Computational interpretation of disease-causing, structural, and non-coding human genetic variants

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Philip Kleinert

Berlin, 2022

Erstgutachter: Prof. Dr. Martin Vingron, Freie Universität, Berlin

Zweitgutachterin: Prof. Dr. Aida Andres, University College London, London

Tag der Disputation: 07.10.2022

Selbstständigkeitserklärung

Name: Kleinert
Vorname: Philip

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. Ich benutze eigene Texte aus Publikationen mit Erstautorschaft, ohne diese explizit kenntlich zu machen. 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.

Datum: _____ Unterschrift: _____

Declaration of authorship

Name: Kleinert
First name: Philip

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. I use my own text from first authorship publications without specifically marking them. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Date: _____ Signature: _____

Zusammenfassung

Obwohl die erste Version der menschlichen Genomsequenz vor zwei Jahrzehnten fertiggestellt wurde, bleibt das Verständnis vieler genomischer Varianten schwer fassbar. Neuartige Erkenntnisse und technologische Fortschritte verbessern die Fähigkeit, Positionen im Genom zu interpretieren. Klinische Anwendungen hinken jedoch der Grundlagenforschung hinterher, da Wissen und bioinformatische Werkzeuge nur eingeschränkt zugänglich sind, um Therapieansätzen und Patienten zugute zu kommen. In dieser Arbeit helfe ich dabei, humangenetische Varianten zu verstehen, indem ich drei voneinander unabhängige Ansätze weiterentwickle.

Um Forschenden Wissen und Zugang zu Varianteninterpretationen zu bieten, entwickle ich ein Tool zur Aufarbeitung und Analyse der gezielten Sequenzierung genomischer Regionen für das Screening von Patientenkohorten am Beispiel eines etablierten Hämophilie A & B MIP-Designs aus der „My Life, Our Future" Initiative. In einem benutzerfreundlichen HTML-Bericht fasst „hemoMIPs“ abgedeckte, unvollständige oder fehlende Regionen, erfasste Varianten, und ihre vorhergesagten genetischen Auswirkungen zusammen. HemoMIPs ist als Open-Source-Tool auf GitHub veröffentlicht und verfügbar.

In einem zweiten Ansatz betrachte ich genomische Strukturvarianten (SVs) und schätze ihre Auswirkungen auf menschliche Phänotypen mithilfe von maschinellem Lernen ab. Modelle werden mithilfe von Menschen und Schimpansen abgeleiteten SVs trainiert. „CADD-SV“ berechnet zusammenfassende Statistiken über verschiedene Variantenannotationen und verwendet Random-Forest-Modelle, um funktionelle SVs zu priorisieren. Die resultierenden CADD-SV-Scores korrelieren mit bekannten pathogenen, seltenen Varianten und somatischen Krebsvarianten. Dieser Ansatz ist als Online-Service sowie als Open-Source-Tool auf GitHub veröffentlicht und verfügbar.

Vor allem die Interpretation von nicht-kodierenden Sequenzabschnitten hinkt der Interpretation von kodierenden Regionen hinterher. In meinem dritten Ansatz konzentriere ich mich auf nicht-kodierende Varianten in Bindungsstellen eines DNA-bindenden Protein (Transkriptionsfaktor CTCF). Hier entwickle ich einen Arbeitsablauf zur Identifizierung menschenspezifischer gewonnener oder verlorener CTCF-Bindungsstellen unter Verwendung von Datensätzen von Affen und Menschen. Varianten werden anhand ihres Einflusses auf die 3D-Genomarchitektur mithilfe von umfassenden Annotationen priorisiert. Die resultierenden Varianten sind in genomischen Regionen angereichert, die die Gehirnentwicklung beeinflussen. Darüber hinaus zeigt eine unabhängige experimentelle Validierung unter Verwendung von Schimpansen, Orang-Utan und menschlichen Zellkulturen und Organoiden eine hohe Überlappung mit diesem Computer gestützten Ansatz.

Abstract

While the first version of the human genome sequence was completed two decades ago, the understanding of many genomic variants remains elusive. Novel insights and technological advances improve the power to interpret genetic alterations in the genome. However, clinical applications lack behind basic research due to reduced accessibility of knowledge and tools to benefit therapeutic outcomes and patients. In this thesis I help improving the interpretation of human genetic variants and increasing accessibility of these tools by using three independent approaches.

To provide insights and access to variant interpretation to researchers and clinicians, I develop a tool to refurbish and analyze targeted sequencing of genomic regions for screening of patient cohorts on the example of an established hemophilia A & B MIP design from the “My Life, Our Future” initiative. In a user-friendly HTML report “hemoMIPs” summarizes covered, incomplete, or missing regions, called variants and their predicted effects. HemoMIPs is published and available as an open-source tool on GitHub.

In a second approach, I look at genomic structural variants (SVs) and estimate their effect on human health and disease using machine learning. Models are trained on human and chimpanzee derived SVs contrasted with matched simulated variants, an approach that has proven powerful for short sequence variants. “CADD-SV” computes summary statistics over diverse variant annotations and uses random forest models to prioritize functional SVs. The resulting CADD-SV scores correlate with known pathogenic, rare population and somatic cancer variants. This approach is published and available as an online scoring service as well as an open-source software on GitHub.

Especially the interpretation of non-coding variants lacks behind coding regions. In my third approach I focus on non-coding variants in binding sites of a widely studied DNA-binding protein (CTCF). Here, I develop a workflow to identify human-specific gained or lost CTCF binding sites using great ape and human datasets. Variants are prioritized for their impact on 3D genome architecture using a comprehensive set of annotations. Candidates are enriched in genomic regions mediating brain development. Further, independent experimental validation using chimp, orang and human NPCs and organoids show high overlap with this computational approach.

Acknowledgment

These projects would not have been possible without the support of various people. First, many thanks to my supervisor, Martin Kircher, always available and incredibly knowledgeable. Many thanks to Ian and Oskar, for moral support during exceptional times and my occasional struggles along the way. Also, thanks to my colleagues at the BIH, Charité. UCSF and the Max Planck Institute for Molecular Genetics. Especially many thanks to Max, Lusi and Sebastian from the kircherlab and Lana Harshman from UCSF for scientific and non-scientific support. Thanks to Alex Pollen for the CTCF samples. Further thanks to my thesis advisors Martin Vingron as well as Birte Kehr for fruitful discussions and comments on the projects. Further thanks to my family for their emotional and the BIH and Charité for the financial support.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

The vast majority of human traits have a genetic component, besides the environmental influences. From height to hair color, from sex to Alzheimer's disease, genetic factors contribute. Deoxyribonucleic acid (DNA) stores genetic information and, like an instructional manual, describes blueprints of every living organism. The combined genetic information in every cell of an organism is called the genome.

When the sequencing of the human genome was completed in 2001 many scientists regarded this as a revolution for basic human research as well as for human health [1]. However, interpretation of genetic variants, in other words, understanding specific sections of the instructions to build a certain organism, is far more complex than initially speculated [2]. This thesis focuses on interpretation of genetic variants in the human genome, as 20 years after the first draft of the human genome, many mysteries in basic genomic research remain unsolved. To better understand the scientific contributions of this work, I will introduce the properties of DNA and how our cells make use of it. I will explain evolutionary mechanisms that lead to species diversity on our planet as well as diseases. I will focus on bioinformatic principles that are used to grasp the enormous size and complexity of genetic datasets. So, please bear with me if you have not understood all terms used so far.

In this thesis, genetic variant interpretation has been studied from three different viewpoints. Therefore, it is split in three general sections. In the "hemoMIPs" section [3], variants in a well-studied disease mechanism (hemophilia) are analyzed and prepared for clinical inspection using a targeted sequencing approach. In the "CADD-SV" section [4], a specific type of variant, so-called, "structural variants" which can span longer sections of the genome, are prioritized for their functional impact. And finally, "CTCF-Evolution" focuses on a specific mechanism of DNA variants that have putative impact on genome function for a wide set of phenotypes.

1.1 Deoxyribonucleic acid

Deoxyribonucleic acid, more commonly referred to as DNA, is a complex molecule that contains the instructions to build and run all known living organisms as well as many viruses. Every eukaryotic cell contains DNA organized in DNA structures called chromosomes in the nuclear core. DNA is a polymer

composed of a double helix chain built from a backbone containing sugar (deoxyribose) linked to one another by a phosphate group. In the center of the double helix, protected from environmental stress, lies the genetic information, encrypted by four nucleobases - cytosine (C), guanine (G) adenine (A) and thymine (T) - that form the alphabet of the cellular instructions (see Figure 1). Both double stranded DNA strains store the same information as the nucleobases are mirrored, always linking pyrimidines (C,T) with purines (A,G). Therefore, positions in the genomic sequence are often referenced to as base pairs. A change in DNA sequence (mutation) may lead to altered functionality based on a variety of mechanisms [5].

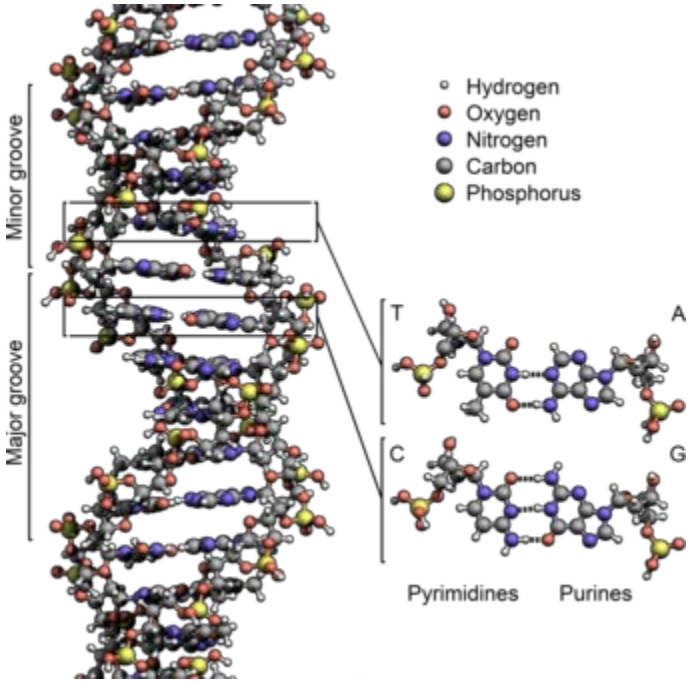


Figure 1: Molecular Structure of DNA with color coded Atoms. DNA is a complex molecule build from Hydrogen, Oxygen, Nitrogen, Carbon and Phosphorus. Heritability derives from information encoded in the sequence of organic bases Thymine (T), Cytosine (C), Adenine (T) and Guanine (G) within the DNA double helix. Due to different number of hydrogen bonds formed, T and A, and C and G match to one another. DNA base pairings are complementary. Information is therefore encrypted twice, on each single strand forming the double stranded DNA double helix as well as in two copies in diploid organisms like humans. Image used from https://en.wikipedia.org/wiki/DNA

1.2 Functional DNA

1.2.1 Coding DNA

Genetic information is stored as a sequence of nucleobases. Some stretches are transcribed by molecular machinery into ribonucleic acid (RNA) that itself can get translated into amino acid chains forming

proteins. Stretches of DNA that are translated into proteins are considered coding-DNA. Three genomic positions form a codon, that is then translated into an amino acid by t-RNAs (see Figure 2). Hence, variation in the sequence of the nucleobases can lead to variation in amino acid sequence and thus the 3D structure and functionality of a protein. Because proteins are responsible for a wide variety of cellular functions such as cellular stability, metabolism and cellular communication, protein variation can lead to distinct biological characteristics (phenotypes) [5]. For instance, genetic variants in proteins controlling the body's ability to form blood clots may lead to increased risk of internal bleeding [6]. This disease, known as hemophilia, as well as the interpretation of the variants involved, are further described in the project section "hemoMIPs".

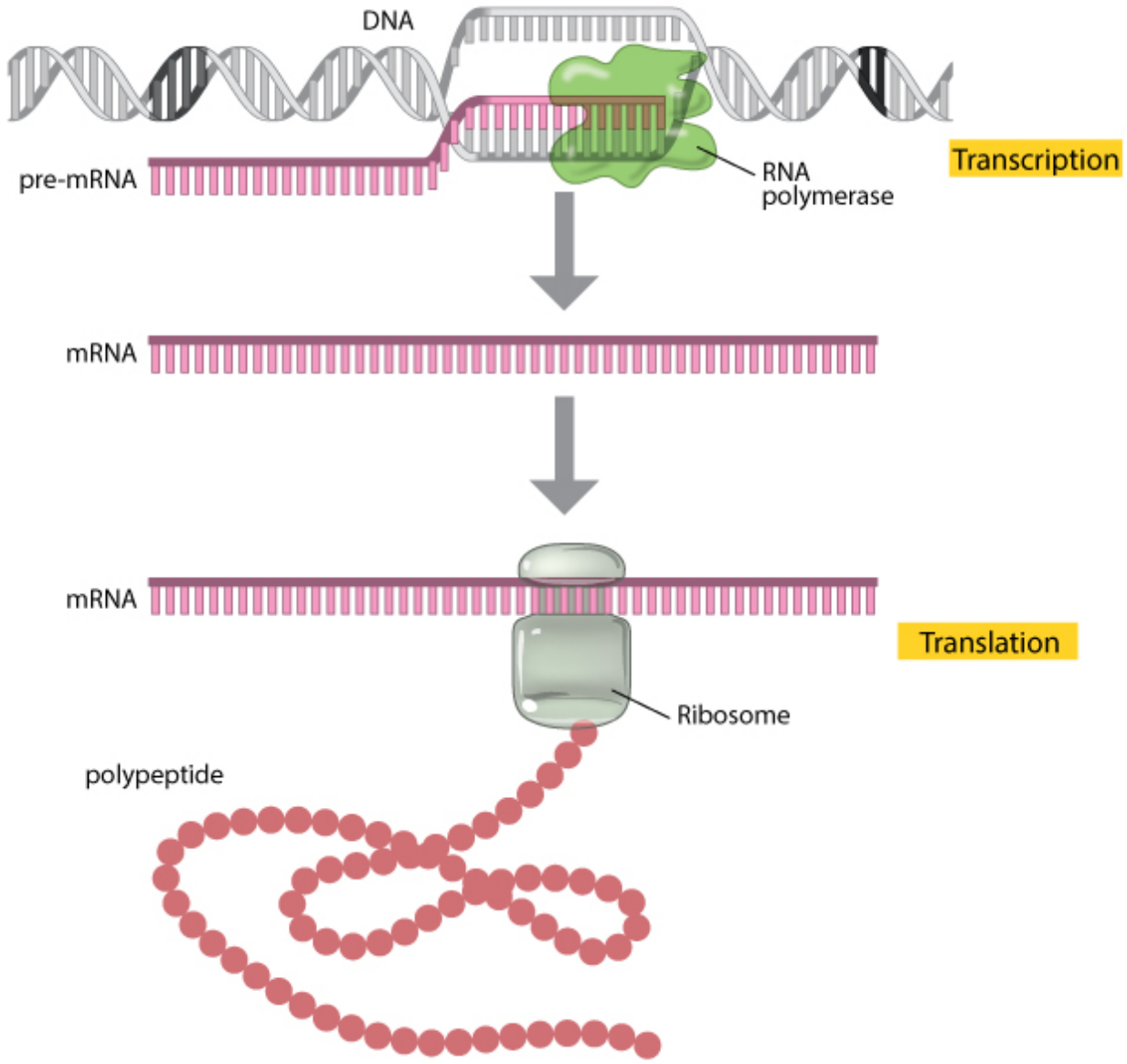


Figure 2: From DNA sequence to protein sequence. DNA is transcribed into messenger RNA (mRNA) that is translated into polypeptide chains (Proteins) by Ribosomes. This process is conducted in every cell in every organism. Proteins are very diverse in their function as they can be involved in metabolic processes, in cellular communication or cellular structure. Picture from 2013 Nature Education (https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/).

1.2.2 Non-coding DNA

Historically DNA was divided into coding DNA that encrypts amino acid chains, and other stretches of DNA originally referred to as junk DNA, long considered meaningless to cellular survival.

However, the function of DNA stretches beyond coding region is far more complex than originally anticipated. Current estimates suggest 8 to 15 % of DNA being "functional" [7]. Non-coding sequences are functional in a wide variety of mechanisms. Cis and trans regulatory elements control the abundance (gene expression) of protein produced from a certain gene [8]. As all cells in the human body contain the same genomic instructions, gene expression is crucial for individual cellular functionality and diversity. Certain genes get activated and deactivated during developmental processes to ensure certain function. Therefore, regulatory elements guarantee the correct expression in time and space of a certain gene. While cis regulatory elements control expression of nearby genes, trans regulatory elements can influence genes distant to the regulatory element. Promoters are sequences typically upstream of the coding region which control gene expression. Enhancers are sequence stretches that, mediated by proteins (e.g., transcription factors), bind to specific sequence motifs within the regulatory element and influence transcription levels of genes. Enhancers can be scattered throughout the genome and do not require proximity to the regulated gene [5].

Further, novel research has shown that the three-dimensional structure of DNA is crucial for its function [9–13]. Therefore, elements that mediate 3D genome architecture hold function. By insulating genetic stretches from surrounding regulatory elements, genomic boundaries can form functional DNA condensates that often are composed of co-expressed genes [14]. The CCCTC-binding factor (CTCF) is a highly conserved protein that mediates 3D genome architecture by binding to specific DNA sequence motifs and therefore forms loops or links DNA stretches to one another. Human variants throughout the genome in CTCF binding sites (BS) are discussed in more detail in the "CTCF evolution" section (see Section 4).

1.3 Variant effects and variant types

As described above, changes in the sequence (genotype) of DNA base pairs impact function and therefore characteristics of an organism (phenotype). Studying this genotype-phenotype relationship is one of the most exciting fields in basic research today. Thousands of scientific papers contribute to the understanding of genetic variants, linking genotype to phenotype [15]. Genetic changes mediate the diversity in life that we see today. Genetic variation can affect individual cells of an organisms as they only

occurred there (somatic variants for instance in cancer tissues) or affect the whole individual and their offspring (germline variants).

1.3.1 Coding sequence

Impact of a genetic variant strongly depends on the sequencing context (for instance falling within coding or non-coding DNA) as well as the type of variant. Single nucleotide variants (SNVs) affect a single base pair in the genome. Especially in coding sequences their effect can be classified depending on the effect they have on protein function. Missense variants lead to the replacement of one amino acid by another. A change in the aminoacid sequence can alter protein structure and function [16]. Change in function does not necessarily correspond to loss of function, as some variants may alter the metabolic properties or location of a protein, rather than rendering it dysfunctional. Hence, altered proteins can also gain functionality. Nonsense variants may lead to complete loss of function. Variants introducing a stop-codon, three base pairs signaling a halt of amino-acid chain formation, lead to shortened proteins that may be non-functional. Short deletions and insertions in coding sequence may lead to frameshift mutations that can offset the translation of the genomic sequence, causing several altered aminoacids and introducing early stop-codons. As protein assembly is based on codons, formed by three base pairs, frameshifts may lead to largely different proteins.

1.3.2 Non-coding sequence

In addition to coding sequence, SNVs can affect non-coding sequence. For instance, Transcription factors (TF), proteins that bind to DNA, make use of a specific sequence motif to find their genomic destination. Certain binding sites (BS) for certain TFs are scattered throughout the genome. Variations in the motif may lead to a change in binding affinity leading to increased or decreased binding of the TF. Changes in TF composition in consequence can lead to changes in gene expression in space and time which may lead to phenotypic differences [17]. Section 4 of this manuscript focuses on non-coding variants in a transcription factor (CTCF) binding motif.

1.3.3 Structural variants

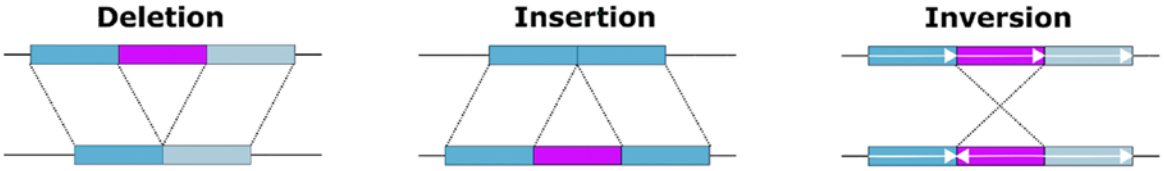


Figure 3: Structural Variants (SVs) in the genome. Deletions, Insertions, and Inversions of 50bp length or more are considered SVs Depiction adapted from Heller et al. [18].

On top of SNVs, variants can span multiple base pairs (bp) in the genome. Structural variants (SVs) are arbitrarily defined as variants affecting more than 50bp. Structural variants can be deletions, insertions, duplications, inversions or other, more complex types of rearrangements (see Figure 3) [19]. SVs in coding sequence can also affect protein function. Rearrangements can for instance lead to fusion proteins composed of previously distinct proteins into one novel protein. Further, large structural variants often affect the 3D genome architecture of a region, as they interfere with regulatory functional setups [20]. Genome architecture describes the three-dimensional shape that the linear DNA molecule is structured in in the cellular core. Boundary regions, regions that form a shift in 3D genome architecture, can be deleted or shuffled around by inversions, leading to altered expression patterns of certain genes [21]. Interpretation of SVs are the focus of section 3 of this manuscript: “CADD-SV”

1.4 DNA sequencing

To identify variants, the genomic sequence of nucleic bases must be determined. Sanger sequencing is an accurate but expensive method to determine DNA sequence that was first introduced in 1977 [22]. Novel methods, termed, “next generation sequencing methods” are faster, with higher throughput and much cheaper. The most widely used technique is based on reversible dye-terminators [23]. In brief, purified DNA is attached on a flow cell, amplified and sequenced by synthesis. Commercial platform, e.g. Illumina Inc. exist. This procedure makes use of fluorophores attached to organic bases. The cluster of amplified DNA strands, fixed on a flow cell is extended base by base. After the incorporation of a base, extension is terminated, a camera detects the type of fluorophore and the corresponding organic base. Afterwards termination is reversed allowing for extension by another base. One by one the sequence is determined. As the human genome consists of 3,100 mega base pairs per haploid genome, shotgun sequencing is applied, sequencing small parts (hundreds of basepairs) of the genome in parallel. Computation software

is needed to assemble the generated small DNA sequence stretches (reads) back into overlapping contigs and finally genomes [24]. A reference genome (for instance human genome build version GrCH38) is used to identify the location of the reads within the human genome (alignment). As sequenced DNA stretches might contain incorrect bases, positions are often covered by multiple reads (coverage). Therefore, identifying variants based on the most prevalent base at a certain position (consensus calling) leads to more accurate results. Overlapping reads allow for the extension of sequence information into longer stretches (contigs). Modern techniques also allow for targeted sequencing of regions of interest. Section 2 of this manuscript "hemoMIPs" focuses on a targeted sequencing approach using Molecular Inversion Probes [3].

1.5 The 3D Genome

DNA in the nucleus is highly organized into chromatin structures in the form of chromosomes. Further, DNA is rolled around histone protein complexes to form nucleosomes, which bind approximately 146bp of DNA. While DNA stretches that are not tightly packed and therefore accessible to transcription factors are called A-compartments, B-compartments consist of highly compressed and therefore inaccessible DNA. On a larger scale, DNA looping brings stretches of DNA into proximity that can be millions of base pairs apart from one another [11]. While looping was first observed in 1878, the function of DNA loops remained obscure for another century [25]. Now, DNA loops are considered crucial to mediate complex expression (the amount of protein being produced in time and space) patterns, as they bring together functional DNA regions such as enhancer and promotor regions or insulate DNA stretches from nearby DNA sequence.

1.5.1 Measuring DNA-Protein interactions

Experimental measures to infer DNA-Protein interaction are used to understand gene regulation as gene expression is mediated by transcription factors (TFs). Chromatin immunoprecipitation (ChIP) combined with next generation sequencing allows for identifying binding sites of specific TFs. With this method, DNA is crosslinked with bound proteins, DNA is sonicated to break the double helix into short segments. Finally specific antibodies are used to filter for the protein of interest. Unlinking, purifying and sequencing the DNA and realigning the reads to the genome leads to aligned sequences peaking at regions where

transcription factors are located [26]. The same concept applies to experiments called CUT&Tag. Among other slight differences, proteins bound to DNA are identified without shearing DNA via sonification. Instead, DNA segments are cut using Tn5 transposase enzymes [27]. Datasets using these experimental procedures were integrated in the projects CADD-SV (Section 3) and "CTCF Evolution" (Section 4).

1.6 Evolution

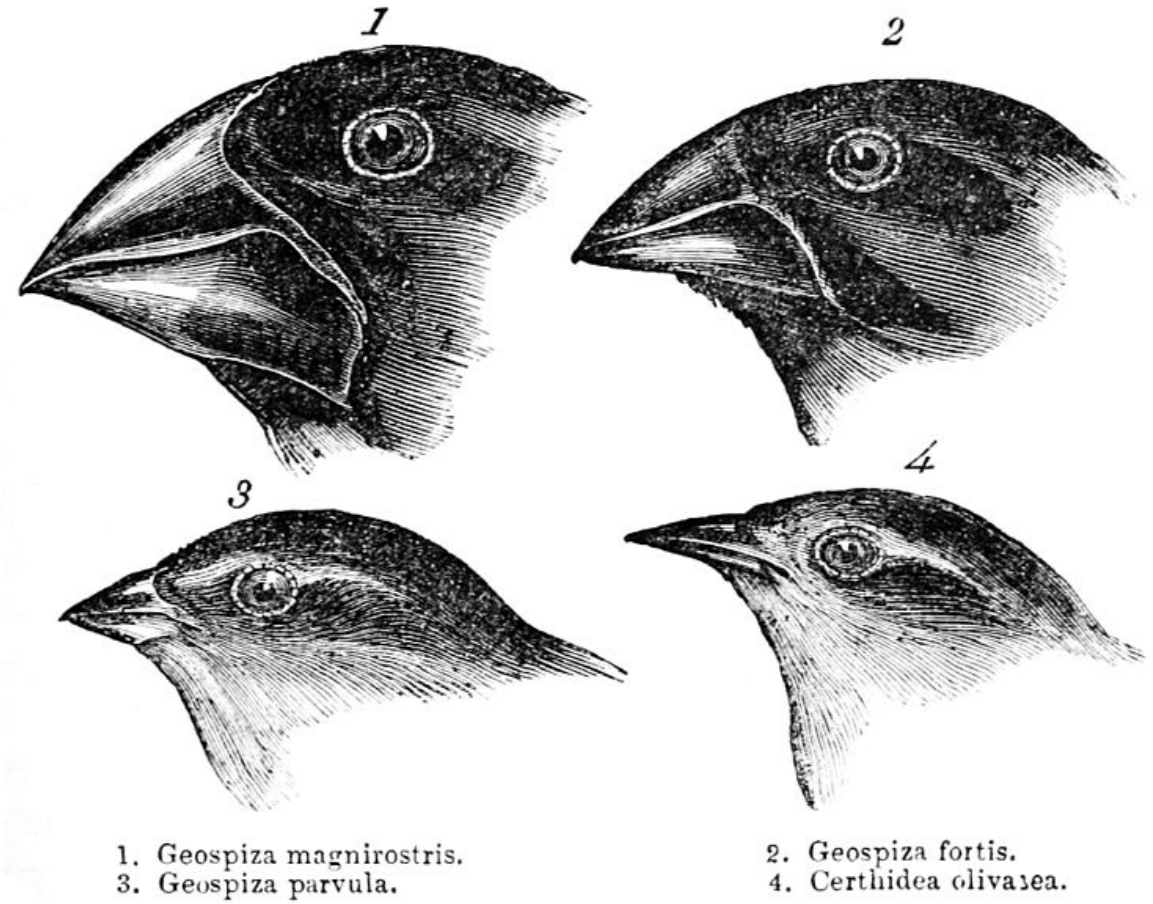


Figure 4: Charles Darwin's drawings of Galapagos finches illustrates the link between genetic variants and function. Due to heritable genetic mutations in their genome, Galapagos finches diverged into subspecies. Their beaks adapted to their environment as specialized beaks are favored (as in have more surviving offspring) that have increased fitness to collect seeds or harvest nuts.

No thesis about variant interpretation should miss mentioning possibly the most influential contribution to science: Charles Darwin's theory of evolution described in "On the Origin of Species" in 1859 [28](see Figure 4). In his seminal work, Darwin describes how speciation happens through constant change and

adaptation that is passed on to future generations. Introducing variants into the gene pool (the set of existing variants in all individuals of a certain species combined) is a natural and desired part of life. The process of naturally occurring variants (mutations) and their heritable effect on fitness is called evolution. All living organisms are subject to evolution. During reproduction, naturally occurring genetic variants get passed on to the offspring. Darwin himself was unaware of the mechanisms of DNA heritability. As described above, genotypic difference may mediate phenotypic difference that may impact fitness. Most variants however that are accumulated are considered neutral with no effect on fitness (yet). Mutations are not directed as they occur mostly random through natural processes, such as through errors during DNA replication[29]. Natural selection, however, may favor or remove certain variants depending on their fitness effect on the carrier[29].

1.6.1 Natural selection

All variants fluctuate in frequency in the population due to genetic drift [30]. Recombination randomly shuffles existing variation from the parents into the offspring, therefore favoring some variants not by their fitness effect but by chance. Positive selection, however, favors variants that influence fitness (for instance reproduction rate) of their carrier. On the other hand, purifying selecting removes variants from the gene pool that are harmful to the fitness of that individual. Purifying selection conserves certain stretches of DNA, as variants decrease the fitness of the carrier and therefore are removed. Some stretches of DNA are so crucial for biological function that they are conserved throughout the evolutionary tree of life. For instance, 16S and 23S ribosomal RNA genes as well as nucleotide binding domains of ABC transporters are so fundamental to cellular survival that they are highly conserved between kingdoms of life as different as Archaea, Bacteria and Eukarya [31].

