To assess the pathogenicity, CNVs were screened for the presence of dosage sensitive genes and the presence of transcription factor specific regulatory elements. Only 3.61 and 46.85 percent copy number variants had exons of dosage sensitive genes and transcription factor specific regulatory elements within themselves respectively. Thus, majority of the copy number variants has a higher likelihood of being benign which is indicative of the dataset since the CNVs in the dataset are representative of normal population. However, the tissue and cell-type

specificity of regulatory elements have not been considered for the screening which is a major limitation. Next, CNVs with 50 percent or more overlaps have been stratified to form copy number variable regions (CNVRs) to see if there are regions across each chromosome that gets more than the usual amount of CNV hits. Majority of the CNVRs had CNV hits of 15 or below except a few which had hits of 100 or more. Surprisingly, some of the maximal CNV hits CNVRs for each chromosome were depleted in segmental duplications but had overrepresentations of different Alu subtypes. In these segmental duplications depleted regions, it is highly likely that Alu elements might be the main repeat architecture sensitizing the genome to undergo more rearrangements which consequently makes the CNVRs get CNV hits of 100 or more.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to start by expressing my deepest gratitude to my graduate advisor, mentor and committee chair Dr. Santhosh Girirajan for taking the time out to guide and advise me on how to proceed with my thesis in the last 6 months. His continuous support, motivation, feedback and patience enabled me to complete my research within this short timeframe. I am glad that I got to learn from an immensely knowledgeable person like Dr. Girirajan.

In addition to my advisor, I would like to extend my heartfelt gratitude to Dr. Anton Nekrutenko who is also part of my thesis committee. Despite being one of the busiest faculty, Dr. Nekrutenko still managed some time to help me with my thesis by giving me insightful comments and ideas on how to do the analysis. Besides Dr. Nekrutenko, I would also like to thank another faculty in my committee Dr. Yifei Huang for also taking out time to advise and encourage me with my research. I am indebted to all my committee members for their constant support and encouragement.

My sincerest thanks also go out to the former chair of the Bioinformatics and Genomics Program Dr. Cooduvalli S. Shashikant for his guidance and advice in the past seven months. In addition, I am also indebted to my fellow lab mates Matthew Jensen, Vijay Kumar, Tanzeen Yusuff, Maitreya Das, Phoebe Ingraham, and Laura Rohan for their constant support and motivation.

Finally, I would like to thank my parents without whom I wouldn't be where I am today. I owe everything to them and without their love, support and encouragement, my existence would be meaningless. My parents are my greatest gift from the Almighty and I can't thank Him enough for that.

INTRODUCTION

Copy number variations represent a subtype of structural variations where a portion of a chromosome is deleted or duplicated resulting in copy number changes of genes within the region (Girirajan et al., 2011). New advances in genomic technologies such as whole exome sequencing and whole genome sequencing have made it easier to discover and characterize the copy number variations in the genome. Due to the high throughput nature of these technologies, the statistical power of these studies has significantly increased too. From these high throughput studies, it is seen that the contribution of copy number variations to human genetic diversity and rare disease is significant (Clancy, 2008). The size of copy number variants varies between small scale variation to large scale variations ranging from as small as a few hundred base pairs to a few million base pairs. However, the smaller copy number variants are much more common in the population than the larger ones (Girirajan et al., 2011). The larger copy number variants are rare in the population but are associated with clinical phenotypes and complex genetic disorders such as 1.5 mbp deletion of 15q13.3 (Sharp et al., 2008). In the events of complex genetic disorders, it has been seen that larger copy number variants are implicated mainly for deleterious phenotypes due to the presence of several genes and regulatory elements (Girirajan et al., 2011).

Both small and large copy number variations have been implicated in pathogenic conditions. An example of small copy number variations would be the rare 9q sub telomere deletion syndrome (Stewart and Kleefstra, 2007). Children with this deletion manifest phenotypes such as multiple congenital anomalies, developmental delay, hypotonia and dysmorphic features (Stewart and Kleefstra, 2007). The size of the CNVs associated with this telomeric deletion syndrome can vary between 200,000 bp till 3,000,000 bp (Stewart and Kleefstra, 2007). However, the

pathogenicity associated with this syndrome is not related to the size of the copy number variants (Stewart and Kleefstra, 2007). The phenotypes manifest because of haploinsufficiency of *EHTM1* gene (Stewart and Kleefstra, 2007). On the other hand, Smith-Magenis syndrome is a syndrome where the disease manifests because of a 3.7 megabase pair chromosomal deletions in chromosome 17p11.2 (Elsea and Girirajan, 2008). The common phenotypes that are associated with Smith Magenis syndrome are distinctive craniofacial and skeletal features, cognitive impairment, global developmental delay and mild to moderate mental retardation (Girirajan et al., 2006). Researchers believe that the symptoms or phenotypes of Smith-Magenis syndrome are due to deletion or loss of function mutations of a specific dosage sensitive retinoic acid-induced gene (*RAI1*) (Elsea and Girirajan, 2008). The deletions that have been found associated with this disease always contained the *RAI1* gene (Elsea and Girirajan, 2008). Interestingly, duplication of the 17p11.2 segment, an extra copy of the *RAI1* gene causes a similar condition known as Potocki-Lupski syndrome (Zhang et al., 2010). This example demonstrates the importance of changes in copy number of segments and how it can lead to deleterious phenotypes in individuals. In addition to copy number variants being associated with the same sets of clinical and deleterious phenotypes, some copy number variants are also implicated towards phenotypic heterogeneity. Some examples of these type of copy number variants are in the regions on chromosomes 15q13.3, 16p11.2, and 15q11 that are associated with phenotypic heterogeneity in the cases of epilepsy, autism, intellectual disability (Girirajan et al., 2011).

It is well known now that specific structural architectures in the genome sensitize segments of the genome to undergo rearrangements more often than usual (Chen et al., 2014). In turn, increased rearrangements increase the likelihood of getting more copy number variants. The structural features include genomic repeats like Alu elements, LINE elements, Segmental

duplications which provide sequence homology that is needed for genomic rearrangements (Cardoso et al., 2016).

Segmental Duplications are stretches of DNA sequences that have nearly identical sequences (90-100%) and can span 10 – 400 kb of genomic DNA (Hastings et al., 2009). They exist in multiple locations in the genome as a result of duplication events that have happened millions of years but not so far back in the past because it can still be seen that they are highly identical and related to each other (Hastings et al., 2009). If the segmental duplications are oriented in the same direction during non-allelic homologous recombination, microdeletions and microduplications can form ranging from few kilobases to megabases in size (Chen et al., 2014). These microdeletions and microduplications can cause syndromic disorders as well. For instance, the 1.4 Mb duplication of the *PMP22* gene and the 22q11.2 deletion syndrome are few of the notable copy number variants mediated by segmental duplications (Chen et al., 2014).

Another class of structural repeats is high copy repeats representing a large fraction of the human genome (Cardoso et al., 2016). One of the most common types of high copy repeats are the retrotransposons such as Alu elements, and L1 LINE elements. Alus are retroelements or retrotransposons that make up almost around 10 percent of the human genome (Chen et al., 2014). They are non-autonomous and use the retrotransposition machinery of L1 elements to hop around the genome by "copy and paste" mechanism. The reason Alu elements can facilitate rearrangements is due to a certain degree of sequence homology between different Alu subtypes that provides the substrate necessary for rearrangements (Chen et al., 2014). Thus, Alu elements are one of the important repeat architectures in mediating both benign and pathogenic copy number variants (Chen et al., 2014). Some of the notable Alu-Alu mediated pathogenic copy number variants are

deletion of *BRCA1* in breast cancer, duplications of *SNCA* gene in autosomal dominant Parkinson's disease among others (Cardoso et al., 2016).

L1 LINE elements are another type of retrotransposons representing almost 20 percent of the human genome. They are known to increase genome instability through mediating rearrangements between genome segments (Chen et al., 2014). 83 percent of the human genome has been found to be susceptible to LINE-LINE mediated rearrangements which consequently increases the likelihood of getting unbalanced structural variants like copy number variants (Cardoso et al., 2016).

In addition to sequence homology mediated Non-Allelic homologous recombination, replication-based mechanisms can give rise to copy number variants as well (Chen et al., 2014). Genomic repeats with high sequence identity in neighboring loci align with each other and the subsequent crossover results in the formation of secondary structures (Chen et al., 2014). The formation of secondary structure causes the replication fork to stall which consequently results in replication resuming from the incorrect locus (Hastings et al., 2009). Thus, in this way, the cascade of replication errors leads to rearrangements and formation of copy number variants. This model of copy number variant formation is known as Fork Stalling and Template Switching (FoSTeS) (Hastings et al., 2009). Like FoSTeS, another mechanism involved in mediating rearrangements are Microhomology mediated break induced repair (MMBIR) (Hastings et al., 2009). The similarity among all the mechanisms mentioned till now is the presence of sequence homology that initiates or stimulates rearrangements. The next mechanism needs little or no homology and is termed Non homologous end joining. This mechanism is one of the main mechanisms for repairing double stranded breaks and for ensuring genomic integrity (Hastings et al., 2009). While repairing the

double-stranded break, deletions or insertions might take place at the joint point (Hastings et al., 2009).

The human genome is found to be heterogenous in terms of copy number variants when considered at the population level. *Itsara et al.* documented the landscape of copy number variants which enabled us to acquire some important insights into the size and the frequency of copy number variants in general population (Itsara et al., 2009). They found out that approximately 65 percent to 80 percent of individuals carry a variant of at least 100kbp in size, 5 to 10 percent of individuals carry a variant that is at least 500kbp in size, and only 1 percent of individuals carry a variant that is larger than 1Mbp in size (Itsara et al., 2009) . The paper further claimed that rare copy number variants are relatively gene-rich when compared to normal copy number variants (Itsara et al., 2009). In another study published in Science, *Sudamant et al.* looked at 159 human genomes to characterize the diversity of human copy number variation. In this study, it was seen that as copy number variants increase in size, their population frequency decreases (Sudmant et al., 2010). Furthermore, 47 percent of the large copy number variants (>50 kbp in length) were found to be common since they were observed in more than 5 percent of the genomes studied. And majority of these large variants were found to overlap segmental duplications that in turn influenced copy number variant frequency. Lastly, a copy number baseline was defined in this study by identifying 173 segmentally duplicated regions that had copy number changes greater than that of reference genome in the majority of the genomes studied (Sudmant et al., 2010).

The pathogenicity associated with copy number variants is dependent on more than one specific mechanism. Disruption of gene function resulting in the manifestation of phenotypes, disruption of epistatic interactions between genes, disruption of chromatin orientation of three-dimensional organization, and interference with regulatory elements are few of the common ways

for manifesting clinical phenotypes (Klopocki and Mundlos, 2011). In many cases, it depends on the presence of genes and regulatory elements in these aberrations. Additionally, changes in copy number can also affect the transcriptional and translational profile in the segment of variable copy number (Klopocki and Mundlos, 2011). However, in many cases, the genic or intergenic copy number variants do not produce deleterious or recognizable phenotypes. This is because some genes can cope or tolerate a change in their copy number (Rice and McLysaght, 2017). There are instances where exons of certain genes have been deleted with no significant changes in the phenotype (Zarrei et al., 2015). Therefore, the role of the gene and their sensitivity to gene dosage determine whether copy number variant genes will exhibit clinical phenotypes or not (Rice and McLysaght, 2017). Thus, the consensus among the scientific community for pathogenicity of copy number variants is attributed to the sensitivity of gene to copy number changes in most cases (Klopocki and Mundlos, 2011). A well-known example of dosage sensitivity in pathogenic copy number variant is Charcot–Marie–Tooth neuropathy. Overexpression of a specific gene *PMP22* (peripheral myelin 22) causes the synthesis of the abnormal myelin sheath which leads to wasting of muscles in lower limbs during early adolescence (Robaglia-Schlupp, 2002). On the other hand, deletion of one of the *PMP22* genes causes low expression which in turn leads to episodic recurrent demyelinating neuropathy (Li et al., 2012).

Copy number variants in non-coding regions have also been found to disrupt regulatory elements that may indirectly result in the manifestation of clinical phenotypes due to perturbation of spatial and temporal gene expression. Pierre Robin syndrome is an example of a pathogenic copy number variant in the non-coding region. In Pierre Robin syndrome, deletions in the regulatory region of *SOX9* gene (17q24) perturbs the transcriptional landscape affecting the expression of *SOX9* gene (Klopocki and Mundlos, 2011). The perturbation of *SOX9* expression results in dysregulation of other genes involved in the development of skeletal muscle including the jaw

(Klopocki and Mundlos, 2011). This example clearly illustrates that the alterations in the regulatory landscape could lead to manifestations of severe phenotypes.

Over the last two decades, due to the advances in next generation sequencing, a big stride has been taken by the research community to accumulate or characterize the nature and pattern of genomic variations in the healthy human population (Girirajan et al., 2011). The availability of genomic variation catalogs in databases has facilitated better study design and better-educated hypothesis for the entire research community. Some of the notable initiatives are the sequencing of the human genome, completion of the 1000 Genomes project, completion of the HapMap project, and the NHLBI Exome Sequencing Project (ESP). This is illustrated by the accumulation of 660 million SNPs in DbSNP which is essentially a repository of single nucleotide polymorphism maintained by NCBI in collaboration with the National Human Genome Research Institute (Manzoni et al., 2016). However, cataloging larger genomic variants have been comparatively slow due to the added complexity of their structure and technological shortcomings. One of the biggest initiatives for characterizing larger genomic variants in healthy human population was undertaken by the Database of Genomic Variation (MacDonald et al., 2013). In the mid-2000s, when it just started, the database had around 1000 copy number variants and few translocations from a few hundred healthy individuals only (MacDonald et al., 2013). However, the copy number variant calls were from low-resolution microarray with low sensitivity. The combination of both low resolution and low sensitivity means the calls had many false positives and false negatives. To overcome this issue, the Database of Genomic Variants kept updating and curating the data (MacDonald et al., 2013). In recent years, due to technological advances of higher resolution microarrays and next generation sequencing, the quality of copy number variant calls has significantly improved. Since the datasets are from multiple experimental methodologies and protocols, the database of genomic variants developed pipelines to standardize the entries in the

database. The initial curation involves assessing the data to identify and remove the false positives from the datasets. Following that, the in-house pipelines of DGV performed multiple filtering steps. Some of the important filtering steps removed variants that incorrectly mapped to the mitochondrial genome or the Y chromosome in female samples, variants that are greater than 3 million base pairs and inversions that are larger than 10 million base pairs, and entries that overlap with gaps in the reference assembly. These are a few of the initial filtering steps that facilitated the standardization of all the entries in the database (MacDonald et al., 2013). Zarrei and colleagues also did a recent curation of the database (Zarrei et al., 2015). It implemented few stringent filters to improve the confidence on copy number variants calls. The main criterions on which the stringent filters were based were the genome-wide assessment and accurate breakpoint resolution. Therefore, studies that used lower resolution arrays, or that had more than one approach or platform were discarded. Datasets from studies that did next generation sequencing, arrays with at least one million probes and arrays with a targeted or custom CNV assay were kept only. Custom copy number probes were kept because the breakpoints were resolved well enough and had high accuracy. Other methodologies were not kept in these stringent datasets because of low sensitivity and low resolution. Thus, most of the copy number variants in this stringent dataset were from next generation sequencing methods and very few high-resolution arrays (Zarrei et al., 2015).

Another huge catalog of copy number variants representing genomic variations in healthy populations has been curated by ExAC. ExAC is a big consortium that had been established by a big group of researchers to aggregate and standardize exome sequencing data from various large sequencing projects. The data from this consortium represents 60,706 unrelated individuals from different ethnic background providing diversity to the datasets. For calling copy number variants on these datasets, various capture methods used in different large-scale participating studies were harmonized using an in-house pipeline developed by the analysis group (Ruderfer et al., 2016).

After standardization, copy number variant calls were generated by using XHMM. Few filtering steps were applied to XHMM calls which involved retaining variant calls that had quality scores of greater than or equal to 60, removing calls that had a frequency of more than 600 times etcetera (Ruderfer et al., 2016). The final datasets had a total of 126, 771 copy number variants that were overlapping GENCODE autosomal coding genes representing a population of 59, 598 individuals of different ethnic background (Ruderfer et al., 2016).

Very few studies have been conducted to characterize the flanking regions of copy number variants. In this study, datasets from both ExAC (126,771 copy number variants) and DGV (24,031 copy number variants) have been leveraged to look at flanking regions of copy number variants and enumerate the abundances of different subtypes of repeat architectures. Knowing the abundances of repeat architectures will give an idea as to what might contribute the most to the formation of these copy number segments. Alu elements (~67.12 percent) were found to be the most abundant repeat architectures at flanking regions of copy number variants followed by segmental duplications (~22.59 percent) and L1 LINE elements (~21.71 percent). Furthermore, AluY, AluSx and AluSx1 were found to be relatively more abundant among other 34 different Alu Subtypes. In addition, the newer AluY subfamilies were found to be the least abundant relatively at the flanking regions. On the other hand, L1 LINE subtypes had a heterogenous pattern and had 114 different L1 subtypes present in the flanking regions. Among those 114 subtypes, 4 specific L1 subtypes L1MB3, L1MB5, L1MB7, L1MB8 were relatively more abundant than the rest. Furthermore, these copy number segments have been screened for the presence of dosage sensitive genes and regulatory elements which might indicate the likelihood of these copy number variants being benign or pathogenic (Klopocki and Mundlos, 2011). Lastly, copy number variants with 50 percent or more overlaps have been stratified to form copy number variable regions (CNVRs) to see if there are regions across each chromosome that gets more than the usual amount of copy

number variant hits. Majority of the CNVRs had hits of 15 or below except a few which had hits of 100 or more. Next, CNVRs with maximal copy number variant hits across each chromosome were screened for the presence of different repeat architectures in order to see which repeat architectures might be contributing the most for causing rise to higher than usual number of copy number variants. Interestingly, some of the maximal CNV hits CNVRs for each chromosome were depleted in segmental duplications but had overrepresentations of different Alu subtypes. In these segmental duplications depleted regions, it is highly likely that Alu elements might be the main repeat architecture sensitizing the genome to undergo more rearrangements which consequently makes the CNVRs get hits of 100 or more. Lastly, the presence of exons of dosage sensitive genes and regulatory elements were also screened to assess the pathogenicity of these regions.

