IDENTIFICATION AND PRIORITIZATION OF PUTATIVE PATHOGENIC STRUCTURAL VARIANTS BASED ON FUNCTIONAL ANNOTATION

Jakob Hertzberg

Dissertation zur Erlangung des Grades eines Doktors der
Naturwissenschaften am Fachbereich der Mathematik und
Informatik der Freien Universität Berlin

Identification and Prioritization of Putative Pathogenic Structural Variants based on Functional Annotation

Detektion und Priorisierung von Potentiell Pathogenen Strukturvarianten anhand von Regulatorischen Annotationen

Jakob Hertzberg

Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) am Fachbereich der Mathematik und Informatik der Freien Universität Berlin

Berlin, 2023

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachter: Prof. Dr. Stefan Mundlos
Datum der Disputation: 05.05.2023

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

[1] Jakob Hertzberg, Stefan Mundlos, Martin Vingron, and Giuseppe Gallone. "TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs." In: *Genome biology* 23.1 (2022), pp. 1–21.

The following publications cover work relevant to the data analyzed in this Ph.D. thesis:

[1] Jonas Elsner, Martin A Mensah, Manuel Holtgrewe, Jakob Hertzberg, Stefania Bigoni, Andreas Busche, Marie Coutelier, Deepthi C de Silva, Nursel Elçioglu, Isabel Filges, et al. "Genome sequencing in families with congenital limb malformations." In: *Human Genetics* 140.8 (2021), pp. 1229–1239.

ACKNOWLEDGEMENTS

During my time as a Ph.D. student, I received tremendous support and had the opportunity to work with many exceptional researchers in fascinating collaborations. First and foremost I would like to acknowledge my supervisor Prof. Martin Vingron for all the advice and discussions that shaped the work in this thesis. I would also like to thank Prof. Stefan Mundlos, who, as a member of my thesis advisory committee, provided invaluable advice throughout my doctoral studies, and Dr. Martin Kircher, the third member of the committee, for his feedback and helpful ideas. During most of the work described in this thesis, I worked closely with Uira Melo, the fourth committee member, whose input and support I am immensely grateful for. I also thank all the group members and researchers I worked with on this and other collaborative projects including but not limited to: Giuseppe Gallone, David Heller, Robert Schöpflin, Verena Heinrich, Hossein Moeinzadeh, Nico Alavi, Emel Comac, Jonas Elsner, Martin Mensa, and Manuel Holtgrewe. Finally, I want to thank my friends and family for their unwavering support and encouragement.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The *Desoxyribonulceic acid* (DNA) is a molecule shared by all living organisms. It is the blueprint for all life on earth. DNA is a polymer present in each individual cell - a sequence drawn from an alphabet of four nucleotides: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The DNA molecule consists of two individual polynucleotides called strands that are identical in sequence but differ in orientation: The forward (*sense*) and the reverse (*antisense*) strand. The two strands are coiled into a double-helix structure in which pairs of nucleotides (A with T, G with C) are connected via hydrogen bonds. This DNA double-helix is further organized into *chromosomes*, larger structures found in the nucleus of eukaryotic cells. The entirety of all chromosomes in a cell is called a *genome*. Its nucleotide sequence is commonly separated into *coding* and *non-coding* regions. *Coding* regions encode amino-acid sequences processed during *transcription* and *translation*. *Non-coding* regions are not actively transcribed but contain regulatory elements such as *promoters* and *enhancers* that control the expression of *genes*.

The basic building block - the nucleotide alphabet - is shared across all eukaryotes but the genomic sequence is highly variable. With increasing evolutionary distance the genetic variability between species tends to increase, as mutations accumulate over generations. Some large-scale events such as genome duplications or ancestral chromosome fusions can also influence the genome composition in comparably short evolutionary time frames. However, these events occur much less frequently than changes affecting the nucleotide sequence of the genome - also called *variants*. Variants can be passed on from a previous generation (*germline* variants) or occur in individual cells after conception (*somatic* variants). They can be categorized by the number of nucleotides they affect and range from *single nucleotide variants* (SNV) and short *Insertions/Deletions* (InDels) to larger genomic alterations, collectively described as *structural variants* (SVs).

Our human genome is *diploid* i.e. it consists of two sets of 23 chromosomes - 22 *autosomes* and 1 *allosome* - one inherited by the father and one by the mother. One set of chromosomes comprises approximately 3.1 billion nucleotides or *base pairs* (bp). Even though this sequence of nucleotides is highly similar between humans (>99% similarity), the average number of genetic variation in a human genome covers approximately 4-5 million bp [1]. This set of variants and the associated alternative sequences (*alleles*) of each individual are also called its *genotype*. The groups of alleles inherited by a single parent are referred to as *haplotypes*. If a variant is present in both haplotypes it is called *homozygous*. Variants inherited by a single parent are called *heterozygous*. Even though the majority of the variants are not disease-causing, they can influence the individual's *phenotype* i.e. their development, physical attributes, and disease susceptibility. Disease-causing variants can act by themselves affecting, for example, single genes causing *monogenic* disorders or in relation to other variants resulting in *multi-factoral/complex* disorders. Monogenic disorders are also referred to as *Mendelian* since they follow recessive and dominant inheritance patterns first established by Gregor Mendel.

SVs are responsible for the majority of altered nucleotides in the human genome [2]. They have been associated with a wide range of diseases such as limb malformations, autism and cancer [2–4]. The genetic variability in the human genome is therefore likely to be as dependent on SVs as on smaller genetic mutations. While SNVs and InDels have been studied extensively - especially those affecting coding sequences - disease-causing mechanisms of SVs remain for the most part unknown. This is largely due to challenges in the detection of SVs. SNVs and small InDels can be reliably identified and are routinely investigated in clinical practice. SVs, however, are much more challenging to detect. They tend to occur in repetitive regions that can not be accurately processed using *short-read sequencing* approaches - the currently most frequently used experimental method to investigate genetic sequences [5]. In addition, SVs are a much more heterogeneous group of variation than SNVs including multiple subgroups with unique variant signatures. *Long-read sequencing* in combination with current variant detection methods have shown great promise to overcome these limitations [6–9]. However, the high cost of these sequencing technologies has limited their application to genetic research. With continuously improving sequencing protocols and machines, the experimental cost has decreased drastically suggesting that long-read sequencing could be included in standard clinical practice more frequently in the future.

The quality and accuracy of sequencing experiments and variant callers are fundamental to the detection of disease-causing variants. Deriving the molecular diagnosis from the detected variation i.e. the prioritization of variants, however, is an equally important and challenging task. Sequencing experiments result in thousands of potential candidate SVs that need to be filtered and prioritized in order to identify in many cases a single disease-causing variant. While a large variety of prioritization methods for SNVs and InDels are available to assess their functional impact, few methods have been developed for the prioritization of potentially disease-causing SVs. In addition, the available current approaches focus almost exclusively on the coding effects of SVs even though several instances of disease-causing yet non-coding SVs are known. Their number is however still marginal in comparison to coding variation. This limits the potential to develop and validate any automated prioritization approach involving non-coding features. To address the lack of data more studies are required that identify SVs in rare disease patients with state-of-the-art methods and investigate non-coding disease-causing mechanisms. This raises the need for dynamic frameworks adaptable to cohort-specific disease contexts that allow to accurately detect SVs with current sequencing technologies and prioritize them with respect to their pathogenic potential.

1.1 RESEARCH OBJECTIVE

In this thesis, we discuss the detection and prioritization of SVs on the example of a patient cohort with congenital limb malformations. We include the entire process from sequencing experiments to the identification of disease-causing candidates. The results are based on a novel pipeline that allows determining a comprehensive SV call set by combining short- and long-read sequencing. Using this pipeline we prioritize SVs based on their potential pathogenic impact with respect to a set of functionally relevant annotations, supported by RNA-seq and Hi-C analysis. We also extend and evaluate a previously introduced method for the automated prioritization of CNVs outperforming current comparable approaches. The novel contributions of this thesis are:

- a **pipeline** that allows for the accurate **detection of SVs** using short- and long-read sequencing technology and their **prioritization based on functional impact**.

- a **flexible annotation framework** using sets of coding and non-coding elements relevant for specific disease contexts that can be applied to **all types of SVs**.

- the **identification of potentially disease-causing candidates** in a cohort of **limb-malformation patients**.

- the extension and evaluation of a **automated prioritization method for CNVs**.

1.2 THESIS OVERVIEW

In this chapter, we gave a brief introduction to the field of genetics and introduced the scope of the thesis. Chapter 2 extends this introduction providing information on methods and technologies that have been used to identify variants and discussing their use in clinical practice. In Chapter 3 we focus on our patient cohort, limb malformations, and relevant previous analyses. Chapter 4 provides a methodical overview of the analyses presented in this dissertation. Chapter 5 presents the results of our variant calling pipelines and the comparison of sequencing technologies and algorithms with respect to their ability to identify SVs. In Chapter 6 we discuss current methods for variant prioritization, evaluate the CNV prioritization method, introduce our manual annotation framework, and the curation of functional annotations for the identification of disease-causing SVs. In Chapter 7 we present the potentially disease-causing variants identified in our patient cohort. Finally, we discuss the results of this thesis in Chapter 8.

2

VARIANT DETECTION AND PRIORITIZATION

The work we present in this thesis can be separated into two parts: First, we employ short- and long-read sequencing in combination with several current calling algorithms to generate a comprehensive call set for each patient. Then we aim to identify coding or non-coding disease-causing candidates among thousands of detected SVs. In this chapter, we provide relevant background information for both parts in the context of current and supporting genetic research. We begin with a brief history of variant detection, and the underlying experimental methods including an overview of current algorithmic approaches for SV calling. Then we discuss the application of genome sequencing in the context of clinical diagnostics and previous work on the prioritization of disease-causing variation.

2.1 A BRIEF HISTORY OF VARIANT DETECTION

2.1.1 *Experimental Procedures*

Large genomic alterations were systematically identified as early as the mid-20th century through *Karyotyping*. Karyotyping is a cytogenetic method in which mitosis is arrested during metaphase and the chromosomes of a cell are arranged and displayed to be viewed under a microscope. These initial studies not only allowed to detect *aneuploidies* i.e. an untypical number of chromosomes but also larger anomalies in individual chromosomes [10, 11]. The potential to identify and characterize these anomalies increased with more sophisticated staining methods. *Fluorescence in situ hybridization* (FISH) introduced the staining and visualization of individual chromosome pairs and targeted sequences [12]. This allowed for the identification of a much greater variety of genomic alterations [13].

In the late 1970s DNA *sequencing* was first introduced by Sanger et al. [14]. The experimental process also know as the *chain-separation method* allowed to determine the nucleotide sequence of a DNA fragment in three steps: chain-terminating *polymerase-chain-reaction* (PCR), size separation by gel electrophoresis and sequence determination. The method defining step i.e. the chain termination PCR is based on the concept of *dideoxyribonucleotides* (ddNTPs) which are added to a growing DNA strand during PCR preventing any further nucleotides to bind and therefore restricting the strand to a specific size. The resulting DNA strands are then assessed in channels of gel electrophoresis. Since the size of the strand determines the speed with which they pass through the medium, the DNA sequence can be reconstructed by concatenating the nucleotides from the bottom up. Using Sanger sequencing, any variant can be identified with nucleotide precision. Due to its accuracy, it is still used as a gold standard and validation for more recent approaches. While it is highly accurate the method is comparably slow and expensive, limiting its use for any large-scale genome analyses.

With the introduction of *microarrays*, multiple regions of the genome could be investigated for variations in parallel albeit only at kilobase (kb) resolution [15]. In addition, microarrays are only able to detect changes in the amount of genomic material i.e. *Copy number variations* (CNVs). By targeting individual regions of the human genome (*probes*), generating complementary DNA (cDNA), and measuring the relative nucleotide sequence abundance in the DNA fragments of an individual (*target*) after hybridization, a genotype for each region can be established. Lower relative nucleotide sequence abundance indicates a lower copy number or a *Deletion* while increased relative abundance indicates a higher copy number e.g. a *Duplication*. Even though microarrays are not able to detect variants at nucleotide resolution, they are still used in clinical practice in combination with targeted gene panels to identify potentially disease-causing CNVs at low cost.

The class of *Next-generation sequencing* (NGS) approaches extends upon the initial sequencing concept, gradually replacing microarrays and cytogentic methods. NGS methods allow for automated, relatively cheap and fast sequence analysis, changing the quantity and resolution of detected variants dramatically [16]. In addition to the identification of variants in all known exons - *whole-exome sequencing* (WES) - entire genomes can be sequenced - *whole-genome sequencing* (WGS). The initial protocol consists of two main steps: clonal amplification of DNA fragments through PCR and massive parallel sequence determination by synthesis. This process produces millions of sequenced DNA fragments or *reads* which are further processed in downstream analysis.

The initial NGS protocol has been adapted in various ways to target individual research questions: *RNA-seq* allows to asses the *transcriptome* - the entirety of all transcribed genomic fragments [17]. Through cross-linking DNA fragments and subsequent digestion and ligation with restriction enzymes, *chromatin conformation capture* (3C) methods can determine the spatial organization of DNA in the nucleus [18]. By combining chromatin immunoprecipitation and sequencing, *ChiP-seq* allows to analyse the binding sites of transcription factors and histone modifications [19]. Single-cell sequencing protocols produce individual DNA or RNA profiles for each cell [20]. With the addition of bisulfite conversion the methylation of cytosines can be analyzed through *whole-genome bisulfite sequencing* (WGBS) [21]. Since their introduction, these NGS technologies any many other adaptions have revolutionized genetic research. Still, for the detection of variants they have several drawbacks: Due to the relatively short read length, NGS technologies are not able to accurately determine the sequences in repetitive regions and especially larger SVs are challenging to detect as they exceed the standard NGS read length. Finally, all NGS technologies rely on PCR which is inefficient in regions of increased *GC-content* [22].

In the last decade, efforts were made to address these limitations with the introduction of *third-generation sequencing technologies*. The first published method, Pacific Biosciences's *single-molecule-real-time* (PacBio-SMRT) sequencing, is a PCR-amplification free technology producing reads in kb-range [23]. During sequencing, circular single-stranded DNA templates are processed individually in so-called *zero-mode waveguiders* (ZMWs) that contain a DNA polymerase. Depending on the sequencing mode DNA templates either pass once or multiple times through the polymerase producing *continuous long reads* (CLRs) or *circular consensus reads* (CCRs), respectively. Errors during PacBio sequencing occur randomly and the majority appear as small InDels in the data. Sequencing a template multiple times i.e. generating CCRs can therefore reduce the error rate significantly from 5-15% to below 1% [24]. However, high-coverage CLRs sequencing experiments are considerably cheaper than CCR and have increased average read length.

A second method, introduced shortly after PacBio-SMRT is *Nanopore sequencing* by Oxford Nanopore Technologies (ONT) [25]. While Nanopore sequencing can also be categorized as a SMRT technology, the approach is fundamentally different in comparison to PacBio-SMRT. Briefly, the technology makes use of small protein pores i.e. *nanopores* in a biological or solid-state membrane with an electrical field that is designed to measure changes in the electric current when molecules pass through them. Since each nucleotide produces a characteristic change in the electric current, this allows for determining the sequence of a DNA fragment that is passing through a nanopore. Similar to PacBio-SMRT CLR reads, the ONT sequencing error rate for long reads is between 5% and 15% depending on the experimental protocol [26]. While the majority of errors during sequencing also appear as small InDels, they are not random but systematic errors in homopolymer regions [27]. Newer sequencing kits in combination with computational methods to detect and eliminate systematic errors aim to reduce the error rate with promising results [28].

Due to their increased read length Nanopore and PacBio sequencing have since their introduction drastically increased variant detection sensitivity in regions that were previously unmappable and have laid the groundwork for the accurate identification of the entire spectrum of SVs [26].

2.1.2 *Computational Approaches*

With new experimental protocols and sequencing technologies new algorithms needed to be developed to leverage their potential. Given the differences between short- and long-read sequencing and even individual long-read technologies, these methods are typically designed for a unique combination of sequencing technology and variant type. However, the initial preprocessing step - the *alignment* - is required by all variant detection methods. For the alignment, the sequencing data is either represented by an *assembly*, a single nucleotide sequence representing an individual's genome, or a collection of reads. Constructing an assembly can reduce certain biases in downstream analysis increasing the accuracy during variant detection but requires considerable resources e.g. long-read technologies with high coverage and in some cases, complementary sequencing experiments for *scaffolding* such as *Hi-C*, a chromosome conformation capture method [26]. Since most analyses - especially those involving multiple individuals - cannot meet these requirements, the far more frequent representation of the sequencing data are *reads*. The reads are passed to an alignment algorithm that conducts a base-wise comparison with a *reference genome* determining *matches* and *mismatches* as well as deletions and insertions. Given the size of the human genome, this alignment process is a computationally challenging task that has been addressed by various algorithms for both short- and long reads. A recent review and evaluation of current alignment methods has been conducted by Alser et al. including *bwa-mem* and *minimap2* - the algorithms used in the work presented in this dissertation for short- and long-read data, respectively [29].

The choice of the reference genome depends on the species and sequencing experiment. For humans, reference genomes are available in multiple versions which are continuously updated. Since the first published version - the result of the human genome project [30] - considerable efforts have been made towards filling unmapped regions or *gaps* in the reference sequence. By taking advantage of recently developed long-read sequencing technologies even the highly repetitive telomeric and centromeric regions have now been sequenced resulting in the first complete human reference genome - the *telomere-to-telomere* reference (T2T) [31]. While the T2T reference allows detecting SVs in previously unmappable regions, it is a very recent development. The majority of variant detection in clinical practice including our analysis is still performed with respect to previous versions of the human reference i.e. *GRCh37* or *GRCh38* since restructuring the existing alignment and variant calling pipelines requires a substantial effort. In addition, annotations of regulatory elements established through complementary sequencing technologies such as ChIP-seq have been determined with respect to a specific reference. This limits any downstream prioritization attempts involving functional annotation to the mappable genomic regions of the corresponding reference version.

The alignment process reveals differences between the investigated individual and the reference. The evidence of these differences is contained in the aligned reads which are encoded in *Sequence Alignment Map* (SAM) or *Binary Alignment Map* (BAM) format. *Variant callers* collect and process the evidence from the SAM/BAM files aiming to identify variant *signatures*, determine the corresponding variant type, and report them in a *variant call format* (VCF) file. Numerous callers are available for specific sequencing experiments, variant types, and specialized purposes such as detecting variants in repeat elements [32]. The identification of SNVs and InDels with short-read sequencing mainly relies on methods employing Bayesian models to determine the most likely genotype at a single position from stacks of aligned reads [33]. A comparison of the currently most used short-read SNV and InDel callers at different sequencing depths has been conducted by Supernat et al [34]. Given the low error rate of short-read sequencing approaches, all callers achieve high accuracy on SNV test sets even in low-coverage regions.

The detection of SVs is, however, more challenging since each SV type has a unique signature in the alignment. SVs are also abundant in repetitive regions and can extend beyond the length of individual reads further complicating the variant calling process. To identify SV signatures three types of evidence are commonly used: *split-read* (SR), *paired-end* (PR), and *read-depth* (RD) evidence. Several complementary approaches have been developed focused on individual types of evidence or combinations of them. The majority of short-read SV callers collect SR and PR evidence. A widely used and highly cited example is *Delly*, which operates based on an undirected, weighted graph that connects paired-end read pairs if they support the same variant signature [35]. Some short-read SV callers e.g. *lumpy* [36] additionally aim to improve their performance for CNVs by supplementing RD evidence. Others like *cn.MOPS* are specifically designed for the detection of copy number changes solely based on coverage information [37, 38]. These specialized implementations lead to unique call sets depending on the choice of the caller. Most short-read pipelines, therefore, employ multiple callers to maximize sensitivity. This, however, requires rigorous downstream filtering to discard any potential artifacts or false-positive calls accumulated over the individual call sets.

Long-read sequencing technologies generate single reads. PR evidence can therefore not be used. In addition, they are often performed with comparably low coverage due to the associated high cost, limiting the use of RD evidence. Thus, long-read SV callers largely rely on SR evidence alone to extract signatures of SVs, as shown in Figure 2.1, from the aligned reads. State-of-the-art and frequently used callers, as indicated by the number of citations, are *Sniffles2*, *SVIM* and *pbsv* [39–41]. Since these algorithms rely on the same type of evidence, the differences in their implementation are less pronounced in comparison to short-read SV callers. However, the clustering of SR signatures and the assignment of individual SV types can still differ considerably. SVIM, for example, is able to determine the original reference location of an inserted segment and can therefore separate interspersed duplications from novel-element insertions while Sniffles and pbsv classify all insertions as either tandem duplications or novel-element insertions. In addition to the divergent assignment of SV signatures and clustering, the callers also include built-in filtering parameters and thresholds that influence the number of reported calls and the proportions of individual SV types. Since no single best-performing method has yet been determined, an ensemble approach, as for the short-read callers, can be used to incorporate the results of several callers increasing the sensitivity but raising the need for a more extensive filtering process.

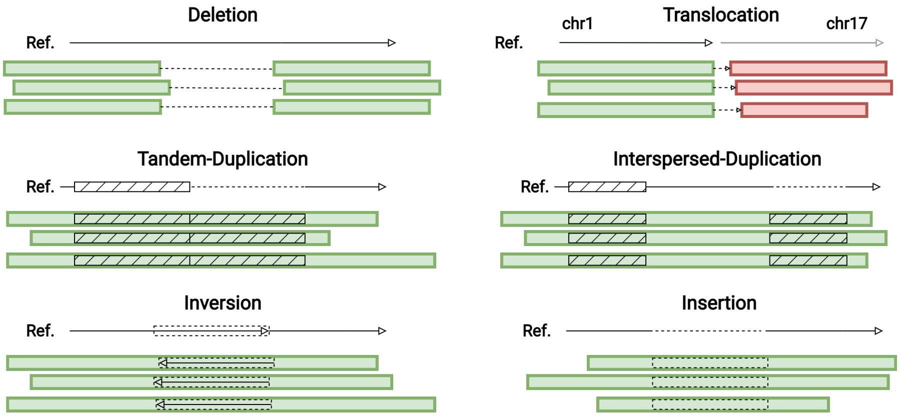


Figure 2.1: **Overview of SV Types based on SR evidence**. The subfigure show illustrations of split-reads aligned to a reference indicating Deletions, Duplications (tandem and interspersed), Inversions, and Insertions.

2.2 FROM SEQUENCING TO CLINICAL DIAGNOSIS

NGS technologies have been used extensively to investigate the human genome for almost two decades. Large-scale studies of populations with thousands of individuals have shed light on the genetic variability in humans, generating comprehensive catalogs of common variation and investigating selective pressure throughout the human genome [1, 42, 43]. In recent years comparable efforts have been published using third-generation sequencing further extending our understanding of genetic variation - especially concerning SVs [44, 45]. *Genome-wide-association-studies* (GWAS) explore the statistical associations between human traits and genotypes by exploiting *linkage disequilibrium* (LD) [46]. These associations, however, often include multiple variants associated with a single phenotype i.e. *complex diseases* and do not allow inferring a direct causative relation. Studying rare disease and inherited disorders in humans requires a more specialized approach. WES and gene panel sequencing have been widely used to study variants affecting known coding regions and have become significant tools used in clinical diagnostics. Rare disease patients are now frequently sequenced and their data is analyzed with the use of various variant calling and interpretation methods to determine the molecular origin of their phenotype [47–49].

WES can offer a substantial increase in successful molecular diagnosis in comparison with previous conventional genetic testing [50]. The diagnostic yield is, however, highly variable depending on the disease context and investigated tissue [51, 52]. With recorded diagnostic yields around 30%, many patients remain without a molecular diagnosis after WES analysis [52]. Since the exome constitutes only 2% of our genome, it is highly likely that the cause of the phenotype for many of these unsolved cases lies in non-coding regions [53]. Several mechanisms have been identified linking non-coding variants to human disease: 1) *Loss- or Gain-of-Function* in target genes of enhancers caused by variants disrupting enhancer sequences [54, 55]. 2) Misexpression of genes through copy number modification of enhancers and other regulatory elements [56]. 3) Perturbation of the chromatin conformation resulting in changes of the regulatory environment through balanced and unbalanced SVs. These perturbations can rewire interactions between enhancers and genes in a process called *enhancer adoption* or *enhancer hijacking* [57]. Given these documented cases of disease-causing variation, it is necessary to include non-coding regions in standard clinical analysis through WGS to maximize the diagnostic yield.

There are, however, several challenges for the large-scale assessment of non-coding variants in the context of clinical diagnostics: First, the non-coding variants identified using WGS greatly outnumber coding variants. On average $\sim 12,000$ variants are identified using WES with 90% present in public databases while a WGS experiment produces approximately ~ 5 million variants [58]. Reducing the number of non-coding variants to a set with few enough to be assessed by a clinician/geneticist, requires rigorous and sophisticated filtering and prioritization approaches. Secondly, our knowledge about the non-coding regions and their biological function is far from complete. In a clinical diagnostic scenario, a single functionally relevant and disease-causing variant among hundreds of thousands needs to be identified. In comparison to coding variants, which can be directly assessed by their effect on transcription and translation i.e. amino-acid sequences and protein function, non-coding variants often affect regions without known biological function. To address this limitation substantial efforts have been made to understand the regulatory involvement of non-coding regions. Examples include the *ENCODE* project, a large-scale collection of ChIP-seq and chromatin accessibility data in hundreds of tissues and cell types [59], studies presenting extensive investigations of chromatin conformation [60–62], *Cap-analysis gene expression sequencing* (CAGE-seq) experiments to identify transcribed enhancers [63] as well as curated sets of experimentally validated regulatory elements [64].

Supported by the continuously extending knowledge of the tissue-specific function of non-coding elements, recent WGS studies show promising results toward the identification of disease-causing variation in larger cohorts [65]. Still, the regulatory function of most identified non-coding elements remains unclear and the potential of WGS to identify disease-causing non-coding variation therefore is not yet exhausted. In addition, short-read sequencing is, as mentioned in the previous section, severely limited with respect to SV calling [66]. Thus, it is likely that even in clinical studies investigating genotype-phenotype relations using WGS SVs are underrepresented. To maximize the sensitivity of variant calling and the diagnostic yield, a combination of WGS and long-read sequencing has to be used. This also requires specialized prioritization approaches for SVs that allow for identifying functionally relevant coding and non-coding variation.

3

SEQUENCING PATIENTS WITH CONGENITAL LIMB MALFORMATIONS

Signaling pathways in limb development are known to be largely conserved but significant morphological differences can be observed across species. This indicates that gene regulation is a major driving factor of limb development. Since the limb is an organ that is easily observable and experimentally modifiable, it is an ideal model to study the mechanisms of gene regulation and has been widely used as such. In our cohort of patients with congenital limb malformations, these mechanisms have likely been disrupted by genetic alterations. Our aim is to identify these disease-causing variations with respect to the potentially disturbed regulatory environments. To provide the necessary background for this analysis we briefly summarize in this chapter the regulatory pathways involved in limb development. Then we discuss a previously conducted short-read WGS analysis including the majority of patients in our cohort which motivated the work discussed in this dissertation. Finally, we present an overview of our extended analysis.

3.1 HUMAN LIMB DEVELOPMENT AND LIMB MALFORMATIONS

The limb development originates from the lateral plate mesoderm. First the *limb bud* is formed, an ectodermal pocket that encloses the proliferating mesenchyme. The development continues along three closely coordinated axes [67]: *proximal-distal* (PD), *anterior-posterior* (AP), and *dorsal-ventral* (DV). The development of each axis is associated with an individual signaling center. The PD-axis is under the control of the *apical ectodermal ridge* (AER) sitting on the tip of the limb bud. During limb development, the AER continues to keep the underlying mesenchyme in a proliferating state, which allows the rest of the limb to grow. The signaling center of the AP-axis is the zone of polarized activity (ZPA), which expresses the *sonic hedgehog gene* (SHH). The DV-axis is controlled by the *WNT family member 7A* (WNT7A). The development of each axis and the coordination between them is tightly controlled by numerous genes and cis-regulatory elements. One prominent example involves the differentiation between hindlimb and forelimb i.e. the*limb identity*. While the underlying regulatory process has not been entirely solved, several key factors have been established based on molecular evidence: The transcription factors TBX5 and TBX4 are expressed in the mesenchyme of the forelimb and hindlimb, respectively, presumably playing a major role in establishing limb identity. Both trigger the expression of the *fibroblast-growth-factor 10* (FGF10) which in turn induces the expression of FGF8 in a feedback loop. Through FGF8 the mesenchymal cells remain proliferated, promoting limb outgrowth. The expression of TBX5 and TBX4 has been found to depend on the sequential rostrocaudal HOX gene expression patterns.

With the introduction of ChIP-seq, RNA-seq, and chromatin conformation methods many such cis-regulatory elements and gene interactions associated with limb development have been identified and investigated both in the context of evolutionary changes [68] and human disease [69]. Since in this dissertation, we focus on the disease-causing mechanism of genetic variation, we only provide a small selection of the evolutionary studies at this point and extend further on limb-associated human disease in the following paragraph: Digit reduction in mammals [70], wing acquisition in bats [71] and limb loss in snakes [72, 73]. A more comprehensive review of limb-related evolutionary studies has been conducted by Petit et al. [74].

The study of the genetic mechanisms underlying congenital limb malformations in humans reaches back to the beginning of the 20th century. After the rediscovery of Mendel's work in 1900, the first human disorder recognized to follow his principles of inheritance was a limb malformation, specifically, the now-called brachydactyly type A1 [75]. Limb malformations are individually rare but overall appear in approximately 1 out of 500 individuals [76]. The prevalence of individual subtypes can vary significantly. They can be caused by environmental factors such as *teratogens* [77] as well as spontaneous and inherited genetic variation. Many of the identified causal variants have been associated with syndromes which include a range of other symptoms in addition to limb malformations [78]. This indicates that the involved genes are not specific to limb development but are also active in other relevant pathways. Non-syndromic or isolated limb malformations are therefore more likely to be caused by variation in *cis*-regulatory elements rather than coding regions, which is supported by individual examples of variants including SVs disrupting non-coding regulatory elements [3, 69]. A prominent disease-causing mechanism of SVs associated with limb malformations is the disruption of *topologically associating domains* (TADs) resulting in a rewiring of relevant enhancer-gene interactions [79, 80]. TADs are windows at the sub-megabase scale with increased interaction frequency determined through chromatin conformation capture. TAD boundaries are significantly enriched for the CCCTC-binding factor (CTCF) which acts as a key insulator protein [81, 82]. Disruption of TAD boundaries through SVs can therefore lead to newly formed interactions between genes and enhancers with potentially disease-causing consequences.

Examples of SVs disrupting non-coding regulatory mechanisms and causing human limb malformations are shown in Figure 3.1.

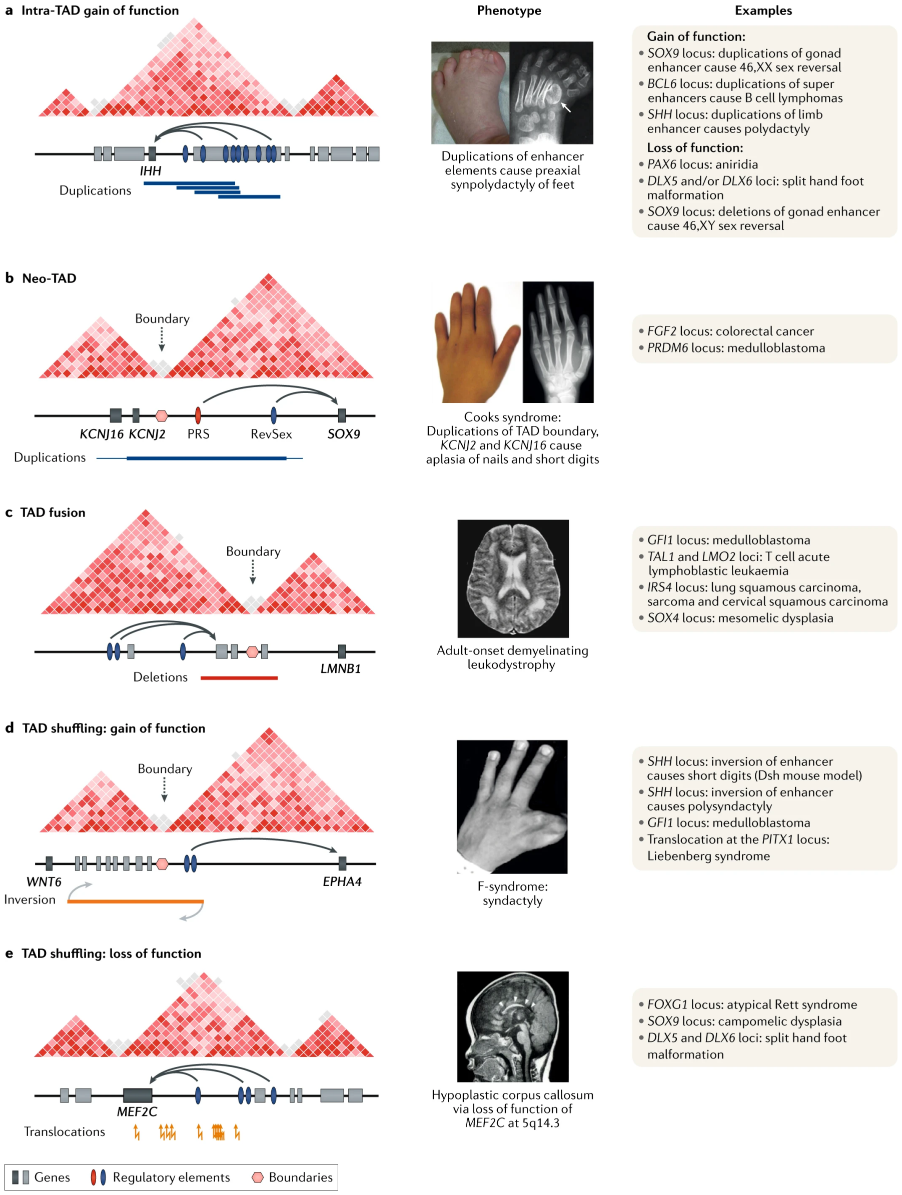

a Intra-TAD gain of function
Phenotype
Examples
IHH
Duplications
Duplications of enhancer elements cause preaxial synpolydactyly of feet
Gain of function:
• SOX9 locus: duplications of gonad enhancer cause 46,XX sex reversal
• BCL6 locus: duplications of super enhancers cause B cell lymphomas
• SHH locus: duplications of limb enhancer causes polydactyly
Loss of function:
• PAX6 locus: aniridia
• DLX5 and/or DLX6 loci: split hand foot malformation
• SOX9 locus: deletions of gonad enhancer cause 46,XY sex reversal
b Neo-TAD
Boundary
KCNJ16 KCNJ2 PRS RevSex SOX9
Duplications
Cooks syndrome: Duplications of TAD boundary, KCNJ2 and KCNJ16 cause aplasia of nails and short digits
• FGF2 locus: colorectal cancer
• PRDM6 locus: medulloblastoma
c TAD fusion
Boundary
LMNB1
Deletions
Adult-onset demyelinating leukodystrophy
• GFI1 locus: medulloblastoma
• TAL1 and LMO2 loci: T cell acute lymphoblastic leukaemia
• IRS4 locus: lung squamous carcinoma, sarcoma and cervical squamous carcinoma
• SOX4 locus: mesomelic dysplasia
d TAD shuffling: gain of function
Boundary
WNT6
EPHA4
Inversion
F-syndrome: syndactyly
• SHH locus: inversion of enhancer causes short digits (Dsh mouse model)
• SHH locus: inversion of enhancer causes polysyndactyly
• GFI1 locus: medulloblastoma
• Translocation at the PITX1 locus: Liebenberg syndrome
e TAD shuffling: loss of function
MEF2C
Translocations
Hypoplastic corpus callosum via loss of function of MEF2C at 5q14.3
• FOXG1 locus: atypical Rett syndrome
• SOX9 locus: campomelic dysplasia
• DLX5 and DLX6 loci: split hand foot malformation
Genes
Regulatory elements
Boundaries

Figure 3.1: **Clinical Examples of Non-Coding SVs in Patients with Limb Malformations.** This figure is adapted from Spielmann et al. 2018. It shows several limb malformations and depictions of the corresponding disease-causing mechanisms.

3.2 PREVIOUS SHORT READ ANALYSIS

We center the discussion of SVs and their potential to cause human disease in this thesis around a cohort of 21 patients (LM01-LM21) with limb malformations. All patients are non-syndromic and exhibit isolated limb malformations. While some symptoms are shared between patients, the combination of symptoms is unique for each individual, with the exception of LM17 and LM18. The patients are a subset of a larger cohort with limb malformations collected by the Department of Hand Surgery of the Katholisches Kinderkrankenhaus Wilhelmstift Hamburg and the Institute of Medical and Human Genetics at the Charité Berlin. The patients were previously tested for known genetic variants using gene panel sequencing and microarrays. Both investigations were unsuccessful in finding disease-causing variations. Elsner et. al, therefore, conducted an analysis using WGS data exploring variations located in coding and non-coding regions of the genome [83].