1.6.2 Primate / Human evolution

As human beings, some research naturally focuses on our own evolution. As part of the phylogenetic tree of primates, *homo sapiens sapiens* first occurred in Africa around 300,000 years ago [32]. Various forms of Homo existed during human evolution, sometimes simultaneously with homo neanderthalensis and Denisovans as the most recent known cousins of modern humans. The split between *homo sapiens* and Neanderthals dates back to about 500,000 years ago [33]. While Neanderthals and Denisovans are extinct today, the closest living relative of humans is the genus pan, consisting of two species: the chimpanzee

and the bonobo. The most recent common ancestor between homo and pan lineages existed about 5 to 8 million years ago [34]. During this time humans accumulated distinct variants of which some are responsible for human success in all habitats on this planet. While most variants are neutral, it remains a contested field of basic research to identify and interpret variants that make humans uniquely human.

1.7 Variant Interpretation

The interpretation of variants as described above is a diverse field of genetic research. Whether and how much impact a certain variant has depends on many factors. Variants are often classified into pathogenic and benign categories for clinical interpretation [35]. Sometimes extending to likely pathogenic, unknown significance and likely benign. Different approaches have been taken to classify novel variants, based on a comprehensive set of annotations describing the variant [36], based on evolutionary conservation [37], based on clinical description of a certain variant [35] in individual or a set of patients or based on experimental readouts [38]. A wide variety of tools have been proposed that make use of one or multiple of these features to describe novel variants still to be classified [39–41]. These approaches are often biased as clinically validated datasets are often limited to a few well-studied genes and computational tools are limited by the ascertainment biases of the training datasets as well as intrinsic biases in the generation of features. Therefore, each approach has advantages and disadvantages.

1.7.1 Challenges

Why is variant interpretation still a growing field of basic research? Interpretation of genetic variants is not trivial. Many challenges remain when trying to understand what consequences a certain genetic variant has on an organism. Genetic variants in the human genome are abundant, even though inter species diversity is low in humans compared to most other species [42]. Identification of the causative variant for a certain disease remains difficult. Many variants are linked to surrounding variants, meaning that they often occur together in various individuals of a population. To link a variant to a phenotype, genome wide association studies (GWAS) are used. Here, the genome of individuals with and without a certain trait are compared using statistical methods to identify variants specific to the group holding the trait [43]. However, these analyses often yield multiple variants in a region that are linked (linkage disequilibrium). Even after associating a region to a known disease, fine mapping is needed to understand the specific variant effect.

As described above, many regulatory variants are poorly understood. Many cellular processes, such as DNA phase separation [44] (describing the process of structures being formed like oil in water) as well as DNA loop extrusion characteristics [45] (Proteins bound to the DNA at two specific locations allow the DNA strand in between to slide along these boundary points) , are novel concepts. Further, many regulatory regions such as enhancers are identified but not linked to their target gene.

Many phenotypes are not caused by a single variant but are mediated by multiple loci of small effect sizes [46]. Only complex combinations might lead to diseases prognosis. Additionally, especially rare disease variants are difficult to identify and interpret as the number of patients to analyze and to understand the genetic background is very limited, sometimes limited to a single patient. In addition, most organisms (humans included) are complex systems consisting of a wide variety of specialized cell types. Variant effects might vary drastically in different cells of the same organism.

1.7.2 Experimental variant interpretation

Novel approaches have been proposed to experimentally validate classified variants. While novel approaches like CRISPR/Cas9 [47], a genetic tool that allows for introduction of variants of interests into a system, revolutionize basic biological research, experimental variant interpretation is still expensive, resource intensive and slow. Massive parallel reporter assays (MPRAs) are set to overcome some of the burdens of experimental validations [48]. Here, gene regulatory activity of individual variants can be measured in a highly parallelized fashion in vivo. Variants are introduced in a reporter construct and linked to a unique DNA barcode. Target gene abundance is measured by sequencing or fluorescence, giving experimental insights into the regulatory activity of a variant. However, this thesis focuses on computational approaches of variant interpretation.

1.8 Computational modeling

1.8.1 Machine learning

More time and cost-efficient variant interpretation is performed in silico. Many computation approaches are based on machine learning (ML) algorithms. ML makes use of existing sample datasets to train a model and learn patterns that enable the model to predict novel inputs. Three approaches exist: supervised

learning feeds the algorithm with example inputs and their desired outputs. The goal is to learn generalized patterns that differentiate inputs into their desired outputs. Unsupervised learning reflects input data without labels. The algorithm's task is to detect structure in the input data that are not provided beforehand. Reinforced learning algorithms strive towards a certain desired output by providing feedback to its previous approaches to do so. Playing a game against an opponent is a classic example of reinforced learning. The CADD-SV section highlights further aspects of machine learning such as hold-out datasets and overfitting.

1.8.2 Software usability

A major focus of this work is usability of the proposed approaches. As they consist of pipelines incorporating existing tools and self-made scripts, managing software dependencies is crucial. To be able to share and further develop existing code, applications like GitHub exist to automate the process of code sharing. Further, many tools require additional libraries for the software to function properly. To accomplish this, package managers like conda have been developed [49]. Conda allows user to install predefined software packages to guarantee stability on various systems. Further, workflow management systems help developers and users alike to automate consecutive processes. Snakemake is a widely used software management tool developed for bioinformatic purposes [50]. Rules define individual calculation steps to be individually run by the user or to be combined in a chain of data processing steps (pipeline).

1.8.3 Summary

Using state of the art workflow and software management tools, I provide software to automate and analyze genomic datasets to prioritize functional variants. I look at known disease mechanisms and provide a tool to analyze and refurbish raw sequencing datasets generated by targeted sequencing (see "hemoMIPs") [3]. I provide a tool to interpret structural variants in the human genome based on a machine learning model that makes use of a wide variety of comprehensive public genomic annotations (see "CADD-SV") [4]. Finally, non-coding variants are prioritized in a functionally focused approach, focusing on 3D genome architecture mediated by a highly conserved transcription factor (see "CTCF-evolution").

2 hemoMIPs – Automated analysis and result reporting pipeline for targeted sequencing data

2.1 Introduction

Patient specific variant detection is of importance for the diagnosis and treatment of various diseases. Identification of disease-causing variants is crucial to understand and suggest personalized therapies for individual carriers. While computational approaches exist to identify variants, an open-source tool to identify variants in predefined stretches of DNA using a specific technology called Molecular Inversion Probes is missing.

Here, I established an automated pipeline to analyze targeted sequencing dataset for variant identification, molecular annotation and quality control. I applied and designed this pipeline specifically for targeted sequencing data from a hemophilia patient cohort.

2.1.1 Molecular Inversion Probes

DNA capture sequencing using Molecular Inversion Probes (MIPs) is a fast and efficient method for targeted sequencing of regions of interest and has been applied in various disease cohorts [51–53]. Targeted sequencing is applied when the region of interest that affects the disease outcome is known. Therefore, only regions of interest are being targeted for sequencing in contrast to whole genome sequencing. A number of protocols exist with minor deviations from the general workflow [54]. The general approach involves designing single stranded DNA probes containing two primer sequences complementary to the region of interest as well as a linker, serving as the backbone to physically link the two primers [55]. These probes are then hybridized to the target DNA and circularized upon polymerase fill-in (elongation of the DNA using a special enzyme called Polymerase that synthesizes chains of nucleic acids) and nick ligation (ligation describing the process of linking the ends of two DNA or RNA strands) (see Figure 5). Degradation of non-circularized molecules enriches the target DNA and sample multiplexing is enabled by using sample specific barcodes in the linker or during an amplification reaction linearizing the DNA and adding required sequencing adapters. Multiplexing using sample specific barcodes (unique identifiers attached to the sequence) is applied to sequence multiple individuals simultaneously and correctly identify variants in each of those.

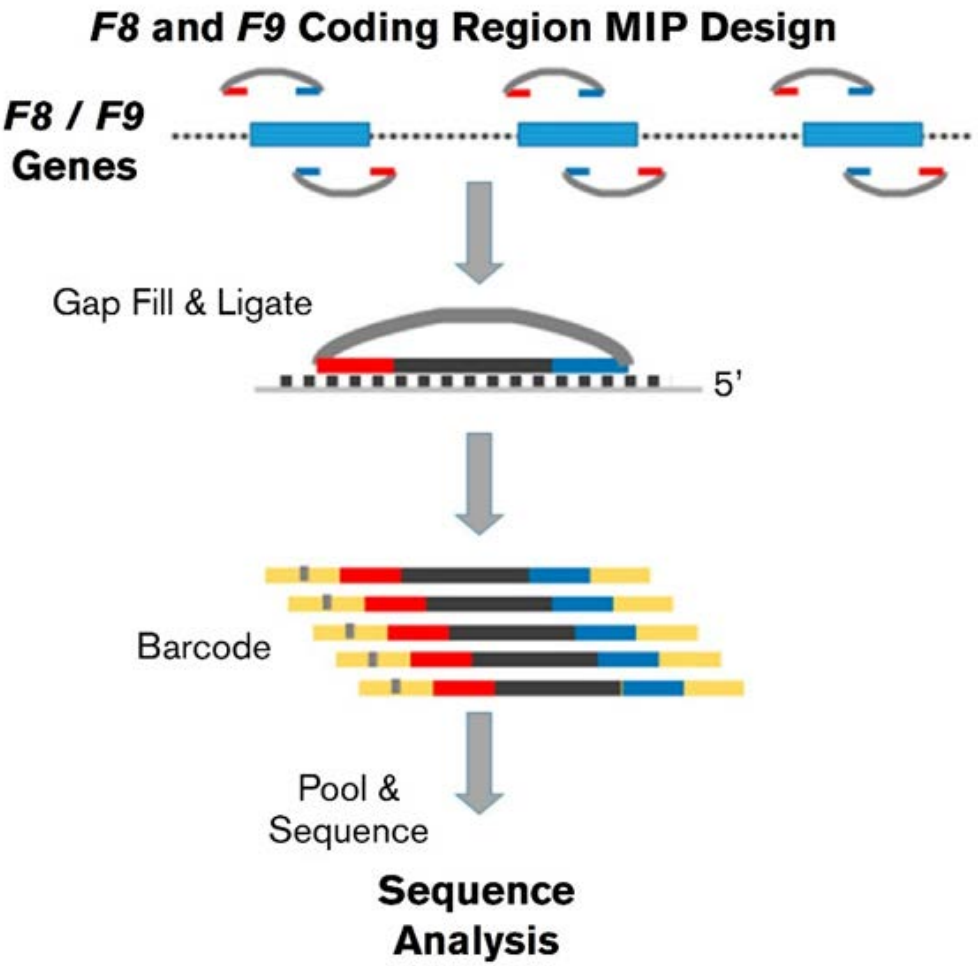


Figure 5: Molecular Inversion Probe workflow: single stranded DNA probes containing two primer sequences (red and blue) complementary to the coding regions of two genes (F8 and F9) as well as a linker (grey), serving as the backbone to physically link the two primers. These probes are then hybridized to the target DNA and circularized upon polymerase fill-in and nick ligation. Degradation of non-circularized molecules enriches the target DNA and sample specific barcodes enable pooled sequencing. Downstream sequencing analysis conducted using the raw sequencing data generated with this approach [56].

2.1.2 Hemophilia

Hemophilia A and B are X-linked (describing the location of the disease on one of the 23 pairs of chromosomes in each human cell, in this case chromosome X) recessive disorders (in contrast to dominant disorders describing the characteristic if one of two "broken" copies is enough to mediate the disease).This results from one or multiple variants out of more than 3,000 known DNA variants in the genes encoding *coagulation factor VIII* (*F8*) and *factor IX* (*F9*), respectively. Determination of the causative genetic variant is important for the patient's reproductive planning, for use in pregnancy and neonatal management, and also to inform risks of neutralizing antibody (inhibitor) formation and bleeding severity.

Therapies targeted to specific patient variants are likely to become more common in the future [56]. The "My Life, Our Future" (MLOF) project is a multisector collaboration developed to provide wide-scale access to free hemophilia genotype analysis for patients in the United States and to create a research repository of associated samples and data to support scientific discovery and treatment advances. For the MLOF initiative, a MIP-based genotyping approach was developed for the *F8* and *F9* genes [56] including more than 450 MIPs.

2.2 Methods

The hemoMIPs pipeline enables hemophilia screening of (typically) 384 patients (derived from 4 96-well-plates) on a single Illumina NextSeq run (a commercial platform that applies next generation sequencing using sequencing by synthesis). Figure 6 outlines the general workflow and the following sections describe data processing and analysis in more detail. All steps are implemented in the workflow management software Snakemake [50] and rely on conda predefined environments to manage software dependencies and easy deployment.

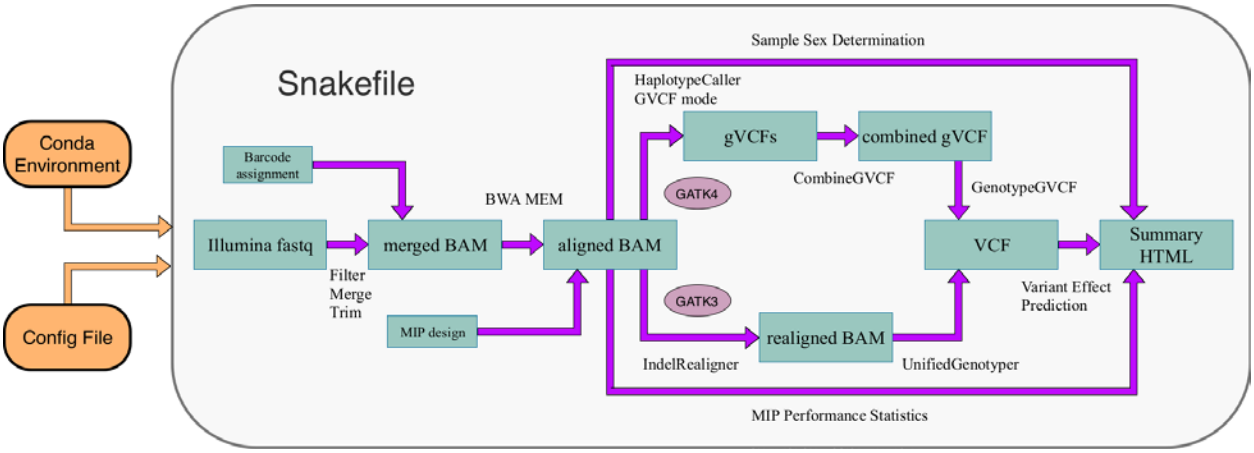


Figure 6: Depiction of the hemoMIPs workflow.The grey outlines describes the steps within the snakemake workflow. Orange boxes name the environment that the workflow is applied in (defined by the dependency software Conda and specified by the Config File). Fastq files generated from Illumina sequencing (containing the raw sequencing read information without location information) is filtered merged and trimmed into BAM files (compressed binary version of the Sequence Alignment/Map format: SAM). Sequenced DNA stretches within the BAM files are aligned to the human genome using an approached called BWA MEM (that uses Burrows Wheeler Alignment to efficiently identify positions in the human genome). From the aligned BAM files either Genome Analysis Toolkit 3 or 4 (GATK) are being used to identify variants compared to the reference genome sequence ultimately leading to variants represented in variant call format (VCF). Here, only variation to the human reference is reported. Sample Sex information as well as MIP Performance Statistics are calculated from the aligned BAM files to subsequently generate the Summary HTML together with the variant information found in the VCF files.

2.2.1 Primary Sequence Processing

The primary inputs are raw FastQ files (containing the identified short DNA sequence (~120 bp) stretches without information about the location within the human genome) from the sequencing run as well as a sample-to-barcode assignment. In primary processing, reads are converted to BAM format (see Figure 6), demultiplexed (storing sample information as read group information), and overlapping paired-end reads (sequencing both ends of a DNA molecule: this process increases the accuracy of the identification of variants as individual positions within the reads might be sequencing errors and therefore incorrect nucleotids) are merged and consensus called [57] (identifying most common nucleotides at individual positions).

2.2.2 Barcode to sample assignment

A two-column tab-separated file is required with the sequencing barcode information. The sample name will be used throughout the processing and reporting. The barcode sequence is assumed to be in the first index read of the Illumina sequencing run (as reads are generated using paired end sequencing) (`I1` FastQ read files created using the Illumina bcl2fastq tool; `bcl2fastq --create-fastq-for-index-reads --use-bases-mask 'Y*,I*,Y*'`). An example for the sample assignment file is provided below:

```
#Seq            Name
CATGCGAGA       Plate_001_01A.1
ACTGGTAGG       Plate_001_01B.2
GCTCCAACG       Plate_001_01C.3
GCGTAAGAT       Plate_001_01D.4
TGACCATCA       Plate_001_01E.5
GGATTCTCG       Plate_001_01F.6
```

2.2.3 MIP design information

Information about the designed MIP probes and their location in the reference genome is needed as a tab-separated text file for the tool *TrimMIParms.py*. The default input file has the following columns: index, score, chr, ext_probe_start, ext_probe_stop, ext_probe_copy, ext_probe_sequence, lig_probe_start, lig_probe_stop, lig_probe_copy, lig_probe_sequence, mip_scan_start_position, mip_scan_stop_position, scan_target_sequence, mip_sequence, feature_start_position, feature_stop_position, probe_strand, failure_flags, gene_name, mip_name. This format is obtained from

MIP designs generated by MIPGEN [55], a tool for MIP probe design available on GitHub (https://github.com/shendurelab/MIPGEN). Alternatively, files containing at least the following named columns can be used: chr, ext_probe_start, ext_probe_stop, lig_probe_start, lig_probe_stop, probe_strand, and mip_name. It is critical, that the reported coordinates and chromosome names match the reference genome used in alignment.

2.2.4 Snakemake configuration

Different aspects of the project (e.g. sequencing run information, reference sequences, local paths, benign variants) can be defined using a central configuration file for snakemake. An example is available at https://github.com/kircherlab/hemoMIPs/blob/master/example_config.yml. Different references and annotations need to be specified in the snakemake config file such as reference genome, Burrows-Wheeler Alignment (BWA) indexed reference genome location, inversion reference, and VEP installation path and cache version. As multiple Illumina lanes can be analyzed simultaneously, the respective run folder names in the local "input" folder and the number of lanes can be set in the "datasets" section of the configuration file.

2.2.5 Known and benign variant information

A list of known and benign variants can be provided in the configuration file, these variants will be shown in gray in the HTML output reports (see Figure 7).

Sample Summary Report

Individual	Sex	Short variants	Incomplete coverage	Deletions (<50% covered)	INT1	INT22	Status
Sample_1	M	F8 (E:26/26) c.*1292A>G [GRCh37 X:154064580 T/C] F8 (I:3/25) c.389-9C>T [GRCh37 X:154221432 G/A]			noINT1	noINT22	OK
Sample_2	F	F8 (E:26/26) c.*1292A>G [GRCh37 X:154064580 T/C] **F8 (E:14/26) c.4264T>C / p.Tyr1422His** [GRCh37 X:154157801 A/G]			noINT1	**INT22-FAILED**	FAILED inversions
Sample_3	F	F9 (E:8/8) c.*687G>A [GRCh37 X:138644917 G/A] F8 (E:26/26) c.*1672G>A [GRCh37 X:154064200 C/T] F8 (E:26/26) c.*1672G>A [GRCh37 X:154064200 C/T] F8 (E:26/26) c.*1292A>G [GRCh37 X:154064580 T/C] **F8 (E:8/26) c.1172G>A / p.Arg391His** [GRCh37 X:154194800 C/T]	X:138645091-138645126 (F9/3-UTR)		noINT1	noINT22	OK
Sample_4	F	F8 (E:26/26) c.*1672G>A [GRCh37 X:154064200 C/T] F8 (E:26/26) c.*1672G>A [GRCh37 X:154064200 C/T] F8 (E:26/26) c.*1292A>G [GRCh37 X:154064580 T/C]			noINT1	noINT22	OK
Sample_5	F	**F9 (E:2/8) c.199G>A / p.Glu67Lys** [GRCh37 X:138619279 G/A] F9 (E:6/8) c.580A>G / p.Thr194Ala [GRCh37 X:138633280 A/G]			noINT1	**INT22-1**	CHECK variants
Sample_6	M	F8 (E:26/26) c.*1672G>A [GRCh37 X:154064200 C/T] F8 (E:26/26) c.*1292A>G [GRCh37 X:154064580 T/C]			noINT1	noINT22	OK

Figure 7: HTML reports are generated for visualization, interpretation and better access to all information collected across the individual workflow steps. Here, a section of report.html shows the obtained genotypes for the demultiplexed samples and highlights potential pathogenic variants, their location in the gene and which exon (E:) (protein coding sequence stretches) or intron (I:) (sequence in between exons) is affected. Additionally incomplete called sites, predefined structural variants (columns INT1 and INT22 referring to inversions of F8-intron 22 and F8-intron 1 which are common causes of severe hemophilia A) and

failed MIPs are reported. The multiplexed samples can be identified via their sampleID. This output is meant to give a general overview over the sample performances.

2.2.6 Alignment and MIP arm trimming

Processed reads are aligned to the reference genome (here GRCh37 build from the 1000 Genomes Project Phase II release) using Burrows-Wheeler Alignment (BWA) 0.7.5 mem [24]. As MIP arm sequence can result in incorrect variant identification (by hiding existing variation below primer sequence), MIP arm sequences are trimmed based on alignment coordinates and new BAM files are created. In this step, MIP design files from MIPgen [55] are used by default. MIP representation statistics (text output file) are calculated from the aligned files. Further, reads aligning to the Y-chromosome-unique probes (*SRY* gene; corresponding sequences should only be obtained from a male individual) are counted for each sample and reported (text output file).

In a separate alignment step, all reads are aligned to a reference sequence file describing only the structural sequence variants as mutant and reference sequences. Results are summarized over all samples with the number of reads aligning to each sequence contig in a text report.

2.2.7 Coverage Analysis and Calling using GATK

Coverage (amount of reads encountered at certain positions) differences between MIPs are handled by down sampling regions of excessive coverage. Variants are genotyped using GATK [58] UnifiedGenotyper (v3.4-46) in combination with IndelRealigner (v3.2-2). Alternatively, GATK v4.0.4.0 HaplotypeCaller is used in gVCF mode in combination with CombineGVCFs and GenotypeGVCFs. Variant annotations of the called (identified) variants, including variant effect predictions and Human Genome Variation Society (HGVS) variant descriptions are obtained from Ensembl Variant Effect Predictor, a tool to annotate and prioritize genetic variants [59].

2.2.8 Reporting

Different HTML reports are generated for visualization, interpretation and better access to all information collected in previous steps. There are two entry points to this information, organized as two different HTML reports – one summarizing all variant calls and MIP performance across samples and the other summarizing per-sample results in an overview table.

The first report (report.html) provides an overview of results for each sample highlighting putative deleterious variants and taking previously defined common/known benign variants out of focus (see Figure 8, gray font). Additional information is provided about potential structural variants and incompletely covered regions. This table also provides an overall sample status field with information about passing and failing samples, as well as flags indicating outlier MIP performances.

The second report (summary.html, See Figure 9) provides a more technical sample and variant summary, per region coverage and MIP performance statistics. This report across samples can be used to assess assay performance (e.g. underperforming MIPs could be redesigned in future assays) and allows identification of suspiciously frequent variants (common variants or systematic errors).

Both reports provide links to individual report pages of each sample. The individual reports (ind_SAMPLENAME.html), provide quality measures like overall coverage, target region coverage, read counts underlying the inferred sample sex and MIP performance statistics (over- or underperforming MIPs in this sample), but most importantly provide detailed information on the identified variants, structural variant call results and regions without coverage (potential deletions).

2.2.9 Reported Tables

In additional to the HTML output files for visualization, results are also presented in computer readable Comma Separated Values (CSV) format files. These CSV files can be joined by either the variant or sample specific identifier columns. The following results are summarized in the respective table files:

ind_status.csv outputs the sample sex inferred from SRY counts, reports outlier MIP performance, number of genotype (GT) calls (genotype being the DNA of an individual at a certain position) , covered sites within the MIP design regions, average coverage, heterozygous sites (describing positions that differ between chromosome pairs in one individual), incompletely covered regions, deletions as well as a textual summary in a sample quality flag (e.g. OK, Failed Inversions, Check MIPs).

variant_calls.csv and *variant_calls_benign.csv* contain all or just benign variants, respectively, with location, genotype, quality scores, allelic depth, coverage and status information.

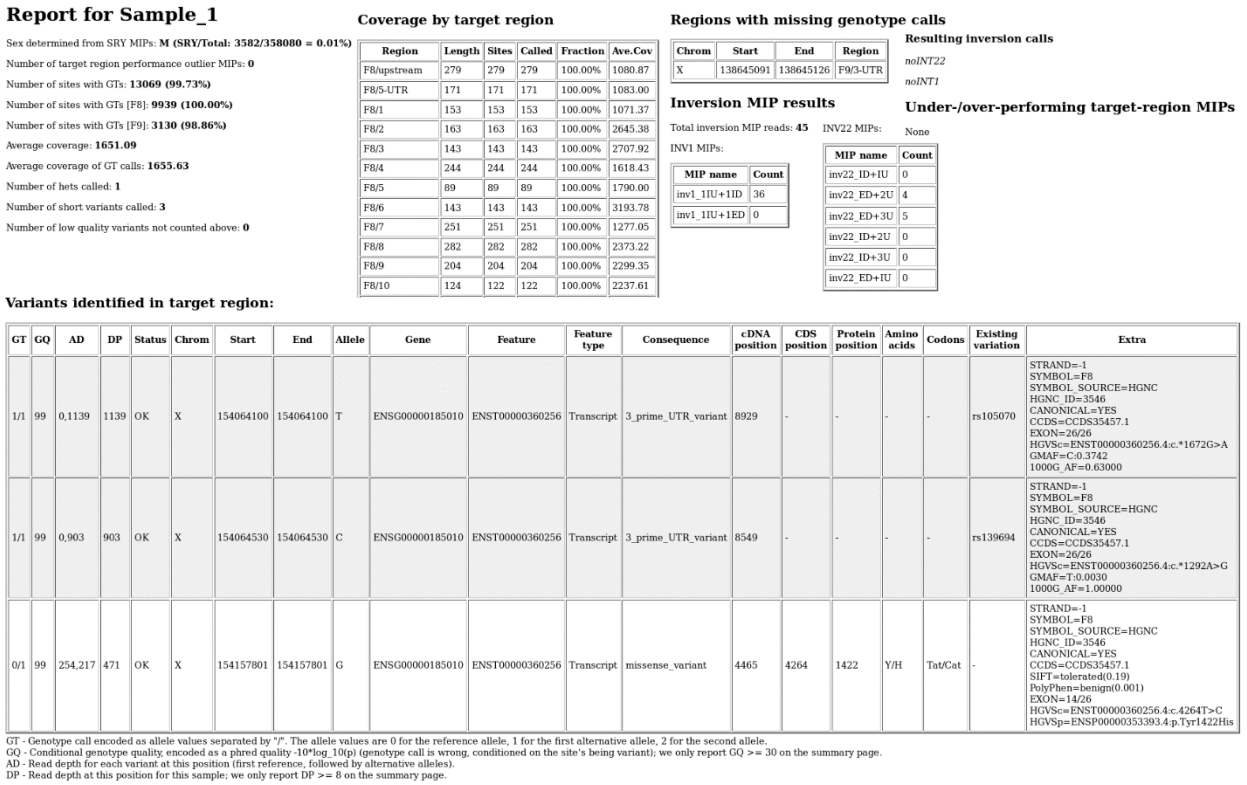
variant_annotation.csv provides additional annotations to called variants based on reference and alternative allele information. These annotations include gene name, exonic location, cDNA and coding DNA sequence (CDS) position, HGVS Transcript and Protein information, variant rsID (unique labels to

identify variants), and 1000G allele frequency (frequency of variants in a worldwide human genetics cohort: 1000 Genomes Project [60]).

inversion_calls.csv contains count results for MIPs targeting predefined structural variants.

2.3 Results

Here I introduce an easy-to-use pipeline to analyze highly imbalanced, targeted, next-generation sequencing data sets generated using MIP experiments. In a user-friendly HTML report, all analysis results including covered, incomplete or missing regions, called variants and their predicted effects are summarized (see Figure 8).