METHODOLOGY

Dataset

The datasets used for my study was downloaded from the FTP server of ExAC and DGV respectively. The total number of copy number variants for the analysis were 150,802 representing the general population.

Initial Summary Statistics of Copy number variant lengths

To get an initial idea of the length of the copy number variants, GNU AWK was used to generate the copy number variant lengths for each copy number variants. Then to get summary statistics of the copy variant lengths, datamash was used to get a sense of the distribution of the data. The summary statistics provided an insight into how the length of the copy number variants is distributed. Histogram and boxplots were drawn in python framework to get a better sense of the distribution of the length of copy number variants.

Screening for Alu Elements at flanking regions of the copy number variants

To look at the flanking regions, 5000 bp upstream and downstream of the copy number variants were extracted from the dataset using AWK. Two new files were created containing the upstream and the downstream flanking regions respectively. The datasets for Alu Elements were downloaded from UCSC genome browser using the group Repeats and track Repeat Masker.

Repeat Masker's dataset was used to screen two different categories of high copy repeats. Repeat Masker's dataset was then filtered to keep only the SINEs for this step of the analysis. Using a custom bash script, the flanking regions of the copy number variants were screened for the presence of different subtypes of ALUs. The output was in the format of a long data where each line represented one specific ALU subtype found in the corresponding flanking regions. To further analyze the output, the data was loaded onto a python framework. In order to calculate the abundance of different types of ALU subtypes, the data was reshaped from long or narrow format to wide format and standardized. A custom python script was used for changing the data shape and creating a standardized counts matrix for enumerating the abundance of different ALU subtypes in the flanking region.

Screening for L1 LINE elements at flanking regions of the copy number variants

The same dataset downloaded from the track Repeat Masker of UCSC genome browser was used for screening of L1 LINE elements at flanking regions of the copy number variants. But this time, the dataset was filtered to keep only the L1 repeat family. Using a custom bash script, the flanking region of the copy number variants were screened for L1 elements. The output was in the format of a long data where each line represented one specific L1 subtype found in the corresponding flanking regions. To further analyze the output, the data was loaded onto a python framework. In order to calculate the abundance of different types of L1 subtypes, the data was reshaped from long or narrow format to wide format and standardized. A custom python script was used to change the data shape and create a standardized counts matrix for each L1 subtypes enumerating the abundance of different L1 subtypes in the flanking region.

Screening for segmental duplications at flanking regions of the copy number variants

The dataset containing positions of segmental duplications in the genome was downloaded from the UCSC genome browser's track Segmental Dups. This dataset was used for screening the presence of segmental duplications at flanking regions of the copy number variants. The counts of segmental duplications at flanking regions for each copy number variants were enumerated and further analyzed in a python framework.

Screening for exons of dosage sensitive genes within Copy Number Variants

The dataset of dosage sensitive genes was downloaded from ftp server of ClinGen Dosage sensitivity Map. The dataset was furthered filtered to keep the genes that had a haploinsufficiency score of 3 or more. Haploinsufficiency of 3 or more for a gene indicates that the dosage sensitivity of the gene is potentially associated with clinical phenotypes. The filtered dataset was then used to screen for the presence of dosage sensitive genes in copy number variant sites.

Screening for Transcription Factor specific regulatory elements at Copy Number Variant Sites

ENCODE Txn factor ChIP dataset was used for screening the presence of transcription factor specific regulatory elements in copy number variants. The dataset contains transcription factor binding sites which have been aggregated from a large collection of ENCODE ChIP-seq data from 91 cell types for 161 transcription factors. MEME-ChIP motif identifier tool was used for identification of DNA binding motifs. In short, the dataset had been clustered and processed by ENCODE pipelines to harmonize transcription factor binding from a huge collection of ChIP seq experiments. The dataset was downloaded from the UCSC table browser. The dataset was filtered

to keep only the transcription factors that had a score of 1000 which indicated the highest signal strength of ChIP seq peaks. Using a custom bash script, the copy number variants were screened for transcription factor specific regulatory elements in the region spanning the copy number variants. The output was loaded onto a python framework where the data was further analyzed. In order to look at the binding profiles of all the transcription factors in copy number variants, a custom python script was used to generate the abundances of Transcription factors specific regulatory elements across all the 150,802 copy number variants.

Copy Number Variants stratified to Copy Number Variable Regions (CNVRs):

Since the mean length of all the copy number variants was 62,000 bp, the copy number variants that had 50 percent or more overlaps were stratified to form copy number variable regions (CNVRs) of 50,000 bp. The number of copy number variant hits were then counted in these copy number variable regions using a custom bash script.

Analysis of Copy Number Variable Regions (CNVRs):

The datasets generated was loaded onto Python framework. The number of hits of copy number variants in the stratified copy number variable regions was visualized using a histogram. The copy number variable regions were grouped by chromosome and then boxplots were generated to visualize the spread of the hits for each chromosome.

The copy number variable regions with the maximum number of copy number variant hit for each chromosome were determined. Using for loops and conditional statements, the genomic coordinates of those copy number variants regions were extracted.

Screening for dosage sensitive genes in copy number variable regions with maximal hits:

The dataset of dosage sensitive genes was downloaded from the ftp site of ClinGen Dosage sensitivity Map. The dataset was furthered filtered to keep the genes that had a haploinsufficiency score of 3 or more. Haploinsufficiency of 3 or more for a gene indicates that the dosage sensitivity of the gene is potentially associated with clinical phenotype. The filtered dataset was then used to screen for the presence of dosage sensitive genes in copy number variable regions with maximal hits.

Screening for ENCODE Transcription factor specific regulatory elements in Copy Number Variable Region (CNVRs) with maximal hits:

ENCODE Txn factor ChIP dataset was again used for screening the copy number variable regions (CNVRs) with maximal hits. The dataset contains transcription factor binding regions that have been aggregated from a large collection of ENCODE ChIP-seq data from 91 cell types for 161 transcription factors. MEME-ChIP motif identifier tool was used for identification of DNA binding motifs. The dataset was downloaded from UCSC Table Browser and filtered to keep only the Transcription factor specific regulatory elements that had a score of 1000. The score defines the strength of the signal from ChIP seq data. The copy number variable regions (CNVRs) were then screened for the presence of Transcription factor specific regulatory elements.

Visualizations:

All the visualizations generated for this study was done in Python framework using Matplotlib and Seaborn.

RESULTS

Initial summary statistics of Copy Number Variant Lengths:

To get a sense of the lengths of Copy Number Variants, summary statistics were generated to get a better representation of the length. There were in total 150,802 copy number variants. The mean for the copy number variant lengths was 62,969.44 base pairs and the median was 103,62 base pairs. Q1 and Q3 for the copy number variant length were 1744 base pairs and 45342 base pairs respectively. And the minimum was 10 base pairs and the maximum was 33,753,656 base pairs. From the initial summary statistics, it could be easily inferred that the distribution is highly skewed. To better visualize the distribution, the histogram was drawn to look at the data.

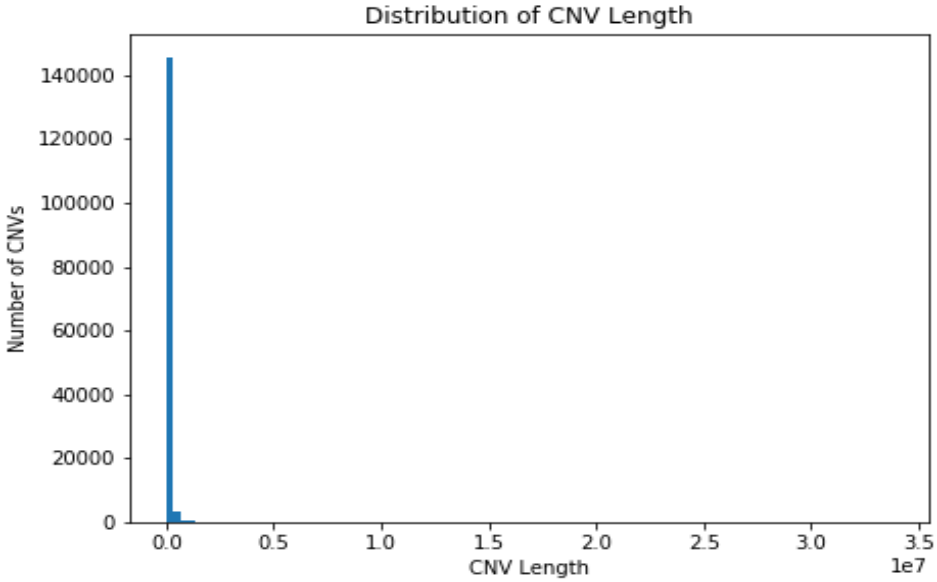

Fig 1: Distribution of Copy Number Variant Lengths

Looking at the initial histogram in fig 1, it is evident that the majority or more than 140,000 of copy number variants had lengths less than 100,000. However, since it is an extremely skewed

distribution, looking at a log-transformed histogram would give a better sense of the data

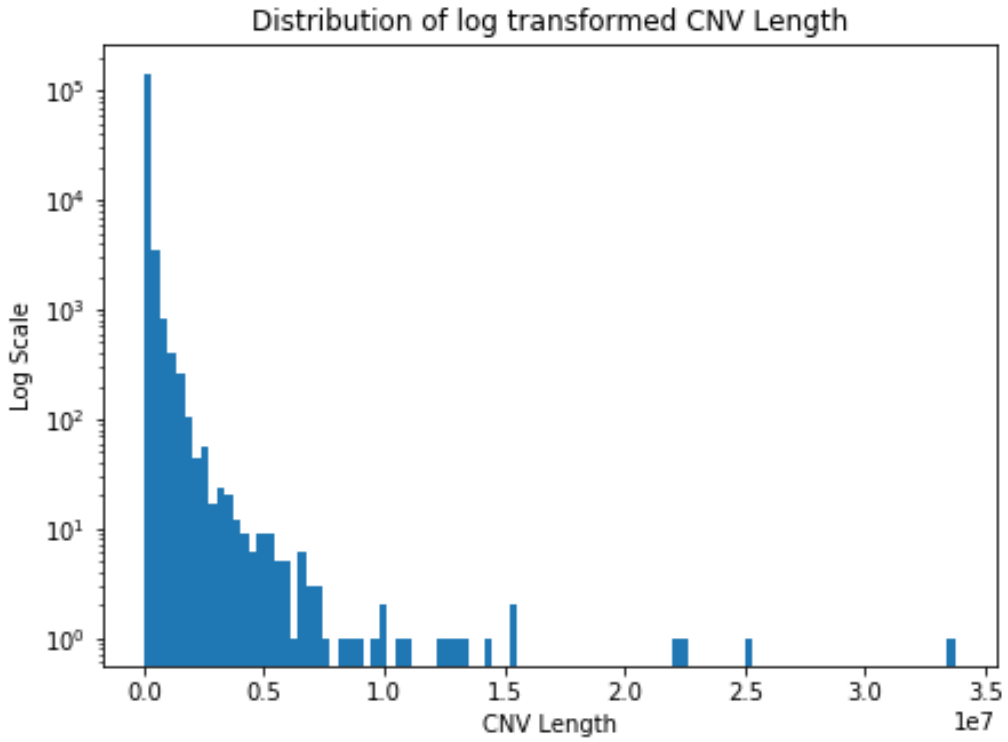

Fig 2: Distribution of Log Transformed Copy Number Variants Lengths

The log-transformed histogram in fig 2 gives a better representation of the distribution of copy number variant length. From this histogram, it can be inferred that there are very few copy number variants that have a length greater than 10,000,000 bp. Thus, this extreme copy number variant lengths causes the non-log transformed histogram to show such an extremely skewed distribution.

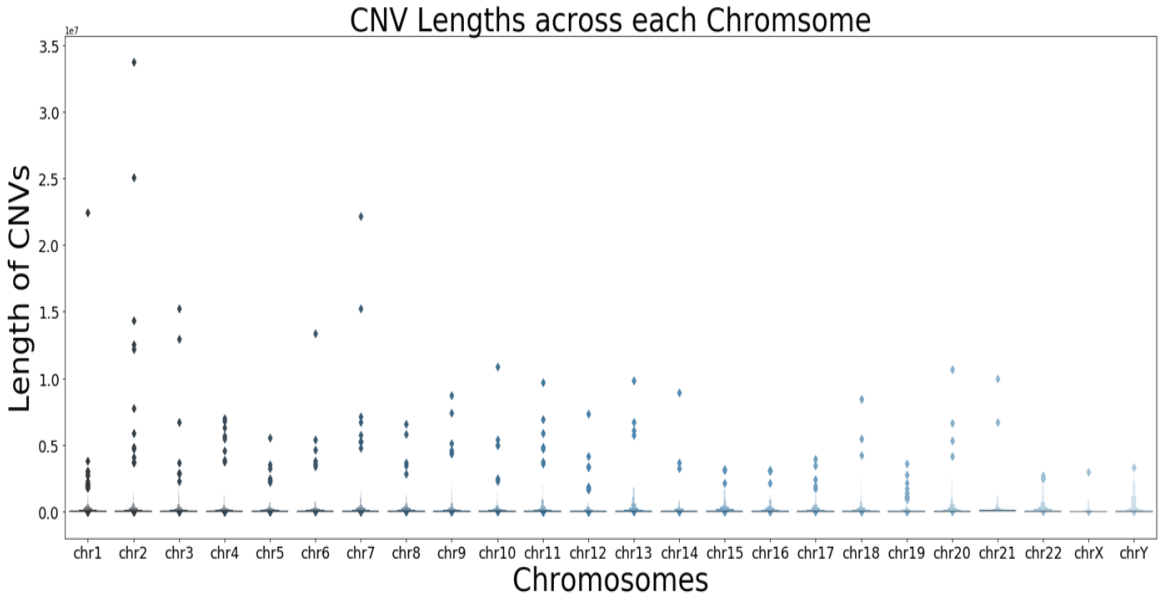

Fig 3: Spread of Copy Number Variant Lengths across each Chromosome

Next, the length of copy number variants was grouped by chromosomes and then their spread was visualized using a boxplot in fig 3. All the chromosomes had the majority of their copy number variant lengths below 100,000. However, every chromosome had few outliers that had lengths of copy number variants over 100,000 base pairs. Moreover, few chromosomes also had copy number variants in excess of 10,000,000 base pairs too.

Screening for the ALU elements at the flanking regions of copy number variants:

Out of 150,802 copy number variants, there were 102,257 copy number variants (67.80 percent) with different type of ALU subfamilies in their upstream flanking regions.

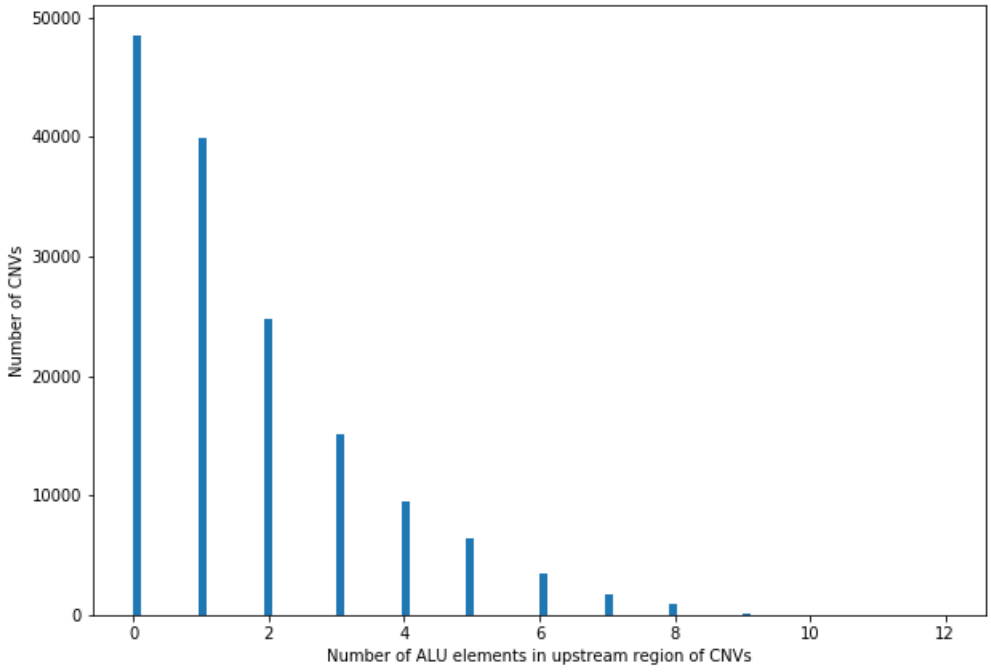

Fig 4: Distribution of Alu elements in upstream regions of Copy Number Variants

From fig 4, it can be inferred that the number of ALU elements present in the upstream regions varied between 1 to 9. However, out of 102, 257 copy number variants, around 39,000 had only one ALU elements followed by 26000 copy number variants that had two ALU elements in their flanking region. Thus, most of the copy number variants had three or less ALU elements present in their upstream flanking regions.

Looking at fig 5 for the presence of ALU elements in downstream flanking regions, the plot has a similar distribution to the previous plot of ALU elements in upstream flanking regions. 100,189 copy number variants had ALU elements in their downstream flanking regions which represented 66.43 percent of the total copy number variants. Out of 100,189 copy number variants, 36,880 had one ALU elements in their flanking regions followed by 25,275 copy number variants

that had two ALUs in their flanking regions. Very few of the downstream flanking regions had eight or nine ALU elements as well.