The Elsner et al. analysis included a total of 69 individuals with limb malformations. For 64 patients, the parents were sequenced allowing for the identification of *de novo* or *shared* variants if a parent showed limb malformations comparable to the patient. The main focus of this analysis was on small variations (SNVs and InDels). However, SVs larger than $1,500$bp were also included. The initial set of high-quality variants was filtered based on *allele frequency* (AF) both with respect to public databases and cohort-specific *allele counts* (AC). Rare variants were split into coding and non-coding variants depending on their overlap with known gene transcripts, then separately annotated and prioritized. Disease-causing coding variants were identified using their predicted effect on the corresponding protein in combination with phenotype information [84]. For non-coding variants, Elsner et al. constructed a framework including multiple cis-regulatory annotations associated with limb development that allows identifying variants potentially disrupting the regulatory environment of genes of interest. For 12 out of the initial 69 patients, candidate variants could be identified. All identified candidate variants are SNVs or InDels acting through coding disease-causing mechanisms.

3.3 EXTENDED ANALYSIS OF THE UNSOLVED CASES

Since short-read sequencing has been shown to capture only a minor proportion of SVs present in the average human genome, it is likely that the majority of SVs including potentially disease-causing variants in the remaining unsolved cases have not yet been detected. Therefore, the sequencing facility at the *Max Planck Institute for Molecular Genetics* (MPIMG) sequenced 21 patients out of the 69 limb malformation cases with PacBio long-read sequencing and we set out to identify a more comprehensive set of SVs based on this new data. We implemented a novel pipeline to process the long reads, reanalyze the original WGS data and perform extensive formatting and filtering steps resulting in a final call set of rare and potentially pathogenic SVs. To assess their functional impact, we curated a set of cis-regulatory annotations relevant to limb development supported by patient-specific Hi-C and RNA-seq analysis and developed an annotation and prioritization framework for both coding and non-coding SVs. Finally, we manually inspected functionally relevant SVs and derived a set of candidate variants for each patient.

4 METHODS

In this chapter, we provide detailed descriptions of the methods underlying the results discussed in this dissertation and relevant previous analyses. First, we provide a summary of a CNV prioritization method - the *TAD-annotation* (TADA) tool - first introduced in the master thesis leading up to and motivating the work done during the Ph.D. In this description, we focus primarily on the methodical details. A more extensive discussion of variant prioritization in general and the available state-of-the-art methods can be found in Chapter 6. We then describe the methods used for the evaluation of TADA's predictive performance in comparison with several current prioritization methods. In the second part of this chapter, we present the pipeline we implemented to process the sequencing data of the limb malformation cohort, call and filter SVs as well as prioritize and visualize them. This includes the changes made to TADA i.e. adjusting it to incorporate all types of SVs and specializing the annotation process to reflect the patients' phenotypes.

4.1 TADA - AUTOMATED PRIORITIZATION OF PATHOGENIC CNVS

The motivation for the development of TADA was to quantify the pathogenic potential of larger genomic alterations based on *machine-learning* (ML) models trained on sets of known pathogenic and benign variants. Current catalogs of annotated pathogenic SVs consist, however, almost exclusively of CNVs due to the limitations of the experimental methods used to detect SVs that were available in the last decades. The most frequently used method to determine larger variants was microarrays which only allow the detection of copy number changes. To increase the predictive performance of the automated prioritization model, we, therefore, decided to limit TADA to CNVs rather than the entire spectrum of SV types. We collected two call sets: A set of common i.e. non-pathogenic CNVs and a set of known pathogenic CNVs. We obtained the pathogenic CNVs from DECIPHER [85]. The common variant set is a compendium from four different data sources [44, 86–88]. After several filtering steps based on AF and overlap between data sets, the size and number matched training set of pathogenic and non-pathogenic CNVs included 6,130 Deletions and 3,410 Duplications (see Hertzberg et al. for details [89]).

We then set out to quantify the impact of CNVs based on the coding and non-coding regulatory elements they affect. This process was centered around a set of TAD boundaries. Their purpose in the annotation process is three-fold: First, TAD boundaries serve as a proxy of regulatory regions across the genome. This allows us to limit the genomic annotations affected by the CNVs to the loci between boundaries and compute features for CNVs with respect to the regulatory environment rather than just the directly affected loci. Secondly, they considerably reduced TADA's computational overhead since the feature computation is performed for TAD-sized windows rather than entire chromosomes. Third, they serve as non-coding annotations themselves. An overview of this TAD-based annotation framework is shown in Figure 4.1.

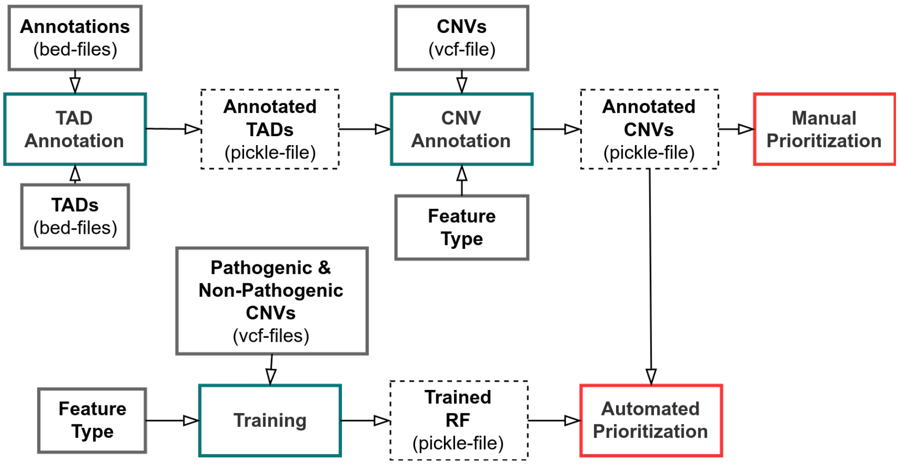


Figure 4.1: **Technical Workflow of TADA**. The figure shows an illustration of the CNV prioritization using the TADA tool. First, TADs are annotated given a set of functionally relevant annotations. The set of TADs is either a default set derived from hESC cells or provided by the user. Then sets of CNVs can be annotated with features derived from the affected annotated TADs. For this, a *Feature Type* needs to be selected i.e. *extended* for the default feature set or *distance* for user-defined annotations. The annotated CNVs are then either directly used in a manual analysis with user-specific filters or an automated analysis based on the pathogenicity score returned by the pre-trained Deletion and Duplication models. Additionally, users can provide two sets of CNVs for training an alternative random forest model with *extended* or *distance* features.

The annotation process requires an extensive catalog of coding and non-coding elements with metrics indicating their potential regulatory importance to best quantity the functional impact of CNVs. We collected dosage sensitivity and intolerance to *Loss-of-Function* scores for genes e.g. *LOEUF* [86, 90], a gene-set associated with developmental disease [85, 86], predicted and validated sets of enhancers [63, 64], CTCF sites [59] and conservation metrics [91]. We first sorted the genomic annotations into their corresponding TAD environment. Then we computed features for CNVs based on the overlapping TADs. The features included the distance to the closest element of each annotation set as well as compound metrics such as the combined haploinsufficiency (HI) i.e. *HI Log-Odds score* across all affected genes. The resulting features are included in the default running mode of TADA. However, we also allowed for user-defined sets of annotations. With this tissue-specific information can be included in the annotation process if needed.

The annotated CNVs can be used in two ways: A manual analysis, filtering based on individual features or they can be processed further to train machine learning models distinguishing between two sets of CNVs. To automatically prioritize pathogenic CNVs, we annotated the size- and number-matched sets of pathogenic and non-pathogenic Deletions and Duplications. For this, we included 14 functional annotation-derived features. Then we split the CNVs 70% 30% in training and test-set and used the training sets to train two separate random forest models for Deletions and Duplications. These random forest models are the basis for the evaluation and comparison with other prioritization methods presented in this dissertation.

4.1.1 *Evaluation of TADA*

ROC-AUC Analysis

For the first comparison of TADA with current prioritization approaches based on ROC-AUC values we used three sets of CNVs: A 5-fold cross-validation (CV) split of our training data i.e. the 70% split of the previously described set of Deletions and Duplications, the CNVs contained in the 30% test-split and a set of ClinVar variants. We collected the ClinVar CNVs from (*https://www.ncbi.nlm.nih.gov/clinvar/* by using the following filter settings: *Type of variation = copy number gain* OR *copy number loss* OR *Deletions* OR *duplications*. We first separated the variants into Deletions and Duplications (73,533 Deletions; 47,022 Duplications) and then into pathogenic i.e. *Pathogenic* and *Likely pathogenic* (11,816 Deletions; 3,880 Duplications) and non-pathogenic i.e. *Benign* and *Likely benign* (13,381 Deletions; 11,609 Duplications). For the evaluation, we only used variants located on autosomes. We also discarded any duplicated variants as previously described [89] and those overlapping with the training data (90% reciprocal overlap) resulting in 17,553 Deletions and 10,062 Duplications.

The *Receiver Operating Characteristic* (ROC) curves are based on two metrics: the *false positive rate* (FPR) on the x-axis and the *true posi-*

tive rate (TRP) also referred to as *sensitivity* or *recall* on the y-axis. The curve itself is computed by selecting varying thresholds on the prediction probability. An ideal classifier would achieve perfect sensitivity without any false positives i.e. an *area under the curve* (AUC) of 1. To compare the performance of TADA with other prioritization methods we, therefore, use the ROC-AUC values as an approximation of their predictive ability.

4.1.2 *F1-Score Analysis*

For the second evaluation of TADA we employ the F1-Score which also allows measuring the predictive performance of classification methods with categorical rather than continuous predictions. We compute the F1-Score as follows:

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FN + FP} \tag{1}$$

with TP as the number of *true positive*, FP *false positive* and FN *false negative* predicitions. All F1-Score computations are based on the ClinVar CNVs previously used for the ROC-AUC Analysis.

4.1.3 *Ranking Analysis*

In the third evaluation we set out to test the calibration of TADA's pathogenicity score in comparison with other prioritization approaches. For this, we generated *test-batches* of *ClinVar* Deletions and Duplications, each containing a single pathogenic variant and 99 benign variants. To account for the size bias between pathogenic and non-pathogenic ClinVar CNVs, we binned the benign variants by size using an *empirical cumulative distribution function* (ECDF) and 60 bins. For each pathogenic variant, we then sampled 99 benign CNVs from the same size bin. This resulted in a total of $3,425$ batches of Deletions and 415 batches of Duplications. Finally, to assess the *ranking ability* of the prioritization methods we first computed pathogenicity scores for all *test-batches*. Then we sorted the 100 CNVs in each batch by the pathogenicity scores separately such that the index of the true pathogenic variants could be used in a performance evaluation. We repeated the sampling process across 30 random seeds.

4.2 ANALYSIS OF THE UNSOLVED CASES

For the analysis of the unsolved limb malformation cases, we implemented a novel pipeline in snakemake (v.7.16.1) consisting of multiple workflows to process the short- and long-read sequencing data, call variants and prioritize them. A *directed acyclic graph* (DAG) of the entire snakemake pipeline is shown in Figure A.1 and a more generalized overview in Figure 4.2. In the following paragraphs, we describe the individual workflows i.e. *rules* of the snakemake pipeline.

First, we present the methodical details of the workflow we implemented to process the PacBio and Illumina data, as well as compare the call sets across callers and technologies (Chapter 5). Then we describe the adjustments to the TADA tool that allows us to perform a limb-malformation-specific annotation. We also provide details on the Hi-C and RNA-seq analysis that contribute to the collection of relevant functional annotations. Finally, we present the methods used to manually inspect functionally relevant variants to distinguish between false- and true positives and the visualization method applied to the filtered set of candidate variants.

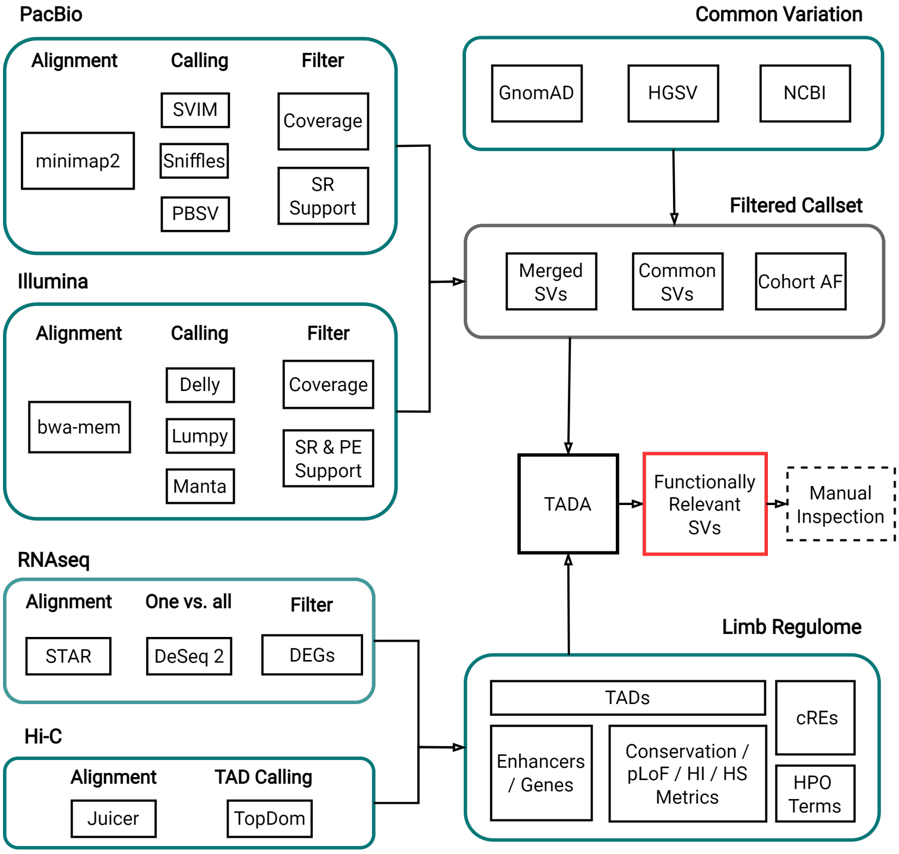


Figure 4.2: **Workflow of the Unsolved Cases Analysis.** We implemented SV calling pipelines for Illumina and PacBio sequencing, then merged, formatted, and filtered the calls based on AF resulting in a final SV call set. To analyze the potential impact of each SV we constructed a set of functionally relevant annotations including the results of patient-specific Hi-C and RNA-seq analysis. Finally, we perform a manual assessment of the remaining candidates with respect to the supporting sequencing evidence and the affected regulatory environment.

4.2.1 *PacBio Processing*

Alignment

First, we aligned the subreads to the reference GRCh38 producing BAM files for each sequencing run. For the alignment, we initially used both the *pbmm2* version of the *minimap2* algorithm [92] and the *nglmr* approach [93]. However, the alignment with *nglmr* was considerably more time-consuming without significant improvement in alignment statistics as measured in coverage, percentage of aligned reads, and within read alignability. Thus, all results are based on the alignment with *minimap2* alone. We generated index files for the subreads of each PacBio run using `pbmm2 index` and the `SUBREAD` preset. Then we aligned the raw data to the GRCh38 reference using `pbmm2 align` (v1.3.0) with the following command:

```
pbmm2 align --preset "SUBREAD" --unmapped --median-filter --strip
     --log-level INFO -j {threads} {index} {input.bam} {output.
    bam}
```

We sorted the aligned reads and again generated index files using `samtools` (v1.9). Finally, we merged all aligned files across runs for each patient into a single *pooled* BAM file with `samtools merge`, if necessary, and added MD tags with `samtools calmd`. We pooled all alignments for the patients sequenced in multiple SMRT cells, generating a single BAM file. We then analyzed the aligned sequencing data using visualizations generated with custom python scripts and computed the coverage by dividing the number of aligned bases by the number of mappable bases in the GRCh38 reference genome (3,088mb).

Variant Calling

We employ three callers to identify SVs based on our aligned long-read data: *SVIM* [40], *Sniffles2* [39] and *PBSV* [41]. For SVIM (v.1.4.2) and Sniffles (v2.0.2) we performed SV calling with single commands as shown below:

```
sniffles --tandem-repeats {tandem_repeats} --input {input.bam} --
    minsvlen 50 --minsupport-auto-mult 0.4 --vcf {output.vcf} --
    threads 4

svim alignment --sample {patient} --max_sv_size 5000000 --segment
    _gap_tolerance 20 --segment_overlap_tolerance 10 --tandem_
    duplications_as_insertions --interspersed_duplications_as_
    insertions --read_names --zmws {working_dir} {input.bam} {
    reference}
```

To call SVs using PBSV (v2.8.0) we used two separate commands `pbsv discover` and `pbsv call`. The tandem repeat file is provided in the PBSV Github repository:

```
pbsv discover --tandem-repeats {tandem_repeats} {input.bam} {
    output.signatures}

pbsv call -j {threads} --types INS,DEL,BND,INV --call-min-reads-
    all-samples 5 --call-min-reads-one-sample 5 --call-min-read-
    perc-one-sample 40 {reference} {input.signatures} {output.vcf
    }
```

We replaced the headers of all resulting VCF files to generalize the format of the patient IDs and normalized the calls using `bcftools norm -c s -d all`.

Formatting

The callers in our pipeline share several similarities in their implementation. All first identify so-called *signatures* from the *CIGAR* strings, encoded information of the read alignment found in the BAM files. The *CIGAR* string includes evidence of deleted or inserted segments inside of individual reads as well as split read information. Depending on the evidence the algorithms attempt to classify the signature as one of the known *SV types*. The number of types depends on the caller. Sniffles2, the most frequently used SV caller based on the number of citations, and PBSV allow identifying 5 major SV types: Deletions (DEL), Insertions (INS), Inversions (INV), Translocations (BND), and (Tandem-)Duplications (DUP). SVIM additionally identifies interspersed Duplications. The frameworks designed to assign an SV type to a signature depend on a set of rules guiding the decision. For example, a read that is aligned to two separate regions mapped to different chromosomes indicates a Translocation. If the two regions are located on the same chromosome and next to each other but mapped in different orientations, the signature indicates an Inversion. Since the set of rules differs depending on the underlying decision framework, the SV types assigned to the signatures are unique to each caller. It should be noted that the variant calls of Translocation i.e. BNDs also refer to Breakends which are detected but unresolved SV signatures in the alignment data. This is the case for both PacBio and short-read callers.

In addition to differences in the assignment of SV types, there are inconsistencies in the output i.e. the VCF files produced by the three callers. VCF files include in addition to the chromosome and start position of an SV call several metrics relevant for the comparison of variants including the SV type, length, end position, mate breakpoints, alternative alleles, and variant IDs. While all callers report these metrics, their data type and definition vary across callers.

To address these inconsistencies between the call sets we implemented a custom script to generalize their format. With the formatting, we explicitly address the following differences between the callers: First, the method-specific assignments of SV types. Second, the deviations in the VCF format. Third, we convert all types of Duplications to Insertions while retaining the originally reported inserted/duplicated sequence for downstream analysis, if possible. For callers that do not return the duplicated sequence for Duplication calls by default, we retrieved the corresponding sequence from the GRCh38 reference genome using the python package `pysam` (v.0.19.0).

Filtering

Most long- and short-read callers employ several built-in filtering mechanics that attempt to reduce the number of *false-positive* SVs while retaining as many *true-positive* calls as possible. SVIM is an exception since it does not filter its output by default but insteads recommends adjusting the threshold on the computed quality score to reduce the number of *false-positives* in a post-processing step. PBSV requires an absolute minimum of 2 supporting reads by default and Sniffles aims to identify potential *false-positives* based on coverage with a default *minimum support multiplier* of 0.1. The number of reported SVs, therefore, also varies greatly between callers solely based on the applied filtering mechanisms.

To exclude biases during the comparison between callers introduced by the unique build-in filtering mechanisms, we designed a common set of filters and if possible adjusted the default calling parameters of each of the three callers accordingly: We restricted the reported SVs to those $\geqslant$ 50bp and supported by a number of reads $\geqslant$ 40% of the long-read coverage. In addition, we discarded all variants located in regions with suspiciously high coverage. These regions mainly contain highly repetitive sequences leading to misalignment and consequently an increased number of *false-positives* even when using long-reads. To determine such regions, we first computed the coverage of 10kb windows across the genome using `bedtools makewindows -w 10000`, `samtools bedcov` and applied custom formatting with `awk`. We then discarded all variants located in regions with at least 5 times the mean coverage.

4.2.2 *Illumina Processing*

Alignment

We applied `bwa-mem` (v0.7.17-r1188) to align the raw sequencing data to the GRCh38 reference genome, compressed the resulting file, and filled in mate coordinates and insert size:

```
bwa mem {reference} -t 16 -Y -v 3 -M -R '@RG\tID:{patient}\tSM:{
    patient}\tPL:ILLUMINA' {input.fq1} {input.fq2} | samtools
    view -b - | samtools fixmate -m - {output.bam}
```

Using `samtools sort` and `samtools markdup` we sorted and marked duplicated reads in the resulting BAM files. We then applied the `gatk BaseRecalibrator` v(4.2.6.1) and `gatk ApplyBQSR` with recommended settings including `dbSNP` and `Mills` reference variant sets for GRCh38. These were mainly included for the SNP and InDel calling and reproduction of the results of the initial WGS analysis. The results of this reproduction are not shown in this dissertation.

Variant Calling

Using the re-calibrated aligned reads as input, we generated individual call sets using Manta (v1.6.0), Delly2 (v0.8.3) and Lumpy (Smoove v0.2.3) with the following commands:

```
python2.7 configManta.py --callRegions grch38_contigs.bed.gz --
    bam {input.bam} --referenceFasta {reference} --runDir {
    working_dir}
python2.7 runWorkflow.py

delly call -q 20 -s 15 -x human.hg38.excl.ts -g {reference} -o {
    output.bcf} {input.bam}

smoove call -x --genotype --name {patient} --outdir {outdir} -f {
    reference} -p {threads} {input.bam}
```

As preparation for the comparison between callers and technologies, we then replaced the headers of all generated VCF files generalizing the format of the patient IDs and normalized the calls using `bcftools norm -c s -d all`.

Filtering & Formatting

In contrast to the PacBio long-read callers, the implementations of the short-read SV callers are much more diverse. A major contribution to these differences is the larger variety of read-based evidence provided by short-read sequencing data. In addition to the *split-read* (SR) evidence that is leveraged by long-read SV callers, short-read data also provides the information of discordant read pairs i.e. *paired-end* (PE) reads. While initial short-read SV callers leveraged only SR evidence, current methods use both types of evidence and some the additional *read-depth* (RD) information for CNV calling. In our pipeline, all callers mainly rely on SR and PE information. However, each of the SV callers uses a unique method to detect SV signatures from the aligned reads: Manta employs a graph-based approach applied to genome segments and performs additional *assembly* of reads supporting variant loci to improve the breakpoint accuracy. In its final processing step, Manta then classifies signatures as a specific SV type. Delly first extracts SV signatures from PE evidence and then refines them using SR evidence. Lumpy centers its SV detection around an abstract breakpoint definition, collecting SR, PE, RD, and, if provided, so-called *generic* evidence e.g. known variation in a population. Similar to the PacBio callers, the short-read SV callers also employ unique built-in filtering mechanisms. This results in significant performances difference even on well-studied cell lines with high coverage short-read data [94].

To account for this variability we applied common filtering steps: We only retain SVs $\geqslant$ 50bp and variants supported by a number of reads $\geqslant$ 40% of the short-read coverage. We also discarded all variants located in regions with suspiciously high coverage as previously described for the PacBio call sets. In addition, we set an upper limit on the SV size filtering all SVs $\geqslant$ 1mb. This additional filter is based on an inspection of the SV calls from Delly and Lumpy after coverage-based filters were applied that revealed a total of $5,704$ and $2,676$ SV calls $\geqslant$ 1mb, not supported by any other caller or long-read sequencing. Since these calls are highly likely to be false positives, predominately called due to misaligned reads in repeats regions, we excluded them from any further analysis. The unfiltered Delly and Lumpy calls with their corresponding size distribution are shown in Figure A.3.

Finally, we converted the filtered VCF files into a generalized format with a custom script allowing for a direct comparison between the Illumina and the PacBio callers. With this, we account for the inconsistencies between the caller outputs such as method-specific assignments of SV types, and general deviations in the VCF format e.g. variant identifiers as well as definitions of end positions and SV length. Since manta returns by default two break points for each Inversion, we also generated single entries for all Manta inversions using the `converInversion.py` script provided in Manta's GitHub repository `https://github.com/Illumina/manta`. We also converted the Duplications of all callers to Insertions while retaining the originally reported inserted/duplicated sequence. If the duplicated sequence was not provided by default, we extract it from the GRCh38 reference genome using the python package `pysam` (v.0.19.0).

4.2.3 *SV Merging Approach*

To merge SV calls across callers and technologies we implemented a custom clustering approach on the basis of *nested containment lists* with the python package `ncls`. Nested containment lists are data structures specifically developed for interval overlap queries [95]. During our approach, we aim to cluster SVs of the same type that potentially represent a single event. The clustering itself is controlled by SV type-specific matching criteria. The criteria for SVs that can be represented as an *interval* i.e. Deletions, Inversions, and Insertions are based on reciprocal overlap. For Insertions we choose their reference position as starting locus and add their length to compute an artificial end position, allowing us to represent them as intervals. The criteria for Translocations are based on distance metrics for both the first and the second i.e. the *mate breakpoint*.

Matching pairs are determined as follows: An *interval* SV x and an SV of the same type y match if there is a reciprocal overlap $\geqslant 50\%$. If both x and y are smaller than 5 kb this threshold is reduced to 10%. Two Translocations are matching if the distance between the first breakpoints is $\leqslant 100$bp and the distance between the two mate breakpoints is also $\leqslant 100$bp. For each resulting SV cluster, the algorithm returns a *representative*. We aimed to define representatives including as much of the potentially affected locus as possible since this will prevent any missed affected regulatory annotation during prioritization. Deletions and Inversions are represented as the maximum interval spanning from the leftmost locus and rightmost locus of all SVs in the cluster. For Insertions, the algorithm returns the leftmost reference locus and the rightmost reference locus without the added insertion sequence length. The Insertion representative also includes the insertion sequence of the initial SV in the cluster. Translocation clusters are represented by the leftmost reference locus of all first breakpoints in the cluster and a list of all mate breakpoints as alternative alleles. Our approach also retains the original identifiers of the SVs in the clusters of all representatives to allow backtracking in downstream analysis.

4.2.4 *Comparison with Catalogs of Common Variation*

To assess the allele frequencies of our call-set with respect to publicly available catalogs of common SVs, we combined multiple data sources (Table 3): A comprehensive collection of SVs from the *GnomAD* consortium including $308,858$ SVs based on $10,847$ unrelated individuals sequenced with paired-end short-read WGS [86], $96,585$ SVs from 15 individuals sequenced with Pacbio long-read sequencing [44]. $111,746$ SVs from 64 individuals deeply sequenced using PacBio CCS [45] and finally a curated set of common SVs $82,288$ from multiple studies from NCBI (nstd186) [96]. All calls were either originally called with respect to the GRCh38 reference or lifted over [97]. We filtered each set individually retaining common variation i.e. SVs with AF$\leqslant 0.01$. For the two long-read data sets we discarded all variants unique to a single individual rather than applying an AF-based filter due to their comparably low cohort size. Additionally, we excluded all SVs < 50bp. We also discarded SVs classified as *CNVs* i.e. unclassified changes in copy-number from the GnomAD catalogs since they could not be compared to the SV types in our cohort call set. Finally, we formatted all public SV catalogs according to the criteria used for both our short-read and long-read call sets. To identify common variation in our cohort, we then merged the SVs detected by all six callers with the call sets of common SVs using the same set of parameters as for the comparison of callers and sequencing technologies. This allowed us to identify and discard any cluster of SVs including at least one variant present in the common SV catalogs.

4.2.5 *RNA-seq Analysis*

The preparation of experimental data was performed by Uirá Souto Melo and the sequencing by the sequencing facility of the MPIMG.. The fibroblast samples of the 21 patients were cultured in DMEM with 10% FBS, 1% L-glutamine, and 1% pen-strep. RNA extraction from fibroblasts was performed in all 21 samples using the RNeasy mini kit (Qiagen, Hilden, Germany). The Poly(A) mRNA capture was done using the KAPA mRNA HyperPrep Kit (KR1352 -v5.17) and sequencing was performed on a HiSeq4000 (Illumina) using a single technical replicate (PE75, 50 million fragments per sample).

We processed the raw sequencing data with a custom *snakemake* pipeline: First, we aligned the reads to the GRCh38 reference using STAR (v.2.7.9a) [98] with the following command:

```
STAR --runMode alignReads --alignIntronMax 1 --readFilesCommand
    zcat --bamRemoveDuplicatesType UniqueIdentical
```

We filtered for reads with a minimum *mapping quality* (MAPQ) of 5. To identify DEGs for each patient we conducted a *one vs. all* analysis with respect to the *UCSC hg38 knownGene* library using a custom R script. Briefly, we removed any data mapped to sex chromosomes and counted the reads of each patient per exon with `summarizeOverlaps`. We then applied the *DESeq2* (v1.26) function `DESeq` computing the difference between read counts of a single sample and the remaining cohort [99]. Finally, we normalized the results using `vst`, computed the logarithmic *Fold-Change* (Log2FC) per gene, and returned the top 50 genes with the highest Log2FC as the list of DEGs for each patient. It should be noted that the low number of replicates likely influences the accuracy of our one vs. all analysis. In addition, the expression of the genes in fibroblast samples does not directly correspond to their expression during limb development which potentially limits the application of our RNAseq results in the prioritization.

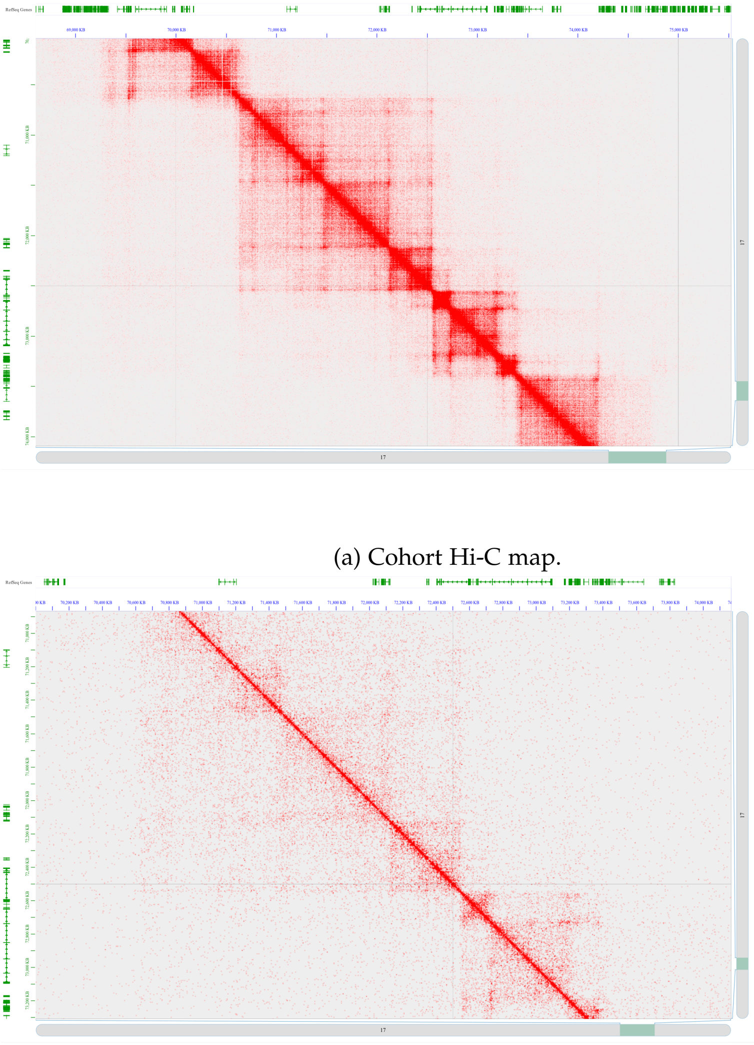
4.2.6 *Hi-C Analysis*

The experimental preparation of the samples and the sequencing were performed by Uirá Souto Melo and sequencing facility at the MPIMG, respectivley. PCR amplification (4–8 cycles) was done using NEBNext Ultra II Q5 Master Mix (New England BioLabs, M0544). The PCR purification and size selection was carried out using Agencourt AMPure XP beads (Beckman Coulter, A63881). Libraries were sequenced in 75bp, or 100bp paired-end runs on a NovaSeq6000 (Illumina) using between 2 and 5 technical replicates.

We conducted the bioinformatic analysis in a custom snakemake pipeline. First, we processed the raw sequencing data using the *Juicer* pipeline (v.1.6) [100] with *bwa* (v0.7.17) for aligning the reads to the GRCh38 reference. We then merged the deduplicated and filtered reads across technical replicates. For each patient, we applied the `pre` function from *Juicer-Tools* (v1.22.01) to generate Hi-C maps including all mapped reads with MAPQ$\geqslant$ 30. In addition, we merged the deduplicated reads across the entire cohort creating a *high-resolution fibroblast Hi-C map*.

To determine TAD boundaries we first generated patient-specific *VC SQRT* normalized contact maps at 25kb resolution for each chromosome using the *Juicer-Tools dump* function and then applied *TopDom* [101]. Currently, no gold standard method for calling TAD boundaries has been determined and large-scale comparisons have observed high variability across callers and data sets [102]. The choice of TAD calling method was therefore not trivial. For our annotation, we required a non-hierarchical TAD caller to define non-overlapping regulatory windows across the genome and *TopDom* has been shown to produce robust results in comparison to other callers for this purpose. Still, the resulting TAD boundaries have been shown to be variable depending on the chosen parameters mainly the `window-size` [102].

To assess this variability within our own Hi-C data we executed TopDom with *window-size* parameters of 5, 10, 20, and 30 using the high-resolution fibroblast Hi-C map. This resulted in $9,356$, $7,042$, $5,801$ and $5,401$ genome-wide *TopDom* entries, respectively. To reflect the high variability between the number of TADs for different *window-sizes* we included two separate TAD sets in our annotation: We performed the initial TAD annotation with respect to the *window-size* 10 TADs and include the *window-size* 5 TADs as an additional non-coding annotation. Both sets are based on the merged cohort Hi-C map which provides more robust TAD calls than patient-specific Hi-C data due to its increased resolution. We illustrate the resolution difference between patient-specific Hi-C maps and the merged cohort map at the *SOX9* locus in Figure 4.3. Since the patient-specific information could still prove valuable for the interpretation of potential SV effects i.e. perturbations of regulatory interactions we included both types of Hi-C data in the visualization of candidate SVs.



(a) Cohort Hi-C map.

(b) Patient Hi-C map.

Figure 4.3: **Comparison of Patient and Cohort Hi-C Map**. The figure shows the resolution of **a)** the merged cohort Hi-C map and **b)** the Hi-C map of the LM01 patient at the *SOX9* locus. Refseq gene annotation are indicated in green.

4.2.7 *PLAC-seq Analysis*

The significant PLAC-seq interactions determined by Yu et al. [103] were not yet available during our analysis. Thus, we implemented custom scripts to reproduce their results in human tissues. First, we retrieved the raw contact information from 4DNA (ID: 4DNESBUE56SA) and lifted the contact pairs over to hg38 [104]. Given the high rate of lost contact pairs during the initial liftover process with a *MinMatch* setting of 0.95 we reduced the threshold to 0.70. We then determined significant *Peak2All* interactions using *FitHiChip* with the following parameters: *BINSIZE=10000 LowDistThr=1000000 BiasType=1 MergeInt=1 QVALUE=0.01*. The final merged set contained 10,351 significant interactions.

4.2.8 *TADA 2.0*

Since TADA was exclusively designed for CNVs, we needed to extend the original framework for the entire spectrum of SVs. Deletions and Duplications are represented as intervals in the annotation process. The inclusion of Inversions in the updated TADA version was therefore trivial. However, Translocations and Insertions require additional processing. In our final call set of limb malformation, many Insertions were originally called as Duplications to allow for a more permissive merging process. For the annotation process, we again assign Duplication labels to Insertions if this SV type is shared across all supporting callers. For template-Insertions, we reasoned that a potential contribution to their functional impact is represented in the inserted sequence. To incorporate this in our annotation process we first align the insertion sequences back to the reference genome (GRCh38) using the command line version of `blat` (v.37) [105]. To improve the run time for individual sequences we include a *.ooc*-file for the GRCh38 reference genome, then formatted each insertion sequence to resemble an entry in a *FASTA* file. We restricted the entries to the first 2kb of each insertion sequence to further reduce the alignment duration. For each sequence we executed blat with the following parameters: `-stepSize=5 -repMatch=2253 -minScore=30 -fastMap 0 -minIdentity=0 -noHead`. Given the output file, we computed an alignment score as the difference between matches and mismatches (including Inserts). For each Insertion sequence, we picked the highest-scoring alignment, extended the position by the initial length of the sequence, and returned the corresponding genomic position. During TADA's annotation process, we then regarded template-Insertions as an SV consisting of two parts: The locus of the Insertion itself ±500bp and the locus of the inserted sequence, if available. For novel sequence insertions, we restrict the representation to the Insertion locus ±500bp. For Translocations, we include both the first and the mate breakpoint ±500bp in the annotation.