Report for Sample_1

Sex determined from SRY MIPs: **M (SRY/Total: 3582/358080 = 0.01%)**
Number of target region performance outlier MIPs: **0**
Number of sites with GTs: **13069 (99.73%)**
Number of sites with GTs [F8]: **9939 (100.00%)**
Number of sites with GTs [F9]: **3130 (98.86%)**
Average coverage: **1651.09**
Average coverage of GT calls: **1655.63**
Number of hets called: **1**
Number of short variants called: **3**
Number of low quality variants not counted above: **0**

Coverage by target region

Region	Length	Sites	Called	Fraction	Ave.Cov
F8/upstream	279	279	279	100.00%	1080.87
F8/5-UTR	171	171	171	100.00%	1083.00
F8/1	153	153	153	100.00%	1071.37
F8/2	163	163	163	100.00%	2645.38
F8/3	143	143	143	100.00%	2707.92
F8/4	244	244	244	100.00%	1618.43
F8/5	89	89	89	100.00%	1790.00
F8/6	143	143	143	100.00%	3193.78
F8/7	251	251	251	100.00%	1277.05
F8/8	282	282	282	100.00%	2373.22
F8/9	204	204	204	100.00%	2299.35
F8/10	124	122	122	100.00%	2237.61

Regions with missing genotype calls

Chrom	Start	End	Region
X	138645091	138645126	F9/3-UTR

Resulting inversion calls

noINT22

noINT1

Inversion MIP results

Total inversion MIP reads: **45**

INV1 MIPs:

MIP name	Count
inv1_1IU+1ID	36
inv1_1IU+1ED	0

INV22 MIPs:

MIP name	Count
inv22_ID+IU	0
inv22_ED+2U	4
inv22_ED+3U	5
inv22_ID+2U	0
inv22_ID+3U	0
inv22_ED+IU	0

Under-/over-performing target-region MIPs

None

Variants identified in target region:

GT	GQ	AD	DP	Status	Chrom	Start	End	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	Extra
1/1	99	0,1139	1139	OK	X	154064100	154064100	T	ENSG00000185010	ENST00000360256	Transcript	3_prime_UTR_variant	8929	-	-	-	-	rs105070	STRAND=-1 SYMBOL=F8 SYMBOL_SOURCE=HGNC HGNC_ID=3546 CANONICAL=YES CCDS=CCDS35457.1 EXON=26/26 HGVSc=ENST00000360256.4:c.*1672G>A GMAF=C:0.3742 1000G_AF=0.63000
1/1	99	0,903	903	OK	X	154064530	154064530	C	ENSG00000185010	ENST00000360256	Transcript	3_prime_UTR_variant	8549	-	-	-	-	rs139694	STRAND=-1 SYMBOL=F8 SYMBOL_SOURCE=HGNC HGNC_ID=3546 CANONICAL=YES CCDS=CCDS35457.1 EXON=26/26 HGVSc=ENST00000360256.4:c.*1292A>G GMAF=T:0.0030 1000G_AF=1.00000
0/1	99	254,217	471	OK	X	154157801	154157801	G	ENSG00000185010	ENST00000360256	Transcript	missense_variant	4465	4264	1422	Y/H	Tat/Cat	-	STRAND=-1 SYMBOL=F8 SYMBOL_SOURCE=HGNC HGNC_ID=3546 CANONICAL=YES CCDS=CCDS35457.1 SIFT=tolerated(0.19) PolyPhen=benign(0.001) EXON=14/26 HGVSc=ENST00000360256.4:c.4264T>C HGVSp=ENSP00000353393.4:p.Tyr1422His

GT - Genotype call encoded as allele values separated by "/". The allele values are 0 for the reference allele, 1 for the first alternative allele, 2 for the second allele.
GQ - Conditional genotype quality, encoded as a phred quality -10*log_10(p) (genotype call is wrong, conditioned on the site's being variant); we only report GQ >= 30 on the summary page.
AD - Read depth for each variant at this position (first reference, followed by alternative alleles).
DP - Read depth at this position for this sample; we only report DP >= 8 on the summary page.

Figure 8: An individual report (ind_Sample_1.html) shows general quality metrics as well as functional annotations of identified variants, the coverage for each targeted region (including regions missing coverage/genotype calls), the counts for MIPs designed to capture structural variants and highlights over- or underperforming MIPs.

2.3.1 GATK3 output

Using the GATK3 version (being used for variant calling and coverage analysis) of hemoMIPs, the MLOF initiative screened 3,000 patients for hemophilia causative variants in 2017, sequencing the *F8* and *F9* genes for about 15% of the total hemophilia A and B population of the United States [56]. All *F8* and *F9* coding regions, splice sites, and upstream (450 bp for *F8* and 300 bp for *F9*) and downstream (1838 bp for

F8 and 1417 bp for *F9*) untranslated sequences were captured using 458 MIP probes, each with about 111 bp in target size. Additional eight probes were designed to capture reference or mutant sequences of large DNA inversions mediated through sequences in *F8* intron 22 or *F8* intron 1 and homologous sequences distal to the *F8* gene, resulting in gene disruptions [61]. Finally, five probes are targeting *SRY* unique sequence to detect patient sex.

Summary Report

Total number of samples: **10**

Total number of sites considered [cov in >50% samples]: **13105**

Sample summary

SampleID	Sex	GTs	%GTs	Ave.Cov	Hets	Variants	VariantList (incl. low quality)
Sample_1	M	13069	99.73%	1953.9188	0	2	X:154063210 C/T, X:154012480 T/C
Sample_2	F	13105	100.00%	3141.4749	1	5	X:138638080 A/G, X:138632217 G/A, X:154064100 C/T, X:154064580 T/C, X:154197744 C/T
Sample_3	F	13069	99.73%	2935.9815	2	5	X:138633980 A/G, X:154064580 T/C, X:154088998 T/C, X:154090034 G/A, X:154158285 G/C
Sample_4	F	13032	99.44%	1799.0597	0	4	X:138633230 A/G, X:138644917 G/A, X:154064201 C/T, X:154064580 T/C
Sample_5	F	13105	100.00%	2218.9992	1	3	X:138645143 G/C, X:154064300 C/T, X:154064530 T/C
Sample_6	M	13105	100.00%	1848.3196	2	3	X:154064200 C/T, X:154064588 T/C, X:154158282 G/C
Sample_7	M	13069	99.73%	3473.7384	0	3	X:154064200 C/T, X:154064580 T/C, X:154157487 -/T
Sample_8	F	13069	99.73%	3046.9438	5	6	X:138633280 A/G, X:138644917 G/A, X:154064200 C/T, X:154064580 T/C, X:154158285 G/C, X:154189406 A/C
Sample_9	F	13105	100.00%	2186.9136	2	4	X:138644917 G/A, X:154064200 C/T, X:154064580 T/C, X:154130395 G/A
Sample_10	M	13105	100.00%	1150.0903	0	4	X:138633287 A/G, X:154064200 C/T, X:154064580 T/C, X:154088781 A/G

Figure 9: An example of the Summary Report (summary.html). This report provides the user an overview of all samples present in the dataset with their inferred sex, genotypes (GT), average coverage (Ave.Cov), number of heterozygous (Hets) and overall variants and the observed variant list with direct links to the individual sample reports.

2.3.2 Causative variants

In 98.4% (2,952/3,000) of patients, the likely causative variant was identified from our results and confirmed using Sanger validation [56]. Of 924 unique variants observed in this hemophilia cohort, 285 novel variants were identified. In cases of severe hemophilia, predicted gene-disrupting variants were common while missense variants dominated for mild-to-moderate disease. Novel hemophilia DNA variants were detected continuously throughout the project, indicating that additional variation likely remains undiscovered [56].

2.3.3 GATK4

I have extended the pipeline to use GATK4 for variant calling and coverage analysis. Results are highly concordant between the two versions, but GATK4 calling is 50 times faster (see also Supplementary Information).

2.3.4 GATK3, GATK4 comparison

The hemophilia datasets perform similar when run either with the GATK3 or GATK4 workflow. However, in low quality genotype calls the performance might vary and a different call set might be obtained. In a reanalysis performed on one of the hemophilia sequencing experiments, the sample specific genotype agreement is above 0.99 (36 different out of 64,308 genotype calls) between the two GATK versions, with high agreement in associated genotype qualities (reflecting the probability of a certain position in the genome being reported correctly) (Figure 10). Therefore, GATK4 was chosen as the standard setting for the workflow as this version maintains support, is 50x faster and can be more easily upgraded as it is still being developed.

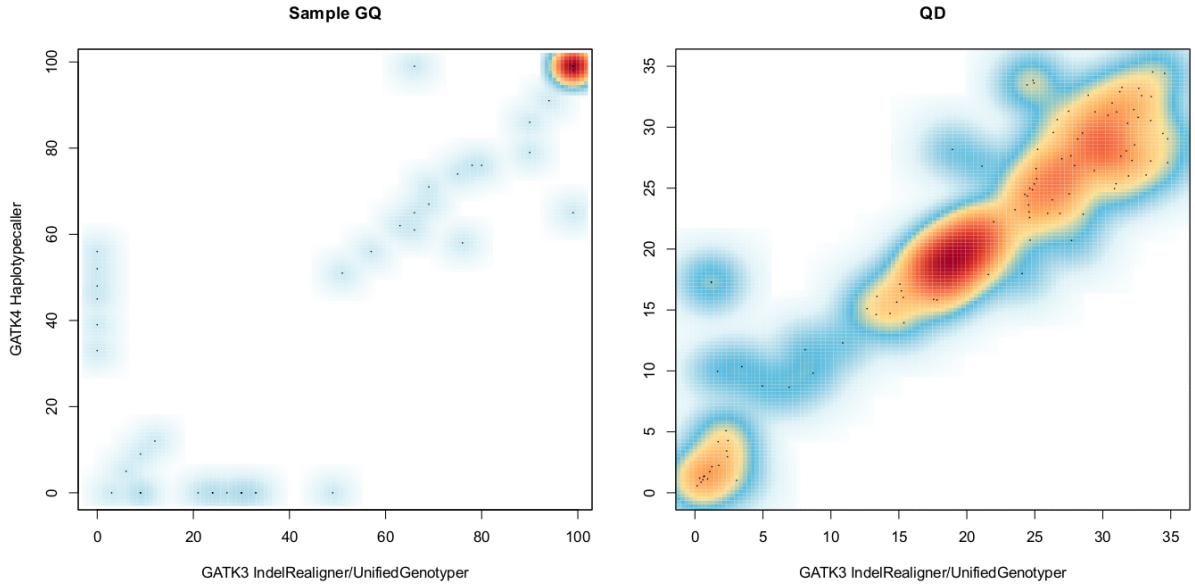


*Figure 10: Comparison of GATK3 and GATK4 results. Heatmap of GATK3 vs GATK4 Genotype Quality (GQ) scores (left) and GATK3 vs. GATK4 QD (Quality by Depth) scores (right). Both scores are on Phred-like scale, expressing the -10*log$_{10}$ likelihood of an incorrect call. While most variants are called with both GATK versions with high confidence (left panel, top right corner), a few variants are missed by either tool. The sample-specific genotype agreement is above 0.99 (36 different out of 64,308 genotype calls). A shifted InDel explains 6 out of 36 different genotypes. Eleven out of the remaining 30 discordant calls are seen below a total read coverage of 3 for one of the callers. Further, among the remaining discordant calls (18 out of 19 being called by GATK3), 14 are low quality calls (GQ < 30).*

2.3.5 Availability and Implementation

HemoMIPs is available on GitHub on https://github.com/kircherlab/hemoMIPs. Its source code is open and available for everyone to download and modify (MIT License). A manual can be found in the main

repository together with example inputs and outputs to run the pipeline. All dependencies are handled by predefined conda environments available in the main repository.

2.4 Discussion

The importance of targeted sequencing approaches to human health is predicted to increase drastically in the coming years, as more and more causative variants and genes are being identified for various diseases [62] [63]. Personalized approaches consider mechanistic differences in disease causation as well as progression. On the example of MIP generated targeted sequencing of hemophilia relevant genes Factor 8 and Factor 9, I show that targeted sequencing can be cost effective, and analysis completely automatized.

2.4.1 Hemophilia as a well-studied example

Hemophilia, as an early candidate example for personalized medicine shows the potential of this approach. Targeted sequencing identifies the diversity of disease-causing variants as well as might be indicative for individual therapies. In this example variant interpretation is directly linked to patient health, as medication for Factor 8 or Factor 9 deficient hemophilia patients differs [64]. However, this approach can be extended to other genomic regions. As more and more genomic variants are being understood and their therapeutic potential is unleashed, targeted sequencing using MIPs might become a common procedure in clinical laboratories.

2.4.2 User friendliness

Further, standardization and automatization are important to provide tools of clinical relevance. I opted for an updated version of a widely used variant calling algorithm, GATK4, to secure future support for this pipeline. Additionally, GitHub, Conda and Snakemake underline the reproducibility and user friendliness of the pipeline as software is managed and open access as well as easy installation guidelines are guaranteed.

The output is generated in a user-friendly manner as well, as CSV format as well as an HTML output readable by all major web browser is provided. Therefore, clinicians can share, compare, and visualize results from individual patients as well as bigger cohorts.

As an open-source and community effort, the hemoMIPs pipeline will continue to evolve with changes in best practice workflows (e.g. provided through GATK) as well as potential novel molecular inversion probe designs and the application to other diseases and genes.

2.4.3 Conclusion

hemoMIPs is an easy and efficient pipeline to analyze MIP target capture data generated on the Illumina sequencing platform. Using an easily adapted Snakemake workflow [50], hemoMIPs performs sample demultiplexing, overlap paired-end merging, alignment using BWA, MIP-arm trimming, variant calling using GATK, coverage analysis and HTML report generation for single end and paired end sequencing datasets. While hemoMIPs was developed to analyze targeted sequencing data of the MLOF Initiative, it can be applied to a broad set of MIP sequencing data sets. Currently various tools and individual pipelines are being used in the genotyping of Molecular Inversion Probe Data. While two pipelines [52,65] are publicly inaccessible, MIPgen tools [55] and bwa-MIPs [66], MIPWrangler [67] stops after alignment and arm trimming. Therefore, hemoMIPs is the first complete analysis workflow that is open source and easy to employ via workflow management.

3 CADD-SV

3.1 Introduction

In addition to Single Nucleotide Variants (SNVs), covering (as the name suggests) just a single position in the genome, structural variants (SVs) also exist. This section of this thesis focuses on the interpretation of SVs. These kinds of variants cover multiple base pairs (oftentimes arbitrarily defined as being at least 50bp in size) and are often classified as deletions, insertions, duplications, inversions and other more complex rearrangements [19]. These various types of SVs therefore delete, insert, duplicate or invert stretches of the genome.

3.1.1 Human SVs

According to a set of SVs discovered in a healthy patient cohort, about 7,439 SVs can be found per individuum spanning a median size of 331 base pairs [19]. Just as for SNVs, genomic diversity is greatest in African populations, decreasing with the population bottlenecks during the Out-of-Africa migration of Europeans and Asians and is lowest in South American ethnic groups. Different types of SVs occur at different frequency in the human genome with deletions being the most prominent, or better described as the most often detected type of SV. These numbers are influenced by the power to detect SVs, which is non-trivial process and is different for each type of SV [68]. SV detection biases are therefore discussed elsewhere[69].

The sum of impacted base pairs by SVs (about 18 Mbp)[70] is greater than for SNVs (about 3.78 Mbp)[71] which is surprising, as using structural variants in the genome as a mediator for phenotypic impact and disease has long been overlooked. Most research and genome analyses have been focused on single nucleotide variants instead. Especially at the beginning of the genomics research era, many phenotypes and diseases could be explained by SNVs as research focused on the detection of these kinds of variants. Deriving mechanistic insights was often straight forward. Genes were rendered dysfunctional by frameshift mutations, stop codons were introduced or amino acid exchanges in protein structures occurred as described above. Most of these variants were SNVs in well-studied protein coding genes (like the Factor 9 gene in hemophilia patients). However, some cases proved to be trickier, showing no SNVs in coding regions, but instead more complex variants in non-coding DNA.

3.1.2 Non-coding DNA

The non-coding segments of the genomes quickly became less junky as novel assays were designed to capture additional layers of information in the genome [72]. New fields of research emerged such as epigenetics, regulatory genomics and genome architecture. Here, non-coding regions proved to be powerful in regulating the expression levels of certain genes in time and space and establishing a diverse set of new potential disease mechanisms.

The challenge of predicting whether a novel variant in the genome will impact the health of the carrier can be extended from SNVs to SVs. However, as these types of variants (SNVs vs SVs) are fundamentally different, and different types of SVs (deletions, insertions, duplications, inversions) have varying impacts, new tools need to be developed that capture the broad spectrum of insights that have been gathered about SVs up until now.

Here I introduce a novel machine learning approach that estimates the impact of SVs on health and disease in the human genome.

3.1.3 Significance of SVs

SVs often span large regions of the genome, covering multiple megabases or even whole chromosome arms [73]. Whole genes can be affected by being deleted, inverted or duplicated in their entirety or in parts (see Figure 11). Further non-coding segments of the genome can be influenced, altering expression levels or genes in time and space. Because of their size, SVs can also alter the 3D genome architecture of a region [74], which is less affected by Single Nucleotid Variants. Breaking regulatory regimes (stretches of DNA being coregulated due to for instance physical proximity) and changing the distance between two functional regions might also influence gene expression.

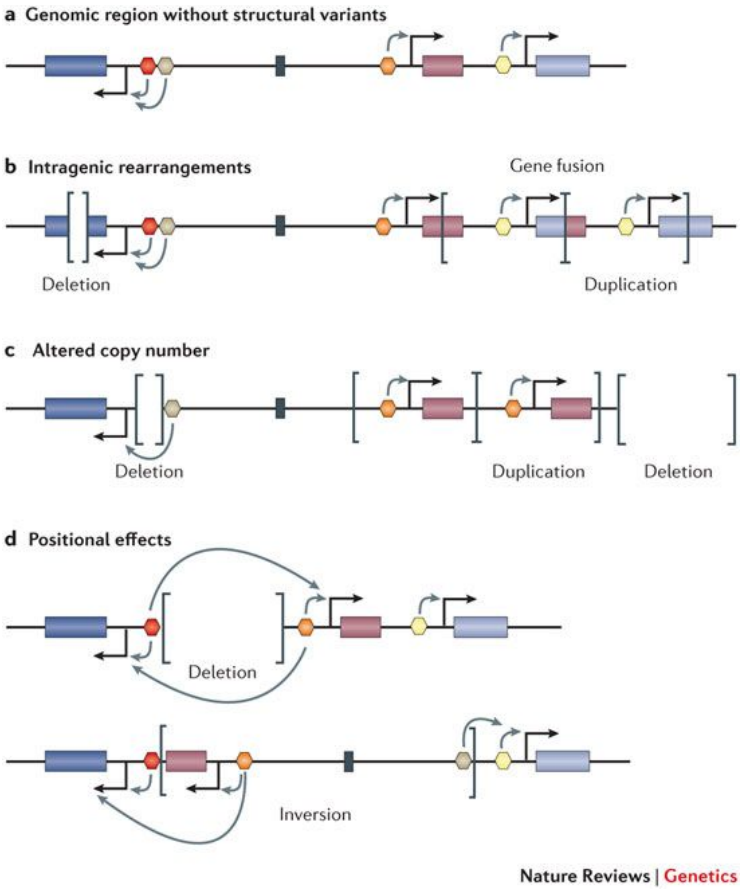


Figure 11: Significance of SVs on human health. Unaffected genomic regions (a) can be deleted or duplicated (b). Copy number alterations (c) can lead to deletion or duplication of regulatory elements (hexagons) or whole or parts of genes (rectangles). Further positional effects, such as deleting or inverting 3D genome architecture boundaries (black). [75]

3.1.4 SV alignment tools and biases

As mentioned above, SVs are classified in various types that differ from one another by their mechanism of creation as well as their mode of phenotypic impact.

Methods primarily designed to detect SNVs were unable to capture the full spectrum of structural rearrangements in the genome [76]. Even though SVs are much larger as SNVs, sometimes spanning multiple megabases, their detection turned out to be challenging due to the nature of short read datasets generated in shotgun sequencing. A deletion can be detected by the loss of reads aligning in the region (copy number loss) and therefore complete absence of alignable reads (homozygous deletions, describing the absence of both copies on the homologous chromosomes) or a drop in coverage (heterozygous deletions, describing the absence of only one of two copies) while duplications can be detected by a gain

of coverage (copy number gain), depending on the number of duplications. Further, reads spanning the break points of the SV can be used to increase the resolution of the size and exact location of the SV and are the only information accessible for inversions, as these SVs are copy number neutral and do not affect coverage [76].

However, due to the mechanistic creation of SVs, breakpoints are often located in repetitive sequence that is not uniquely alignable to the reference genome. In addition, many bioinformatic pipelines discard reads that cannot be aligned to the reference genome due to too many errors in the alignment process, leading to the loss of crucial information of the breakpoint reads necessary to detect SVs [77]. In recent years many pipelines have been developed to call SVs from short read sequencing data, each looking at individual signatures of the certain classes of SVs.

However, a comparison of SV detection pipelines discovered a surprisingly small overlap of SVs called by all tools, showing the biases generated by weighing information about SV presence differently [69]. Recent advances using population data and new technologies such as long read sequencing assays [78] or even SV detection using microscopes [79,80] greatly improved the detection of SVs and therefore catapulted SV interpretation into the focus of evolutionary, genomic as well as clinical researchers.

3.1.5 SVs and clinical significance

Just like SNVs SVs can impact phenotypes and cause diseases. Various papers already link specific SVs to human disease [21,74,80,81].

SNVs are less likely to impact the 3D genome architecture of a region compared to large rearrangements. A prime example of human diseases mediated by SVs are limb malformations studied by the research group of Prof. Dr. Mundlos in Berlin. The researchers showed that deletion or inversion of a genomic region involved in early embryonic development of limbs can impact healthy formation of the limb even without affecting the implicated genes themselves. For example, Spielmann et al [20] established a model of SVs rearranging the regulatory regime (see Figure 12) organized in Topological Associated Domains (TADs). TADs are genomic segments of the genome that are held in close proximity by regulatory architecture proteins such as *CTCF* or *cohesin*. Aspects of the role of CTCF in human brain evolution is discussed in Chapter 4. Genes situated in the same TADs are often coregulated and coexpressed by a common set of regulatory elements such as enhancers [82]. However, breaking the boundaries of TADs by

deletion or inversion can lead to reformation of the DNA stretch and therefore novel gene expression due to contact with a different set of regulatory elements. These deletions and inversions, originally found in human disease phenotypes (see Figure 12) that were unexplained previously were shown to be causal in mouse models.

Detecting and differentiating the disease-causing variant from a variant that has no, or very little functional consequence remains challenging.

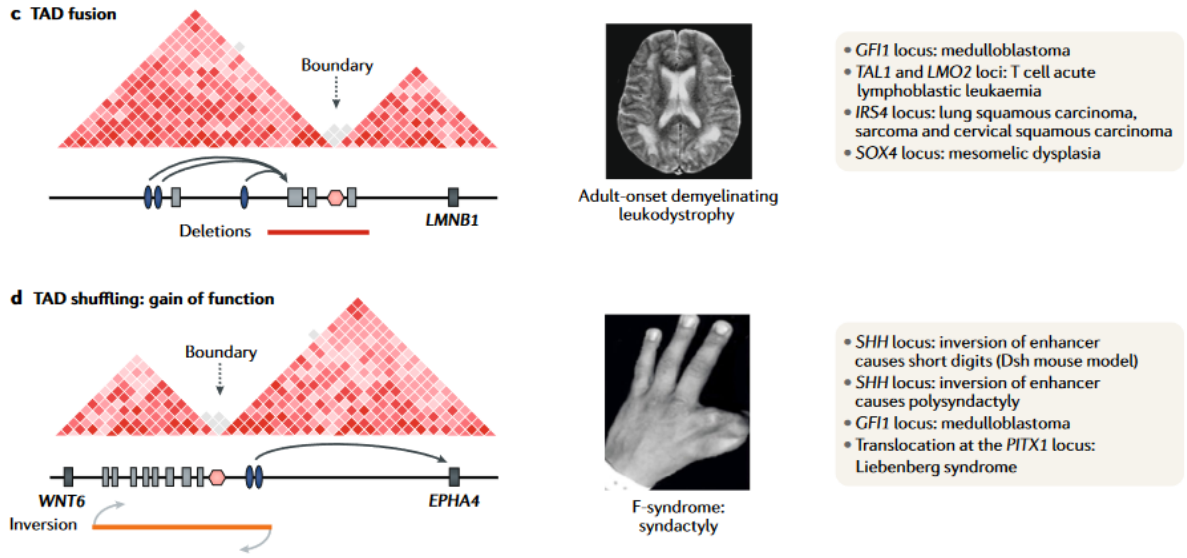


Figure 12: Depiction of disease-causing loci in the human genome. SVs (deletion top, inversion bottom) are shown to affect phenotypes (left). Red triangles represent a heatmap of regions in contact to one another along the DNA sequence. Boundaries can be affected by SVs leading to expression changes of nearby genes, that in consequence alter phenotypes.[20]

3.1.6 Feature set motivation

To be able to comprehensively interpret novel structural variants, a set of features needs to be gathered that represent the wide spectrum of current biological and medical insights into putative functionality and pathogenic mechanisms. The following paragraphs introduce the motivation behind the selection of these features. Further details about the feature transformation as well as a table with all features and their respective categories (see Table 1) can be found in the results section.

3.1.6.1 Integrated Scores

CADD-score [83] as well as LINSIGHT score [84] are widely used scores to estimate the effect of single nucleotide variants in the human genome. Both scores integrate a variety of features to estimate functionality on a base pair resolution throughout the genome. While CADD, as the small variant model for CADD-SV, uses evolutionary derived variants to train a machine learning model, LINSIGHT trains on loci under selective pressure vs neutral regions and just like CADD infers functionality, integrating a wide set of genomic features. Both scores summarize variant effects in a single score beyond coding variants genome-wide, which make them highly informative for CADD-SV. Both methods have been extensively tested on various datasets and were shown to be powerful in prioritizing functional variants on various datasets. CADD-SV calculates a max (most pathogenic) value of the scores over the span and flank of an SV, the sum of all present scores, as well as the amount of top 10% scores from the score distribution of both integrated scores over the span and flank regions of an SV. All transformations give insight into presence as well as abundance of functionally informative base pairs within the span of the SV (see Methods Section 3.2).

3.1.6.2 Species conservation and constraint

Conservation scores provide further information about the functionality of variants. Regions in the genome that undergo little change throughout an evolutionary trajectory are considered crucial for the general functionality of an organism. Purifying selection maintains the exact genomic sequence in species that diverged many million years ago as changes might render the coding or non-coding stretches dysfunctional. Hence, highly conserved genome stretches are a strong indication for functional sequence.

3.1.6.2.1 PhastCons

CADD-SV uses PhastCons (PHylogenetic Analysis with Space/Time models on CONServation) scores, a Hidden Markov model that estimates the likelihood of a given base pair to be part of a conserved region, based on multiple genome alignments [37]. Three scores are used, derived from three different multiple genome alignments. PhastCons20way uses a multiple sequence alignment of 20 vertebrate species, PhastCons30way uses 30 vertebrate species, and PhastCons100way, which uses mammals, birds, fish and other species groups to infer conserved DNA stretches. The three different scores represent different evolutionary timescales of conservation. CADD-SV summarizes these scores over span and flank of an SV

by extracting the max (most conserved) position as well as the sum of all top 10% scores within the individual score distribution (see Methods Section 3.2).

3.1.6.2.2 GERP

Further CADD-SV makes use of GERP (Genomic Evolutionary Rate Profiling) which uses simulations to estimate the strength of selection per base pair throughout the human genome by comparing the expected number of naturally occurring substitutions to the observed rate using multiple species alignments [85]. GERP also represents species conservation and is powerful detecting short-lived functional sequence stretches that arose recently in the human genome. CADD-SV transforms the GERP annotation into max (most conserved) values and the sum of the top 10% scores, as well as the amount of top 10% conserved position within a given SV (see Methods Section 3.2).

3.1.6.2.3 Syntenic regions

To infer the impact of a novel SV on the conservation of gene order in its genomic context, I use information about syntenic information from synteny mapper [86]. Synteny describes the maintenance of blocks of genomic regions in the same order throughout evolution. Breaking co-expressed genes apart in their physical proximity might interfere with the function of these genes. Purifying selection maintains the localization of certain genomic ensembles. CADD-SV calculates the distance or overlap of the SV to the next conserved syntenic block [86].

3.1.6.2.4 Ultra-conserved regions

Finally, CADD-SV uses ultra-conserved regions inferred from a 120 species multiple sequence alignment that highlights genomic stretches that are maintained by very strong purifying selection (inferred using GERP) throughout the entire vertebrate lineage [87]. CADD-SV summarizes this annotation, which is unlike GERP and PhastCons not a score but represents genomic loci in a bed-format, as number of ultra-conserved elements overlapping, number of bp overlapping, as well as fraction of SV being ultra-conserved.

3.1.6.3 Epigenetic and regulatory activity

Epigenetic and regulatory annotations may provide hints for several potential non-coding disease mechanisms and are crucial for a meaningful genome-wide score of SV pathogenicity. Epigenetics, as described above, are heritable genome alterations that are not encoded in the DNA-sequence itself. These sets of features impact functionality by changing the expression levels of genes in time and space by defining when and where a genomic sequence is being accessible to the molecular machinery. CADD-SV uses various public datasets that fall within this category. The caveat of epigenetic annotations is that they, unlike species constraint metrics, are variable in different cell lines and cell-types throughout the human body. Therefore, some experimental assay derived datasets were used from a specific cell line which is stated in each paragraph.

3.1.6.3.1 ENCODE

The Encyclopedia of DNA elements (ENCODE) was a large initiative that followed on the Human Genome Project, intended to identify functional elements in the human genome. In various phases datasets were generated and made publicly available [8]. Some members of the ENCODE Consortium considered 80% of the human genome to be functional [8]defining function from molecular activity as containing RNA expression, histone modifications, DNaseI hypersensitive sites or transcription factor binding sites. The definition of functionality however is now without controversy: expression of pseudogenes, for instance, is considered by many scientists as a transcribed stretch of the human genome that is not functional.

3.1.6.3.2 DNase-seq

I use DNase-seq peaks from ENCODE to infer accessibility of genomic regions. In this assay DNase I, an endonuclease that cuts DNA, is used to cut regions in the genome that are not tightly packed. Hence, these regions are considered as open chromatin, providing access to enzymes and transcription factors to interact with the DNA. The DNase experiments were conducted on A549, an adenocarcinoma epithelial cell line, widely used as a model cell line in basic research as they are well characterized and easy to culture [88]. CADD-SV uses the max values as well as the sum over all encountered ENCODE-provided accessibility scores.