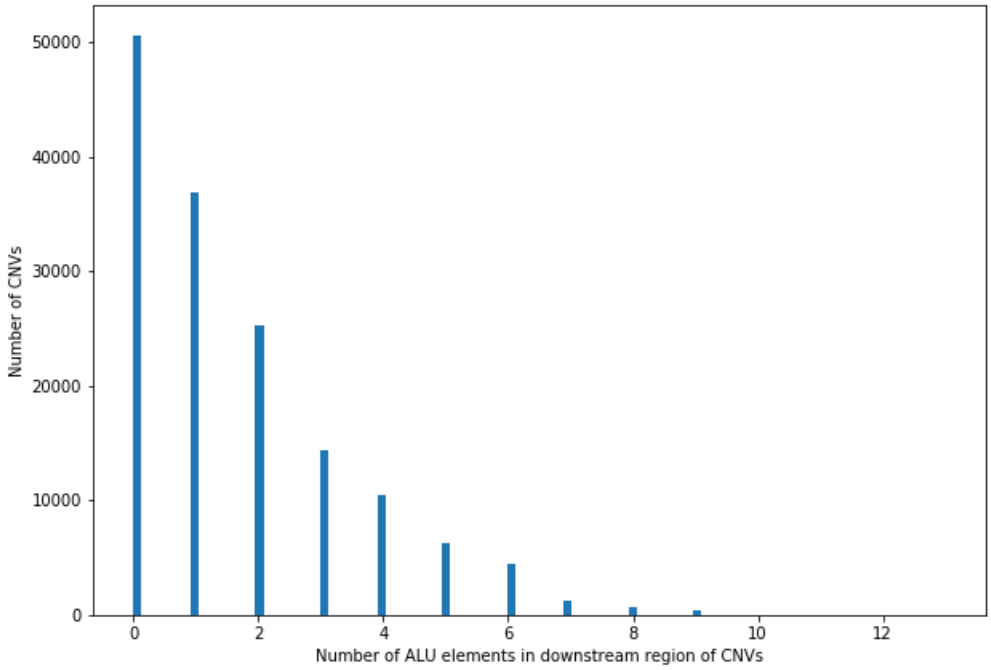

Fig 5: Distribution of Alu elements in downstream flanking regions of Copy Number Variants

The next step was to look at the abundance of different ALU subfamilies in the upstream and the downstream flanking regions of the copy number variants. A standardized counts matrix was generated to enumerate the presence of different ALU subfamilies for both upstream and downstream flanking regions.

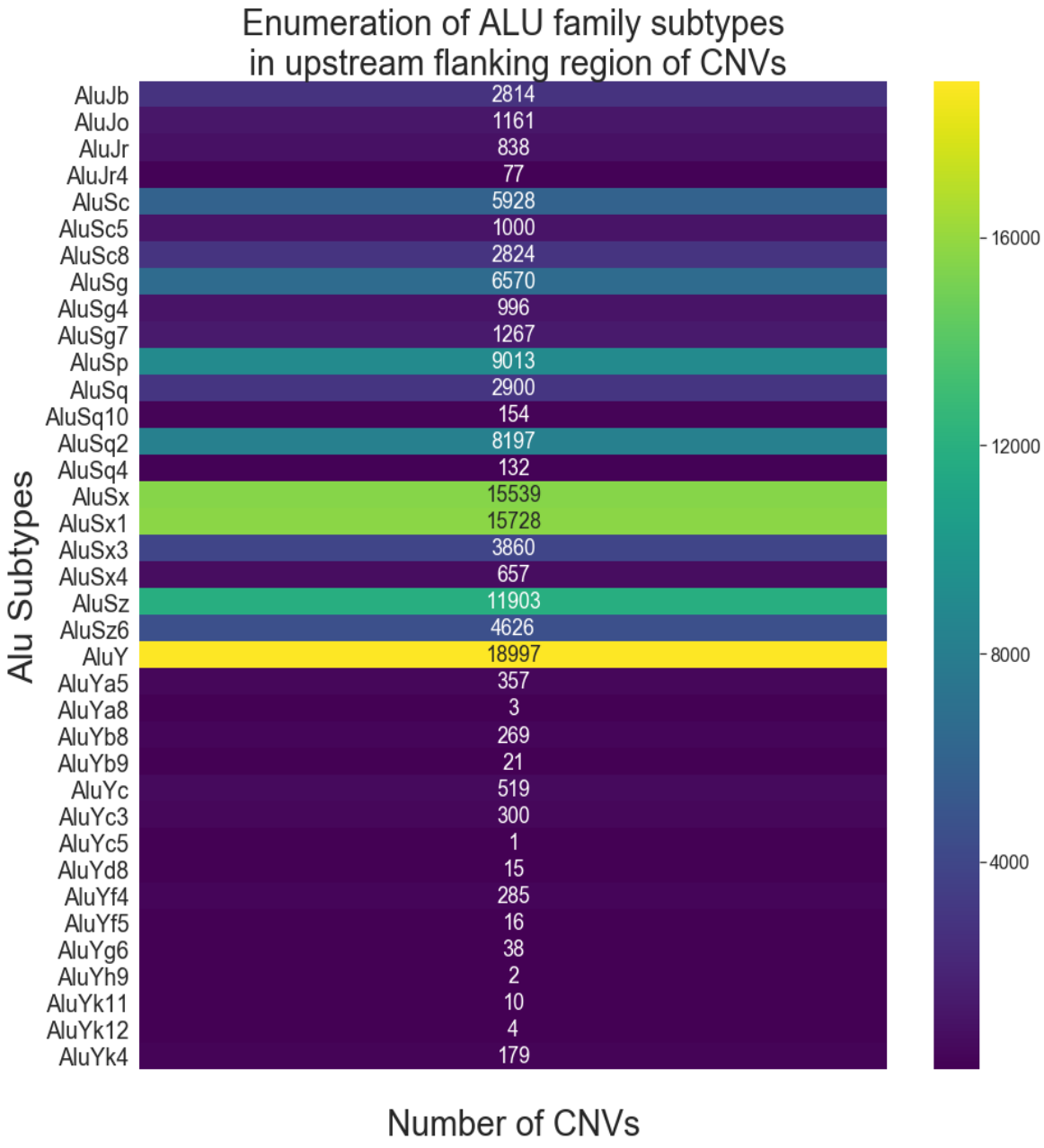

Fig 6: Enumeration of Alu subtypes in upstream flanking regions of Copy Number Variants

There were 37 different ALU family subtypes found in the upstream flanking regions of copy number variants. Looking at the standardized counts matrix and heatmap for counts of ALU family subtypes, it can be easily inferred that AluY subtypes are relatively the most abundant at the upstream flanking regions of the copy number variants. In addition, AluSx and AluSx1 are

relatively more abundant among other ALU subtypes present in the upstream flanking regions. However, the rest are not as abundant as these three ALU subtypes. The least common subtypes are AluYc5, AluYh9, AluYk12 and AluYb9 with abundances of 1, 2, 4 and 21 respectively. These Alus are relatively newer subtypes evolutionarily (Konkel et al., 2015)

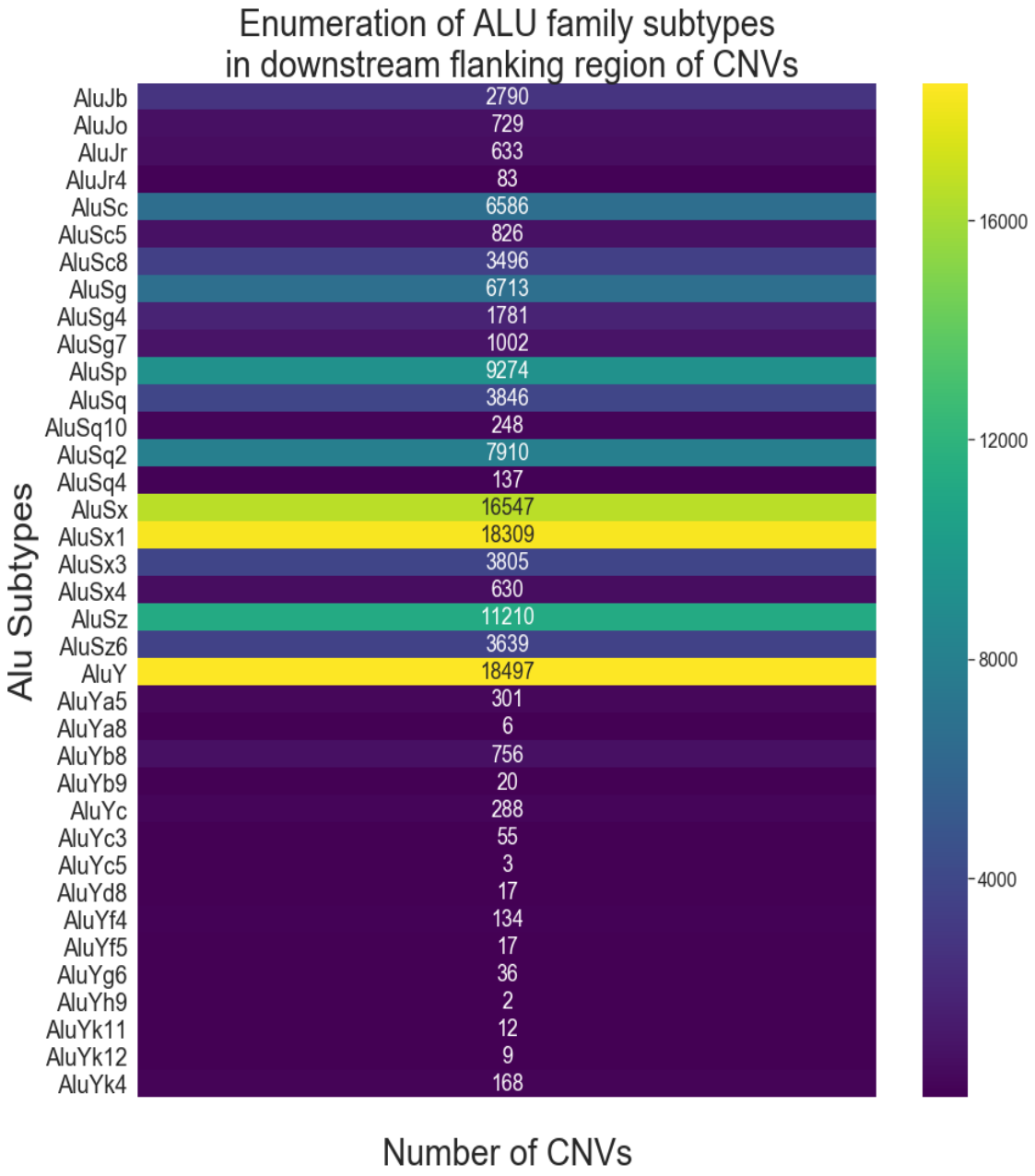

Fig 7: Enumeration of Alu subtypes in upstream flanking regions of Copy Number Variants

The downstream flanking regions had similar counts matrix for Alu family subtypes. There were also 37 different Alu family subtypes in the downstream flanking regions of copy number variants as well. AluY subtype was relatively the most abundant among other subtypes followed by AluSx1, and AluSx. However, in the case of downstream flanking regions, AluSx1 is more abundant in downstream flanking regions than in upstream flanking regions with a difference of 2581. The least abundant Alu subtypes are AluYh9, AluYc5, AluYa8 and AluYb9. Three of the four least abundant subtypes were common in both upstream and downstream flanking regions with the only differences being AluYk12 and AluYa8. These again show that younger AluY subfamilies are relatively much less abundant compared to other subtypes (Konkel et al., 2015).

48545 (32.19 percent) and 50613 (33.56 percent) copy number variants did not have Alu elements in their upstream flanking regions and downstream regions respectively. Out of those copy number variants, 3281 (6.76 percent) and 3268 (6.45 percent) copy number variants had segmental duplications in these upstream regions and downstream regions respectively. L1 elements were relatively more abundant as 11644(23.98 percent) and 13484(26.64 percent) copy number variants had them in upstream and downstream regions respectively.

Screening for L1 LINE elements at the flanking regions of copy number variants:

31,594 L1 elements were found in the upstream flanking regions which accounts for 20.95 percent of the total copy number variants in the dataset. Of these 31,594 L1 elements, 16587 had 1 L1 elements, 8508 had 2 L1 elements, 4123 had 3 L1 elements, 1499 had 4 L1 elements and 595 had 1 L1 element in the downstream flanking regions respectively.

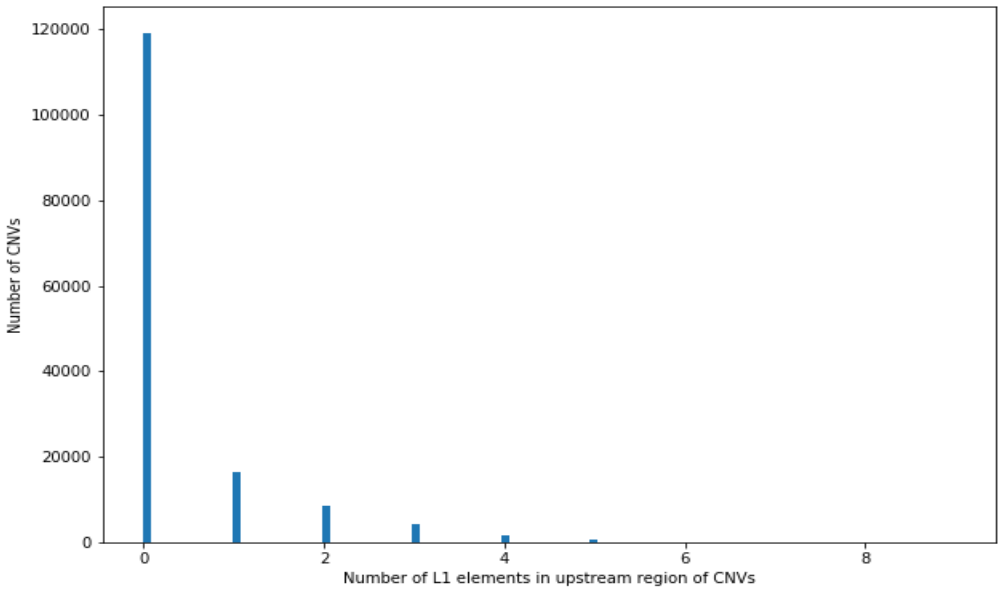

Fig 8: Distribution of L1 LINE elements in upstream flanking regions of Copy Number Variants

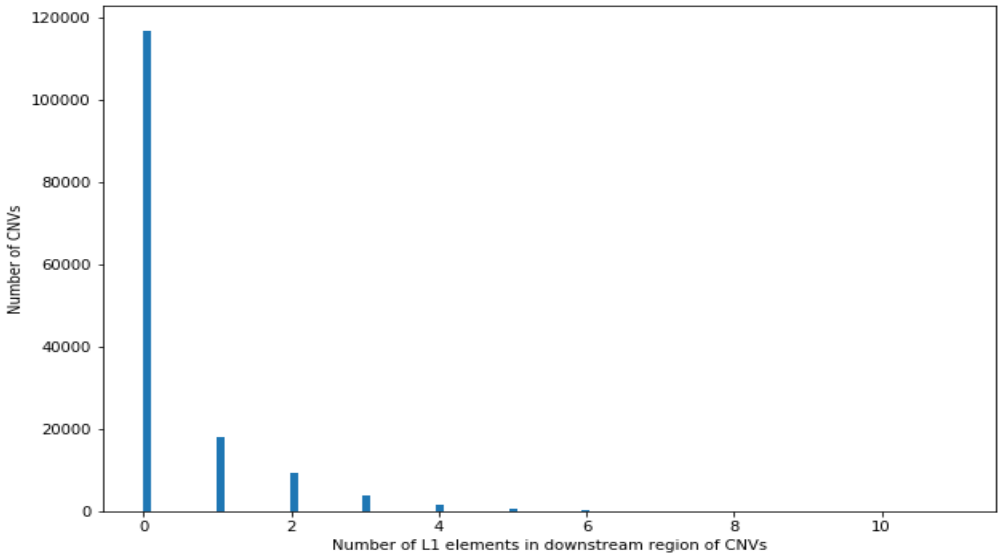

Fig 9: Distribution of L1 LINE elements in downstream flanking regions of Copy Number Variants

The abundance of L1 elements in downstream flanking regions of copy number variants is similar to the abundance of L1 elements in upstream flanking regions. There are 33889 copy number variants (22.47 percent) with L1 elements in their flanking regions. Out of 33889 copy number variants, 17958 copy number variants had 1 L1 element, 9413 copy number variants had 2 L1 elements, 3802 copy number variants had 3 L1 elements, 1718 copy number variants had 4 L1 elements and 486 copy number had 5 L1 elements. There are about 119,208 (79.05 percent) and 116,913 (77.53 percent) copy number variants without any L1 elements in both upstream and downstream flanking regions respectively which indicate L1 elements are relatively not that abundant at the flanking regions of copy number variants.

Out of those copy number variants with no L1 LINE elements in their flanking regions, 82307 (69.04 percent) and 79784 (68.24 percent) copy number variants had various subtypes of Alu elements in these upstream regions and downstream regions respectively. Segmental duplications were also found as 6625 (5.56 percent) and 6755 (5.77 percent) copy number variants had them in upstream and downstream regions of these copy number variants respectively.

To enumerate the abundances of different L1 subtypes in the upstream region of copy number variants, a standardized counts matrix was generated and heatmap was used to visualize the differences in counts between different L1 subtypes. In total, there were 114 different L1 elements present in upstream flanking regions. Only 3 L1 elements had abundance greater than 1000. Those three L1 elements were L1MB7, L1MB3 and L1MB4 with abundances of 1967, 1514 and 1022 respectively. There were quite a few with only a count of one. Those were L1M6, L1ME5, L1MEa, L1MEg1, L1MEg2 and L1P4c. In total, 111 L1 elements had an abundance of less than 1000.

