A BIOINFORMATIC WORKFLOW FOR ANALYZING WHOLE GENOMES IN RARE MENDELIAN DISEASE

by

Madeline Hazel Couse

B.Sc., The University of Waterloo, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2017

Abstract

The vast majority of the human genome (˜98%) is non-coding. A symphony of non-coding sequences resides in the genome, interacting with genes and the environment to tune gene expression. Functional non-coding sequences include enhancers, silencers, promoters, non-coding RNA and insulators. Variation in these non-coding sequences can cause disease, yet clinical sequencing in patients with rare Mendelian disease currently focuses mostly on variants in the ˜2% of the genome that codes for protein. Indeed, variants in protein-coding genes that can explain a phenotype are identified in less than half of patients with suspected genetic disease by whole exome sequencing (WES). With the dramatic reduction in the cost of whole genome sequencing (WGS), development of algorithms to detect variants longer than 50 bp (structural variants, SVs), and improved annotation of the non-coding genome, it is now possible to interrogate the entire spectrum of genetic variation to identify a pathogenic mutation.

A comprehensive pipeline is needed to analyze non-coding variation and structural variation from WGS. In this thesis, I developed and benchmarked a bioinformatics workflow to detect pathogenic non-coding SNVs/indels and pathogenic SVs, and applied this workflow to unsolved patients with rare Mendelian disorders. The pipeline detected ˜80-90% of deletions, ˜90% of duplications, ˜65% inversions, and ˜50% of insertions in a simulated genome and the NA12878 genome. The pipeline captured the majority of known pathogenic non-coding single nucleotide variant (SNVs) and insertion deletions (indels), and effectively prioritized a spiked-in known pathogenic non-coding SNV. Several interesting candidate variants were detected in patients, but none could be convincingly implicated as pathogenic.

The bioinformatic workflow described in this thesis is complementary to sequencing pipelines that analyze only protein-coding variants from whole genomes. Application of this workflow to larger cohorts of patients with rare Mendelian diseases should identify pathogenic non-coding variants and SVs to increase diagnostic yield of clinical sequencing studies, assist management of genetic diseases, and contribute knowledge of novel pathogenic variants to the scientific community.

Preface

This thesis comprises unpublished work performed by the author. All analyses of non-coding variation and structural variation described in this paper were performed by the author. Patients with hereditary sensory and autonomic neuropathy were recruited by Gabriella Horvath, who also helped in interpretation of variants. Analysis to rule out pathogenic exonic variants was performed by Farah Zahir, past member of the Friedman lab, assisted by Clara Van Karnebeek, Casper Shyr, and Maja Tarailo-Graovac of the Treatable Intellectual Disability Endeavour (TIDE) BC team. Patients with Aicardi syndrome were recruited by Cristina Dias. Sequencing, as well as analysis and interpretation of exonic variants were the collaborative effort of Jan Friedman, Cristina Dias, Steven Jones, Farah Zahir, and Yaoqing Shen. Allison Matthews, Jill Mwenifumbo, and Phillip Richmond from Wyeth Wasserman's lab provided insight into variant interpretation and structural variant analysis. Shaun Jackman assisted with interpretation of the *DNMT1* tandem duplication.

This research was covered by the UBC Research Ethics Boards under the project "Genetic Alterations in Rare Diseases", certificate number H09-01228.

Figure 1.1 was adapted from *Expert Review in Molecular Medicine*, 17, Philip Cowie, Elizabeth A. Hay, Alasdair MacKenzie, The Noncoding Human Genome and the Future of Personalised Medicine, p.4, Copyright (2015), with permission from Cambridge University Press. Figure 1.3 was reproduced from *Cell*, 161:5, Dario G. Lupianez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hlya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas et al., Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions, p.1014, Copyright (2015), with permission from Elsevier. Figure 1.4 was adapted from *Nature Methods*, 17, Monya Baker, Structural variation: the Genome's Hidden Architecture, p.133, Copy-

Table of Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

CRE cis-regulatory element
WGS whole genome sequence/sequencing
TFBS transcription factor binding site
TSS transcription start site
TAD topologically associated domain
BCA balanced chromosomal abnormality
lncRNA long non-coding RNA
snoRNA small nucleolar RNA
miRNA microRNA
siRNA short interfering RNA
WES whole exome sequence/sequencing
HGMD Human Gene Mutation Database
CNV copy number variant
SV structural variant
CADD combined annotation dependent depletion
SNV single nucleotide variant
Indel insertion deletion
FATHMM functional analysis through hidden Markov model
CMA chromosomal microarray analysis
PE paired-end
RD read-depth
RP read-pair
SR split read
AS de novo assembly
HSAN hereditary sensory and autonomic neuropathy
DGV Database of Genomic Variants
GIAB Genome in a Bottle
1000G 1000 Genomes Project
PPV positive predictive value
TP true positive
FN false negative

FP false positive
DECIPHER Database of Genomic Variation and Phenotype in Humans
SRS short-read sequencing
LRS long-read sequencing
PacBio Pacific Biosciences
SMRT single molecule real-time

Acknowledgements

For funding, I would like to thank NSERC for the Canada Graduate Scholarship Masters Program award, and my supervisor, Jan Friedman.

Many thanks to Jan Friedman for his uncompromising intellect, kind support, and sage advice. I feel very fortunate to have had him as a supervisor. Thank you to the past and present members of the Friedman lab. Thank you to my committee members, Wyeth Wasserman and Inanc Birol, for taking the time to advise me on my thesis.

Thank you to Green College for providing me with a warm and open-hearted community for my first two years in Vancouver. Thank you to my ever-supportive parents and siblings. Big thanks to my twin sister, Margot Couse, for the countless hours spent mumbling and grumbling with me over the phone. Thanks to my lovely friends in Vancouver and elsewhere. And finally, thank you to Lucian Go, for putting up with me while I wrote this thesis.

Chapter 1

Introduction

1.1 The Human Genome

The first draft of the human genome was published over fifteen years ago [29]. The announcement was greeted with excitement and high expectations. With the sequence of the human genome, the blueprint for human life, would scientists lay bare the mysteries encoded in our DNA? The human genome was seemingly simpler and yet more complex than previously imagined. The genome encodes a mere ˜21,000 protein coding genes, less than one quarter the number that had been predicted. On the other hand only ˜2% of the genomes ˜3 billion bases comprise protein-coding genes. Clearly, there was more to the genome than the simplistic and deterministic genome-as-blueprint concept. Indeed, scientists uncovered a symphony of cis-regulatory elements (CRE) residing in the non-coding genome: promoters, enhancers, silencers, insulators, and various classes of non-coding RNAs. These elements respond to dynamic environmental cues to tune gene expression through space and time, guided by the genomes intricate spatial architecture. As our understanding of the non-coding genome and the overall genomic structure has deepened, so too has our knowledge of their contribution to human disease.

Though we have begun to unravel the role of the non-coding genome in shaping phenotypes, this knowledge has not yet been harnessed in clinical sequencing pipelines for diagnosing disease. These pipelines focus on the 2% of DNA in the protein-coding regions of the genome. This thesis will attempt to integrate structural and CRE analysis into a whole genome sequence (WGS) analysis pipeline for prioritizing regulatory and structural variants in rare Mendelian disease.

1.2 Genome Regulation and Dysregulation

Gene expression regulation begins at the cell surface, where ligand-receptor interactions initiate signal transduction cascades that eventually activate

transcription factors [35]. Such activated transcription factors bind CREs, which then interact locally or distally with RNApolII at a gene promoter to induce or repress gene expression (fig.1.1). Above the linear sequence of DNA, the 3D topology of the genome directs and constrains interactions of CREs with their cognate promoters to orchestrate gene expression.

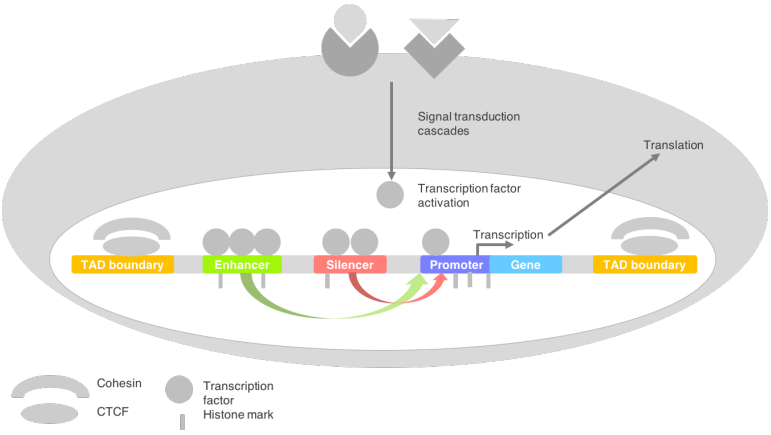


Figure 1.1: The flow of information in a cell, greatly simplified. Modified from Cowie et al. 2015, with permission from Cambridge University Press. [11]

Variants in CREs may confer disease risk by altering transcription factor binding sites (TFBS), disrupting regulatory domains, or influencing susceptibility of a sequence to epigenetic modification [38]. Many examples of pathogenic variants residing in or disrupting regulatory sequences have been identified in recent years. Indeed, a recent paper compiled 453 different non-coding variants associated with Mendelian diseases [52]. CRE structure and function, as well as examples of diseases resulting from their disruption, are described below.

1.2.1 Promoters

A promoter is the sequence upstream of a gene transcription start site (TSS) that is bound by the transcriptional apparatus to initiate gene expression. A core promoter refers to the ˜50 bases that bind the transcription pre-initiation complex, which includes RNA polymerase II and basal transcription factors [11]. Meanwhile, the non-core promoter may be kilobases in length and encompass TFBS that are necessary for tissue-specific gene expression. Mutations in promoter sequences can introduce or remove TFBS

to change levels of gene expression.

For example, a constitutional mutation in the promoter of *telomerase reverse transcriptase (TERT)*, which encodes a catalytic subunit of telomerase, was found to segregate with the disease in a family with hereditary melanoma [23]. *TERT* is upregulated in 90% of cancers and is associated with immortality in cancer cells. The germline mutation was located 57 bp upstream of the transcription initiation site. Mutations occurring somatically 124 and 146 bp downstream of the transcription initation site were also found in tumors of several unrelated patients. These germline and somatic mutations were found to create a CCGGAA/T binding motif for Ets/TCF transcription factors that resulted in increased *TERT* expression.

1.2.2 Enhancers and Silencers

Enhancers and silencers are short (50-1500 bp) CREs that are bound by transcription factors to upregulate or downregulate promoter activity, respectively. Their activity is orientation and distance-independent [36]. Indeed, enhancers may target promoters at a distance as far as one megabase; this was observed fortuitously in a mutant mouse, *sasquatch*, generated by random integration of a transgene into an intron of *LMBR1*, 1 Mb away from the *Sonic hedgehog gene (SHH)* [31]. The transgene disrupted a long-range enhancer that regulates developmental *SHH* expression and causes preaxial polydactyly (PPD). Dysregulated *SHH* is ectopically expressed in the anterior margin of the mouse limb bud, causing the limb to develop preaxial digits. Point mutations in this long-range enhancer in both mice and humans also cause PPD.

1.2.3 Topologically Associated Domains

Our DNA is bound by proteins, including histones and transcription factors, as well as non-coding RNAs. This assemblage of molecules, termed chromatin, is folded like origami into the nucleus. It is thought that interactions between distal CREs and genes are mediated by folding of the intervening chromatin in regulatory domains termed topologically associated domains (TADs) (fig. 1.2). TADs are discrete genomic regions about ˜1 Mb in size that interact with themselves at a higher frequency than with the rest of the genome. TADs are bordered by regions with low interaction frequency, called TAD boundaries [37]. The TAD boundaries are associated with insulator binding factor CTCF, housekeeping and tRNA genes, and SINE elements[13]. Dixon et al performed Hi-C experiments in mouse embryonic

stem (ES) cells, human ES cells, and human IMR90 cells and showed that TAD boundaries are largely conserved between mice and humans (75.9% of mouse boundaries are boundaries in humans, compared to 29.0% expected by chance), and, further, TAD boundaries are largely invariant across cell types.

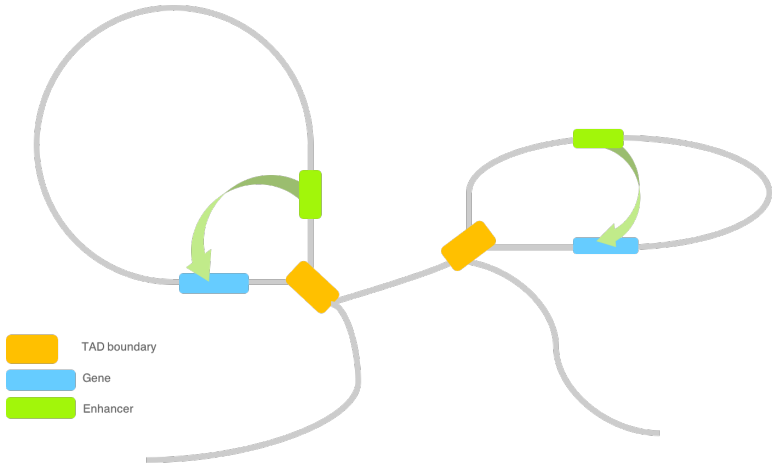


Figure 1.2: DNA is partitioned into TADs. TAD boundaries prevent regulatory elements in one TAD from interacting with genes in neighbouring TADs.

Disruptions to TADs can produce disease. In several families with rare limb malformations, genomic rearrangements were observed at the *WNT6 /IHH/EPHA4/PAX3* locus [32]. All of these rearrangements disrupted a TAD boundary, either telomeric or centromeric to the *EPHA4* TAD (fig. 1.3). Disruption of either boundary caused ectopic interactions of a 150 kb region within the EPHA4 TAD with neighboring TADs. This region was shown to contain a cluster of enhancers driving limb expression of *EPHA4*; in mutants, this region targeted *PAX3, WNT6,* and *IHH*, resulting in ectopic expression and limb malformations in mice and humans.

Translocations and inversions are referred to as balanced chromosomal abnormalities (BCAs) if they do not cause any net gain or loss of genetic material. TAD disruptions were recently shown to cause long-range genetic regulatory changes in 7% of a developmental anomaly cohort with BCAs studied by whole genome sequencing [47]. Interestingly, BCA breakpoints in eight patients impacted a TAD encompassing one particular gene, *MEF2C*, which lies in the critical region of the 5q14.3q15 microdeletion syndrome. The breakpoints in these patients were all at difference loci within the TAD, and

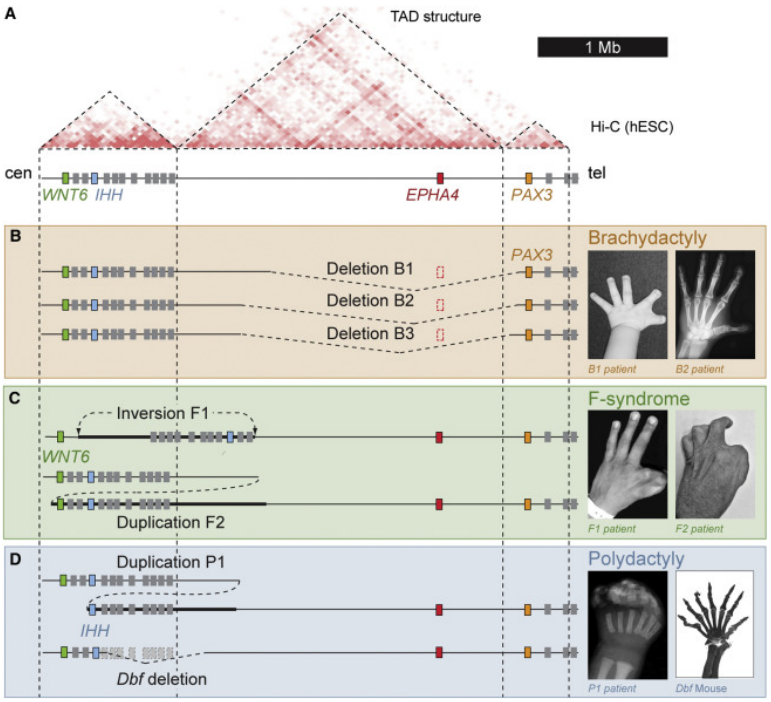


Figure 1.3: Different Limb Pathogenic Structural Variations in Human and Mouse Map to the *EPHA4* TAD (A) Hi-C profile around the *EPHA4* locus in human ESCs (Dixon et al., 2012). Dashed lines indicate the *EPHA4* TAD and boundaries. Cen, centromeric; tel, telomeric.(BD) Schematic of structural variants (left) and associated phenotypes (right). (B) Brachydactyly-associated deletions in families B1, B2, and B3. Note thumb and index finger shortening with partial webbing in a child (B1 patient) and adult (B2 patient). (C) F-syndrome-associated inversion in family F1 and duplication in family F2. Note similar phenotypes of index/thumb syndactyly. (D) Polydactyly-associated duplication (P1) and deletion in the doublefoot (Dbf) mouse mutant. The radiograph of the patients hand and the skeletal preparation of the Dbf/+ mouse show similar seven-digit polydactyly. Reproduced from Figure 1 of Lupiez et al., 2015, with permission from Elsevier. [32]

none affected the adjacent TAD boundaries, but all still changed *MEF2C* expression and likely caused the developmental anomalies in these patients.

1.2.4 Non-coding RNAs

Much of the non-coding genome is transcribed, and the role of these non-coding transcripts in regulating gene expression is well appreciated [17] . Regulatory RNAs can be divided into long non-coding RNAs (lncRNA) and small non-coding RNAs including small nucleolar RNA (snoRNA), microRNA (miRNA), and short interfering RNA (siRNA). snoRNAs are involved in chemically modifying other RNA, such as transfer RNA and ribosomal RNA, while siRNAs and miRNAs repress gene expression after transcription [6]. Small RNAs are associated with many regulatory roles, notably in brain development. Indeed, small RNA dysregulation is associated with neurodevelopmental and neurodegenerative disorders, and with brain cancer [6]. For instance, microdeletions at 15q11.2 resulting in loss of the paternal copy of SNORD116 snoRNAs cause Prader-Willi syndrome [15]. Heterozygous mutations in the seed region of the miRNA *MIR96* cause nonsyndromic progressive hearing loss [40].

lncRNAs are regulatory elements that are gene-like in structure, with promoters, introns and exons. lncRNA expression is highly tissue-specific; these RNAs are implicated in the regulation of cell maintenance and fate, particularly in the brain [6]. *De novo* translocations disrupting *LINC00299* are implicated in neurodevelopmental disability [55].