3.1.6.3.3 RNA-seq

To measure the presence and abundance of RNA of transcribed regions of the genome, CADD-SV uses an RNA-seq dataset from ENCODE conducted on GM12878 (a female fibroblast cellline). RNA-seq is an assay that uses reverse transcribed RNA molecules followed by next generation sequencing in biological samples [89]. The output gives insight into expression levels of genes as well as other forms of transcribed RNA molecules such as long non-coding RNAs. CADD-SV uses max scores as well as the sum of all values over the affected region.

3.1.6.3.4 Histone modifications

Molecular modifications of the tails of histones (Figure 13), molecular complexes that pack DNA like a spool, are an additional way for an organism to regulate DNA expression by communicating with cellular factors or altering chromatin structure. Various modifications are known with varying cellular functionality. Methylation or acetylation of various tails from the different core proteins mediate regulation of expression or accessibility of genomic DNA stretches. CADD-SV uses H2AFZ, H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3 and H4K20me1 tracks from ENCODE [8]. The number behind the H stands for one of the four core proteins that form the histone tetramere, the letter and number after this represents the aminoacid (K stands for Lysin) and its corresponding number in the aminoacid chain of the protein, while the last part stands for the modification itself: me1, a single methylation group; me2, demethylation; me3, trimethylation; and ac acetylation. I used experiments conducted on IMR90 (human fibroblasts isolated from lung tissue) or HepG2 (human liver cancer) cell lines. All histone modifications are integrated into CADD-SV as a max and sum value over the affected region.

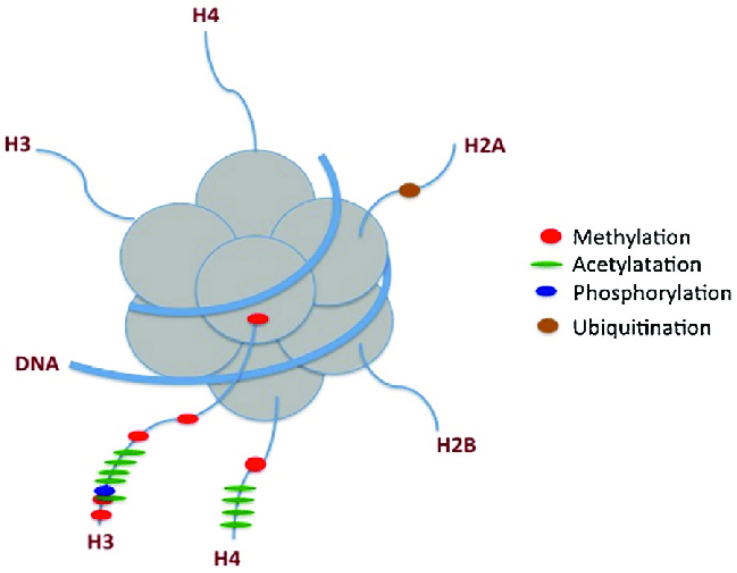


Figure 13: Schematic representation of the histone complex and potential modifications [90].Individual Histone molecules are shown in grey, DNA bound by the histone shown in blue, histone tails (outreaching ends of the protein sequence) are represented in thin blue lines, being identified by the individual histone molecules (H3/H4/H2A or H2B). Putative modifications of the histon tails are shown in red, green, blue and brown.

3.1.6.3.5 chromHMM

chromHMM is a software that utilizes a multivariate Hidden Markov Model to characterizes the chromatin state of DNA. Based on various feature annotations chromHMM[91]. provides a genome-wide biological characterization of genomic regions into 25 functional categories such as active promotor, strong/weak enhancer or heterochromatin [91].

3.1.6.3.6 ReMap

To specifically look at proteins that interact with the DNA to modulate transcription (so-called transcription factors) CADD-SV uses a summary statistic from the ReMap project [92]. Here, manually curated public datasets derived from ChIP-seq or DAP-seq experiments are used to infer presence of one or more transcription factors on DNA stretches. For an introduction into the experimental procedures see section 4.2.5.

3.1.6.3.7 GC content

The amount of Guanine and Cytosine (GC) pairing in the DNA is informative about the binding energy and stability of the DNA double strands. It further correlates with biological function, for example gene bodies contain a higher proportion of GC compared to the genomic background. Further CpG islands, particularly GC rich regions, are shown to have regulatory activity and are often found in promotor regions [93].

3.1.6.4 Population and disease constraint

Conservation scores give insight into sequence constraint over long evolutionary time spans. However, shorter timespans are better represented by human derived population metrics. Purifying selection within human populations, for instance, can be inferred by the absence of variation in a given genome position in large, putative healthy population cohorts such as gnomAD[19]. In addition, linking variants to human diseases can be informative.

3.1.6.4.1 pLi

The absence of loss of function variants in a gene in large, putative healthy human datasets is indicative of the essentiality of this gene. In contrast, the presence of loss of function variants in a gene in a healthy individual might indicate non-functionality. CADD-SV uses pLi (Probability of being loss of function intolerant) [94], a gene-based score that characterizes genes on a score from 0 to 1, with 1 meaning highly intolerant to loss of function variants.

3.1.6.4.2 Constrained Coding Regions

Havrilla et al. use 123,136 genomes from putative healthy individuals to identify human coding regions under purifying selection [95]. Unlike pLi this score is not a per gene metric but identifies highly constrained (absence of variation) regions within genes. The score ranges from 0 to 100, with 0 representing the presence of a variant in at least one healthy individual and 100 representing the most constrained regions.

3.1.6.4.3 Haploinsufficiency

Due to the diploid structure of the genome, all genes are present with at least two copies in every cell. However, for some genes, maintaining one functional copy is not sufficient (haploinsufficiency). This mechanism describes for example dominant genetic diseases where rendering one allele dysfunctional is enough to mitigate a disease phenotype. The Deciphering Developmental Disorders study has published a score that estimates the likelihood per gene to be haploinsufficient based on a large cohort of children with developmental disorders and their parents [96].

3.1.6.4.4 Missense badness, PolyPhen-2, and Constraint (MPC score)

An additional exonic score based on a human population dataset (ExAC) that is integrated in CADD-SV is the Missense badness , PolyPhen2 [97] and Constraint metric (MPC), ranging from 0 to 5, with higher values expressing increasing deleteriousness[98]. MPC estimates the expected variation in exonic sequence using sequence specific mutation rates and compares this from the observed variation in ExAC [98].

3.1.6.5 3D Genome organization

The 3D architecture of the genome is crucial for maintaining its function. Especially structural variants can interfere with the genome organization as large stretches of DNA are affected. I utilize various independent metrics of 3D genome architecture described in this section.

3.1.6.5.1 CTCF binding sites

The CTCF protein plays a crucial role in defining DNA loops and therefore often functions as a "communication" mediator or boundary for regulatory elements [99]. SVs overlapping CTCF binding sites might disrupt gene regulation by rewiring enhancer contacts or changing the proximity of co-regulated genes. Here, I use a ChIP-seq assay dataset from 19 cell lines that describes binding sites of CTCF [100]. ChIP-seq is an experimental approach to determine protein occupancy on DNA. CADD-SV integrates this dataset by using the number of binding sites encountered, the number and fraction of base pairs defined as

binding sites as well as the distance of the next binding site to the breakpoints of the SV (being 0 when a CTCF binding site is present within the SV).

3.1.6.5.2 Directionality Index

Directionality index is a term derived from Hi-C experiments, i.e. proximity ligation experiments unveiling DNA segments that are close in physical space. In brief, directionality index values represent the directionality of DNA contacts in a given region of the genome. With very high or very low values representing this position to be specifically contacting regions in a specific direction along the DNA strand, suggesting the presence of a contact boundary in the other direction. This index provides information about the structure of a given genomic region. Extreme values may indicate highly structured DNA stretches, where maintaining the 3D architecture might be crucial for functionality [101]. CADD-SV uses max and min values of processed datasets from GENOMEGITAR that summarizes Hi-C experiments from multiple cell lines [102].

3.1.6.5.3 Enhancer-Promotor Links

Many regulatory elements like enhancers communicate with the Transcription Start Site (TSS) by getting into close proximity. FOCS (FDR-corrected OLS with Cross-validation and Shrinkage), a method to identify enhancer promotor contacts, uses variable sources of experimental input datasets to infer links based on correlated activity patterns of promotors and enhancers [103]. Breaking this link might lead to misregulation of genes in time and space. CADD-SV integrates this score by calculating the overlap of an enhancer-promotor link as well as the distance from the breakpoints to an existing link.

3.1.6.5.4 Frequently Interacting Regulatory Elements (FIRE)

I use a dataset from Schmitt et al. that defines frequently interacting regulatory elements (FIRE) that, as the name suggests, are often in close contact with each other [104]. This metric is derived from contact maps of 14 human tissues and seven cell types. CADD-SV uses max and min scores from five cell lines: GM12878

(a lymphoblastoid cell line), MSC (human mesenchymal stem cells), MES (embryonic stem cells), IMR90 (fibroblast cell line) and h1 (embryonic stem cells).

3.1.6.6 Gene and element annotation

Most annotations that were previously described are genome-wide scores that emphasize the mission of CADD-SV to provide features that extend beyond the coding sequence of the genome. The function of some non-coding DNA stretches, previously considered junk DNA in its entirety, is still challenging to predict. However, I also provide CADD-SV with gene and element annotation to add information about coding sequence specifically. I use human gene model annotations from ENSEMBL [105], containing the coordinates of genes, transcripts, exons, start codons, stop codons, 3- and 5-prime untranslated regions as well as coding sequence (CDS).

3.1.7 Machine Learning

"Machine learning describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks."[106] One example is the classification of variants as pathogenic or benign. Training datasets can be used to infer the status of novel data points (generalization). In Classification models are trained to correctly predict labels from pre-assigned labeled datasets. Problems are often described by minimization of a loss function on a training example. Optimizing this loss function improves the model and therefore the accuracy of the predictions.

One of the main problems of machine learning approaches is to stop the trained model from overfitting. Here, the model learns intrinsic features of the labelled training sets by heart to near perfection but is unable to predict new datasets well. To estimate model performance, a hold-out set of the original training dataset is held back to test the model on the same type of labels using an unseen dataset (from the model's point of view). This set is often referred to as holdout or test set. In addition, to validate the capacity of the model beyond biases in the existing dataset, an additional, independent validation set can be used that, for instance, is a labelled dataset where the labels derive from a different kind of assay [107].

3.1.7.1 Supervised learning (clinically labelled)

Using supervised learning approaches has its advantages as insights are easier to interpret as the outcome measures the distance to existing, known labels. However, biases in the labelled dataset will be learned by the model and therefore propagate into the predicted results. Classic examples are image classifiers distinguishing horses from dogs where all horse imagines were retained from a labelled dataset containing the imprint of the horse farm and therefrom dog pictures with imprints being classified as horses. Or certain images primarily taken at night make a classifier distinguish between day and night pictures instead of the signatures of horses and dogs.

Choosing an unbiased dataset is therefore crucial to maintain meaningful interpretations. CADD-SV uses an unbiased evolutionarily motivated approach that was first developed for SNVs to score the impact of deletions, insertions and duplications in the human genome.

3.2 Methods

The following pages describe the methodology behind CADD-SV, a machine learning approach that classifies putative pathogenic and benign SVs.

3.2.1 Training dataset

Using an unbiased training dataset is crucial to get meaningful insights into a problem instead of propagating a bias into the results. Existing datasets that label pathogenicity of SVs exist, however they are heavily biased towards very large SVs, towards coding regions in the genome and particularly towards well studied genes. In addition, these datasets are sparse, even though having increased in size during the timeline of this project as research turned towards improved detection and interpretability of SVs. ClinVar provides a reference dataset composed of 3,262 deletions, 82 duplications and 78 insertions that are labelled as “pathogenic” and few SVs labelled as “benign”. These SVs are particularly large, compared to SVs in the general population, enriched in transcription start sites as well as having a high proportion of well-studied genes with high gene-pathogenicity scores. This is due to researchers and clinicians tending to focus on certain well understood mechanisms, diseases, or their favorite genes (see Figure 14). Especially small SVs, impacting non-coding functional sequence segments as well as genes not in the research focus could be wrongly classified using a classifier trained on labeled pathogenic and benign ClinVar annotations.

3.2.1.1 Evolutionary set

Instead of relying on labelled datasets from ClinVar, I opted for an evolutionarily motivated dataset. A related approach was first applied by Kircher et al. for short variants called CADD – Combined Annotation Dependent Depletion. Here, evolutionarily fixed variants between chimpanzee and humans are considered to be non-disease causing and therefore benign or neutral as millions of years of purifying selection would have removed variants that are disadvantages for the carrier. In contrast to CADD for shorter sequence variants, I do not only use fixed variants in humans but also consider fixed SVs in the chimpanzee genome (see Figure 15). The motivation behind this is, that evolutionary deleted sequence in the human genome is not present in the reference genome and therefore absent in experimental read outs as well. However, sequence deleted in the chimpanzee can be mapped back to the human genome

containing all possible information over the sequence span. These variants are spread over the entire genome while being slightly more common in repetitive regions around the telomeres.

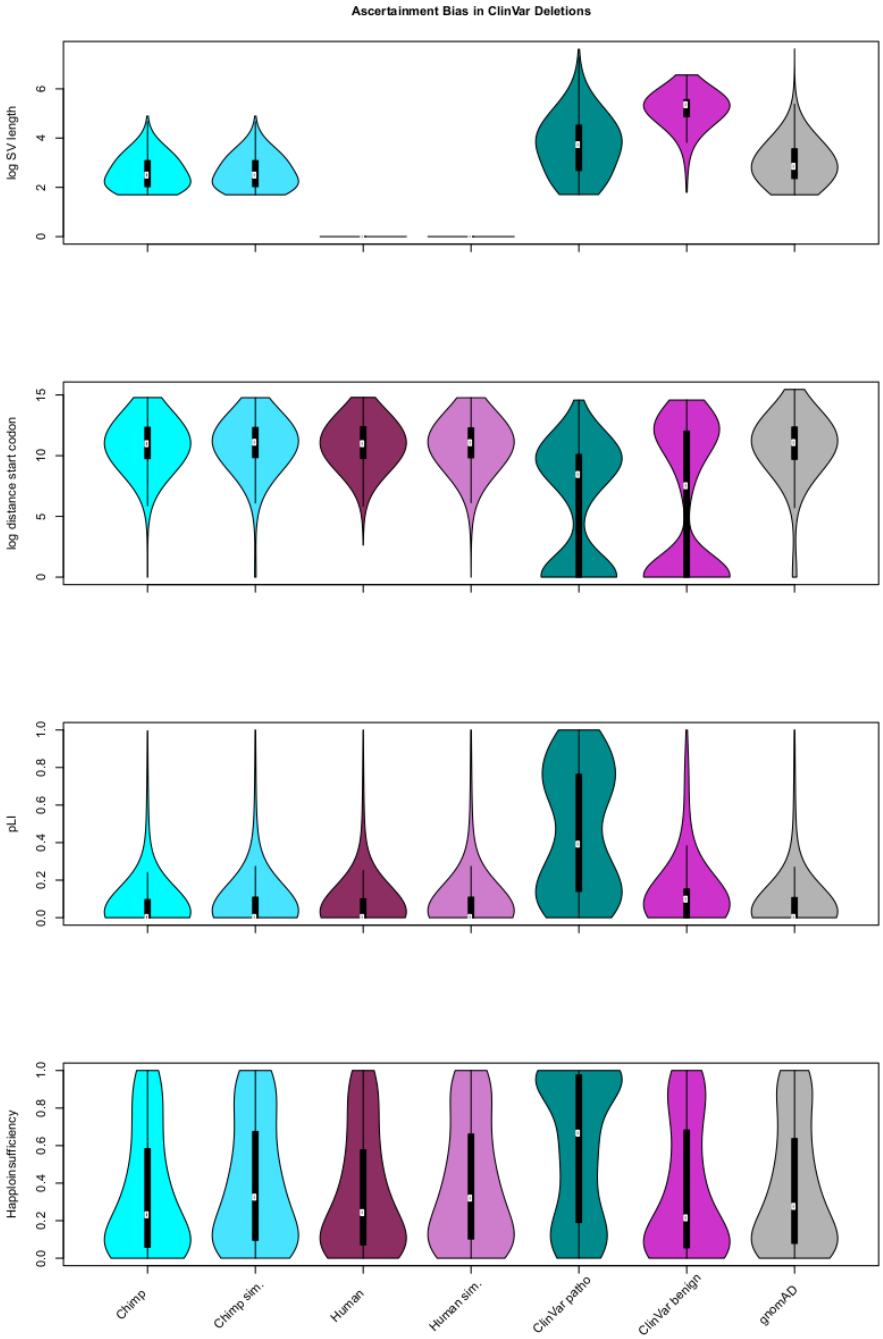


Figure 14: Ascertainment bias in labeled deletion datasets. To make accurate predictions using machine learning it is crucial to have an unbiased dataset to train on. ClinVar pathogenic or benign labelled deletions are hand curated and individually verified but are biased towards very large deletions and are clustering around well-studied genes (as shown in the excess of high pLI and Haploinsufficiency scores). Our evolutionary derived dataset however does not suffer from these kinds of ascertainment bias and is similar to the occurrence of deletions discovered in a population cohort (gnomAD-SV v2.0)

These regions, for reasons explained above, are in the following named as proxy-neutral, as I consider them a representative unbiased example of neutral variants.

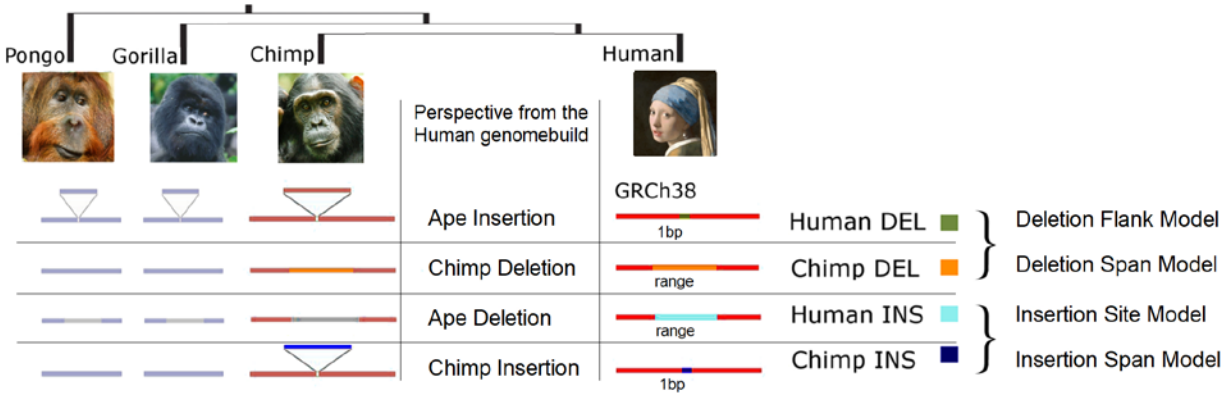


Figure 15: Motivation of Training Dataset for the CADD-SV framework. Human and chimpanzee derived Structural Variants are considered to be neutral or beneficial if they reached fixation. Therefore, previously identified human and chimpanzee derived SVs [108] are used as proxy-neutral training dataset. Top panel describes the evolutionary relationship between Orang Utan (Pongo), Gorilla, Chimp and Humans, with Chimps being humans closest relative. Left Panel depicts DNA sequences from all species with Insertions being described as triangle-structures while Deletions are faded in grey. Centered below the human depiction are genome variants shown as present in the human genome build. Human DEL: are sequence stretches absent in the human genome and therefore are interpreted as inserted sequence in all ape genomes. Chimp DEL is depicted as absent sequence in chimp genomes while being present in the other species. Human INS: present in the human genome build but described as deletions in all other apes; Chimp INS: insertion in the chimp genome while being absent in humans. Each set of SVs is used to train an individual model that uses either Flank or Span or Insertion Site as training data.

3.2.1.2 Proxy-deleterious set

To provide the classifier with an unbiased set of putative pathogenic variants, I use simulated variants randomly distributed throughout the sections/parts/regions in the genome where the proxy-neutral set is derived from. Therefore, encountering variants by chance that contain functional DNA. This set will be referred to as proxy-deleterious. The simulated set is matched in size and length distribution to the proxy-neutral set for all used deletion and insertion sets.

The mislabeling in both datasets should not be neglected as many proxy-deleterious variants do not fall within functional sequence and some proxy-neutral variants are putatively functional, as humans and chimpanzees do show some phenotypic differences. However, the depletion of functional variants in such neutral sets has proven sufficient to powerfully predict the effects of short variants[83].

3.2.2 Model

As mentioned above, various methods exist in machine learning, each powerful for certain tasks and having potential pitfalls. Here, I use Random Forrest classifiers using the R Package “randomforest" as well as the standard implementation of linear regression.

3.2.2.1 Random Forest Classifier

Random forest classifiers are named after a forest of decision trees bundled together to have a multitude of potential paths leading to the desired classification. Apart from classifying they can also be used for non-binary outputs.

3.2.2.2 R Package

I used the R package (v3.5.1) “randomforest” to train the random forest models contrasting proxy-neutral and proxy-deleterious sets. I followed the best practices of the manual adjusting the parameters as recommended [109].

3.2.2.3 Parameters

Random Forest models can be optimized by adjusting the depth and number of trees in the forest. I optimized the model using a hyper parameter search for the number of trees (ntree) and the depth using (maxnodes) and (nodesize). I considered up to 1000 trees and nodesize and maxnode parameter values of 10, 50, 100, 250, 500, 1000 each. Parameters were optimized individually using fixed values of 100 for the other parameters.

3.2.3 Feature transformations and annotation

CADD-SV uses a wide variety of coding and non-coding genomic features to make predictions on SV pathogenicity genome-wide. All features need to be summarized, as most of them have base pair resolution while SVs span by definition at least 50 bp, sometimes megabase pairs. The CADD-SV framework relies on custom shell and R scripts, wrapped in snakemake rules, that query annotation files,

extract regions of interest using tabix, and summarize features using AWK or BEDtools [110]. Feature values are stored in BED (Browser Extensible Data) format.Further, information about the pipeline is provided below. All features and their respective transformations are summarized in Table 1.

3.2.4 Implementation

CADD-SV is implemented in Snakemake, using conda for dependency management. CADD-SV was designed to be applicable for bioinformaticians and clinicians alike. The source code for the framework is available for download on GitHub (*https://github.com/kircherlab/CADD-SV/*). Conda and Snakemake guarantee easy installation procedures as well as stability through dependency management. Further, I implemented CADD-SV to be time and memory efficient, while being highly parallelizable for application on a cluster-network. A set of 1,000 short SVs can be scored on a regular laptop in 13 minutes using 600MB of memory. However, in contrast to all competing tools, CADD-SV jobs are highly parallelizable, strongly improving time-performance. In addition to the source code, a webservice (*https://cadd-sv.bihealth.org/*) allows for online scoring of SVs in a BED-like format as well as for obtaining results for different human genome builds (GRCh38; NCB16 & GRCh37 through automated coordinate liftover). In addition, pre-scored variants from cohorts such as gnomAD or ClinVar can be queried online including all feature annotations. For better interpretability, feature outlier values are color-coded based on their Z-scores.

3.2.5 Model Performance Assessment

To assess the performance of machine learning algorithms it is important to define test and validation sets. As many approaches suffer from overfitting, a hold-out dataset can be used to measure model performance. Here, I randomly withheld 10% of the annotated SVs as holdout-set. Validation sets are independent datasets not used in training of the model. To validate CADD-SV performance I use a wide variety of independent datasets.

Model performance was estimated using Area under the Receiver Operator Characteristic Curve using the R Package PRROC [111].

3.2.6 Validation sets

CADD-SV was designed to not use curated SV sets in training, it does not derive features from clinical datasets such as ClinVar or OMIM, and it does not use gnomAD-SV allele frequencies as features either. Therefore, CADD-SV can be validated using those datasets.

3.2.6.1 ClinVar

Pathogenic and benign annotations for clinical SVs [35] were downloaded from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar) on June 24th, 2021. Only variants with pathogenic or benign labels, of at least 50bp length and annotated as deletion (pathogenic n = 3262, benign = 33), duplication (pathogenic n=82, benign n = 4) or insertion (pathogenic n = 78, benign n = 18) are considered. Further, to increase the number of pathogenic insertions, unique pathogenic insertions (n = 39) reported by Hancks et al.[112] and Gardner et al. [113] were added. Area Under the Receiver Operating Characteristic (AUROC) metrics are calculated using the PRROC R-package [111].

To show the clinical benefit of prioritization of SVs using CADD-SV, I use genotyped SVs from the 1000 Genomes project [114] and add one (randomly selected) labelled pathogenic SV found in ClinVar into the reported set of individual specific SVs. From the 1000 Genomes' SV events, I consider Alu and Line1 SVs to be insertions. I report the rank of the pathogenic SVs within the complete SV set.

3.2.6.2 Putative healthy population cohorts

Germline SVs identified from healthy individuals over various populations [19] were downloaded from gnomAD-SV release v2.0 (https://gnomad.broadinstitute.org/downloads). Allele frequency values of common and ultra-rare SVs are determined across all available populations. Common variants are defined as minor allele frequency greater 0.05, ultra-rare variants are defined as singletons.

3.2.6.3 Non-coding SVs

To assess CADD-SVs ability to prioritize functional stretches of DNA, I use healthy population SVs from gnomAD-SV containing a genome wide association study (GWAS) linked SNV. I assume that presence of

an association with a functional trait can be seen as a proxy for functional SVs. The GWAS catalog was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/gwasCatalog.txt.gz. Further, I use deletions and insertions reported to be associated with changes in gene expression patterns as well as SVs under natural selection [115], both hinting towards functional stretches of DNA that are beyond coding effects.

3.2.6.4 Cancer

Somatic SVs (n = 95,749) from cancer patients were obtained from the International Cancer Genome Consortium [116] at https://dcc.icgc.org/api/v1/download?fn=/PCAWG/consensus_sv/final_consensus_sv_bedpe_passonly.icgc.public.tgz. In addition, insertions reported in cancer genomes were taken from Qian et al. (n = 18) [117]. To assess the performance of CADD-SV beyond coding regions, I use non-coding SVs (n = 687) that are known to impact human gene expression in data from the GTEx consortium [118].

3.2.7 Tool comparison

CADD-SV performance on various validation sets was compared to existing tools SVScore [119], AnnotSV [120], StrVCTVRE [121], and the TAD-fusion-score [122] using standard parameters. TAD-fusion only scores deletions and was primarily developed to identify 3D genome alteration. As SVScore and TAD-Fusion scores were not available for the current genome build GRCh38, UCSC liftover [123] was used to transfer SV coordinates and respective scores.

3.3 Results

3.3.1 Training dataset

Machine Learning methods strongly rely on the quality of training datasets to yield meaningful predictions. Using clinical databases such as CinVar or HGMD to curate an annotated training dataset is challenging for SNVs or small indels, where it requires a careful matching of pathogenic and benign variants in genomic regions and effect classes [36,84]. This seems currently impossible for SVs. The ClinVar dataset [35] is very sparse for SVs, i.e. only few (3,262 deletions, 82 duplications and 78 insertions) and mostly very large SVs (mean size of 106 kb for deletions) are being annotated. This is insufficient for an insightful training dataset, especially as population-derived SVs are much smaller in genomic size (mean of 7.4 kb). Further, when compared to large population SV sets [19] strong biases towards high effect variants and clustering around well studied genes are apparent (Figure 14). Therefore, I opt for an unbiased evolutionary set of SVs obtained from comparisons in the great ape lineage [108]. A key strength of this approach is that the model is trained on a larger training set of 19,113 deletions and 26,823 insertions and duplications that does not suffer from the ascertainment bias inherent to curated sets.

3.3.1.1 Combined Annotation Dependent Depletion

This is motivated by the Combined Annotation Dependent Depletion (CADD) framework, an approach that has proven powerful in the interpretation of SNVs and short indels [83]. In CADD-SV, I assume that millions of years of purifying selection removed SVs that are deleterious, i.e. have a negative impact on human or chimpanzee reproductive success. Thus, fixed SVs in humans or chimpanzees can be classified as proxy-neutral. In contrast, variants of the same size randomly drawn from the human genome are likely to contain a significant number of deleterious variants by chance. While many of the random variants will be neutral, an unknown but considerable fraction would likely be deleterious. For simplicity, I refer to these variants as proxy-deleterious. The contrast between the proxy-neutral and proxy-deleterious variant sets, i.e. the relative paucity of deleterious, phenotypically influential genome alterations in the proxy-neutral set and the resulting differences in their annotation features, is the core characteristic of what I then model as SV deleteriousness

3.3.2 Feature Annotation

A diverse set of annotations (see Methods and Table 1) was integrated to gain predictive, genome-wide models for prioritizing structural variants of phenotypic effect. While many annotations are readily available for SNVs, informative and computational efficient statistics need to be created to summarize annotations over the span of SVs. Further, distance measures can retain information about the vicinity of the impacted DNA sequence. For this purpose, I developed an automated SV annotation pipeline using the workflow management system Snakemake [50] that combines BEDtools[110] and tabix [124] with customized bash and R scripts [109]. I integrate not only coding information such as gene models but also a wide variety of regulatory annotations retrieved from ENCODE [8], such as histone modifications or DNA accessibility. In addition, I make use of functional and evolutionary scores [36,37,84,85] as well as information about the 3D architecture of the genomic region derived from Hi-C experiments [102,104,125].