4.2.9 *Manual Inspection of SVs*

We employed two methods to visualize and manually inspect SVs: 1) *Samplot* depictions of the SV loci using the aligned Illumina and PacBio reads [106] and 2) a custom script developed by Nico Alavi at the MPIMG allowing to inspect the read support of Insertions and Translocations using CIGAR string. Samplot can be applied to short- and long-read data but is limited to variants that can be represented as an interval spanning multiple base pairs. For Deletions and Duplications, Illumina complements the PacBio reads with additional visual RD evidence. We, therefore, generated *Samplot* figures including the coverage and read alignment of Illumina and PacBio data for all Deletions, Duplications, and Inversions contained in the set of functionally relevant SVs. The majority of the Translocations in the call-set of functionally relevant SVs are supported by Illumina callers only (91.78%) indicating a high rate of *false-positives* of this variant type likely called due to ambiguous read pair alignments. We attempt to reduce that rate by inspecting the long-read sequencing evidence at each breakpoint location. While we are aware that the calls were not directly supported by PacBio callers, we reason that individual long-read SR evidence should at least indicate the presence of *true-positive* calls, and the lack of any supporting evidence strongly suggest *false-positive* variant calls. In a similar fashion, we inspect long-reads at Insertion loci. Due to their increased length, PacBio-derived SR evidence at these breakpoints can better distinguish *true-positives* from *false-positive* Insertions. In summary, we collected aligned PacBio reads around Insertions and Translocations ±500bp and visualized the corresponding CIGAR strings. This allowed us to identify the concordance between split reads at the breakpoint loci.

In order to efficiently inspect the around 200 SVs for each patient, we implemented a custom web application that combines the visualization approaches and allows iterating through an SV call set. For each SV the application retrieves a visualization of the sequencing evidence depending on the variant type. The user then assigns *true-positives* and *false-positive* labels, generating a CSV file with annotated variant information including the manual inspection decision.

4.2.10 *Visualization of Candidate SVs*

We implemented a custom visualization approach that is based on the python package *coolbox* [107]. *Coolbox* allows to visualize genomic data in the majority of formats used in standard bioinformatic analyses including Hi-C data derived from Juicer analyses. It offers both *Juypter* notebook support and command-line capabilities. First, We slightly modified the original implementation to allow specific gene groups to be highlighted among visualized data from an Ensembl GTF-file. We then implemented a script that retrieves all annotations of the *Limb Regulome*. This includes two sets of Hi-C data - the fibroblast Hi-C map merged across all patients and a patient specific Hi-C map - as well as overlapping coding and non-coding annotations. Finally, we iterate over candidate variants, producing visualizations of the affected loci ±400kb. For each patient, we generate a final report including all visualization in *PDF*-format.

5 DETECTION

The first step towards a successful molecular diagnosis and identifying the corresponding disease-causing mechanisms is the accurate detection of variation from sequencing data. To maximize the SV detection sensitivity we combine short- and long-read sequencing in the analysis of the limb malformation cohort. Figure 5.1 shows an overview of the variant detection pipeline involved in this part of the project. We implemented two separate workflows to process short- and long-read sequencing, call SVs and filter them.

In the following sections, we will present the results generated by our pipeline for each technology and conduct a comparison across the corresponding callers. With this, we aim to identify the unique properties i.e. potential advantages and disadvantages of the methods with respect to their ability to detect SVs, and demonstrate the increased detection sensitivity of our *ensemble approach* i.e. the combination of multiple SV callers. We then conduct an analysis of the call set merged across short- and long-read data, allowing us to highlight significant differences between the technologies. In the third section of this chapter, we discuss the results of the filtering approaches we use to reduce the merged call set to the rare and potentially pathogenic variants based on catalogs of common variation and the allele frequency of SVs in our own cohort. Finally, we present the results of the entire variant detection pipeline - the reduced set of rare variant calls that serves as the basis for our prioritization of potentially pathogenic SVs.

5.1 PACBIO

The PacBio CLR sequencing data for each of our 21 patients is based on cultured fibroblast cells. Since the data has not yet been published we briefly describe the experimental setup here. All experimental work was performed by Uirá Souto Melo and the Seqcore facility at the MPIMG. The high molecular weight (HMW) DNA was extracted from the fibroblasts with a smart DNA prep kit (Analytik Jena). Quality control (QC) steps were performed on the DNF-467 Genomic DNA 50kb Analysis Kit using a 5200 Fragment Analyzer system (Agilent). For library preparation, the DNA was sonicated using the Megaruptor 3 shearing kit and the Megaruptor 3 instrument (Diagenode; parameters 20µg, HMW-DNA; Speed: 3). The corresponding QC was performed with the DNF-464 High Sensitiv-

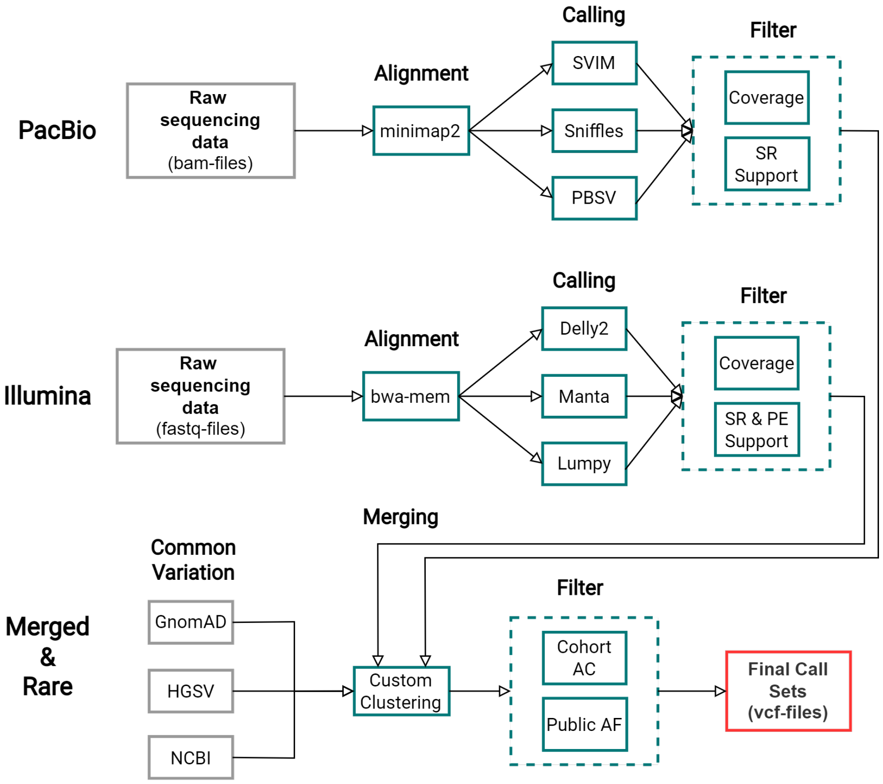


Figure 5.1: **The SV Detection Pipeline.** The detection pipeline includes two workflows: One for Illumina short-read sequencing and one for PacBio long-read sequencing. After the alignment of the sequencing data to a reference genome, we employ three callers for each technology to detect SVs. The call sets then undergo formatting, quality control, and filtering steps designed specifically for this combination of SV callers. Finally, we combine the filtered SVs of the technologies across all six callers into a single call set comparing the merged variants with public databases and computing cohort-wide allele frequencies to determine and ultimately discard common variation.

ity Large Fragment 50kb kit and size selection using the BluePippin Size-Selection System (Sage Science) with range selection mode (*BPstart*= 30kb; *BPEnd*= 80kb) and a library input of 3–5µg. All sequencing was done with Sequel II systems. The majority of patients were sequenced on a single SMRT cell (16 patients) with some exceptions: LM01 (8 cells), LM10 (2 cells), LM11 (5 cells), LM12 (5 cells), LM14 (5 cells), LM15 (3 cells) and LM20 (7 cells).

5.1.1 *PacBio Alignment*

To investigate any potential issues during alignment or sequencing we first set out to analyze the quality of our aligned reads. This analysis included the coverage, read length, within-read align ability, nucleotide distribution, and GC content. The coverage and read length are shown in Figure 5.2. The results of the remaining alignment analysis are shown in the appendix (Figure A.2).

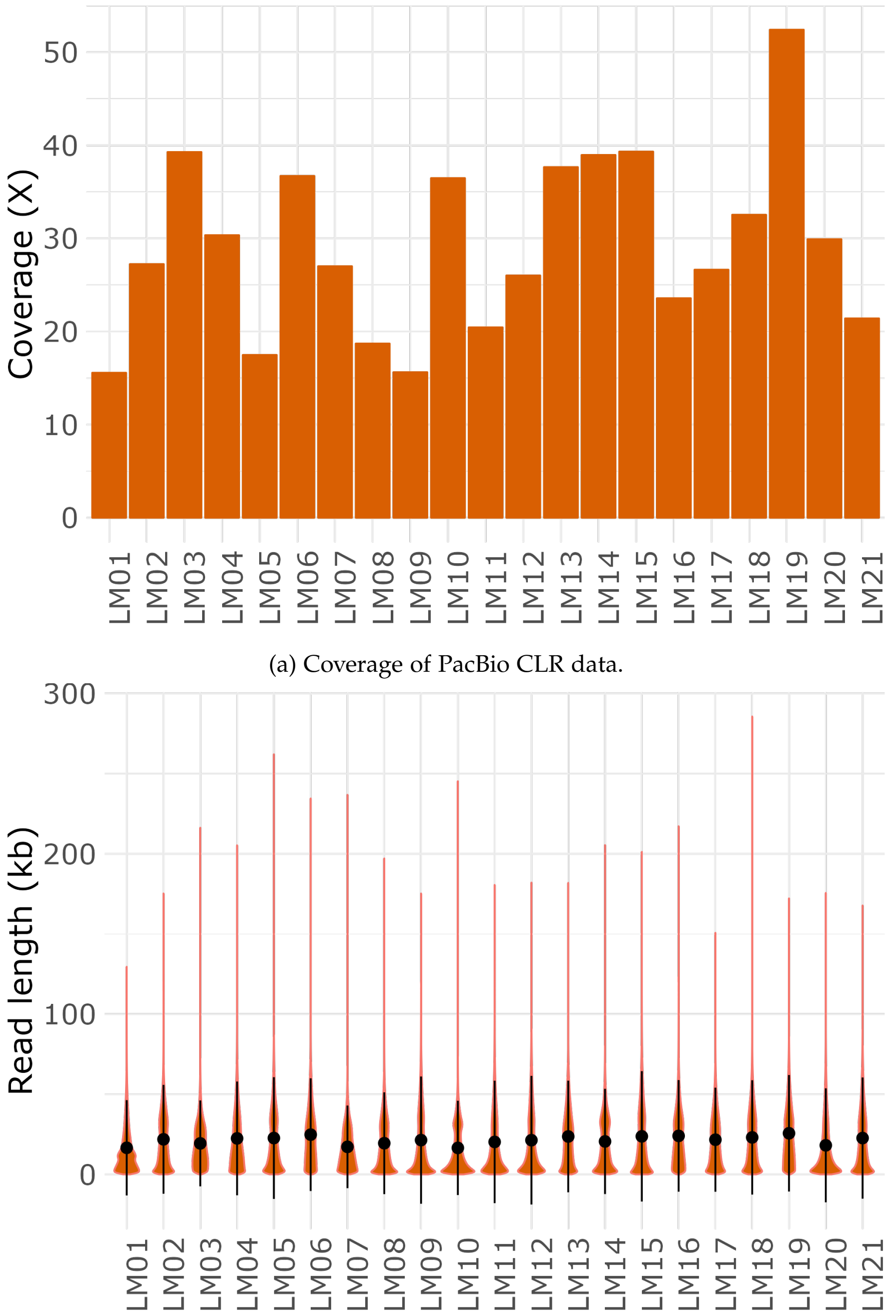
The number of aligned bases varies considerably across patients (between 48,093mb for LM01 and 121,467mb for LM15) as does the number of aligned reads (between 2,596,664 for LM05 and 7,528,967 for LM03). To achieve reasonable sensitivity in SV calling we required at least 15X coverage in each of our samples. Coverages beyond this threshold only marginally increase the performance of the callers in our pipeline [40]. The computed coverages for the patients in our cohort are between 15.6X and 52.4X with a cohort-wide average of 29.21X, satisfying the minimal threshold of 15X.

The mean length of aligned reads also differs signficantly across the patients (between 16,580bp for LM10 and 24,850bp for LM06) with a cohort-wide average of 21,342bp. Most reads are smaller than 40kb indicated by the highest observed third quartile of 35.14kb. Unaligned reads are generally shorter than aligned reads (cohort-wide average 5.43kb). For all patients, we can also observe outliers of aligned reads $\geqslant$ 100kb with a maximum observed length of 262kb (LM05). The cohort-wide mean of within-read align ability is 85.86% indicating that the majority of reads are almost fully aligned. Based on the measured read length, coverage, and within read-align ability the assessment of the aligned data indicates successful sequencing runs that should allow detecting a wide range of SVs.

5.1.2 *PacBio SV Calling*

To maximize the sensitivity during variant calling we employed three long-read SV callers: SVIM, Sniffles2 and PBSV [39–41]. We adjusted the calling parameters to reflect our thresholds for read support (number of supporting reads $\geqslant$ 40% *coverage*) and SV size ($\geqslant$ 50bp). We then formatted the filtered call sets to allow for a direct comparison between callers addressing the inconsistencies of the three SV callers in terms of SV type assignment and VCF-file output (see Methods for details). Finally, we visualized the filtered and formatted call sets and size distributions for SVIM, Sniffles2, and PBSV as shown in Figure 5.3. The visualizations allow us to highlight the general properties of the call sets and unique differences between callers which we discuss in the following paragraph.

The majority of identified SVs (mean across callers and samples) are Insertions (59.8%) followed by Deletions (38.9%), a small proportion of Inversions (0.5%) and Translocations (0.7%). The proportion of variant types remains robust across callers with the exception of an increased number of Inversions identified by Sniffles2 in a subset of



(a) Coverage of PacBio CLR data.

(b) Read Length Distribution of PacBio CLR data

Figure 5.2: **Alignment Statistics of the PacBio data.a)** shows the coverage (X) of each patient and **b)** the read length distribution in kb. Whiskers indicate the 25th and 75th percentile and dots the median.

patients. While the mean number of SV calls is similar for all three callers: 16, 279 (SVIM), 17, 199 (Sniffles2), and 16, 254 (PBSV), there are considerable differences across patients (up to 6, 548 SVs) and when comparing individual patients across callers (up to 2, 927 SVs).

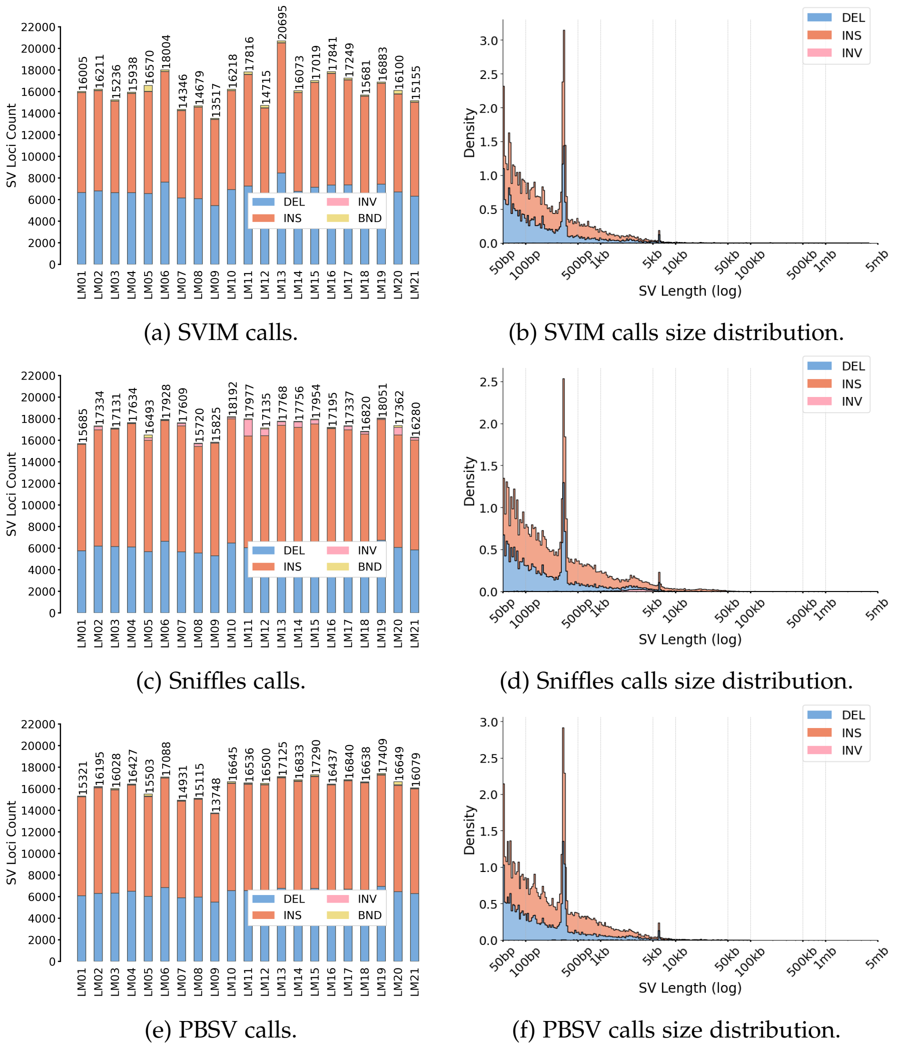
Most identified variants are between 50 and 500bp ($> 75\%$ for all callers). We can observe two prominent peaks in the size distribution of all callers at 300bp and 1kb that also have been described in previous studies [44, 45]. The peaks correspond to groups of transposable elements - the first at 300bp to *ALU*- and the second at 6kb to *LINE1*-elements. With increasing size, the number of variants decreases drastically for all callers.

Sniffles identifies a higher number of large SVs ($\geqslant$ 100kb). This has a significant influence on the cumulative length of the call set. While the call sets of SVIM (183mb) and PBSV (175mb) are of similar length, SVs identified by Sniffles span a total of 976mb. This is largely due to a set of 185 large ($\geqslant$ 100 kb) SVs unique to the Sniffles call set with a cumulative length of 620 mb. The majority of the large SVs are Inversions (100 out of 185) with a maximum length of 122 mb. SVIM, in comparison, identifies a total of 76 SVs $\geqslant$ 100 kb - none of which are Inversions - with a cumulative length of 2.3 mb.

This subgroup of unique Sniffle calls is an example of the implementation differences between callers during the SV-type assignment that directly influence the generated call sets. SVIM uses a threshold (default 100 kb) as an upper bound for SVs to be classified as Inversions or Deletions indicated by SR evidence. All SVs larger than the set threshold are classified as Translocations. Sniffles, however, does not apply such a distinction leading to the inflated number of large Inversions seen in our comparison. The underlying cause for the signatures of these large events is likely the presence of repeat elements located at multiple loci in the same chromosome. Segments of SRs can be aligned to two locations of the same repeat e.g. the left side of the first repeat location and the right side of the second location. Thus, indicating a Deletion or Inversion depending on the read orientation between the two loci. While these calls are highly likely to be false-positive, there is no other signal than the length to distinguish them from true-positive SVs. Thus, we retain these unique variants in our variant detection pipeline to avoid discarding any potential true-positive SV calls.

5.1.3 *Custom SV Merging*

A major part of this thesis revolves around the comparison of SV calls from different callers and technologies. To analyze and visualize the agreement between call sets systematically we needed to merge all corresponding SVs into a single set and assess the amount of shared identified variation. Since this merging process is not only required for the PacBio analysis but for all downstream comparisons between callers, technologies, and public databases, the choice of the under-



(a) SVIM calls.

(b) SVIM calls size distribution.

(c) Sniffles calls.

(d) Sniffles calls size distribution.

(e) PBSV calls.

(f) PBSV calls size distribution.

Figure 5.3: **PacBio SV Calls and Size Distributions.** The left-side figures show the number of SV calls grouped by SV type for SVIM, Sniffles and PBSV, respectively. Variant types are Insertions (INS), Deletions (DEL), Translocation (BNDs) and Inversions (INV). The entries on the X-axis are sample identifiers and the Y-Axis shows the number of SV Loci. The total number of SVs for each patient is shown on top of each bar. The right-side figures show the corresponding size distribution again grouped by SV type for each caller.

lying matching algorithm and criteria is not trivial. The majority of current large-scale sequencing experiments compute the overlap or distance between variants depending on the SV type and match two individual SVs if specific reciprocal overlap or distance criteria are met [44, 45, 86]. These thresholds on overlap or distance directly control the trade-off between *sensitivity* and *precision*. Since no standard has yet been established, the number of SVs in a sample can therefore change considerably between studies even though the initial call set might be the same.

Several more sophisticated methods have been proposed in recent years that allow matching SVs across callers or genotype SVs based on the support in individual sequencing technologies [108–110]. However, the algorithms are either designed for a specific set of callers or would require extensive adjustment to support the callers and technologies we use in our pipeline. We, therefore, implemented a custom approach that allows merging variants across callers and technologies. Briefly, the approach clusters SVs of the same type if type-specific matching criteria are fulfilled and returns a single representative for each cluster. A detailed description of the approach and the matching criteria is provided in Chapter 4.

5.1.4 *Comparison of PacBio Callers*

Using our custom merging approach we generated a combined PacBio call set for each of our patients. With this, we aim to quantify the previously observed differences between callers further highlighting the importance of combining individual callers to increase sensitivity. We computed the number of variants grouped by SV type identified by a single, two, or all three callers. The results are shown in Figure 5.4. Of the variants in our cohort-wide call set 62.97% are shared by all three callers. This proportion changes significantly when testing for individual SV types. For Deletions and Inversions, the proportion of shared variation across callers is 70.02% and 61.2%, respectively. Roughly 4.8% of Inversions are identified by all callers and 4.63% of Translocations are shared. We can clearly observe the previously reported increased proportion of Inversions uniquely identified by Sniffles2 in Figure 5.4d. In addition, we observe significant differences in the number of variants shared by two callers depending on the SV type: Sniffles2 and PBSV identify 12.19% Insertions that are not detected by SVIM, a proportion more than three times the number of Deletions identified only by SVIM and PBSV. However, for Translocations, the shared proportion of variants between Sniffles2 and SVIM (16.53%) is considerably higher than the number of variants uniquely identified by PBSV and Sniffles.

While the PacBio callers share a considerable proportion of detected SVs, many calls are unique to single callers. This highlights the need to employ multiple SV callers to increase sensitivity even for long-read sequencing data. However, variants supported by a single caller in many cases could prove to be false positives. A hard threshold on caller support would likely significantly reduce the number of false positives. However, we reason that any hard thresholds discard potentially true positives as well and can even increase the bias introduced by the shared evidence for SV detection between callers. Thus, we retain all calls identified by a single caller at this stage of the pipeline.

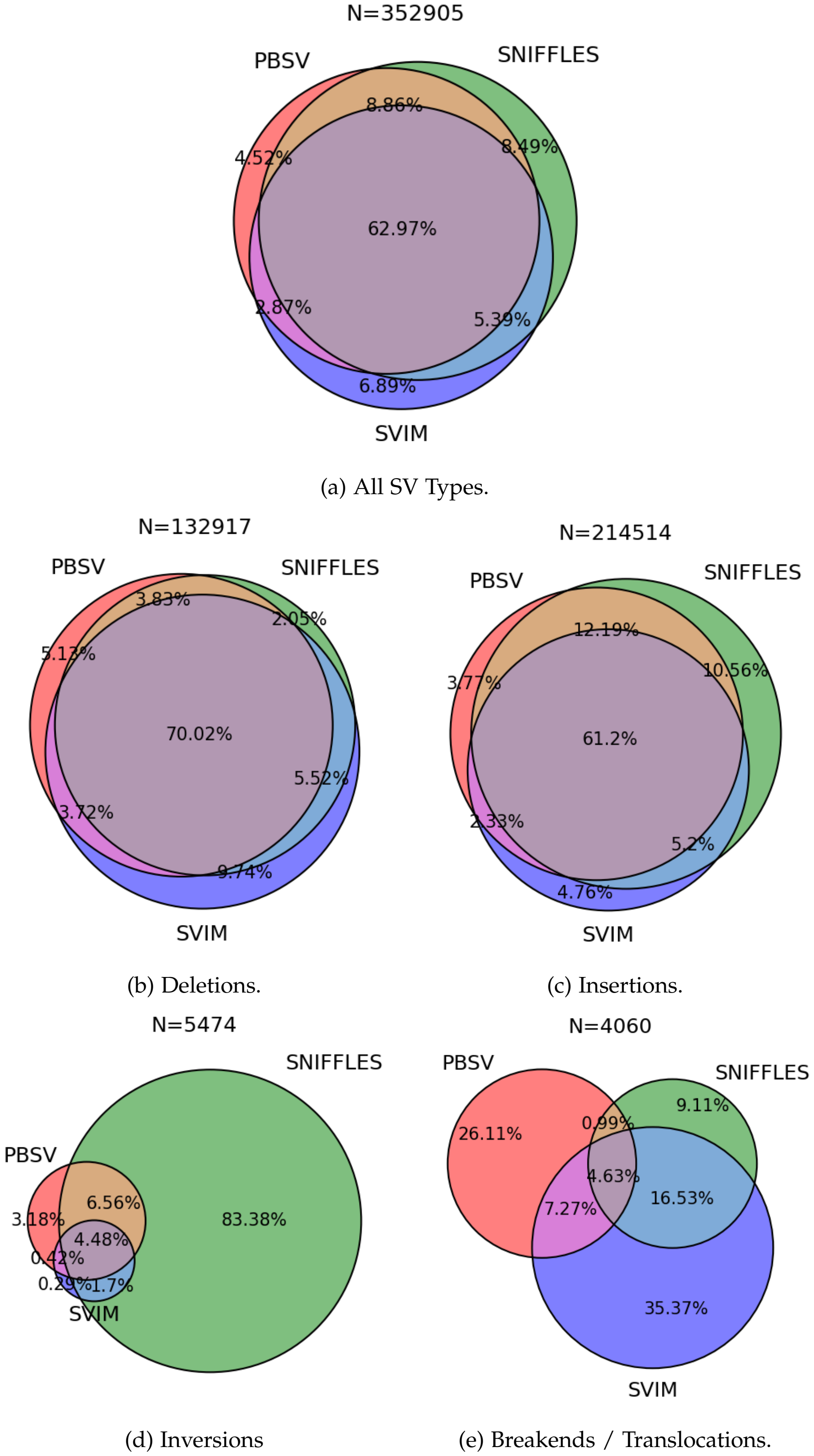
5.2 ILLUMINA

The WGS data has already been investigated for disease-causing variation in a previous analysis by Elsner et al. [83]. With the exception of two cases, all cases remain without a molecular diagnosis. Since the analysis was mainly focused on SNVs/InDels and coding SVs $\geqslant 1500$ bp, a more thorough analysis of SVs could reveal the disease-causing variation in the unsolved cases. While we expect the majority of SVs to be detected by PacBio, the information in the short reads can serve as additional evidence supporting individual SV calls and in some cases allows for a more accurate determination of breakpoint positions due to the low base-wise error rate. We therefore also include the short-read Illumina data in our analysis. We implemented several novel processing steps as part of our pipeline, updating callers and adding alternative methods not used in the Elsner et al. analysis as well as implementing custom SV filtering (see Methods for details). While we had access to the initially processed data, our novel pipeline allowed us to conduct a comparison between callers and technologies previous to any significant post-processing. The underlying raw sequencing data and the experimental processing, however, are the same as in the initial WGS analysis and are described in the corresponding publication [83].

5.2.1 *Alignment*

First, we investigated the quality of the raw reads using *fastqc* [111]. Fastqc allows the assessment of a variety of quality metrics for raw fastq files including sequence quality and length, per-base quality, over-represented sequences, and sequence Duplication levels. The report produced for the samples in our cohort showed no indication of irregularities based on the built-in quality control mechanisms. An example is that the Phred-scaled base quality does not drop below 20.

After the quality assessment, we aligned the reads for each patient to the GRCh38 reference using *bwa-mem* (see Methods for details) [112]. We then computed as described for the PacBio data the coverage of the aligned reads to ensure a successful downstream analysis. Figure 5.5 shows the resulting coverage values for all patients which varies between 26X and 43X with a cohort-wide average of 35X. In comparison to the PacBio data, the Illumina coverage is considerably more stable across patients. This is likely due to the more established experimental protocol which allowed sequencing each patient in a single run.



(a) All SV Types.

(b) Deletions.

(c) Insertions.

(d) Inversions

(e) Breakends / Translocations.

Figure 5.4: **Comparison Between PacBio long-read SV Callers.** The total number of SVs after merging for each type is shown on top of the individual Venn diagrams. The size of a circle is scaled by the number of variants identified by the corresponding caller.

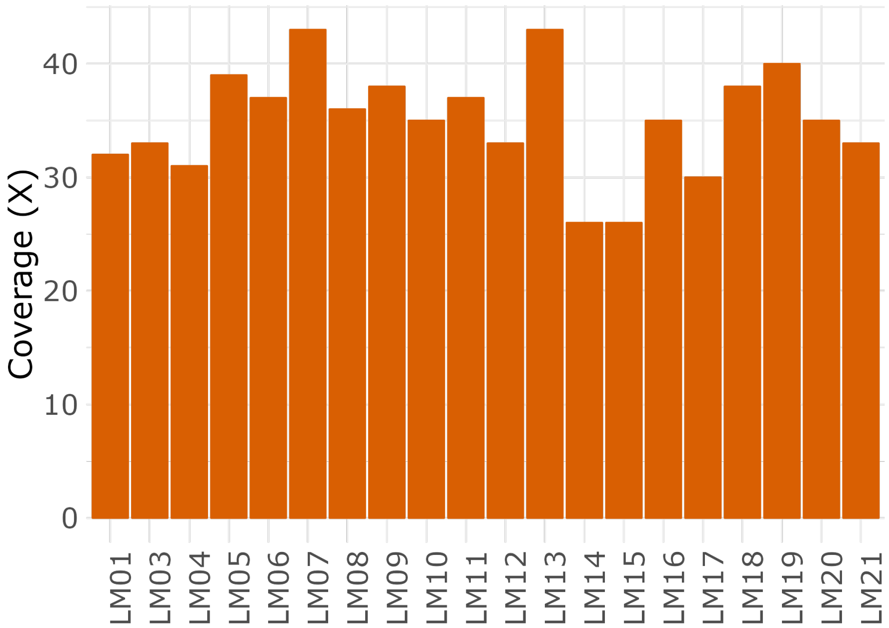
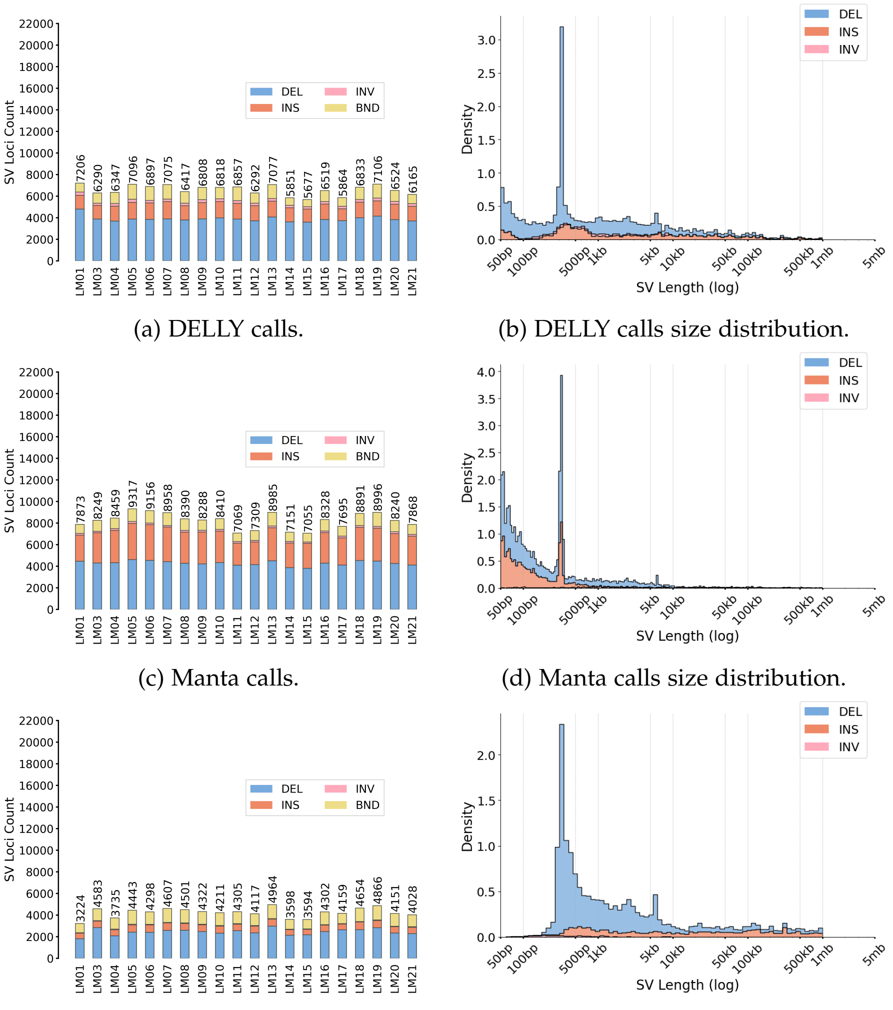


Figure 5.5: **Coverage of the Illumina WGS Data.** The Y-Axis shows the total number of aligned bases divided by the number of all mappable bases in the GRCh38 reference genome for each patient on the X-Axis.

5.2.2 *SV Calling*

We employ three SV callers for the short-read data in our pipeline to maximize the detection sensitivity: 1) Delly [35], 2) Lumpy [36] and 3) Manta [113]. Both Lumpy and Manta have not been applied in the analysis by Elsner et al. and could provide previously undetected SV calls potentially including a functionally relevant candidate variant. We processed the SV calls in a similar fashion as the PacBio call sets: First, we filtered all calls based on read support (min. number of support reads ⩾ 40%) and SV size (⩾ 50bp). Then we formatted all VCF files accounting for the inconsistencies between the output of individual callers. Details on the filtering and formatting process are provided in Chapter 4. Finally, we visualized the number of filtered and formatted calls grouped by SV type and the corresponding size distributions as shown in Figure 5.6. The visualizations allow us to identify the unique properties of the SV callers which we present and discuss in the following paragraph.

The majority of SVs identified from the short-read data are Deletions with an average across patients and callers of 56.32%. For Delly and Manta, Insertions are the second most frequently identified SV type with an average of 27.43% followed by Translocations (14.14%) and In-



(a) DELLY calls.

(b) DELLY calls size distribution.

(c) Manta calls.

(d) Manta calls size distribution.

(e) Lumpy calls.

(f) Lumpy calls size distribution.

Figure 5.6: **Illumina SV Calls and Size Distributions.** The figures on the left side show the number of SV calls grouped by SV type for DELLY, Manta and Lumpy, respectively. The variant types are the same as for the PacBio data: Insertions (INS), Deletions (DEL), Translocations (BNDs) and Inversion (INV). The X-axes show the sample IDs and the Y-Axis the number of SVs with the total number indicated on the top of each bar. The figures on the right show the corresponding size distribution of SV calls grouped by SV type for each caller. In these figures the X-axis are also sample IDs and the Y-Axis shows the number of SVs. Each SV type is plotted individually.

versions (2.81%). Lumpy identifies a higher proportion of Translocations (26.64%) than Insertions (14.47%) but a comparable proportion of Inversions (1.15%). The total number of SVs varies significantly across callers with a cohort-wide mean of 6,586, 8,234, and 4,233 for Delly, Manta, and Lumpy, respectively. The maximum difference between patients for individual callers is 1,941 SVs in the Manta Call set between patients LM15 and LM19.

Similar to the size distributions of the PacBio call sets we can observe a peak at $\sim$ 300bp corresponding to *ALU* elements and a peak at $\sim$ 6kb corresponding to *LINE1* elements. The individual size distributions of the short-read callers, however, vary significantly. While the majority of Delly SV calls are small variants $\leqslant$ 500bp (50.07%), there is still a high number of large SV calls $\geqslant$ 100kb (3,899 out of 105,231) and SVs $\geqslant$ 5kb (21.94%) even after discarding any SVs $\geqslant$ 1mb (see Methods for details). In comparison, in Manta's call set 82.25% of the SVs are $\leqslant$ 500bp and 5.00% $\geqslant$ 5kb. We observe the strongest deviation from any previously shown size distributions for Lumpy. The majority of Lumpy calls are $>$ 500 bp (63,50%). This is reflected in the cumulative length of the call sets which span 1,634 mb, 587 mb, and 2,461 mb for Delly, Manta, and Lumpy calls, respectively. The Lumpy call set, therefore, has a cumulative length of more than 4.2 times higher than Manta while reporting on average 1.95 times fewer variants.