1.3 Next Generation Sequencing and Clinical Studies

The estimated cost to sequence the first draft of the human genome via Sanger sequencing was $300 million
(`https://www.genome.gov/sequencingcosts/`). Since the introduction of next-generation sequencing a decade ago, that figure has dropped dramatically, outpacing Moores law for computing costs. Remarkably, the cost of sequencing the genome using Illuminas HiSeq X Ten machines (after the system cost) has broken the US$1000 goal set by the National Human Genome Research Institute; this represents a 10,000 fold price reduction compared to 2004 [57]. With the human genome sequence known and increasingly affordable next-generation sequencing technology available, the number of Mendelian disease genes that have been identified increased from 100 to 3000

within a decade [29]. Simultaneously, the number of sequenced genomes has exploded. Large-scale sequencing projects like the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), and the ESP (NHLBI GO Exome Sequencing Project, URL: `http://evs.gs.washington.edu/EVS/`) have made hundreds of thousands of exomes and whole genomes publicly available, benefitting clinical sequencing projects greatly with databases of control exomes/genomes from many populations. Although single-gene testing remains an appropriate diagnostic approach for non-heterogeneous disorders with distinct phenotypes, whole exome sequencing (WES) and WGS are the approaches of choice for investigating heterogeneous known or suspected Mendelian disorders.

1.3.1 Exome Sequencing

WES is a sequencing method that targets the protein-coding regions of the genome. It is the current method of choice for clinical sequencing studies and has been the most commonly used tool for Mendelian disease gene discovery[61]. WES interrogates less than 2% of the genome, yet it covers ˜85% of known disease-causing variants [57]. However, this proportion is subject to ascertainment bias, as the search for disease-causing variants has largely been limited to exonic analysis. WES studies on children with birth defects or neurodevelopmental disorders have usually reported diagnostic rates of ˜25-28% [59] but up to 50% with more stringent patient selection criteria [44].

There are several limitations to the use of WES for identifying genetic variation. First, WES captures exons in a non-uniform manner, resulting in insufficient coverage of some genes. A recent study [14] with a TruSeq capture kit demonstrated that an average of 10% of exons had a minimum coverage of fewer than 10 reads, despite an overall mean coverage of over 80X. Of these exons, a quarter resided in genes harboring known or likely disease-causing Human Gene Mutation Database (HGMD) variants. In addition, WES lacks reliability and sensitivity in detecting small copy number variants (<100kb), large indels (<50bp), and other complex structural variants (SVs) [53]. Finally, WES does not capture the vast majority of the non-coding genome.

1.3.2 Whole Genome Sequencing

Due in large part to its higher cost and challenges in data interpretation, WGS has been performed far less frequently than WES for clinical diagno-

sis. However, WGS covers both the coding and non-coding genome more uniformly than WES, and, further, can detect SVs across a wide range of sizes with single base resolution.

The advantages of WGS over WES have been demonstrated in a number of studies. In a cohort of 50 patients with severe intellectual disability who remained undiagnosed after microarray analysis and exome sequencing, Gilissen et al[20] detected 8 *de novo* copy number variants (CNVs), which play an important role in neuropsychiatric disorders. Carss et al[7] used WES and WGS to identify rare pathogenic variants in a phenotypically and genetically heterogeneous cohort of 722 individuals with inherited retinal disease. Forty five of the individuals unsolved by WES underwent WGS, and pathogenic variants not detectable by WES were identified in a further six individuals. For 3 of these patients, this was due to coverage in WGS at a variant location that was absent from the bait in the exome capture kit. Two other individuals had a large deletion and one individual had a large indel not called by WES. In 605 patients who underwent WGS, a total of 33 SVs (31 deletions and 2 tandem duplications) were identified with precise breakpoint resolution, which would not have been possible with WES. Finally, the authors demonstrated the superior uniformity of coverage in GC rich regions afforded by WGS in comparison to WES, with the identification of compound heterozygous mutations in one individual in the first exon of *GUCY2D*, with a 76% GC content. This exon was not covered in the WES capture kit.

1.4 Prioritizing Non-coding SNVs and Indels

The whole genome sequence of any individual varies from the reference human genome at millions of sites. Even after filtering variants that occur at polymorphic frequencies in normal populations from WGS of patients with rare Mendelian diseases, the variant list can contain hundreds of thousands of non-coding and coding variants. Filtering for variants in regulatory regions is one strategy to reduce the search space in order to identify the one or two pathogenic non-coding variants responsible for a Mendelian disease. Another is to use computational predictions of functionality based on models derived from diverse sets of genome annotations and known pathogenic variants. These scores can help to prioritize genetic variants by providing likelihoods that a variant is deleterious or functional. Two scores that apply to both coding and non-coding variants are the CADD score and the FATHMM score.

1.4.1 Combined Annotation Dependent Depletion (CADD) score

The CADD score integrates diverse genome annotations of fixed or nearly fixed alleles with simulated alleles to score the deleteriousness of any possible single nucleotide variant (SNV) or small insertion-deletion (indel) [28]. Deleteriousness, corresponding to a reduction in organismal fitness, correlates with molecular functionality and pathogenicity. The CADD score is derived from 63 distinct annotations including conservation metrics such as GERP and phastCons, regions of DNAse hypersensitivity, transcription factor binding, and protein-level scores such as SIFT and PolyPhen. All 8.6 billion possible SNVs in the genome are given a CADD score. The CADD scores are phred-scaled from 1 to 99, where variants in the highest 10% of all scores are assigned scores of 10 or greater, variants in the highest 1% are scored 20 or greater, variants in the highest 0.1% 30 or greater, and so on.

1.4.2 Functional Analysis Through Hidden Markov Model (FATHMM) score

FATHMM is a machine learning approach integrating 46 sequence conservation, histone modification, transcription factor binding site, and open chromatin annotations to assess the functional consequences of non-coding and coding variants [50]. Unlike CADD, FATHMM uses an algorithm to weight different annotations according to relevance, and according to the paper, outperforms CADD in predicting functional consequences of non-coding variants. FATHMM scores for all possible SNVs are available, and range from 0 to 1. Scores of greater than 0.5 indicate that a variant is likely to be functional.

1.5 Structural Variation

1.5.1 SV classes and detection from NGS

SVs are balanced or unbalanced genomic rearrangements that affect more than 50 bp of genomic sequence. Although SVs are estimated to impact more than 1% of each human genome, versus 0.1% for SNPs [45], SVs are under-ascertained in clinical sequencing studies, mainly due to the shortcomings of SV calling with WES. Pathogenic SVs are typically identified clinically by karyotyping or chromosomal microarray analysis (CMA); however, WGS affords a finer resolution and broader scope of SV identification.

SVs that can be detected from WGS are insertions, deletions, duplications, inversions, and inter/intrachromosomal translocations (fig. 1.4). Although the full spectrum of SVs can, in principle, be identified from WGS, SV detection in practice is limited by the length of the reads; WGS read length is typically 50-400 bp, shorter than most SVs. SV detection is further limited in repetitive regions of the genome, which are known to be variable in structure but to which short reads cannot be uniquely mapped [41]. As such, a number of strategies have been developed to identify SV signatures from short-read paired-end (PE) sequencing.

1.5.2 Algorithms for identifying SVs

There are four different approaches designed to identify SVs from PE NGS: read-depth (RD), read-pair (RP), split-read (SR), and *de novo* assembly (AS) (fig. 1.5). RD methods are based on deviation of sequencing depth from the local genomic average and are used to detect copy number variants (CNVs). RP, SR, and AS methods use sequence signatures from PE reads to identify SVs. PE sequencing produces reads sequenced from both ends of DNA fragments, termed read pairs. On average, the length between read pairs will be consistent (e.g. 350 bp). RP methods are based on read pairs with insert-sizes or orientations that are inconsistent with expected values. A read that aligns to two separate locations in the reference genome and whose mate maps uniquely is termed a SR. SRs allow single base-pair resolution of breakpoints. AS methods order reads and merge them into larger fragments, called contigs, to reassemble the original sequence without the use of a reference. A method that incorporates all methods should be able to detect the broadest spectrum of SVs with high sensitivity.

1.6 Rare Disease Cohorts

1.6.1 Hereditary Sensory and Autonomic Neuropathy (HSAN)

HSAN is a group of clinically and genetically heterogeneous disorders of the peripheral nervous system. Our cohort is comprised of eight patients in six families with early-onset HSAN whose symptoms include loss of pain sensation, developmental delay, and gastrointestinal complications. WGS was performed on these patients, and my colleagues performed an analysis of rare exonic SNVs and indels. A genetic diagnosis was only reached in two patients.

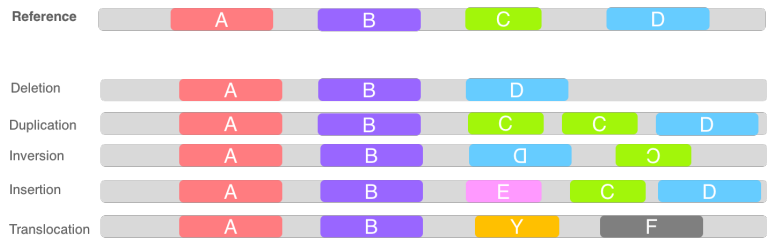


Figure 1.4: SV classes that are detectable by WGS. Modified from Baker, 2012, with permission from Nature Publishing Group. [5]

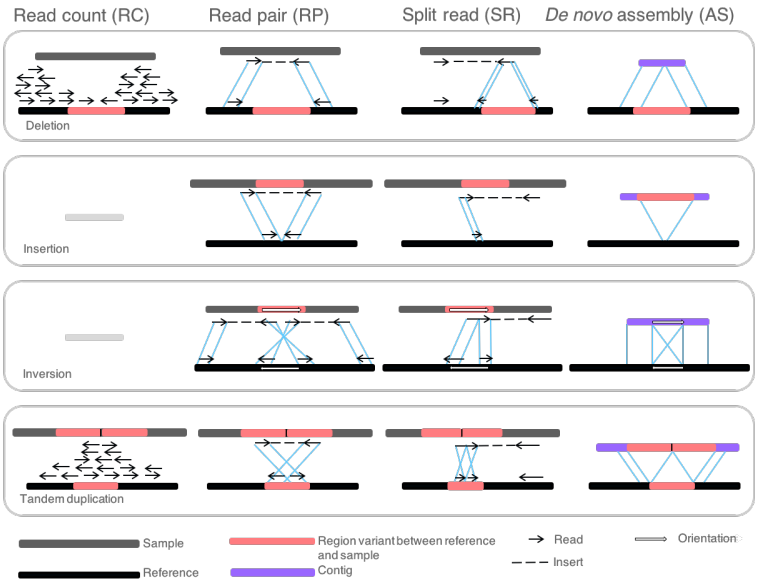


Figure 1.5: PE read signatures for SVs from read count (RC), read-pair (RP), split-read (SR), and de novo assembly (AS) methods. Modified from Tattini et al 2015, with permission under the Creative Commons Attribution License [56].

1.6.2 Aicardi Syndrome

Aicardi syndrome is an extremely rare neurodevelopmental disorder characterized by chorioretinal lacunae, agenesis of the corpus callosum, and infantile spasms. The disorder has been described almost exclusively in females or in boys with Klinefelter syndrome (XXY), with a risk to siblings of less than 1% [2]. It is, therefore, hypothesized to be caused by a dominant, male lethal *de novo* mutation in a gene on the X chromosome. Chromosome microarray analysis and exome sequencing of patients with Aicardi syndrome have not revealed the genetic cause for the disorder. Our Aicardi syndrome cohort is comprised of 9 patients recruited from across Canada and the United States. Analysis of protein-coding variations in these patients has not revealed a candidate gene.

1.7 Thesis Rationale and Objective

WES can be used to investigate SNVs and small (<50 bp) indels in protein-coding genes. WES is currently the technology of choice for clinical sequencing studies, but SNVs and indels in protein-coding genes that can explain a phenotype are identified in less than 50% of patients with suspected genetic disease by WES. There is growing interest in using WGS for clinical diagnosis now that WGS is more affordable and functional annotation of the genome is improving. In addition, numerous methods have been developed for calling SVs from WGS data.

I hypothesize that SNVs and indels that disrupt regulatory sequences and SVs in coding or non-coding sequences are a cause of genetic disease in patients who remain undiagnosed after CMA and WES (or coding SNV/indel analysis of WGS). The objectives of this thesis are to 1) develop and benchmark a bioinformatics workflow for detection of pathogenic non-coding SNVs/indels and pathogenic SVs, and 2) to use this workflow to analyze the WGS of unsolved patients recruited from in-house HSAN and Aicardi syndrome studies to test my hypothesis.

Chapter 2

Methods

Figure 2.1 illustrates the bioinformatics workflow created to identify regulatory variants and SVs from WGS. The steps taken in the pipeline and the benchmarking performed to validate the tools are described below. Software versions and parameters are listed in table 2.1.

2.1 Sequencing and Alignment

2.1.1 Whole Genome Sequencing

HSAN samples were sequenced following the manufacturers recommendations on Illumina HiSeq2000 machines in Corey Nislows lab at the University of British Columbia, Vancouver, B.C., Canada. Aicardi samples were sequenced following the manufacturers recommendations on Illumina HiSeq2000 machines at the Michael Smith Genome Sciences Centre, Vancouver, B.C., Canada. Chastity-failed reads, or reads with a high frequency of low quality bases at the beginning of the read, were excluded from further analyses. HSAN samples were sequenced to 30X with 100 bp paired-end reads. WGS was performed on samples from nine Aicardi syndrome patients to 30X with 125 bp paired-end reads. Of these nine patients, two were additionally sequenced as trios; affected tissue from these probands was sequenced to 100X with 100 bp paired-end reads and their parents were sequenced to 30X with 100 bp paired-end reads. Many reads from one trio probands affected tissue, however, were dropped due to low quality, leading to a coverage of 30X.

2.1.2 Alignment and SNV and Indel calling

Fastq files were aligned with SpeedSeq, a wrapper with a modular architecture for performing rapid, parallelized whole-genome alignment and variant calling. Default parameters were used. SpeedSeq uses BWA mem to align raw reads to a reference genome, SAMBLASTER to rapidly mark duplicate reads, Sambamba to sort reads and covert SAM to BAM, and FreeBayes to

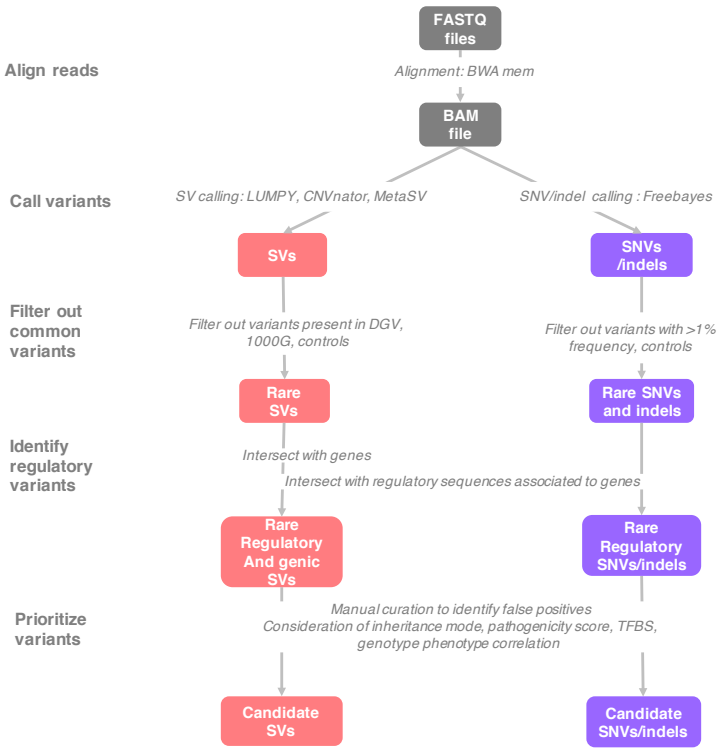


Figure 2.1: Bioinformatic workflow for transforming raw sequence reads to rare regulatory variants. The workflow for processing SVs from a BAM file is illustrated in red on the left branch, while the purple branch on the right illustrates the workflow for processing SNVs and indels from a BAM file.

call SNVs/indels. SpeedSeq also extracts split reads and discordant reads pairs for downstream SV calling with LUMPY. HSAN samples were aligned to the human genome build 19, or hg19. Aicardi samples were aligned to the human genome build 38, or hg38, to take advantage of its improvements in the X chromosome assembly.

2.2 Structural Variant Calling and Benchmarking

2.2.1 MetaSV Consensus SV Caller

MetaSV is a consensus SV caller that was used to incorporate SV calls from multiple tools to provide a high-sensitivity SV set [42]. This is useful as there is no single SV calling tool that optimally detects all SVs across a range of sizes. MetaSV provides support for output of CNVnator, an RD approach [1], Breakdancer, an RP approach [8], Pindel, an SR and RP approach [60], LUMPY, an SR and RP approach [30], and Manta, an SR and RP approach [9]. These tools were selected to benchmark SV calling. MetaSV performs intra- and inter-tool merging for SVs with significant overlap, and then performs local assembly at SV breakpoints as an additional line of evidence and to refine breakpoints. Finally, SVs are genotyped and annotated. SVs detected by multiple tools (or multiple lines of evidence in one tool, e.g. SR and RP or SR and RD) are considered to be high-confidence, or PASS. SVs detected by only one tool or one line of evidence, e.g. just RD, are low-confidence, or LOWQUAL. MetaSV is also augmented with a soft-clip-based method for detecting insertions. A soft-clipped read refers to the SR and its uniquely mapped mate. Candidate insertion intervals are generated from soft-clipped reads, which are then assembled to generate insertion locations.