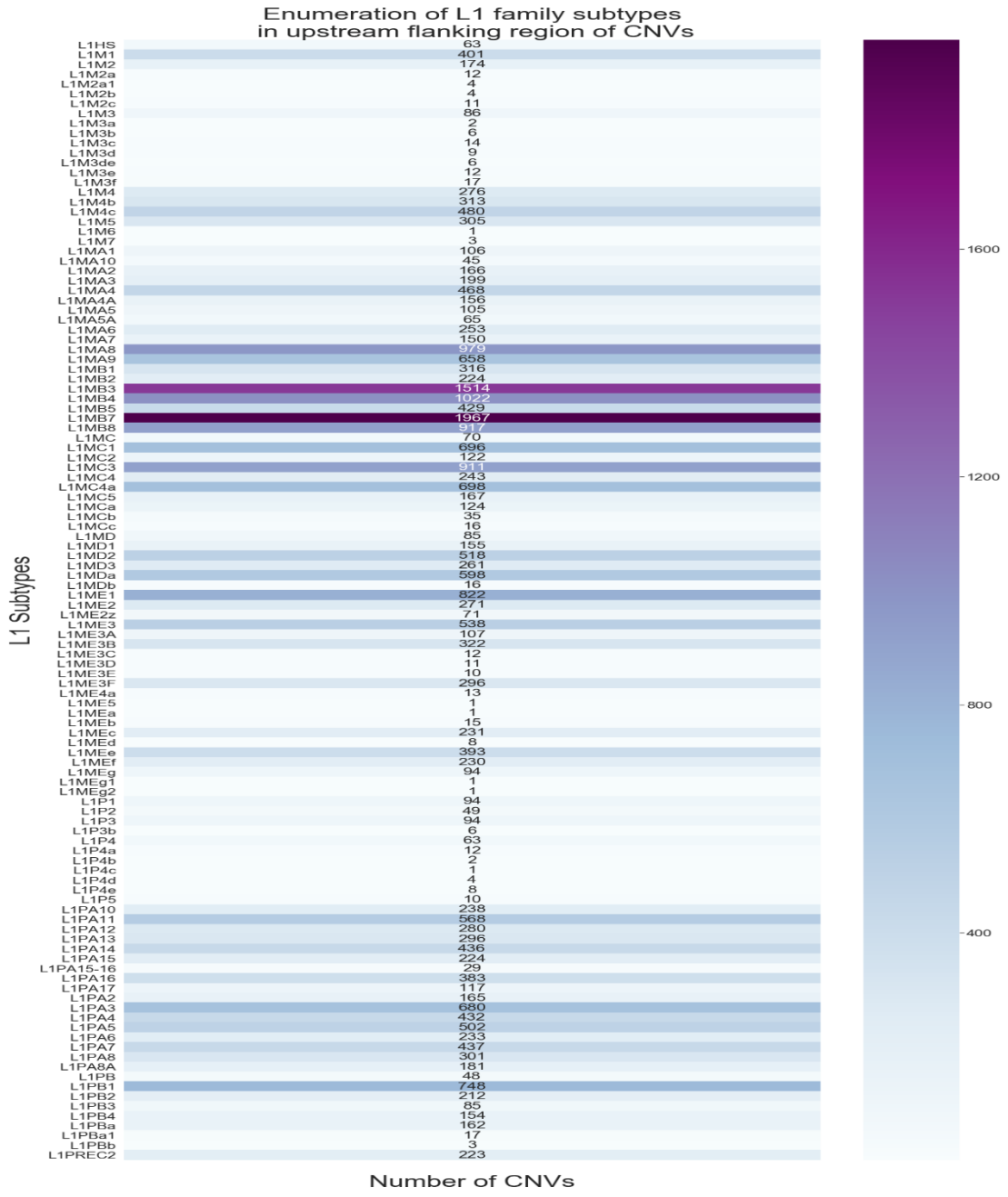

Fig 10: Enumeration of L1 subtypes in upstream flanking regions of Copy Number Variants

The L1 abundance at the downstream flanking regions of the copy number variants was pretty similar to the upstream flanking regions with very subtle differences. There were 114 different L1 elements present at the downstream flanking regions as well. Out of those 114 Alu

elements, 108 Alu elements had abundances of less than 1000 counts and only 6 had abundances of greater than 1000. Those were L1MB3, L1MB7, L1MB8, L1ME1, L1PA4 and L1PB2. And the least abundant ones were L1M6, L1P, L1P3b, L1P4d. The least abundant L1 elements were not like the ones in the upstream flanking regions except L1M6.

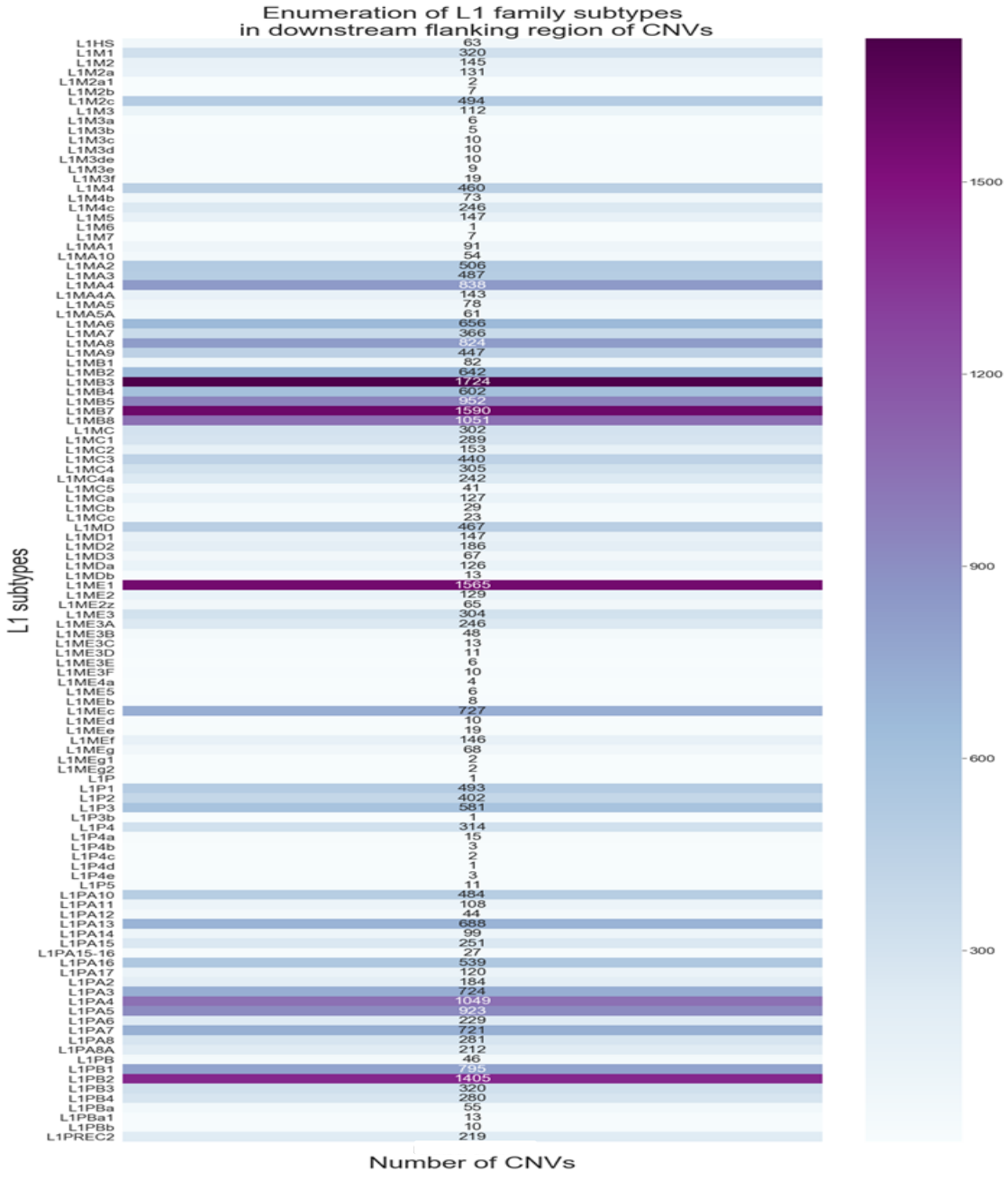

Fig 11: Enumeration of L1 subtypes in downstream flanking regions of Copy Number Variants

Screening for the abundance of segmental duplications at the flanking regions of copy number variants:

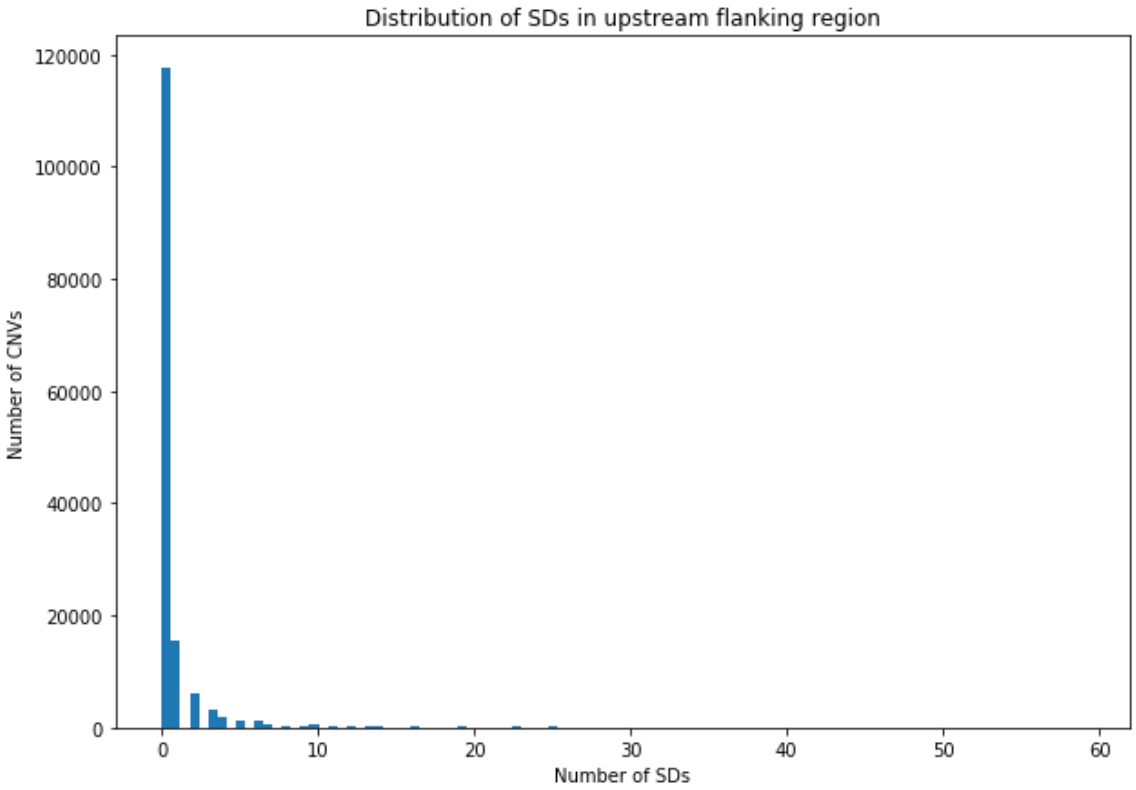

Fig 12: Distribution of Segmental Duplications in upstream flanking regions of Copy Number Variants

117,544 copy number variants did not have segmental duplications in their upstream flanking regions which accounted for 77.80 percent of total copy number variants in the dataset. The remaining copy number variants had a varying number of segmental duplications present in the upstream flanking regions. Since most of the copy number variants do not have segmental duplications, this histogram is highly skewed and does not represent the variability in the counts of segmental duplications in the remaining copy number variants.

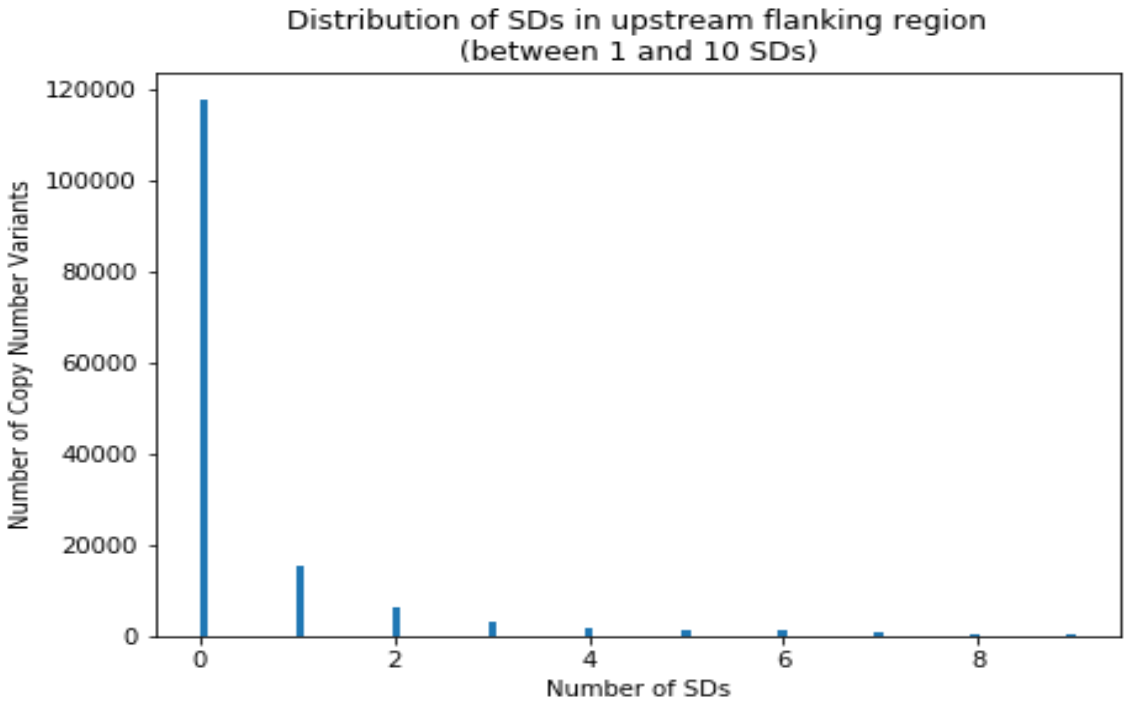

Fig 13: Distribution of SDs in upstream flanking regions (between 1 and 10 SDs)

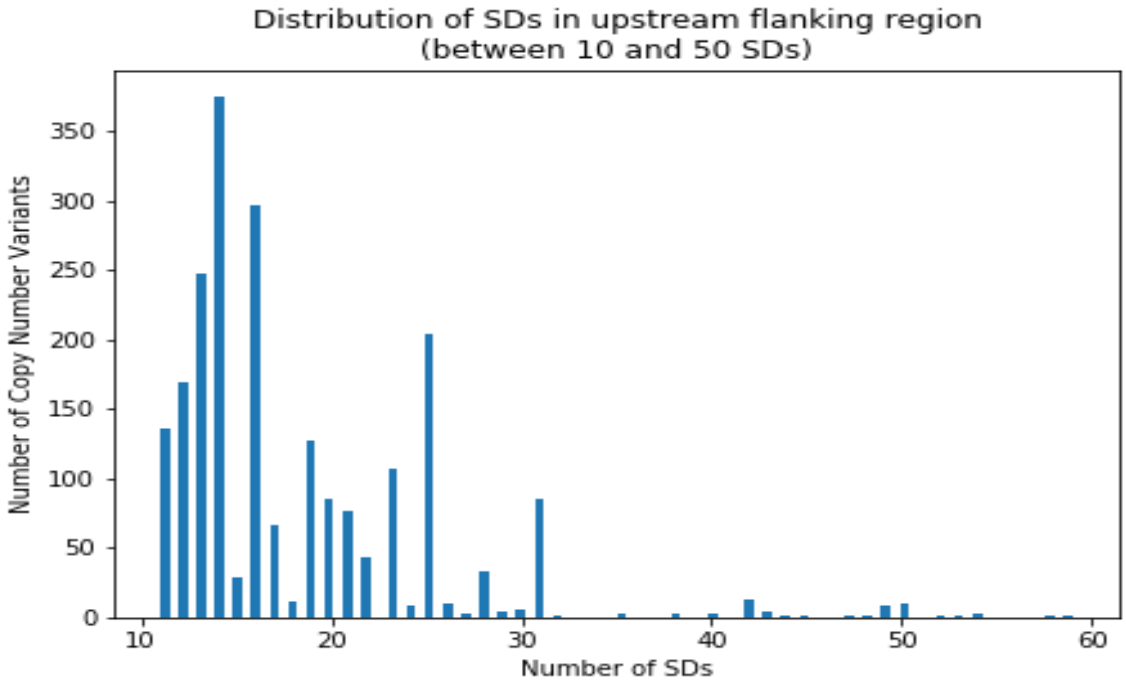

Fig 14: Distribution of SDs in upstream flanking regions (between 10 and 60 SDs)

The figures above show two different plots that demonstrate the variability in the number of segmental duplications at upstream flanking regions of copy number variants. It is evident that copy number variants have a variable number of segmental duplications in their upstream flanking regions ranging from as low as 1 to as high as 59. There were 30,457 (91.58 percent) copy number variants that had between 1 and 10 segmental duplications. And there were 2801(9.19 percent) and 16 (0.052 percent) copy number variants that had greater than 10 segmental duplications and greater than 50 segmental duplications respectively.