Overall, our analysis of the filtered and formatted call sets generated by the short-read callers indicate significant differences in the proportion of SV types and size distributions. Several of these differences can be directly linked to unique processing steps of the algorithms. An example is the increased proportion of small SVs, especially Insertions, reported by Manta which uses both PE and SR evidence to build a signature graph while Delly relies on signatures initially only identified with PE evidence and refines them with SR evidence. This limits its potential to identify smaller SVs. In addition, Manta performs an assembly of reads at breakpoints allowing to more accurately detect SVs larger than the read size and sequence resolve them resulting in an increased number of Insertions around 300 bp not observed in the call sets of the other short-read SV callers. While the significant lack of small variation reported by Lumpy can not directly be linked to the model itself since it should detect SVs from SR, PE, and RD evidence with equal weight, a previous comprehensive comparison of short-read SV callers reports a similar lack of sensitivity for smaller SVs [94]. This indicates an issue in the implementation that requires further analysis.

5.2.3 *Comparison of Callers*

To quantify the differences between the short-read callers and assess their contribution towards a comprehensive call set, we merged the SVs for each patient with the previously described matching criteria into a single short-read call set and computed the proportion of shared variation. The resulting call set includes a total of 177,335 SVs with an average number of 8,867 SVs per patient. The proportion of shared variation is shown in Figure 5.7.

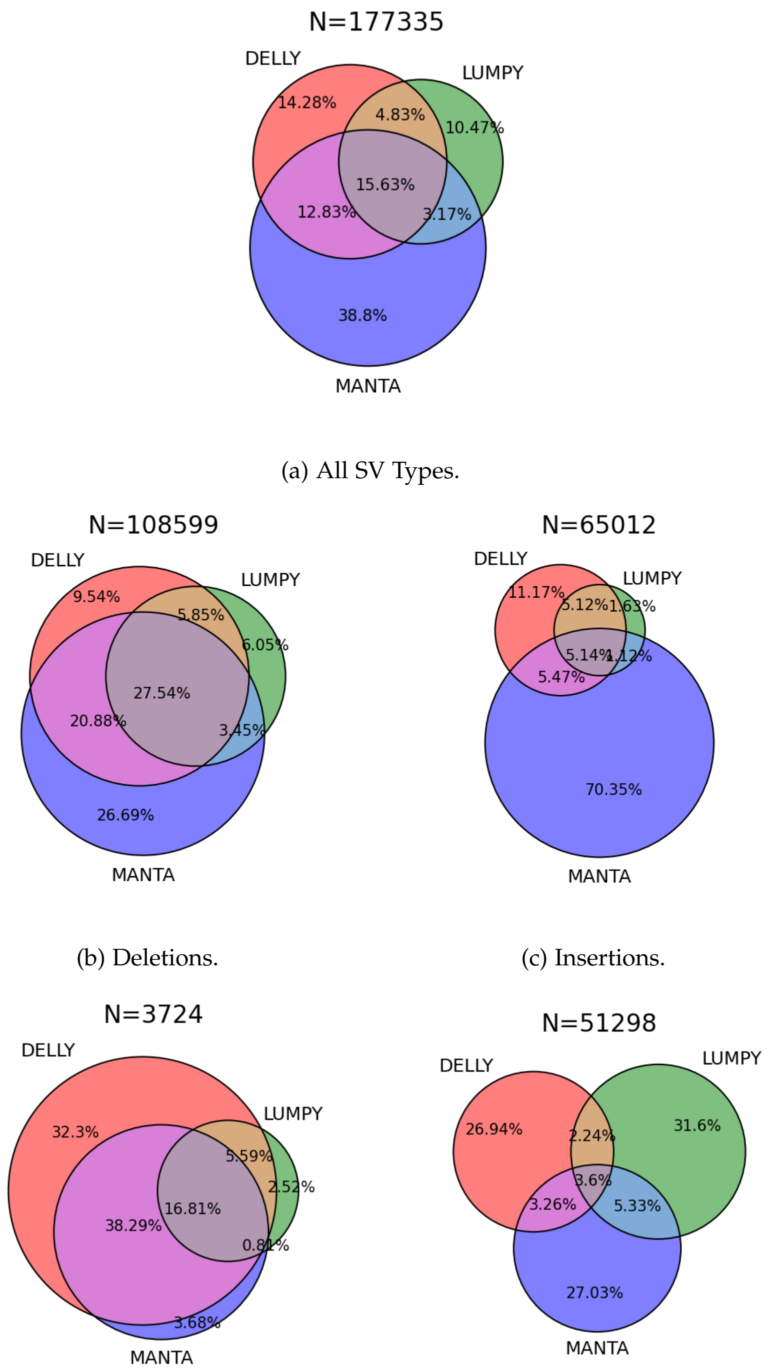
Of the merged SV calls 15.63% are identified by all three callers. This proportion varies depending on the SV type with a higher proportion of shared Deletions (27.54%) and Inversions (16.81%) than Insertions and Translocations (5.14% and 3.60%, respectively). Particularly Manta identifies a large proportion of unique Insertions (70.35%) and Deletions (26.69%). The vast majority of these unique Insertions are small variants ($98,71 \leqslant 500 bp$) also indicated in the differences of the previously presented SV size distributions. Lumpy identifies the smallest proportion of unique SVs across all variant types (10.44%) followed by Delly which uniquely identifies 14.28% of the merged call set. The proportions of shared variation between two callers are higher for Manta and Delly with 20.88% Deletions and 38.29% Inversions than for Lumpy in combination with any of the other two callers (less than 6% for all variant types).

The analysis shows that the overlap between the call sets of the Illumina short-read callers is significantly less than between the long-read SV callers. Thus, the combination of multiple callers is necessary to achieve high detection sensitivity allowing for a more thorough analysis of SVs in the limb malformation patients. In addition, many of the detected variants - especially those < 1500 bp have not been part of the initial analysis conducted by Elsner et. al and could represent functionally relevant candidate variants.

5.3 technology comparison

Few cohorts have been investigated using both long-read and short-read sequencing with the aim to identify potentially disease-causing variation. Even in current public reference databases of long-read sequencing data, the SV catalogs are based on less than 100 samples [44, 45]. We, therefore, have a unique opportunity to analyze the agreement between the SV calls from the two sequencing technologies in a larger cohort and in a setting similar to potential future clinical practice. Other studies have previously performed comparisons using gold-standard SV call-sets consisting of validated variation such as *Genome-in-a-Bottle* (GIAB) or deeply sequenced trio data to derive technology-specific performance metrics [114, 115]. Since we lack such a set of validated SVs for our cohort, we conduct a more explorative analysis (Figure 5.8). With this, we aim to highlight the benefits of combining multiple technologies and callers to maximize the SV detection sensitivity.

We base the comparison on a call set of SVs merged across sequencing technologies and the corresponding callers using the previously



(a) All SV Types.

(b) Deletions.

(c) Insertions.

(d) Inversions

(e) Breakends / Translocations.

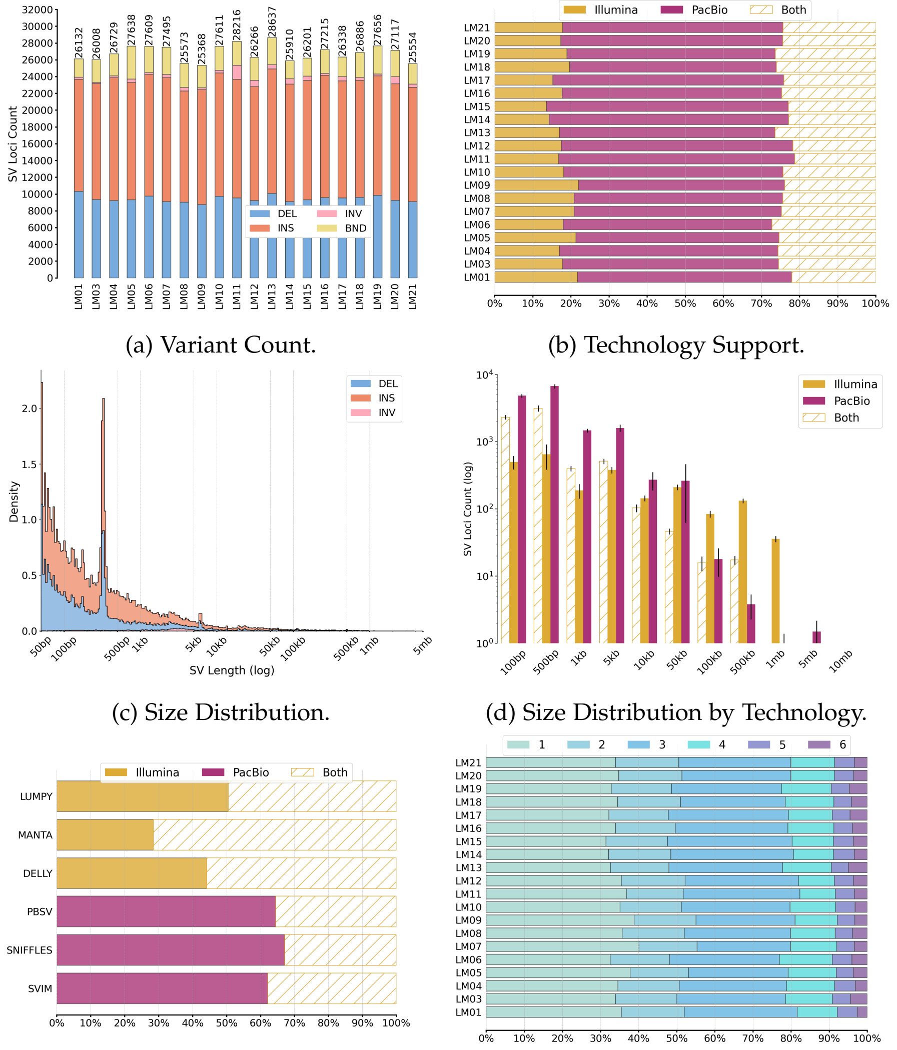
Figure 5.7: **Comparison Between Illumina short-read SV Callers.** The total number of SVs after merging for each SV type is shown on top of the individual Venn diagrams. The size of the circles is scaled by the number of variants of the SV type identified by the corresponding caller.

described merging strategy. The total number of merged variants stratified by SV type and size distributions are shown in (Figure 5.8a, Figure 5.8c). Our merging approach results in a call set of 536, 159 variants with a mean number of 25, 808 SVs per patient. The majority of SVs are Insertions (52.56%) followed by Deletions (35.24%), Translocations (10.48%), and Inversions (1.72%). The proportions of SV types remain generally stable across patients with a *standard deviation* (SD) for all SV types < 2.00%. There are, however, some exceptions e.g. the increased proportion of Inversions for the LM11 patient.

To assess the contribution of the technologies to the merged call set, we computed the proportion of variants identified by a single or both technologies (Figure 5.8b), the corresponding size distributions (Figure 5.8d), caller support (Figure 5.8f) and the proportion of SVs identified by a short- or long-read caller also detected by callers of the other technology (Figure 5.8e). Since no short-read data was available for the LM02 sample, we limit the analysis to the remaining 20 patients.

In our first analysis of this merged call set, we grouped the callers by technology and measured the proportion of variants supported by short-read and long-read sequencing. We observe that the majority of SVs are uniquely supported by long-read sequencing (57.41%). In contrast, only 18.15% of the SVs are only detected by short-read callers and 24.43% by callers of both technologies. These proportions remain robust across patients (SDs: 2.30, 3.00, and 1.59 for long-read, short-read, and calls supported by both technologies, respectively) but vary depending on the SV type (Figure A.4): We observe 41.64% shared Deletions, 18.25% shared Insertions, 6.1% Inversions and 0.67% Translocations. While long-read callers identify a large proportion of unique Deletions (42.53%), Insertions (77.00%), and Inversions (59.97%), short-read callers detect a total of 50, 920 (90.41%) unique Translocations. This increased number of short-read specific Translocations is likely a direct consequence of the read length: The SR evidence from short-read sequencing allows to detect breakpoints of SVs larger than individual reads but especially for Insertions these signatures can not be classified as a specific SV type leading to an inflated number of unresolved Breakends. Since there is no clear differentiation between Translocations and potential Insertions which would allow us to merge across the two SV types, we retain all the unique Breakends for further filtering in downstream analysis.

In the second part of our analysis, we focused on the size distribution of the initial merged call set stratified by technology support. The combined set of SVs spans a total of 2, 429mb. A major contribu-



(a) Variant Count.

(b) Technology Support.

(c) Size Distribution.

(d) Size Distribution by Technology.

(e) Technology Support of Callers.

(f) Caller Support.

Figure 5.8: **Comparison Between the short- and long-read SV Call-Sets.** This figure shows the results of our analysis comparing callers and technologies: **a)** SV counts of the merged call-set grouped by SV type for each patient on the X-Axis. **b)** The proportion of SVs supported by either PacBio, Illumina, or both technologies for each patient. **c)** The size distribution for each SV type of the merged-call set. **d)** The technology support for SVs binned by size. **e)** The proportion of SVs detected by a caller that is unique to the corresponding sequencing platform. **f)** The merged-call set of each patient stratified by caller support.

tion (1,750mb) to the cumulative length stems from a small proportion (0.80%) of large SVs (⩾ 100kb). The majority of SVs, however, is ⩽ 500bp (75.47%). We observe that even though calls unique to PacBio sequencing are more abundant, their cumulative size (905mb) is considerably smaller than the combined length of short-read specific calls (1,345 mb). This is largely due to the increased number of large SVs ($\geqslant$ 100kb) detected by Delly and Lumpy not supported by any of the long-read callers with a combined length of $1,076$mb. Notably, there is no shared variation $\geqslant$ 500kb in the entire merged call set indicating a potentially increased proportion of false-positive SV calls beyond this size threshold.

We further investigate the agreement between technologies for each individual caller. Long-read callers identify on average a large proportion of unique SVs (64.59%) - especially in comparison with short-read callers (44.14%) reflecting our previous results. The number of detected SVs unique to the corresponding sequencing technology remains stable across long-read callers ($SD = 2.022$) which is not the case for short-read callers ($SD = 9.27$): Manta detects the highest proportion of SVs shared with long-read callers (71.54%) in comparison to Delly (55.58%) and Lumpy (49.49%). A large contribution to this set are Insertions of which 87.03% are variants shared across technologies. In comparison, only 42.90% and 39.50% of the Insertions detected by Delly and Lumpy, respectively, are detected by a long-read caller.

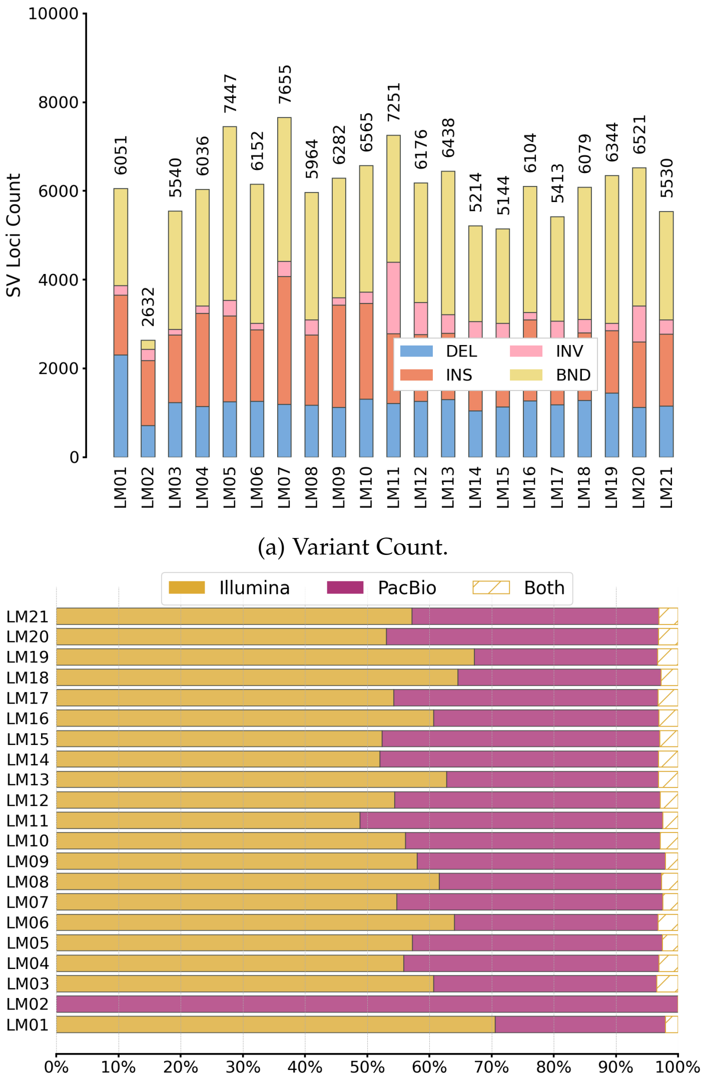
Finally, we stratify the call set depending on the number of supporting callers. We observe a major proportion of SVs that are uniquely detected by a single caller (34.72%). Across all patients, the number of SVs tends to decrease with an increasing number of supporting callers. However, there is a spike in the SVs supported by three callers (28.93%). This spike is caused by the increased overlap between the approaches of the same sequencing technology - especially the long-read SV callers.

With our analyses, we are able to reveal significant differences between the call sets of short- and long-read SV callers: First we confirm the expected significantly increased SV detection sensitivity through the addition of long-read sequencing data. Interestingly, short-read sequencing also provides a non-negligible proportion of unique SVs indicating the benefit of an approach leveraging both short- and long-read data. These sequencing technology-specific call sets follow considerably different size distributions as indicated by our second analysis. This is likely due to the type of read evidence available for each sequencing technology. PE evidence specifically allows to detect discordant read pairs across large distances which are potentially not identified by SR evidence derived from long-read sequencing data. The third analysis shows that the differences between technologies are also caller dependent. Manta, for instance, produces a call set much more similar to the PacBio callers with a higher proportion of Insertions and SVs $\leqslant$ 500bp as well as a reduced proportion of large ($\geqslant$ 1mb) SVs. Delly, in contrast, focuses predominately on PE evidence limiting its potential to identify SR signatures also present in long-read sequencing data. Overall, our pipeline allows detecting SVs with high sensitivity through the combination of short- and long data and multiple SV callers. While this is desirable for clinical diagnostic pipelines, since it increased the probability to include all potentially disease-causing variants, it also requires more rigorous filtering approaches to distinguish not only pathogenic from benign but false- from true-positive variant calls.

5.4 ALLELE FREQUENCY FILTERING

Many of the 25, 808 SVs we detect in each patient are common i.e. appear in multiple individuals in a population. Given the rare phenotype of our patient, it is likely that the disease-causing SVs we aim to identify are not among them. To distinguish between the common and therefore likely benign and rare, potentially disease-causing variants we need to determine their *allele frequency* (AF). While for SNVs and InDels extensive collections of variants annotated with their allele frequency computed in multiple populations are available for comparison, catalogs of common SVs are much less abundant - especially concerning SVs detected through long-read sequencing. Rather than just comparing our call sets with public databases we, therefore, use a two-step approach: First, we collect catalogs of SVs including information on their AF determined in previous studies of large cohorts. Second, we make use of the information contained in our own cohort, computing the AFs of SVs across all 21 patients.e Using this approach we aim to significantly reduce the total amount of variants and identify a set of *singletons* i.e. rare SVs present in a single patient. Given the patient-specific phenotypes in our cohort, we reason, that the disease-causing variation is among this set of unique variation. However, even patients with similar but not identical symptoms can share the same disease-causing SVs. We therefore also conduct a separate investigation using SVs detected in more than one individual in our downstream analysis.

To determine known common SVs in the call sets of the limb malformation patients, we first combined SVs from four catalogs of common variation [44, 45, 86, 96]. We filtered each dataset based on SV size and AF, if available, or a corresponding metric to discard all rare SVs. Then we formatted the filtered call set according to the same standard used for the short- and long-read variants detected with our own pipeline. The final set of combined common SVs included a total of 190, 205 variants. We then merged this set with the SVs of each patient and excluded any cluster containing common variation (see Methods for details of the filtering, formatting and merging process). The filtered i.e. rare call-set is shown in Figure 5.9a. We retain a total of 126, 538 rare SVs with a mean of 6, 026 per patient. The majority of remaining SVs are Translocations (43.59%) followed by Insertions (28.89%), Deletions (20.93%), and Inversions (6.58%).



(a) Variant Count.

(b) Technology Support.

Figure 5.9: **Call-Set After Comparison and Filtering using Common Variation.** **a)** shows the variant count grouped by SV type for each patient. **b)** shows the technology support of the remaining variants. **c)** shows the caller support.

The comparison with public SV catalogs of common variation resulted on average in a 76.65% reduction of SVs per patient. However, the number of filtered variants highly depends on the SV type. Deletions and Insertions are abundant in the public data sets and therefore likely represent a high proportion of common variation in the human population. Inversions on the other hand are much less frequent with a total of 480 variants in the combined set of common variation. Unresolved Breakends are not represented at all in any of the common SV catalogs and even though the GnomAD call set contains Translocations, none pass the AF threshold of 0.01. Thus, the number of Translocations in our cohort after filtering remains entirely unchanged and the number of Inversions is only slightly reduced (987 filtered out of 9,302). The initially high proportion of Breakends detected by short-read SV callers is therefore also reflected in the overall number of remaining variants supported by the individual sequencing technologies with 57.07% unique to short-read sequencing (Figure 5.9b). The contribution of short-read specific calls to the filtered call set is especially apparent when comparing the cohort-wide average with the number of remaining SVs in the patient LM02. As previously mentioned no WGS data was available for this patient limiting the merged call set to PacBio SVs. Of the initial 19,511 variants merged across the three long-read SV callers 2,632 rare SVs remained.

We, therefore, required an additional assessment of AF including Translocations and Inversions. Given the considerable number of samples in our own cohort especially in comparison with the current publicly available PacBio cohorts, we reasoned that we can employ the AF or *alelle count* (AC) in our cohort as additional evidence for the identification of common and likely benign variation. To compute the cohort-AC we employ the same merging approach previously used for the comparison with public databases. However, in this analysis, we retain sample-specific information in the process, such that we are able to back-trace SV clusters to individual patients. Based on this information we generate a cohort-wide call set that consists of 52,211 SV cluster representatives annotated with their corresponding AC, sample IDs, and genotypes. We then stratify SVs based on AC: *Shared* SVs are detected in all patients, *Major* SVs are detected in $\geqslant 50\%$ of the samples, *Poly* SVs are detected in $< 50\%$ but more than one sample and *Singletons* are detected in a single individual. An overview of the cohort call set grouped by these AC criteria is shown in Figure 5.10.

Across all SV types, we observe on average 75.43% Singleton, 15.61% Poly, 8.48% Major, and 0.48% Shared variants. For Translocations, we observe the highest proportion of SVs detected in two or more individuals (40.53%) which is to be expected as the previous AF filter could not be applied to this variant type. However, for Inversions, which were also under-represented in the public call sets of common variation, the proportion of Singletons in our cohort is still high (93.36%). 30.87% and 20.25% of the filtered Deletions and Insertions are present in more than one individual, respectively. This

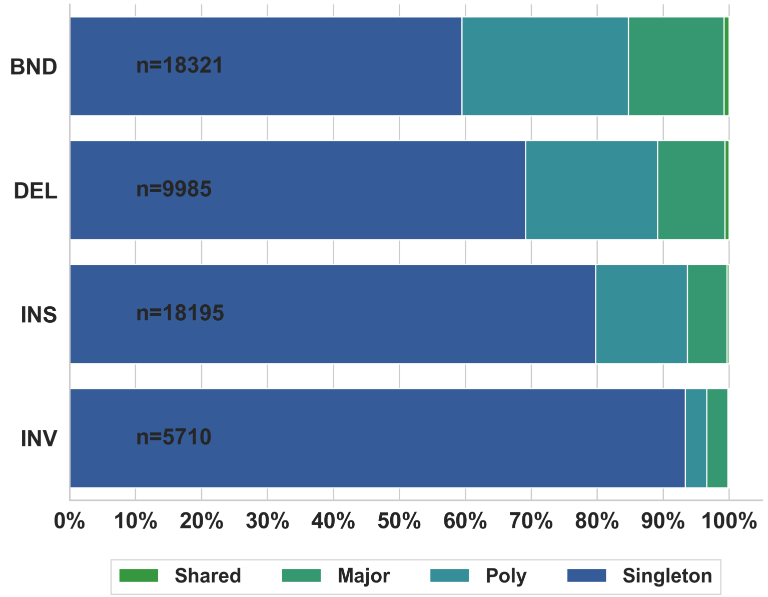
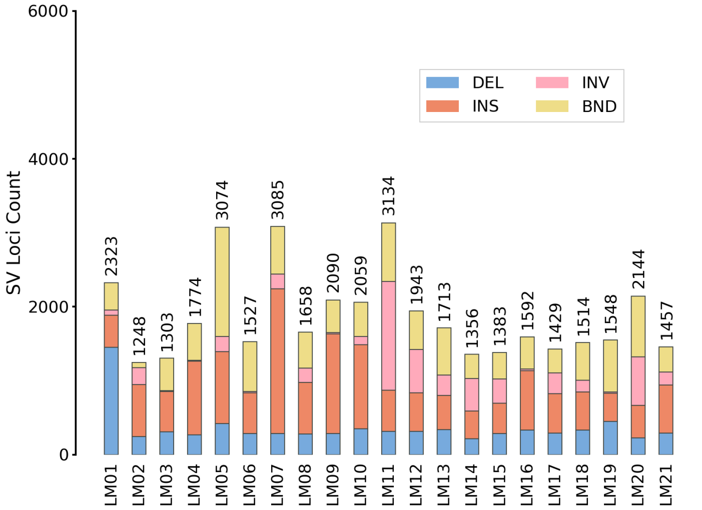


Figure 5.10: **Filtered SVs Stratified by Cohort Allele Count.** The figure shows the proportion of variants stratified by SV type and cohort allele count.

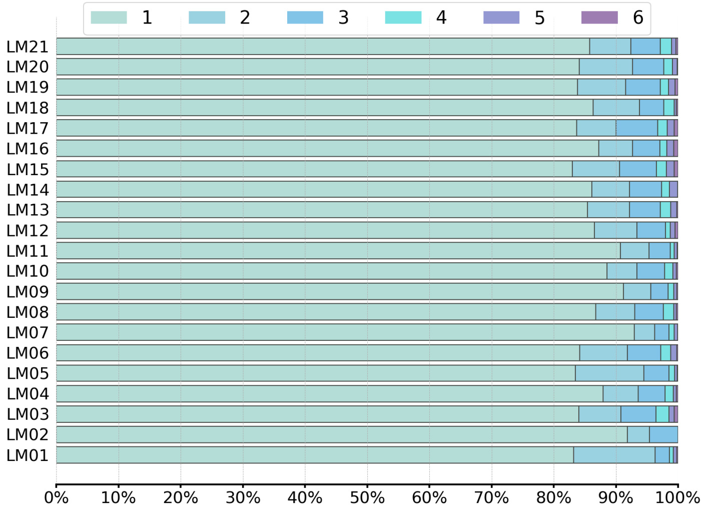
indicates that the cohort-specific AC provides additional information on common SVs that are not listed in any of the publicly available SV catalogs. It should be noted, that this shared variation in our call set is not only includes SV common in the population but are also highly influenced by biases in the SV callers and alignment process. Regardless, the common variants are unlikely to include any potentially disease-causing candidates. We, therefore, set a strict filter, retaining only singleton SVs for further analysis. The resulting final call-set is shown in Figure 5.11.

The call set included 39,354 Singletons with an average of 1,874 per patient. The majority of Singletons are Insertions (38.40%). We observe a high proportion of Translocations (28.51%) followed by Deletions (19.75%) and Inversions (13.34%). As in the unfiltered call set, most Translocations are short-read specific i.e. are not detected by any Pacbio Caller (76.71%). A marginal proportion of variants is supported by both short- and long-read callers (3%) while most SVs are derived from long-read sequencing (62.06%). Finally, we observe an overwhelming majority of SVs supported by a single caller (86.83%). This indicates a potentially high rate of *false-positives*.

In comparison to the initial SV call set merged across callers and technologies, we achieved a reduction of 92.73% per patient using public databases and AF information derived from our own cohort. Comparable pipelines frequently include, as previously discussed, additional thresholds on caller or technology support. Given the high proportion of calls supported by a single caller in our final call set, applying similar filters would therefore reduce the number of variants considerably. This would also reduce the amount of manual investigation needed in any further prioritization. However, while we agree that such a threshold would decrease the number of *false-positive*, we argue that strict thresholds set without further inspection of the calls would potentially exclude disease-causing variation as well. Thus, we retain all SVs of this final call set for the second part of our pipeline - the prioritization of potentially pathogenic variation using functional annotation.



(a) Variant Count.



(b) Caller Support.

Figure 5.11: **Final Call-set after Cohort AC Filtering. a)** Variant Count grouped by SV type for each patient. **b)** Variants stratified by caller support.

6 PRIORITIZATION

The fundamental aim of variant prioritization is the assisted identification of disease-causing variation and ultimately reducing the amount of required experimental validation while increasing the number of successful clinical diagnoses. Even though the majority of methods have been developed for the prioritization of SNPs and InDels, a growing number of prioritization approaches have been published in recent years designed specifically for SVs. These SV prioritization methods can be approximately grouped into two categories depending on their underlying framework: *Automated* and *Semi-Automated* approaches. Automation in this context means that the method is capable of reducing an initial set of variants to a number of candidates without supervision to be then inspected by a clinician or geneticist. Current methods achieve this by supplying a summary pathogenicity score produced by machine-learning (ML) models that allows to *rank* SVs by their pathogenic potential. The ranking can then serve as a recommendation system for the clinician. *Automated* methods are agnostic to the disease or tissue context and provide a more general recommendation. The *semi-automated* approaches on the other hand largely focus on annotation and visualization of SVs using disease-specific information often including phenotype descriptions of a patient. While these approaches can be used in combination with user-defined thresholds to reduce the number of variants, their application involves in most cases a more extensive manual analysis.

In this chapter, we summarize the current state of SV prioritization separated into automated and semi-automated methods. For each group of approaches, we then present the related work as developed in this thesis. First, we discuss the previously developed CNV prioritization approach TADA which we evaluated as part of this Ph.D. work. Secondly, we present an adaption of TADA to include all SVs and limb-malformation-specific prioritization. This includes the construction of a *Limb Regulome* - a combination of functionally relevant genomics annotations including the patient-specific RNAseq and Hi-C analysis.

6.1 AUTOMATED PRIORITIZATION OF SVS

Several machine learning models have been introduced in recent years for the prioritization of disease-causing SVs. The models can be separated depending on the origin of their training data. Most approaches rely on call sets of known pathogenic SVs [116–118]. These models are trained to distinguish between pathogenic and benign SVs in rare disease patients and are exclusively designed to assess the effect of SVs on coding regions. This is primarily due to the lack of known pathogenic SVs that act through non-coding related mechanisms limiting the potential to train any type of machine learning model to prioritize non-coding SVs. Other approaches that include non-coding annotation, therefore need to rely on alternative call sets of SVs. The *supervised machine learning framework* SVFX, for instance, employs somatic SVs as a proxy of pathogenicity [119]. In addition, it provides a model for common disease in humans for which an increased number of associated non-coding yet not necessarily pathogenic SVs is available.

Another alternative to the use of annotated pathogenic SVs is the use of conserved variation across species. This allows training models that identify deleterious rather than pathogenic SVs [120, 121]. These *evolutionary* models are applicable to all types of SVs, assigning a score that indicates the selective pressure on individual variants. This score can be understood as a general assessment independent of the patient's phenotype and affected cell type as well as inheritance pattern. In contrast, the majority of models trained on known pathogenic coding SVs are also able to include patient-specific information e.g. their symptoms encoded as *Human-Phenotype-Ontoloy* (HPO) terms. Their computed summary pathogenicity scores then reflect not only the functional effect of an SV on a gene but also its significance to the patient's disease. This application is potentially more in line with a typical clinical scenario where the phenotype and a selection of associated genes are in many cases known. However, disease-causing variations in genes not yet associated with the phenotype are more challenging to detect with these approaches. Evolutionary models are in this case more suited to detect disease-causing SVs among the high-ranked deleterious SVs since their summary score does not rely on phenotype-specific knowledge. The selection of an automated prioritization approach, therefore, should be chosen depending on the suspected molecular mechanisms.

6.1.1 *Evaluation and Performance Comparison of TADA*

TADA is an example of an automated prioritization approach focused exclusively on CNVs. The concept of TADA was initially developed during the master thesis preceding this dissertation. We provide a summary of the method in Chapter 4. Briefly, TADA consists of two random forest classifiers trained to distinguish between pathogenic and benign Deletions and Duplications based on a set of features quantifying the changes to the affected regulatory environment. Although the models were trained during the master thesis, we conducted a rigorous comparison of TADA with current prioritization approaches as part of the Ph.D. work. The method itself and the comparisons were published in a single manuscript [89]. In the following section, we describe the in total four analyses performed to evaluate TADA's predictive performance in comparison with other prioritization methods:

- A *ROC-AUC* score-based analysis across three call sets to assess the classifiers' general classification ability
- A *F1-Score* analysis including methods without continuous pathogenicity metrics
- A *Ranking* analysis to investigate the methods' ability to identify a single pathogenic variant from a large background of benign CNVs
- A *developmental-disease* analysis based on two patients with known pathogenic Duplications.

ROC-AUC Analysis

For the initial analysis, we assessed TADA's predictive performance for multiple thresholds set on its summary pathogenicity across three call sets: First, a 5-fold *cross-validation* (CV) split of the original training set. Second, the test-set split. Third, a set of pathogenic and benign *ClinVar* CNVs. We first computed *ROC-AUC* values for the Duplication and Deletion model and each data set. The evaluation of the Deletion model results in ROC-AUC scores of 0.8379 (*5-CV*), 0.8059 (*Test-Split*), and 0.8865 (*ClinVar*). The ROC-AUC values for the Duplication model are 0.8069 (*5-CV*), 0.7868 (*Test-Split*) and 0.8424 (*ClinVar*). We then compare our predictive performance to SVFX and SVScore [119, 120]. The results are shown in Figure 6.1.

The SVFX framework allows training classifiers on individual variant sets to identify pathogenic CNVs based on functionally relevant annotations. The annotation and prioritization are therefore conceptually similar to the TADA framework. However, TADA is trained on size-matched data, while SVFX employs a normalization method to account for the size bias between pathogenic and non-pathogenic variants. There are several practical limitations to this method leading to overestimated performance metrics mainly driven by data leakage between the training and test set. To reduce this bias and allow for a sensible performance comparison, we trained an SVFX model on our own size-matched training data. We then compute ROC-AUC scores for all three test data sets: 0.7836 (*5-CV*), 0.7613 (*Test-Split*) and 0.8311 (*ClinVar*) for the SVFX Deletion model. For the Duplications the ROC-AUC scores are 0.7613(*5-CV*), 0.7575 (*Test-Split*) and 0.7384 (*ClinVar*). Thus, TADA outperforms SVFX across all three test sets.

The second comparable method, SVScore calculates the mean of the ten highest *Combined Annotation Dependent Depletion* (CADD) scores [122] in the interval affected by a CNV. It, therefore, does not allow for retraining on our data which excludes a performance comparison based on the $5 - CV$ data. Instead, we applied the method to the *Test-Split* and *ClinVar* CNVs with default parameters. Then we normalized the scores for each set of variants to a range between 0 and 1. This allows for a direct comparison to the pathogenicity scores computed by TADA and SVFX. The resulting SVScore ROC-AUC value for the Deletion model are 0.6909 (*Test-Split*) and 0.8771 (*ClinVar*). For Duplications, the ROC-AUC scores are 0.7079 (*Test-Split*) and 0.8582 (*ClinVar*). While TADA outperforms SVScore on the *Test-Split* data, the difference of ROC-AUC scores for the *ClinVar* variants is less pronounced. SVScore performs marginally better than TADA on *ClinVar* Duplications.

We reason that this increased performance for *ClinVar* Duplications is likely caused by the underlying size bias between pathogenic and non-pathogenic *ClinVar* variants. All CNVs larger than 1 Mb are given a score of 100 by SVScore. Thus, it effectively labels all large CNVs as pathogenic. In data sets with a high number of large pathogenic CNVs and with many benign CNVs < 1mb, such as ClinVar, the method, therefore, performs particularly well. We argue, that this likely leads to an underestimation of small yet pathogenic CNVs. To investigate this further, we conducted an additional analysis stratifying the Deletions contained in the test sets i.e. *Test-Split* and *ClinVar* by size into three groups: *Small* (< 50kb), *Medium* (< 100kb), *Medium-Large* (< 1mb) and $>= 1mb$. We used Deletions rather than Duplications in this comparison due to the increased number of pathogenic variants. We measure the performance of TADA, SVFX and SVScore for all size groups using ROC-AUC values. The results are shown in supplementary Figure A.5. TADA outperforms both SVFX and SVScore across all size groups with the exception of *Large* Deletions, indicating that TADA's is less reliant on the size difference between pathogenic and non-pathogenic variants.