2.2.2 VarSim Paired-End Read and SV Simulation

To validate the sensitivity and specificity of the SV calling approach, VarSim was used to simulate paired-end reads and SVs from a reference genome and validate the results of alignment and variant calling [43]. VarSim is able to simulate deletions, insertions, tandem duplications, and inversions. Translocation simulations are not currently available but are planned for a future version. VarSim inserts variants, e.g., previously reported SVs from the Database of Genomic Variants (DGV), into a user-specified reference genome. DGV is a curated catalogue of SVs from control individuals obtained using microarrays and NGS [34]. VarSim then uses ART to simulate reads in FASTQ format for secondary analysis. ART is a set of tools that

uses error models or quality profiles from real sequencing data to simulate synthetic reads [24]. After alignment and variant-calling through the user-defined pipeline, VarSim compares the called variants to the SV truth set used as input for simulating reads, breaking down sensitivity and precision by SV type and size.

Using VarSim, 30X NGS 2x100 bp paired-end reads were generated from hg19 using variants from DGV. Reads were then processed using SpeedSeq, and SVs called with Pindel, CNVnator, Manta, Breakdancer, Lumpy, and MetaSV.

2.2.3 SV Benchmarking with Biological Data: NA12878 WGS

The CEU NA12878 sample has been analyzed extensively by the Genome in a Bottle (GIAB) Consortium (`http://jimb.stanford.edu/giab`) in order to characterize its high-confidence SNPs, indels, and homozygous reference regions. Preliminary benchmark deletions and insertions have also been called [46]. Deletions were downloaded from `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/svclassify_Manuscript/Supplementary_Information/Personalis_1000_Genomes_deduplicated_deletions.bed` and insertions were downloaded from `ftp://ftptrace.ncbi.nih.nlm.gov/giab/ftp/technical/svclassify_Manuscript/Supplementary_Information/Spiral_Genetics_insertions.bed`. The deletions were called by the Personalis approach (`http://www.personalis.com/assets/files/posters/ashg2013/An_integrated_approach_for_accurate_calling.pdf`) , which is a consensus method that uses CNVnator, BreakDancer, Pindel, and BreakSeq. This set of deletions was further refined by pedigree analysis of 16 family members and PCR validation [46]. The deletion call-set also includes deletions called by the 1000 Genomes Project pilot phases, which were validated by assembly or other independent technologies such as CMA. Insertions were called using Spiral Genetics Anchored Assembly.

The NA12878 genome sequenced to 50X by the Platinum Genomes project (https://www.illumina.com/platinumgenomes.html) was downloaded from the BaseSpace Sequence Hub. FASTQ files were aligned using SpeedSeq and SVs were called using LUMPY, CNVnator, and metaSV. Deletions and insertions were extracted from the final SV call set and converted to BED format using a custom bash script. Using intersectBED, deletions were considered true positives if they shared a reciprocal overlap of 50% with a benchmark deletion. In other words, deletion A must overlap deletion B by at least a fraction of 1/2, and deletion B must overlap deletion A by

at least a fraction of 1/2. Using windowBED, insertions were considered true positives if a benchmark insertion resided within 10 bp of the insertion point.

2.3 Annotations

2.3.1 Gene Lists

For the HSAN study, candidate genes were defined from the list of 50 genes in Gene Dxs hereditary neuropathy panel (`https://www.genedx.com/test-catalog/available-tests/hereditary-neuropathy-panel/`). For the Aicardi syndrome study, all RefSeq genes (1087 with unique HUGO Gene Nomenclature Committee names) on the X chromosome were analyzed.

2.3.2 Identification of Regulatory Sequences in the Human Genome

Publicly available databases were used to identify human regulatory sequences. The FANTOM5 genome-wide, tissue- and cell-specific atlas of enhancers was downloaded from `http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed`. Vista enhancers, which have been tested for biological activity in transgenic mice, were downloaded from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/`).
miRNA, snoRNA, and miRNA binding sites were also downloaded from UCSC .The 2500 bp sequence upstream of each RefSeq gene (used as a proxy for promoters), and RefSeq 3 and 5 untranslated regions (UTRs) of genes were downloaded using the UCSC table browser (`https://genome.ucsc.edu/cgi-bin/hgTables`). Genomic regions deemed ultrasensitive to variation based on conservation and presence of transcription factor binding sites were downloaded from the supplementary information of Khurana et al [27]. TADs identified in human embryonic stem cells by HiC (combined replicates) were downloaded from the Ren lab Hi-C website at `http://chromosome.sdsc.edu/mouse/hi-c/download.html`.

2.3.3 Association of Regulatory Sequences to Known Genes

UTRs and promoters regulate the genes in which they reside or to which they are proximal. Distal elements, however, do not always target the closest gene. Andersson et al. [3] performed pairwise correlations between RNA expression from FANTOM5 enhancer elements and transcription start sites

(TSSs) to associate enhancer function with genes. Vista enhancers, Ultra-sensitive regions, and non-coding RNAs were associated to genes lying within the same TAD using intersectBED; if a regulatory region lies within the same TAD as a gene, there is evidence to suggest the two interact [37]. TAD boundaries flanking candidate gene-containing TADs were extracted using a custom python script that defines the downstream and upstream TAD boundaries for that gene (see Appendix).

2.4 Benchmarking Regulatory SNV and Indel Detection

To determine the sensitivity of the regulatory region annotations in identifying true pathogenic non-coding variants, the list of pathogenic non-coding variants identified by Smedley et al [52] was intersected with the relevant regions on hg19. Using intersectBed, 3UTR variants were intersected with RefSeq 3UTRs, 5UTR variants with RefSeq 5UTRs, promoter variants with the regions 2500 bp upstream of RefSeq genes, RNA genes with RefSeq genes, miRNA variants with the UCSC wgRNA track (snoRNAs and miRNAs), and enhancer variants with FANTOM5, Vista enhancers, and Khurana ultra-sensitive regions.

To compare the CADD and FATHMM scores for pathogenic non-coding variants versus a random set of genomic variants, FATHMM and CADD scores were computed for all pathogenic non-coding SNVs (343) and a set of 343 randomly sampled rare variants from one HSAN patient. Pathogenic non-coding indels (110) were excluded, as FATHMM does not support scores for indels. To simulate the efficacy of the SNV/indel annotation and prioritization pipeline for a heterogeneous Mendelian disorder, a variant in the promoter region of *LDLR* was selected and inserted into the vcf file of one of the HSAN patients. This particular variant (hg19 chr 19:11200073C¿T), present in the variant list compiled by Smedley et al, was chosen because it is associated with familial hypercholesterolemia (MIM 143890), a heterogeneous phenotype associated with variants in *APOA2, ITIH4, GHR, GSBS, EPHX2,* and *LDLR.*

2.5 Filtering and Annotation of SNVs and Indels

1000G, dbsnp147, esp6500, ExAC, the Haplotype Reference Consortium, and Kaviar databases, as well as refGene annotations, segmental duplications, and functional scores were downloaded from ANNOVAR using the an-

notate_variation.pl script (e.g., annotate_variation.pl -buildver hg19 -downdb -webfrom annovar exac03 humandb/). Conserved TFBS and ENCODE TFBS were downloaded from UCSC using the same script (e.g., annotate_variation.pl -buildver hg19 -downdb tfbsConsSites humandb/). Using ANNOVARvariants_reduction.pl script, variants present at a frequency of greater than 1% in 1000G, dbsnp147, esp6500, ExAC, the Haplotype Reference Consortium, and Kaviar were filtered out. Using the ANNOVAR table_annovar.pl script, rare variants were annotated with refGene annotations, segmental duplications, functional scores (CADD and FATHMM-MKL), conserved TFBS as annotated by TRANSFAC, ENCODE TFBS, regulatory regions associated with genes, and TADS containing candidate genes. Upon manual inspection of variants in IGV and UCSC, variants that were found to have rs numbers were excluded.

Indel annotation can vary by software, and therefore coordinates between known indels and variants may differ. This can result in common indels not being filtered out. Indels that appeared to be technical artefacts based on their visual presence (or the presence of similar indels) in samples from other cohorts upon inspection in IGV, were also excluded. After manual curation, FATHMM and CADD scores were used to prioritize SNVs; variants with a CADD score greater than 15 were flagged, as were variants with a FATHMM non-coding score of greater than 0.5, as these scores above these thresholds are considered to indicate functional variants. Genotype-phenotype correlation, inheritance mode, and presence of TFBS were then used to further narrow down candidate pathogenic variants.

2.6 Filtering and Annotating SVs

2.6.1 VCF to BED format

In order to compare SV intersections with common SVs and with regulatory regions or genes, SV vcf files were converted to BED format using a custom shell script where the SVLEN is extracted and added to the start coordinate to find the end coordinate of the SV. For insertions, the end coordinate was equal to the start coordinate plus one, regardless of the SVLEN.

2.6.2 Comparison to SV Control Databases

Methods for SV comparison to DGV and 1000G were modified from Hehir-Kwa *et al* 2016 [22]. A reciprocal overlap was used to match deletions,

duplications, and inversions, rather than comparison of SV length and center.

2.6.3 Comparison to DGV

For HSAN, the SV calls were compared to the 2016-05-15 hg19 release of the database of genomic variants (DGV) (`http://dgv.tcag.ca/dgv/docs/GRCh37_hg19_variants_2016-05-15.txt`). For Aicardi syndrome, the SV calls were compared to the 2016-08-31 hg38 release of DGV (`http://dgv.tcag.ca/dgv/docs/GRCh38_hg38_variants_2016-08-31`). Deletions were compared to DGV entries where varianttype was equal to CNV and variantsubtype was deletion or loss. Duplications were compared to DGV entries where varianttype was equal to CNV and variantsubtype was duplication, gain, or tandem duplication. For inversions, entries with varianttype OTHER and variantsubtype inversion were used. Insertions were compared to entries where varianttype was equal to CNV and variantsubtype was insertion or novel sequence insertion or mobile element insertion. For SVs other than insertions, variants with a reciprocal overlap of at least 80% with a DGV entry were filtered out using bedtools subtractBed. Insertions were filtered out with windowBed if a DGV variant resided within 500 bp from the insertion point.

2.6.4 Comparsion to 1000G

The SV calls for HSAN were compared to the SV release of phase 3 of the 1000 Genomes Project for hg19 (`ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/phase3/intergrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz`). Annotations for 1000G SVs aligned to hg38 are not yet available. Deletions were compared to 1000G events where SVTYPE matched DEL, CNV, DEL_ALU, DEL_HERV, DEL_LINE1, or DEL_SVA. Duplications were compared to SVTYPES matching DUP or CNV. Inversions were compared to records with SVTYPE=INV. Insertions were compared to SVTYPES ALU, LINE1, or SVA. As described above for DGV, matching was done based on overlap of the patient SV, or within a 500 bp window for insertions.

2.6.5 Translocations

Unfortunately, MetaSV does not handle the SVs annotated as BND (breakpoint) by Lumpy, which represent translocations and insertions. As such,

these annotations are filtered out of the final MetaSV VCF output. Instead, a custom python script was used to extract entries annotated as BND by Lumpy. 95% confidence intervals surrounding the breakends for IMPRECISE variants were used to determine bed start and end coordinates. Translocation bed files were then concatenated to the main SV bed file and sorted using BEDtools sort.

2.7 Identifying Genic and Regulatory SVs

Rare SVs were intersected with genes and regulatory elements using intersectBED.

2.8 HSAN Modifications to Workflow

Variants present in the HSAN patients for whom a genetic diagnosis had been made by exome sequencing, HSAN3 and HSAN4, were subtracted from the other HSAN patients. For SNVs and indels, this was done using the Annovar annotate_variation.pl script using the variant lists from HSAN3 and HSAN4 as filters. Similarly, HSAN3 and HSAN4 SVs were filtered out using subtractBED with a reciprocal overlap of 0.8, comparing like SV types.

2.9 Aicardi Syndrome Modifications to Workflow

We hypothesized that the mutation causing Aicardi syndrome was a *de novo* mutation on the X chromosome. As such, in trio probands, putative *de novo* SNVs/indels were identified by filtering out parent SNV/indels using Annovars annotate_variation.pl script. Similarly, parent SVs were filtered out using subtractBed with a reciprocal overlap of 0.8, comparing like SV types. To filter out technical artefacts, parent variants were also filtered out from all 9 Aicardi genomes.

2.9.1 Counting Variants in Common

Regulatory regions, as described in Methods 2.3.2, were selected for those residing on the X chromosome using Unix command grep chrX. A custom python script was used to count the number of patients in which chromosome X regulatory regions were mutated by SNVs/indels in the nine Aicardi

genomes and between the two trio probands for *de novo* SNVs/indels. Regulatory regions with variants in 9, 8, 7, 6, 5, or 4 patients were manually inspected in IGV and the UCSC genome browser.

Tool	Version	Command-line arguments (if not specified, default was used)
VarSim[1] Art	0.6.3 03.11.14	-read_length 100 vc_num_snp 3000000 vc_num_ins 100000 -vc_num_del 100000 vc_num_mnp 5000 vc_num_complex 5000 sv_num_ins 2000 sv_num_del 2000 sv_num_dup 200 sv_num_inv 100 sv_percent_novel 0.01 vc_percent_novel 0.01 mean_fragment_size 350 sd_fragment_size 50 vc_min_length_lim 0 vc_max_length_lim 49 sv_min_length_lim 50 sv_max_length_lim 1000000 nlanes 5 totalcoverage 30
SpeedSeq[2] BWA Samblaster Sambamba FreeBayes	0.10 0.7.15-r1140 0.1.22 0.5.9 0.9.21	-
Pindel[3]	0.2.5b8	-T 10 -w 5
CNVnator[4]	0.3.2	bin size: 100 -his 100 -stat 100 -partition 100 -calling 100
Breakdancer[5]	1.3.6	-m 1000000000 -r 2
Manta[6]	0.29.6	-m local -j 10
LUMPY[7]	0.2.13	-
MetaSV[8]	0.5.4	-min_support_ins 2 -max_ins_intervals 500000
ANNOVAR[9]	2016Feb01	variants_reduction.pl:aaf_threshold 0.01 table_annovar.pl: -protocol refGene,cadd*,fathmm*,genomicSuperDups, tfbsConsSites,wgEncodeRegTfbsClustered, wgRna,targetScanS [plus regulatory region bed files specific to cohort] Not available for hg38
BEDtools[10]	2.24.0	-

Table 2.1: Software and parameters used

[1] `https://github.com/bioinform/varsim`
[2] `https://github.com/hall-lab/speedseq`
[3] `https://github.com/genome/pindel`
[4] `https://github.com/abyzovlab/CNVnator`
[5] `https://github.com/genome/breakdancer`
[6] `https://github.com/Illumina/manta`
[7] `https://github.com/arq5x/lumpy-sv`
[8] `https://github.com/bioinform/metasv`
[9] `http://annovar.openbioinformatics.org/en/latest/`
[10] `http://quinlanlab.org/tutorials/bedtools/bedtools.html`

Chapter 3

Results

A bioinformatic workflow was constructed to analyze WGS from patients with rare Mendelian diseases. Several components of this workflow, including SV detection, regulatory variant detection, and pathogenic regulatory variant prioritization, were tested to benchmark the performance of the pipeline. Finally, the pipeline was tested on a cohort of patients with HSAN and a cohort of patients of Aicardi syndrome and identified two candidate variants of interest. The results are described in detail below.

3.1 Benchmarking Against Simulated Data

3.1.1 Comparison of SV callers

First, SV calls from metaSV based on input from Pindel, CNVnator, Breakdancer, Manta, and LUMPY were compared to the set of calls for each tool individually (fig. 3.1, deletions). metaSV calls were divided into metaSV_all and metaSV_PASS. metaSV_all represents the union of SVs called using any number of SV tools or approaches, in other words, calls made by all five SV tools and integrated by metaSV. SVs called by only one approach, for example just by RD, are considered to be low-confidence. SVs called using two or more tools/approaches are considered to be high-confidence. metaSV_PASS represents only high-confidence calls. Sensitivity is measured by the number of true positive SV calls divided by the total number of true SVs (eqn. 3.1). Precision, or positive predictive value (PPV), is measured by the proportion of true positive SV divided by all SVs called (eqn 3.2). The F1 score is the harmonic mean of PPV and sensitivity (eqn. 3.3).

$$Sensitivity = TP/(TP + FN) \tag{3.1}$$

$$PPV = TP/(TP + FP) \tag{3.2}$$

$$F1 = 2TP/(2TP + FP + FN) \tag{3.3}$$

Where TP = true positive, FN= false negative, FP=false positive

As expected, none of the individual tools was as sensitive as the set of metaSV calls together (metaSV_all). Surprisingly, LUMPY achieved F1 scores similar to the metaSV_PASS calls across all deletion sizes and types, and similar sensitivity to metaSV_all calls (Fig. 3.1). LUMPY was therefore was selected for more in-depth analysis of SV data.

3.1.2 LUMPY and CNVnator Versus All Other Callers

Figure 3.1 shows that LUMPY detected deletions as accurately as the consensus of 5 SV callers. However, because we are searching for rare disease-causing SVs of all types, it is important to consider the sensitivity of the methods in more depth. metaSV_all achieved a higher sensitivity than LUMPY for deletions (fig. 3.1). Some SVs, such as large deletions and large duplications, are missed by SR and RP methods but are identifiable by a RD approach.

I predicted that LUMPY and CNVnator together, merged by metaSV, would achieve a sensitivity equaling metaSV_all because LUMPY and CNVnator together combine SR, RP, and RD methods. While LUMPYs sensitivity for deletions was nearly equivalent to metaSV_all, it dropped noticeably for deletions greater than 500,000 bp (fig 3.1), while CNVnators sensitivity increased, indicating that SR and RP methods are not as sensitive as an RD method in detecting larger deletions. Indeed, the sensitivity of LUMPY and CNVnator together for deletions was 79% vs. 80% for metaSV_all (Fig. 3.2, Table 3.1). For duplications, LUMPY and CNVnator together were equivalent to metaSV_all at 89.4% sensitivity (Fig. 3.3, Table 3.2), and LUMPY and CNVnator together had slightly lower sensitivity for inversions at 63.5% vs. 66.7% for metaSV_all (Fig. 3.4, Table 3.3). CNVnator does not detect inversions, so this reduction must be due to some decrease in sensitivity introduced in either the merging process or local assembly step when LUMPY and CNVnator are input to metaSV. Although the addition of CNVnator calls decreased overall SV detection specificity for deletions and duplications, reflected by a decrease in the F1 score, it is important for capturing SVs that go undetected by LUMPY.