The distribution of segmental duplications in downstream flanking regions was very similar to the distribution of segmental duplications in upstream flanking regions. There were 34,638 (22.97 percent) copy number variants that had segmental duplications in their downstream flanking regions. Out of 34,638 copy number variants, 31,582 (91.17 percent) copy number variants had segmental duplications between 0 and 10. And 3056 (8.82 percent) copy number variants had more than 10 segmental duplications.

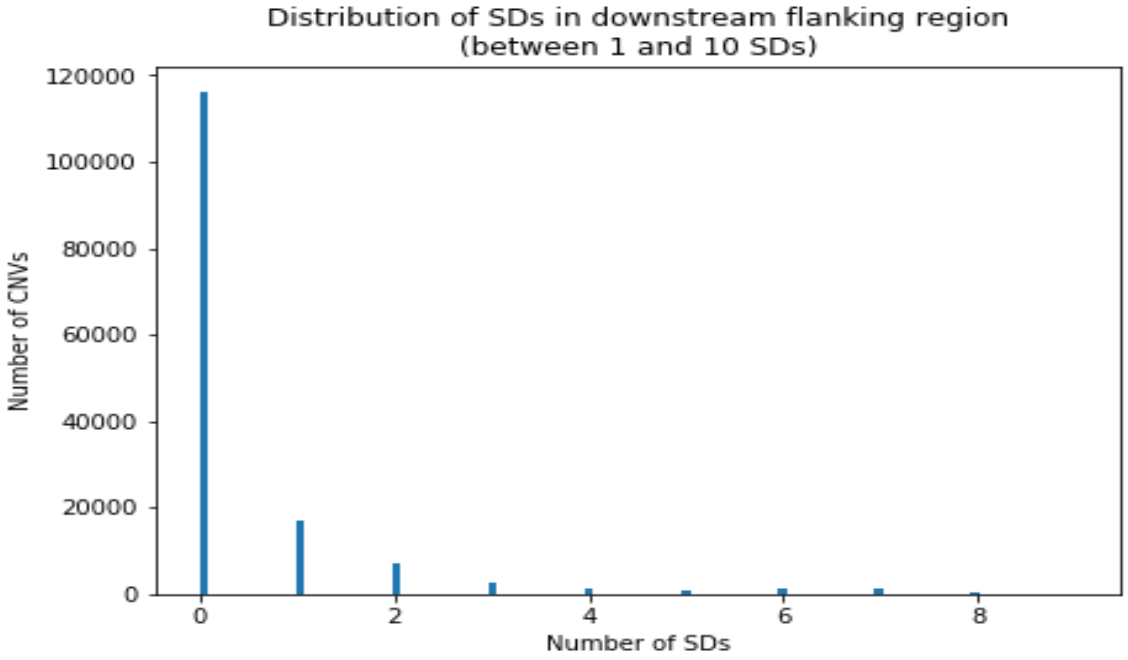

Fig 15: Distribution of SDs in downstream flanking regions (between 1 and 10 SDs)

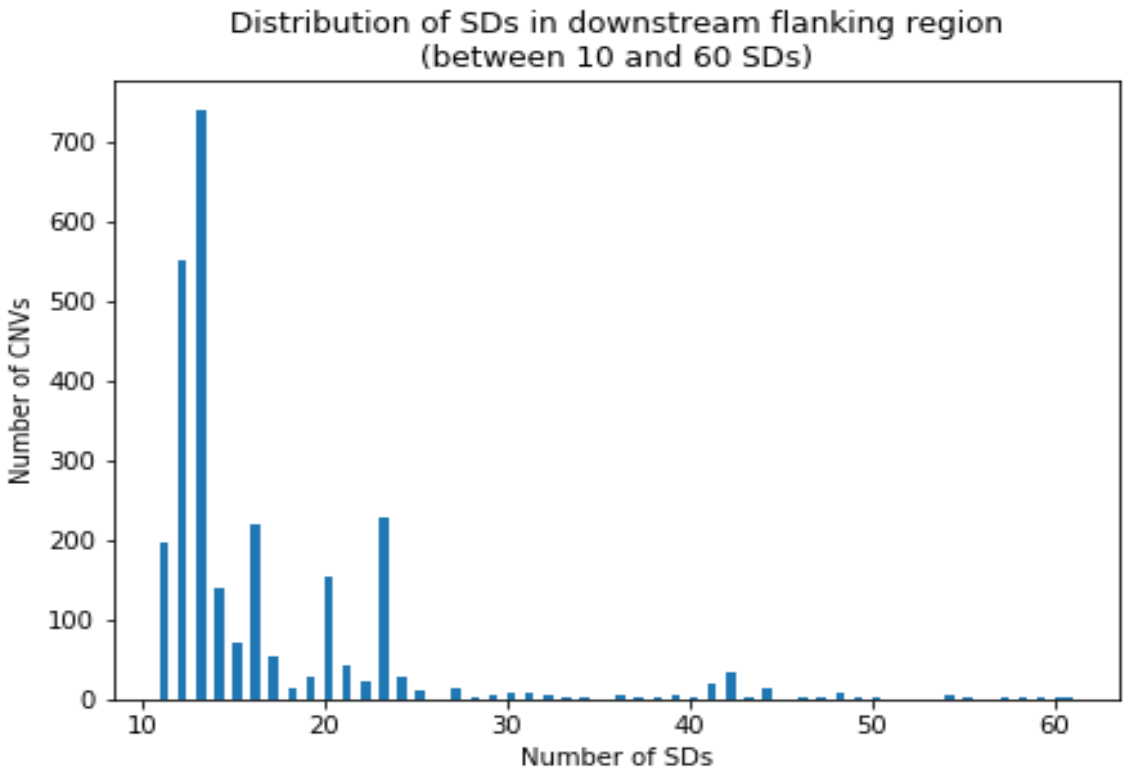

Fig 16: Distribution of SDs in downstream flanking regions (between 10 and 60 SDs)

As mentioned before, there were 117544 copy number variant coordinates with no segmental duplications in the upstream flanking regions. It was necessary to check if these 117544 copy number variants had Alu elements in their upstream flanking regions. Out of those 117544 copy number variants, 80921 (68.84 percent) were found to have different subtypes of Alu elements in their flanking regions. Similarly, there were 116164 copy number variants with no segmental duplications in their downstream flanking regions. Out of those 116164 copy number variants, 78968 (67.97 percent) were also found to have different subtypes of Alu elements in their downstream flanking regions. Different L1 elements were also found to be in the flanking regions of copy number variants with no segmental duplications. Out of 117544 upstream flanking regions, 25244 (21.47 percent) had L1 elements and out of 116164 downstream flanking regions, 25877 (22.27 percent) had L1 elements in them.

Screening for exons of Dosage Sensitive genes within copy number variants:

775 out of 1016 curated dosage sensitive genes were present in these copy number variants. However, out of 150,802 copy number variants, only 5444 (3.61 percent) copy number variants had dosage sensitive genes with haploinsufficiency score of 3 and more. Therefore, only 5444 copy number variants might have the potential to be pathogenic due to the presence of dosage sensitive genes.

Screening for transcription factor specific regulatory elements within copy number variants:

80,152 copy number variants did not have any regulatory elements which account for 53.15 percent of total copy number variants. Out of the remaining 70,650 copy number variants, 68,414 (96.83 percent) variants had 100 or less transcription factor specific regulatory elements and 2236 (3.17 percent) copy number variants had 100 or more transcription factor specific regulatory elements.

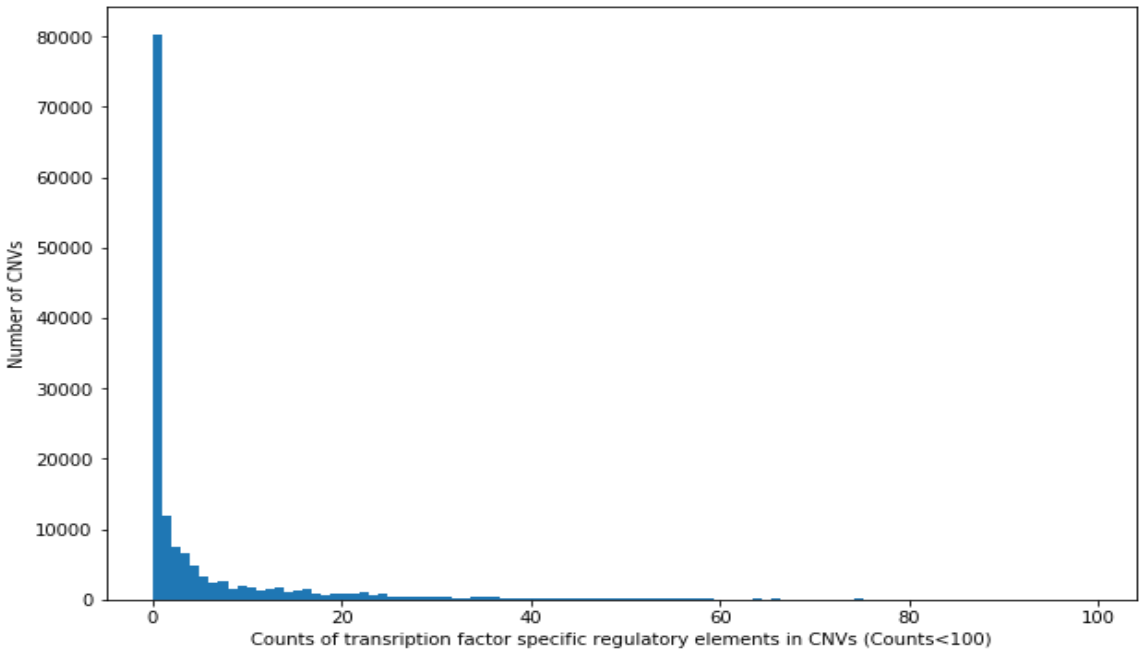

Fig 17: Counts of transcription factor specific regulatory elements across copy number variants (Counts < 100)

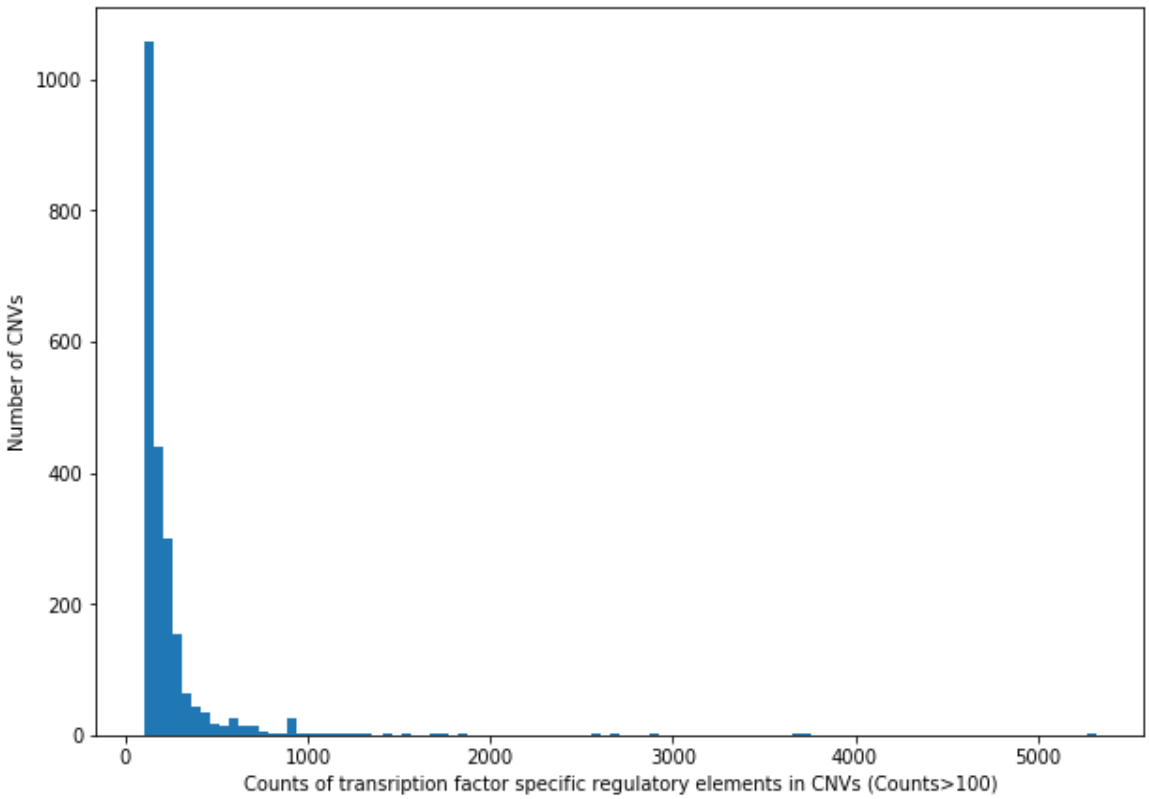

Fig 18: Counts of transcription factor specific regulatory elements across copy number variants (counts > 100)

Out of 160 Transcription factors in the TXN factor ChIP dataset, 152 were found to have transcription factor specific regulatory elements in these copy number variant segments. However, the majority of their abundances were not as high except for a few. Transcription factors that had high abundances were *CTCF, CEBPD, EP300, FOXA1, MAFK, MAX, PHF8, POLR2A RAD21, RUNX3, SMC3, SPI1,* and *TEAD4*. They all were present in more than 10, 000 copy number variants. CTCF was the highest among them with binding sites in 56415 (37.40 percent) copy number variants. Next, *POLR2A* and *RAD21* were the only transcription factor that had binding sites in more than 20,000 copy number variants.

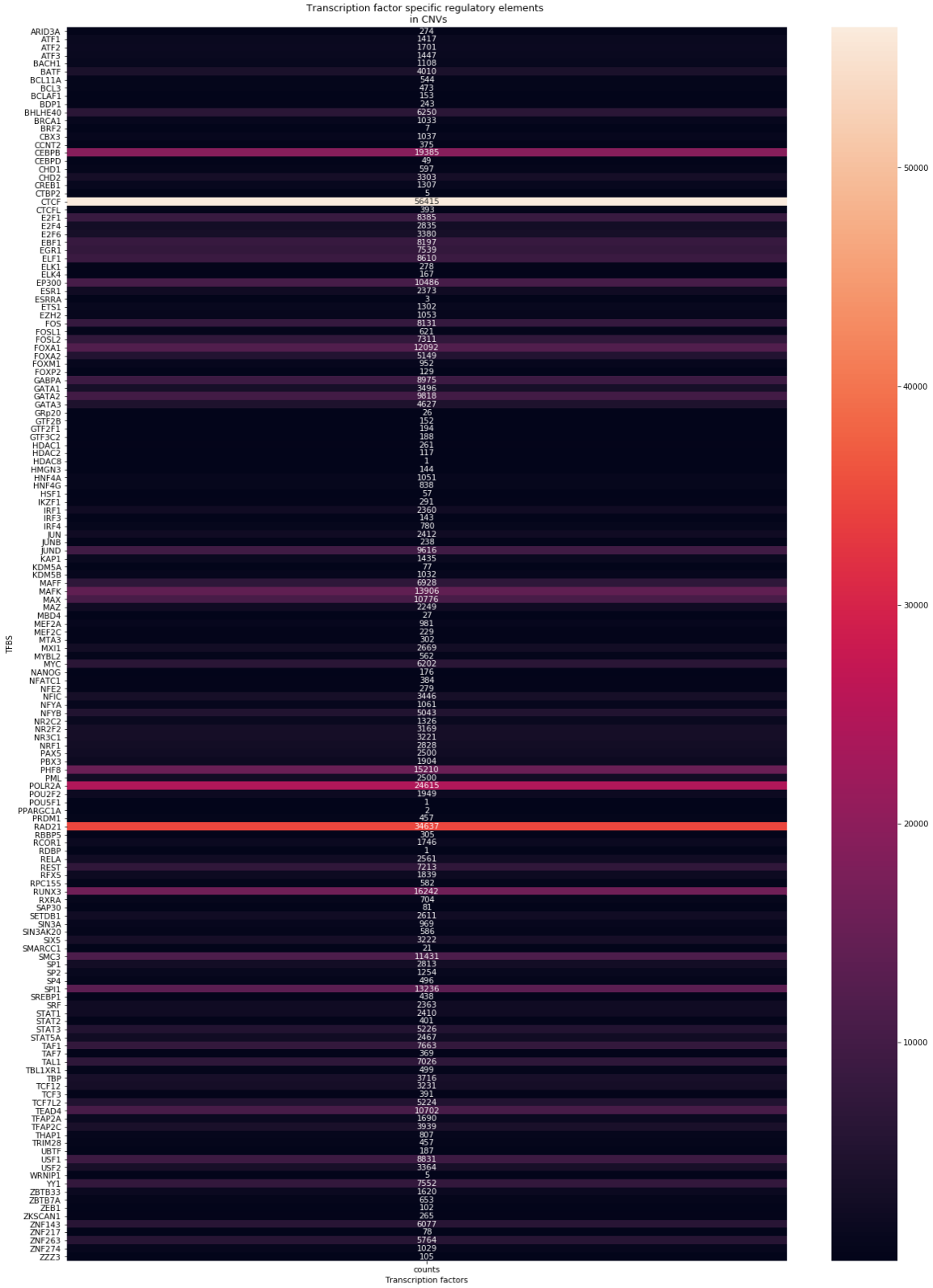

Fig 19: Enumeration of transcription factor specific regulatory elements in copy number variants

Copy Number Variant hits stratified to Copy Number Variable Regions (CNVRs):

There are 25340 copy number variable regions that had one or more copy number variant hits. However, majority of them had copy number hits below 15. The mean and median for copy number hits in copy number variable regions (CNVRs) were 5.11 and 1.0 respectively. The mean and median gave an initial idea of how skewed the distribution is. Looking at fig 20 representing histogram of copy number variant hits against the number of copy number variable regions, it can be concluded that more than 20,000 copy number variable regions had hits below 15 or less. However, from this initial histogram, it could not be concluded whether there were CNV hits of 100 or more in copy number variable regions or not. Therefore, looking at fig 21 representing histogram of copy number variant hits greater than 150 would give better representation for the distribution of maximal hits. Looking at this histogram makes it clear that there are some copy number variable regions (CNVRs) with hits greater than 150.