In an additional ROC-AUC-based analysis we compared TADA to CADD-SV [121]. CADD-SV is a recent adaptation of the original CADD method aimed at the prioritization of deleterious SVs. It, therefore, is not a direct competitor of TADA but provides a similar although more evolutionary centered prediction. We used the pre-trained CADD-SV classifier on the *Test-Split* and *ClinVar* variants. To allow a comparison with TADA's pathogenicity score we used the

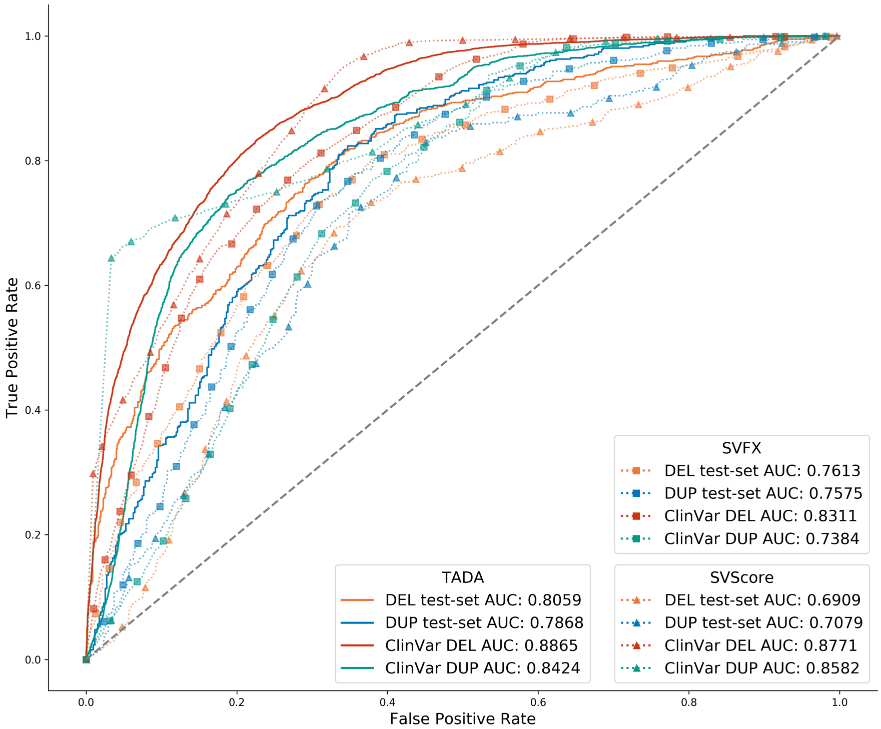


Figure 6.1: **Classification Performance of TADA, SVFX and SVScore.** The figure shows the ROCs for *Test-Split* and *ClinVar* variants and the corresponding ROC-AUC scores.

maximum of *span* and *flank* raw scores and additionally employed a min-max-normalization for each variant set. We then compute ROCs for both test sets. The results are shown in supplementary Figure A.6. While TADA outperforms CADD-SV for both *ClinVar* and *Test-Split* CNVs, the difference in performance, especially for the *Test-Split* Duplications, is marginal (0.0012).

F1-Score Analysis

In the second analysis, we compared TADA to the Ensembl Variant Effect Predictor (VEP) [123]. VEP is a method preferentially used for the annotation and prioritization of SNPs and InDels as it allows assess individual changes in amino-acid sequences. However, it also allows to annotate CNVs with regulatory annotation and returns an *IMPACT* rating reflecting the potential pathogenicity of a variant. The *IMPACT* rating is categorical and therefore not directly comparable to the continuous summary score of TADA. Briefly, VEP groups variants in four categories: (*HIGH*, *MODERATE*, *LOW* and *MODIFIER*). To allow for a comparison with TADA, we defined *HIGH* or *MODERATE* CNVs as pathogenic and *LOW* as well as *MODIFIER* variants as non-pathogenic. We then computed F1-Scores, a macro-averaged metric of *precision* and *recall* for the *Test-Split* and *ClinVar*

	DUP Test Set	DEL Test Set	ClinVar DEL	ClinVar DUP	ClinVar DEL (<1Mb)	ClinVar DUP (<1Mb)
TADA	**0.73**	**0.74**	**0.73**	0.53	**0.69**	0.42
SVScore	0.43	0.46	0.67	**0.83**	0.66	**0.54**
VEP	0.47	0.42	0.69	0.59	0.63	0.43

Table 1: **Classification Comparision of TADA, SVScore and VEP.** The performance is measured in macro averaged F1 scores on Deletions and Duplications of the test split as well as ClinVar variants. F1-scores in bold indicate the best-performing method for the individual variant set.

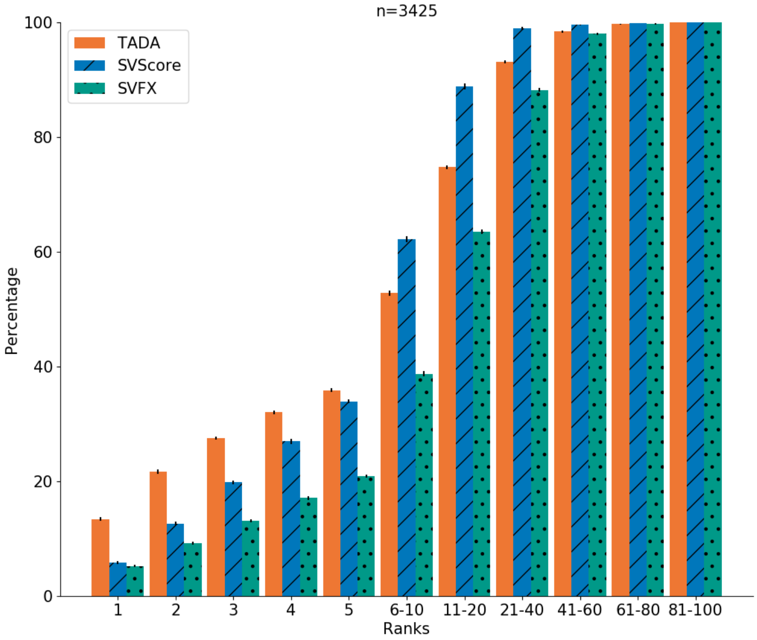
CNVs. To account for the unbalanced size distribution of *ClinVar* variants we also computed a separate F1-Score for CNVs < 1mb. We then compared the results to F1-Scores based on TADA and SVScore predictions. In order to transform the continuous scores into two categories, we classified all variants with a TADA pathogenicity score higher than 0.5 as pathogenic. For SVScore we used the 90th percentile of the scores in an individual call-set to distinguish between pathogenic and non-pathogenic CNVs following the recommended threshold with the highest reported performance [120]. The results of the comparison are shown in 1. TADA outperforms SVScore and VEP for both Deletions and Duplications of our *Test-Split* variants and *ClinVar* deletions. For *ClinVar* Duplications SVScore classifies 84% of the variants correctly which is the best macro-averaged F1 score amongst all three tools. However, as previously discussed, the high performance can be explained by the dependency on the size difference between pathogenic and non-pathogenic *ClinVar* variants.

Ranking Analysis

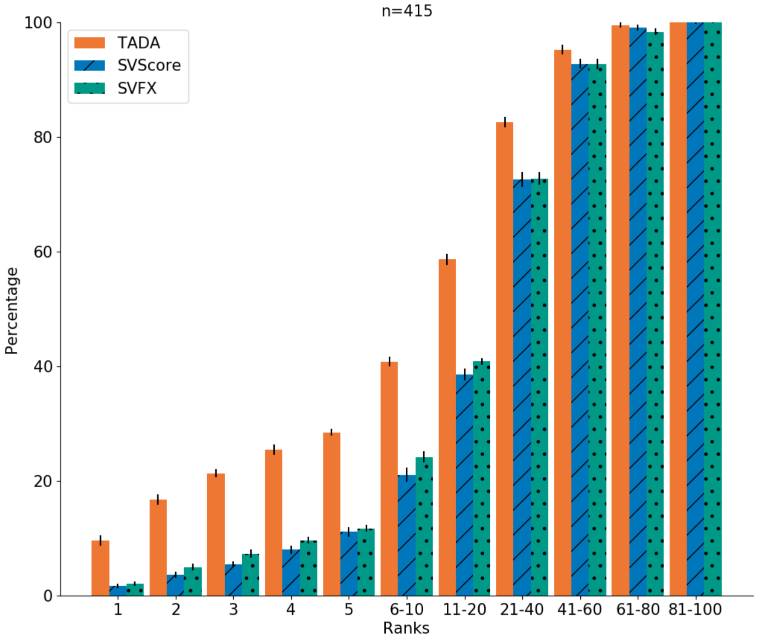
The ROC-AUC and F1 score comparison rely on distinct thresholds separating pathogenic from non-pathogenic CNVs. While this provides a generalized indication of the predictive performance it does directly reflect the application in clinical practice. In a typical scenario, the clinician ultimately has to identify a single pathogenic variant from a large background of non-pathogenic CNVs. A classifier designed to assist in the molecular diagnosis of rare disease patients should therefore be able to assign a well-calibrated pathogenicity score. This score would allow ranking all CNVs detected in the patient such that the true pathogenic variant is placed as high as possible reducing the number of manual inspections. To test the calibration of our classifiers, we compute the fraction of true positives and the mean predicted value. With a *perfectly calibrated* pathogenicity score these values would be equal. We visualized the calibration of TADA's random forest classifiers in supplementary Figure A.7. The results indicate that the TADA's summary pathogenicity scores are well-calibrated.

To test the calibration of TADA's pathogenicity score in comparison with alternative prioritization methods we conducted an additional *ranking analysis* (see Methods for details). Briefly, we generated test batches of 100 CNVs containing a single pathogenic variant and computed its rank based on the pathogenicity scores. This allowed us to determine the method-specific proportion of true pathogenic CNVs and the corresponding standard deviation for all ranks between 1 to 100. The resulting performance is shown in Figure 6.2. The TADA deletion classifier outperforms both SVFX and SVScore. In 35.9% of the batches, the true pathogenic variant is placed the top 5 ranks compared to 20.9% (SVFX) and 33.9% (SVScore). However, SVScore places more pathogenic CNVs in the ranks below the 10th rank. In the duplication ranking analysis, TADA outperforms SVFX and SVScore with respect to all ranks. 28.66%, 11.7% and 11.06% of true pathogenic Duplications from 415 batches are placed among the first 5 ranks by TADA, SVFX and SVScore, respectively.

In an additional analysis, we assessed TADA's rankings ability with respect to rare variants. Typically in pipelines such as the one we developed for the limb malformation cohort, initial filtering is included to discard any common variation. The remaining rare variants are then further investigated using prioritization methods e.g. TADA. To simulate this scenario we generate again batches of 100 CNVs including a single true pathogenic *ClinVar* variant and 99 rare variants with AF < 0.01 [86]. In this ranking analysis, SVscore performs marginally better than TADA placing the true pathogenic variants in 4.6% of the batches among the first 5 ranks. In a manual inspection of the true pathogenic variants, we observed that SVScore assigns low scores to all CNVs not affecting coding regions while TADA also considers non-coding variants as potentially pathogenic. To test the effect of non-coding features on the ranking ability with rare variants, we trained a separate model using only coding features and performed the ranking analysis again. With the coding model TADA outperformed SVScore across all ranks. The difference in performance is likely due to the investigator bias in the *ClinVar* CNV set since it almost exclusively includes pathogenic coding variants. However, we reason current prioritization methods need to be able to assess non-coding variation as well given the growing evidence for pathogenic CNVs acting through non-coding related mechanisms [57]. The current trained TADA models therefore also include features accounting for the non-coding regulatory environment.



(a) Ranking Performance for Deletions.



(b) Ranking Performance for Duplications.

Figure 6.2: **Ranking Performance Comparison of TADA's Deletion and Duplication Classifiers.** For each bin we computed the percentage of variants placed among the corresponding rank or ranks. Black bars indicate the standard variation based on 30 random sampling runs.

Developmental Disease Patient Analysis

In a final evaluation, we measured TADA's performance on two individuals with developmental disease (DD). Both patients were part of a previous in-depth analysis exploring the potential of Hi-C to resolve disease-causing Duplications [124]. To simulate the application of TADA on the two patients (DD1 and DD2) in a scenario where the pathogenic Duplications are not yet known, we *spiked in* the two known pathogenic Duplications into the initially detected set of CNVs of DD1 and computed summary pathogenicity scores. The results are visualized in Figure 6.3. TADA is able to identify both disease-causing Duplications as pathogenic and assigned higher pathogenicity scores than the 90th percentile (0.4336). The DD2 Duplication was placed on rank 2 (0.7986) of all detected CNVs. However, the DD1 Duplications was ranked considerably lower (0.5865).

To identify the driving factor behind TADA's classification process we investigated the regulatory environment of the pathogenic Duplications. Both Duplications are located in proximity to the SOX9 locus overlapping TAD boundaries, multiple genes, and FANTOM5 enhancers. However, while the DD2 Duplication directly affects the SOX9 gene locus, the DD1 Duplication is located outside its coding region. The direct overlap with SOX9, a highly haploinsufficient gene (0.9981 p(HI)), likely drives TADA's classification process for the DD2 Duplication given previous indications on TADA's reliance on coding information. Since the DD2 Duplication is likely acting through increased gene dosage effects the prediction accurately reflects potential pathogenic effects. The suggested disease-causing mechanism for the DD1 by Melo et al. is the formation of a novel chromatin domain (*Neo-TAD*) including copies of KCNJ2, KCNJ16 as well as SOX9 enhancers leading to misexpression of KCNJ2 [124]. While we can observe an increased pathogenicity score for the DD1 Duplication, we suspect that it is driven by the proximity to SOX9 alone rather than reflecting the complex rearrangements causing the patient's phenotype.

This analysis of DD patients indicates that TADA is able to identify pathogenic CNVs from a background of benign variation but largely depends on coding information. We performed a permutation-based test of feature importance to investigate the extent of TADA's reliance on coding information. We measured the decrease in performance on the *5-CV* set after permuting clusters of features determined through partial correlation analysis (see Hertzberg et al. for details [89]). The six most important features i.e. those with the highest decrease in accuracy after permutation are all coding-related. Similar to the previous ranking ability comparison of TADA and SVScore this is largely due to the investigator bias in currently available data sets of validated/known pathogenic variation. Limited by the lack of knowledge

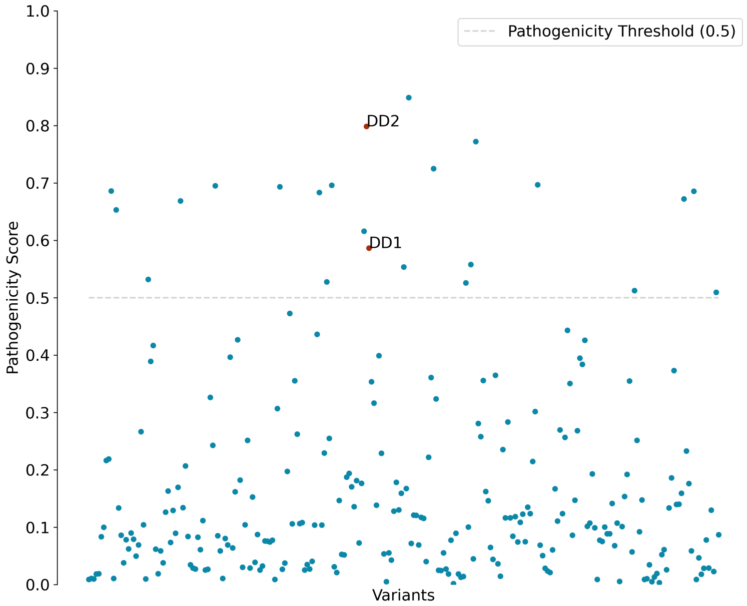


Figure 6.3: **Pathogenicity of Duplications in DD-Patients.** The figures shows the computed summary pathogenicity scores of TADA for all Duplications detected in the DD-Patients. The true pathogenic variants are marked in red. The dotted line indicates a threshold of 0.5 distinguishing pathogenic from non-pathogenic CNVs.

about the regulatory function of non-coding regions, the majority of case studies that have been conducted so far were focused on variation affecting genes. Thus, validated pathogenic variants are almost exclusively coding. While we expect that with the growing number of case studies revealing pathogenic non-coding variation, TADA's automated classification process becomes more reliable for such variants, the application of its current models should be restricted to coding variation. This also applies to all other machine learning-based prioritization methods trained on known pathogenic variation.

The previous analysis of the patients in our limb malformation cohort included an investigation of large SVs hitting genes of interest without resulting in any identified disease-causing candidates. It is therefore likely that the true pathogenic SVs are acting through non-coding regulatory mechanisms. This should also be reflected in our prioritization approach. Since the machine learning classifiers of TADA and comparable *automated* approaches underestimate non-coding effects, we therefore instead focus on a *semi-automated* approach that allows leveraging all available information including relevant non-coding regulatory annotations.

6.2 SEMI-AUTOMATED PRIORITIZATION OF SVS

Semi-automated prioritization methods do not employ machine learning models for direct classification and therefore are not restricted by the lack of known non-coding pathogenic SVs. They focus on the assisted inspection of individual variants rather than the assignment of summary pathogenicity scores. However, some of the methods we assign to this category provide features that reduce the number of SVs to a set associated with the patient's phenotype or relevant regulatory elements. Thus, we chose to describe the category as *semi-automated* rather than *manual*. Current tools include visualization-based methods that allow inspecting the regulatory environment of SVs [125, 126], annotation frameworks leveraging phenotype and hereditary information for larger cohorts [127], predictive approaches for the potential perturbation of regulatory interactions [128] and disease-specific models aimed predominantly at TAD-boundary hitting CNVs and Inversions [129].

For our purpose i.e the identification of disease-causing SVs in limb-malformation patients we require a method that combines multiple aspects of these approaches: An annotation for all types of SVs incorporating limb-specific regulatory elements, a corresponding set of filters that allow reducing the number of SVs to functionally relevant candidates and a visualization approach for a final manual inspection. While each of the currently available tools could prove to be

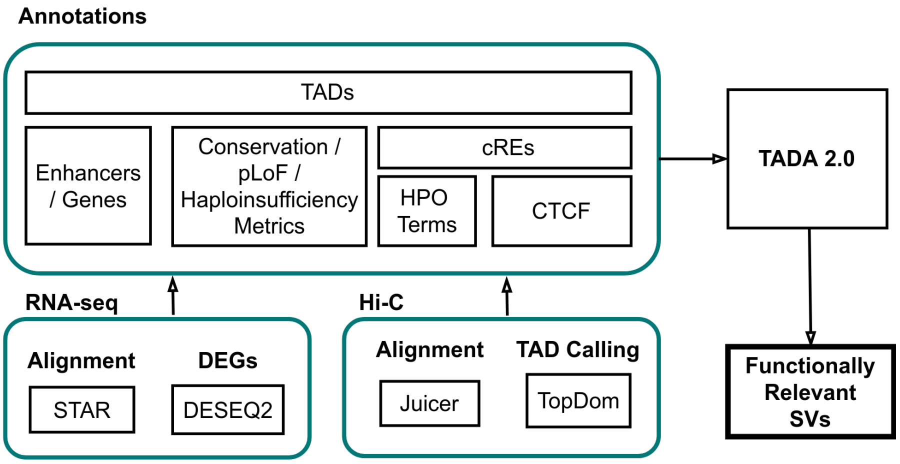


Figure 6.4: **Functional Annotation based Prioritization of SVs.** We center our annotation process around an extended version of TADA capable of annotating all types of SVs. As input, we collect a comprehensive collection of disease-context-specific regulatory elements including the annotations derived from our RNA-seq and Hi-C analysis.

beneficial for this process, they would require extensive adaptions - especially when applied in an ensemble approach. Thus, we set out to develop a novel prioritization method i.e. TADA *2.0* fitted to the limb malformation cohort that allows us to tune filtering steps during prioritization and include it in our *snakemake* pipeline (Figure 6.4). As the basis for this prioritization method, we use the TADA annotation framework. It is therefore an approach that directly depends on the information contained in collections of regulatory annotations.

6.2.1 *Limb Regulome*

To reflect the regulatory environment impacted by SV as accurately as possible, we set out to collect a comprehensive set of limb-development-specific annotations i.e. a *Limb Regulome*. We include regulatory elements from publicly available data sources as well as patient-specific annotations derived from our Hi-C and RNA-seq analyses. In the following paragraphs, we describe each set of annotations split into coding and non-coding, their origin, and the corresponding experimental analyses.

Coding Annotations

We first collected gene annotations from ENSEMBL (v.104). Then we discarded any duplicated entries based on gene symbols and annotated them with metrics describing their *haploinsufficiency* and *Loss-of-Function* intolerance if the gene was present in the corresponding databases [85, 86]. We also collected Exon annotations as well as

5′ and 3′-prime UTRs, Start- and Stop-Codons from the ENSEMBL resource. For exons we reduced the initially available information to the genomic loci and corresponding gene symbols, discarding any duplicated entries. We also retained the list of developmental-disease-associated genes i.e. *DDG2P-genes* used in the initial TADA annotation set [85]. In addition, we collected four sets of genes specific to the disease context of our cohort: 1) genes correlated with limb development based on a separate scRNA study (citation needed) 2) genes known to be associated with limb development derived from previous work by the Mundlos AG at the Max-Planck-Institute for molecular genetics [130] 3) phenotype associated genes and 4) differently expressed genes (DEGs). To generate a list of genes associated with phenotypes for each patient we first matched the original descriptions provided by the investigating clinicians to individual HPO terms [131]. Using a mapping from HPO terms to phenotypes we then extracted the corresponding genes (`http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt`). The DEGs are the results of our pipeline to process the patient-specific RNA-seq data. We provide a detailed description of the process in Chapter 4.

Non-Coding Annotations

We collected cell-type specific active enhancers determined using *cap analysis gene expression* (CAGE) from FANTOM5 [63]. Given the lack of limb-specific enhancers in this specific resource, we used the entire collection during annotation. We annotate all FANTOM5 enhancers with aggregated base-wise conservation scores calculated across 100 species [91]. Similarly, we downloaded the entire set of human *candidate regulatory elements* (cRE) from ENCODE [132]. While many regulatory elements contained in this set are likely not relevant to limb development they could indicate regions of potential interest that would warrant further investigation. We also collected a set of experimentally validated enhancers from VISTA [64]. To reflect the limb-development associated regulatory landscape we collected a set of cREs identified in mouse embryonic limb tissue (14.5 days - ENCODE ID: *ENCFF890IPV*), lifted the coordinates over to GRCh38 (188, 109 elements remaining) and split them into *Promoter-Like-Signatures* (PLS), *distal Enhancer-Like-Signatures* (dELS) and *proximal Enhancer-Like-Signatures* (pELS). Given the lack of publicly available limb-specific CTCF sites in humans, we collected ENCODE ChIP-seq derived CTCF narrow peaks from *fibroblast* samples (ENCFF882YMD). To associate enhancers to potential gene targets in a limb-development-specific context we included the data from a recent PLAC-seq experiment performed using embryonic mouse tissues [103] (see Methods for details). Finally, we included TAD boundaries derived from our Hi-C analysis both as proxies for windows of increased regulatory interactions and as individual annotations (see Methods for details).

6.2.2 *Adaption of TADA for the Limb Malformation Cohort*

With the collected *Limb Regulome* we then set out to update and extend TADA's annotation framework. Briefly, we first adjusted the initial variant processing to include all types of SVs and implemented customs scripts that allow realigning the inserted sequence of Insertion calls revealing the original location in the reference (see Methods for details). This allowed us to include not only the position of Insertions in the annotation process but also the regulatory environment of the inserted sequence. We then adjusted the original TADA features for a limb malformation-specific analysis as described in the following paragraph.

The *Limb Regulome* provides several new opportunities to extend the original TADA features. For each set of novel (and original) annotations, we include two SV features: The distance to the closest element inside the same TAD environment and the bp-overlap aggregated over all overlapping elements. In addition, we compute the distance to the two sets of TAD boundaries called by *TopDom* for the cohort Hi-C map with *window-sizes* 5 and 10. We also return the total number of genes and enhancers overlapping with an SV and several metrics quantifying the regulatory importance of the impacted elements: Minimum *Loss-of-Function* intolerance and maximum predicted haploinsufficiency either of all overlapping genes or the closest gene if no overlap was found. Similarly, we report the maximum conservation of all overlapping or the closest FANTOM5 enhancer. To account for SVs overlapping multiple haploinsufficient genes we compute the compound *Haploinsufficiency Log-Odds Score* as introduced by Huang et al. [90]. Finally, we compute the exon overlap as the proportion of overlapping exonic base pairs vs. total exonic base pairs of each overlapping gene and report the highest proportion. The features and their definition are summarized in table 2.

6.2.3 *Annotation of the Rare SVs*

The call sets returned by the first part of our pipeline include up to 3,143 variants per patient and still likely include a high number of *false-positives* indicated by the rate of variants uniquely detected by a single caller. Avoiding the application of a hard threshold on caller support, we first want to determine which SVs are functionally relevant with respect to the patients' phenotypes and the more general context of limb development. Given the *Limb Regulome* and updated TADA framework, we are now able to annotate SVs with features corresponding to the affected limb-development-specific regulatory environment. For the annotation process, we first generate configuration files for each patient which include the locations of the set

Feature	Description	Distance (bp)	Overlap (bp)
Human PLS	Promoter-like signatures in human tissues and cell-types	Yes	Yes
Human dELS	Distal enhancer-like signatures in human tissues and cell-types	Yes	Yes
Human pELS	Proximal enhancer-like signatures in human tissues and cell-types	Yes	Yes
Fibroblast CTCF	CTCF peaks in all human tissues and cell-types	Yes	Yes
Mouse Limb PLS	Promoter-like signatures in mice embryonic limb	Yes	Yes
Mouse Limb dELS	Distal enhancer-like signatures in mice embryonic limb	Yes	Yes
Mouse Limb pELS	Proximal enhancer-like signatures in mice embryonic limb	Yes	Yes
Fibroblast CTCF	CTCF peaks in fibroblast	Yes	Yes
Stop Codon	ENSEMBL stop codon annotations	Yes	Yes
Start Codon	ENSEMBL start codon annotations	Yes	Yes
5_UTR	ENSEMBL 5′-UTRs codon annotations	Yes	Yes
3_UTR	ENSEMBL 3′-UTRs codon annotations	Yes	Yes
Gene	ENSEMBL gene annotations	Yes	Yes
FANTOM5	CAGE-based enhancers in human tissues	Yes	Yes
VISTA	Experimentally validated enhancers in human tissues	Yes	Yes
DDG2P	Genes associated with developmental-disease	Yes	Yes
Limb Gene	Genes associated with limb-development	Yes	Yes
scRNA Gene	Genes correlated with limb-development based on scRNA-seq analysis	Yes	Yes
Phenotype Gene	Gene associates with the patient's phenotype	Yes	Yes
DEG	Differentially expressed gene derived from RNA-seq analysis	Yes	Yes
TAD Boundary (10w)	TopDom TAD boundaries called for the cohort with windows-size 10	Yes	No
TAD Boundary (5w)	TopDom TAD boundaries called for the cohort with windows-size 5	Yes	No
scRNA Gene Correlation	Correlation with limb-development of the closest scRNA gene	n/a	n/a
Min. Gene LOEUF	Min. LoF intolerance score of overlapping genes or the closest gene.	n/a	n/a
Number of affected Genes	The total number of affected genes	n/a	n/a
Number of affected Enhancers	The total number of affected enhancers	n/a	n/a
Max. Gene HI	Min. haploinsufficiency score of overlapping genes or the closest gene.	n/a	n/a
HI Log Odds	Compound score describing the combined haploinsufficiency for all overlapping genes	n/a	n/a
Max. Exon Overlap	Max. Proportion of exons hit among overlapping genes	n/a	n/a
Max. Enhancer conservation	Max. conservation of overlapping or closest FANTOM5 enhancer	n/a	n/a

Table 2: **Updated TADA Features for a Limb-development specific Annotation.** This table shows the identifier we use for each feature, a more detailed description and if annotation elements are used for overlap/distance computations.

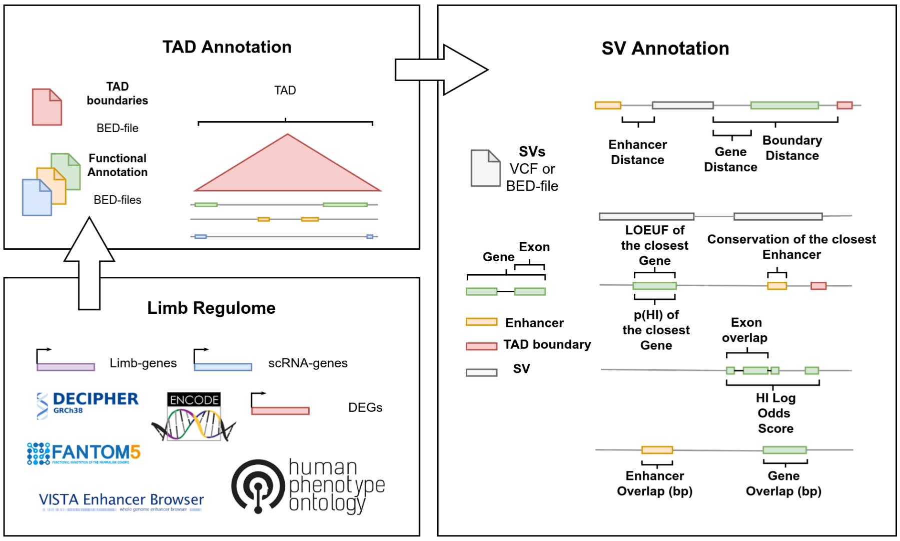
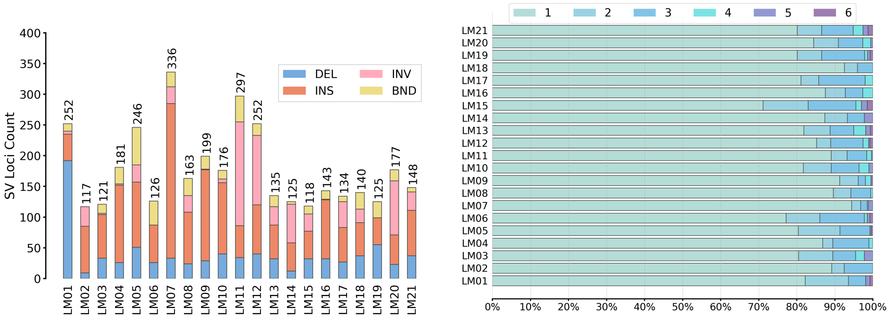


Figure 6.5: **Limb-Development Specific Annotation using TADA**. The illustration depicts the workflow of the extended TADA annotation framework. We updated the process to include all types of SVs and collected a comprehensive set of regulatory annotations including disease-specific sets of genes and enhancers. This version of TADA returns annotated SVs without further automated prioritization.

of previously presented regulatory elements. We then perform the annotation with respect to TAD boundaries determined by *TopDom* with *window-size* 10 derived from the cohort Hi-C map. An overview of this process is shown in Figure 6.5. The annotation process provides detailed information of overlap and distance to limb development and phenotype-associated regulatory elements. To determine which SVs are functionally relevant, we implemented a filtering approach centered around four sets of genes: 1) Phenotype associated genes derived from HPO terms (*Phenotype-Genes*), 2) Genes associated with limb development identified in previous studies (*Limb-Genes*), 3) Genes correlated with limb-development based on a scRNA analysis ((*scRAN-Genes*)) and 4) Top 50 DEGs determined using the patient-specific RNA-seq analysis (*DEGs*). The cohort-wide sets contain 658 (*Limb-Genes*) and 1,401 (scRNA-Genes). Since the symptoms and therefore the HPO terms are unique to almost all patients the number of genes contained in the *Phenotype-Genes* set varies considerably across the cohort (between 21 and 1,325 Genes). An exception is patient LM21 for which no phenotype descriptions are available.

For each group of genes, we identify TADs (*TopDom window-size* 10) containing at least one gene of interest. We then combine the gene-set-specific TADs (*Limb-TADs*)and discard any duplicates. To reduce the initial set of rare variants to those of potential function relevance we filter for SVs overlapping with any of *Limb-TADs*. For each of the



(a) SV Counts Stratified by Type.

(b) Supporting Number of Callers.

Figure 6.6: **Call-Set of Functionally Relevant SVs. a)** shows the number of SV stratified by type for each patient (X-Axis). **b)** shows the proportion of SVs supported by a single up to all 6 callers.

remaining SVs, we assess the overlap with coding and non-coding annotations located in the affected TAD environments. We filter for SVs overlapping with enhancer annotations (FANTOM5, VISTA, cREs), Gene-bodies (Phenotype-, Limb-, scRNA- or DEGs), Start/Stop Codons, UTRs or TAD boundaries to retain only variants for which potential disease-causing mechanism are interpretable given our *Limb Regulome*.

The set of functionally relevant SVs contains a total of 3,518 variants with a mean of 188.38 SVs per patient. The number of SVs per patient stratified by type and the caller support are shown in Figure 6.6. Of the remaining SVs the majority are Insertions (44.27%) followed by Deletions (22.51%), Inversions (17.22%), and Translocations (11.13%). 70.88% of the functionally relevant SVs are uniquely detected based on PacBio long-read evidence and 85.16% supported by a single caller. Only a marginal proportion of variants is supported by both short-read and long-read callers ($< 2\%$). Overall, there are several notable differences to the singleton call-set of our detection pipeline: the proportion of Translocations has dropped considerably (17.38%), there is a higher percentage of PacBio-specific calls (8.82%), and the proportion of small SVs ($\leqslant$ 500bp) has decreased by 4.51%.

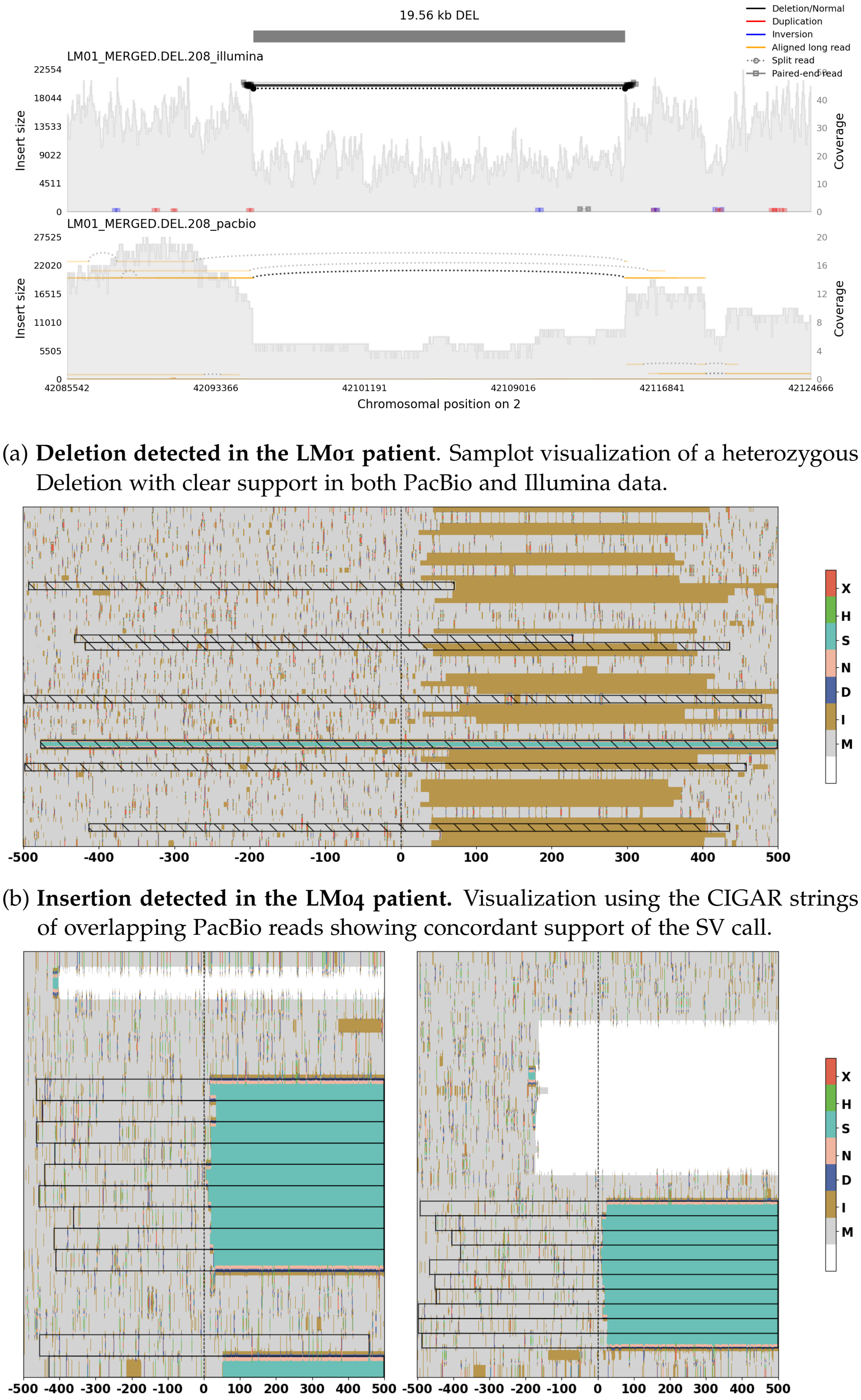
6.2.4 *Sequence-based Validation of Annotated SVs*

Each SV in the remaining set of functionally relevant calls represents a potential candidate variant. However, the call sets still exceed a *manageable* number of variants expected by genetics for a detailed manual inspection. Many of the remaining SVs are supported by a single caller and haven't undergone any further quality-based filtering after the initial SV calling. To investigate the number of *false-positive* SVs among the set of functionally relevant calls, we conduct a sequence-based manual inspection (see Methods for details). Briefly, we generated visualizations of all Deletions, Duplications, and Inversions using *samplot* [106]. For Translocations and Insertions, we employed an in-house method developed by Nico Alavi at the MPIMG. Examples of the visualization are shown in Figure 6.7. Using a custom web application we then inspected the functionally relevant SVs of all patients discarding any potential *false positives*. Through manual inspection, we reduced the set of SVs for the entire cohort to 400 with an average of 19 variants per patient. With this, we have achieved a reduction of $> 99\%$ in comparison to the initially merged call-set as shown in Figure 6.8. 45.95% of the remaining SVs or *candidates* are Deletions, 37.21% Insertions , 11.45% Translocationss and 0.84% Inversions. The majority of the SVs were detected by PacBio callers (64.25%), 19.74% by Illumina callers, and 16.00% by both technologies. Of the PacBio callers, SVIM supported the highest number of final candidates (255), followed by Sniffles (228) and PBSV (206). Illumina callers supported on average (177) fewer variants than PacBio callers ranging from 100 (LUMPY) to 157 SVs (MANTA). While the proportion of calls supported by single callers has decreased considerably in comparison to the set of functionally relevant calls 35.52% of the *true-positive* SVs have still been uniquely detected by one caller. In stark contrast to the set of functionally relevant SVs, 70.25% of the candidates are $\leqslant$ 500bp with 4 candidate SVs $\geqslant$ 100kb across the entire cohort indicating a high rate of *false-positive* calls among medium and large variants.