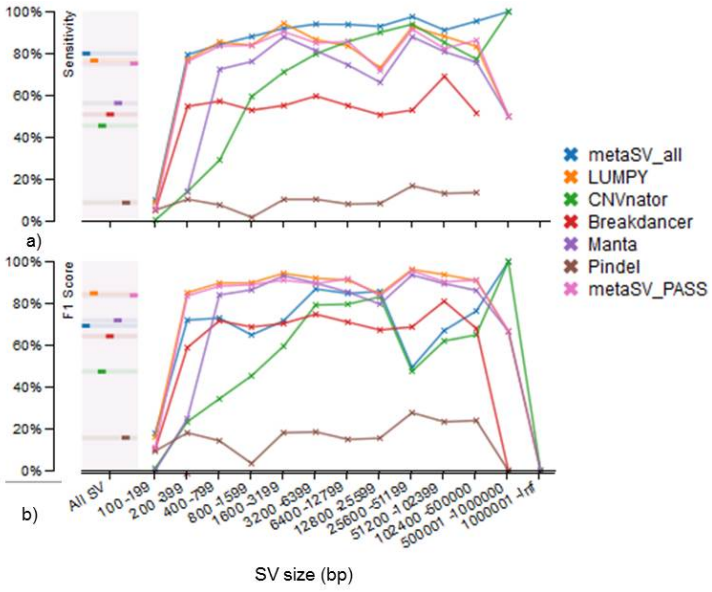


Figure 3.1: Deletion detection a) sensitivity and b) F1-score of 5 SV callers and metaSV consensus calls on 30X 100 bp paired-end simulated reads. metaSV_all represents SVs called using only one tool or approach (LOWQUAL), as well as SVs called using two or more tools/approaches. metaSV_PASS represents only SVs called using two or more tools/approaches. The lines on the rectangle on the left of the graph indicate the sensitivity or F1 score of each approach for detecting all sizes of SVs. Inf=infinity.

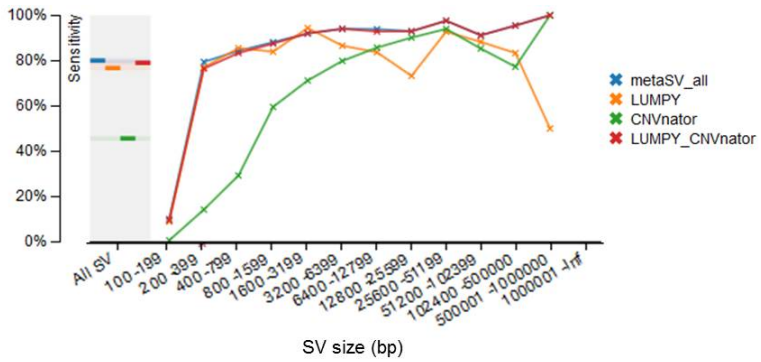


Figure 3.2: Deletion detection sensitivity of metaSV (combining 5 SV callers), LUMPY alone, CNVnator alone, and LUMPY and CNVnator together, with 30X 100 bp paired-end simulation. metaSV_all represents SVs called using only one tool or approach (LOWQUAL), as well as SVs called using two or more tools/approaches. The lines on the rectangle on the left of the graph indicate the sensitivity of each approach for detecting all sizes of SVs. Inf=infinity. Note that the lines for metaSV_all (blue) and LUMPY_CNVnator overlap.

Tool	Called	True Positives	F1	Sensitivity
MetaSV_all	2319	1418	69.3	80.0
LUMPY	1430	1359	84.9	76.7
CNVnator	1635	808	47.4	45.6
LUMPY_CNVnator	2193	1400	70.6	79.0

Table 3.1: Deletion detection sensitivity for 30X 100 bp pair-end simulation

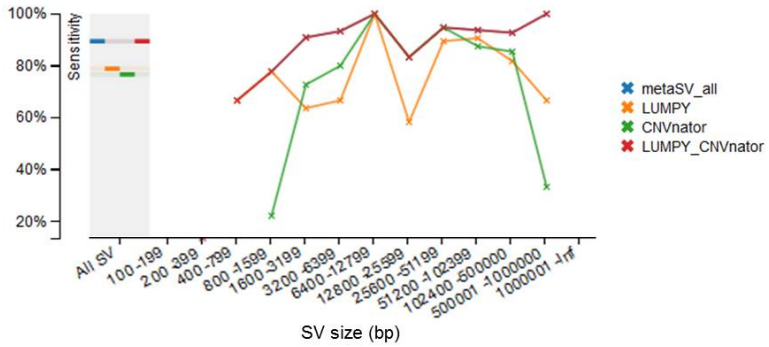


Figure 3.3: Duplication detection sensitivity of metaSV (combining 5 SV callers), LUMPY alone, CNVnator alone, and LUMPY and CNVnator together, with 30X 100 bp paired-end simulation. metaSV_all represents SVs called using only one tool or approach (LOWQUAL), as well as SVs called using two or more tools/approaches. The lines on the rectangle on the left of the graph indicate the sensitivity of each approach for detecting all sizes of SVs. Inf=infinity. Note that the lines for metaSV_all (blue) and LUMPY_CNVnator overlap.

Tool	**Called**	**True Positives**	**F1**	**Sensitivity**
MetaSV_all	762	161	34.2	89.4
LUMPY	221	142	70.8	78.9
CNVnator	702	138	31.3	76.7
LUMPY_CNVnator	592	161	41.7	89.4

Table 3.2: Duplication detection sensitivity for 30X 100 bp pair-end simulation

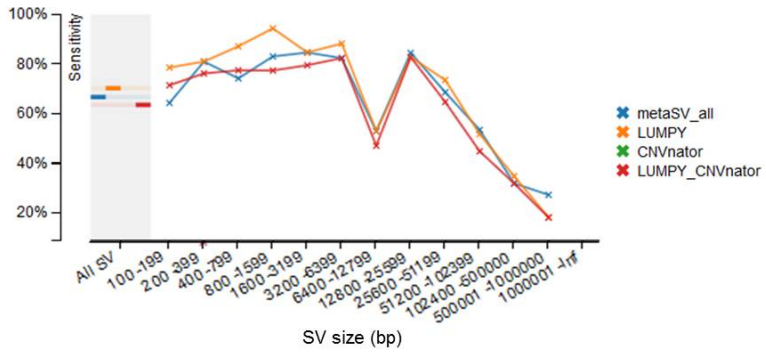


Figure 3.4: Inversion detection sensitivity of metaSV (combining 5 SV callers), LUMPY alone, CNVnator alone, and LUMPY and CNVnator together, with 30X 100 bp paired-end simulation. metaSV_all represents SVs called using only one tool or approach (LOWQUAL), as well as SVs called using two or more tools/approaches. The lines on the rectangle on the left of the graph indicate the sensitivity of each approach for detecting all sizes of SVs. Inf=infinity.

Tool	Called	True Positives	F1	Sensitivity
MetaSV_all	367	338	77.4	66.7
LUMPY	359	356	82.2	70.2
CNVnator	0	0	0	0
LUMPY_CNVnator	328	322	77.1	63.5

Table 3.3: Inversion detection sensitivity for 30X 100 bp pair-end simulation

Finally, Breakdancer, Pindel, and Manta are capable of detecting insertions but captured a mere 0.5% of simulated insertions, calling 225 insertions with only 9 true positives. Here, metaSVs insertion calling algorithm achieved a sensitivity of 10.3%, which, while low, is a 20-fold increase over the other tools. Similarly, the F1-score is ˜15 times higher for the metaSV algorithm than for metaSV using the 5 SV callers. Further investigation revealed that VarSim was incorrectly processing insertion variants from the test set. Intersecting the test set of insertions generated by metaSVs insertion boosting algorithm with the truth set of insertions using windowBED

revealed 468 true positive insertions, as opposed to the 188 as judged by VarSim. Further, intersection of LUMPY SVs where SVTYPE=BND detected an additional 596 true positives, 163 of which overlap with metaSV INS calls. LUMPY marks large intra/inter-chromosomal insertions as BND (it cannot detect large novel insertions, however, as reads will be unmapped, and does not support small insertion detection). In total, the pipeline detected 1029 out of 1829 insertion breakpoints. The sensitivity for insertion detection was therefore about 49%. When using a window of 10 bp around insertion breakpoints, the sensitivity was 56%.

Based on the results of benchmarking on simulated data, the SV calling pipeline was chosen to include LUMPY and CNVnator as input to MetaSV, with the insertion-boosting algorithm set to true. Interestingly, LUMPY performed substantionally better than Pindel in detecting deletions, with a sensitivity of 76.7% versus 8.80%. On the other hand, the authors of the metaSV paper reported sensitivities of 84.6% and 92.8% for LUMPY and Pindel respectively[42]. This is perhaps due to suboptimal parameter settings in this thesis, as default settings were used, where as the authors of metaSV used non-default settings. Further, the authors reported SV detection sensitivity for a genome with 50X coverage, rather than the 30X coverage used in this thesis.

3.2 Benchmarking Against Biological Data: WGS from NA12878

3.2.1 Deletion Detection

After performing benchmarking on simulated data to determine the most sensitive combination of tools for calling SVs, SV analysis was performed on the NA12878 50X genome from Platinum Genomes using LUMPY, CNVnator, and metaSV. Deletions and insertions called using this approach were compared to the svclassify deletion and insertion truth sets, respectively. 2,394 true positive deletions were called from a total of 2,676 truth set deletions (Table 3.4). The deletions called ranged in size from 50 to 139,619 bp, with 75% of deletions under 1000 bp in length, and 66% under 500 bp. Deletion detection sensitivity was 90% for all deletions called and 88% for high-confidence deletions, with a near doubling in F1 score from 0.43 for all deletions to 0.76 for high-confidence deletions. The average size of true positive deletions was almost three times the size of false negative deletions (1,451 bp vs. 533 bp). The modal size of the false negative deletions was

53 bp, while the modal size of the true positive deletions was 314 bp. The distribution of deletion sizes (fig 3.5) displays peaks at 300 bp and ˜6000 bp, consistent with *Alu* elements and L1/LINES, respectively.

3.2.2 Insertion Detection

The 28 true positive insertions ranged in size from 12 to 242 bp. Insertion detection sensitivity was 47% for all insertions called, and 47% for high-confidence insertions, with an increase in the F1-score from 0.03 for all insertions to 0.05 for high-confidence insertions (Table 3.4). The average size of true positive insertions was 74 bp, while the average size of false negative insertions was 100 bp.

SV Type	**Called**	**True Positives**	**Sensitivity**	**F1**
Deletion	8381	2394	89.5	43.3
High-confidence deletion	3510	2342	87.5	75.7
Insertion	2258	32	47.0	3.0
High-confidence insertion	1286	32	47.0	5.0

Table 3.4: Deletion and insertion sensitivity and F1 scores for SV calling on the 50X NA12878 genome.

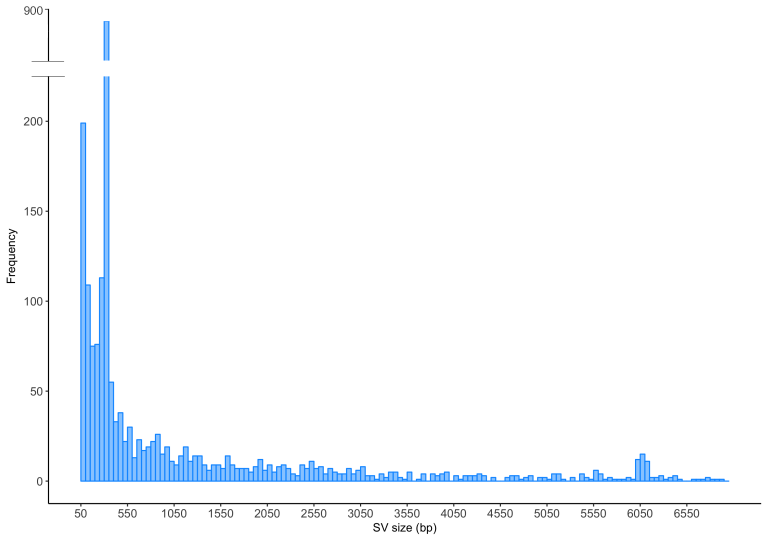


Figure 3.5: Size distribution of smaller deletions detected in NA12878. The number of deletions detected is plotted against the size of the deletion. The x-axis maximum was set to 7,000 to highlight the peaks at 300 bp and 6,000 bp. Maximum deletion size was 139,619 bp.

3.3 Genomiser Non-Coding Mendelian Variants

3.3.1 Detection of pathogenic non-coding variants

To determine if the non-coding pathogenic SNVs and indels compiled by Smedley et al. (2016) could be detected by the analysis pipeline, these variants were intersected with their respective regulatory regions. This filtering process identified all miRNA and RNA gene pathogenic variants, 97% of 5UTR variants, 93% of 3UTR variants, 87% of promoter variants, and 12% of enhancer variants (fig. 3.6).

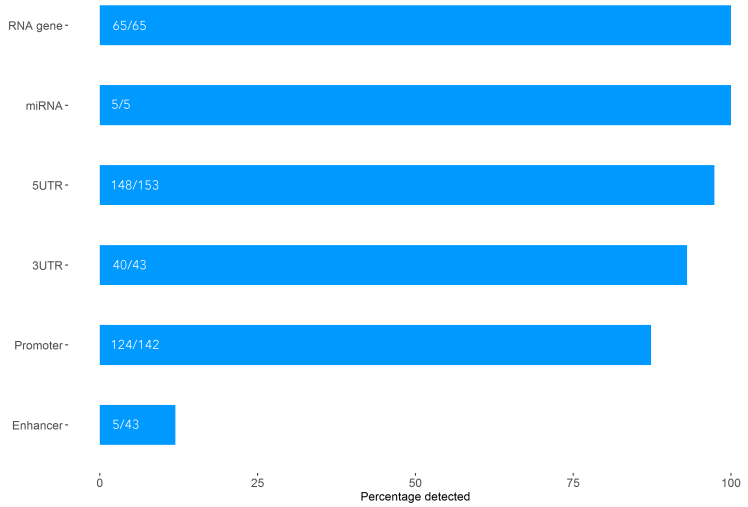


Figure 3.6: Percentage of Genomiser pathogenic non-coding variants compiled by Smedley et al. (2016) identified by the corresponding regulatory region annotations.

3.3.2 CADD and FATHMM Scores for Non-Coding Variants

Next, the CADD and FATHMM scores for 343 pathogenic non-coding SNVs from the 453 SNVs and indels from the Smedley list were compared to a set of 343 randomly sampled rare variants. The 110 pathogenic indels were not examined, as FATHMM does not support scores for indels. For pathogenic non-coding SNVs, the median FATHMM score was 0.94, while for random rare variants it was 0.1 (fig. 3.7). Pathogenic non-coding SNVs had a median CADD score of 13, while the median CADD score for random rare variants was 2 (fig. 3.8). As expected, both FATHMM and CADD score distributions for pathogenic non-coding variants and random rare variants were very significantly different (two-sample Kolmogorov-Smirnov test, p-value <2.2x10-16)

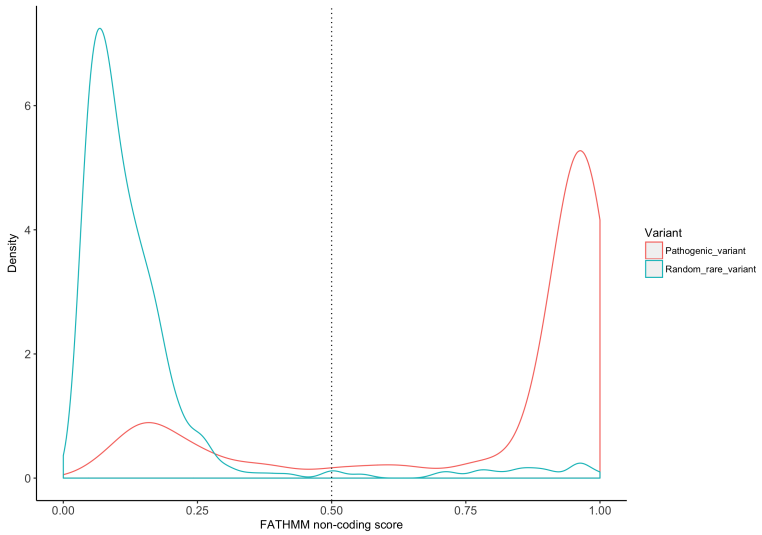


Figure 3.7: FATHMM non-coding score distributions for pathogenic non-coding SNVs compiled by Smedley et al. (2016) (turquoise) and random rare variants (pink). FATHMM scores greater than 0.50 are considered to indicate a likely functional variant.

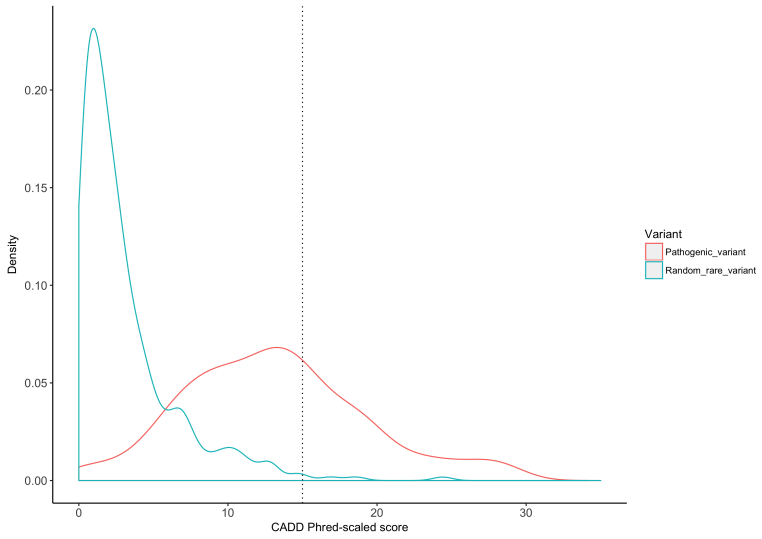


Figure 3.8: CADD score distributions for pathogenic non-coding SNVs compiled by Smedley et al. (2016) (turquoise) and random rare variants (pink). CADD scores greater than 15 are considered likely to indicate a deleterious variant.