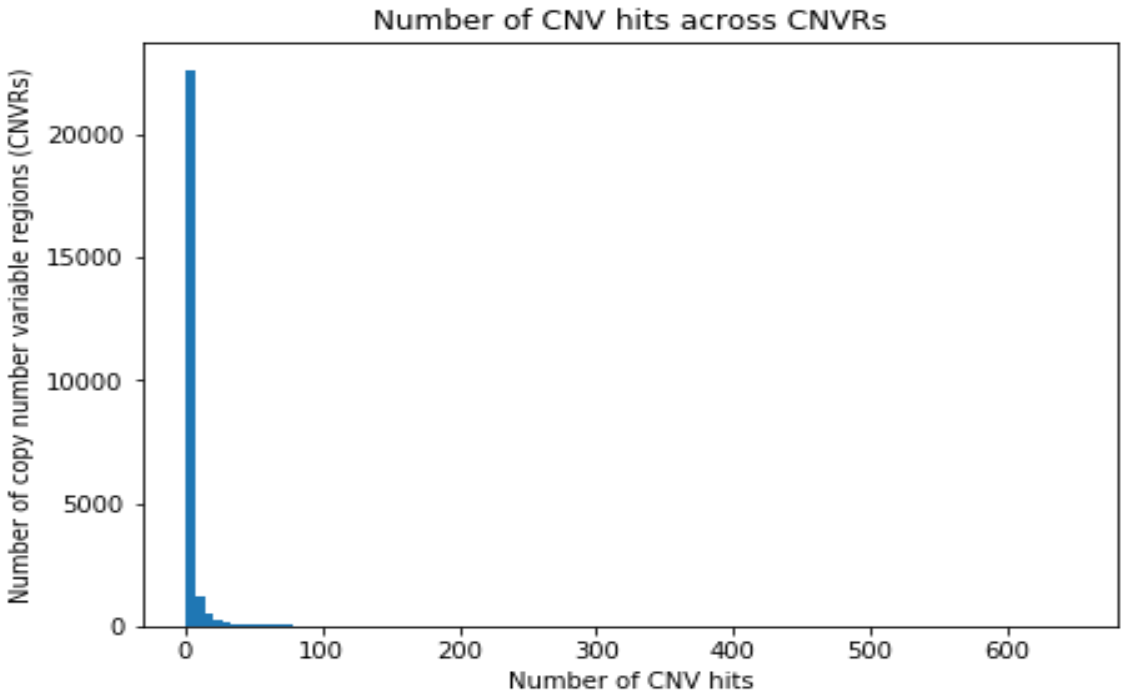

Fig 20: Number of Copy Number Variant Hits across Copy Number Variable Regions

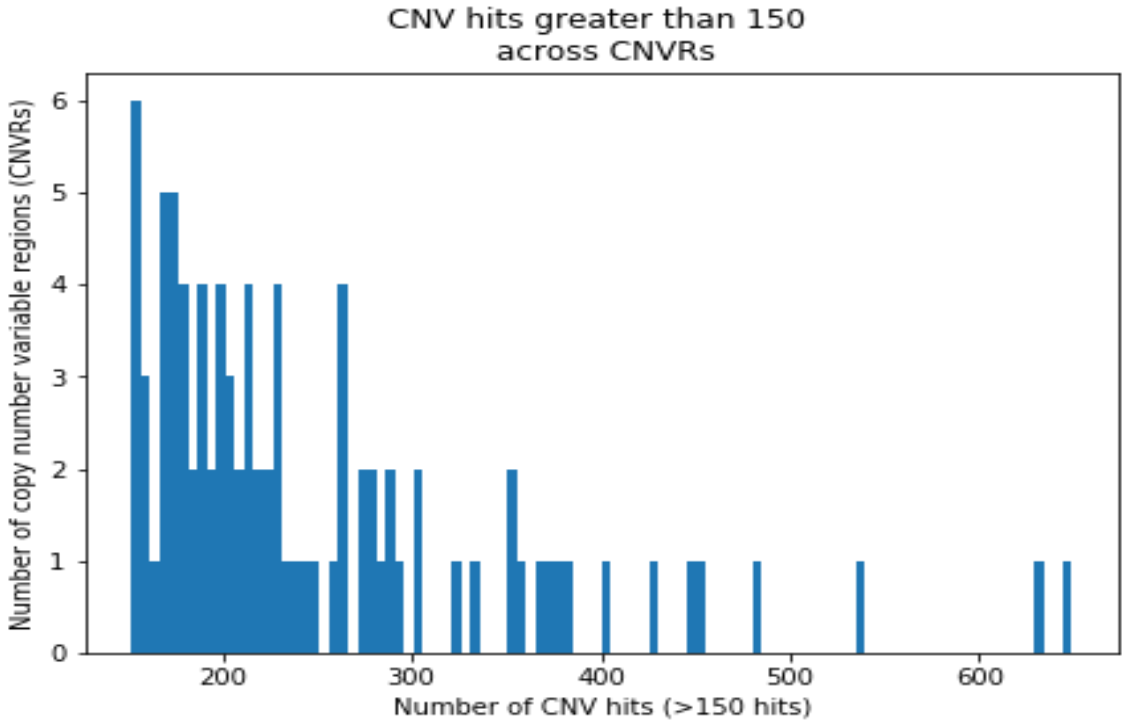

Fig 21: Number of Copy number variant hits greater than 150 across Copy Number Variable Regions (CNVRs)

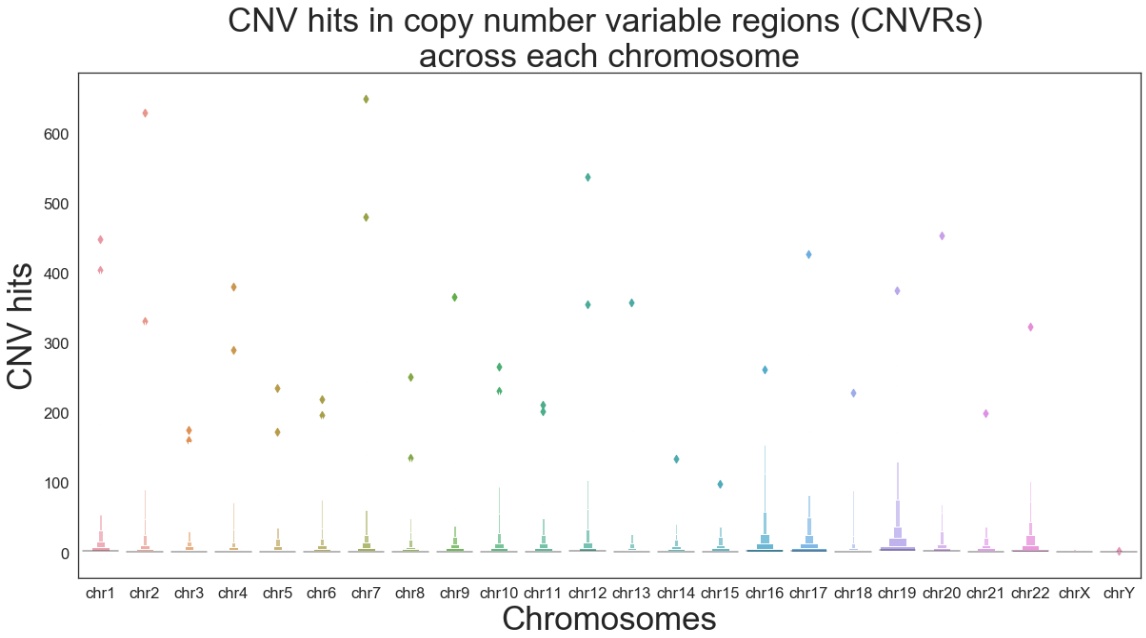

Fig 22: Spread of Copy Number Variant hits in Copy Number Variable Regions across each Chromosome

The boxenplot in fig 22 above shows the spread of copy number variant hits in copy number variable regions (CNVRs) across each chromosome. From the boxenplot, it can be inferred that there are some copy number variable regions (CNVRs) in each chromosome that are getting much higher number of copy number variant hits than other CNVRs in each chromosome. Looking at the genomic repeat features of these copy number variable regions with maximal hits across each chromosome might tell us what makes them so susceptible to copy number variants.

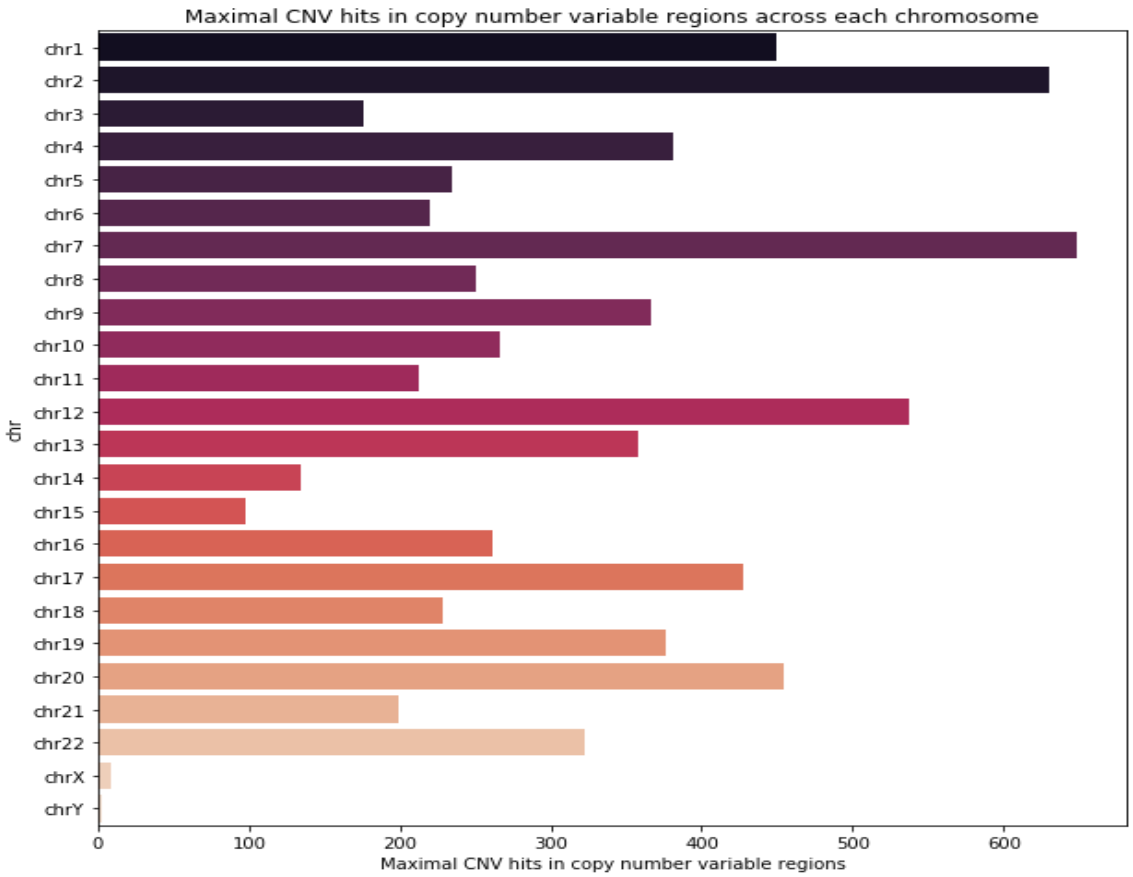

Fig 23: Maximal Copy Number Variant hits in copy number variable regions across each chromosome

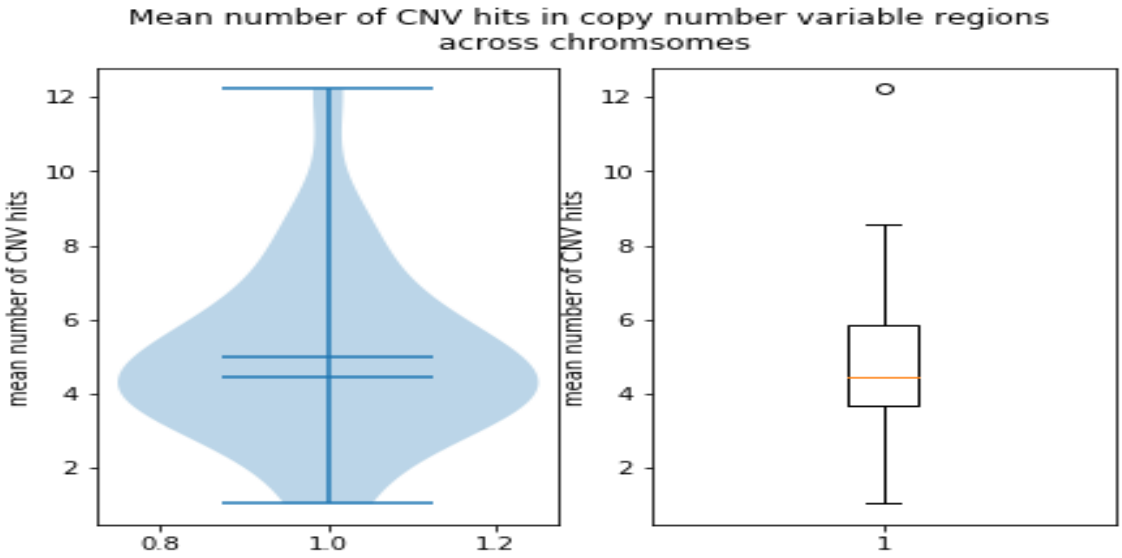

Fig 24: Distribution of Mean Number of Copy Number Variant hits in copy number variable regions across chromosomes

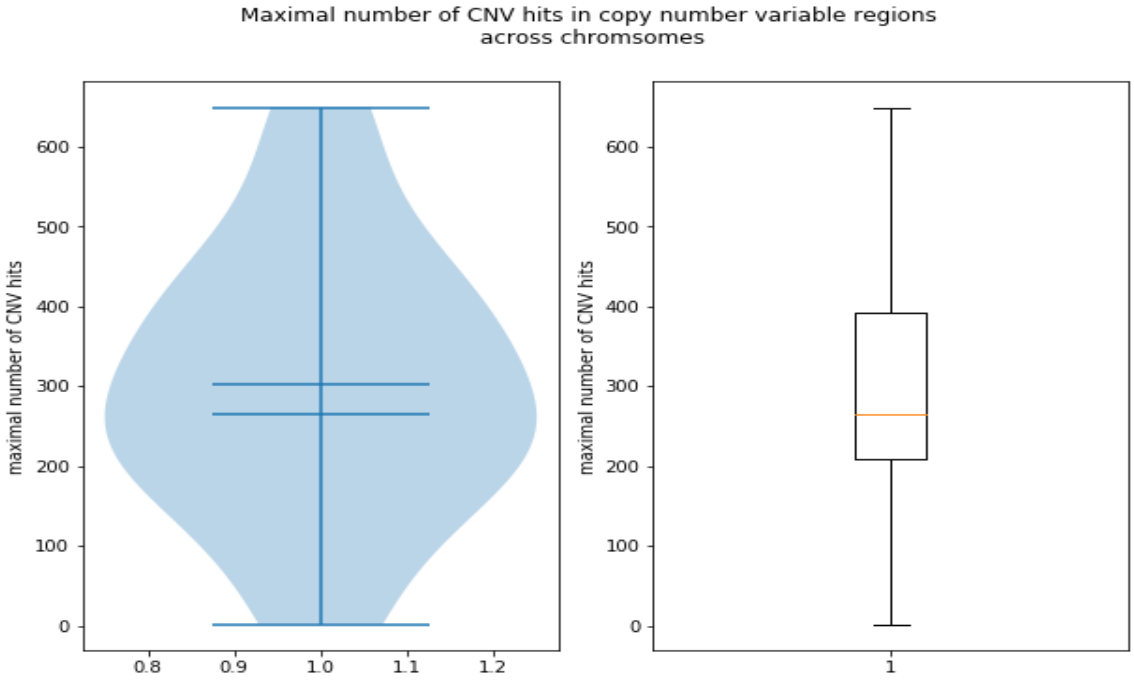

Fig 25: Distribution of Maximal Number of Copy Number Variant hits in copy number variable regions across chromosomes

All the copy number variable regions with maximal copy number variant hit across each chromosome had at least 100 hits. The top 3 copy number variable regions with maximal hits were from chromosome 7, 2 and 12 with maximal hits of 649, 630 and 538 respectively. Chromosome X and Chromosome Y were the exceptions because the dataset did not have enough representation of chromosome X and Y copy number variants due to filtering steps. Majority of the copy number variable regions across each chromosome had maximal hits between 200 and 400 as suggested by the boxplot and the violin plot. In addition, most of the copy number variable regions across each chromosome had mean copy number variant hits between 3.75 and 6 hits.

Screening for Alu elements in copy number variable regions with maximal hits across each chromosome:

Since these copy number variable regions get the most copy number variants, it is important to look at the repeat architecture in these regions. Microhomology elements such as different Alu subtypes are one of the most essential genomic features that contribute to the formation of copy number variants. There are on average 17.48 Alu elements and 22 different Alu family subtypes present within these copy number variable regions (CNVRs). Almost all the maximal hits region for each chromosome had enrichment for AluYs with only the exception of chromosome 3, 4, 7, 15, 17, and 18. AluSx, AluSx1, AluSg, and AluSq were relatively abundant in these copy number variable regions as well.