6.3 CANDIDATE SVS IN THE LIMB REGULOME

With the reduced set of candidate SVs, we set out to investigate which functional annotations are affected. First, we conducted an exploratory analysis, computing the variant counts across functional annotations stratified by the four gene sets. We grouped several of the annotations into a total of 8 categories: *TADs* (domains containing at least one gene of the corresponding gene-set), *TAD boundaries* (both window-size 10 and 5), *Gene* (gene bodies of the four gene-sets), *Exon* (exons of the gene-sets), *cREs* (human and mice cREs in limb and other tissues), *CTCF* (fibroblast CTCF binding sites), *UTR* (3′ and 5′-UTRs) and *S Codon* (Start- and Stop Codons). The results are shown in Figure 6.9. It should be noted that the variant counts in the figure do not reflect the unique number of SVs overlapping with individual categories since the gene set are not mutually exclusive.

We observe the highest number of candidate SVs in TADs containing *scRNA-Genes*. This is to be expected as this set also includes the highest number of genes. Of these SVs many overlap with functionally relevant annotations i.e. scRNA-genes and cREs. Notably, we also observe a smaller proportion of coding SVs variants located in



(a) **Deletion detected in the LM01 patient**. Samplot visualization of a heterozygous Deletion with clear support in both PacBio and Illumina data.

(b) **Insertion detected in the LM04 patient.** Visualization using the CIGAR strings of overlapping PacBio reads showing concordant support of the SV call.

(c) **Translocation detected in the LM05 patient.** CIGAR-based visualization of the first and the mate breakend of the Translocation indicating the presence of the variant.

Figure 6.7: **Examples of Visualizations for True-Positive SVs**. **a** shows an example for the Samplot visualizations to inspect RD, SR, and PR evidence for Deletions, Duplications and Inversions. **b** and **c** are examples of the custom CIGAR-string-based visualization for translocations and Insertions. Each row represents a single overlapping read. The CIGAR abbreviations are color-coded: M (Match), I (Insertion), D (Deletion), N (Alignment Gap), S (Soft-clipped), H (Hard-clipped), X (Alignment Difference). Sequences with a MAPQ $<$ 5 are indicated with hatched and split-read alignments with standard rectangles.

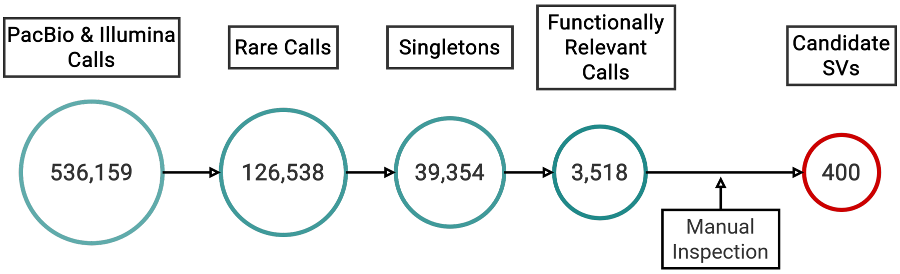


Figure 6.8: **SV Count through Filtering Steps.** The Figure shows the number of SVs across all patients after the comparison with public databases, cohort frequency computation, annotation and manual inspection.

exonic regions of scRNA genes. A similar number of SVs is overlapping directly with TAD boundaries and CTCF sites potentially perturbing regulatory environments of scRNA genes. For known limb-development-associated genes and their corresponding TADs we observe a similar distribution and the second-highest number of overlapping SVs. The majority of these variants are non-coding overlapping with cREs, introns of genes and UTRs. However, a small proportion is also directly affecting the coding sequence of limb genes. For phenotype-associates genes and DEGs we observe a considerably lower number of overlapping SVs. All of these are affecting non-coding annotations including cREs, gene bodies, and TAD boundaries. We can observe a peak i.e. a notably high variant count for a single patient in TADs containing phenotype-associated genes, also reflected in the number of SVs overlapping with gene bodies and cREs. This is due to the high number of phenotype-associated genes of the LM12 patient $(1,325)$ - a consequence of the more general phenotypic description. In comparison, the average of phenotype-associated genes is 172. None of the SVs located in TADs containing a phenotype-associated gene are coding variants.

The exploratory analysis of candidate SVs reveals several variants of interest overlapping coding and non-coding annotations with potentially disease-causing implications. Through the extensive filtering process, the remaining variant set can now be analyzed manually with respect to the affected regulatory environment.

6.3.1 *Prioritization of Shared Variation*

With the exception of two patients (LM17 and LM18) in our cohort, all phenotypes descriptions are unique. This suggests the presence of disease-causing variants that are not shared across individuals. However, even in patients with similar albeit not identical phenotypes the same variant can be pathogenic depending on the penetrance and *variable expressivity* of the corresponding disorder. A disorder is said

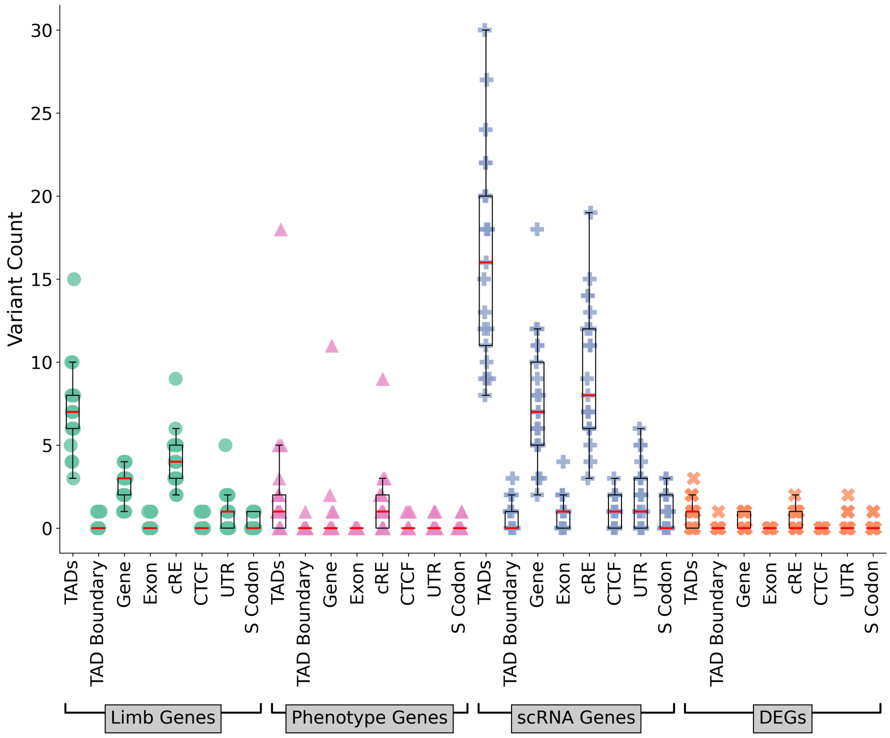


Figure 6.9: **Functionally Relevant SVs stratified by overlapping Annotations.** The Figure shows the number of SVs for each patient overlapping with functional annotations grouped by gene category. Boxes are limited by the 25th and 75th percentile. Median variant counts are indicated in red.

to have a *reduced* or *incomplete* penetrance if certain symptoms are not present in a patient but observed in others with the same genetic variant. While penetrance mainly refers to the absolute proportion of patients with a specific set of symptoms, variable expressivity describes the overall variability in the range of symptoms associated with a disorder. If a disorder, therefore, has increased variable expressivity/*reduced* penetrance, patients can share the pathogenic variants while exhibiting different or no symptoms at all.

To investigate the potential of such shared pathogenic SVs in our cohort we conducted an additional SV prioritization focused on variation detected in 2 or 3 individuals. While it is unlikely that any variant shared between 3 individuals causes divergent phenotypes, we aimed to increase the sensitivity of the analysis with the more conservative threshold. For the prioritization of shared SVs, we used a modified version of our previous prioritization pipeline: First, we selected all SVs in the rare call set detected in 2 and 3 individuals (6,972 SVs). We then converted Insertions back to Duplications, if all initial calls supported the Duplication SV type assignment. For the remaining Insertions, we aligned the inserted sequences to the reference genome, as previously described, retrieving their respective genomic origin, if possible. In the singleton analysis, we collected patient-specific gene sets derived from HPO terms and the RNA-seq analysis. To include a similar set of annotations during the prioritization of shared variation, we generated two gene sets: Phenotype-associated and differentially expressed genes shared by 2 or more individuals. Combining the *shared* gene sets and the remaining annotations of the *Limb Regulome* we again employed TADA for the SV annotation. We then selected any SV located in TADs (*TopDom window-size 10*) containing a gene of interest as described in the patient-specific analysis. The set of functionally relevant shared variation included a total of 569 SVs.

To ensure that we only consider *true-positive* variants in further analysis, we applied adapted versions of the two visualization approaches allowing us to manually inspect the sequencing evidence of individual SVs: We generated *Samplot* figures including Illumina and PacBio reads from all samples in which the SVs were detected for Deletions, Duplications, and Inversions. For Translocations and Insertions, we employed the previously described approach visualizing the aligned PacBio reads and corresponding CIGAR strings at break-point loci adapted to include data from multiple samples. We then manually assessed the support of each functionally relevant variation using the web application. Given the detection of the SVs in several samples, we would expect a comparably high proportion of *true-positives*. However, the majority of SVs we visually inspected were ambiguous variant calls likely due to the highly repetitive content of the genomic regions (an Example is shown in Figure 6.10). In addition, we also labeled any variant calls as *false-positives* if insufficient read support was present in individual samples. The remaining call set included a total of 45 *true-positive* shared SVs supported by read evidence in each associated sample.

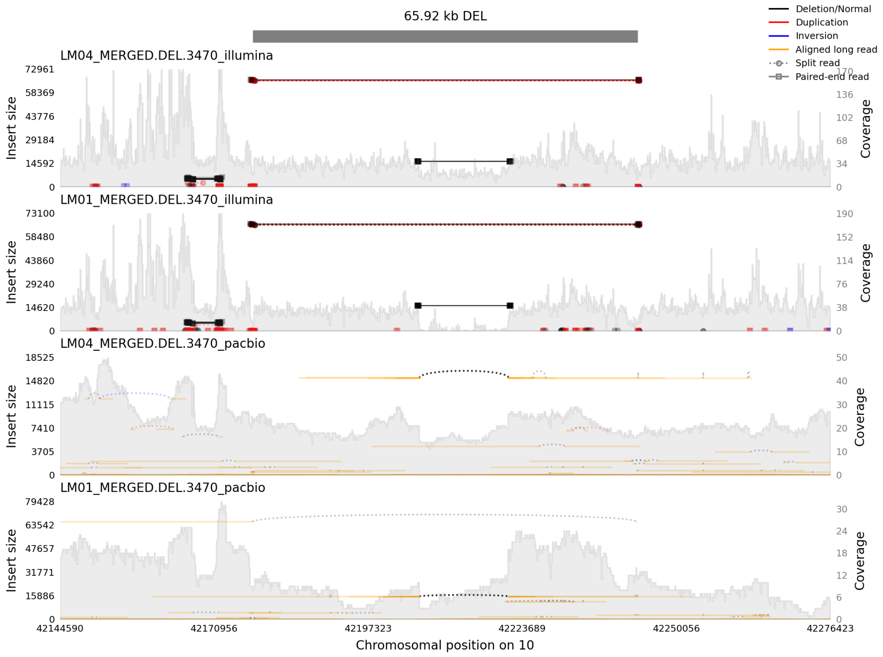


Figure 6.10: **Likely False-Positive Shared Deletion.** This Figure shows the Samplot visualization for a large Deletion detected in two samples (LM01 and LM04). While both short- and long-technologies support the presence of a smaller Deletion as indicated by SR and RD evidence at this locus, only Illumina includes PR evidence supporting the large 65.92kb Deletion. Read pairs located at the same position also indicate a Duplication. Thus, the Deletion call is likely a false-positive.

7 CANDIDATES

The remaining patient-specific and shared variants are validated through manual inspection and affect regulatory elements in the *Limb Regulome*. While this indicates a potential involvement in disease-causing mechanisms, further inspection by geneticists is required to identify strong candidates and speculate on the biological relevance of the overlapping regulatory annotations. However, there are specific challenges to this part of the prioritization for non-coding variants. When inspecting coding variants, the SV type, gene, and affected exon could provide sufficient evidence to speculate on the disease-causing mechanism. Non-coding SVs require a more extensive visualization as the interpretation is not solely dependent on the directly affected regulatory element but on the potentially disrupted interactions with any associated genes. Any non-coding SV, therefore, needs to be analyzed with respect to their entire surrounding regulatory environment. For example, SVs disrupting TAD boundaries can potentially rewire several gene-enhancer interactions. Any gene located in the affected TAD environment and the gene-enhancer interactions, therefore, have to be considered in the manual inspection. This raises the need for methods to visualize SVs and their entire regulatory environment, including annotations relevant to the patient's disease.

There are several methods currently available that address this need with varying degrees of flexibility: the *integrative genomics viewer* (IGV) is a dynamic framework allowing the combination of a wide variety of bioinformatic file formats such as BAM-, BED- and bigWig files [133]. While IGV is a powerful method suitable to visualize any type of genomic data, its application is mainly limited to an interactive interface which would require separate instances for each patient-specific set of regulatory annotations and SVs. In addition, IGV is not able to visualize Hi-C data as a heatmap which is a frequently used representation to investigate TAD structures and potential effects on interaction frequencies at variant loci.

An alternative is the *UCSC Genome Browser*, a web-based application including a comprehensive collection of genomic annotations, a variety of metrics quantifying mutational constraint or conservation, and catalogs of pathogenic and common variation [134]. While it is possible to create user-specific *track-hubs* for the UCSC Browser, a subgroup of our *Limb Regulome* is patient-specific and the visualization should therefore be generated using an application running on a local server. This is possible with an adapted version of the UCSC browser i.e. the *Genome Browser in a Box*. However, similar to IGV this would require several instances of an interactive interface to generate visualization for each patient which is challenging to include in our dynamic *snakemake* framework.

To allow for a dynamic and local visualization process of the patients' SVs in their regulatory environment, we, therefore, implemented a custom script that relies on a hub of patient-specific annotations and returns individual images for each SV without an interactive interface. In this chapter, we briefly describe this approach and present the generated visualizations. The method is part of our overall pipeline and can potentially be modified to any disease context if sufficient functionally relevant annotations are available. To illustrate the potential manual investigation using our visualizations, we present examples of candidate SVs identified in the limb malformation cohort and the affected regulatory environments. Similarly, we apply the visualization method to the filtered shared variation and present the resulting candidates.

7.0.1 *Targeted Visualization Approach*

Our approach is based on *coolbox*, a python package that allows visualizing genomic data as tracks at genomic loci (see Methods for details). We use in our visualization the *Limb Regulome* i.e. coding non-coding annotations described in Table 2 including the gene sets associated with the patients' phenotypes or limb development. Based on the annotations we construct a *browser* with the following features/tracks for each variant loci:

- Hi-C heatmaps derived from the high-resolution cohort and the patient-specific data
- TAD boundaries (*TopDom*) called with window-sizes 5 and 10
- Ensemble gene annotations and transcripts color-coded with respect to the phenotype and limb-associated genes
- Tracks of overlapping cREs, Enhancers, UTRs, Stop- and Start-Codons or CTCF sites
- True-positive and functionally relevant SV calls in the patient

Examples of these visualizations with potentially disease-causing variations are shown in the next section.

7.0.2 *Final Candidates*

The primary function of the visualization is to assist geneticists during the manual inspection of candidate variants. In this section, we therefore only present candidate variants and affected regulatory regions without further speculation on the disease-causing mechanisms. We also passed these candidates on to geneticists familiar with limb malformations. To focus on potentially interpretable candidates in this section, we limit the initial set of candidate variants to a subgroup of examples excluding SVs affecting intronic regions of scRNA, phenotype or Limb genes not overlapping additional regulatory elements.

LM01

The patient was presented with a mirror image polydactyly. In our analysis, we detected a total of 39 *true-positive* SVs overlapping with relevant annotations in the *Limb regulome*. Among these candidates, we identified a 19.56kb Deletion hitting a TAD boundary (window size 5 and 10), multiple FANTOM5 enhancers, and dELS derived from limb tissue in Mice (Figure 7.1). The affected regulatory region includes the *PKDCC* gene which enables on-membrane spanning protein tyrosine kinase activity and is associated with several limb abnormalities (HP:0100491, HP:0002813, HP:0002814, HP:0002817). While no non-coding pathogenic SVs are known, multiple pathogenic nonsense and splice-donor SNPs have been previously observed in patients with rhizomelic limb shortening and dysmorphic features. The variant was detected by Illumina SV callers but was not discussed in any previous analysis of this patient. Notably, the Deletion is not part of any PacBio call-set even though there is clear visual support as shown in Figure 6.7a. The variant also has not been filtered out during the post-processing steps of our pipeline given that there is no similar reported SV in the initial call sets of any PacBio caller.

LM02

In this patient with a triphalangeal thumb on the right and thumb hypoplasia on both hands, we identified 15 candidate variants. The results are based on PacBio sequencing alone as no short-read WGS data was available for this patient. We detected a small 51bp Deletion affecting a UTR of the *PDX1* gene. While *PDX1* is not associated with the patient's phenotype or limb development it is located in the same TAD environment as limb-gene *CDX2*. The affected UTR is linked to the *CDX2* via a significant interaction derived from the PLAC-seq data.

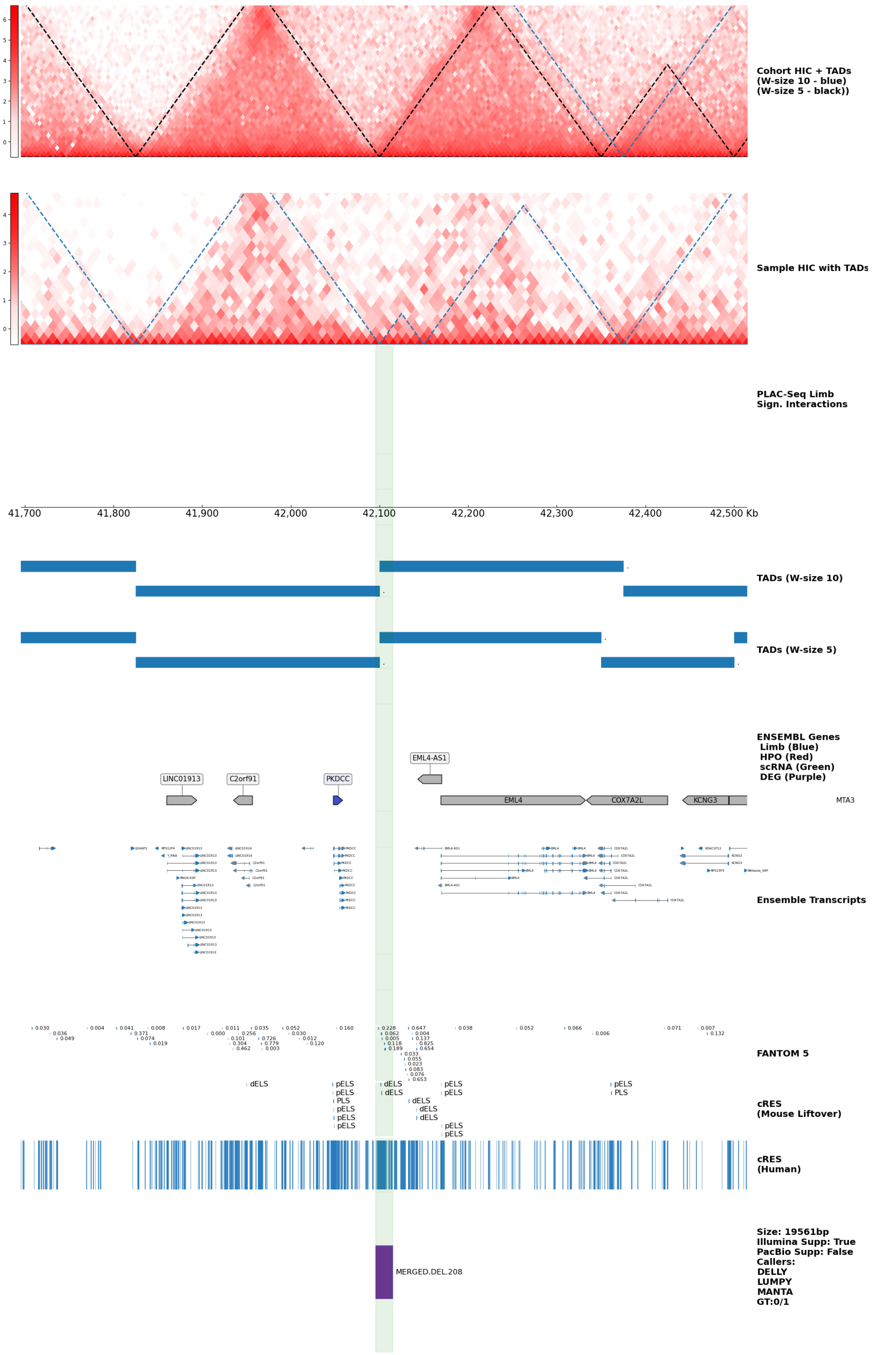


Figure 7.1: **Candidate Deletion detected in LM01.** The variant overlaps a TAD boundary and several regulatory elements in close proximity to a limb-development-associated gene.

LM03

Among the 43 *true-positive* and functionally relevant SVs in this syndactyly patient, we identified several interesting candidates: a 612bp Deletion detected in the intronic region of a DEG which is in the proximity of the phenotype associated gene *ZNF81*. Interestingly, the

Hi-C data at this locus indicates a much larger homozygous Deletion, which was not accurately detected by any of the callers likely due to low coverage at the breakpoint positions. The affected regulatory environment is shown in Figure 7.2. The smaller 612bp Deletion was initially detected by PBSV and Sniffles i.e. it was unique to the PacBio call set. In addition, we identified a coding SV (51bp Deletion) located in the exonic region of a scRNA gene *UNC5B* and a 86bp Insertion overlapping with a significant PLAC-interaction associated with the phenotype gene *IFT81*.

LM04

In this patient with oligodactyly of both hands and syndactyly of toes 4 and 5 on both feet we detected 22 candidate SVs. An example is a 51bp Insertion only detected by Sniffles which affects a cREs linked through the PLAC-seq data to the limb-development associated gene *C2CD3*. The gene is involved in centriolar distal appendage assembly. Coding SNPs have been identified in patients with orofaciodigital Syndrome XIV.

LM05

In the LM05 patient, we detected 31 *true-positive* SVs overlapping with functionally relevant annotations. An example is a homozygous Translocation between chr8 and chr12 identified by all three PacBio callers. The Translocation is located in the coding region of scRNA gene *UHRF2* which is involved in cell cycle regulation. Another prominent SV is an 11kb Duplication detected by all callers with the exception of Lumpy. The Duplication directly overlaps several cREs, UTRs a CTCF site, and a Stop Codon. The affected TAD also includes the limb development associated gene *DUSP3*.

LM06

The 17-1946 patient presented a wide range of limb malformations including cutaneous finger syndactyly, aplasia/hypoplasia of the middle phalanges of the hand, and tibial deviation of the 2nd toe. Among the 23 candidate SVs, we detected 3 variants that affect multiple relevant regulatory elements: a 55kb Deletion initially called by Illumina SV callers but with visual support in PacBio data. The heterozygous Deletion overlaps several regulatory elements including a CTCF site, FANTOM5 enhancer, and multiple cREs. It also is in close proximity to the limb development associated gene *SNX10*. Pathogenic small variants in the SNX10 gene which is involved in intracellular trafficking have so far only been linked to osteoporosis. We also identified a PacBio-specific 333bp Insertion located in the same TAD environment as the limb development gene *PHOSPHO1* and DEG *ABI3* overlapping with a UTR and cREs. Finally, Sniffles detected a 51bp Inser-

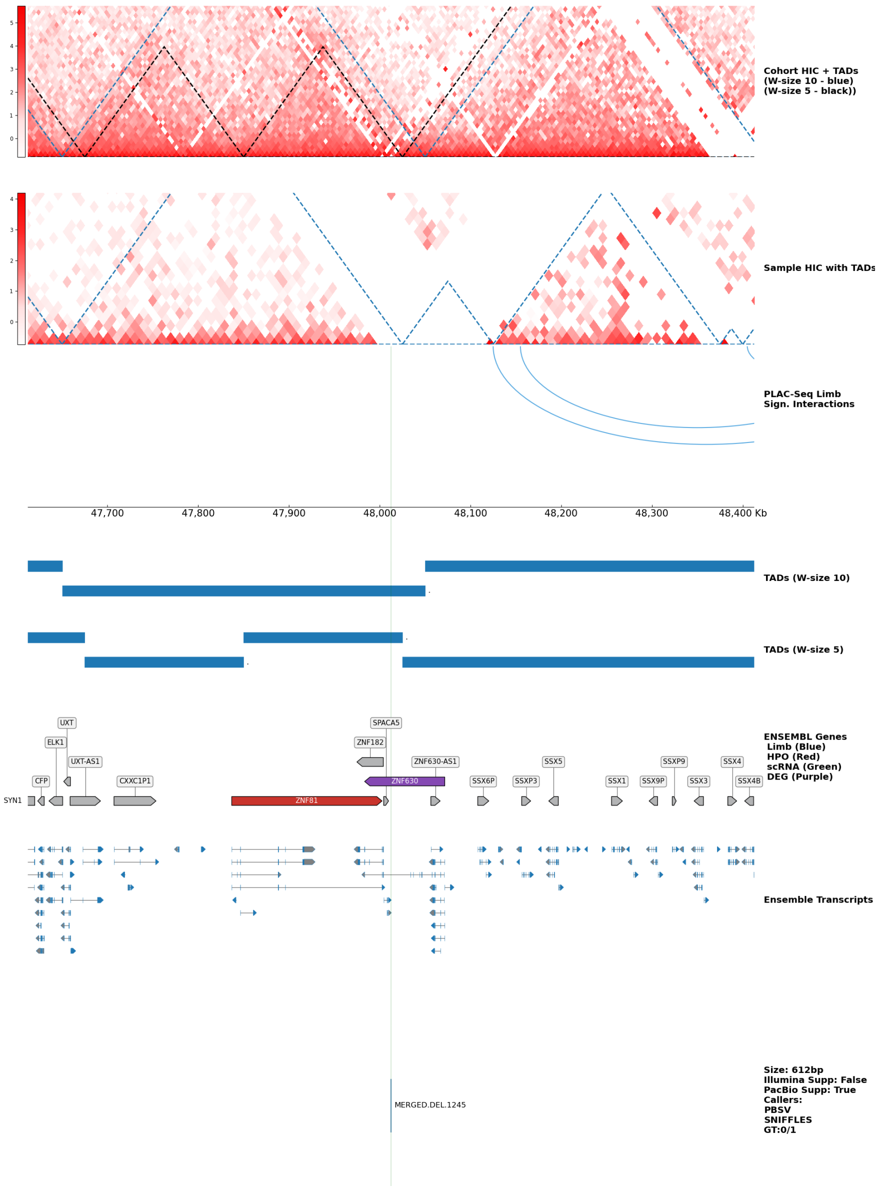


Figure 7.2: **Candidate Deletion detected in LM03**. While the actually reported variant is located in the intronic region of a DEG (ZNF630) it does not overlap with any other annotations. The Hi-C data, however, indicates a larger homozygous deletion at the same locus.

tion affecting multiple regulatory elements in the same TAD as the phenotype-associated gene (*SCN1B*) and the limb gene *LGI4*.

LM07

In this patient with radius and thumb aplasia in both hands, we detected 16 candidate SVs. Among them is an 81bp Deletion detected by 5 callers and therefore supported by both technologies. The Deletion overlaps with a cRE and a 3′ UTR of the limb gene *WNK1*. However, so far no pathogenic SNPs or SVs have been identified in limb malformation patients that affect *WNK1* which is suspected to be controlling the transport of sodium and chloride ions. In addition, we 4 callers detected a 15.85kb Duplication upstream of the scRNA gene *ANK2* affecting multiple FANTOM5 enhancers including two with increased conservation scores of 0.493 and 0.450 and cREs.

LM08

We detected a comparatively low number of 12 candidate SVs in the 18-5002 patient which was presented with a preaxial hand polydactyly. One of the candidate SVs is a 176bp Insertion detected by Sniffles that is located in the intronic region of the limb development gene *SULF2* and also directly affects a cRE. *SULF2* encodes for an endosulfatase acting on heparan sulfate. However, no pathogenic SNPs or SVs are known that directly linking it to limb malformations.

LM09

The patient was presented with a radius and thumb aplasia. We detected a total of 14 candidate SVs. Among them, a 2.1kb Deletion was detected by both PacBio and Illumina callers. The Deletion is hitting the intronic region of the limb gene *RARB* which encodes for a member of the thyroid-steroid hormone receptor superfamily of nuclear transcriptional regulators. It also directly overlaps with a cRE and interestingly with a ribosomal pseudogene (RNA5SP126) which we can be seen in the transcript track of our visualization (Figure A.8).

LM10

In this patient with syndactyly in toes $1-3$ in the right foot and $2-3$ in the left foot, we identified 23 candidate variants. Delly and Manta detected a 328bp Deletion that is visually supported in the PacBio data as well. The Deletion overlaps with a TAD boundary (window size 5 and 10) of the TAD environment containing *ARCN1* and *DDX6*, both phenotype associated-genes.

LM11

The LM11 is one of the two patients that have been previously solved in the WGS analysis. The patient was presented with cutaneous syndactyly of the fingers III-IV and the toes II-III. During the analysis of SNPs and InDels a paternally inherited frameshift variant in *ALDH1A2* was identified which is a direct target of *HOXD13* and involved in vertebrate digit development. Regardless, we included the patient in our analysis and also identified 16 SVs that affect functionally relevant annotations and passed our manual inspection. The SVs include a 23.09kb heterozygous Duplication detected by PacBio and Illumina Callers overlapping several FANTOM5 enhancers, CTCF sites, and cREs. The Duplication is located in the same window-size 5 TAD annotation derived from the cohort Hi-C map as the phenotype-associated gene *BCOR*. However, in this visualization, the Hi-C heatmap also indicates several substructures between the Duplication and the phenotype gene (Figure 7.3).

LM12

The phenotypic description of this patient included a cleft Lip/Palate, skeletal muscle atrophy, and lissencephaly. The more general set of corresponding HPO terms resulted in the highest number of phenotype-associated genes across all patients. Thus, there is also a comparably higher number of SVs overlapping TADs containing a phenotype-associated gene (Figure 6.9). Among the 35 candidate SVs, we identified a potentially disease-causing 121 bp Deletion and a 57bp Insertion. The Deletion is located in the intronic region of the phenotype-associated *COL5A1* gene which encodes an alpha chain for the low abundance fibrillar collagens. In addition, the Deletion directly overlaps with a cRE. The SV call is uniquely supported by PacBio and therefore has not been part of any previous analysis. This is also the case for the 57bp Insertion which overlaps a cRE located in the intronic region of the phenotype-associated gene *TGFA* encoding a growth factor involved in cell proliferation, differentiation, and development.

LM13

In this patient with radius aplasia and thumb hypoplasia, we detected 24 candidate SVs. Among them is a 1kb Deletion located in a TAD environment containing the limb development-associated gene *IHH* which encodes a preproprotein that plays a role in bone growth and differentiation. Variations in *IHH* have been previously determined to be the cause of brachydactyly type A1. The Deletion overlaps with a TAD boundary identified in the patient Hi-C data and a cRE. Another candidate variant, a 2kb Insertion is hitting an exon of the limb-gene *IFT88*.

LM14

This patient was presented with forearm reduction defects. Through our singleton analysis, we detected 16 candidate SVs including two Deletions located in the TAD environment of scRNA-gene *CDC37L10*(Figure 7.4). The larger 180kb Deletion overlaps with the entire gene also affecting several CTCF sites, UTRs, FANTOM5 enhancers, and cREs. Notably, the variant is supported by 5 callers and called as homozygous Deletion while the patient's Hi-C does not reflect this.

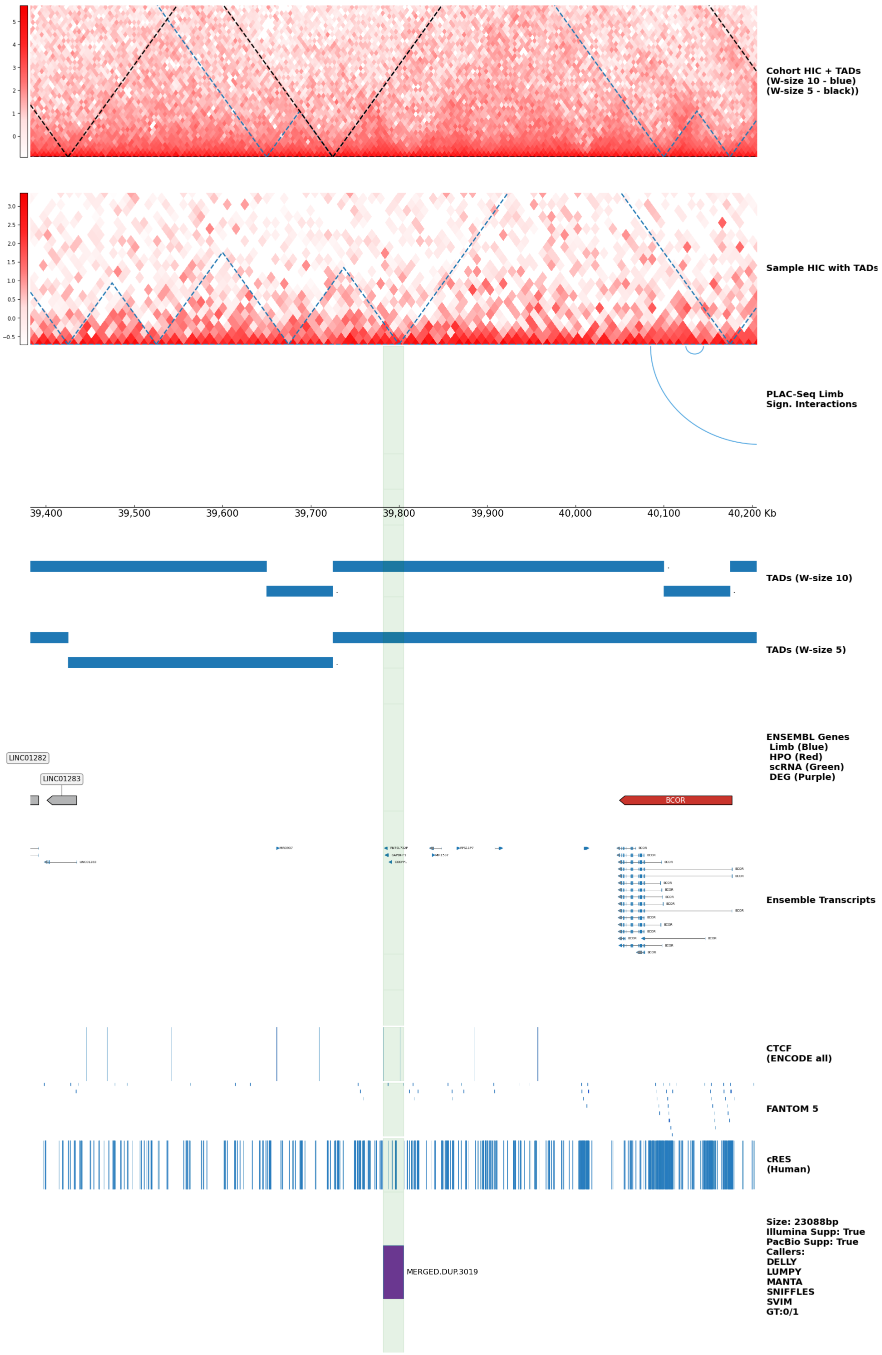


Figure 7.3: **Candidate Deletion detected in LM11**. The 20.09kb Duplication is supported by 5 callers and overlaps several relevant regulatory elements in the same TAD annotation as a phenotype-associated gene.