Table 1: Features used in the CADD-SV model and their respective transformations. All features are categorized into 6 categories: "Integrated scores", "Species conservation and constraint", "Population and disease constraint", "Epigenetic and regulatory activity", "3D genome organization", "Gene and element annotation"

Feature	**Transformation**	**Category**
CADD	max,sum 10% top quartile distribution	Integrated score
PhastCons100	max,sum 10% top quartile distribution	Species conservation and constraint
PhastCons30	max,sum 10% top quartile distribution	Species conservation and constraint
PhastCons20	max,sum 10% top quartile distribution	Species conservation and constraint
CCR	sum	Population & disease constraint
chromHMM_1-25	sum	Epigenetic & regulatory activity
nr_ctcf_BS	sum of BS, BP, fraction, distance	3D genome organization
Directionality Index, FIRE	max, min	3D genome organization
DNase-seq	max, sum	Epigenetic & regulatory activity
H2AFZ	max, sum	Epigenetic & regulatory activity
H3K27ac	max, sum	Epigenetic & regulatory activity
H3K27me3	max, sum	Epigenetic & regulatory activity
H3k36me3	max, sum	Epigenetic & regulatory activity
H3K4me1	max, sum	Epigenetic & regulatory activity
H3K4me2	max, sum	Epigenetic & regulatory activity
H3K4me3	max, sum	Epigenetic & regulatory activity

H3K79me2	max, sum	Epigenetic & regulatory activity
H3K9ac	max, sum	Epigenetic & regulatory activity
H3K9me3	max, sum	Epigenetic & regulatory activity
H4K20me1	max, sum	Epigenetic & regulatory activity
RNA.seq	max, sum	Epigenetic & regulatory activity
Enhancer Promotor links	distance, fraction	3D genome organization
FIRE	max, min	3D genome organization
Percent GC	fraction	Epigenetic & regulatory activity
Exon	fraction, count bp, distance	Gene & element annotation
Transcript	fraction, count bp	Gene & element annotation
Gene	fraction, count bp, distance	Gene & element annotation
Start codons	fraction, count bp, distance	Gene & element annotation
Stop codons	fraction, count bp	Gene & element annotation
3' UTR	fraction, count bp	Gene & element annotation
5' UTR	fraction, count bp	Gene & element annotation
CDS	fraction, count bp	Gene & element annotation
GERP	max, fraction, sum top 10% quantile	Species conservation and constraint
A549 Hi-C	nested, unnested, distance, boundary overlap	3D genome organization
Caki2 Hi-C	nested, unnested, distance, boundary overlap	3D genome organization
escTAD	distance, boundary overlap	3D genome organization
microsyn_intra	distance, boundary overlap	Species conservation and constraint
pLI	max	Population & disease constraint
remapTF	max	Epigenetic & regulatory activity
f5_enhancers	count bp, fraction	Gene & element annotation
DDD_HaploInsuf	max	Population & disease constraint
deepC saliency	max	3D genome organization
ultra conserved regions	nr of elements, count bp, fraction	Species conservation and constraint
LINSIGHT	max	Integrated score
MPC	max	Population & disease constraint

3.3.2.1 Span and Flank Annotation

All SVs are annotated over the full span of the event as well as 100 bp up- and downstream (Figure 16). For insertions, the span of novel SVs only contains the site of integration and CADD-SV does not derive

features from the inserted sequence. While deletions directly remove putatively functional sequence, insertions and duplications interfere with molecular function by integration of additional sequence, e.g. disrupting regulatory interactions by increasing distance[126] or introducing frameshifts into coding sequence. I incorporate this in the CADD-SV modelling by deriving features from the deleted sequence (span), annotating the context of the SV (flank) and including distance features in the model. Across SV ranges, I mostly annotate max values, mean values and the amount of high impact values above the top 90th percentile of an annotation. Additionally, span and flank models use genomic distances to certain feature coordinates (e.g. genes, exons, and enhancers). All features and their transformation are described in Table 1.

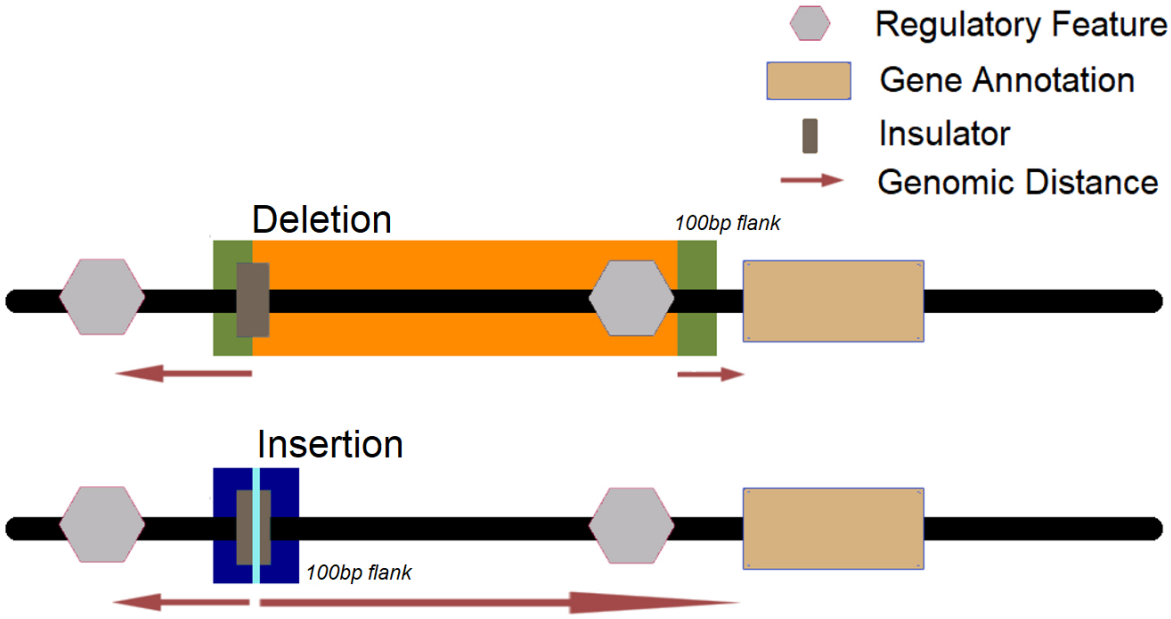


Figure 16: Depiction of implementation of the four models generated from the proxy-neutral and proxy deleterious variant sets. Deletion model predicting the span shown in orange, predicting the flank shown in green. Insertion model prediction the insertion site shown in turquoise, prediction the insertion flank shown in blue. While deletion of a novel sequence provides information about the deleted sequence in the human genome build, insertions rely more on site of integration. Therefore, flanking regions to the SVs are taken into account. Informative stretches are shown as hexagons (regulatory), gene annotations or Insulators (blocks). Genomic distances are considered for all models, depicted as red arrows.

3.3.2.2 Feature normalization

To ease later interpretation of feature impact, all features are *Z*-score transformed (mean 0, standard deviation of 1) using the annotation value distributions of the same type of SV from healthy individuals

reported in gnomAD [19]. This transformation serves primarily the interpretability of the model and does not negatively affect model training or create issues with using gnomAD variants for model validation, as the same transformation is applied for both training class labels.

3.3.3 Model Training

3.3.3.1 Parameters

For the random forest models, I limit the number and depth of the decision trees based on a hyperparameter search (Figure 17; explored ranges for ntree = {25, 50, 75, 100, 200, 500, 1000}, nodesize = {10, 50, 100, 250, 500, 1000}, maxnodes = {10, 50, 100, 250, 500, 1000}, while one parameter was optimized, the other parameters were set to 100).

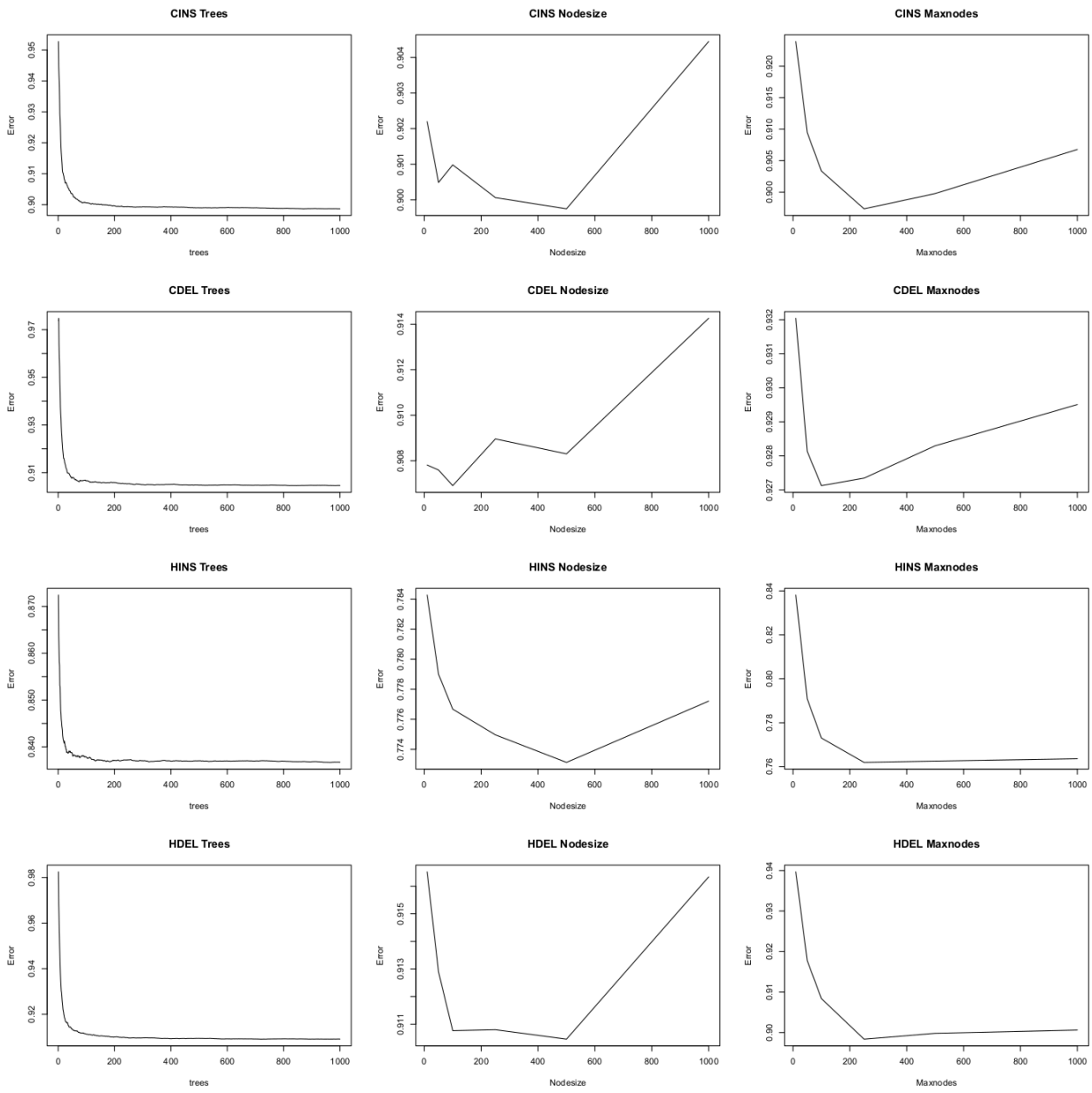


Figure 17: Hyperparameter search for all Random Forest models. I explored the mean of squared errors for a random subset of 1000 SVs for the parameters "nodesize" and "maxnodes" for both values of n= {10, 50, 100, 250, 500, 1000} and chose parameters minimizing error and overfitting (by inferring validation performance). The number of trees (ntree = {25, 50, 75, 100, 200, 500, 1000}) for each model was chosen based on observing no further improvement of error by increasing the number of trees. CINS: Chimp Insertions, CDEL: Chimp Deletions, HINS: Human Insertions, HDEL: Human Deletions

SV mediated pathogenicity depends on the type of SV. I implement separate models for deleted (DEL), inserted (INS), or duplicated (DUP) sequence. Due to the lack of training data for inversions and translocations, I can currently do not train models for these variant types. Using the described training data sets, I train four types of models (Figure 15). I train models of human-derived deletion (human DEL) and insertion events (human INS) against respective sets of equally sized events drawn across the genome.

Further, models based on chimp insertion (chimp INS) and deletions events (chimp DEL) are trained. Here, I project the events onto the human reference sequence and use the human annotations. While the human events are also manifested in the human reference, the chimp events allow us to use human annotation unaffected by an actual SV event. Hence, chimp DEL models are similar to how I would score new events observed in an individuals' genome aligned to the human reference sequence. In contrast, no human annotation for human derived deletions can be obtained over the span of the deletion as experimental readouts and conservation score are not available for the missing sequence. Similarly, chimp INS provide an insertion model based on events that did not impair human annotations or biochemical readouts.

3.3.4 Model implementation

To score novel SVs in the human genome, I exploit this relationship by training the span of novel deletions with the chimp DEL set and train the sequence 100bp up- and downstream of the breakpoints using the human DEL set. As the inverse applies for insertions and duplications, i.e. chimpanzee insertions do not span sequence in the human genome build while human derived insertions do, I use the chimp INS set for the insertion site and the human INS set for the up- and downstream sequence. Duplications are scored using the full sequence span of the duplicated locus, hence using the chimp DEL model for the span and human INS model for the up- and downstream sequence. The final score is calculated from the maximum (more deleterious) value of both models. See Figure 18 for the CADD-SV workflow.

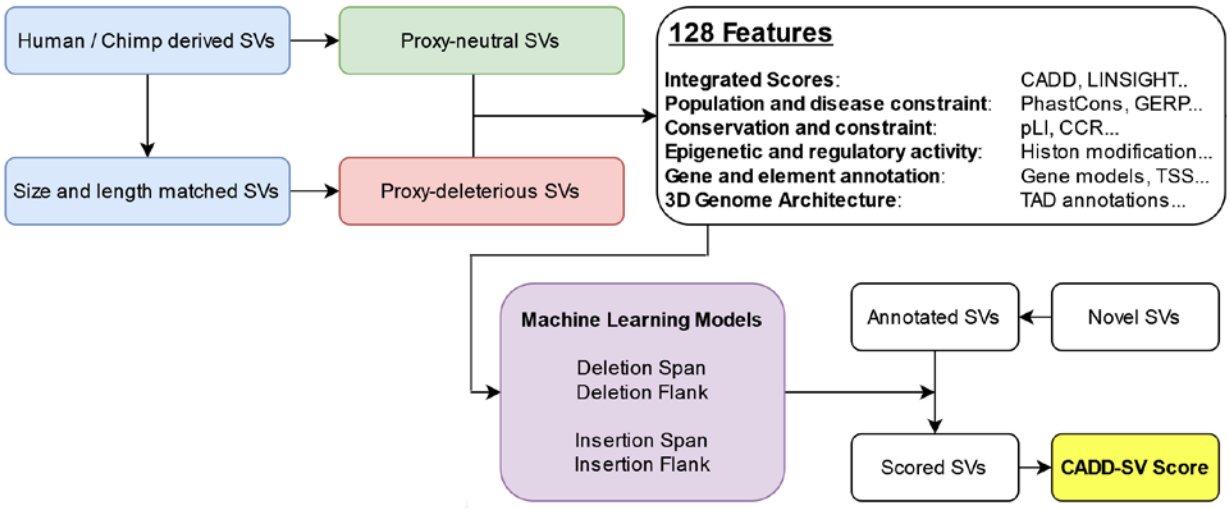


Figure 18: CADD-SV workflow. Evolutionary SVs (proxy-neutral) contrasted with Size and length matched simulated variants used as proxy deleterious training dataset. Next, various informative features are annotated and transformed (see Methods and Table

1) across span or flank of the variants to train multiple Random Forest classifiers. Models are used to score user provided novel SVs. For this purpose, variants are annotated, features transformed, and models applied. The maximum value of the flank and span model scores is used as the raw model score.

3.3.5 Logistic Regression compared to Random Forest Models

I trained both logistic regression models as well as random forest models. The latter show increased holdout performance as well as validation set performance (Figure 19) and I only describe the random forest models further.

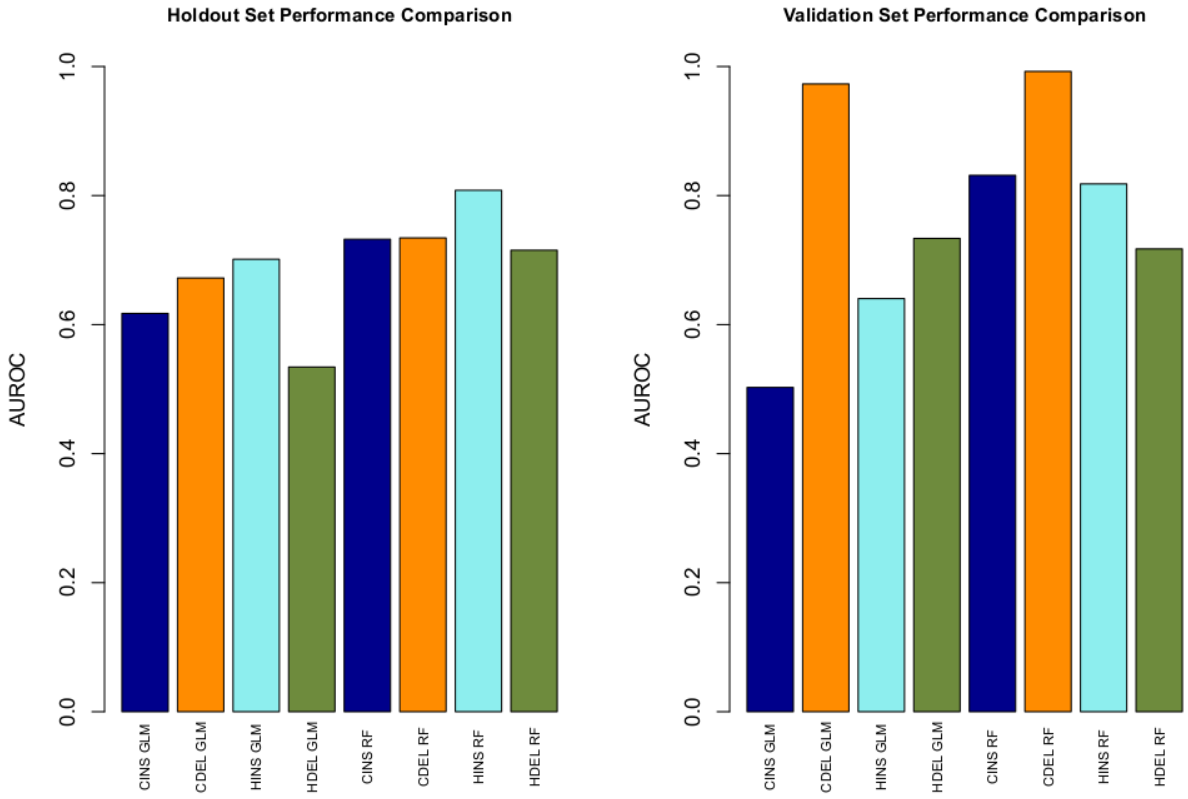


Figure 19: Model comparison of Random Forrest (RF) classifiers and generalized linear models (GLM) trained using the R GLM package. I validated the performance of both classifiers using 10% randomly sampled holdout data (left) as well as one of the validation sets (labelled pathogenic SVs from ClinVar vs. common SVs in gnomAD, right) CINS: Chimp Insertions (blue), CDEL: Chimp Deletions (orange), Human Insertions (turquoise), Human Deletions (green).

I opted for holdout performance validation over cross validation as the choice of training data allows for a sufficiently large training set. The holdout shows that all four model types differentiate between the proxy-benign and proxy-pathogenic sets (Fig 2A). Considering the anticipated mislabeling in our training data, specifically in the randomly drawn SVs as described above, the holdout performance will however not be representative for our models' performance in scoring actual pathogenic versus benign variants.

Here, I only look for a non-random model performance and the relative ranking of the INS, DEL and DUP models.

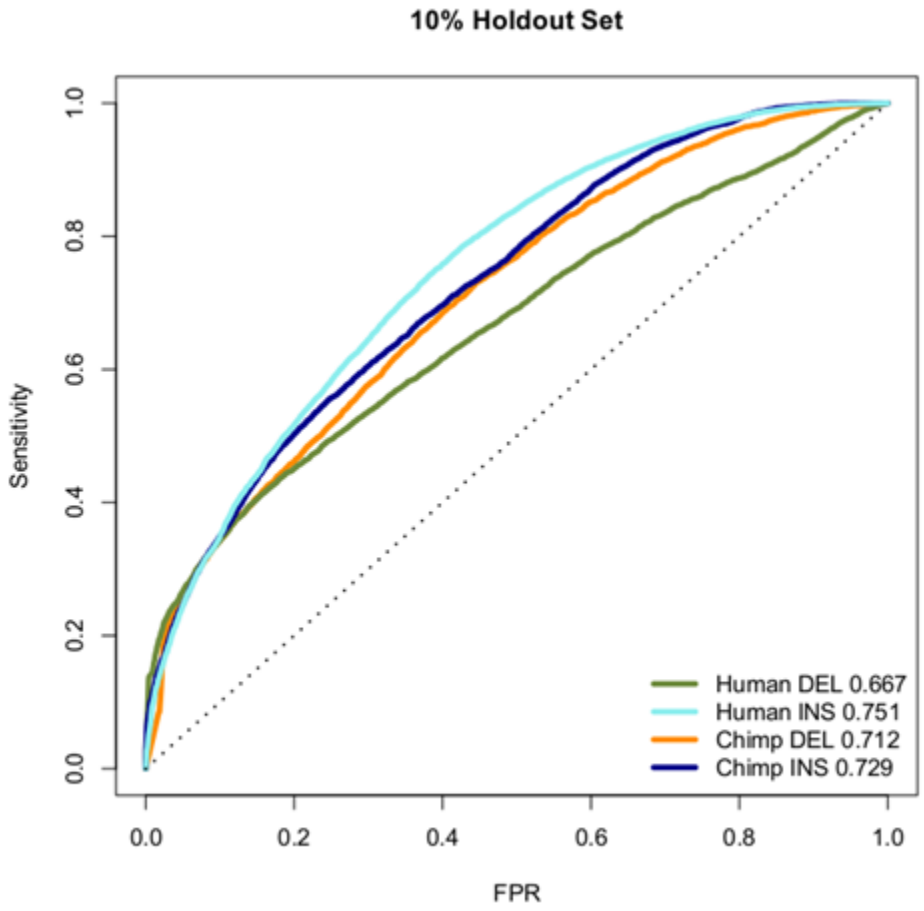


Figure 20: All four models show a non-random separation of the two classes in a random 10% holdout. Performance is measured as sensitivity over false positive rate (FPR). Note that all training datasets contain a high amount of mislabeled SVs, as a majority of proxy-deleterious SVs is likely to be neutral. Human Deletion Model shown in green, Human Insertion Model in turquoise, Chimp Deletion Model in orange, Chimp Insertion Model in blue.

3.3.5.1 Performance of Random Forest Models

The model score distribution for the holdout data as depicted in a ROC-Curve is available in Figure 20 for the proxy-pathogenic and proxy-benign SV sets. I see a significant shift (see Figure 21) with a bimodal distribution in the proxy-pathogenic variants, with the smaller mode corresponding to the potentially pathogenic variants in the randomly drawn set.

3.3.6 Phred Scoring

For better interpretation, I also provide a Phred-scaled transformation of the model score relative to a healthy population cohort, i.e. a $\log_{10}$ score derived from the proportion of variants with a greater or equal score in the genomAD-SV set. The CADD-SV scores on the Phred scale range from 0 (potentially benign) to 48 (potentially pathogenic), indicating the position of the novel variant within the gnomAD-SV score distribution. For example, a score above three corresponds to the top 50%, 10 corresponds to the top 10%, 20 to the top 1% and 30 to the top 0.1% of scores observed from gnomAD-SV.

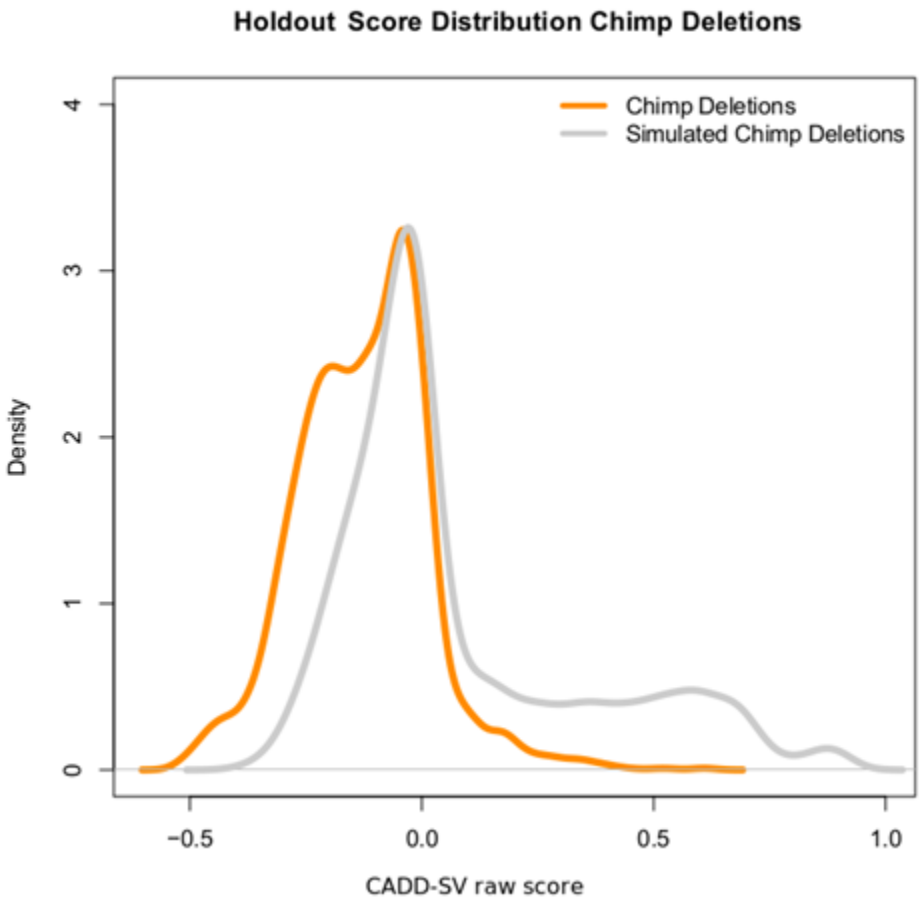


Figure 21: Model prediction scores of the chimpanzee deletion model are shifted towards high impact SVs in the simulated set of chimpanzee deletions (grey) in the hold-out dataset.

3.3.7 Feature importance

I analyzed feature contributions in our random forest models using the R package randomForest [127]. To ease interpretation, I categorized model features into six groups ("Integrated scores", "Species conservation and constraint", "Population and disease constraint", "Epigenetic and regulatory activity",

"3D genome organization", "Gene and element enrichment"; Table 1). Models benefit highly from features in the groups of "Species conservation and constraint" (incl. GERP, PhastCons, phyloP scores) and "Integrated scores" (i.e. summaries of CADD SNV and LINSIGHT scores) in differentiating between the contrasted SV sets. Regulatory annotations as well as 3D genome architecture features contribute to a smaller extent but are present within the top 20 most important features of all models (e.g. ReMap transcription factor occupancy, TAD annotations, enhancer-promotor links and ChromHMM states). Distance features (such as distance to coding sequence) are particularly prevalent in the human DEL flank model, where for a reference altered by the deletion event these features become informative. Major feature contributions for all models are available Figures 22-25.

Figure 22: Feature contributions of the human deletion (HDEL) flank model. Features are grouped and color coded respectively. Conservation scores (orange) are most informative to the HDEL model

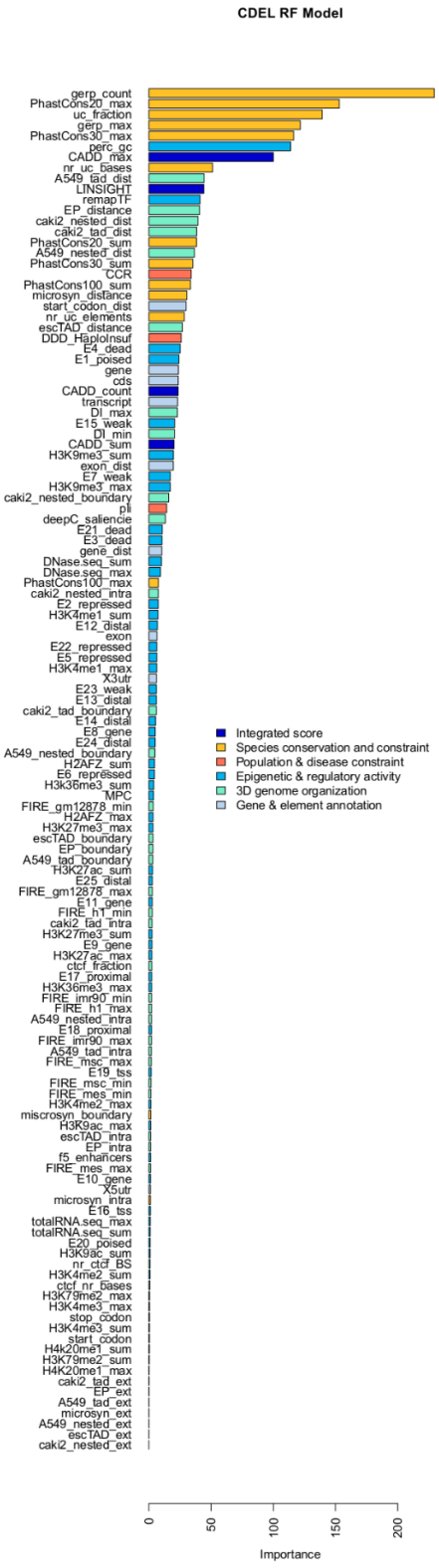


Figure 23: Feature contributions of the chimpanzee deletion (CDEL) span model. Features are grouped and color coded respectively. Conservation features are most informative to the CDEL model.