3.3.3 Spike-in of a Pathogenic SNV

In order to determine the efficacy of the SNV/indel calling pipeline in prioritizing pathogenic variants, a variant in the promoter region of *LDLR* associated with familial hypercholesterolemia was selected and inserted into the WGS vcf file of one of the HSAN patients. After filtering for rare variants and removing variants present in the two diagnosed HSAN patients, 76,761 SNVs and indels remained in the test vcf file. Of these, only two (including the spike-in) impacted regulatory regions associated with familial hypercholesterolemia (table 3.5). Manual inspection of the variants in IGV and the UCSC genome browser revealed that the *APOA2* variant, annotated as a complex SNV, was composed of two common SNPs (rs3829793 and rs149905240). Thus, the true pathogenic variant in the *LDLR* promoter was the only remaining candidate variant. Interestingly, this variant was annotated as falling in the 5UTR and not the promoter, where it actually lies. The variant possesses pathogenic FATHMM and CADD scores and falls within a conserved V$SREBP1_02 motif bound by sterol regulatory element binding transcription factor 1. Consistency checks for correct in-

Locus(Hg19)	Ref	Alt	Regulatory region	FATHMM	CADD	TFBSCons sites
1:161194396	GTGAC	CTGAG	APOA2 promoter	.	.	.
19:11200073	C	T	LDLR 5'UTR	0.9992	15.16	SREBP1_02

Table 3.5: Regulatory variants associated with familial hypercholesterolemia genes identified in a patient genome with variant chr 19:11200073C<T spiked in.

heritance mode (autosomal dominant, in this case) and genotype-phenotype correlation would prioritize this variant as a candidate pathogenic mutation.

3.4 HSAN Analysis

Our lab performed WGS on eight patients with childhood-onset HSAN. Coding sequence analysis of these patients revealed pathogenic exonic variants in two patients [1]. The phenotypes for the undiagnosed HSAN patients are listed in table 3.6.

3.4.1 HSAN SNVs and Indels

Prior to filtering, the genome of each HSAN patient differed from the reference sequence by an average of ˜4,301,000 SNVs and indels. Filtering out variants from the two solved patients as well as common variants reduced this to ˜73,000 SNVs/indels (range 64,027-81,973). On average, 7 regulatory variants associated with hereditary neuropathy genes were found per patient prior to manual inspection in IGV, with the majority found in gene promoters, the longest of the regulatory sequences (fig 3.9).

[1] After the analysis for this thesis was completed, my colleagues discovered a pathogenic exonic variant in an additional HSAN patient, HSAN2, upon re-analysis of coding sequence variants using an improved analysis pipeline.

Symptom	**HSAN1-c1, HSAN1-c2, HSAN1-c3**	**HSAN2**	**HSAN5**	**HSAN6**
Positive intradermal histamine skin test for HSAN	✓	✓	✓	✓
Insensitivity to pain	✓	✓	✓	✓
Developmental delays	-	✓	✓	✓
Recurrent vomiting	-	-	✓	✓
Other	Anhidrosis	Mild white matter disease	Chronic lung disease from aspiration	Severe behavioural problems

Table 3.6: HSAN patient phenotypes. A checkmark indicates the presence of a symptom or test result. A minus symbol indicates the absence of a symptom or test result.

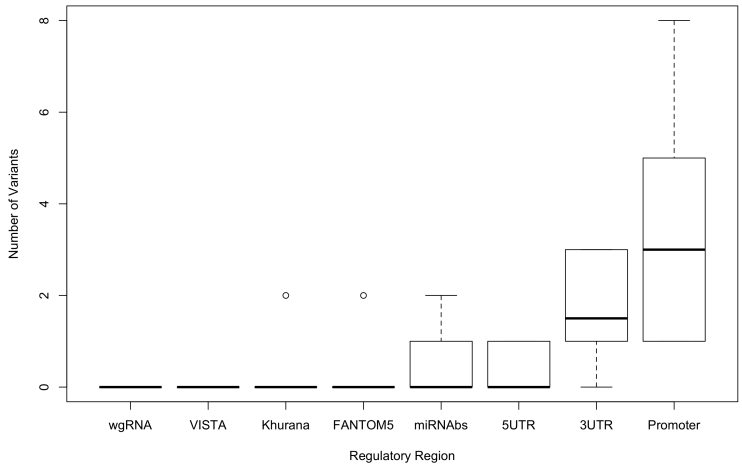


Figure 3.9: Average number of regulatory variants associated with hereditary neuropathy genes in patients with HSAN (n=6). Bold lines indicate the median number of variants in a particular regulatory region. The upper and lower hinges correspond to the 25th and 75th percentiles. Upper and lower whiskers extend the hinge to the highest value within 1.5xinterquartile range (IQR) and from the hinge to the lowest value within 1.5xIQR, respectively. The circles represent outliers, with values greater or less than the limits of the upper and lower whiskers, respectively. The regulatory variants are categorized by the regulatory region into which they fall. wgRNA=miRNA and snoRNA, VISTA=Vista enhancer, Khurana= Khurana et al (2015) ultra-sensitive region, FANTOM5=FANTOM enhancer, miRNAbas = miRNA binding site, 5UTR=5'UTR, 3UTR=3'UTR, Promoter =2500 bp upstream of transcription start site.

Only one regulatory SNV in one patient was scored as potentially pathogenic and fit the predicted inheritance pattern for the proband (*de novo* mutation). A heterozygous variant at a conserved locus (chr17:15165831, hg19) of the *peripheral myelin protein 22 (PMP22)* 5'UTR, 58 bases downstream of the transcription start site (NM_153321) was found in this patient (fig.3.10).

The variant has pathogenic FATHMM non-coding (0.92) and CADD scores (15.3). Further, it lies in HA-E2F1, GR, CTCF, Pol2, GATA-1, and E2F6_(H-50) TFBS binding sites, as annotated by ENCODE. *PMP22* is a hereditary neuropathy gene, the expression of which is critical for normal peripheral nerve myelination [39]. Duplications in *PMP22* cause Charcot-Marie-Tooth disease type 1A (CMT1A, a hereditary motor and sensory neuropathy), deletions cause Hereditary Neuropathy with Liability to Pressure Palsies (HNPP), and point mutations can cause either [39]. This particular patient is a boy with mild white matter disease, learning disabilities, and diminished pain sensation. Although the phenotype of this patient is not typical of either classical CMT1A or HNPP, *PMP22* can cause a broad range of phenotypes and is dosage-sensitive. We performed Sanger sequencing on this patient and both of his parents and found that the variant was inherited from the unaffected mother. We, therefore, concluded that it is probably not responsible for his severe phenotype.

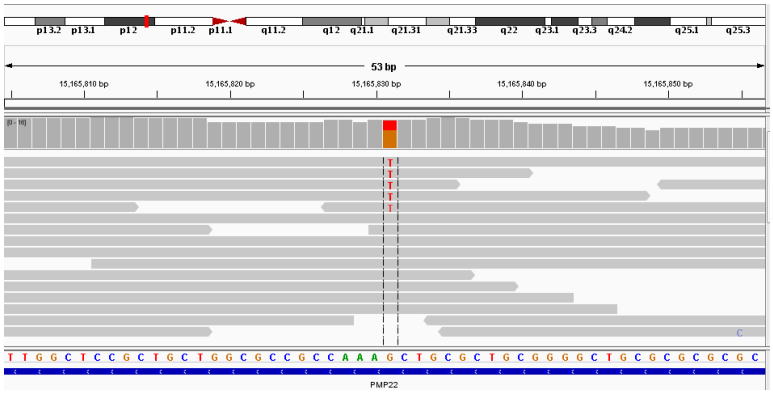


Figure 3.10: IGV screen shot of a heterozygous variant at a conserved locus of the PMP22 5UTR. The alternate allele (T) is supported by 5 reads, while the reference allele (G) is supported by 9 reads.

3.4.2 HSAN SV Analysis

On average, 7,627 SVs were detected in each HSAN patient. The majority were deletions (55%), followed by duplications (23%), insertions (21%) and inversions (1%) (fig. 3.11). 51% of these SVs were found in DGV or 1000G. Of the SVs found in DGV or 1000G, 74% were high-confidence SVs. Of the SVs not found in DGV or 1000G, 26% were high-confidence SVs. After sub-

traction of SVs in the two diagnosed patients from the other HSAN patients and intersection with hereditary neuropathy regulatory regions and genes, ˜1 high-confidence SV and ˜6 low-confidence SVs were found per patient, on average. As mentioned in the Methods, metaSV does not handle LUMPY SVs of type BND, representing translocations and insertions, so these were considered separately. 2,745 BNDs on average were found per patient. After subtraction of BNDs from the two diagnosed patients, on average ˜3 BNDs were found to intersect with hereditary neuropathy regulatory regions and genes.

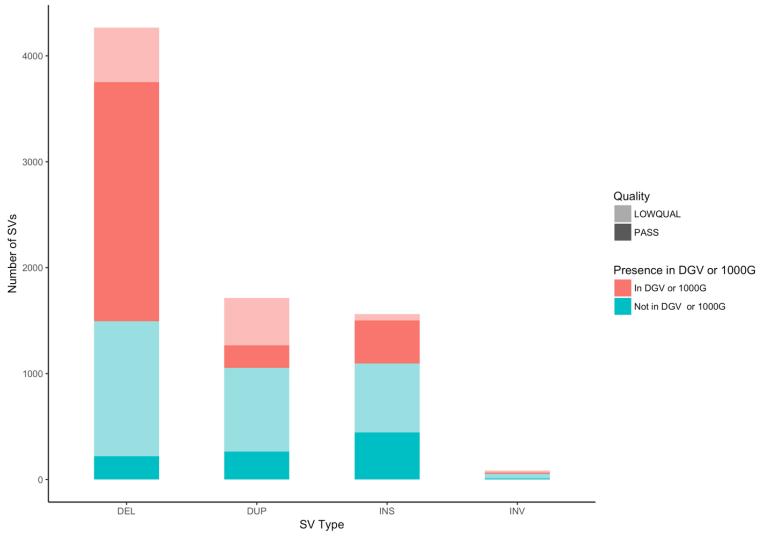
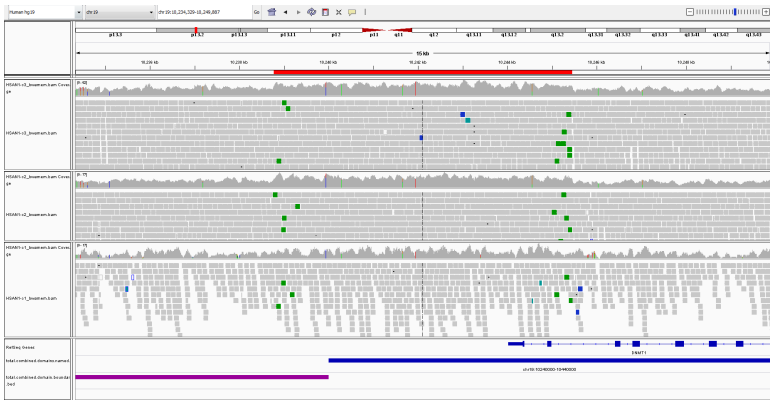


Figure 3.11: Comparison of HSAN SVs to DGV and 1000G (n=8). The SV numbers are averages. Presence in DGV or 1000G was based on whether the SV shared at least 8% reciprocal overlap with a DGV or 1000G variant of the same type. SVs are further subdivided by quality, indicated by the transparency of the bars. DEL=deletion, DUP=duplication, INS=insertion, INV=inversion. LOWQUAL=low confidence SV called by only one approach. PASS= high-confidence SV called by two or more approaches.

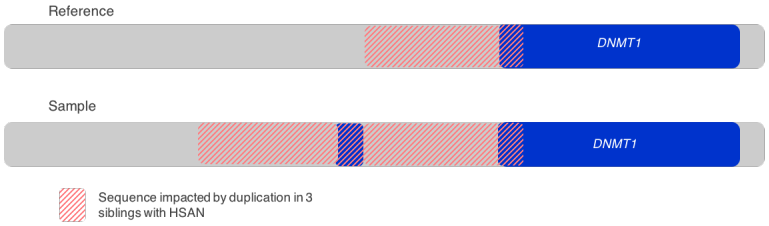
A rare, heterozygous 6,666 bp SV duplicating the last two exons of *DNA methyl transferase 1 (DNMT1)* was found in three affected siblings from one family (fig. 3.12). The duplication (chr19: 10238761- 10245460, hg19)

was called using RD, SR, and RP methods in one sibling, and RD alone in the other two. The breakpoints of this SV overlap *AluSz* SINE elements, and input of breakpoint split reads to BLAT revealed 88 bp of homology at the junction. This duplication is not present in DGV or 1000G. The region is overlapped by deletions greater than 100,000 bp in DGV, but not by duplications. In the Database of Genomic Variation and Phenotype in Humans using Ensembl Resources (DECIPHER), the region is overlapped by a number of duplications ranging in size from 3 million to 6 million bp.

Analysis of this duplication in IGV indicates that it is a tandem duplication that does not disrupt the *DNMT1* coding sequence. The reads at the boundaries of the duplication show discordant orientation; the read pairs map in an outward orientation rather than inward, which is diagnostic of a tandem duplication (green reads, fig.3.12.a.) (`http://software.broadinstitute.org/software/igv/interpreting_pair_orientations`). The duplication is not inverted; an inverted duplication would be marked by overlapping read pairs (blue and teal) at the SV boundaries. The duplicated sequence includes *DNMT1* exons 40 and 41 and is adjacent to the *DNMT1* gene, thus creating a pair of exons orphaned from *DNMT1* by the intervening duplicated non-coding sequence. A schematic of the duplication is shown in figure 3.12b. The duplication extends ˜1 kb into the TAD boundary downstream of the gene (purple bar, fig. 3.12a). While the RD, SR, and RP signals of this SV are diagnostic of a high-confidence tandem duplication, it should be noted that it has not been confirmed by Sanger sequencing. Interpretation of the pathogenicity of this variant will be considered in detail in the Discussion (4.4.2).



(a) Figure 3.12a



(b) Figure 3.12b

Figure 3.12: IGV screencap of a heterozygous 6,666 bp tandem duplication spanning the last two exons of *DNMT1* in three siblings. a) Reads at this locus in IGV for each sibling are shown. The red bar near the top of the figure represents the span of the tandem duplication. The dashed vertical line indicates the midpoint of the duplication. Reads are colored by pair orientation: green reads indicate reads whose mate is mapped in the opposite orientation to that expected. Blue and teal reads indicate read pairs whose mates overlap. The duplication is evidenced by the increased RD relative to the adjacent regions, and the green discordant read pairs at the boundaries of the duplication. The blue bar at the bottom of the figure represents a TAD. The TAD boundary is illustrated in purple. b) Linear depiction of the *DNMT1* tandem duplication (not to scale). *DNMT1* is illustrated in blue. The sequence in the reference genome that is duplicated in the siblings is depicted by diagonal pink lines.

3.5 Aicardi Syndrome Analysis

3.5.1 Aicardi Syndrome SNV and Indel Analysis

Prior to filtering, each Aicardi syndrome patient had an average of 4,525,000 SNVs and indels in comparison to the reference sequence. Filtering out variants from parents as well as common variants reduced this to 102,000 SNVs/indels per patient on average (range 75,846-126,700). Interestingly, the number of *de novo* variants identified in the two trio probands differed substantially: 170,977 in one patient (6,948 on the X chromosome) and 3,545 (135 on the X chromosome) in the other. This is likely due to the fact the high coverage in the former patient (100X) identifies many inherited SNVs and indels that are not picked up in the parents due to insufficient coverage (30X). A mean of 91 regulatory variants was found on the X chromosome in each of the Aicardi patients, with the majority falling in promoters (fig 3.13).

3.5.2 Aicardi Syndrome SV Analysis

On average, 800 X chromosome SVs were detected in each patient with Aicardi syndrome (fig. 3.1.4). 42% of these SVs were found in DGV. Of the SVs found in DGV, 44% were high-confidence SVs. Of the SVs not found in DG V, 6% were high-confidence SVs. After subtraction of SVs from the parent genomes and intersection with X chromosome regulatory regions and genes, 269 SVs were found per patient, on average. 428 X chromosome BNDs on average were found per patient. After subtraction of BNDs present in the four parent genomes, on average ˜80 X chromosome BNDs were found to intersect with X chromosome regulatory regions and genes. The number of inversions in one patient genome was dramatically higher than the number of inversions in all other genomes (2,112 X chromosome inversions vs. an average of 54 standard deviation=60, in the other patients). Further, the number of insertions in this patient was zero. This patient was therefore excluded from SV analyses. The high number of inversions was reflected by a high proportion of blue and teal reads (indicating overlapping mate pairs) evident in IGV throughout the entire genome of this patient (indicating overlapping mate pairs). This was also evident in the genome of this patient aligned to hg19 from FASTQ by my colleagues, using a different pipeline. It is unclear what the origin of these discordant read pairs is. It may be a technical artifact originating from the sequencing in this patient.