LM15

This patient was presented with several limb malformations including hypoplasia of the proximal phalanx of the 2nd and 3rd toe and radial deviation of the 4th and 5th finger. We identified a total of 28 candidate SVs including a 10kb Deletion located in proximity to the phenotype-associated gene *PTPN11*. The Deletion overlaps with several cREs and a CTCF site derived from fibroblast. While the initial call was unique to the Illumina callers there is clear evidence in the long-read sequencing data for this variant

LM16

For the LM16 patient with partial syndactyly in fingers 3 and 4 of both hands, we detected 12 potentially disease-causing variants. Among them is a 3.75kb heterozygous Deletion supported by both short- and long-read callers. The Deletion directly overlaps with the limb-gene *SGLP1* involved in the apoptotic signaling pathway. While it does overlap with a cRE there is no direct overlap with any coding sequence.

LM17 & LM18

Both LM17 and LM18 were initially described as patients of a subcohort with a shared limb malformation i.e. absent radius. We detected 25 (LM17) and 15 (LM18) candidate variants in our singleton analysis. However, we did not detect any shared functionally relevant variation that passed the manual inspection. Among the candidate variants from our singleton analysis for LM17 is a 50bp Insertion hitting a TAD boundary and cRE of the TAD environment containing limb-gene *RPL26*. The same TAD also contains the differentially expressed *RNF222*. For LM18 we identified among other candidates a Translocation affecting a cRE derived from embryonic limb tissue in mice in the same TAD environment as limb-gene *PKDCC*. We also detected two Deletions hitting an exon of scRNA-gene *PHF20* and limb-gene *CSF1*, respectively.

LM19

In this brachydactyly patient we detected 21 candidate SVs including a 52 bp Insertion affecting a cRE upstream of the phenotype-associated gene *GJA*. The gene is a member of the connexin family and is involved during embryonic development. Pathogenic coding SNPs have mainly been detected in patients with oculodentodigital dysplasia. The reported Insertion is uniquely identified by SVIM. However, it has passed the manual inspection suggesting it is a *true-positive* variant call.

LM20

For the oligodactyly patient LM20, we identified 22 *true-positive* and functionally relevant SVs. Among them is a Translocation supported by all PacBio callers located in a TAD environment containing the limb-gene *PLK4*. The Translocation affects a cRE element - specifically, a PLS derived from embryonic mice limb tissue. Interestingly, the PLS is linked through a significant PLAC-seq interaction to the limb-gene *INTU* which is located in the upstream TAD environment.

LM21

For this patient no phenotypic information was available. We, therefore, conducted the prioritization solely based on the known limb development associated with and scRNA gene as well as the DEGs identified for this patient. In total, we detected 28 *true-positive* and functionally relevant SVs. Interestingly, among them, there are three large heterozygous Deletions spanning 488kb and 250kb. Both passed our manual inspection and are supported by short-read and long-reads (Figure 7.5).

7.0.3 *Shared Candidates*

We initially identified 45 *true-positive* shared candidate SVs. Of these variants, 15 SVs are located in TADs containing a limb development-associated gene affecting cREs and TAD boundaries. A similar proportion (16 SVs) are detected in TADs with scRNA genes again including a single coding variant. Several SVs are also hitting TADs containing *shared* phenotype-associated genes. However, these variants would only be relevant for further inspection if the genes are also associated with the phenotypes of the patients in which the variants were detected. This is not the case for any of the candidate SVs. We also do not observe any variants located in TADs containing a *shared* DEG. Among the detected shared variation in functionally relevant regulatory environments is an Insertion found in patients LM4 and LM6. The 2.4kb Insertion hits the 3'-UTR of the limb development gene *ENPP1*. We also detect a larger 6kb Deletion affecting multiple cREs in the TAD environment of scRNA gene *PRTG*. Most of the identified candidate SVs (41 out of45) are smaller (< 500bp) than these examples.

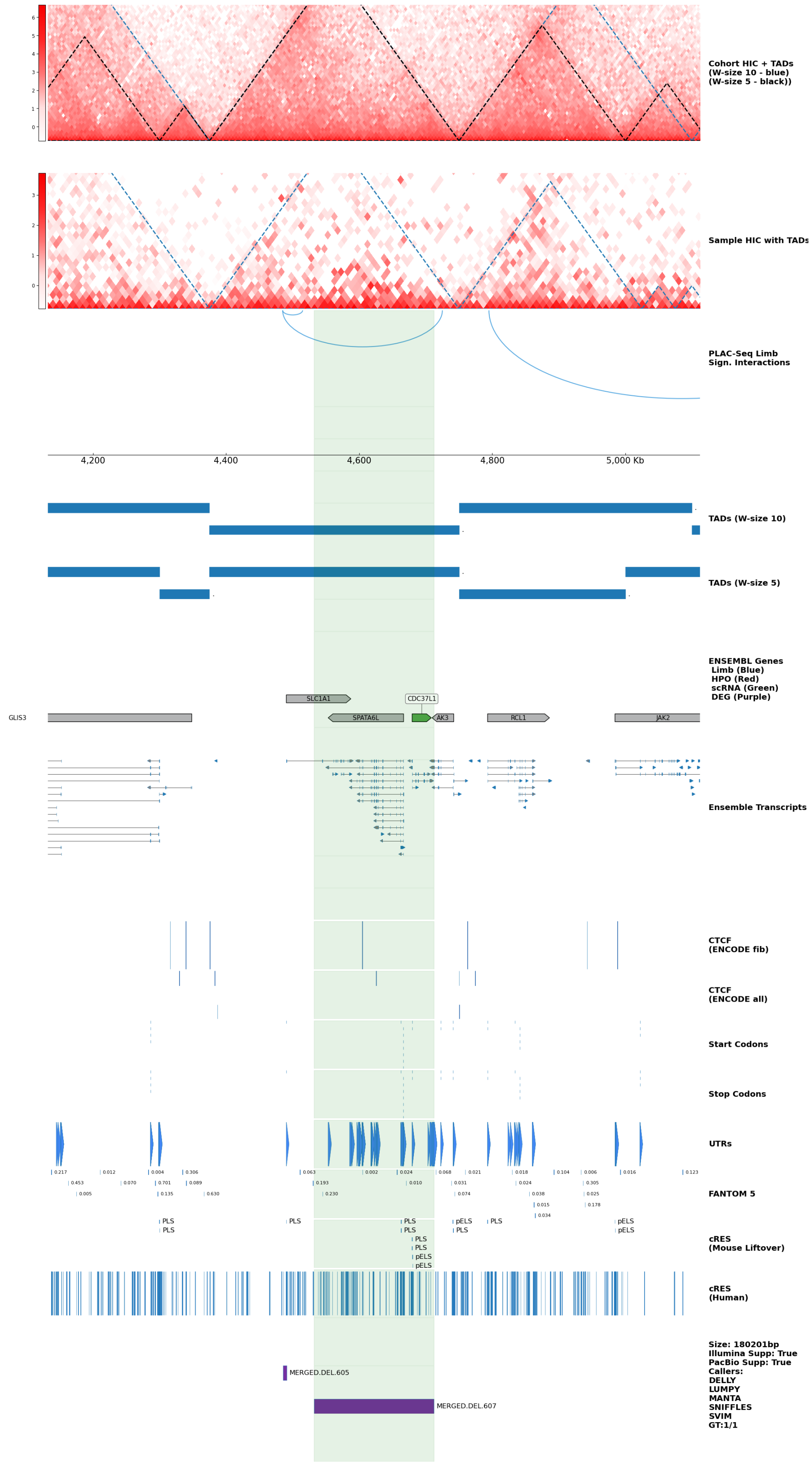
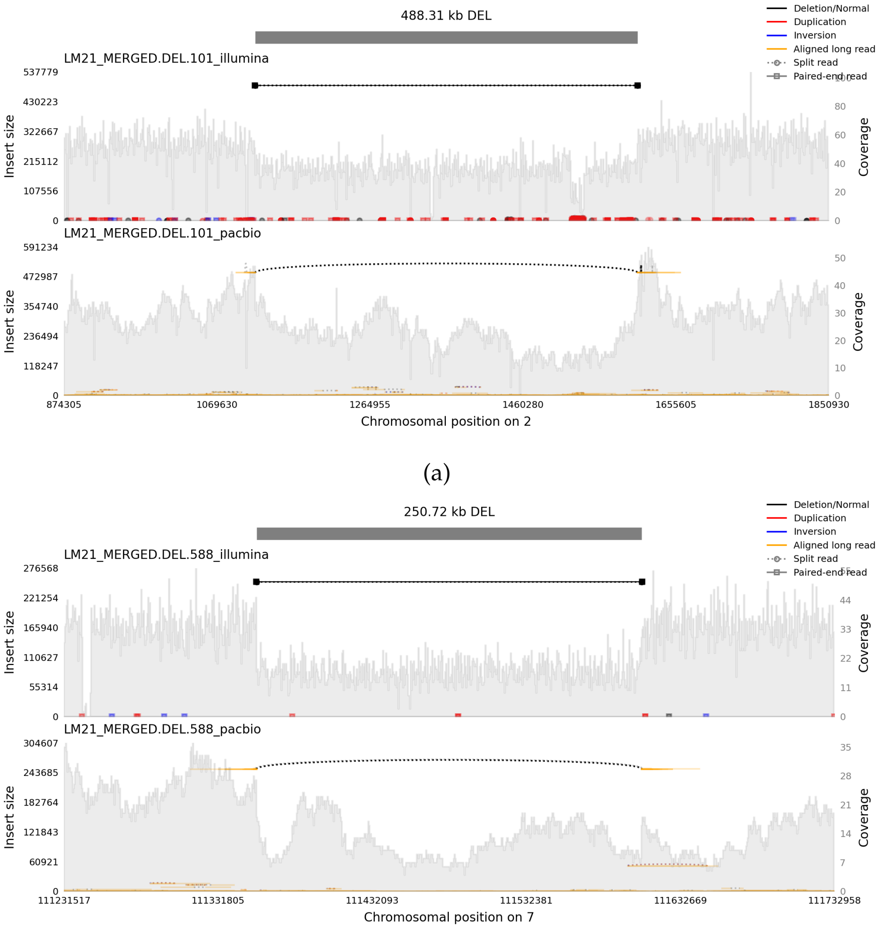


Figure 7.4: **Candidate Deletions detected in LM14**. The larger Deletion directly affects the scRNA and several non-coding annotations derived from humans and embryonic limb tissue from mice.



(a)

(b)

Figure 7.5: **Samplots of Large Deletions in the LM21 Patient. a)** shows the visualization of the 488kb Deletion detected by initially 4 callers including SVIM and all three Illumina methods. **b)** shows the 250kb Deletion called by Illumina callers but supported in the visualization by PacBio reads.

8 DISCUSSION

The investigation of genetic alterations is a fundamental part of clinical diagnostics. Standard experimental procedures used to identify pathogenic variation and reach a molecular diagnosis commonly include microarrays, gene panel sequencing, and WES. These approaches and the corresponding downstream analysis are limited in two significant aspects: 1) they predominantly focus on small genetic variation i.e. SNVs/InDels due to the underlying experimental technologies. Short-read sequencing specifically has been shown to under-perform with respect to the detection of SVs since they are abundant in highly repetitive regions. The majority of SVs are therefore not routinely investigated even though they have been shown to be a significant contributor to human disease. 2) They mainly assess the pathogenic potential of the detected variation with respect to coding sequence which only accounts for approximately 2% of the genome. In recent years a growing number of non-coding pathogenic variants including SVs have been identified indicating the need to expand the standard clinical procedures to the entire genome. However, the quantification of the functional impact in non-coding regions requires extensive knowledge of the involved regulatory elements. Several efforts have been made to curate large-scale catalogs of predicted and validated regulatory elements [59, 63, 64]. In addition, Hi-C and comparable technologies have allowed exploring tissue- and cell-type-specific chromatin conformations which can be disrupted by SVs potentially leading to disease [57, 61, 62]. These resources offer the potential to quantify the functional impact of non-coding variation with respect to a specific disease context and affected tissue i.e. a *functional annotation-based prioritization*.

In this thesis, we discuss the detection and prioritization of SVs on the example of a patient cohort with limb malformations. The patients have been previously analyzed with standard genetic testing and short-read sequencing but remain without a molecular diagnosis. Thus, we set out to further identify potentially pathogenic SVs in an extensive analysis. We performed PacBio long-read sequencing for all patients and additionally conducted RNA-seq and Hi-C experiments to investigate the patients' transcriptomes and chromatin conformations. To process the data we implemented a novel pipeline that combines short- and long-read sequencing detecting SVs using a total of six callers, and prioritizes them based on a disease-specific set of functional annotations. With our pipeline, we were able to reduce the overall set of detected SVs by $\geqslant$ 99% identifying sets of candidates for each patient that affect regulatory environments relevant to the patients' phenotypes.

Further, we conducted an elaborate evaluation of TADA - a machine learning approach for the prioritization of disease-causing CNVs - illustrating the potential of an automated prioritization method agnostic to the disease context. The evaluation showed superior performance of TADA in comparison to multiple current prioritization methods based on ROC-AUC and F1-scores as well as a ranking analysis with respect to the majority of test sets. The underlying training data of TADA includes, however, almost exclusively pathogenic coding variants. Any application is therefore currently limited to variants affecting genes. With an increasing number of validated disease-causing non-coding variants, we aim to expand the classifiers to reliably quantify the functional impact in all genomic regions.

Through a rigorous comparison of SV callers and technologies, we investigated the capabilities and limitations of long- and short-read SV callers. As expected short-read callers are able to detect fewer variants than their long-read counterparts. They also detect a higher rate of likely *false-positives* as indicated by the minor proportion of shared variation between callers and large $\geqslant$ 100kb variants that could not be confirmed in a manual inspection. The majority of *false-positive* SVs $\geqslant$ 100kb were called by Delly and Lumpy. The Manta call set, however, is more similar to PacBio callers and includes the highest number of candidate SVs among all short-read callers. This suggests a more robust performance of this specific approach likely due to the assembly of breakpoint supporting reads that increases the number of identified Insertions and reduces the number of unresolved Breakends. The long-read SV callers shared a significantly higher proportion of variants with similar SV type and size distributions. This is likely due to the shared focus on SR evidence while short-read callers include several types of read evidence. However, the number of manually inspected *true-positive* functionally relevant SVs still differs with SVIM outperforming the other two callers. The choice of SV caller, therefore, influences the call set and the determined candidate SVs significantly, especially with respect to short-read SV callers. Ensembl approaches involving multiple callers account for this variability increasing the sensitivity during SV detection but also potentially decreasing precision. Therefore they should be used in combination with rigorous quality assessment and filtering of *false-positives*.

Following the comparison, we assessed the allele frequency of the detected SVs with respect to public catalogs of common variation. While discarding clusters containing common SVs allowed us to reduce the number of Deletions and Insertions significantly, all Translocation and a high proportion of Inversion remained. Only through a second filtering step employing the allele count quantified in our own cohort, we were able to identify shared instances of Translocations. Such an analysis is however only possible with a sufficient number of samples. While Translocations are less frequent than Deletions or Insertions in the population, they still represent potentially disease-causing variation that should be included in clinical diagnostic frameworks. Especially in analyses based on short-read experiments i.e. the majority of current genetic testing, callers detect an overwhelming amount of unresolved Breakends. Without public resources to determine common and likely benign Breakends/Translocations and reduce the set of candidate SVs, the prioritization of SVs will likely remain limited to other SV types or needs to rely on hard caller/technology support thresholds. In our analysis we avoided these thresholds, suspecting that they also exclude *true-positives* and potentially disease-causing variation. This was confirmed in our manual inspection of functionally relevant SV. The majority of the remaining *true-positive* calls were supported by a single caller and would have been discarded purely based on caller support. In addition, for many SVs that were initially detected by callers of a single technology we were able to determine supporting reads of both sequencing technologies in the manual analysis. While this illustrates the limitations of hard thresholds on caller/technology support, the manual inspection is not feasible for larger cohorts. In addition, most patients are not currently investigated using both short- and long-read data. Any manual analysis would therefore rely on short-read data alone which likely limits its capabilities to determine *true-positive* SV - especially for Insertions and Translocations. This highlights the need for automated methods that potentially learn from multiple sequencing technologies to determine *false-positive* SV calls in short-read data.

We collected multiple sets of relevant coding and non-coding regulatory elements associated with limb development for a semi-automated functional annotation-based prioritization approach. Among them, we used the TAD boundaries derived from the patient-specific Hi-C experiments and DEGs from a *one vs. all* comparison of the RNA-seq data. We modified TADA to include all types of SVs and developed a new set of features corresponding to the disease context of our cohort. The annotation and subsequent filtering allowed us to reduce the SVs to those affecting regulatory environments relevant to limb development including several candidate SVs for each patient. This part of the pipeline can potentially be adapted to other disease contexts if TAD boundaries and sets of associated regulatory elements are available. Finally, we included visualization approaches in our pipeline that allow for the inspection of individual candidate variants. For each patient, we report multiple examples of potentially disease-causing candidates.

The successful identification of disease-causing SVs in the limb malformation cohort suggests the potential of our pipeline and specifically long-read sequencing for assisting in clinical diagnostics. There are however several limitations that should be addressed: The current high cost of long-read sequencing limits the application of the pipeline to research projects with the necessary resources. Any application in a clinical setting will require modifications toward a short-read-specific analysis. While we are able to determine sets of candidate SVs that are both confirmed in a manual inspection and affecting relevant regulatory elements, we did not perform any additional validation of the variants. Any presented candidates should therefore be tested in further analysis. This is particularly evident by the fact that we identify several candidate variants for the patients LM11 and LM15. For both patients, disease-causing small variants were already identified in the WGS analysis conducted by Elsner et al. [83]. The candidate SVs should therefore be considered with care and certainly require further investigation by geneticists familiar with the patient's disease. In our analysis, we have not compared the performance of our pipeline to any other approaches. This is due to the lack of methods that allow for a disease-context-specific prioritization of SVs. Still, a comparison could be performed with similar automated and semi-automated approaches. Even though several instances of disease-causing complex SVs are known we have not analyzed any in our analysis since not all callers are able to detect them. Also, our annotation framework is not currently able to process complex SVs. This could be considered in potential extensions of our pipeline.

Our in-depth analysis can serve as an example for future studies focused on the prioritization of SVs specifically those affecting non-coding sequences in genetic research and clinical diagnostics. With a continuously growing number of investigations on the tissue-specific function of non-coding elements, our functional annotation-based prioritization can be extended to a wide range of human diseases and could prove to be a valuable approach potentially increasing the number of successful diagnoses.

9

TERMS AND ACRONYMS

AC	Allele Count
AER	Apical Ectodermal Ridge
AF	Allele Frequency
AP	Anterior-Posterior
ALU	Arthrobacter luteus
AUC	Area Under the Curve
BAM	Binary Alignment and Map
BND	Translocation / Breakend
bp/kb/mb	Base-pairs/Kilobase-pairs/ Megabase-pairs
BWA	Burrows Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CAGE-Seq	Cap Analysis Gene Expression Sequencing
CCR	Continous Consensus Read
cDNA	Complementary DNA
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CLR	Continous Long Read
CNV	Copy Number Variant
cRE	Candidate Regulatory Element
CTCF	CCCTC-binding factor
CV	Cross Validation
DD	Developmental Disease
DDG2P	Development Disorder Genotype - Phenotype Database
ddNTP	Dideoxynucleotide
DECIPHER	Database of Genomic Variation and Phenotype in Humans using Ensembl Resources
DEG	Differentially Expressed Genes

DEL	Deletion
dELS	Distal Enhancer like Signature
DUP	Duplication
DV	Dorsal-Ventral
ECDF	Emperical Distribution Function
ELS	Enhancer like Signature
ENCODE	Encyclopedia of DNA Elements
FANTOM	Functional Annotation of the Mouse/Mammalian Genome
FC	Fold Change
FGF8/FGF10	Fibroblast Growth Factor
FISH	Fluorescence in Situ Hybridization
GIAB	Genome In A Bottle
GnomAD	Genome Aggregation Database
GRCh37/GRCh38	Genome Reference Consortium Human Build 37/38
GWAS	Genome Wide Association Study
hESC	Human Embryonic Stem Cells
HI	Haploinsufficiency
Hi-C	High-throughput Chromosome Conformation Capture
HMW	High Molecular Weight
HOX	Homeobox
HPO	Human Phenotype Ontology
InDel	Insertion/Deletion up to 50bp
INS	Insertion
INV	Inversion
KCNJ2/KCNJ16	Potassium Inwardly Rectifying Channel Subfamily J Member 2/16
LD	Linkage Disequilibrium
LINE	Long Interspersed Nuclear Elements
LOUEF	Loss-of-Function Observed/Expected Upper Bound Fraction

LM01-LM21	Limb Malformation Patient 1-21
ML	Machine Learning
MPIMG	Max Planck Institute for Molecular Genetics
NCBI	National Center for Biotechnology Information
NCLS	Nested Containment List
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PBSV	PacBio Structural Variant Calling Tools
PCR	Polymerase-Chain-Reaction
PD	Proximal-Distal
pELS	Proximal Enhancer like Signature
PLS	Promotor like Signature
PR	Paired Read
QC	Quality Control
RD	Read Depth
ROC	Receiver Operator Curve
scRNA	Single Cell RNA
SD	Standard Deviation
SHH	Sonice Hedgehog Gene
SOX9	SRY-Box Transcription Factor 9
SMRT	Single Molecule Real Time
SNV	Single Nucleotide Variants
SR	Split Read
SV	Structural Variant
SVIM	Structural Variant Indentification Method
TAD	Topologically Associating Domain
TADA	TAD Annotation
TBX4/TBX5	T-box transcription factor 4/5

T2T	Telomere to Telomere
UCSC	University of California Santa Cruz
UTR	Untranslated Region
VCF	Variant Call Format
VEP	Variant Effect Predictor
WGBS	Whole Genome Bisulfite Sequencing
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WNT7A	Wingless-Type MMTV Integration Site Family, Member 7A
ZMW	Zero Mode Waveguide
ZPA	Zone of Polarized Activity

A

SUPPLEMENTARY FIGURES

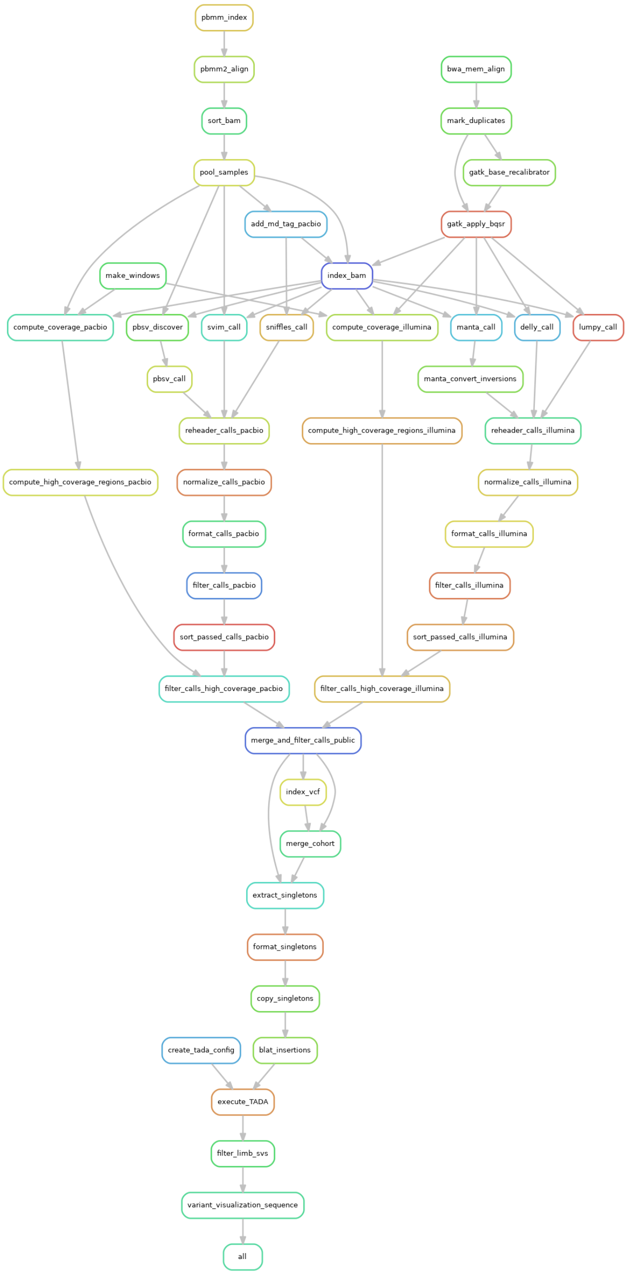
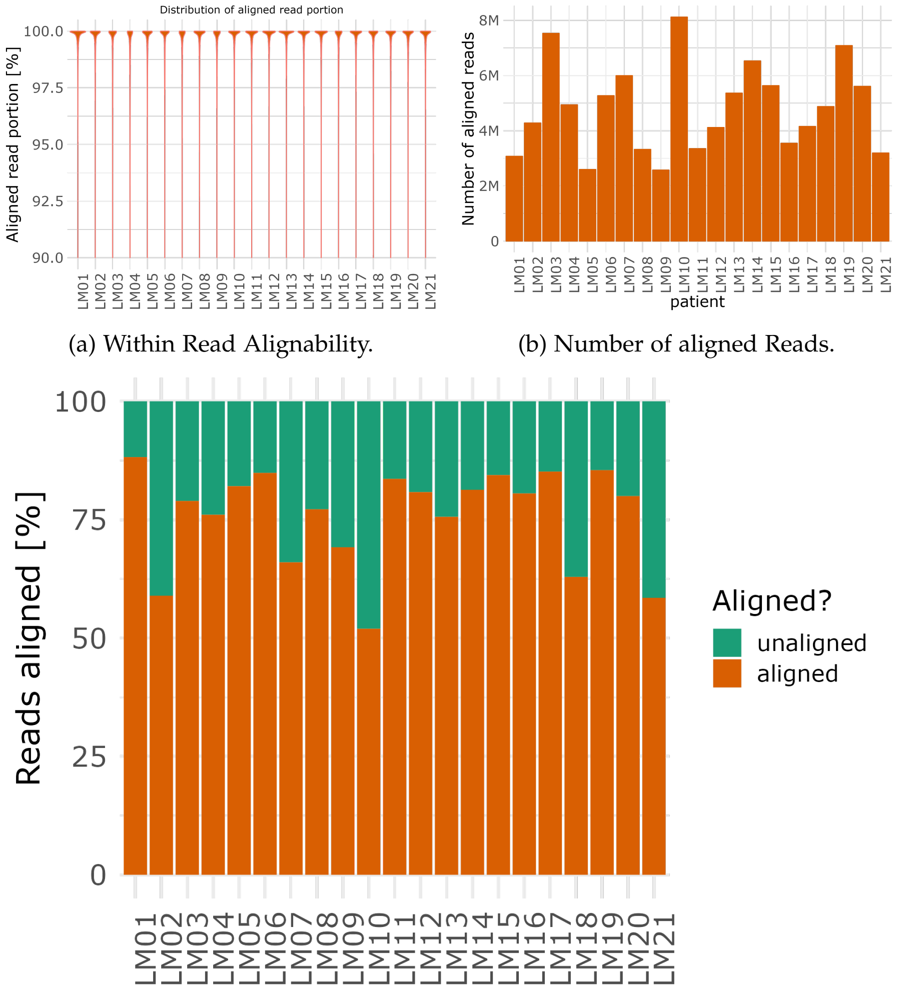


Figure A.1: **DAG of the snakemake pipeline.** The figure shows the rules involved in the snakemake pipeline starting from the alignment of short- and long-read sequencing data to the visualization of functionally relevant SVs with respect to the disease-context for manual inspection.

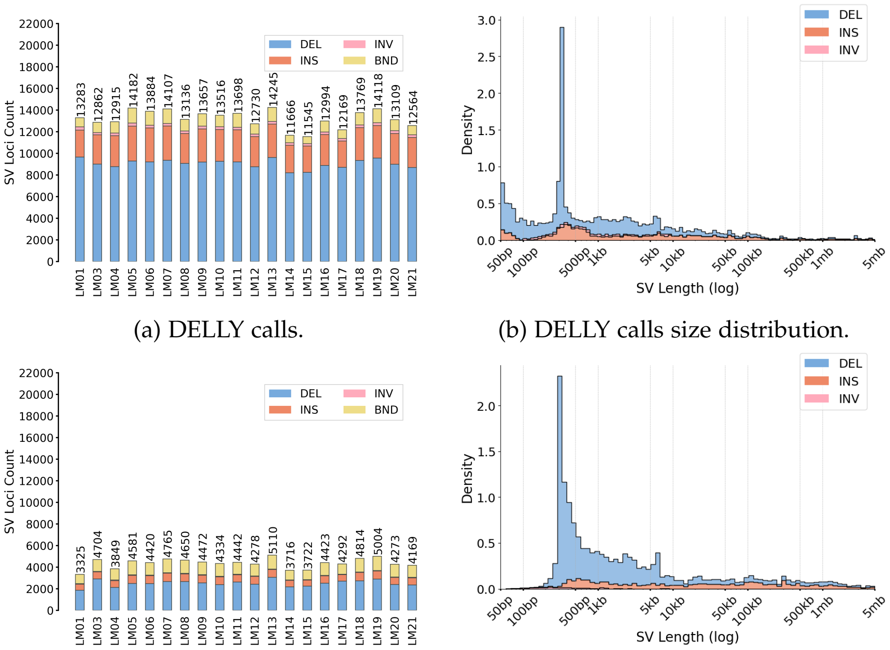


(a) Within Read Alignability.

(b) Number of aligned Reads.

(c) The Fraction of Aligned Reads.

Figure A.2: **Additional Alignment Statistics of the PacBio data.** **a** shows the fraction of aligned reads for each patient. **b** shows the within read align ability and **c** the number of aligned reads.



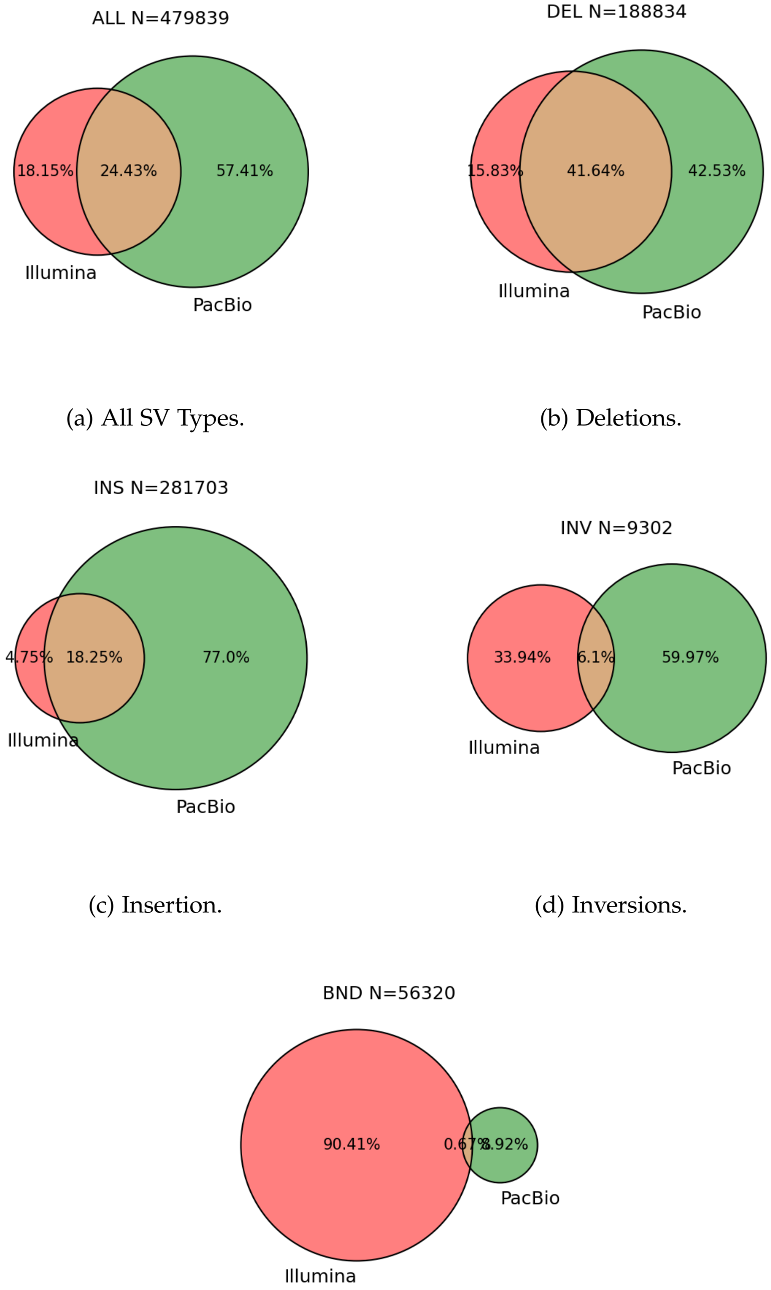
(a) DELLY calls.

(b) DELLY calls size distribution.

(c) Lumpy calls.

(d) Lumpy calls size distribution.

Figure A.3: **Illumina SV Calls and Size Distributions of Unfiltered Delly and Lumpy calls**. The figures on the left side show the number of SV calls grouped by SV type for DELLY and Lumpy, respectively. X-axes show the sample IDs and the Y-Axis the number of SVs with the total number indicated on the top of each bar. The right side figures show the corresponding size distribution.



(a) All SV Types.

(b) Deletions.

(c) Insertion.

(d) Inversions.

(e) Breakends / Translocations.

Figure A.4: **Technology Support stratified by SV Type**. The total number of SVs for each type is shown on top of the corresponding Venn diagram. The individual circles are scaled by the number of variants.

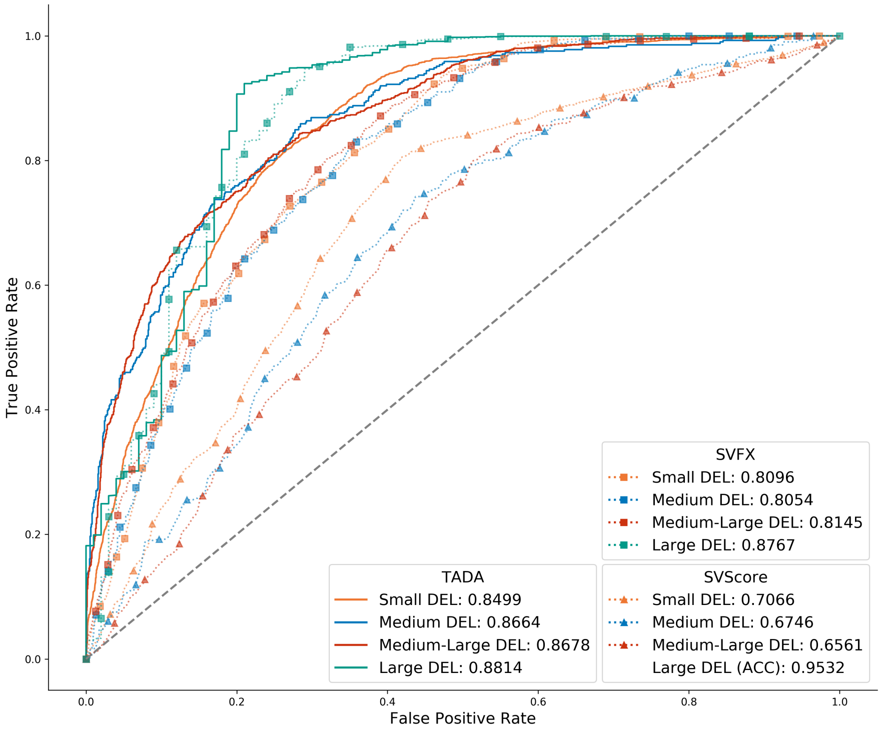


Figure A.5: **Performance Comparison of TADA, SVScore and SVFX on Size-Stratified ClinVar Deletions.** The figure shows the ROC-Curves of all three tools computed for ClinVar variants stratified by size intro four categories: Small (< 50kb), Medium (< 100kb), Medium-Large (< 1mb), Large (>= 1mb).