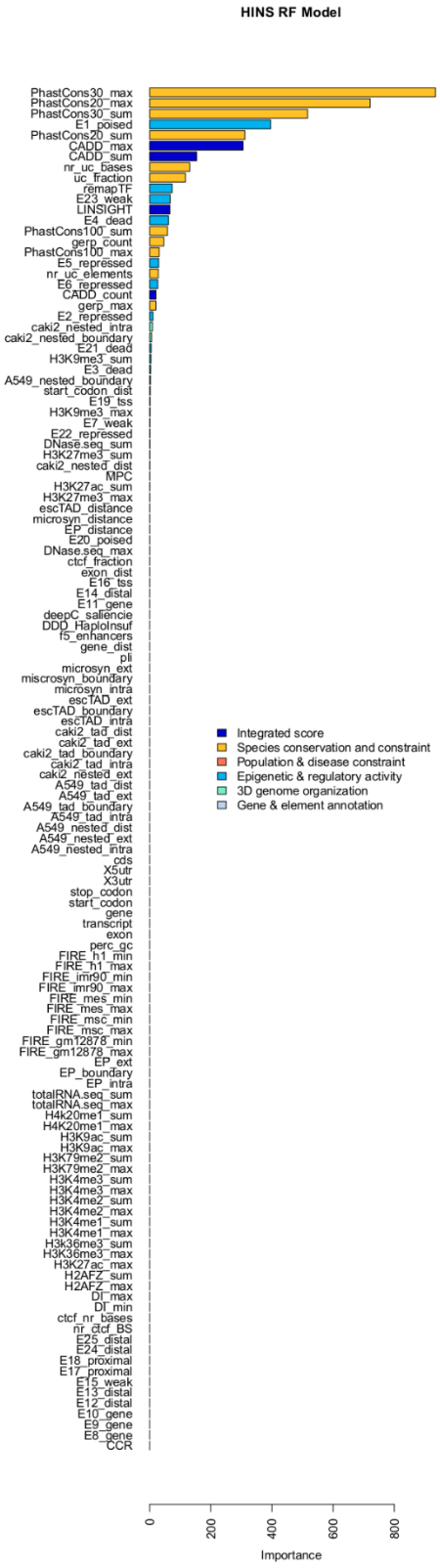


Figure 24: Feature contributions of the human insertion (HINS) flank model. Features are grouped and color coded respectively. PhastCons Score (conservation, yellow) are most informative to the model.

Figure 25: Feature contributions of the chimpanzee insertion (CINS) span model. Features are grouped and color coded respectively. Conservation scores (PhastCons, yellow) as well as Integrated scores (LINSIGHT and CADD, blue) are most informative to the CINS model

3.3.8 Validation set performance

To validate the general applicability of the framework, I use multiple lines of evidence (Figure 26A) to substantiate the results of the holdout performance. I look at known pathogenic variants from ClinVar (Figures 26B, 26D-F), I show that SVs occurring in healthy populations are under negative selection and therefore high CADD-SV scores enriched for singletons events (Figure 26C), I analyze variants from the International Cancer Genome Consortium (Figures 26D-F), and SVs affecting gene expression (Figures 26D-F). Thereby, I show that CADD-SV can be used to prioritize both pathogenic germline and somatic structural variants.

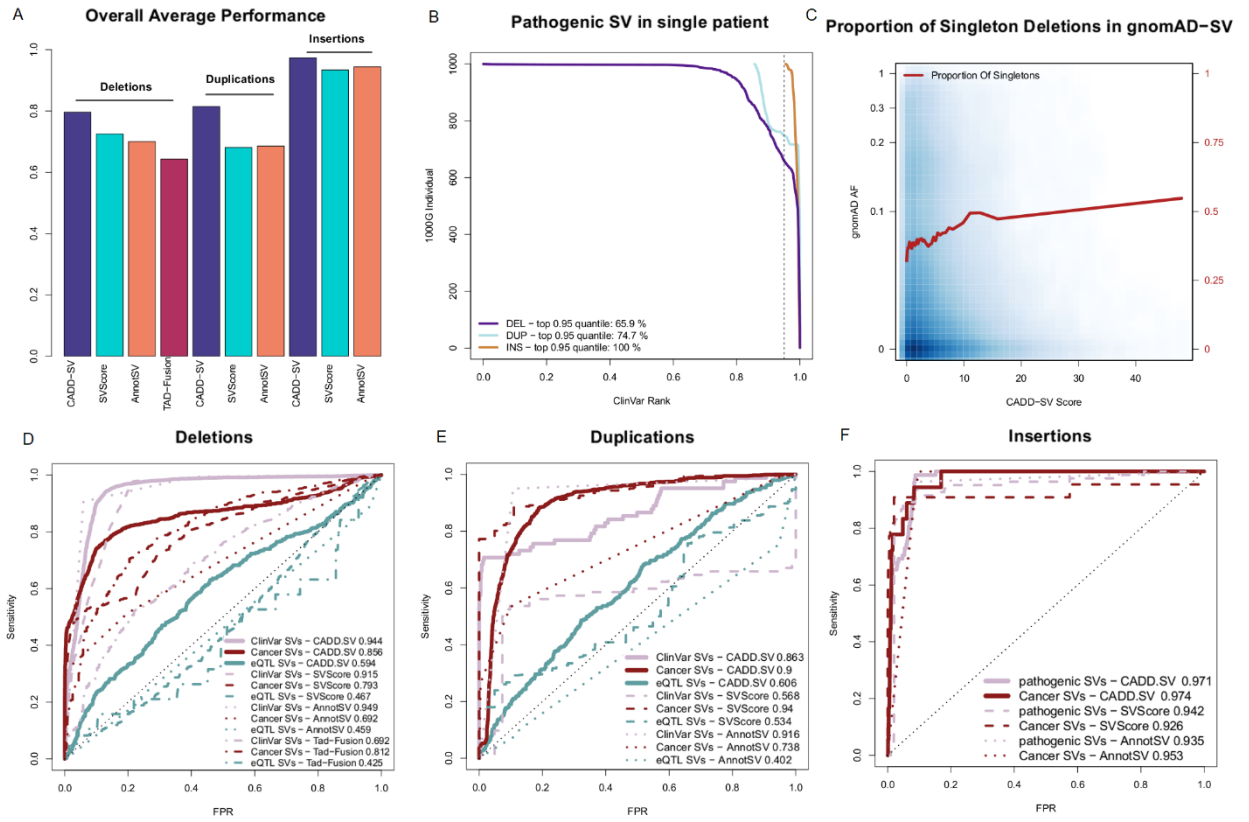


Figure 26: Validation set performance of the Random Forest models. ***A)*** *Summary of the performance of CADD-SV scores compared to SVScore, AnnotSV and TAD-Fusion scores across three validation sets (pathogenic variants, cancer variants and putative eQTL SVs) for deletions, duplications and insertions.* ***B)*** *Rank of ClinVar pathogenic SVs added to SVs of healthy individuals from the 1000 Genomes Project. CADD-SV prioritizes the pathogenic SVs over the other SVs in a single simulated patient, scoring pathogenic variants in the top fifth percentile of deletions, duplications and insertions for 65.9%, 74.7% and 100% of simulated variant sets, respectively.* ***C)*** *CADD-SV score distribution as a function of gnomAD allele frequency. Higher CADD-SV values represent an increased likelihood to be deleterious. In the deleterious tail of the score distribution, there is an excess of singletons (shown in red; bin size 0.025), which hints at negative selection against deleterious deletions.* ***D-F)*** *CADD-SV performance of various validation sets compared to common gnomAD SVs (AF ≥ 0.05). Performance is measured as sensitivity over false positive rate (FPR). CADD-SV is able to identify ClinVar pathogenic SVs (n=3262 deletions, 82 duplications and 78 insertions, pale red) as well as SVs reported in the ICGC cancer cohort (n=52,677 deletions, 42,972 duplications and 18 insertions, dark red) from common SVs*

*in gnomAD. Further, CADD-SV can identify non-coding SVs that are associated with differences in gene expression (turquoise). CADD-SV scores (solid lines) are compared to SVScore (dashed lines), AnnotSV (dotted lines) and TAD-Fusion (dashed and dotted lines) for deletions **(D)**, duplications **(E)** and insertions **(F)**.*

3.3.8.1 Clinvar

I collected pathogenic SVs from ClinVar (n=3262 deletions, 82 duplications and 78 insertions). To look at how CADD-SV prioritizes pathogenic variants among all SVs identified in single individuals (including rare and singleton events). I applied this dataset in two separate analysis steps.

3.3.8.1.1 Relative rank of ClinVar Variant within Individuum

I added each one clinically characterized SV from ClinVar into sets of structural variants found in presumed healthy individuals from the 1000 Genomes Project [128]. I assessed the performance of CADD-SV by looking at the pathogenic variants' rank among all observed SVs. I found that in 65% of cases the ClinVar deletion is within the top fifth percentile of all ranks (Figure 26B). Clinically labelled insertions and duplications were even more enriched among the top candidates. In 100% of individuals for insertions and 75% of individuals for duplications do these events fall within the top fifth percentiles of scores.

3.3.8.1.2 Prioritization of clinical variants in cohort

Further, I contrasted the complete sets of pathogenic SVs from ClinVar with a matched number of common SVs from gnomAD (AF ≥ 0.05, Figures 26D-F). CADD-SV correctly identifies a vast majority of the known pathogenic SVs with an Area Under the ROC Curve (AUROC) of 0.944 for deletions (Fig 3D). CADD-SV performs comparable to the existing tools SVScore [119] with an AUROC of 0.915 and AnnotSV [120] with an AUROC of 0.949. It outperforms TAD-Fusion score [122], which has an AUROC of 0.692, but was primarily designed to detect 3D-architecture alterations. Finally, I compared to StrVCTVRE [121], which was designed to score exonic variants specifically, and cannot score all of these variants. However, CADD-SV outperforms StrVCTVRE on prioritizing exonic ClinVar deletions from a background of exonic gnomAD-SV deletions (Figure 27).

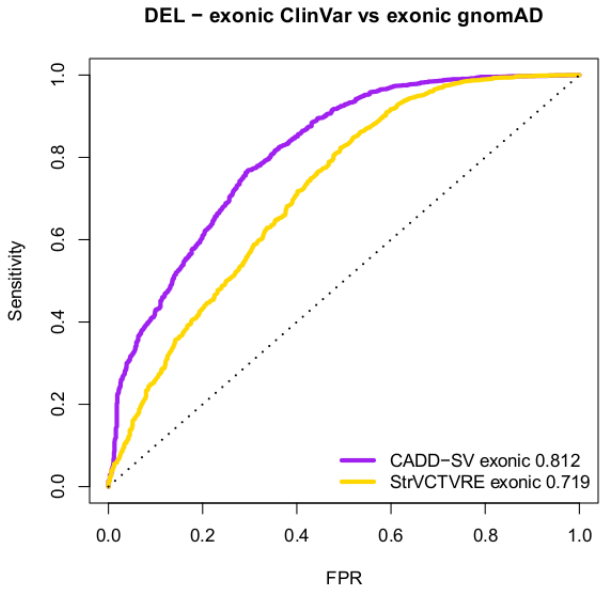


Figure 27: CADD-SV (purple) and StrVCTVRE (yellow) performance compared on ClinVar exonic sequences vs gnomAD-SV common exonic SVs. Shown is a ROC-Curve for the exonic SVs only. CADD-SV slightly outperforms StrCTVRE in correctly identifying pathogenicly-labeled SVs (n=3183).

3.3.8.2 Healthy population cohort

Variants reported in the gnomAD-SV database are considered largely benign as this cohort consists of healthy individuals, not excluding potential complex or late-onset diseases [19]. While being devoid of embryonal lethal variants, healthy datasets can contain pathogenic or haploinsufficiency variants that are expected to be under purifying selection and are therefore rare in allele frequency

3.3.8.2.1 Purifying selection in CADD-SV pathogenic tail

I assessed the distribution of CADD-SV scores in SVs from the gnomAD SV call-set. Allele frequency (AF) values are significantly decreased in the pathogenic tail of the CADD-SV score distribution compared to the benign tail (top/bottom fifth percentile CADD-SV scores, two-sided Wilcoxon rank sum test, p-value < 10^{-16}). I reason that CADD-SV is able to prioritize deleterious variants in healthy individuals as these variants would be under negative selection and removed from the gene pool. Accordingly, the proportion of singleton deletions amongst the top fifth percentile CADD-SV scores (pathogenic tail) is 1.3 times higher than the average of the full SV set (Figure 26C). This observation is striking for deletions but less

pronounced in the insertion and duplication SV sets (Figure 28). I note that in the top fifth percentile, 35% of deletions are coding variants classified as "Loss of Function" by the gnomAD consortium[19] compared to 0.3 % of variants scored in the remainder of the CADD-SV score distribution.

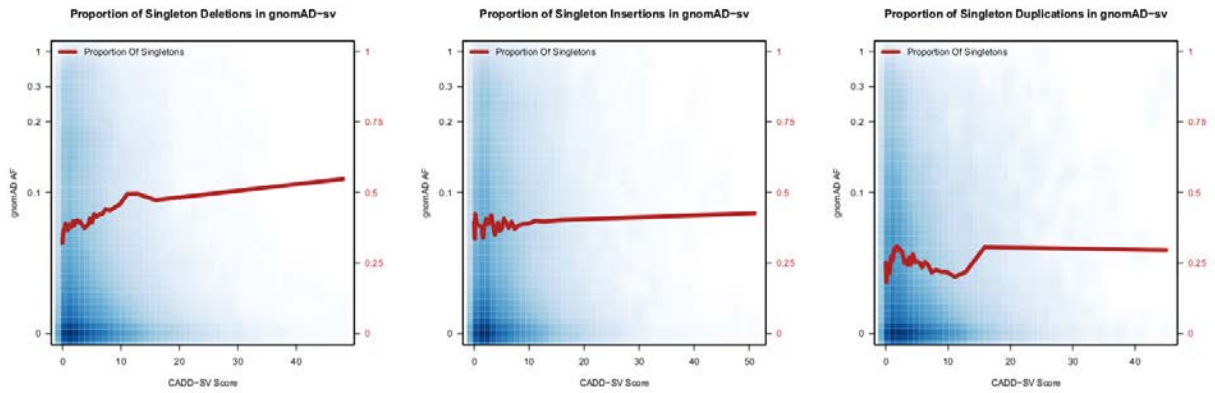


Figure 28: Proportion of singleton insertions and duplications in the gnomAD-SV data set of putative healthy individuals. The pathogenic CADD-SV score tail (≥20, being the top 1% most pathogenic) is enriched in singletons, suggesting purifying selection against SVs with high CADD-SV scores. However, this effect is less pronounced in insertions and duplications compared to the deletions.

3.3.8.2.2 Length effect

Further, the average deletion length of gnomAD-SV variants in the top 5% of scores is six times longer compared to the rest of the distribution, suggesting that longer deletions are more likely to be functional as they affect more sequence. However, short (less than 100bp) and high scoring (top fifth percentile) deletions are 1.1 times more likely to be singletons compared to short deletions, suggesting that CADD-SV prioritizes SVs beyond length. In addition, I detect high frequency deleterious variants in the pathogenic tail, speculating that these variants could be phenotypically functional variants and potentially beneficial for carriers.

3.3.8.2.3 GWAS associations in pathogenic tail

Table 2: Top 10 CADD-SV scores in gnomAD-SV with number of overlapping GWAS SNVs. Chr, start, end describe the genomic location. Phred scaled CADD-SV Score in "score", raw and span score of the model as well as number of overlapping GWAS SNVs labelled.

chr	start	end	Score	raw score	raw span	raw flank	GWAS
13	25165903	25531417	45	0.994923027	0.994923	0.028845	8
13	50567002	51538041	45	0.994706653	0.994707	0.169125	105
17	30657307	33838598	48	0.995145249	0.995145	-0.08501	291
17	34812000	36251000	44	0.994688107	0.994688	-0.19966	262
2	85889000	87236000	44	0.994688107	0.994688	-0.01483	80

2	119820000	123792000	45	0.994706653	0.994707	0.070122	247
2	130754220	133138212	48	0.995145249	0.995145	-0.01371	30
20	8105854	23795733	48	0.995145249	0.995145	0.004245	1223
5	159698170	166152496	44	0.994688107	0.994688	-0.12854	262
8	82680857	87501720	48	0.995145249	0.995145	-0.00586	146

I showed that rare variants are strongly enriched in the most pathogenic tail of the CADD-SV distribution (Figure 26C). Further, the tail of the CADD-SV pathogenic score distribution is strongly enriched in SVs containing GWAS identified SNVs, suggesting the presence of functional genomic regions (Figure 29). Containing a GWAS hit is not equal to being a potentially pathogenic SV, as many recorded associations are towards non-disease traits such as body height or longevity. However, it provides evidence that CADD-SV is able to prioritize functional stretches of sequence in the genome without using the GWAS catalog as an input itself. The top ten gnomAD-SV variants contain an average of 265 GWAS associated SNVs (Table 2). I further investigated the shortest (mean length of 225,336 bp) five top scoring variants (CADD-SV Phred score ≥ 35) and found all of them to be ultra-rare (AF ≤ 0.0009), with three out of five being singletons (Table 3).

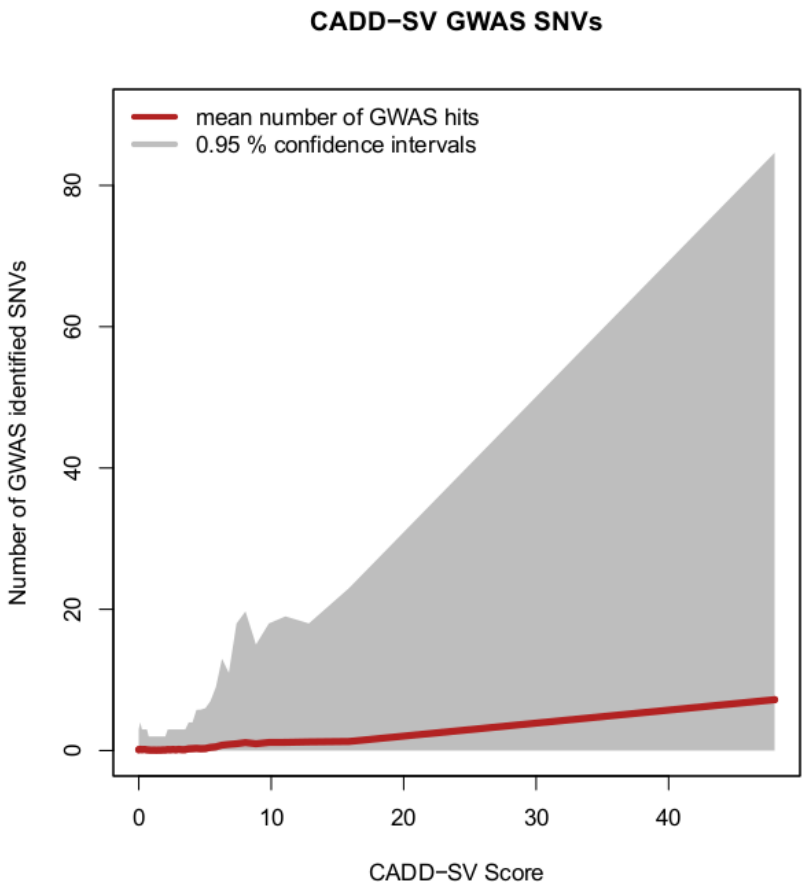


Figure 29: CADD-SV score distribution as a function of number of GWAS identified SNVs per deletion from gnomAD-SV. The average number of GWAS associated SNVs increases drastically especially among high scoring SVs, suggesting an enrichment of functional variants in the pathogenic tail of the CADD-SV score distribution. Note that CADD-SV is a Phred-scaled score distribution (log_{10} scale) with high values corresponding to high pathogenicity.

Table 3: CADD-SV score outliers from gnomAD-SV (length < 200kb) with GWAS overlap, ClinVar SNV overlap and allele frequency

Chr	start	end	score	GWAS	ClinVar	GnomAD-SV
10	96394432	96633999	36	9	none	1
16	21594700	21748000	35	8	pathogenic	20
16	28353000	28610100	35	96	pathogenic	2
4	73004055	73231324	35	18	pathogenic	1
7	106922194	107171638	35	13	none	1

3.3.8.2.4 ClinVar associations in pathogenic tail

Further, three out of five variants overlap multiple ClinVar curated pathogenic variants, belonging to two autosomal recessive disease genes and one autosomal dominant disease gene.

3.3.8.2.5 Recessive pathogenic variants

The two recessive diseases are Batten disease mediated by mutations in CLN3 [129], see Figure 30, and hearing loss mediated by mutations in OTOA [130], see Figure 31.

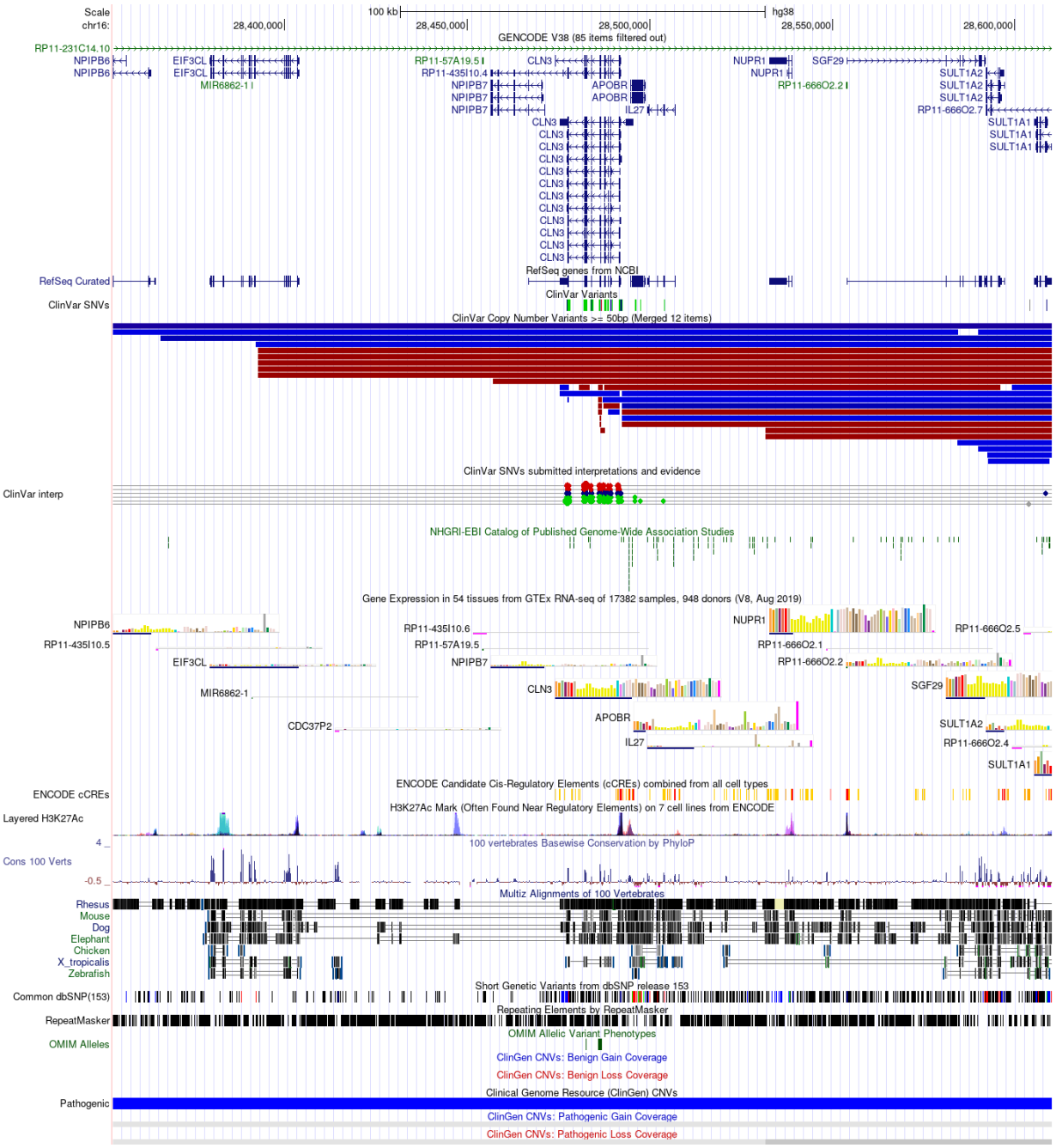


Figure 30: UCSC Genome Browser tracks of a region (chr16:28353000-28610100) deleted in two individuals present in the gnomAD-SV cohort. Various genes are affected, with CLN3 being identified as causing Batten disease, a fatal disease of the nervous system. Various positions of this SV are highly conserved among 100 Vertebrate Genomes, giving CADD-SV power to detect this SV with a high score. This SV is not a singleton, suggesting a recessive disorder. In some cases, Batten disease can also have a late onset of disease symptoms, potentially explaining the presence of this SV in a healthy cohort.

Figure 31: UCSC Genome Browser tracks of a region (chr16:21594700-21748000) deleted in 20 individuals present in the gnomAD-SV cohort. Two genes are affected, with OTOA being identified as autosomal recessive disease causing severe hearing loss, when rendered dysfunctional by ClinVar annotated SNVs within the OTOA gene body. Further, various positions of this SV are highly conserved among 100 Vertebrate Genomes, giving CADD-SV power to detect this SV with a high score. Unlike other putative pathogenic SVs, this SV is not a singleton, suggesting a recessive disorder or reduced purifying selection on phenotypes such as hearing loss

3.3.8.2.6 Autosomal dominant pathogenic variant

The one autosomal dominant neurodevelopmental disorder is Chopra-Amiel-Gordon syndrome, mediated by mutations in ANKRD17 [131], see Figure 32.

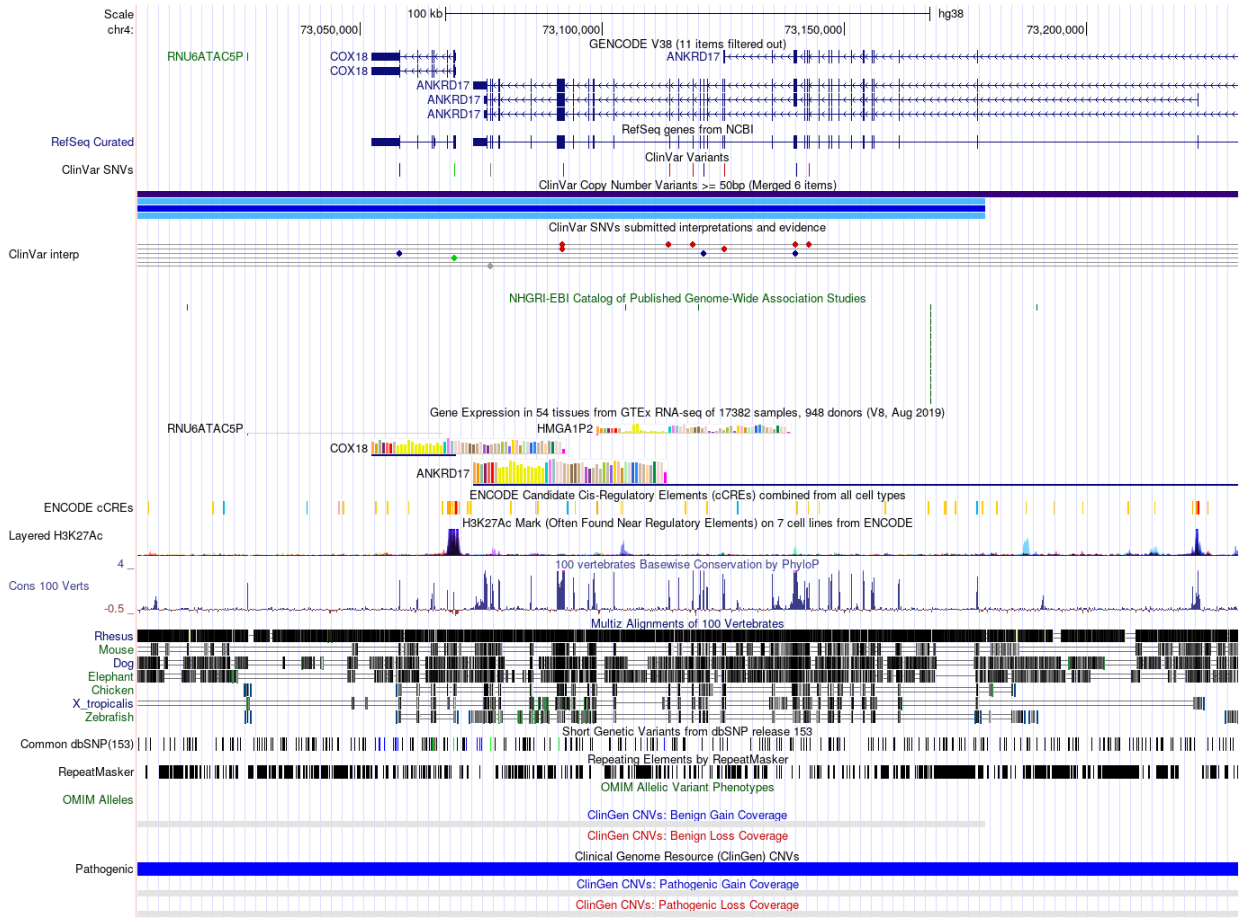


Figure 32: UCSC Genome Browser tracks of a region (chr4:73004055-73231324) deleted in one individual present in the gnomAD-SV cohort. Two genes are affected, with ANKRD17 being identified as autosomal dominant disease-causing Chopra-Amiel-Gordon syndrome (CAGS) with various pathogenic ClinVar SNVs being annotated within the gene body of this gene. CAGS patients are characterized by developmental delay and intellectual disability ranging in severity from moderate to severe. Various positions of this SV are highly conserved among 100 Vertebrate Genomes, giving CADD-SV power to detect this SV with a high score.