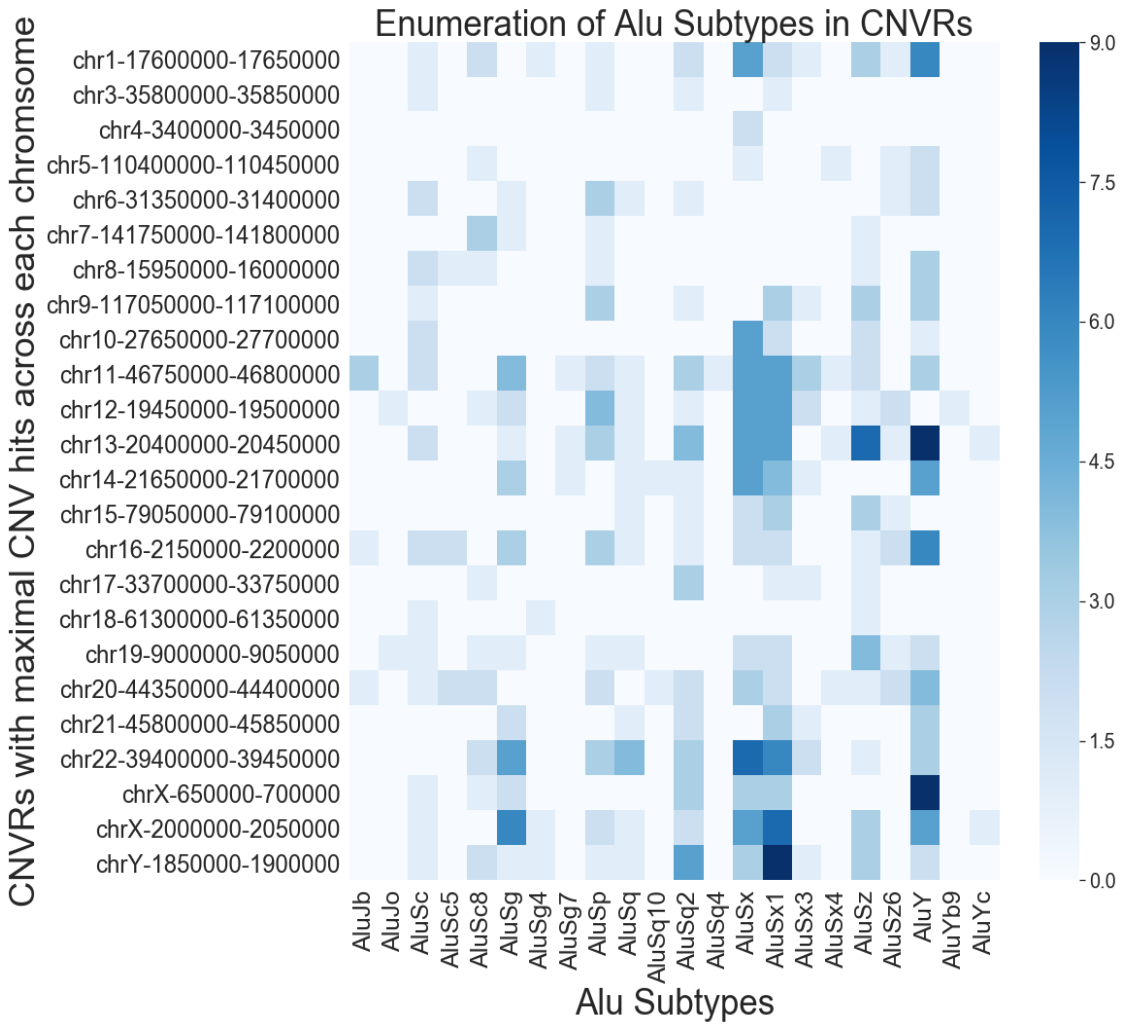

Fig 26: Enumeration of Alu Subtypes in Copy Number Variable Regions with Maximal Copy Number Variants hits across each chromosome

Screening for L1 LINE elements in copy number variable regions with maximal hits across each chromosome:

L1 elements are another group of repeat elements that are implicated in the formation of copy number variants. On average, there are 5.6 L1 elements and 43 different L1 element subtypes within these copy number variable regions (CNVRs). It is clear from this heatmap below that L1 elements have a heterogenous pattern and are much less abundant than Alu elements.

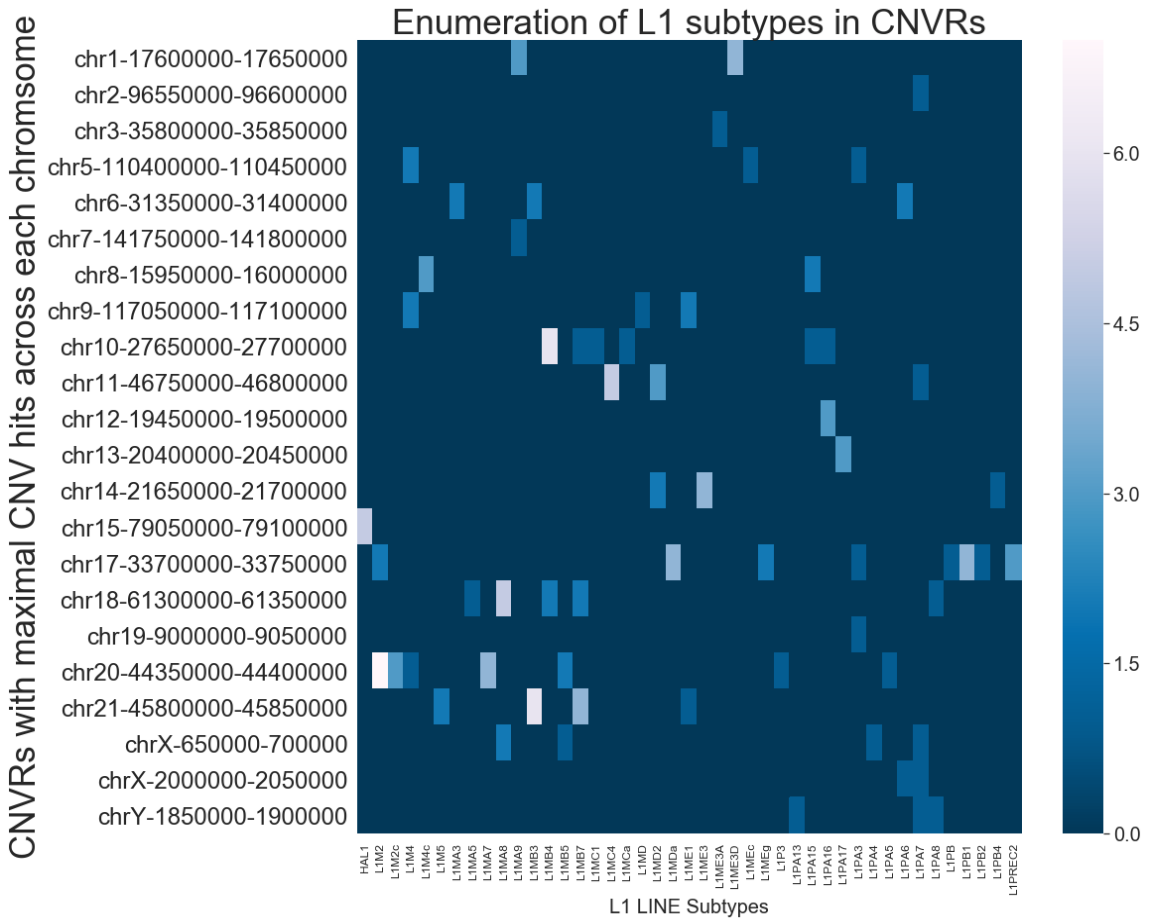

Fig 27: Enumeration of L1 LINE Subtypes in Copy Number Variable Regions with Maximal Copy Number Variants hits across each chromosome

Screening for segmental duplications in copy number variable regions with maximal hits:

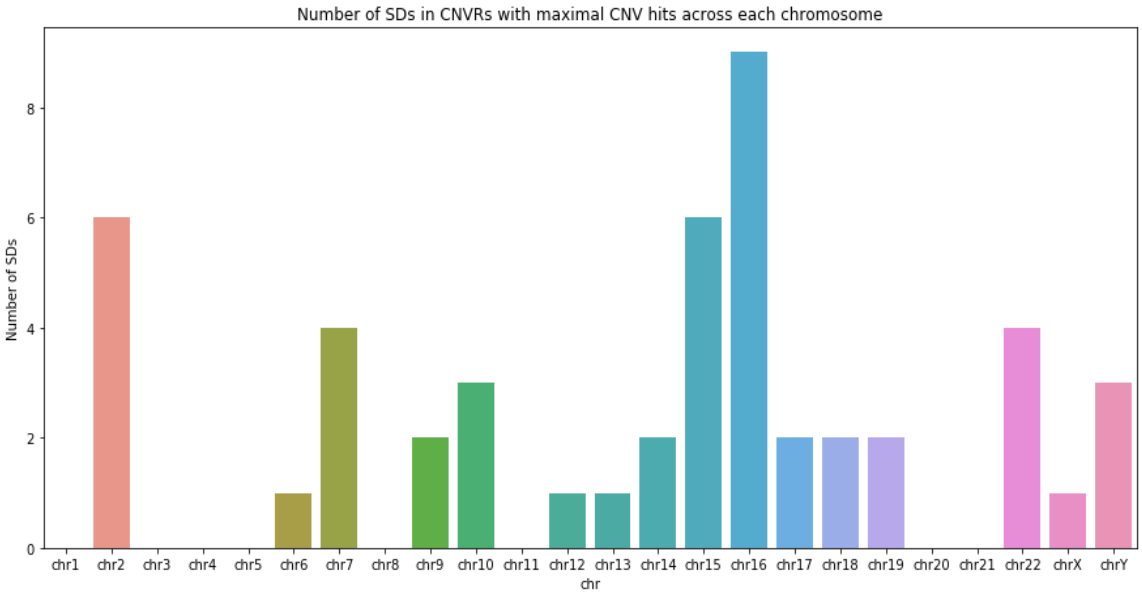

Fig 28: Number of Segmental duplications in copy number variable regions with maximal Copy Number Variant hits across each chromosome

There were seven maximal hits copy number variable regions (CNVRs) that did not have any segmental duplications at all. However, these copy number variable regions are getting hits of greater than 100 copy number variants. So, even though low copy repeats like segmental duplications are not present, overrepresentations of high copy repeats like Alu elements might facilitate the formation of copy number variants within these regions. An exception is the CNVR of chromosome 2. Although this specific CNVR is getting copy number variant hits of 620, no Alu subtypes were present within the segment. On the contrary, interestingly, 6 segmental duplications were present which might indicate the increased likelihood of segmental duplications being responsible for the heightened propensity of this specific region to copy number variants.

Screening for exons of dosage sensitive genes in copy number variable regions with maximal hits:

Dosage sensitive genes with haploinsufficiency score of 3 or more were screened for their presence in these copy number variable regions with maximal copy number variant hits. Looking at the data, maximal hits copy number variable regions did not have any dosage sensitive genes except chromosome 16 which had *PKD1* gene with a haploinsufficiency score of 3. Thus, except chromosome 16, neither of the maximal hits copy number variable regions had high potential to be pathogenic.

Screening for transcription factor specific regulatory elements in copy number variable regions with maximal hits:

Copy number variants can disrupt binding of transcription factors to regulatory elements leading to alteration of the cascades of interactions that may consequently lead to clinical phenotypes. So, it is necessary to look at the transcription factor specific regulatory elements. 43 different transcription factors were found in these copy number variable regions. *CTCF*, *RAD21*, *POLR2A, STAT3, ZNF263, and MAFK* transcription factor binding sites were found in many of these copy number variable regions with maximal hits. Alterations of these transcription factor binding sites can perturb crucial interactions that may be detrimental in the context of a particular cell type.

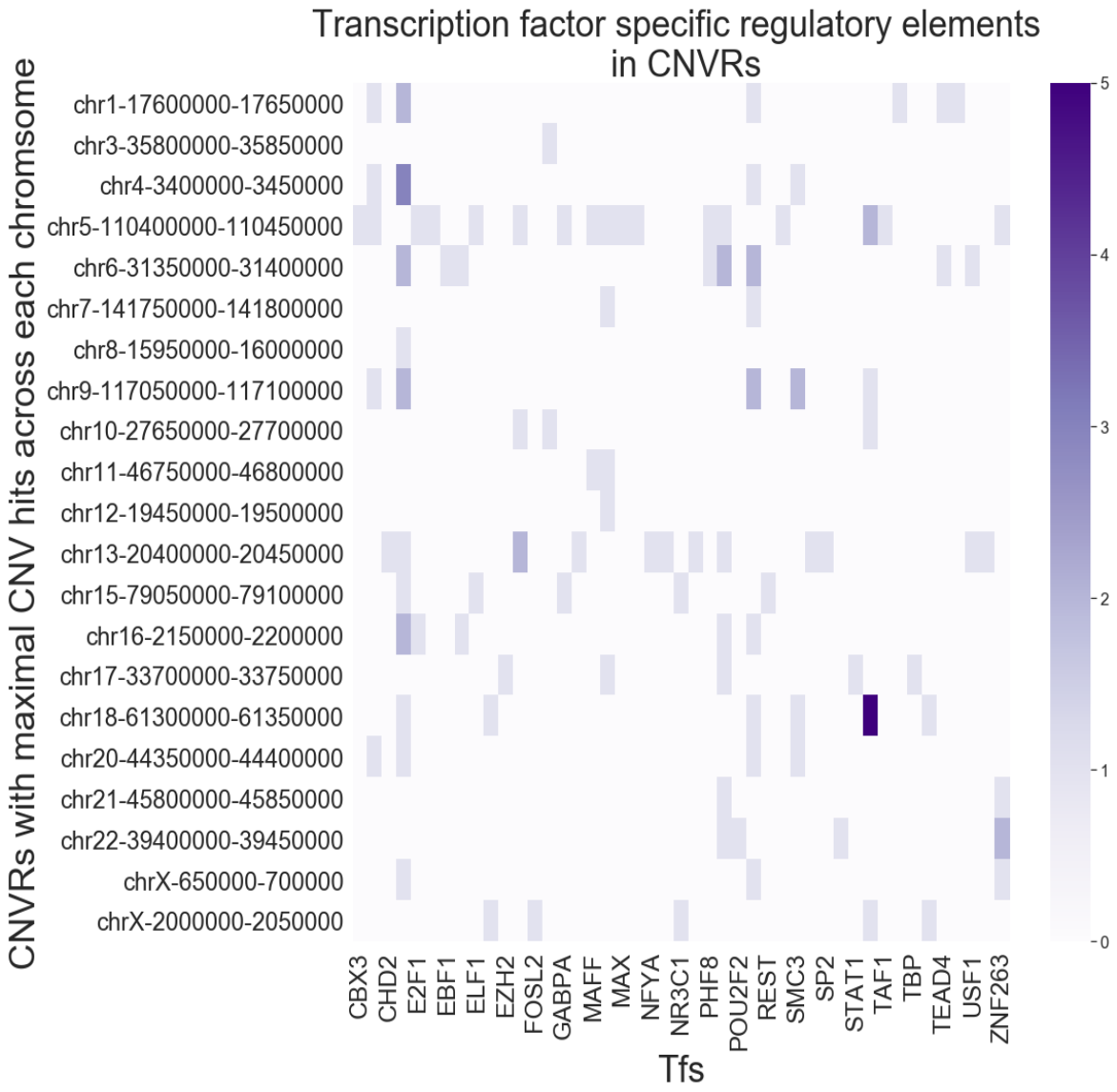

Fig 29: Transcription factor specific regulatory elements in Copy Number Variable Regions with maximal hits across each chromosome

DISCUSSION

Copy number variants are classed as aberrations that modify the number of copies of segments in the genome by deletions or duplications. They are a source of genomic variation that contribute to diversity in the human genome (Girirajan et al., 2011). It has been known for some time that the architecture of the genome dictates the formation of these copy number variants (Hastings et al., 2009). The presence of repeated sequences such as high identity segmental duplications, Short Interspersed Nucleotide Sequence, Long Interspersed Nucleotide Sequences make the genome more prone to undergo rearrangements (Hastings et al., 2009). These repeated architectures in the genome are the key elements driving non allelic homologous recombination, fork stalling and template switching (FoSTeS), and microhomology-mediated break-induced replication (MMBIR) which serve as the main mechanisms implicated in causing deletions and duplications in the genome resulting in copy number changes (Hastings et al., 2009). There have been few studies done to understand the mechanistic context of how these deletions and duplications give rise to structural variations in the genome (Arlt et al., 2012), (Hastings et al., 2009). However, very few studies have been conducted to explore the flanking regions of copy number variants which might give insights on the formation signature of copy number variants. In this study, data generated by next generation sequencing and high-resolution microarray studies have been used to explore the flanking regions of 150,802 copy number variants. The dataset that represents the majority (84.20 percent) of the total data has been taken from ExAC. ExAC is a consortium of researchers that aggregated and standardized exome sequencing data from large sequencing projects and called single nucleotide variants and copy number variants on these datasets (Ruderfer et al., 2016). The final dataset represented a reference atlas of copy number variants from whole exome sequencing of 59,898 individuals across diverse global ethnicities

(Ruderfer et al., 2016). Multiple filtering steps have been applied by ExAC to increase confidence on the variant calls and to resolve breakpoints (Ruderfer et al., 2016). In addition to this dataset, data from the database of genomic variations have been used. Database of genomic variation is a good repository for small to large scale genomic variations in the normal population (MacDonald et al., 2013). The dataset that has been used in this study from this source was curated extensively recently and only involved copy number variant calls that were from high coverage next generation sequencing study and high-resolution microarrays (Zarrei et al., 2015). Both the dataset combined represent a comprehensive and an exhaustive catalog of copy number variants depicting a wide spectrum of human genetic variation found in the general population. However, there are some limitations associated with these datasets. Firstly, calling copy number variants from exome sequencing study using hidden Markov method like XHMM is not always accurate and since these copy number variants are relatively rare in the genome, there will always be some level of inaccuracy with the calls (Belkadi et al., 2015). In addition, even though these datasets have undergone multiple processing steps for harmonization, variability in the sequencing library protocols and exome capture methods may still contribute noise to the datasets (Belkadi et al., 2015). Lastly, copy number variants from X and Y chromosome are under-represented in these datasets due to not passing the required threshold in filtering steps (Ruderfer et al., 2016).