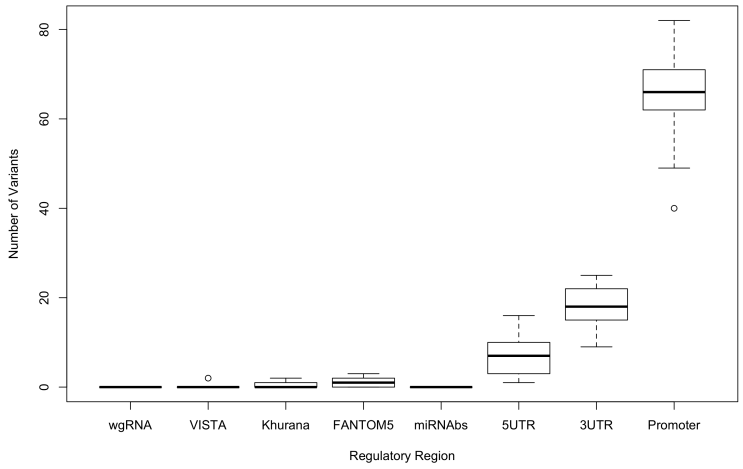


Figure 3.13: Average number of regulatory variants associated with X chromosome genes in patients with Aicardi syndrome (n=9). Bold lines indicate the median number of variants in a particular regulatory region. The upper and lower hinges correspond to the 25th and 75th percentiles. Upper and lower whiskers extend the hinge to the highest value within 1.5xinterquartile range (IQR) and from the hinge to the lowest value within 1.5xIQR, respectively. The circles represent outliers, with values greater or less than the limits of the upper and lower whiskers, respectively. The regulatory variants are categorized by the regulatory region into which they fall. wgRNA=miRNA and snoRNA, VISTA=Vista enhancer, Khurana= Khurana et al (2015) ultra-sensitive region, FANTOM5=FANTOM enhancer, miRNAbas = miRNA binding site, 5UTR=5'UTR, 3UTR=3'UTR, Promoter =2500 bp upstream of transcription start site.

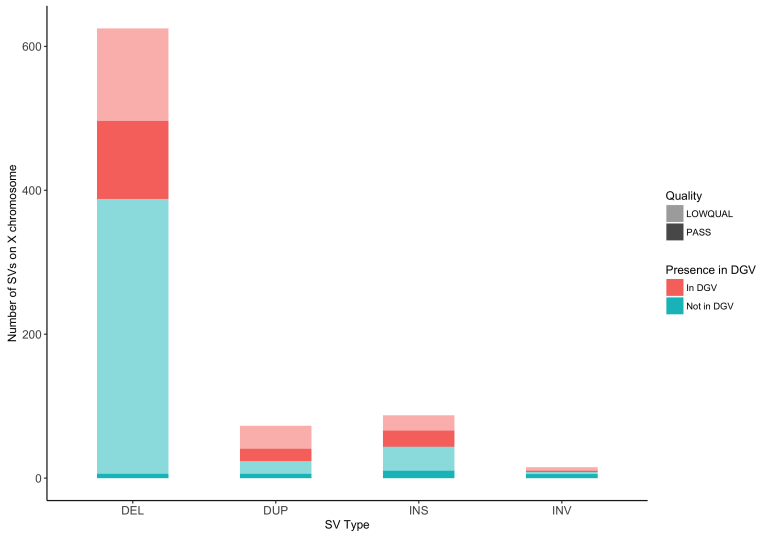


Figure 3.14: Comparison of Aicardi X chromosome SVs to DGV (n=9). The SV numbers are averages. Presence in DGV was based on whether the SV shared at least 80% reciprocal overlap with a DGV variant of the same type. SVs are further subdivided by quality, indicated by the transparency of the bars. DEL=deletion, DUP=duplication, INS=insertion, INV=inversion. LOWQUAL=low confidence SV called by only one approach. PASS= high-confidence SV called by two or more approaches.

3.5.3 Aicardi syndrome *de novo* variant analysis

De novo regulatory SNVs, indels, and SVs on the X chromosome were identified in the two trio probands. While there were several regulatory regions harboring putative de novo SNVs or indels in the trio probands, inspection of these variants in IGV indicated they were false positives. Thus, no candidate *de novo* regulatory SNVs or indels were identified that overlapped in the two trio probands or with any of the other 7 Aicardi syndrome genomes. *MID1, IL1RAPL2*, and *TFDP3* were found to be overlapped by rare *de novo* putative SVs in the two probands. All are putative *de novo* low quality duplications called by CNVnator. Manual inspection of these variants in IGV did not support their existence; the RD distributions in the probands looked very similar to that of the parents, despite these duplications not

being called in the parents by CNVnator. The putative pathogenicity of these variants is discussed further in the Discussion (4.5.1).

Chapter 4

Discussion

4.1 Overview

The aim of this thesis was to identify non-coding variants and SVs from WGS in patients with rare Mendelian diseases. More specifically, the objectives were to 1) develop and benchmark a bioinformatics workflow for detection of pathogenic non-coding SNVs/indels and pathogenic non-coding or coding SVs, and 2) to use this workflow to analyze WGS data of unsolved patients recruited from in-house HSAN and Aicardi syndrome studies to identify candidate pathogenic variants.

A bioinformatic workflow was constructed to identify putative functional regulatory variants from raw sequence data. SV calling was benchmarked against SV truth sets from a simulated genome and the NA12878 genome. A compilation of CRE annotations was selected to filter for functional variants and permit comparison to known pathogenic non-coding variants. Finally, the workflow was applied to HSAN and Aicardi syndrome patient genomes. The workflow successfully detected and prioritized rare regulatory variants and SVs. Several interesting candidate variants were detected, but none could be convincingly implicated as pathogenic in these patients.

4.2 Summary of Findings

4.2.1 VarSim Benchmarking Results and Limitations

Five SV callers (Pindel, CNVnator, Breakdancer, Manta, and LUMPY) and a consensus SV caller (metaSV) were used to call variants on a simulated genome containing known deletions, duplications, inversions, and insertions. LUMPYs F1 score for deletions, duplications, and inversions was comparable to the consensus set of high-confidence calls generated by metaSV from the input of all five callers. LUMPY, a SR and RP approach, and CNVnator, a RD approach, together with metaSV, a consensus caller that performs local assembly of candidate regions, had a sensitivity equivalent to the combination of LUMPY, CNVnator, Pindel, Breakdancer, Manta, and metaSV

together. Although insertion detection appeared to be poor, this seems largely to reflect errors in varSims comparison method between truth and test SVs of this type. Altogether, LUMPY, CNVnator, and metaSV had a deletion detection sensitivity of 80%, a tandem duplication detection sensitivity of 89%, an inversion detection sensitivity of 64%, and an insertion detection sensitivity of 49%.

Simulated genomes, however, do not accurately reproduce artefacts such as chimeric molecules and reads from poorly assembled genomic regions that can confound SV calling in biological genomes. The SV calling approach was therefore evaluated on the NA12878 genome, an extremely well-studied genome in which SVs have been characterized.

4.2.2 NA12878 Benchmarking Results and Limitations

SVs were called from the NA12878 genome by analyzing a 50X coverage data set from Illumina Platinum Genomes. SVs were called using LUMPY, CNVnator, and metaSV. Based on comparison to truth SV calls from svclassify, the estimated deletion detection sensitivity was 90% with an F1 score of 0.43. The modal size of false negative deletions was 53 bp, while the modal size of true positive deletions was 314 bp, indicating that the SV pipeline is limited in its ability to call small deletions. Insertion detection sensitivity was 47% for all insertions called and also for high-confidence insertions alone, indicating that only high-confidence insertions were true positives. These results were consistent with the results from benchmarking against simulated data. The improvement in deletion detection sensitivity is most likely attributable to the fact that the NA12878 genome was sequenced to 50X, while the simulated genome had a coverage of 30X.

Analyses of these gold standard SVs is subject to ascertainment bias. The SVs in the deletion truth set were called from PE short-read sequencing (SRS) data using SR, RD, and RP methods. Indeed, CNVnator was one of the tools used to call deletions in the Personalis and 1000G set, so this truth set is biased towards CNVnator, which was one of the tools used in this thesis. Insertions in the truth set were called using an AS approach from SRS data, which are limited in detecting long insertions. Indeed, the maximum insertion size in the truth set was only 353 bp. It is likely that larger insertions do exist in NA12878 and that the SV truth sets themselves were biased towards SVs that are detectable by the approaches I used in my research. Therefore, the sensitivity and F1 scores obtained in this study are probably overestimates of the true values. The limitations to SV detection from SRS are discussed in more detail below.

4.2.3 Limitations to SV Calling from SRS

The set of SVs detectable with SRS approaches is limited as reads are generally shorter than most SVs (50-400 bp). Indeed, the human genome consists of 50-69% repetitive sequences, patterns of DNA sequence that occur in multiple copies of the genome[12] and 5% of the genome cannot be uniquely mapped with 100 bp read length [21]. This repetitive sequence is composed of transposable elements, low complexity regions, and pseuodogenes. These regions present a challenge for SRS reference genome-based alignment, which is the method of choice in clinical sequencing studies due to its cost effectiveness and low per-base error rate .

Long-read sequencing (LRS) can overcome the SRS alignment challenges by spanning SVs and repetitive regions by several kilobases or more. Indeed, in one study characterizing SVs in a personal genome using a combination of SRS and Pacific Biosciences(PacBio) LRS, SRS approaches to SV calling identified only 57% of SVs identified using the long-read approaches [16]. PacBio single molecule real-time (SMRT) sequencing offers read lengths of 10 kb on average, with some reads close to 100 kb [10]. PacBio reads can be used to assemble whole genomes *de novo* or scaffold assembly from SRS, thereby improving completeness and considerably improving SV detection.

Recently, PacBio LRS data derived from two functionally haploid genomes was analyzed to identify SVs [25]. The haploid genomes were obtained from hydatidiform moles. Hydatidiform moles are abnormalities of human pregnancy that form from fertilization by sperm of an enucleated egg or by loss of maternal chromosomes post-fertilization [26]. Some hydatidiform moles are diploid due to subsequent duplication without cytokinesis of the fertilizing sperm; functionally, they can be considered haploid as they lack allelic variation, as is true based on the analyses these authors performed. ˜20,500 SVs were identified from each genome (˜13,000 insertions, ˜7,500 deletions, 47 inversions). Half of the inserted or deleted sequences consisted of tandem repeats or complex arrays of different repeat classes. The authors also created a pseudo-diploid genome by down-sampling the genomes of the two individuals and combining them. Interestingly, the study found that the sensitivity of SV detection from this pseudo-diploid genome was less than half that in either haploid genome due to difficulties in detecting heterozygous SVs, regardless of coverage. 83% of SVs reported in the study had not been described in previous SV studies, including 1000G. Of particular relevance to my study, analysis of SRS data from the two haploid genomes using LUMPY and WHAM, SR and RP methods, respectively, identified only 10% of variants identified by LRS technology. Clearly, SV detection from SRS

is seriously limited by short read length, especially in repetitive regions of the genome. Thus, benchmarking with GIAB gold standard deletion and insertion SV call-sets is biased towards SVs detectable by SRS.

Unfortunately, GIAB gold standard sets of duplications, inversions, and translocations are not available for NA12878. This is likely due to the computational difficulties in detecting these SV types and the difficulty in obtaining orthogonal validation. As a consequence, many SV tools are only benchmarked against deletions and/or insertions. LUMPY, for example, is benchmarked against deletions, duplications, inversions and translocations from simulated data but only against deletions from NA12878 [30]. Indeed, that LUMPY was trained on the NA12878 deletions biases its performance in the analysis performed in this thesis.

4.3 Genomiser Non-Coding Mendelian Variants

4.3.1 Detection of Pathogenic Non-Coding Variants

As expected, detection of pathogenic non-coding SNVs and indels was high for non-coding genic or proximal regulatory non-coding variants, with detection rates ranging from 87-100%. Variants missed by RefSeq UTR and promoter annotations would be picked up by Ensembl gene predictions, which are more comprehensive. This was apparent by looking at the Ensembl gene prediction track in the UCSC genome browser.

Pathogenic enhancer variant detection by intersection with FANTOM5 enhancers, Vista enhancers, and ultra-sensitive regions was poor, at 12%. Enhancers have been predicted using a range of methods, including enhancer RNA expression, EP300 binding sites, RNA polymerase II binding sites, DNase I hypersensitivity sites, and histone modification patterns, but there is little consistency in enhancer predictions based on different technologies [19]. A consensus set of enhancers integrating annotations should be more comprehensive than enhancers predicted using any one technology, like the FANTOM5 or VISTA sets. Until such a comprehensive truth set of enhancers exists, detection of enhancer variants will be limited in sensitivity, as reflected in the benchmarking performed here. Massively parallel functional assays of enhancers will also contribute knowledge of enhancers with biological function [4].

4.3.2 FATHMM and CADD Scores of Pathogenic Non-Coding Variants

With respect to categorization of non-coding variants as pathogenic or not, a FATHMM score cutoff of 0.5 categorizes 80% of true positive variants as pathogenic while scoring 5% of random rare variants as false positives. A CADD score cutoff of 15 categorizes 35% of true positive variants as pathogenic while scoring 1% of random rare variants as false positives. To categorize 80% of true positive variants as pathogenic, a CADD score of 8 would have to be used; this would identify 9% of random rare variants as false positives. Utilizing the FATHMM non-coding score identifies fewer false positives than the CADD score with an equivalent rate of true positives, making it a more reliable indicator of the pathogenicity of non-coding variants.

4.3.3 Spike-in of a Pathogenic Non-Coding SNV

One of the difficulties in analyzing non-coding variants is their sheer number. Even after filtering for rare variants, an average of 64027-126700 SNVs and indels remained in each of our HSAN and Aicardi syndrome patients. The efficacy in reducing this number by filtering for non-coding variants in candidate gene-associated regulatory regions was demonstrated by spiking a known pathogenic variant in the *LDLR* promoter into a patient variant file. Filtering for rare regulatory variants associated with familial hypercholesterolemia genes detected only the true pathogenic variant and one additional variant that was excluded after manual inspection. The spiked-in variant was therefore successfully prioritized as the best candidate pathogenic variant. That this was the only true rare regulatory variant associated with familial hypercholesterolemia in this genome indicates that the pipeline described in this thesis identifies a limited number of rare regulatory variants associated with each gene. Indeed, this was seen in the HSAN genomes, where an average of 7 rare regulatory variants associated with a list of 50 candidate genes were identified per patient. In Aicardi syndrome patients, an average of 91 rare regulatory variants associated with 1087 genes on the X chromosomes were identified. The number of rare regulatory variants identified is, of course, limited by the annotations used in the pipeline. The true number of such variants is likely larger.

4.4 HSAN Analysis

4.4.1 HSAN SNV and Indel Analysis

A heterozygous SNV identified in the *PMP22* 5UTR of one HSAN patient was a good candidate based on FATHMM and CADD scores and its presence within TFBSs. Indeed, a previous study explored the potential contribution of variants in the highly conserved *PMP22*5 region to gene expression, concluding that rare variation in this region may alter *PMP22* dosage and contribute to the clinical variability of CMT1A and HNPP [51]. The genotype-phenotype correlation for this patient was poor, but regulatory variants are not necessarily expected to recapitulate phenotypes caused by coding variants. For instance, coding variants in *PTF1A* cause syndromic pancreatic agenesis with neurological symptoms, while *PTF1A* enhancer mutations cause an isolated pancreatic anomaly [49]. However, Sanger sequencing of the proband and parents revealed that the variant was inherited from the unaffected mother and was therefore likely to be benign. This emphasizes the importance of a trio study design in sequencing studies, which can reduce the number of candidates by factoring in inheritance mode when the parents phenotypes are known. Unfortunately, funds for WGS were limited to sequencing only the probands in this cohort.

4.4.2 HSAN SV Analysis

An average of 7,627 SVs was detected in each of the WGS data sets from HSAN patients. Only 50% of these were high-confidence calls. Of the SVs found in DGV or 1000G (51%), 76% were high-confidence calls, while of the SVs not found in DGV or 1000G, only 26% were high confidence calls. This indicates that the set of rare SVs has a higher false positive rate than those that intersect with common SVs. The frequency of rare structural variation identified here is therefore inflated by false positive low-confidence calls. With this in mind, after subtracting SVs present in the two diagnosed HSAN patients and intersecting variants with regulatory regions, an average of only 10 putative rare genic and regulatory SVs were identified in each patient. Manual inspection and genotype-phenotype correlation narrowed this down to one candidate SV in the three affected siblings of one family.

A rare, heterozygous 6,666 bp tandem duplication affecting the last two exons of *DNMT1* was identified in these three siblings. Autosomal dominant SNVs in the targeting sequence domain of *DNMT1* cause HSAN1E. While these siblings all have mild HSAN phenotypes, with some insensitivity to pain and mild anhidrosis, HSAN1E is characterized by hearing loss,

dementia, and sensory loss. It is typically adult-onset. The possibility that these siblings may go on to develop further symptoms, such as hearing loss, in the future cannot be excluded.

Interestingly, a recent case report describes a *DNMT1* SNV in a patient with childhood-onset HSAN1E and some phenotypic similarities to these siblings: intermittent shooting pain in feet in childhood, repeated infections, and insensitivity to pain [18]. The patient developed deafness in adulthood. In spite of these parallels, the phenotype of the siblings we studied is not typical of HSAN1E, as confirmed by the clinical expert who phenotyped them.

The molecular pathology of HSAN1E is likely due to reductions in *DNMT1* enzymatic activity resulting from mis-folding of the *DNMT1* protein [54]. The duplication I found, on the other hand, is not predicted to disrupt the protein. However, the orphaned exons may be translated, if alternative splicing occurs between the full gene and the orphaned exons, which could be could be further elucidated with RNA-seq. If the gene is mistranslated, protein folding could be impaired. The SV does extend ˜1kb into the adjacent TAD boundary, which is 240 kb in length, however disease-causing TAD disruptions described in the literature fully delete, invert, or duplicate a TAD boundary. It is therefore unlikely that this small duplication interferes with *DNMT1* regulation.

That the phenotypes of the siblings are not a close match to documented HSAN1E cases and that neither the gene nor gene regulation are confidently predicted to be affected makes it difficult to assess the pathogenicity of this SV. Similarly, the mode of inheritance of HSAN in these patients is unknown. One might expect an autosomal recessive mode of inheritance given that all three siblings are affected, but it is possible that the mode of inheritance is autosomal dominant with one parent affected. These siblings are adopted, and little information about the biological parents is available, complicating the interpretation of this SV. The duplication of *DNMT1* I found must, therefore, be classified as a variant of uncertain significance (VUS) until new clinical evidence arises to support the likelihood that the SV is either benign or pathogenic. For instance, if symptoms typical of HSAN1E develop in these three siblings in adulthood, this would provide evidence for the pathogenicity of the duplication. The presence of this duplication in similarly affected patients would also support pathogenicity. Discovery of this SV in unaffected individuals in population databases would support the hypothesis that it is benign.