3.3.8.2.7 Additional Healthy Cohort Dataset

Further, CADD-SV is able to prioritize an additional set of SVs [115] associated with expression changes (see Figure 33A and 33C) as well as SVs under natural selection (see Figures 33B and 33D), with most Phred scores exceeding a value of 10 (top 10%) and many above 20 (top 1%) or even 30 (top 0.1%). This all supports that CADD-SV is able to prioritize functional stretches of DNA genome-wide and beyond exonic regions.

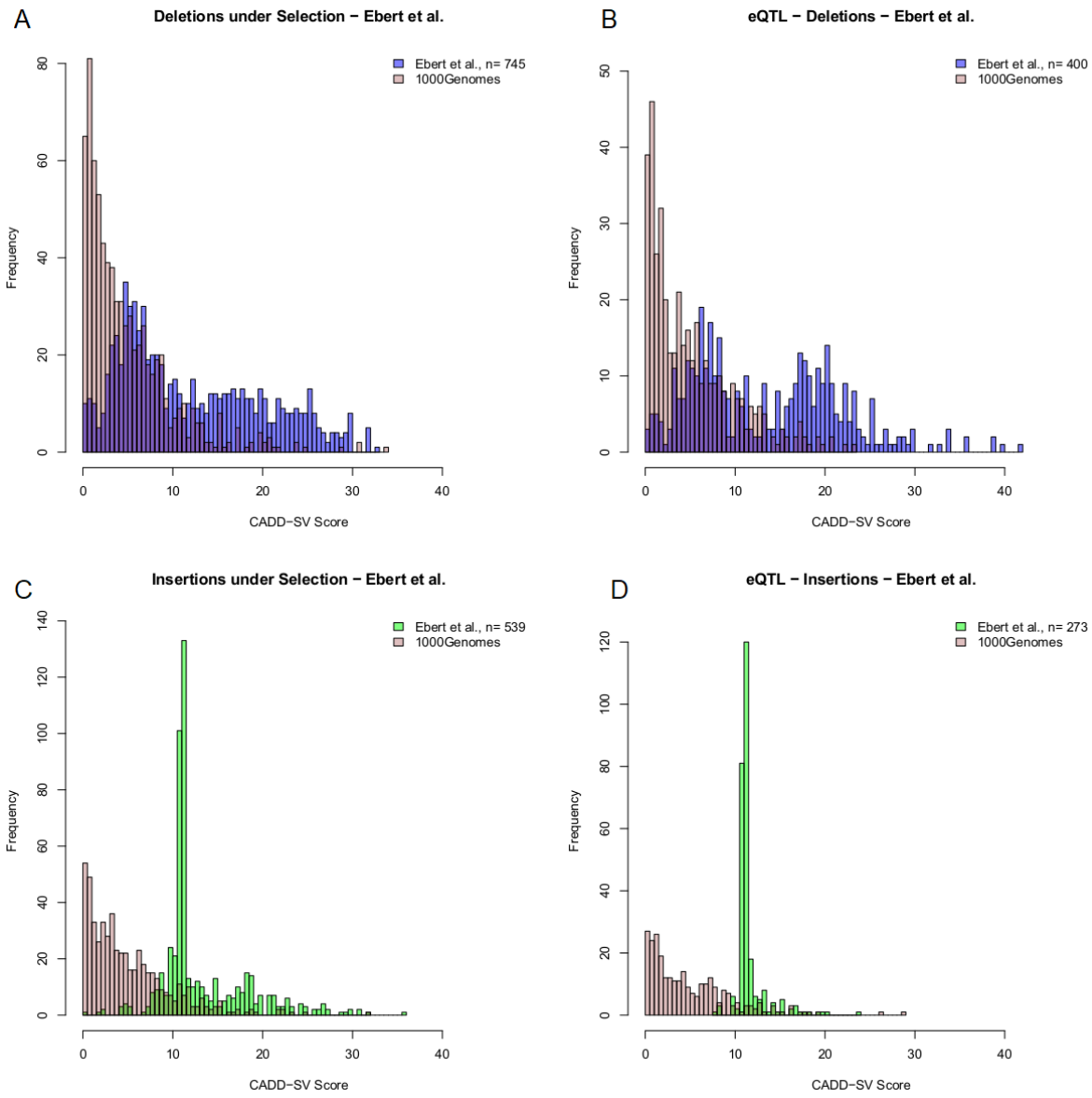


Figure 33: Functional deletion and insertion SVs annotated in Ebert et al. 2021. CADD-SV prioritizes both SVs under natural selection as well as expression associated SVs in this data set. Shown are score distributions for the functional set (deletions in blue, insertions in green) against the same number of randomly drawn SVs from 1000Genomes. Note that CADD-SV is a Phred scaled score distribution with high values corresponding to high pathogenicity.

3.3.8.3 SVeQTLs

To test the ability to prioritize functional variants beyond coding regions, I use a set of non-coding SVs known to alter the expression of genes. Here, I look at 387 deletions and 300 duplications that were shown to affect expression levels of nearby genes and are therefore considered eQTLs by the GTEx consortium [118]. I compare them against common variants (AF ≥ 0.05) from gnomAD in a ROC curve analysis (Figures

26D-F). Even though less pronounced compared to ClinVar or the cancer-derived SVs, CADD-SV is able to differentiate the two classes of SVs (AUROC 0.598 for deletions and 0.635 for duplications, respectively) outperforming existing methods SVScore (AUROC 0.467 for deletions and 0.534 for duplications), AnnotSV (AUROC 0.459 for deletions and 0.402 for duplications) and TAD-Fusion score (AUROC 0.425 for deletions).

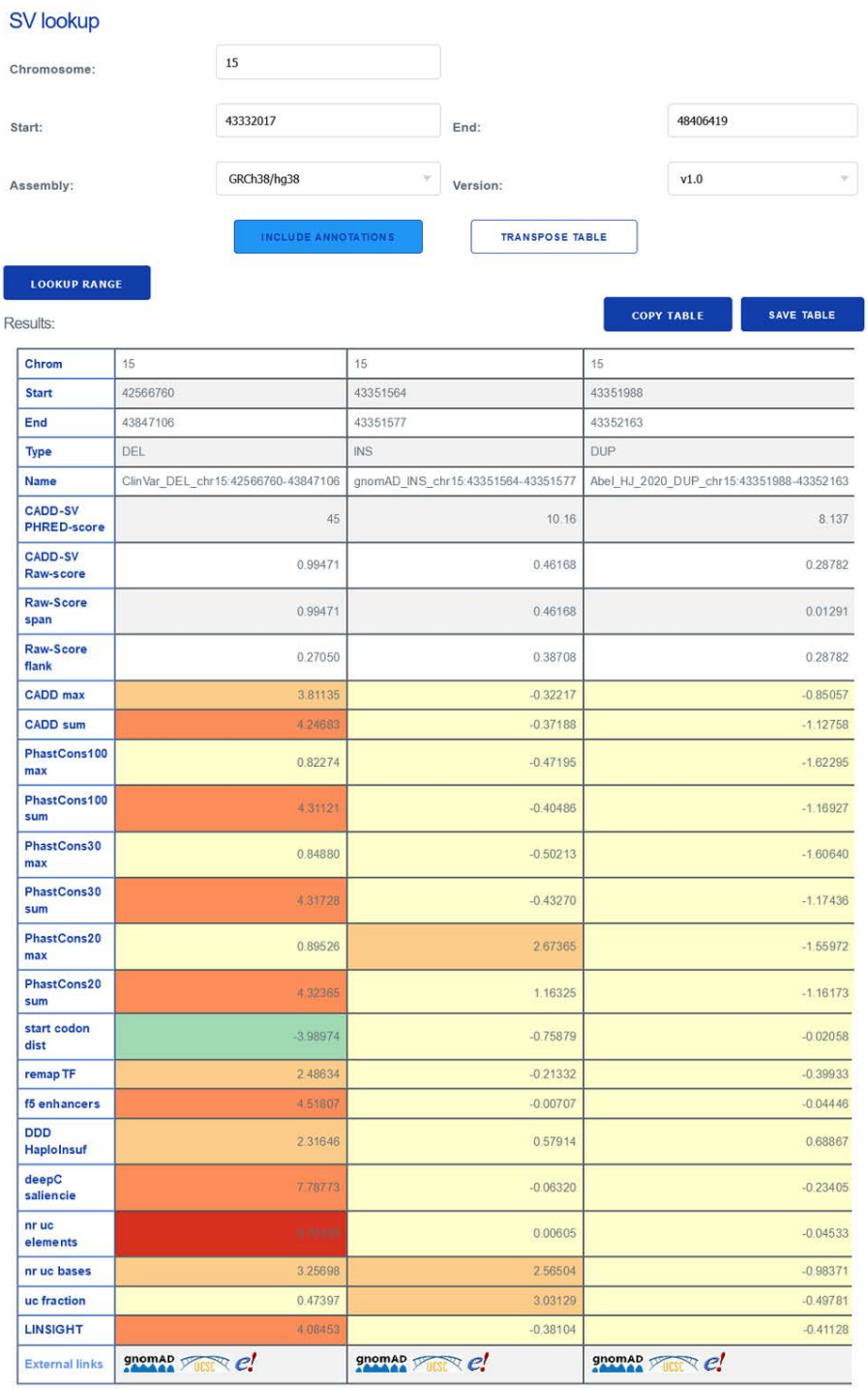
3.3.8.4 Somatic Cancer variants

I assessed the performance of CADD-SV on somatic variants and the power to identify deleterious cancerogenous variants (n=52,677 deletions, 42,972 duplications and 18 insertions) using SV variants from cancer patients in the International Cancer Genome Consortium [116] as well as insertions reported in Qian et al. [117]. I find an enrichment of SVs detected in cancer patients in the pathogenic tail of the distribution compared to SVs from a healthy cohort (two-sided Wilcoxon rank sum test, p-value $<10^{-16}$). CADD-SV enriches the cancer-derived SVs from common gnomAD-SVs in a ROC Curve analysis (Figures 26D-F, AUROC values of 0.848, 0.933, and 0.975 for deletions, duplications, and insertions, respectively), outperforming available tools on this task and supporting the claim that CADD-SV prioritizes functional somatic SVs.

3.3.9 Implementation and Interpretation

3.3.9.1 Phred Scaling and Z-score transformation

To make scores easier to interpret and as outlined above, I Phred-scale CADD-SV raw scores among all SVs reported in gnomAD-SV. For example, a value of 30 represents that 99.9% of variants reported from healthy individuals are scoring lower than the variant under consideration. In addition, all feature annotations are used and reported after *Z*-score transformation according to the features' value distribution observed for gnomAD-SV variants. This allows users to inspect the individual features for extreme values easily (additionally color coded for visual inspection on the webserver). For instance, a conservation feature value of four represents an outlier value of four standard deviations away from the gnomAD mean of that specific annotation.



Chrom	15	15	15
Start	42566760	43351564	43351988
End	43847106	43351577	43352163
Type	DEL	INS	DUP
Name	ClinVar_DEL_chr15:42566760-43847106	gnomAD_INS_chr15:43351564-43351577	Abel_HJ_2020_DUP_chr15:43351988-43352163
CADD-SV PHRED-score	45	10.16	8.137
CADD-SV Raw-score	0.99471	0.46168	0.28782
Raw-Score span	0.99471	0.46168	0.01291
Raw-Score flank	0.27050	0.38708	0.28782
CADD max	3.81135	-0.32217	-0.85057
CADD sum	4.24683	-0.37188	-1.12758
PhastCons100 max	0.82274	-0.47195	-1.62295
PhastCons100 sum	4.31121	-0.40486	-1.16927
PhastCons30 max	0.84880	-0.50213	-1.60640
PhastCons30 sum	4.31728	-0.43270	-1.17436
PhastCons20 max	0.89526	2.67365	-1.55972
PhastCons20 sum	4.32365	1.16325	-1.16173
start codon dist	-3.98974	-0.75879	-0.02058
remap TF	2.48634	-0.21332	-0.39933
f5 enhancers	4.51807	-0.00707	-0.04446
DDD HaploInsuf	2.31646	0.57914	0.68867
deepC saliencie	7.78773	-0.06320	-0.23405
nr uc elements	3.70289	0.00605	-0.04533
nr uc bases	3.25698	2.56504	-0.98371
uc fraction	0.47397	3.03129	-0.49781
LINSIGHT	4.08453	-0.38104	-0.41128
External links	gnomAD UCSC e!	gnomAD UCSC e!	gnomAD UCSC e!

Figure 34: The CADD-SV webserver can score custom SV sets, but it can also be used for direct lookup of pre-scored deletions, duplications and insertions from gnomAD, ClinVar, as well as call-sets from Abel et al. [132] and Beyter et al. [133]. For a given SV, the website provides the combined model scores as well as annotation values normalized to the range in the healthy gnomAD cohort (Z-score). This enables users to identify interesting variants from color-highlighted extreme feature values and not just by the combined CADD-SV score. Further, the website provides direct links for each SV to external resources like gnomAD, Ensembl or the UCSC Genome Browser.

3.3.9.2 CADD-SV webserver

Such noticeable values are highlighted by color-coding on the CADD-SV website (Figure 34) for the pre-scored variant sets. Generally, CADD-SV scores with or without annotation information are available from our command line tool as well as on the webserver for direct variant interpretation. Our online services include region lookups of existing SV datasets, coordinate transfers between human genome builds, the download of pre-scored datasets and annotations, a simple API for the retrieval of pre-scored variants as well as the online scoring of novel SV datasets. Coordinate ranges and variants of other genome builds (i.e. GRCh37/hg19 and NCBI36/hg18) can be used on the webserver and are automatically lifted to GRCh38 coordinates (providing the original coordinates in the variant's name field).

3.3.9.3 CADD-SV command line tool

The CADD-SV framework can be cloned and used from GitHub (https://github.com/kircherlab/CADD-SV/) and is available as a Supplemental Code file. All external data sets used are publicly available under the locations specified in the Methods.

3.4 Discussion

I present the CADD-SV framework, an unbiased and powerful tool for the annotation and prioritization of deleterious structural variants [4].

3.4.1 Biases

Various biases can affect the outcome of SV interpretation. In the following section I discuss putative biases to the CADD-SV framework and SV detection. SV-calling algorithms affect the detection of SVs and therefore their correct interpretation, ClinVar validation sets are under ascertainment biases as, among other biases, well studied genes are overrepresented.

3.4.1.1 SV calling

Structural variant calling is prone to biases towards certain types of SVs, as for example the signal to detect deletions is vastly different compared to signals of duplication or inversions [134]. Further, the exact annotation of SV breakpoints is often limited, e.g. due to their frequent positioning in repetitive sequence [76]. Apart from these universal limitations, changes in the application of arrays and sequencing technologies over the last decades have affected available SV sets.

3.4.1.2 Ascertainment biases

In previous works it seems underappreciated how much the historic and functional ascertainment imprinted on potential training and validation sets affects machine learning. Specifically, the ClinVar-annotated SVs are comparably large and clustered around well-studied genes. Using an alternative source for the training data, the CADD-SV approach is not confounded, and performance can be evaluated broadly, as no allele frequency features nor ClinVar annotations are included in the features or otherwise considered when building the training sets. The number of labelled SVs to validate the performance of CADD-SV is still limited though. Assessing the performance on duplications and insertions is limited though, as the number of known pathogenic events is small and strongly biased towards coding sequence. I anticipate that future datasets will provide a better opportunity to test and interpret models for duplications and insertions.

3.4.2 Mechanistic diversity

Estimating functional effects of SVs is highly complex due their size (involving different molecular targets) but also due to different mechanistic types of SVs (e.g. deletion, insertion, duplication or inversion of sequence). Thus, deleteriousness effects cannot just result from the sequence alteration, but also from interactions with the sequence context. For example, sequences shielding gene regulation (e.g. TAD boundaries) can be deleted between coding sequences or non-functional sequence can be inserted, interfering with an existing regulatory unit. Therefore, I model each SV type (deletions, insertions and duplications) separately, and I use the sequence span as well as the flanking sequence regions to capture putative pathogenic effects comprehensively. Further, I integrate distance features and a large set of annotations covering both coding and non-coding effects. This allows CADD-SV high predictive performance on known disease variants from ClinVar, which often cover coding sequence and stand-out by their gene model annotations and gene scores such as pLI [94] or Deciphering Developmental Disorders' Haploinsufficiency score[96]. Extending this to other previously described disease mechanisms for pathogenic non-coding variants [20], CADD-SV makes use of sequence conservation [37], enhancer element annotations [92,135] and enhancer links [103], molecular assay readouts such as RNase-seq or ChIP-seq, as well as information about 3D interactions from the Hi-C directionality index [102,104] or computational predictions such as deepC [125].

3.4.3 Limitations

CADD-SV aided SV interpretation is limited in various ways. Mechanistically SV occurrence and interpretation is very diverse as multiple types of SVs exist. Further, annotated datasets are limited by their experimental layouts as well as their human influence.

3.4.3.1 SV types

Inversions and translocations are particularly hard to assess as they are copy number neutral and their impact is often mediated by proximity of certain functional elements to one another or functional entities such as TADs being broken or reshuffled rather than deleting or inserting functional sequence directly. To our knowledge, there is no training dataset sufficient in size and curation to capture the complexity of these events. Therefore they are currently not part of the CADD-SV framework and can not be scored. As no single model could capture the mechanistic diversity of the three currently considered SV types

(insertions, deletions, and duplications), CADD-SV reports normalized model scores and features through relative ranks as well as *Z*-scores (i.e. values reported as standard deviation away from the mean) based on SVs from a large cohort of healthy individuals.

3.4.3.2 Aided Interpretation

Mechanism of deleteriousness can vary widely within and between SV types as SVs often influence big portions of the genome. To summarize this diversity in a single score can mislead as the same score within and between SVs can feature widely different modes of pathogenicity. Therefore, Phred-scaled model scores aid by providing an intuitive interpretation and feature normalization enables users to inspect extreme values for the different annotations directly, visually highlighting certain annotations and hinting at potential pathogenic mechanisms beyond the final CADD-SV score.

3.4.3.3 Labelled datasets used for validation

Especially for rare variants, clinical databases like ClinVar or OMIM have incomplete coverage. CADD-SV does not use these databases to derive features as I do not want it to be intrinsically limited to previously known disease genes or to reflect the historic ascertainment that imprints on these databases [136,137]. I recognize that computationally distinguishing functional variants from pathogenic variants is difficult and that available curated data sources like ClinVar and OMIM can still be used in downstream interpretation of the results. Evaluating SVs experimentally will provide insights into disease mechanisms that are currently not represented.

3.4.4 SV size

In contrast to other tools, length is not a feature of CADD-SV. However, I assume that SV length would be a good indicator of SV impact, as long SVs are more likely to affect coding regions or generally functional annotations. SV length itself might be a confounder too, as long benign SVs might be misinterpreted solely for their length and not for their actual genomic signatures. As the contrasting datasets used in training the CADD-SV framework are matched in SV length, length as a feature does not contribute to the model. However, some genomic feature transformation such as the sum of all intersected annotation values or the number of bases above a certain threshold, correlate inevitably with length but are bound to functional annotations being present across the span. With increased length there is an increased (random) chance to overlap annotations.

3.4.5 Existing tools

AnnotSV [120] is powerful and efficient in annotating novel SVs with a wide set of annotations. However, validation of AnnotSV on ClinVar is biased as AnnotSV uses overlap of novel SVs with labelled SVs from ClinVar as a feature. Further, it categorizes SVs in five bins from benign to pathogenic instead of a continuous score. CADD-SV is powerful in detecting functional SVs. Across multiple data sets, I highlight the increased predictive power of CADD-SV compared to AnnotSV, SVscore [119] and TAD-Fusion [122]. I could only provide a limited comparison to StrVCTVRE [121], which is designed to score only exonic variants. A comparison of SVFX [138] was not possible, as the package is not easily deployed and explicitly normalizes features on a specific training data set. Its released ClinVar variant models are based on a model trained on the same variant set used in our validation assessment.

3.4.6 Outlook

The feature integration implemented by CADD-SV can easily be extended using additional annotations. Currently, I use features derived from experiments conducted in specific cell-types (e.g. GM12878, H1, A549, CAKI2). More comprehensive or additional cell-types can be included in updated versions. Further, CADD-SV does not make us of the inserted sequence itself. Therefore, future versions of CADD-SV could make use of sequence-based prediction models in addition to reference annotations, e.g. to predict open reading frames, repeat content, presence of transcription factor binding sites or the likelihood of the novel inserted sequence being of open or closed chromatin. This might be powerful in assessing inserted sequence function beyond the surrounding genomic context of the insertion event. In addition, specific mechanistic events such as gene-fusion predictions are not part of our features. CADD-SV can only estimate the effect of such events based on already considered feature values like the distance to genes.

3.4.7 Conclusion

In summary, CADD-SV integrates rich sets of annotations in predictive models of SV effects. CADD-SV is built from machine learning models with an unbiased training using evolutionary-derived and putative benign variants that underwent millions of years of purifying selection. These variants are contrasted with a background set of the same size and length, encountering deleterious events by chance. I show that our approach is able to model and score deletions, insertions as well as duplications. CADD-SV is available as open-source resource as well as webservice application for direct online scoring. CADD-SV provides a

userfriendly interpretation of obtained annotations by normalizing all feature values using the gnomAD putative healthy population cohort. I validate the CADD-SV models using clinically annotated, non-coding or population germline SVs as well as somatic SVs reported in cancer patients. To highlight the potential of CADD-SV, I applied our tool to functional SVs identified from selection screens, QTL studies or variants identified across many, supposedly healthy individuals. Most of the top-scored variants in the healthy population dataset are singletons, suggesting purifying selection on these SVs, and I was able to pinpoint pathogenic variants in multiple cases.

4 CTCF Evolution

4.1 Introduction

4.1.1 Mechanistic focused variant interpretation

Variant interpretation plays a large role in personalized medicine. In this section I will focus on mechanistic insights of variant interpretation and how variants can be assessed depending on their function. To do so, I look at a transcription factor, CCCTC-binding factor (CTCF), that is mainly known for its role in 3D genome architecture mediation [139].

4.1.2 Role of CTCF

CTCF is a transcription factor that is involved in regulation of gene expression, though it is best known for its contribution to regulate genome architecture. CTCF was originally discovered as a negative transcriptional regulator of the *c-myc* gene in chicken [140]. Its binding motif was described as three spaced repeats of CCCTC. Upon binding to this motif CTCF mediates the formation of DNA loops. Further CTCF can anchor DNA to laminar structures [141]. CTCF forms homodimers, which, colocalized with cohesin, a ring-structured Protein, loops the DNA in the form of a "loop extrusion" mechanism [142]. Cohesin functions as a circular structure around one or two double stranded DNA regions halted by bound CTCF (see Figure 35). CTCF binding can be inhibited by CpG methylation of the core motif CCGCGNGGNGGCAG [99]. CTCF therefore brings functional stretches of DNA like enhancers and promotors into proximity which can increase or decrease transcription of RNA molecules and therefore regulate gene expression. CTCF binding affinity and functionality depends on motif conservation. Genetic variants can increase or decrease binding affinity and therefore impact chromatin looping and ultimately gene regulation. There are estimated to be an average of 55,000 CTCF bound motifs in various human cell lines [100]. Assessing the function and importance of these motifs therefore requires bioinformatic guidance.

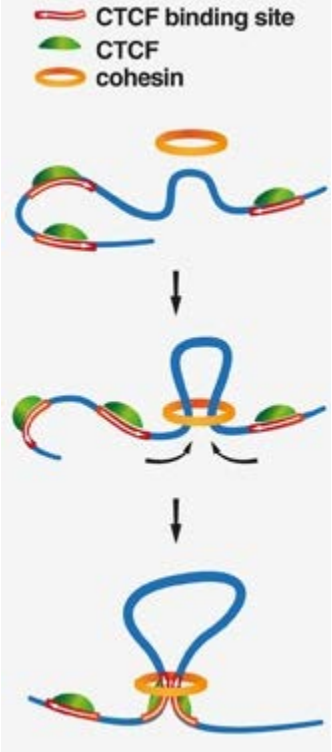


Figure 35: Role of CTCF in 3D genome architecture. CTCF (green) bound to specific sequence (red) along the DNA (blue). Aided by a ring-shaped Protein (Cohesin), bound CTCF- DNA complexes are locked within the cohesion ring structure, stopping DNA from extrusion through the ring. A loop is formed. [143]

4.1.3 Archaic derived variants.

As described above, variants occur randomly at varying rates over time and are kept or discarded from the gene pool by natural selection depending on their functional impact. In the CADD-SV section of this manuscript I used derived chimp and human structural variants. Here I focus on human derived SNVs and InDels that arose on the human lineage after the split from the chimpanzee and bonobo, as well as variants that occur after the split from archaic humans i.e. Neanderthals and Denisovans. These are extinct species (or subspecies) of the homo lineage that split from the modern human line about 500,000 years ago. Both, Neanderthals and Denisovans contribute (on average 1-4 %) DNA to each modern human genome from ancient interbreeding events [33]. What is known about these modern human cousins stems from archeological findings as well as genome sequences from fossils. Behavioral and cultural information is sparse. Further, the reason for extinction is unknown with researchers speculating about climatic events as well as modern human competition and replacement [144]. One of the unsolved questions about modern human success is the contribution of our cognitive abilities and how those compare to archaic humans.

Neanderthals are shown to have similar (or slightly larger) sized skulls [145] and therefore brain volume. Denisovan brain size is unknown due to lack of archeological findings (see Figure 36).

Figure 36: Modern Human (left) and Neanderthal (right) skulls. Image via Wikimedia Commons/DrMikeBaxter

4.1.4 Brain evolution

Human success is often described as a consequence of our unique brain properties. Human evolution led to highly increased brain volumes as well as cognitive function. Various genes have been described that underwent rapid evolution since divergence to other primates such as MCPH1, ASPM [146] or NOVA1 [147]. While MCPH1 and ASPM contribute to larger brain volumes and accumulated many DNA changes after the split from chimpanzees, the human NOVA1 variant emerged after the split from archaic humans. Modern human variants in NOVA1 influence synaptic activity, glutaminergic signaling as well as electrophysiological properties in human organoids compared to archaic versions of the same gene [147]. The same study showed that genetic variants in brain developing genes show effects in human organoids.

However, the effect of CTCF to mediate phenotypic changes in the human lineage is unknown. Further, the role of CTCF for human evolution in general, and in particular on brain evolution, is not well understood. To shed some light on human specific CTCF evolution I use human derived variants (SNVs and Indels) and apply various filtering steps to prioritize brain relevant functional CTCF gain or loss variants. In

this project I use comprehensive annotations to prioritize CTCF peaks that play a role in human specific brain evolution.

4.2 Methods

4.2.1 Identification of human derived changes

As I focus on variants that are important for human specific brain evolution, I use variants reported to have arisen throughout human evolution. I use variants (SNVs and InDels) identified from fixed differences from chimpanzees to identify human specific changes in the genome [83]. The set contains 14.9 million variants and can be downloaded from (https://krishna.gs.washington.edu/download/CADD-development/v1.6/training_data/GRCh37/). These variants arose since the split from chimpanzee in the human lineage and are therefore target of about 4.1 million years of natural selection [148].

4.2.2 Recent human derived changes

To further differentiate variants in the human lineage that arose more recently, I use archaic human genomes from Neanderthals [149] and Denisovans [150] as outliers. Variants that differ from archaic human genomes as well as outgroups are considered under natural selection for less than 700,000 years. This set contains 321,820 variants. Variant sets were downloaded from (cdna.eva.mpg.de/neandertal/altai/catalog/). In the following sections I refer to candidate lists derived from archaic genome variants as “recent human” evolutionary candidates.

4.2.3 Prioritizing functional CTCF sites

The variant lists described above are considered mostly neutral variants as variants would be removed by purifying selection if they are harmful for the carrier. To identify variants relevant for CTCF mediated brain evolution, I further filter all variants in various steps. These steps are designed to prioritize functional variants with a high likelihood to affect genome architecture mediated by CTCF binding.

4.2.4 Open Chromatin

Variant lists were filtered for accessibility to the genomic machinery by using DNase hypersensitive sites (see 3.1.6.3.2) downloaded from ENCODE (https://www.encodeproject.org/files/ENCFF788SJC/). BED files were intersected using *BEDtools intersect* from BEDtools [110]. All files used were transferred to the human genome build GRCh37 where necessary using liftOver [123].

4.2.5 CTCF ChIP-seq peaks

To identify variants overlapping functional CTCF sites, human, great ape and macaque CTCF ChIP-seq peaks were used (see Figure 37). Filtering was conducted differently for CTCF gain and CTCF loss candidates. While CTCF gain variants are required to overlap a CTCF peak in the human but not in the ape datasets, CTCF loss variants are required to be present in at least one outgroup dataset while missing in the human CTCF ChIP-seq peak annotation.