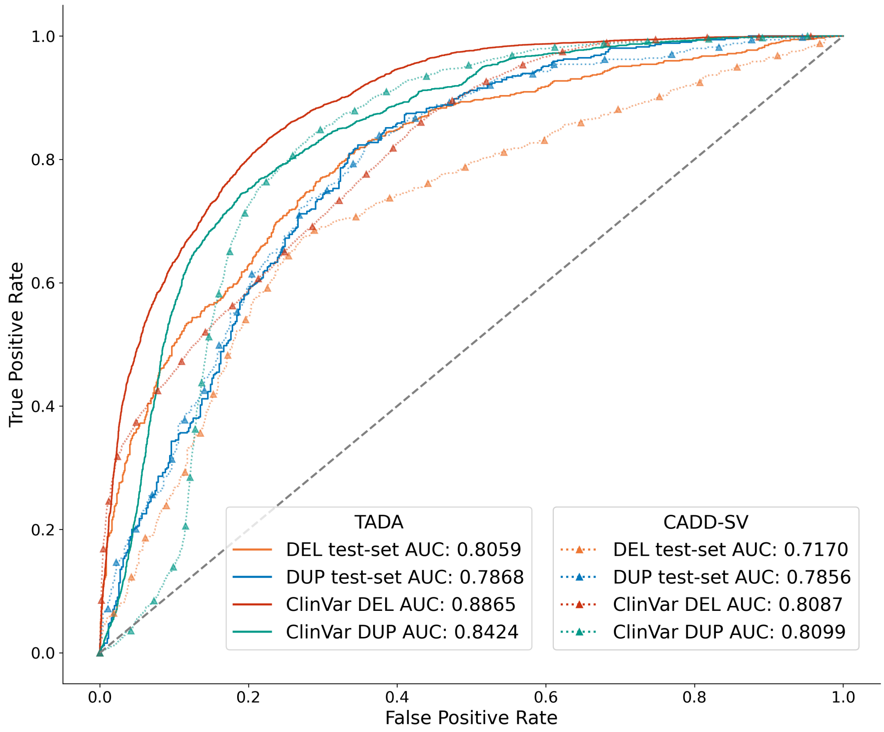


Figure A.6: **Performance Comparision of TADA with CADD-SV.** The figure shows the ROC-Curves of TADA and CADD-SV for the Test-split and ClinVar data.

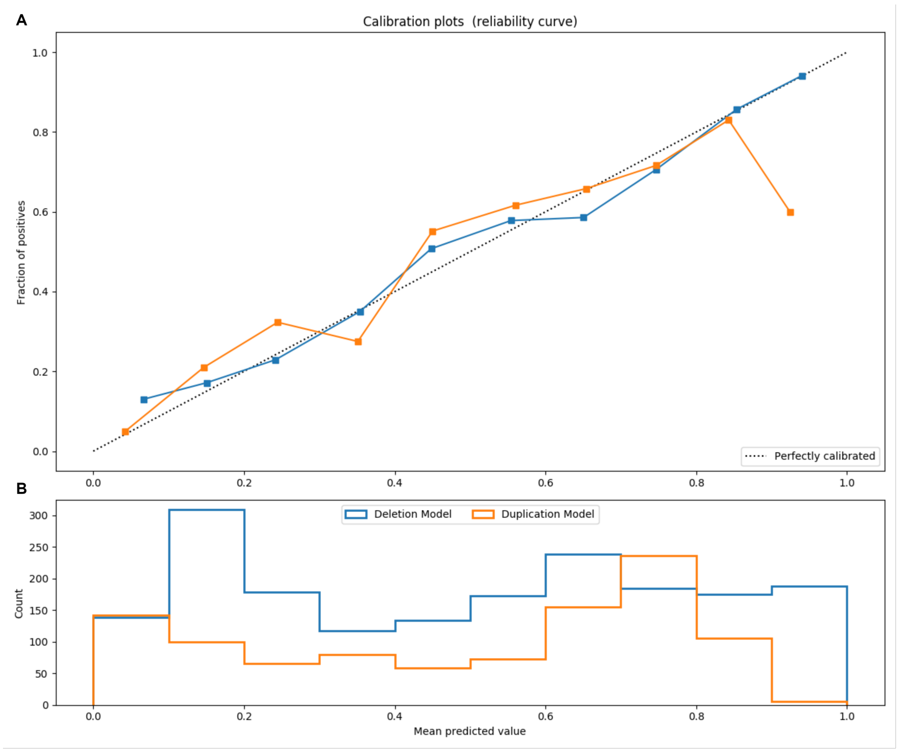


Figure A.7: **Calibration of the Predicted Class Probabilities for the Deletion and Duplication Model. A** shows the fraction of positives vs the mean predicted value. **B** shows the absolute count of variants predicted over mean predictive values. The dotted line in the upper plot indicates perfect calibration.

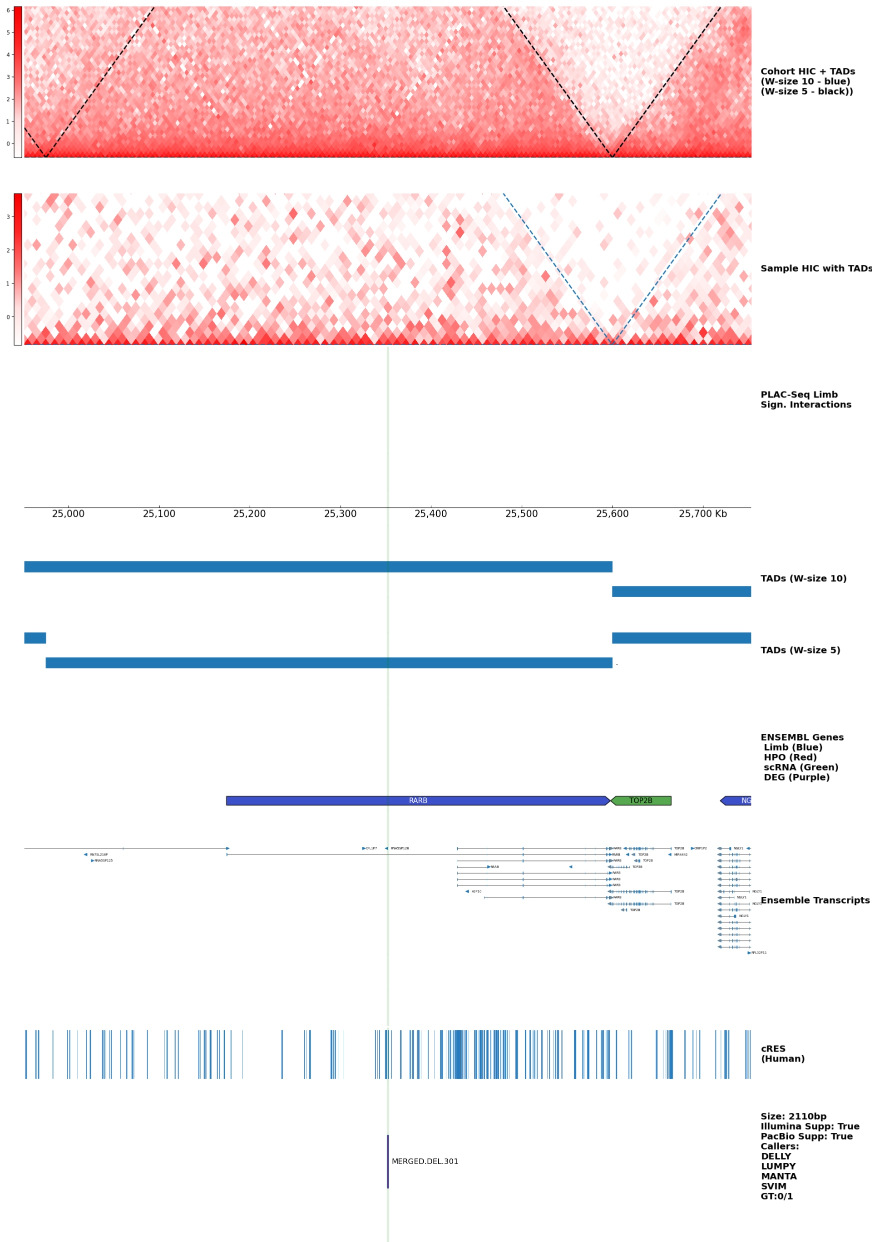


Figure A.8: **Candidate Deletion detected in LM08**. The 2.1kb Deletion is supported by 3 callers and overlaps cREs as well as a ribosomal pseudogene.

B

SUPPLEMENTARY TABLES

Database	Reference	Download Link	Download Date	Number of SVs
GnomAD	GRCh38	https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/vcf/nstd166.GRCh38.variant_call.vcf.gz	18.05.2021	308858
Audano et al. 2019	GRCh38	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20181025_EEE_SV-Pop_1/VariantCalls_EEE_SV-Pop_1/	18.05.2021	96585
Ebert et al. 2021	GRCh38	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel_alt.vcf.gz	05.10.2021	111331
Ebert et al. 2021 (Inversions)	GRCh38	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_inv.tsv.gz	05.10.2021	417
NCBI Common	GRCh38	https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/vcf/nstd186.GRCh38.variant_call.vcf.gz	18.06.2021	82288

Table 3: Data sources of common SVs.

C

ABSTRACT

Current genetic testing of patients performed to identify the molecular cause and potentially drive therapy decisions is predominately focused on small variation affecting coding regions due to the limitations of the underlying experimental methods. Long-read sequencing approaches have been shown to overcome these limitations allowing the detection of the entire spectrum of larger genomic alterations i.e. structural variants (SVs) with an unprecedented resolution potentially revealing previously undetected disease-causing mechanisms. In this thesis, we discuss the potential of long-read sequencing in combination with a functional annotation-based framework to identify non-coding pathogenic SVs in a cohort of limb malformation patients. In the process, we developed a pipeline that combines short- and long-read sequencing data, filters the detected SVs based on allele frequency, and applies an extensive functional annotation-based prioritization resulting in sets of candidate SVs for all involved patients. We also conduct a comprehensive comparison of callers and technologies highlighting the superior performance of long-read sequencing for SV detection and an evaluation of an automated prioritization method indicating superior performance to comparable approaches. The results of this thesis suggest the potential of performing an extended analysis of SVs as part of clinical diagnostics workflows and the relevance of non-coding functional annotation during variant prioritization.

D

ZUSAMMENFASSUNG

Der Standard der Detektion von potentiell krankheitsversachenden Mutationen in klinischen Untersuchungen ist momentan beschränkt durch die angewandten Sequenziermethoden (Illumina Sequenzierung) auf Varianten kleiner als 50bp, die direkt codierende Elemente im Genom beeinflussen. Größere Mutationen auch Strukturvarianten (SVs) genannt, werden in den meisten Analysen nicht betrachtet. Die Sequenziertechnologien der dritten Generation (PacBio, Nanopore) ermöglichen eine akkurate Bestimmten von SVs und haben das Potential vorher unerforschte pathogene Mutationen zu entdecken. Wir untersuchen dieses Potential anhand einer Patientenkohorte mit Entwicklungsstörungen, die sich durch Veränderungen der Extremitäten äußern. Für diese Analyse haben wir eine Pipeline entwickelt, die es erlaubt, sowohl Illumina als auch PacBio Daten zu verarbeiten, zu filtern und mit relevanten nicht codieren regulatorischen Elementen zu annotieren. Mithilfe der Pipeline identifizieren für jeden betroffenen Patienten mehrere seltene Mutationen, die regulatorische Elemente betreffen, welche mit dem Krankheitsbild der Patienten assoziiert sind. Wir demonstrieren außerdem die Vorteile von PacBio in einem ausführlichen Vergleich mit Illumina Daten im Bezug auf die Detektion von SVs und evaluieren eine Methode für die automatisierte Klassifizierung von pathogenen Variationen. Die Ergebnisse dieser Arbeit zeigen, dass die dezidierte Analyse von nicht codierenden SVs zur Entdeckung von vorher nicht erkannten und potentiell krankheitsverursachenden Mutationen führen kann. Die Analyse und die damit zusammenhängende Pipeline können dadurch als Grundlage für zukünftige genetische Untersuchungen verwendet werden.

BIBLIOGRAPHY

[1] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (2015), p. 68.

[2] Joachim Weischenfeldt, Orsolya Symmons, Francois Spitz, and Jan O Korbel. "Phenotypic impact of genomic structural variation: insights from and for human disease." In: *Nature Reviews Genetics* 14.2 (2013), p. 125.

[3] Malte Spielmann and Stefan Mundlos. "Looking beyond the genes: the role of non-coding variants in human disease." In: *Human molecular genetics* 25.R2 (2016), R157–R165.

[4] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, et al. "Strong association of de novo copy number mutations with autism." In: *Science* 316.5823 (2007), pp. 445–449.

[5] Todd J Treangen and Steven L Salzberg. "Repetitive DNA and next-generation sequencing: computational challenges and solutions." In: *Nature Reviews Genetics* 13.1 (2012), pp. 36–46.

[6] David Gordon, John Huddleston, Mark JP Chaisson, Christopher M Hill, Zev N Kronenberg, Katherine M Munson, Maika Malig, Archana Raja, Ian Fiddes, LaDeana W Hillier, et al. "Long-read sequence assembly of the gorilla genome." In: *Science* 352.6281 (2016), aae0344.

[7] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, et al. "Assembly and diploid architecture of an individual human genome via single-molecule technologies." In: *Nature methods* 12.8 (2015), p. 780.

[8] Jeong-Sun Seo, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, Han Cao, Ji-Young Yun, Jihye Kim, et al. "De novo assembly and phasing of a Korean human genome." In: *Nature* 538.7624 (2016), p. 243.

[9] Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. "Resolving the complexity of the human genome using single-molecule sequencing." In: *Nature* 517.7536 (2015), p. 608.

[10] PatriciaA Jacobs, AG Baikie, JA Strong, et al. "The somatic chromosomes in mongolism." In: *The Lancet* 273.7075 (1959), p. 710.

[11] DG Harnden, AH Cameron, VM Crosse, and OH Wolh. "A new trisomic syndrome." In: *The Lancet, London* 1 (1960), pp. 787–790.

[12] Jeffrey M Levsky and Robert H Singer. "Fluorescence in situ hybridization: past, present and future." In: *Journal of cell science* 116.14 (2003), pp. 2833–2838.

[13] Lars Feuk, Andrew R Carson, and Stephen W Scherer. "Structural variation in the human genome." In: *Nature Reviews Genetics* 7.2 (2006), pp. 85–97.

[14] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[15] Sabina Solinas-Toldo, Stefan Lampel, Stephan Stilgenbauer, Jeremy Nickolenko, Axel Benner, Hartmut Döhner, Thomas Cremer, and Peter Lichter. "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances." In: *Genes, chromosomes and cancer* 20.4 (1997), pp. 399–407.

[16] Sara Goodwin, John D McPherson, and W Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies." In: *Nature Reviews Genetics* 17.6 (2016), pp. 333–351.

[17] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[18] Elzo De Wit and Wouter De Laat. "A decade of 3C technologies: insights into nuclear organization." In: *Genes & development* 26.1 (2012), pp. 11–24.

[19] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology." In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[20] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. "Single-cell sequencing-based technologies will revolutionize whole-organism science." In: *Nature Reviews Genetics* 14.9 (2013), pp. 618–630.

[21] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. "Human DNA methylomes at base resolution show widespread epigenomic differences." In: *nature* 462.7271 (2009), pp. 315–322.

[22] Yen-Chun Chen, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. "Effects of GC bias in next-generation-sequencing data on de novo genome assembly." In: *PloS one* 8.4 (2013), e62856.

[23] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. "Real-time DNA sequencing from single polymerase molecules." In: *Science* 323.5910 (2009), pp. 133–138.

[24] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, et al. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome." In: *Nature biotechnology* 37.10 (2019), pp. 1155–1162.

[25] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. "Continuous base identification for single-molecule nanopore DNA sequencing." In: *Nature nanotechnology* 4.4 (2009), pp. 265–270.

[26] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. "Long-read human genome sequencing and its applications." In: *Nature Reviews Genetics* 21.10 (2020), pp. 597–614.

[27] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. "Nanopore sequencing and assembly of a human genome with ultra-long reads." In: *Nature biotechnology* 36.4 (2018), pp. 338–345.

[28] Yao-Ting Huang, Po-Yu Liu, and Pei-Wen Shih. "Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing." In: *Genome biology* 22.1 (2021), pp. 1–17.

[29] Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D Singer, et al. "Technology dictates algorithms: recent developments in read alignment." In: *Genome biology* 22.1 (2021), pp. 1–34.

[30] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. "The human genome browser at UCSC." In: *Genome research* 12.6 (2002), pp. 996–1006.

[31] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. "The complete sequence of a human genome." In: *Science* 376.6588 (2022), pp. 44–53.

[32] Eugene J Gardner, Vincent K Lam, Daniel N Harris, Nelson T Chuang, Emma C Scott, W Stephen Pittard, Ryan E Mills, Scott E Devine, 1000 Genomes Project Consortium, et al. "The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology." In: *Genome research* 27.11 (2017), pp. 1916–1929.

[33] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. "Genotype and SNP calling from next-generation sequencing data." In: *Nature Reviews Genetics* 12.6 (2011), pp. 443–451.

[34] Anna Supernat, Oskar Valdimar Vidarsson, Vidar M Steen, and Tomasz Stokowy. "Comparison of three variant callers for human whole genome sequencing." In: *Scientific reports* 8.1 (2018), pp. 1–6.

[35] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. "DELLY: structural variant discovery by integrated paired-end and split-read analysis." In: *Bioinformatics* 28.18 (2012), pp. i333–i339.

[36] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. "LUMPY: a probabilistic framework for structural variant discovery." In: *Genome biology* 15.6 (2014), pp. 1–19.

[37] Mingfu Zhu, Anna C Need, Yujun Han, Dongliang Ge, Jessica M Maia, Qianqian Zhu, Erin L Heinzen, Elizabeth T Cirulli, Kimberly Pelak, Min He, et al. "Using ERDS to infer copy-number variants in high-coverage genomes." In: *The American Journal of Human Genetics* 91.3 (2012), pp. 408–421.

[38] Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arne Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. "cn. MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate." In: *Nucleic acids research* 40.9 (2012), e69–e69.

[39] Moritz Smolka, Luis F Paulin, Christopher M Grochowski, Medhat Mahmoud, Sairam Behera, Mira Gandhi, Karl Hong, Davut Pehlivan, Sonja W Scholz, Claudia MB Carvalho, et al. "Comprehensive structural variant detection: from mosaic to population-level." In: *BioRxiv* (2022).

[40] David Heller and Martin Vingron. "SVIM: structural variant identification using mapped long reads." In: *Bioinformatics* 35.17 (2019), pp. 2907–2915.

[41] *Pacific Biosciences.* pbsv. `https://github.com/PacificBiosciences/pbsv`.

[42] Bjarni V Halldorsson, Hannes P Eggertsson, Kristjan HS Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O Ulfarsson, Gunnar Palsson, Marteinn T Hardarson, Asmundur Oddsson, Brynjar O Jensson, et al. "The sequences of 150,119 genomes in the UK biobank." In: *Nature* 607.7920 (2022), pp. 732–740.

[43] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. "The mutational constraint spectrum quantified from variation in 141,456 humans." In: *Nature* 581.7809 (2020), pp. 434–443.

[44] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. "Characterizing the major structural variant alleles of the human genome." In: *Cell* 176.3 (2019), pp. 663–675.

[45] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. "Haplotype-resolved diverse human genomes and integrated analysis of structural variation." In: *Science* 372.6537 (2021), eabf7117.

[46] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. "10 years of GWAS discovery: biology, function, and translation." In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[47] Kelly D Farwell, Layla Shahmirzadi, Dima El-Khechen, Zöe Powis, Elizabeth C Chao, Brigette Tippin Davis, Ruth M Baxter, Wenqi Zeng, Cameron Mroske, Melissa C Parra, et al. "Enhanced utility of family-centered diagnostic exome sequencing with inheritance model–based analysis: results from 500 unselected families with undiagnosed genetic conditions." In: *Genetics in Medicine* 17.7 (2015), pp. 578–586.

[48] Yaping Yang, Donna M Muzny, Jeffrey G Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, et al. "Clinical whole-exome sequencing for the diagnosis of mendelian disorders." In: *New England Journal of Medicine* 369.16 (2013), pp. 1502–1511.

[49] Daniel Trujillano, Aida M Bertoli-Avella, Krishna Kumar Kandaswamy, Maximilian ER Weiss, Julia Köster, Anett Marais, Omid Paknia, Rolf Schröder, Jose Maria Garcia-Aznar, Martin Werber, et al. "Clinical exome sequencing: results from 2819 samples reflecting 1000 families." In: *European Journal of Human Genetics* 25.2 (2017), pp. 176–182.

[50] Lisenka ELM Vissers, Kirsten Jm Van Nimwegen, Jolanda H Schieving, Erik-Jan Kamsteeg, Tjitske Kleefstra, Helger G Yntema, Rolph Pfundt, Gert Jan Van Der Wilt, Lotte Krabbenborg, Han G Brunner, et al. "A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology." In: *Genetics in Medicine* 19.9 (2017), pp. 1055–1063.

[51] David R Adams and Christine M Eng. "Next-generation sequencing to diagnose suspected genetic disorders." In: *New England Journal of Medicine* 379.14 (2018), pp. 1353–1362.

[52] Kyle Retterer, Jane Juusola, Megan T Cho, Patrik Vitazka, Francisca Millan, Federica Gibellini, Annette Vertino-Bell, Nizar Smaoui, Julie Neidich, Kristin G Monaghan, et al. "Clinical application of whole-exome sequencing across clinical indications." In: *Genetics in Medicine* 18.7 (2016), pp. 696–704.

[53] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. "The sequence of the human genome." In: *science* 291.5507 (2001), pp. 1304–1351.

[54] Michael N Weedon, Ines Cebola, Ann-Marie Patch, Sarah E Flanagan, Elisa De Franco, Richard Caswell, Santiago A Rodríguez-Seguí, Charles Shaw-Smith, Candy HH Cho, Hana Lango Allen, et al. "Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis." In: *Nature genetics* 46.1 (2014), pp. 61–64.

[55] Laura A Lettice, Taizo Horikoshi, Simon JH Heaney, Marijke J van Baren, Herma C van der Linde, Guido J Breedveld, Marijke Joosse, Nurten Akarsu, Ben A Oostra, Naoto Endo, et al. "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly." In: *Proceedings of the national academy of sciences* 99.11 (2002), pp. 7548–7553.

[56] Eva Klopocki, Silke Lohan, Francesco Brancati, Randi Koll, Anja Brehm, Petra Seemann, Katarina Dathe, Sigmar Stricker, Jochen Hecht, Kristin Bosse, et al. "Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis." In: *The American Journal of Human Genetics* 88.1 (2011), pp. 70–75.

[57] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. "Structural variation in the 3D genome." In: *Nature Reviews Genetics* 19.7 (2018), pp. 453–467.

[58] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. "Ac-

curate whole human genome sequencing using reversible terminator chemistry." In: *nature* 456.7218 (2008), pp. 53–59.

[59] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414 (2012), p. 57.

[60] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." In: *Cell* 159.7 (2014), pp. 1665–1680.

[61] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." In: *Nature* 485.7398 (2012), p. 376.

[62] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. "Chromatin architecture reorganization during stem cell differentiation." In: *Nature* 518.7539 (2015), p. 331.

[63] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. "An atlas of active enhancers across human cell types and tissues." In: *Nature* 507.7493 (2014), p. 455.

[64] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. "VISTA Enhancer Browser—a database of tissue-specific human enhancers." In: *Nucleic acids research* 35.suppl_1 (2006), pp. D88–D92.

[65] Ernest Turro, William J Astle, Karyn Megy, Stefan Gräf, Daniel Greene, Olga Shamardina, Hana Lango Allen, Alba Sanchis-Juan, Mattia Frontini, Chantal Thys, et al. "Whole-genome sequencing of patients with rare diseases in a national health system." In: *Nature* 583.7814 (2020), pp. 96–102.

[66] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. "Structural variant calling: the long and the short of it." In: *Genome biology* 20.1 (2019), pp. 1–14.

[67] Rolf Zeller, Javier López-Ríos, and Aimée Zuniga. "Vertebrate limb bud development: moving towards integrative analysis of organogenesis." In: *Nature Reviews Genetics* 10.12 (2009), pp. 845–858.

[68] Justin Cotney, Jing Leng, Jun Yin, Steven K Reilly, Laura E DeMare, Deena Emera, Albert E Ayoub, Pasko Rakic, and James P Noonan. "The evolution of lineage-specific regulatory activities in the human embryonic limb." In: *Cell* 154.1 (2013), pp. 185–196.

[69] Julia E VanderMeer and Nadav Ahituv. "cis-regulatory mutations are a genetic cause of human limb malformations." In: *Developmental Dynamics* 240.5 (2011), pp. 920–930.

[70] Kimberly L Cooper, Karen E Sears, Aysu Uygur, Jennifer Maier, Karl-Stephan Baczkowski, Margaret Brosnahan, Doug Antczak, Julian A Skidmore, and Clifford J Tabin. "Patterning and post-patterning modes of evolutionary digit loss in mammals." In: *Nature* 511.7507 (2014), pp. 41–45.

[71] Walter L Eckalbar, Stephen A Schlebusch, Mandy K Mason, Zoe Gill, Ash V Parker, Betty M Booker, Sierra Nishizaki, Christiane Muswamba-Nday, Elizabeth Terhune, Kimberly A Nevonen, et al. "Transcriptomic and epigenomic characterization of the developing bat wing." In: *Nature genetics* 48.5 (2016), pp. 528–536.

[72] Francisca Leal and Martin J Cohn. "Loss and re-emergence of legs in snakes by modular evolution of Sonic hedgehog and HOXD enhancers." In: *Current Biology* 26.21 (2016), pp. 2966–2973.

[73] Evgeny Z Kvon, Olga K Kamneva, Uirá S Melo, Iros Barozzi, Marco Osterwalder, Brandon J Mannion, Virginie Tissières, Catherine S Pickle, Ingrid Plajzer-Frick, Elizabeth A Lee, et al. "Progressive loss of function in a limb enhancer during snake evolution." In: *Cell* 167.3 (2016), pp. 633–642.

[74] Florence Petit, Karen E Sears, and Nadav Ahituv. "Limb development: a paradigm of gene regulation." In: *Nature Reviews Genetics* 18.4 (2017), pp. 245–258.

[75] William Curtis Farabee. "Hereditary and Sexual Influences in Meristic Variation: A Study of Digital Malformations in Man." PhD thesis. Harvard University, 1903.

[76] Natalie C Butterfield, Edwina McGlinn, and Carol Wicking. "The molecular regulation of vertebrate limb patterning." In: *Current topics in developmental biology* 90 (2010), pp. 319–341.

[77] Rolf Zeller. "The temporal dynamics of vertebrate limb development, teratogenesis and evolution." In: *Current opinion in genetics & development* 20.4 (2010), pp. 384–390.

[78] Georg C Schwabe and Stefan Mundlos. "Genetics of congenital hand anomalies." In: *Handchirurgie· Mikrochirurgie· Plastische Chirurgie* 36.02/03 (2004), pp. 85–97.

[79] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." In: *Cell* 161.5 (2015), pp. 1012–1025.

[80] Katerina Kraft et al. "Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations." In: *Nature Cell Biology* 21.3 (Mar. 2019), pp. 305–310. ISSN: 1476-4679. URL: https://doi.org/10.1038/s41556-019-0273-x.

[81] Boyan Bonev and Giacomo Cavalli. "Organization and function of the 3D genome." In: *Nature Reviews Genetics* 17.11 (2016), p. 661.

[82] Zhilian Jia, Jingwei Li, Xiao Ge, Yonghu Wu, Ya Guo, and Qiang Wu. "Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection." In: *Genome biology* 21 (2020), pp. 1–24.

[83] Jonas Elsner, Martin A Mensah, Manuel Holtgrewe, Jakob Hertzberg, Stefania Bigoni, Andreas Busche, Marie Coutelier, Deepthi C de Silva, Nursel Elçioglu, Isabel Filges, et al. "Genome sequencing in families with congenital limb malformations." In: *Human Genetics* 140.8 (2021), pp. 1229–1239.

[84] Damian Smedley, Julius OB Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J Buske, Nicole L Washington, et al. "Next-generation diagnostics and disease-gene discovery with the Exomiser." In: *Nature protocols* 10.12 (2015), pp. 2004–2015.

[85] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. "DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources." In: *The American Journal of Human Genetics* 84.4 (2009), pp. 524–533.

[86] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, et al. "A structural variation reference for medical and population genetics." In: *Nature* 581.7809 (2020), pp. 444–451.

[87] Matthew Aguirre, Manuel A Rivas, and James Priest. "Phenome-wide burden of copy-number variation in the UK biobank." In: *The American Journal of Human Genetics* 105.2 (2019), pp. 373–383.

[88] Jeffrey R MacDonald, Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W Scherer. "The Database of Genomic Variants: a curated collection of structural variation in the human genome." In: *Nucleic acids research* 42.D1 (2014), pp. D986–D992.

[89] Jakob Hertzberg, Stefan Mundlos, Martin Vingron, and Giuseppe Gallone. "TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs." In: *Genome biology* 23.1 (2022), pp. 1–21.

[90] Ni Huang, Insuk Lee, Edward M Marcotte, and Matthew E Hurles. "Characterising and predicting haploinsufficiency in the human genome." In: *PLoS genetics* 6.10 (2010), e1001154.

[91] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." In: *Genome research* 15.8 (2005), pp. 1034–1050.

[92] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences." In: *Bioinformatics* 34.18 (2018), pp. 3094–3100.

[93] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, and Michael C Schatz. "Accurate detection of complex structural variations using single-molecule sequencing." In: *Nature methods* 15.6 (2018), pp. 461–468.

[94] Daniel L Cameron, Leon Di Stefano, and Anthony T Papenfuss. "Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software." In: *Nature communications* 10.1 (2019), pp. 1–11.

[95] Alexander V Alekseyenko and Christopher J Lee. "Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases." In: *Bioinformatics* 23.11 (2007), pp. 1386–1393.

[96] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. "Database resources of the national center for biotechnology information." In: *Nucleic acids research* 36.suppl_1 (2007), pp. D13–D21.

[97] Angela S Hinrichs, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. "The UCSC genome browser database: update 2006." In: *Nucleic acids research* 34.suppl_1 (2006), pp. D590–D598.

[98] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[99] Michael Love, Simon Anders, and Wolfgang Huber. "Differential analysis of count data–the DESeq2 package." In: *Genome Biol* 15.550 (2014), pp. 10–1186.

[100] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas SP Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. "Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments." In: *Cell systems* 3.1 (2016), pp. 95–98.

[101] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. "TopDom: an efficient and deterministic method for identifying topological domains in genomes." In: *Nucleic acids research* 44.7 (2015), e70–e70.

[102] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. "Comparison of computational methods for the identification of topologically associating domains." In: *Genome biology* 19.1 (2018), pp. 1–18.

[103] Miao Yu, Nathan R Zemke, Ziyin Chen, Ivan Juric, Rong Hu, Ramya Raviram, Armen Abnousi, Rongxin Fang, Yanxiao Zhang, David U Gorkin, et al. "Integrative analysis of the 3D genome and epigenome in mouse embryonic tissues." In: *bioRxiv* (2022).

[104] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. "The 4D nucleome project." In: *Nature* 549.7671 (2017), pp. 219–226.

[105] W James Kent. "BLAT—the BLAST-like alignment tool." In: *Genome research* 12.4 (2002), pp. 656–664.

[106] Jonathan R Belyeu, Murad Chowdhury, Joseph Brown, Brent S Pedersen, Michael J Cormier, Aaron R Quinlan, and Ryan M Layer. "Samplot: a platform for structural variant visual validation and automated filtering." In: *Genome biology* 22.1 (2021), pp. 1–13.

[107] Weize Xu, Quan Zhong, Da Lin, Ya Zuo, Jinxia Dai, Guoliang Li, and Gang Cao. "CoolBox: a flexible toolkit for visual analysis of genomics data." In: *BMC bioinformatics* 22.1 (2021), pp. 1–9.

[108] Timothy Becker, Wan-Ping Lee, Joseph Leone, Qihui Zhu, Chengsheng Zhang, Silvia Liu, Jack Sargent, Kritika Shanker, Adam Mil-Homens, Eliza Cerveira, et al. "FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods." In: *Genome biology* 19.1 (2018), pp. 1–14.

[109] Caralyn Reisle, Karen L Mungall, Caleb Choo, Daniel Paulino, Dustin W Bleile, Amir Muhammadzadeh, Andrew J Mungall, Richard A Moore, Inna Shlafman, Robin Coope, et al. "MAVIS: merging, annotation, validation, and illustration of structural variants." In: *Bioinformatics* 35.3 (2019), pp. 515–517.

[110] Jouni Sirén, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas A Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, et al. "Pangenomics enables genotyping of known structural variants in 5202 diverse genomes." In: *Science* 374.6574 (2021), abg8871.

[111] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. *FastQC*. Babraham Institute. Babraham, UK, Jan. 2012.

[112] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." In: *arXiv preprint arXiv:1303.3997* (2013).

[113] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications." In: *Bioinformatics* 32.8 (2016), pp. 1220–1222.

[114] Justin M Zook, Nancy F Hansen, Nathan D Olson, Lesley Chapman, James C Mullikin, Chunlin Xiao, Stephen Sherry, Sergey Koren, Adam M Phillippy, Paul C Boutros, et al. "A robust benchmark for detection of germline large deletions and insertions." In: *Nature biotechnology* 38.11 (2020), pp. 1347–1355.

[115] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. "Multi-platform discovery of haplotype-resolved structural variation in human genomes." In: *Nature communications* 10.1 (2019), pp. 1–16.

[116] Azza Althagafi, Lamia Alsubaie, Nagarajan Kathiresan, Katsuhiko Mineta, Taghrid Aloraini, Fuad Al Mutairi, Majid Alfadhel, Takashi Gojobori, Ahmad Alfares, and Robert Hoehndorf. "DeepSVP: integration of genotype and phenotype for structural variant prioritization using deep learning." In: *Bioinformatics* 38.6 (2022), pp. 1677–1684.

[117] Andrew G Sharo, Zhiqiang Hu, Shamil R Sunyaev, and Steven E Brenner. "StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants." In: *The American Journal of Human Genetics* 109.2 (2022), pp. 195–209.

[118] Daniel Danis, Julius OB Jacobsen, Parithi Balachandran, Qihui Zhu, Feyza Yilmaz, Justin Reese, Matthias Haimel, Gholson J Lyon, Ingo Helbig, Christopher J Mungall, et al. "SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing." In: *Genome medicine* 14.1 (2022), pp. 1–13.

[119] Sushant Kumar, Arif Harmanci, Jagath Vytheeswaran, and Mark B Gerstein. "SVFX: a machine learning framework to quantify the pathogenicity of structural variants." In: *Genome biology* 21.1 (2020), pp. 1–21.

[120] Liron Ganel, Haley J Abel, FinMetSeq Consortium, and Ira M Hall. "SVScore: an impact prediction tool for structural variation." In: *Bioinformatics* 33.7 (2017), pp. 1083–1085.

[121] Philip Kleinert and Martin Kircher. "A framework to score the effects of structural variants in health and disease." In: *Genome research* 32.4 (2022), pp. 766–777.

[122] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. "A general framework for estimating the relative pathogenicity of human genetic variants." In: *Nature genetics* 46.3 (2014), p. 310.

[123] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. "The ensembl variant effect predictor." In: *Genome biology* 17.1 (2016), p. 122.

[124] Uirá Souto Melo, Robert Schöpflin, Rocio Acuna-Hidalgo, Martin Atta Mensah, Björn Fischer-Zirnsak, Manuel Holtgrewe, Marius-Konstantin Klever, Seval Türkmen, Verena Heinrich, Ilina Datkhaeva Pluym, et al. "Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases." In: *The American Journal of Human Genetics* 106.6 (2020), pp. 872–884.

[125] Jacob D Spector and Arun P Wiita. "ClinTAD: a tool for copy number variant interpretation in the context of topologically associated domains." In: *Journal of human genetics* 64.5 (2019), p. 437.

[126] Barbara Poszewiecka, Victor Murcia Pienkowski, Karol Nowosad, Jérôme D Robin, Krzysztof Gogolewski, and Anna Gambin. "TADeus2: a web server facilitating the clinical diagnosis by pathogenicity assessment of structural variations disarranging 3D chromatin structure." In: *Nucleic Acids Research* (2022).

[127] Manuel Holtgrewe, Oliver Stolpe, Mikko Nieminen, Stefan Mundlos, Alexej Knaus, Uwe Kornak, Dominik Seelow, Lara Segebrecht, Malte Spielmann, Björn Fischer-Zirnsak, et al. "VarFish:

comprehensive DNA variant analysis for diagnostics and research." In: *Nucleic acids research* 48.W1 (2020), W162–W169.

[128] Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C Brown, A Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R Hughes. "DeepC: predicting 3D genome folding using megabase-scale transfer learning." In: *Nature methods* 17.11 (2020), pp. 1118–1124.

[129] Victor Sanchez-Gaya and Alvaro Rada-Iglesias. "POSTRE: a tool to predict the pathological effects of human structural variants." In: *bioRxiv* (2022).

[130] *Mundlos AG: Development and Disease*. `https://www.molgen.mpg.de/Development-and-Disease`.

[131] Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. "The human phenotype ontology in 2017." In: *Nucleic acids research* 45.D1 (2017), pp. D865–D876.

[132] Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, et al. "New developments on the Encyclopedia of DNA Elements (ENCODE) data portal." In: *Nucleic acids research* 48.D1 (2020), pp. D882–D889.

[133] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. "Integrative genomics viewer." In: *Nature biotechnology* 29.1 (2011), pp. 24–26.

[134] Donna Karolchik, Angie S Hinrichs, and W James Kent. "The UCSC genome browser." In: *Current protocols in bioinformatics* 40.1 (2012), pp. 1–4.

SELBSTÄNDIGKEITSERKLÄRUNG

Name: Jakob Hertzberg

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Berlin, 2023

Jakob Hertzberg