4.4.3 HSAN Analysis Limitations

Detection of regulatory variants associated with HSAN is limited by both the candidate gene list and the regulatory region annotations. This pipeline will not detect variants in novel HSAN disease genes, nor will it detect variants in as-yet unannotated regulatory regions. Furthermore, lower-than-expected coverage (<30X) was obtained in two of the HSAN patients included in this study, limiting variant detection. It is, therefore, possible that undiagnosed patients possess variants in genes not on the candidate gene list or in regulatory regions that are not currently annotated. It is also possible that pathogenic variants are present in these data sets that were not called due to insufficient coverage. Lastly, the possibility that these patients are phenocopies, in other words that the origin of their disease is environmental and not genetic, cannot be excluded.

To search for regulatory variants in additional genes related to HSAN, a tool like Phenolyzer could be used to compile a list of genes related to known HSAN genes by protein-protein interactions, sharing a gene family or biological pathway, or transcriptionally regulating another gene [58]. If more funds become available, patients with low coverage should be re-sequenced, and if possible, all of the parents should be sequenced by WGS.

4.5 Aicardi Syndrome Analysis

4.5.1 Aicardi Syndrome SNV/Indel and SV Analysis

Non-coding and SV analysis of Aicardi syndrome genomes failed to reveal a candidate *de novo* X chromosome variant in these patients. CNVnator called *de novo* duplications of *IL1RAPL2, MID1*, and *TFDP3* in both trio probands. *IL1RAPL2* is about 1 mbp and is an orphan interleukin receptor that was identified in fetal brain tissue [48]. *MID1* is about 400 kb and is a microtubule-associated protein; *MID1* mutations cause Opitz syndrome, a midline malformation syndrome (OMIM:300000). *TFDP3* is less than 2000 bp in length and is a transcription factor that suppresses E2F1-induced apoptosis-dependent P53. It is ubiquitously expressed in human tissues, including brain [33]. Interestingly, the two probands have putative duplications that overlap a small region (2,600 bp) that encompasses *TFDP3*. Given that *TFDP3* is only 1,679 bp long, it is far less likely for the two probands to both have SVs in this gene by chance than it is in *MID1* or *IL1RAPL2*. However, all these SVs are low-confidence, as they have been only called by CNVnator, which has an F1 score for duplications of only

47.4% as measured by VarSim. Further, visual inspection in IGV did not support the presence of these duplications, as the read depth distributions of the probands appeared very similar to that of their parents. It is therefore very likely that these are false positives. Furthermore, these SVs are not found in any of the other probands, making it unlikely that they cause Aicardi syndrome, if it is indeed a genetically homogenous condition.

4.5.2 Aicardi Syndrome Limitations

There are several reasons that a causative mutation might not have been identified in the Aicardi syndrome patients. First, it is possible that a causative variant is located in a genomic region that is invisible to most alignment methods, such as a segmental duplication, tandem repeat or other poorly mapped genomic region. Second, it is possible that the causative mutation is an SV that cannot be detected from SRS data due to its length or repetitive content. Third, the variant may be in a regulatory region that is not annotated, or insufficiently annotated, within the pipeline. Fourth, it is possible that the causative mutation is somatic, and the read frequency of the variant allele was too low to be detectable in the samples studied. Indeed, some clinical features of Aicardi syndrome are patchy, which can be indicative of mosaicism. To test this hypothesis, we could sequence affected tissue from additional patients at high coverage, e.g. 100X, and compare this to blood WGS from the same patients. Lastly, the chorioretinal lacunae, or holes in the retina, and agenesis of the corpus callosum suggest that cells carrying the causative variant die. In this case, a mosaic mutation would not be detectable by sequencing DNA from viable tissue.

4.6 Conclusions and Future Directions

4.6.1 Summary

This thesis benchmarked and tested a bioinformatic workflow for identifying pathogenic regulatory variants and SVs from WGS in rare Mendelian disease. While typical sequencing pipelines analyze SNVs and indels in the exonic regions of the genome, this workflow extends WGS analysis to cover the full spectrum of genetic variation. The SV calling pipeline, validated against a simulated genome, detected 80% of deletions, 89% of tandem duplications, 64% of inversions, and ˜50% of insertions. On experimental data, the pipeline detected 90% of deletions and 47% of insertions. The SV truth sets used for benchmarking likely only represent about 10% of structural

variation in the genome due to difficulties in calling SVs in repetitive and GC rich genomic regions from SRS data. Nevertheless, use of the analysis pipeline I developed will detect many pathogenic SVs and increase diagnostic yield from clinical sequencing studies.

Extending the analysis to non-coding regulatory regions identified only 0.14 variants per candidate gene in the HSAN study and 0.08 variants per X-chromosome gene in the Aicardi syndrome study. This number of variants is manageable for a bioinformatician to analyze manually. In a heterogeneous disorder like intellectual disability, which is associated with over 1000 different genes, the pipeline would identify more than 100 regulatory variants. In this case, a FATHMM threshold of 0.5 would filter out 95% of rare SNVs, with a sensitivity of ˜80% for detecting pathogenic SNVs. Even a FATHMM threshold of 0.2, with a sensitivity of 90%, would only filter out 84% of rare SNVs. Manual inspection of indels, on the other hand, revealed many to be sequencing or algorithmic artefacts. A large database of in-house whole genome sequences, all processed through the same pipeline, would filter many of these technical artifacts out.

4.6.2 Comparison to a study analyzing WGS from patients with a heterogeneous disease

Given that only six HSAN patient genomes were studied and that, given its extreme rarity and phenotypic homogeneity, Aicardi syndrome is likely caused by pathogenic variants of just one gene [2], it is unsurprising that no pathogenic non-coding variants or SVs were discovered. Relevant to this thesis is a WGS study of 722 individuals with inherited retinal disease, a heterogeneous disorder, 537 pathogenic alleles were identified in 404 individuals [7]. This equated to a diagnostic rate of 56%. Of the pathogenic alleles, only 31 were deletions, 2 were tandem duplications, and 3 were SNVs in regulatory regions. In other words, 4% of individuals in the cohort possessed pathogenic deletions, 0.3% possessed pathogenic tandem duplications, and 0.4% possessed synonymous or regulatory region SNVs/indels. If we assume that the prevalence of pathogenic SVs and non-coding SNVs in inherited retinal disease is generalizable to other heterogeneous disorders, SRS would reveal one patient in 25 with a pathogenic deletion, 1 patient in 333 with a pathogenic tandem duplication, and one patient in 240 with a pathogenic non-coding mutation. This study of retinal disease only looked at non-coding mutations in introns of candidate genes and SVs disrupting exons of candidate genes, and so the rate of discovery of these mutations might be increased by expanding the search space. As with the limitations to the

HSAN and Aicardi syndrome studies, the unsolved portion of this cohort may have pathogenic variants in repetitive regions, regions of poor coverage, or genes not on the list of inherited retinal-disease associated genes. Further, the inheritance may be oligogenic or influenced by environmental factors. In spite of these limitations to whole genome SRS analysis and a candidate-gene based approach, this study demonstrated the value of WGS in identifying pathogenic non-coding variants and SVs.

4.6.3 Future directions

This thesis described the limitations to SV identification from SRS. LRS is expensive, but there is an affordable alternative: 10X GemCode Technology is a library preparation and analysis method that leverages droplet microfluidics and molecular barcoding to construct 40-200 kb linked reads (pseudo-long reads) from short reads. Short reads from many long DNA molecules are each tagged with molecular barcodes unique to the long fragment of origin, giving the ability to link distant segments into a single contig (`https://www.10xgenomics.com/`). Like PacBio SMRT sequencing, pseudo-long read sequencing allows assembly of repetitive regions but is more affordable. Given the added cost of these technologies, a practical approach for clinical diagnosis might be to perform LRS only after all other analyses are exhausted. We have obtained funding to perform 10X GemCode library preparation and Illumina sequencing on an Aicardi syndrome patient. This may reveal variants undetected in Aicardi syndrome patients by SRS.

Several months after designing the pipeline described in this thesis, a paper was published describing a tool, Genomiser, for identifying pathogenic SNVs and indels in Mendelian disease [52]). The paper presents a regulatory Mendelian mutation (REMM) score for prioritizing variants. The score is based on machine learning from a set of 453 known pathogenic non-coding variants and is claimed to be superior to FATHMM and CADD scores in prioritizing pathogenic non-coding variants. Genomiser harnesses TAD boundaries and FANTOM5 enhancers in prioritizing variants, as well as patient phenotypes. However, Genomiser does not handle SVs, as it is limited to input from vcf files containing SNVs and indels. In the future, it may be enlightening to test Genomiser on our unsolved genomes. In addition, a novel reference-free k-mer based algorithm, RUFUS, is in development for *de novo* mutation detection from trios and quartets (`https://github.com/jandrewrfarrell/RUFUS`). We are in the process of testing this software on Aicardi syndrome trios.

4.6.4 Conclusions

The bioinformatic workflow described in this paper is complementary to sequencing pipelines that analyze only protein-coding variants from whole genomes. Benchmarking against simulated and real whole genome data, as well as known pathogenic SNVs and indels, validated its utility in detecting variants across the entire spectrum of genetic variation. Application of this workflow to larger cohorts of patients with rare Mendelian diseases should identify pathogenic non-coding variants and SVs, increasing diagnostic yield of clinical sequencing studies, assisting management of genetic diseases, and contributing knowledge of novel pathogenic variants to the scientific community.

Appendix A

Python script

```python
# Author: Madeline Couse
import csv
from sys import argv


#This script takes a as input a bed file of candidate genes with four columns: chromosome, start, stop, and gene name. The script outputs the upstream
# and downstream (left and right) TAD boundaries for each gene.


script,input_genes,output_TAD_boundaries=argv

tad_list=[]

#Make a list of TADs, where each TAD in the list is a sublist with chr, start, stop, and tad name entries
with open("total.combined.domains.named.bed", 'rb') as g:
            tad_reader = csv.reader(g,delimiter = "\t",quoting=csv.QUOTE_NONE)
            for tad in tad_reader:
                tad_list.append(tad)

fieldnames=['chr','start','stop','name']

tads_for_genes=open(output_TAD_boundaries, 'wb')

tad_writer=csv.writer(tads_for_genes, delimiter = "\t",quoting=csv.QUOTE_NONE)

#Iterate through gene list. For each gene, iterate through the TAD list to find which TAD encompasses the gene.
# I.e. The gene and TAD are on the same chromosome, the start coordinate of the TAD is less than the start coordinate of the gene,
# and the end coordinate of the TAD is greater than the end coordinate of the gene.
#To extract the TAD boundary downsteam of the gene, select the end coordinate of the TAD downsteam of the gene,
# and the start coordinate of the TAD encompassing the gene (boundary_left).
#To extract the TAD boundary upstream of the gene, select the end coordinate of the TAD encompassing the gene,
# and the start coordinate of the TAD upstream of the gene (boundary_right).

with open(input_genes, 'rb') as f:
    gene_reader = csv.DictReader(f,delimiter = "\t",quoting=csv.QUOTE_NONE,fieldnames=fieldnames)
    for gene in gene_reader:
        pos = 0
        for tad in tad_list:
            if gene["chr"] == tad[0]:
                if int(gene["start"]) >= int(tad[1]) and int(gene["stop"]) <= int(tad[2]) :
                    chr = gene["chr"]
                    if tad_list[pos] == tad_list[-1]:
                        boundary_left = gene["chr"] + ":" + tad_list[pos-1][2] + "-" +  tad_list[pos][1] + ",left_tad:" + gene["name"]
                        boundary_right = gene["chr"] + ":" + tad_list[pos][2] + "-" + tad_list[pos][2] + ",right_tad:" + gene["name"]
                        tad_writer.writerow([chr,(tad_list[pos][1]),tad_list[pos][1],boundary_left])
                        tad_writer.writerow([chr,(tad_list[pos][2]),tad_list[pos][2],boundary_right])
                    else:
                        if tad_list[pos+1][0] == chr and tad_list[pos-1][0] == chr:
                            boundary_left = gene["chr"] + ":" + tad_list[pos-1][2] + "-" +  tad_list[pos][1] + ",left_tad:" + gene["name"]
                            boundary_right = gene["chr"] + ":" + tad_list[pos][2] + "-" + tad_list[pos+1][1] + ",right_tad:" + gene["name"]
                            tad_writer.writerow([chr,(tad_list[pos-1][2]),tad_list[pos][1],boundary_left])
                            tad_writer.writerow([chr,(tad_list[pos][2]),tad_list[pos+1][1],boundary_right])
                        #If TAD is at the leftmost end of chromosome, set left TAD boundary equal to start coordinate of gene-containing TAD
                        elif tad_list[pos+1][0] != chr:
                            boundary_left = gene["chr"] + ":" + tad_list[pos][1] + "-" +  tad_list[pos][1] + ",left_tad:" + gene["name"]
                            boundary_right = gene["chr"] + ":" + tad_list[pos][2] + "-" + tad_list[pos+1][1] + ",right_tad:" + gene["name"]
                            tad_writer.writerow([chr,(tad_list[pos][1]),tad_list[pos][1],boundary_left])
                            tad_writer.writerow([chr,(tad_list[pos][2]),tad_list[pos+1][1],boundary_right])
                        #If TAD is at the rightmost end of chromosome ,set right TAD boundary equal to end coordinate of gene-containing TAD
                        else:
                            boundary_left = gene["chr"] + ":" + tad_list[pos-1][2] + "-" +  tad_list[pos][1] + ",left_tad:" + gene["name"]
                            boundary_right = gene["chr"] + ":" + tad_list[pos][2] + "-" + tad_list[pos][2] + ",right_tad:" + gene["name"]
                            tad_writer.writerow([chr,(tad_list[pos][1]),tad_list[pos][1],boundary_left])
                            tad_writer.writerow([chr,(tad_list[pos][2]),tad_list[pos][2],boundary_right])
                else:
                    pass
                    pos = pos + 1
            else:
                pass
                pos = pos + 1

tads_for_genes.close()
```

Figure A.1: This script takes a BED file of TADs from combined replicates of human embryonic stem cells [13], and a BED file with gene coordinates and names, and outputs the TAD boundaries flanking the TAD in which a gene resides. This script does not yet take into account genes that reside inside TAD boundaries.

Bibliography

[1] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984, June 2011.

[2] J. Aicardi. Aicardi syndrome. *Brain and Development*, 27(3):164–171, Apr. 2005.

[3] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jrgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Mller, T. F. Consortium, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar. 2014.

[4] C. C. Babbitt, M. Markstein, and J. M. Gray. Recent advances in functional assays of transcriptional enhancers. *Genomics*, 106(3):137–139, Sept. 2015.

[5] M. Baker. Structural variation: the genome's hidden architecture. *Nature Methods*, 9(2):133–137, Feb. 2012.

[6] G. Barry. Integrating the roles of long and small non-coding RNA in brain function and disease. *Molecular Psychiatry*, 19(4):410–416, Apr. 2014.

[7] K. J. Carss, G. Arno, M. Erwood, J. Stephens, A. Sanchis-Juan, S. Hull, K. Megy, D. Grozeva, E. Dewhurst, S. Malka, V. Plagnol, C. Penkett, K. Stirrups, R. Rizzo, G. Wright, D. Josifova, M. Bitner-Glindzicz, R. H. Scott, E. Clement, L. Allen, R. Armstrong, A. F.