Three different repeat architectures (Alu elements, L1 LINE elements, and segmental duplications) were characterized at both upstream and downstream flanking regions of copy number variant calls. Knowing the abundance of these repeats in the flanking regions may help to predict which repeat elements are contributing the most to the formation of copy number segments. From my analysis, it was seen 67.81 percent of copy number variants had different subtypes of Alu elements present in their upstream flanking region. Similarly, 66.43 percent of copy number variants also had different subtypes of Alu elements present in their downstream flanking regions.

In addition, in many cases, the flanking regions had more than one Alu elements present and up to 9 Alu elements in some cases. This reflects the enrichment of Alu elements in flanking regions of copy number variants which consequently increases the likelihood of aberrant recombination between these elements. Thus, based on the abundance inferences, AluY, AluSx and AluSx1 may be considered important mediators in the formation of copy number variants. Interestingly, almost all the young AluY subfamilies had very low abundances at the flanking regions. The relatively lower abundances of younger AluY subfamilies might be because of not accumulating enough copy number in the genome as a result of their recent emergence (Konkel et al., 2015). A limitation of this analysis is not looking in-depth at the context of evolutionary dynamics for the repeat architectures and whether they correlate with the abundances of repeat architectures at the flanking regions of copy number variants.

In addition, 6.76 percent and 6.45 percent of copy number variants with no Alu subtypes in their flanking regions had segmental duplications in these upstream regions and downstream regions respectively. Lastly, L1 elements were relatively more abundant as 23.98 percent and 26.64 percent of copy number variants with no Alu subtypes in flanking regions had them in upstream and downstream regions respectively. The remaining copy number variants might have other elements with sequence homology in their flanking regions that are facilitating the rearrangements. Or these variants might have arisen by non-homologous end joining which requires little to no sequence homology to facilitate the formation of these variants (Hastings et al., 2009).

In the case of L1 elements, only 21 and 22.47 percent of copy number variants had L1 elements in their upstream and downstream flanking regions respectively. There were few copy number variants that had more than 1 and up to 5 L1 elements in their flanking regions. In addition, L1MB3, L1MB5, L1MB7, L1MB8 were found to be relatively abundant in both upstream and

downstream flanking regions of copy number segments. These four L1 LINE subtypes are evolutionarily old and have pieces of evidence of existing even before the divergence of the primates (Yang et al., 2014). In addition, L1PB2 and L1ME1 were relatively abundant in downstream flanking regions but not in the upstream flanking regions. These two subtypes are relatively younger compared to the other four subtypes mentioned above. These kinds of reflects the heterogeneity in the pattern observed at the flanking regions of copy number variants for L1 elements.

For segmental duplications, only 22.05 and 22.47 percent of copy number variants had portions of segmental duplications in their upstream and downstream flanking regions respectively. Majority of these upstream and downstream flanking regions had portions of 1 to 5 high identity segmental duplications. Interestingly, there were quite a few flanking regions where the number of segmental duplications was in excess of 10 and up to 59 in some. That reflects the variability in the number of segmental duplications present at the flanking regions of copy number variants. In addition, regions with over representations of segmental duplications are considered highly susceptible to rearrangements and hence have a higher likelihood of giving rise to copy number variants (Kim et al., 2008). That is why regions with over representations of segmental duplications are considered hotspots for copy number variants (Kim et al., 2008). So, the copy number variants in the dataset that have overrepresentations of segmental duplications in their flanking regions are probably representative of regions that are hotspots for copy number variants. Lastly, the copy number variants that did not have segmental duplications had a high abundance of Alu elements in both upstream and downstream flanking regions. 68.64 and 67.97 percent of those upstream and downstream flanking regions with no segmental duplications had various subtypes of Alu elements in those regions. This again reflects the abundance of Alu elements at both upstream and downstream flanking regions.

Studies suggest that the presence of exons of dosage sensitive genes within the variant increases the likelihood of a copy number segments being pathogenic (Rice and McLysaght, 2017). Therefore, in order to gauge the pathogenicity, these variants were screened for the presence of dosage sensitive genes. Only 3.61 percent of copy number variants had exons of dosage sensitive genes present in them. This percentage demonstrates that most of these copy number segments have a higher likelihood of being benign. In addition to that, copy number variants may disrupt transcription factor binding sites or regulatory elements which in turn can cause the manifestation of diseased phenotypes by perturbation of spatial and temporal gene regulation (Haraksingh and Snyder, 2013). These non-coding regions are key players in the interaction of enhancers with promoters of genes and modulation of gene expression. Therefore, these non-coding regions indirectly modulate the transcription of nearby by genes. Perturbation of interaction of these regulatory elements with promoters or modification of chromatin interactions are one of the main mechanisms by which copy number segments are touted to affect the transcriptional landscape (Klopocki and Mundlos, 2011). Therefore, the goal was to screen which transcription factor's regulatory elements are present in these copy number variants. *CTCF* was found to be most abundant in these copy number variants followed by *CEBPD, EP300, FOXA1, MAFK, MAX, PHF8, POLR2A RAD21, RUNX3, SMC3, SPI1,* and *TEAD4*. Some of the transcription factors like for instance *RUNX3, FOXA1, CTCF* have been implicated in causing diseases, and thus perturbation of these transcription factor's regulatory elements might result in some clinical phenotype (Vecellio et al., 2015), (Sahu et al., 2011), (Gregor et al., 2013). However, many of these transcription factors mentioned are ubiquitous and are expressed in various tissues in the body. Thus, how the copy number variants would affect the pathway cascades of these regulatory elements and whether the perturbation would be enough to elucidate phenotype is yet to be fully understood. Additionally, *CTCF* is regarded as one of the essential factors in chromatin

organization and chromatin regulation and thus serves important functions in every cell type (Ohlsson et al., 2010). Recently, perturbation of *CTCF* binding sites has been associated with intellectual disability, microcephaly and growth retardation and it signifies the importance of proper functioning of *CTCF* (Gregor et al., 2013). However, these manifestations are observed in few people and it is still hard to implicate copy number variants for these manifestations (Gregor et al., 2013). One of the significant limitations of this screening is cell and tissue specificity of regulatory elements for the transcription factors. It has not been considered, and many of these elements' functions are thus subject to change and more so during developmental processes.

Next, using the same dataset, copy number variants that had 50 percent or more overlap were stratified to copy number variable regions (CNVRs) for enumerating the number of hits each copy number variable regions get. The primary purpose behind this step was to find out regions in each chromosome that were getting the maximum number of copy number variant hits. 25,340 copy number variable regions were found to have one or more copy number variant hits. When the number of hits is visualized, it can be clearly inferred that most of these copy number variable regions have hits below 15 or less. However, there are very few copy number variable regions that have copy number variant hits of 100 and more. These few copy number variable regions might be biologically significant since majority of the other regions have hits below 15. Next, all the copy number variable regions were grouped by chromosomes. It was seen that CNVRs with maximal copy number variants hits for each chromosome had hits of at least 100 or more. Looking at the repeat architectures for these regions might help determine what makes them so prone to getting higher number of copy number variants. In addition to that, looking at the genomic features might also indicate as to whether they might be benign or pathogenic.

These maximal hit CNVRs for each chromosome had representations of different Alu subtypes. AluY was relatively enriched and was present in majority of these regions. AluSx, AluSx1, AluSg and AluSq were also found to be relatively enriched in these regions. Overrepresentations of different Alu subtypes have been linked with making regions more susceptible to undergo recombination more often and consequently increasing the possibility of formation of copy number variants (Cardoso et al., 2016). Thus, the presence of different Alu subtypes in these regions may contribute to the formation of multiple copy number segments in these regions. Interestingly, the different subtypes of Alus that were found across these CNVRs were mostly relatively older Alus except AluYb9 and AluYc which were only found in CNVRs of chromosome 12,13 and X. Thus, older subtypes of Alus are more prevalent in these CNVRs than younger CNVRs. In addition, the presence of L1 elements has been known to increase genomic instability and has been implicated in the formation of copy number variants (Cardoso et al., 2016). These maximal hits regions were therefore screened for the presence of different L1 subtypes as well. It was clear that L1 elements were relatively less enriched compared to Alu elements and no specific L1 elements seem to be relatively enriched across all these regions. Segmental duplications are another class of repeat architecture that are implicated as a causative factor in the formation of many syndromic disorders (Cardoso et al., 2016). Strikingly, there were seven maximal hit copy number variable regions that did not have any segmental duplications in them. Instead, most of these regions had high representations of different Alu subtypes in them. Thus, the absence of segmental duplications in these regions signifies the importance of high representations of different Alu subtypes. Furthermore, in these segmental duplications depleted regions, it is highly likely that Alu elements might be the main repeat architecture sensitizing the genome to undergo more rearrangement and consequently increasing the likelihood for the formation of larger number of copy number segments. An exception was the CNVR of chromosome 2. Although this specific CNVR was getting copy number variant hits of 620, no Alu subtypes were present within the

segment. On the contrary, interestingly, 6 segmental duplications were present which might indicate the increased likelihood of segmental duplications being responsible for heightened propensity of this specific region to copy number variants.

To assess the likelihood of pathogenicity of these maximal hit copy number variable regions, these regions were screened for dosage sensitive genes with haploinsufficiency score of 3 and more. Only one of the maximal hit copy number variable regions on chromosome 16 had a dosage sensitive gene which was *PKD1* gene. The absence of dosage sensitive genes in these regions illustrate the low likelihood of these regions being pathogenic. Additionally, these regions were also screened for the presence of transcription factor specific regulatory elements. *CTCF, RAD21, POLR2A, STAT3, ZNF263, SMC3, RUNX3 and MAFK* were the most prevalent in these regions. The regulatory elements that were present in these regions were similar to regulatory elements that were enriched in the copy number variants. Dysregulation of *CTCF*, and *RUNX3* have all been associated with disease (Gregor et al., 2013) (Vecellio et al., 2015). Thus, having these regulatory elements increase the likelihood of these copy number variable regions being pathogenic since perturbation of interaction of these regulatory elements might result in the manifestation of phenotypes. Having said that, the tissue and cell-type specificity of these regulatory elements have not been considered which is a major limitation. From these findings, it can be said that the majority of copy number variable regions with maximum hits across each chromosome are most likely benign. Moreover, that is indicative of the dataset since copy number variants were called from a general population of normal individuals who do not exhibit life-threatening phenotypes.

REFERENCES

Arlt, M., Wilson, T., and Glover, T. (2012). Replication stress and mechanisms of CNV formation. Current Opinion In Genetics & Development *22*, 204-210.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q., Antipenko, A., Shang, L., Boisson, B., Casanova, J., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings Of The National Academy Of Sciences *112*, 5473-5478.

Cardoso, A., Oliveira, M., Amorim, A., and Azevedo, L. (2016). Major influence of repetitive elements on disease-associated copy number variants (CNVs). Human Genomics *10*.

Carvalho, C., and Lupski, J. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics *17*, 224-238.

Chen, L., Zhou, W., Zhang, L., and Zhang, F. (2014). Genome Architecture and Its Roles in Human Copy Number Variation. Genomics & Informatics *12*, 136.

Clancy, S. (2008). Copy Number Variation. Nature Education *1*, 95.

Elsea, S., and Girirajan, S. (2008). Smith–Magenis syndrome. European Journal Of Human Genetics *16*, 412-421.

Girirajan, S., Campbell, C., and Eichler, E. (2011). Human Copy Number Variation and Complex Genetic Disease. Annual Review Of Genetics *45*, 203-226.

Girirajan, S., Vlangos, C., Szomju, B., Edelman, E., Trevors, C., Dupuis, L., Nezarati, M., Bunyan, D., and Elsea, S. (2006). Genotype–phenotype correlation in Smith-Magenis syndrome: Evidence that multiple genes in 17p11.2 contribute to the clinical spectrum. Genetics In Medicine *8*, 417-427.

Gregor, A., Oti, M., Kouwenhoven, E., Hoyer, J., Sticht, H., Ekici, A., Kjaergaard, S., Rauch, A., Stunnenberg, H., and Uebe, S. et al. (2013). De Novo Mutations in the Genome Organizer CTCF Cause Intellectual Disability. The American Journal Of Human Genetics *93*, 124-131.

Haraksingh, R., and Snyder, M. (2013). Impacts of Variation in the Human Genome on Gene Regulation. Journal Of Molecular Biology *425*, 3970-3977.

Hastings, P., Lupski, J., Rosenberg, S., and Ira, G. (2009). Mechanisms of change in gene copy number. Nature Reviews Genetics *10*, 551-564.

Itsara, A., Cooper, G., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R., Myers, R., Ridker, P., and Chasman, D. et al. (2009). Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. The American Journal Of Human Genetics *84*, 550-551.

Kim, P., Lam, H., Urban, A., Korbel, J., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. Genome Research *18*, 1865-1874.

Klopocki, E., and Mundlos, S. (2011). Copy-Number Variations, Noncoding Sequences, and Human Phenotypes. Annual Review Of Genomics And Human Genetics *12*, 53-72.

Konkel, M., Walker, J., Hotard, A., Ranck, M., Fontenot, C., Storer, J., Stewart, C., Marth, G., and Batzer, M. (2015). Sequence Analysis and Characterization of Active HumanAlusubfamilies Based on the 1000 Genomes Pilot Project. Genome Biology And Evolution evv167.

Li, J., Parker, B., Martyn, C., Natarajan, C., and Guo, J. (2012). The PMP22 Gene and Its Related Diseases. Molecular Neurobiology *47*, 673-698.

MacDonald, J., Ziman, R., Yuen, R., Feuk, L., and Scherer, S. (2013). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Research *42*, D986-D992.

Manzoni, C., Kia, D., Vandrovcova, J., Hardy, J., Wood, N., Lewis, P., and Ferrari, R. (2016). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings In Bioinformatics *19*, 286-302.

Ohlsson, R., Bartkuhn, M., and Renkawitz, R. (2010). CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. Chromosoma *119*, 351-360.

Rice, A., and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. BMC Biology *15*.

Robaglia-Schlupp, A. (2002). PMP22 overexpression causes dysmyelination in mice. Brain *125*, 2213-2221.

Ruderfer, D., Hamamsy, T., Lek, M., Karczewski, K., Kavanagh, D., Samocha, K., Daly, M., MacArthur, D., Fromer, M., and Purcell, S. (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nature Genetics *48*, 1107-1111.

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., Sankila, A., Turunen, J., Lundin, M., and Konsti, J. et al. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. The EMBO Journal *30*, 3962-3976.

Sharp, A., Mefford, H., Li, K., Baker, C., Skinner, C., Stevenson, R., Schroer, R., Novara, F., De Gregori, M., and Ciccone, R. et al. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nature Genetics *40*, 322-328.

Stewart, D., and Kleefstra, T. (2007). The chromosome 9q subtelomere deletion syndrome. American Journal Of Medical Genetics Part C: Seminars In Medical Genetics *145C*, 383-392.

Sudmant, P., Kitzman, J., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., and Eichler, E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. Science *330*, 641-646.

Vecellio, M., Roberts, A., Cohen, C., Cortes, A., Knight, J., Bowness, P., and Wordsworth, B. (2015). The genetic association ofRUNX3with ankylosing spondylitis can be explained by allele-specific effects on IRF4 recruitment that alter gene expression. Annals Of The Rheumatic Diseases *75*, 1534-1540.

Yang, L., Brunsfeld, J., Scott, L., and Wichman, H. (2014). Reviving the Dead: History and Reactivation of an Extinct L1. Plos Genetics *10*, e1004395.

Zarrei, M., MacDonald, J., Merico, D., and Scherer, S. (2015). A copy number variation map of the human genome. Nature Reviews Genetics *16*, 172-183.

Zhang, F., Potocki, L., Sampson, J., Liu, P., Sanchez-Valle, A., Robbins-Furman, P., Navarro, A., Wheeler, P., Spence, J., and Brasington, C. et al. (2010). Identification of Uncommon Recurrent Potocki-Lupski Syndrome-Associated Duplications and the Distribution of Rearrangement Types and Mechanisms in PTLS. The American Journal Of Human Genetics *86*, 462-470.

Appendix

Coordinates of CNVRs with maximum CNV hits

Table 1: Coordinates of CNVRs with maximum CNV hits

Chromosome	Start	End	CNV hits	Cytoband
chr1	17600000	17650000	449	1p36.3
chr2	96550000	96600000	630	2q11.1
chr3	35800000	35850000	176	3p22.3
chr4	3400000	3450000	381	4p16.3
chr5	110400000	110450000	235	5q22.1
chr6	31350000	31400000	220	6p21.33
chr7	141750000	141800000	649	7q34
chr8	15950000	16000000	251	8p22
chr9	117050000	117100000	366	9q32
chr10	27650000	27700000	266	10p12.1
chr11	46750000	46800000	212	11p11.2
chr12	19450000	19500000	538	12p12.3
chr13	20400000	20450000	358	13q12.1
chr14	21650000	21700000	134	14q11.2
chr15	79050000	79100000	98	15q25.1
chr16	2150000	2200000	262	16p13.3
chr17	33700000	33750000	427	17q12
chr18	61300000	61350000	229	18q21.33
chr19	9000000	9050000	376	19p13.2
chr20	44350000	44400000	454	20q13.12
chr21	45800000	45850000	199	21q22.3
chr22	39400000	39450000	323	22q13.1
chrX	650000	700000	8	Xp22.3
chrX	2000000	2050000	8	Xp22.3
chrY	1850000	1900000	2	Yp11.32