Brady, J. Carmichael, M. Chitre, R. H. H. Henderson, J. Hurst, R. E. MacLaren, E. Murphy, J. Paterson, E. Rosser, D. A. Thompson, E. Wakeling, W. H. Ouwehand, M. Michaelides, A. T. Moore, T. Aitman, H. Alachkar, S. Ali, L. Allen, D. Allsup, G. Ambegaonkar, J. Anderson, R. Antrobus, R. Armstrong, G. Arno, G. Arumugakani, S. Ashford, W. Astle, A. Attwood, S. Austin, C. Bacchelli, T. Bakchoul, T. K. Bariana, H. Baxendale, D. Bennett, C. Bethune, S. Bibi, M. Bitner-Glindzicz, M. Bleda, H. Boggard, P. Bolton-Maggs, C. Booth, J. R. Bradley, A. Brady, M. Brown, M. Browning, C. Bryson, S. Burns, P. Calleja, N. Canham, J. Carmichael, K. Carss, M. Caulfield, E. Chalmers, A. Chandra, P. Chinnery, M. Chitre, C. Church, E. Clement, N. Clements-Brod, V. Clowes, G. Coghlan, P. Collins, N. Cooper, A. Creaser-Myers, R. DaCosta, L. Daugherty, S. Davies, J. Davis, M. De Vries, P. Deegan, S. V. V. Deevi, C. Deshpande, L. Devlin, E. Dewhurst, R. Doffinger, N. Dormand, E. Drewe, D. Edgar, W. Egner, W. N. Erber, M. Erwood, T. Everington, R. Favier, H. Firth, D. Fletcher, F. Flinter, J. C. Fox, A. Frary, K. Freson, B. Furie, A. Furnell, D. Gale, A. Gardham, M. Gattens, N. Ghali, P. K. Ghataorhe, R. Ghurye, S. Gibbs, K. Gilmour, P. Gissen, S. Goddard, K. Gomez, P. Gordins, S. Grf, D. Greene, A. Greenhalgh, A. Greinacher, S. Grigoriadou, D. Grozeva, S. Hackett, C. Hadinnapola, R. Hague, M. Haimel, C. Halmagyi, T. Hammerton, D. Hart, G. Hayman, J. W. M. Heemskerk, R. Henderson, A. Hensiek, Y. Henskens, A. Herwadkar, S. Holden, M. Holder, S. Holder, F. Hu, A. Huissoon, M. Humbert, J. Hurst, R. James, S. Jolles, D. Josifova, R. Kazmi, D. Keeling, P. Kelleher, A. M. Kelly, F. Kennedy, D. Kiely, N. Kingston, A. Koziell, D. Krishnakumar, T. W. Kuijpers, D. Kumararatne, M. Kurian, M. A. Laffan, M. P. Lambert, H. L. Allen, A. Lawrie, S. Lear, M. Lees, C. Lentaigne, R. Liesner, R. Linger, H. Longhurst, L. Lorenzo, R. Machado, R. Mackenzie, R. MacLaren, E. Maher, J. Maimaris, S. Mangles, A. Manson, R. Mapeta, H. S. Markus, J. Martin, L. Masati, M. Mathias, V. Matser, A. Maw, E. McDermott, C. McJannet, S. Meacham, S. Meehan, K. Megy, S. Mehta, M. Michaelides, C. M. Millar, S. Moledina, A. Moore, N. Morrell, A. Mumford, S. Murng, E. Murphy, S. Nejentsev, S. Noorani, P. Nurden, E. Oksenhendler, W. H. Ouwehand, S. Papadia, S.-M. Park, A. Parker, J. Pasi, C. Patch, J. Paterson, J. Payne, A. Peacock, K. Peerlinck, C. J. Penkett, J. Pepke-Zaba, D. J. Perry, V. Pollock, G. Polwarth, M. Ponsford, W. Qasim, I. Quinti, S. Rankin, J. Rankin, F. L. Raymond, K. Rehnstrom, E. Reid, C. J. Rhodes,

M. Richards, S. Richardson, A. Richter, I. Roberts, M. Rondina, E. Rosser, C. Roughley, K. Rue-Albrecht, C. Samarghitean, A. Sanchis-Juan, R. Sandford, S. Santra, R. Sargur, S. Savic, S. Schulman, H. Schulze, R. Scott, M. Scully, S. Seneviratne, C. Sewell, O. Shamardina, D. Shipley, I. Simeoni, S. Sivapalaratnam, K. Smith, A. Sohal, L. Southgate, S. Staines, E. Staples, H. Stauss, P. Stein, J. Stephens, K. Stirrups, S. Stock, J. Suntharalingam, R. C. Tait, K. Talks, Y. Tan, J. Thachil, J. Thaventhiran, E. Thomas, M. Thomas, D. Thompson, A. Thrasher, M. Tischkowitz, C. Titterton, C.-H. Toh, M. Toshner, C. Treacy, R. Trembath, S. Tuna, W. Turek, E. Turro, C. Van Geet, M. Veltman, J. Vogt, J. von Ziegenweldt, A. V. Noordegraaf, E. Wakeling, I. Wanjiku, T. Q. Warner, E. Wassmer, H. Watkins, A. Webster, S. Welch, S. Westbury, J. Wharton, D. Whitehorn, M. Wilkins, L. Willcocks, C. Williamson, G. Woods, J. Wort, N. Yeatman, P. Yong, T. Young, P. Yu, A. Webster, and F. L. Raymond. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *The American Journal of Human Genetics*, 100(1):75–90, Jan. 2017.

[8] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9):677–681, Sept. 2009.

[9] X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Kllberg, A. J. Cox, S. Kruglyak, and C. T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, Apr. 2016.

[10] C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. R. Cramer, M. Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank, and M. C. Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, Dec. 2016.

[11] P. Cowie, E. A. Hay, and A. MacKenzie. The noncoding human genome and the future of personalised medicine. *Expert Reviews in Molecular Medicine*, 17, 2015.

[12] A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics*, 7(12), Dec. 2011.

[13] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012.

[14] C. Du, B. N. Pusey, C. J. Adams, C. C. Lau, W. P. Bone, W. A. Gahl, T. C. Markello, and D. R. Adams. Explorations to improve the completeness of exome sequencing. *BMC Medical Genomics*, 9:56, 2016.

[15] A. L. Duker, B. C. Ballif, E. V. Bawle, R. E. Person, S. Mahadevan, S. Alliman, R. Thompson, R. Traylor, B. A. Bejjani, L. G. Shaffer, J. A. Rosenfeld, A. N. Lamb, and T. Sahoo. Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in PraderWilli syndrome. *European Journal of Human Genetics*, 18(11):1196–1201, Nov. 2010.

[16] A. C. English, W. J. Salerno, O. A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D. I. Ritter, C. R. Beck, C. F. Davis, M. Dahdouli, S. Ma, A. Carroll, N. Veeraraghavan, J. Bruestle, B. Drees, A. Hastie, E. T. Lam, S. White, P. Mishra, M. Wang, Y. Han, F. Zhang, P. Stankiewicz, D. A. Wheeler, J. G. Reid, D. M. Muzny, J. Rogers, A. Sabo, K. C. Worley, J. R. Lupski, E. Boerwinkle, and R. A. Gibbs. Assessing structural variation in a personal genometowards a human reference diploid genome. *BMC Genomics*, 16(1), Apr. 2015.

[17] M. Esteller. Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12):861–874, Dec. 2011.

[18] R. Fox, J. Ealing, H. Murphy, D. P. Gow, and D. Gosal. A novel DNMT1 mutation associated with early onset hereditary sensory and autonomic neuropathy, cataplexy, cerebellar atrophy, scleroderma, endocrinopathy, and common variable immune deficiency. *Journal of the peripheral nervous system: JPNS*, 21(3):150–153, Sept. 2016.

[19] T. Gao, B. He, S. Liu, H. Zhu, K. Tan, and J. Qian. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics*, page btw495, Aug. 2016.

[20] C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. M. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, R. Leach, R. Klein, R. Tearle, T. Bo, R. Pfundt, H. G. Yntema, B. B. A. de Vries, T. Kleefstra, H. G. Brunner, L. E. L. M. Vissers, and J. A. Veltman. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347, July 2014.

[21] R. L. Goldfeder, J. R. Priest, J. M. Zook, M. E. Grove, D. Waggott, M. T. Wheeler, M. Salit, and E. A. Ashley. Medical implications of technical accuracy in genome sequencing. *Genome Medicine*, 8:24, 2016.

[22] J. Hehir-Kwa, T. Marschall, W. P. Kloosterman, L. C. Francioli, J. A. Baaijens, L. Dijkstra, A. Abdellaoui, V. Koval, D. T. Thung, R. Wardenaar, B. Coe, P. Deelen, J. d. Ligt, E.-W. Lameijer, F. v. Dijk, F. Hormozdiari, E. E. Eichler, P. d. Bakker, M. Swertz, C. Wijmenga, G.-J. v. Ommen, E. Slagboom, D. Boomsma, G. o. t. Netherlands, A. Schoenhuth, K. Ye, and V. Guryev. A high-quality reference panel reveals the complexity and distribution of structural genome changes in a human population. *bioRxiv*, page 036897, Jan. 2016.

[23] B. Heidenreich, P. S. Rachakonda, K. Hemminki, and R. Kumar. TERT promoter mutations in cancer development. *Current Opinion in Genetics & Development*, 24:30–37, Feb. 2014.

[24] W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, Feb. 2012.

[25] J. Huddleston, M. J. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. S. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C.-S. Chin, J. Korlach, R. K. Wilson, and E. E. Eichler. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, page gr.214007.116, Nov. 2016.

[26] P. A. Jacobs, C. M. Wilson, J. A. Sprenkle, N. B. Rosenshein, and B. R. Migeon. Mechanism of origin of complete hydatidiform moles. *Nature*, 286(5774):714–716, Aug. 1980.

[27] E. Khurana, Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. S. Evani,

P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. H. Gm, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. M. Lipkin, D. G. MacArthur, G. Marth, D. Muzny, T. H. Pers, G. R. S. Ritchie, J. A. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu, . G. P. Consortium, E. T. Dermitzakis, H. Yu, M. A. Rubin, C. Tyler-Smith, and M. Gerstein. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*, 342(6154):1235587, Oct. 2013.

[28] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Mar. 2014.

[29] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, Feb. 2011.

[30] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15:R84, 2014.

[31] L. A. Lettice, S. J. H. Heaney, L. A. Purdie, L. Li, P. d. Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. d. Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735, July 2003.

[32] D. Lupiez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–1025, May 2015.

[33] Y. Ma, Y. Xin, R. Li, Z. Wang, Q. Yue, F. Xiao, and X. Hao. TFDP3 was expressed in coordination with E2f1 to inhibit E2f1-mediated apoptosis in prostate cancer. *Gene*, 537(2):253–259, Mar. 2014.

[34] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(Database issue):D986–D992, Jan. 2014.

[35] A. MacKenzie, B. Hing, and S. Davidson. Exploring the effects of polymorphisms on cis-regulatory signal transduction response. *Trends in Molecular Medicine*, 19(2):99–107, Feb. 2013.

[36] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7:29–59, 2006.

[37] N. Matharu and N. Ahituv. Minor Loops in Major Folds: EnhancerPromoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS Genet*, 11(12):e1005640, Dec. 2015.

[38] A. Mathelier, W. Shi, and W. W. Wasserman. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76, Feb. 2015.

[39] M. McGrath. Charcot-Marie-Tooth 1a: A narrative review with clinical and anatomical perspectives. *Clinical Anatomy*, 29(5):547–554, July 2016.

[40] . Menca, S. Modamio-Hybjr, N. Redshaw, M. Morn, F. Mayo-Merino, L. Olavarrieta, L. A. Aguirre, I. del Castillo, K. P. Steel, T. Dalmay, F. Moreno, and M. . Moreno-Pelayo. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics*, 41(5):609–613, May 2009.

[41] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kura, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Sttz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, and J. O. Korbel. Mapping copy number variation by population scale genome sequencing. *Nature*, 470(7332):59–65, Feb. 2011.

[42] M. Mohiyuddin, J. C. Mu, J. Li, N. B. Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam. MetaSV: an accurate and inte-

grative structural-variant caller for next generation sequencing. *Bioinformatics*, 31(16):2741–2744, Aug. 2015.

[43] J. C. Mu, M. Mohiyuddin, J. Li, N. B. Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, 31(9):1469–1471, May 2015.

[44] A. C. Need, V. Shashi, Y. Hitomi, K. Schoch, K. V. Shianna, M. T. McDonald, M. H. Meisler, and D. B. Goldstein. Clinical application of exome sequencing in undiagnosed genetic conditions. *Journal of Medical Genetics*, pages jmedgenet–2012–100819, May 2012.

[45] A. W. Pang, J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq, D. F. Conrad, H. Park, M. E. Hurles, C. Lee, J. C. Venter, E. F. Kirkness, S. Levy, L. Feuk, and S. W. Scherer. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11:R52, 2010.

[46] H. Parikh, M. Mohiyuddin, H. Y. K. Lam, H. Iyer, D. Chen, M. Pratt, G. Bartha, N. Spies, W. Losert, J. M. Zook, and M. Salit. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, 17:64, 2016.

[47] C. Redin, H. Brand, R. L. Collins, T. Kammin, E. Mitchell, J. C. Hodge, C. Hanscom, V. Pillalamarri, C. M. Seabra, M.-A. Abbott, O. A. Abdul-Rahman, E. Aberg, R. Adley, S. L. Alcaraz-Estrada, F. S. Alkuraya, Y. An, M.-A. Anderson, C. Antolik, K. Anyane-Yeboa, J. F. Atkin, T. Bartell, J. A. Bernstein, E. Beyer, I. Blumenthal, E. M. H. F. Bongers, E. H. Brilstra, C. W. Brown, H. T. Brggenwirth, B. Callewaert, C. Chiang, K. Corning, H. Cox, E. Cuppen, B. B. Currall, T. Cushing, D. David, M. A. Deardorff, A. Dheedene, M. D'Hooghe, B. B. A. de Vries, D. L. Earl, H. L. Ferguson, H. Fisher, D. R. Fitz-Patrick, P. Gerrol, D. Giachino, J. T. Glessner, T. Gliem, M. Grady, B. H. Graham, C. Griffis, K. W. Gripp, A. L. Gropman, A. Hanson-Kahn, D. J. Harris, M. A. Hayden, R. Hill, R. Hochstenbach, J. D. Hoffman, R. J. Hopkin, M. W. Hubshman, A. M. Innes, M. Irons, M. Irving, J. C. Jacobsen, S. Janssens, T. Jewett, J. P. Johnson, M. C. Jongmans, S. G. Kahler, D. A. Koolen, J. Korzelius, P. M. Kroisel, Y. Lacassie, W. Lawless, E. Lemyre, K. Leppig, A. V. Levin, H. Li, H. Li, E. C. Liao, C. Lim, E. J. Lose, D. Lucente, M. J. Macera, P. Manavalan, G. Mandrile, C. L. Marcelis, L. Margolin, T. Mason, D. Masser-

Frye, M. W. McClellan, C. J. Z. Mendoza, B. Menten, S. Middelkamp, L. R. Mikami, E. Moe, S. Mohammed, T. Mononen, M. E. Mortenson, G. Moya, A. W. Nieuwint, Z. Ordulu, S. Parkash, S. P. Pauker, S. Pereira, D. Perrin, K. Phelan, R. E. P. Aguilar, P. J. Poddighe, G. Pregno, S. Raskin, L. Reis, W. Rhead, D. Rita, I. Renkens, F. Roelens, J. Ruliera, P. Rump, S. L. P. Schilit, R. Shaheen, R. Sparkes, E. Spiegel, B. Stevens, M. R. Stone, J. Tagoe, J. V. Thakuria, B. W. van Bon, J. van de Kamp, I. van Der Burgt, T. van Essen, C. M. van Ravenswaaij-Arts, M. J. van Roosmalen, S. Vergult, C. M. L. Volker-Touw, D. P. Warburton, M. J. Waterman, S. Wiley, A. Wilson, M. d. l. C. A. Yerena-de Vega, R. T. Zori, B. Levy, H. G. Brunner, N. de Leeuw, W. P. Kloosterman, E. C. Thorland, C. C. Morton, J. F. Gusella, and M. E. Talkowski. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, 49(1):36–45, Jan. 2017.

[48] T. R. Sana, R. Debets, J. C. Timans, J. F. Bazan, and R. A. Kastelein. Computational identification, cloning, and characterization of IL-1r9, a novel interleukin-1 receptor-like gene encoded over an unusually large interval of human chromosome Xq22.2-q22.3. *Genomics*, 69(2):252–262, Oct. 2000.

[49] G. S. Sellick, K. T. Barker, I. Stolte-Dijkstra, C. Fleischmann, R. J. Coleman, C. Garrett, A. L. Gloyn, E. L. Edghill, A. T. Hattersley, P. K. Wellauer, G. Goodwin, and R. S. Houlston. Mutations in PTF1a cause pancreatic and cerebellar agenesis. *Nature Genetics*, 36(12):1301–1305, Dec. 2004.

[50] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, page btv009, Jan. 2015.

[51] E. Sinkiewicz-Darol, D. Kabziska, I. Moszyska, and A. Kochaski. The 5' regulatory sequence of the PMP22 in the patients with Charcot-Marie-Tooth disease. *Acta Biochimica Polonica*, 57(3):373–377, 2010.

[52] D. Smedley, M. Schubach, J. O. B. Jacobsen, S. Khler, T. Zemojtel, M. Spielmann, M. Jger, H. Hochheiser, N. L. Washington, J. A. McMurry, M. A. Haendel, C. J. Mungall, S. E. Lewis, T. Groza, G. Valentini, and P. N. Robinson. A Whole-Genome Analysis Framework for

Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *The American Journal of Human Genetics*, 0(0), Aug. 2016.

[53] D. J. Stavropoulos, D. Merico, R. Jobling, S. Bowdin, N. Monfared, B. Thiruvahindrapuram, T. Nalpathamkalam, G. Pellecchia, R. K. C. Yuen, M. J. Szego, R. Z. Hayeems, R. Z. Shaul, M. Brudno, M. Girdea, B. Frey, B. Alipanahi, S. Ahmed, R. Babul-Hirji, R. B. Porras, M. T. Carter, L. Chad, A. Chaudhry, D. Chitayat, S. J. Doust, C. Cytrynbaum, L. Dupuis, R. Ejaz, L. Fishman, A. Guerin, B. Hashemi, M. Helal, S. Hewson, M. Inbar-Feigenberg, P. Kannu, N. Karp, R. H. Kim, J. Kronick, E. Liston, H. MacDonald, S. Mercimek-Mahmutoglu, R. Mendoza-Londono, E. Nasr, G. Nimmo, N. Parkinson, N. Quercia, J. Raiman, M. Roifman, A. Schulze, A. Shugar, C. Shuman, P. Sinajon, K. Siriwardena, R. Weksberg, G. Yoon, C. Carew, R. Erickson, R. A. Leach, R. Klein, P. N. Ray, M. S. Meyn, S. W. Scherer, R. D. Cohn, and C. R. Marshall. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *npj Genomic Medicine*, 1:15012, Jan. 2016.

[54] Z. Sun, Y. Wu, T. Ordog, S. Baheti, J. Nie, X. Duan, K. Hojo, J.-P. Kocher, P. J. Dyck, and C. J. Klein. Aberrant signature methylome by DNMT1 hot spot mutation in hereditary sensory and autonomic neuropathy 1e. *Epigenetics*, 9(8):1184–1193, Aug. 2014.

[55] M. Talkowski, G. Maussion, L. Crapper, J. Rosenfeld, I. Blumenthal, C. Hanscom, C. Chiang, A. Lindgren, S. Pereira, D. Ruderfer, A. Diallo, J. Lopez, G. Turecki, E. Chen, C. Gigek, D. Harris, V. Lip, Y. An, M. Biagioli, M. MacDonald, M. Lin, S. Haggarty, P. Sklar, S. Purcell, M. Kellis, S. Schwartz, L. Shaffer, M. Natowicz, Y. Shen, C. Morton, J. Gusella, and C. Ernst. Disruption of a Large Intergenic Noncoding RNA in Subjects with Neurodevelopmental Disabilities. *The American Journal of Human Genetics*, 91(6):1128–1134, Dec. 2012.

[56] L. Tattini, R. DAurizio, and A. Magi. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3, June 2015.

[57] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, Sept. 2014.

[58] H. Yang, P. N. Robinson, and K. Wang. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, 12(9):841–843, Sept. 2015.

[59] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*, 369(16):1502–1511, Oct. 2013.

[60] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov. 2009.

[61] X. Zhang. Exome sequencing greatly expedites the progressive research of Mendelian diseases. *Frontiers of Medicine*, 8(1):42–57, Mar. 2014.