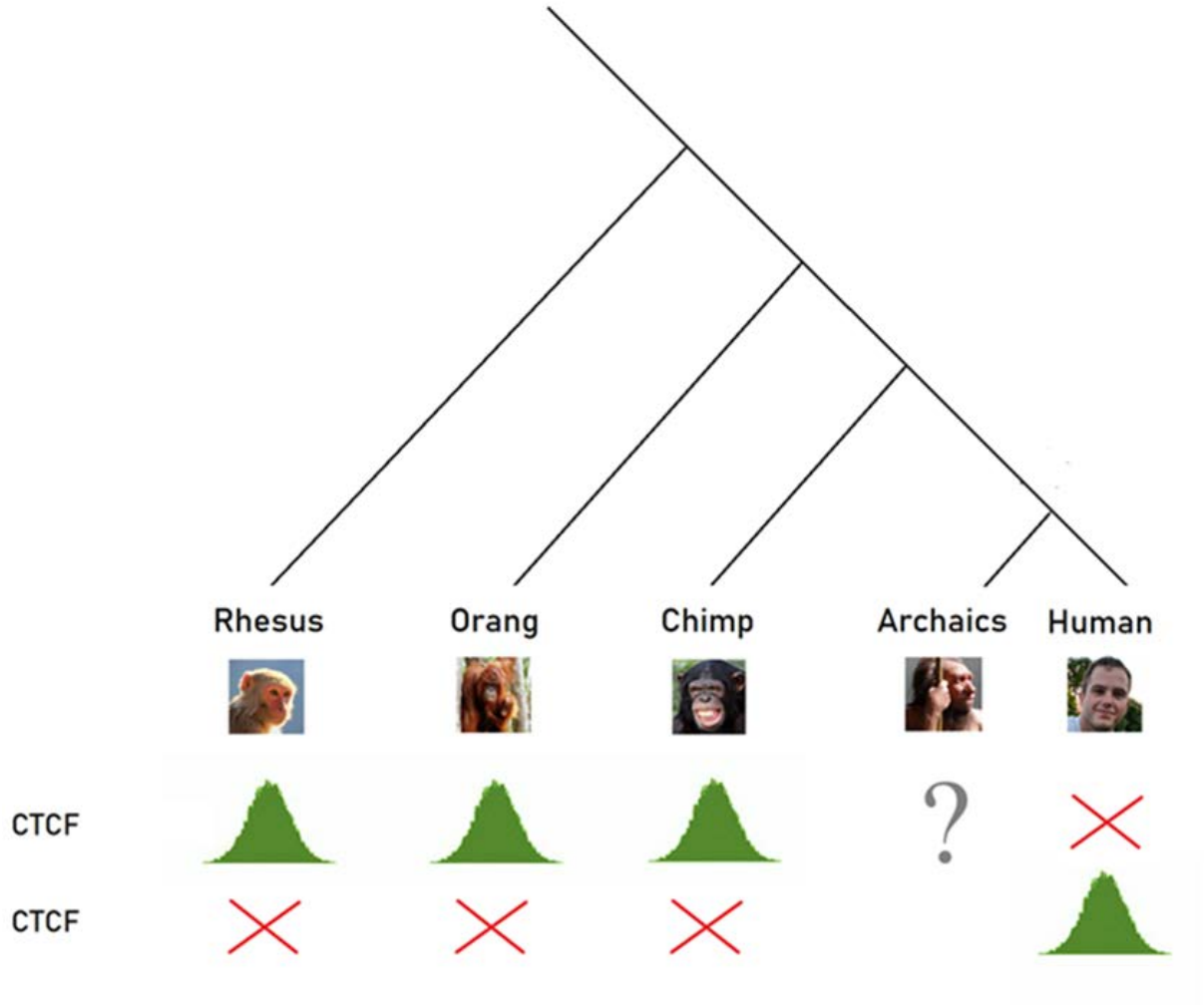


Figure 37: Differential CTCF sites in an evolutionary comparison. Depiction of evolutionary relationship of great apes and maccaca mulatta (Rhesus). Archaic CTCF peaks are not available. Human CTCF loss variants are considered when presence of CTCF loci in outgroup species is observed while the same locus shows no peak in human CTCF abundance. Human CTCF gain is defined as presence of human CTCF while no CTCF peak can be observed in outgroup species.

4.2.5.1 Human CTCF peaks

Human CTCF binding sites were inferred using ChIP-seq peaks from experiments conducted in 19 different human cell lines (downloaded from http://genome.cshlp.org/content/suppl/2012/08/28/22.9.1680.DC1/Table_S2_Location_of_ChIP-seq_binding_positions_in_19_cell_lines.txt).

The original set of 77,811 peaks was further filtered for CTCF binding sites to be present in at least 3 cell lines [100].

4.2.5.2 Ape CTCF peaks

To infer presence or absence of CTCF binding sites in the evolutionary past, I use great ape ChIP-seq datasets to compensate for the absence of biological experiments in archaic genomes. Datasets were downloaded from EBI [151] for *Pongo pygmaeus pygmaeus (Orang)* lymphoblast cells (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1511/E-MTAB-1511.processed.21.zip/do1256_CTCF_LCL_07729upstate_ppyEB185JC_CRI01.fq.sam.bam), *chimpanzee troglodytes lymphoblast* cells (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1511/E-MTAB-1511.processed.18.zip/do1285_CTCF_LCL_07729upstate_ptr18359_CRI01.fq.sam.bam) as well as *Macaca mulatta (Rhesus)* lymphoblast cells (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1511/E-MTAB-1511.processed.26.zip/do1279_CTCF_LCL_07729upstate_mml173-02_CRI01.fq.sam.bam). Peaks were called using best practice guidelines for the JAMM ChIP-seq alignment and peak calling pipeline. (https://github.com/mahmoudibrahim/JAMM/wiki/ChIP-Seq-Alignment-and-Processing-Pipeline). All files were lifted to the human genome build GRCh37.

4.2.6 Motif Scan

To identify the impact of derived variants on the CTCF core motif (see Figure 38), I use FABIAN21 from GeneCascade (https://www.genecascade.org/FABIAN21/). This tool estimates the impact of a variant on binding affinity to the position weight matrix (PWM). PWMs represent importance of individual base pairs for the binding affinity of factors to specific sequences. Therefore, variants are prioritized that increase or decrease binding affinity of CTCF to the novel sequence. Human gained variants were filtered for presence of increased binding affinity, while lost variants were filtered for decreased binding affinity.

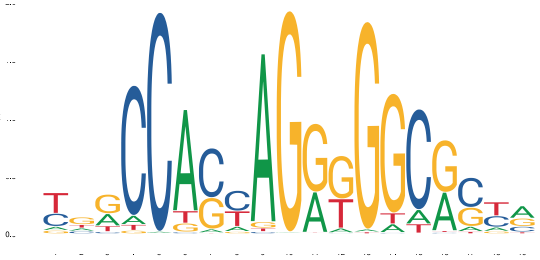


Figure 38: CTCF binding motif (Position Weight Matrix [PWM] representation). Nucleotide bases are represented as A, C, G, T, with their respective size representing importance of individual bases at certain positions within the motif.

4.2.7 Coding Proximity

To further filter candidate variants for functionality, I require variants to be in 75 kb proximity to coding sequence. Gene coordinates [105] were downloaded from ftp://ftp.ensembl.org/pub/release-96/gff3/homo_sapiens/Homo_sapiens.GRCh37.96.chr.gff3.gz.

4.2.8 3D Genome Structure

To infer presence of genome structure, I use annotation of A/B compartments [152] from GM12878 cell lines (ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE63nnn/GSE63525/suppl/GSE63525%5FGM12878%5Fsubcompartments%2Ebed%2Egz). Compartmentalization of the genome hints towards gene regulation by architectural properties. Candidates are filtered to be present in 75 kb proximity to a compartment border to specifically enrich in CTCF sites that mediate 3D genome architecture.

4.2.9 Candidate lists

I apply the listed filtering steps to human variants derived after the split from the chimp as well as human variants derived after the split from archaic humans. Further, two filtering workflows are applied to both derived variant lists to account for the potential functional mechanism of losing or gaining CTCF binding sites. Therefore, CTCF gain variants are filtered for absence of CTCF peaks in the great apes and presence of CTCF peaks in humans as well as increased predicted CTCF binding affinity by FABIAN21. CTCF loss variants are filtered for presence of CTCF peaks in the apes, absence in humans as well as decreased predicted binding affinity. The workflow is shown in Figure 39.

4.2.9.1 Annotating prioritized CTCF sites

In addition to the filtering steps, I annotate the candidate lists with information that provides further insight into the potential candidates.

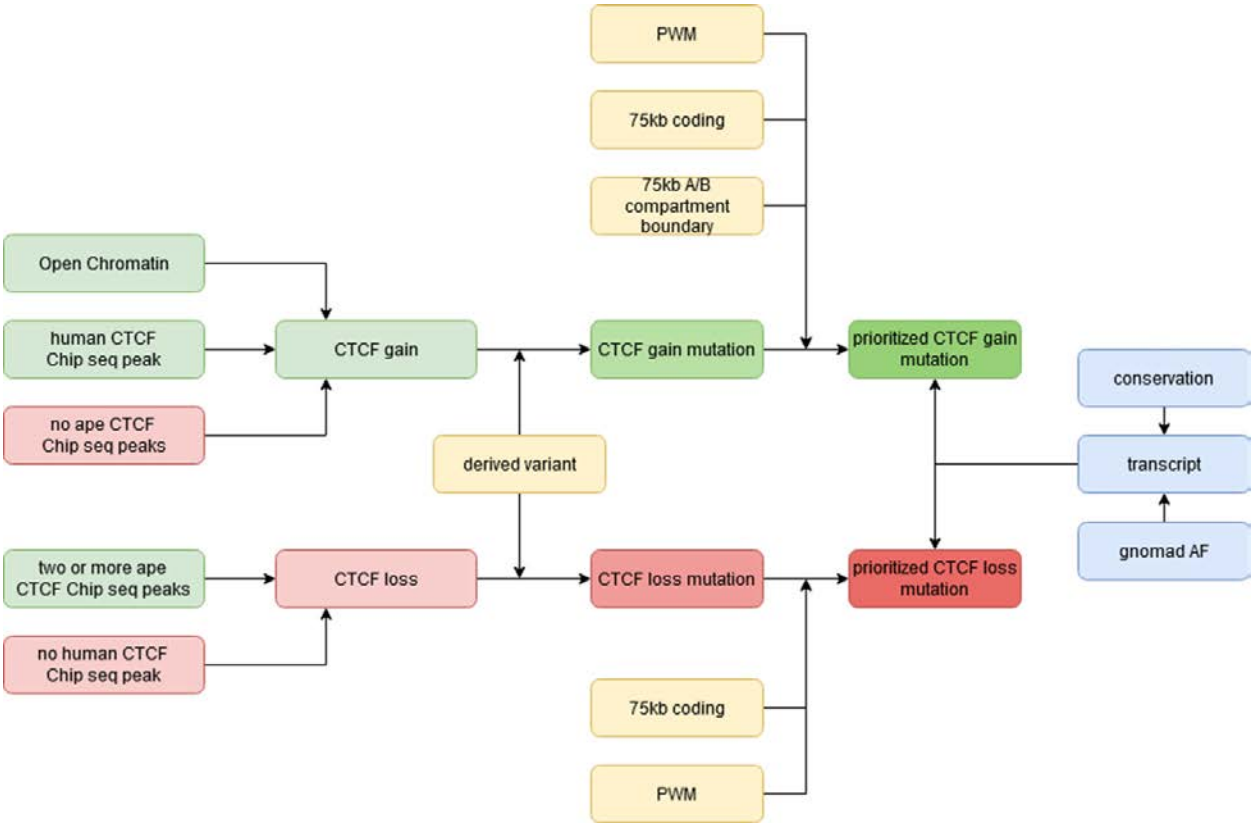


Figure 39: Depiction of the workflow of the CTCF prioritization pipeline. CTCF gain workflow on top (dark green), CTCF loss workflow bottom (dark red). CTCF gain and loss sites are identified using logic depicted in Figure 37. In addition, presence of open chromatin is required for the CTCF gain site. The four sets of derived variant lists (shown in yellow; human gain, human loss, human recent gain, human recent loss) are intersected with the CTCF gain/loss sites respectively to identify mutations within differential CTCF sites. Further annotations are used to prioritize CTCF gain mutations (yellow). Finally additional annotations (blue) are obtained for further inspection.

4.2.9.1.1 Conservation

I also include information of the general conservation of the 100bp region centered on the candidate variant. I calculate the mean PhastCons score [37] as a proxy for functionality as elements of crucial function tend to fall in highly conserved genomic sequence.

4.2.9.1.2 Transcript information

In addition to filtering for proximity to coding sequence, I also include names of all transcripts within the 75 kb up or downstream of the candidate site. Genes were further inspected for links to brain development or evolution from the literature.

4.2.9.1.3 gnomAD Allelefrequency

To give information about presence and abundance of variants in the human population, I annotate all candidate lists with gnomAD allele frequency, if applicable [153].

4.2.9.1.4 Presence of Fantom5 enhancer

Enhancer annotations derived from Fantom5 [154] were used to infer regulatory activity of the region. Fantom5 enhancer coordinates were downloaded from Zenodo (https://zenodo.org/record/556775#.Xkz3G0oo-70). Files were lifted using UCSC liftOver [123] and intersected using BEDtools [110]. Presence or absence of known enhancers are encoded in the candidate output table as 1 or 0, respectively.

4.2.9.2 Gene Ontology Enrichment Analysis

All genes from all four final candidate sets were tested for Gene Ontology enrichment using tissue-specific expression analysis (TSEA) as well as cell-type specific expression analysis (CSEA)-tools version 1.0 [155]. A list of gene symbols (HUGO) was uploaded to the webservice (http://genetics.wustl.edu/jdlab/tsea/ and http://genetics.wustl.edu/jdlab/csea/). Candidate lists are intersected with tissue and cell-type expression lists and tested for enrichment using Fisher's Exact test with Benjamini-Hochberg correction for multiple testing.

4.2.9.3 Validation using NPC and brain organoid CTCF peaks

In addition to the computational prioritization of functional, brain relevant CTCF variants, experiments were conducted by collaboration partners at the University of San Francisco (UCSF) to validate the prioritized variants. CUT&Tag [27] experiments to identify activate CTCF binding sites in brain development were carried out in Neural Progenitor Cells (NPC). NPCs are model cell lines derived from pluripotent stem cells that resemble human brain characteristics [156]. NPC CUT&Tag was carried out on human and chimp derived NPCs. Further, CTCF peaks in chimp, orang and human brain organoids were acquired using CUT&Tag. Data processing was conducted at UCSF on respective reference genomes (sample information is provided in Table 4). All CTCF peak files were lifted to human GRCh37 coordinates using UCSC liftOver.

Table 4: Sample Information for CUT&Tag experiments conducted on human, chimp and gorilla NPCs and organoids

Name	**Species**	**Source**	**Genome reference**
humanNPC	human	NPCs	GRCh37
chimpNPC	chimp	NPCs	Clint_PTRv2/panTro6
humanOrganoid	human	NPCs dissociated from organoids	GRCh37
chimpOrganoid	chimp	NPCs dissociated from organoids	Clint_PTRv2/panTro6
orangOrganoid	orangutan	NPCs dissociated from organoids	ponAbe2

4.3 Results

4.3.1 CTCF gain variants

Out of 14.9 million variants, 34,528 intersect with a CTCF peak found in at least 3 cell lines (73,732 peaks). The vast majority of those are found in open chromatin (33,769). Out of 321,820 variants arising after the split from archaic humans, 1,223 fall within a CTCF peak in 3 cell lines. 1,202 of those intersect with regions of open chromatin. After applying filters described above for motif impact, 75 kb coding and 3D genome architecture boundary proximity, 211 and 6 candidates remain in the human derived gain variant sets after the split from chimp and archaic humans respectively (see Figure 40).

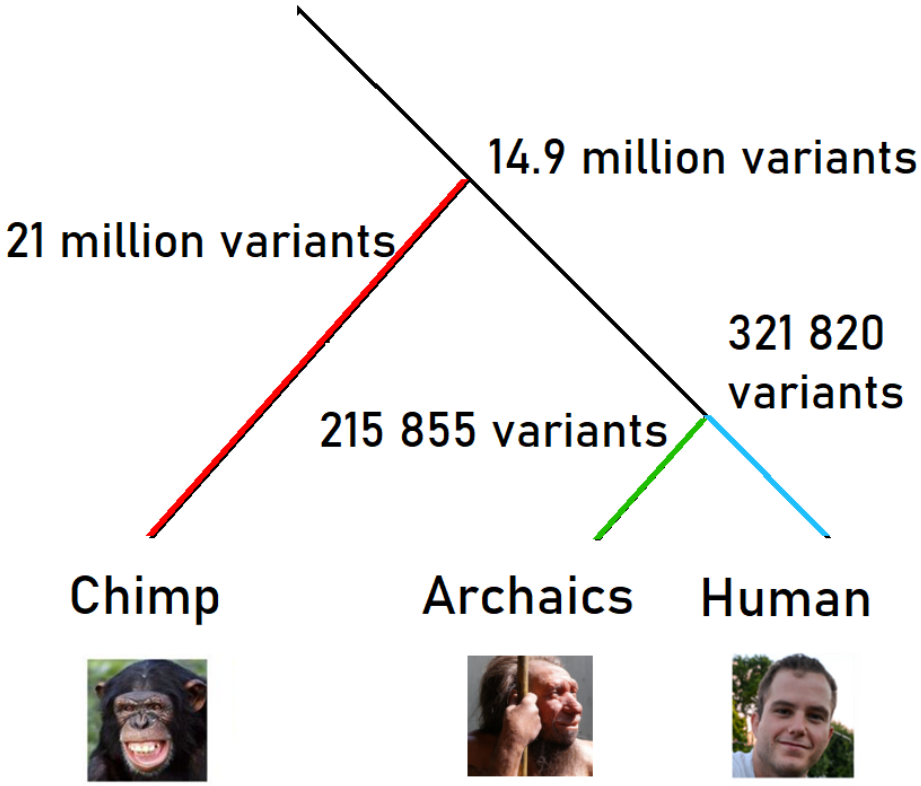


Figure 40: Depiction of evolutionary relationships of chimps, archaics and modern humans. Modern human variants were identified using either chimp or archaic humans as outgroups. Therefore, variants arising in the more recent human past can be identified (light blue).

4.3.2 CTCF loss variants

660 variants derived from chimpanzees fall within CTCF peaks in apes while showing no CTCF binding in human cell lines. 25 recent human evolutionary variants fulfill this criterion. After applying filters described above for motif impact, 75 kb coding proximity 527 and 18 candidates remain in the human derived loss variant sets after the split from the chimp-human ancestor and archaic humans, respectively.

4.3.3 Validation

4.3.3.1 CUT&Tag experiments

CUT&Tag results show high overlap with existing CTCF datasets suggesting experimental success. The number of peaks per cell are further consistent with expected numbers (29,451 CTCF sites in human NPCs, 17,584 in chimp NPCs, 31,510 in human organoid NPCs, 28,804 in chimp organoid NPCs, 19,228 in orang organoid NPCs).

11,925 human NPC CTCF binding sites are human specific, while 9,160 human organoid peaks are human specific (see Figure 41). On the contrary, 11,682 peaks are absent in human NPCs but present in Chimp or Orang. 7,031 peaks are absent in human organoids compared to great ape organoids (see Figure 42).

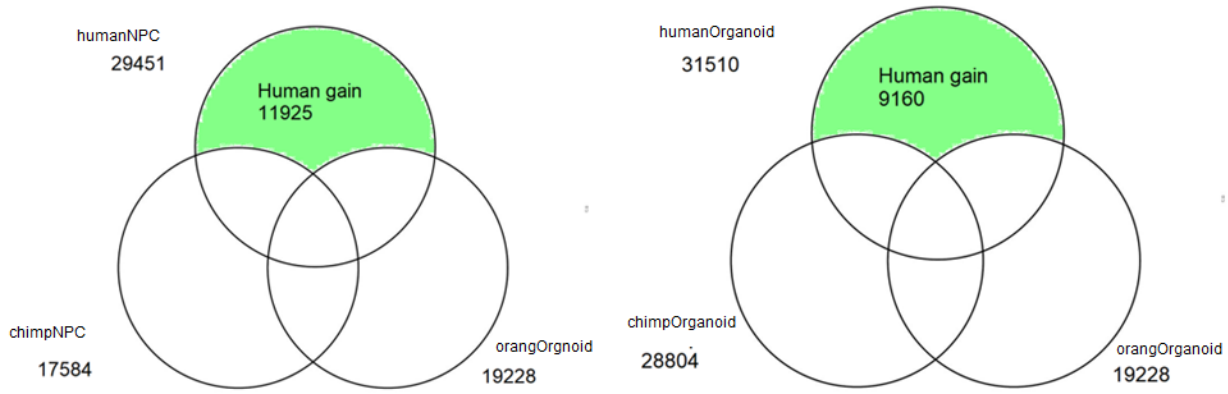


Figure 41: Results from the CUT&Tag experiments, highlighting human gain variants (green). Circles represent all variants obtained from experiments in human-, chimp-, orang- Neural Progenitor Cells (NPCs) or organoids.

4.3.3.1.1 Gain candidates

Further, CTCF CUT&Tag experimental results support our computational approach. Peaks are enriched in the human recent gain set (0.83 % of candidate variants show peaks in human experiments compared to 0.39 % peaks in apes) and the human gain set (0.53 % peaks vs. 0.44 % in apes). All of these values are significant ($p < 0.05$) using a one sample t-test.

Only considering variants that show a CTCF peak in human NPCs and organoids while being absent in all ape experiments, 8 out of 211 candidates remain in the human CTCF gain candidate list (see Table 5). No recent human variant fulfills the strict criteria.

Table 5: Candidate list of human CTCF sites that were gained after the split from chimpanzee.

chr	start	End	alt	ref	enhancer	gnomadAF	PhastCons
chr15	98964963	98964964	C	T	1	0.000574	0.178319
chr15	98965131	98965132	G	T	0	.	0.048224
chr3	48315198	48315199	C	G	0	.	0.20499
chr3	48315253	48315254	T	G	0	.	0.388409
chr3	48315261	48315262	G	A	0	.	0.409313
chr3	58810075	58810076	G	C	0	.	0.175747
chr4	40982541	40982542	GAGCCAACTGCAG	G	0	.	0.013988
chr9	71175209	71175210	C	T	1	.	0.0025

4.3.3.1.2 Loss candidates

Peaks are depleted in the human recent loss (0.33 % compared to 0.65 %) and human loss sets (0.23 % vs 0.31 %). All of these values are significant ($p < 0.05$) using a one sample t-test. In other words: CUT&Tag experiments confirm that gain sets are enriched with human-specific CTCF peaks, and the loss sets are depleted with human-specific CTCF sites.

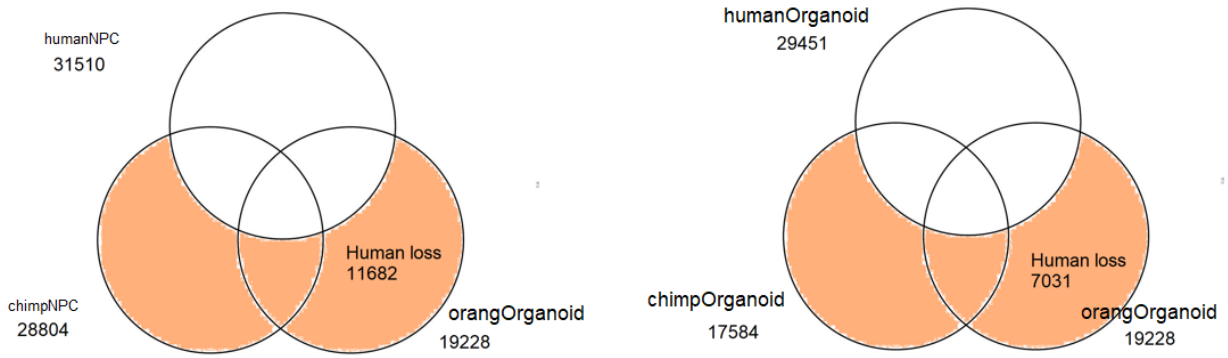


Figure 42: Results from the CUT&Tag experiments, highlighting Human loss variants (orange). Circles represent all variants obtained from experiments in human-, chimp-, orang- Neural Progenitor Cells (NPCs) or organoids.

Taking experimental readouts into account, 21 prioritized loss candidates show CTCF binding in chimp NPCs as well as chimp and orang organoids, while absent in human experimental readouts (see Table 6).

Table 6: Candidate list of human CTCF sites that were lost after the split from chimpanzee.

chr	start	end	alt	ref	enhancer	gnomadAF	PhastCons
chr10	5952519	5952520	T	C	0	.	0.003037
chr11	69521299	69521300	A	G	0	.	0.290729
chr15	43622270	43622271	C	G	0	.	0.3332
chr17	73893726	73893727	C	CT	0	.	0.197172
chr17	73893787	73893788	C	T	0	.	0.00275
chr17	73893842	73893843	C	G	0	.	0.019202
chr5	1.31E+08	1.31E+08	A	C	0	.	0.170924
chr5	1.31E+08	1.31E+08	A	G	0	.	0.164691
chr5	1.31E+08	1.31E+08	A	G	0	.	0.190525
chr8	1.02E+08	1.02E+08	T	G	0	.	0.179167
chr1	15586145	15586146	C	T	0	.	0.352513
chr1	15586172	15586173	C	CT	0	.	0.315178
chr1	2.28E+08	2.28E+08	G	A	0	.	0.07116
chr1	2.28E+08	2.28E+08	GGAGC	G	0	.	0.07116
chr1	2.28E+08	2.28E+08	G	C	0	.	0.079073
chr10	43704791	43704792	A	G	0	3.19E-05	0.017816
chr10	1.02E+08	1.02E+08	T	G	0	.	0.07833
chr12	1.1E+08	1.1E+08	A	G	0	.	0.072849
chr12	1.22E+08	1.22E+08	T	C	0	.	0.047422
chr19	8461735	8461736	T	C	0	.	0.004179
chr5	1.51E+08	1.51E+08	A	G	0	.	0.134286

One variant fulfils the criteria for the recent human loss prioritized variant list:

Table 7: Candidate list of recent human CTCF sites that were lost after the split from archaic humans.

chr	start	End	alt	ref	enhancer	gnomadAF	PhastCons
chr12	1.1E+08	1.1E+08	A	G	0	.	0.072849

4.3.4 Annotations

4.3.4.1 Gene Ontology Enrichments

No molecular or biological function was enriched in the analysis. The combined gene list is enriched for genes expressed in the developing brain (using CSEA tool [157], Bonferroni corrected p= 0.001). Genes overlapping developing brain expression patterns are: *CA12, SLCO3A1, ABCG1, LPAR1, BTG2, CPS1, PPP3CA, GRIK4, LANCL1, DMRT2, SLC9A5, HTR6, RAP1GAP, CAPN13, HRH2, CA2, GLTP, RBM20, ACVR1C, FGF1, FOSB, TBC1D12, BBX, GSN, SNAI3, TFEC, ABCC5, PRKCB, CDC14B, GCH1, WIPF3, FAM124A*

4.3.4.2 Enhancer overlap, Allelefrequency, Conservation

Eleven prioritized human gain variants and 28 prioritized human loss variants overlap known Fantom5 enhancers. Both equal 5% of prioritized variants. No enhancers are found in the recent evolutionary variant sets. The vast majority, 179 (85 %) prioritized human gain variants and 449 (85 %) prioritized human loss variants are absent in the gnomAD population cohort. Most variants are rare with one variant (chr2:121438101_TGCGCGCGCATGTGCGTGTGT>T) however reaching a global allele frequency of 33 %.

All candidates show increased PhastCons scores compared to the background distribution (randomly shuffled variants undergoing the same filtering procedure) (see Figure 43). This observation is particularly strong in sites of recent human evolution.

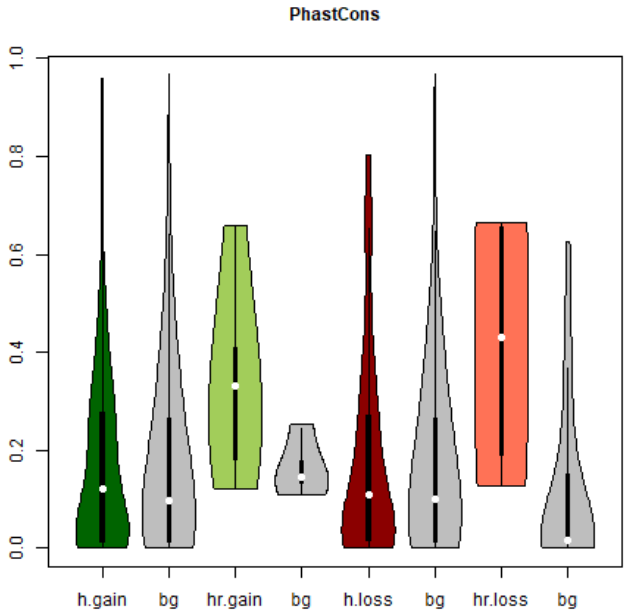


Figure 43: Evolutionary constraint on human gain (dark green), human recent gain (light green), human loss (red) and human recent loss (orange) compared to their respective background distributions. Median values are shown as white dots.

4.4 Discussion

Here I prioritize and further interpret functional variants with a specific focus on a DNA binding factor (CTCF) involved in 3D genome architecture [11,20,139]. Variants disrupting or creating CTCF binding motifs potentially alter genomic 3D structure and therefore gene regulation [143]. I focus on variants arising during human-specific evolution since the split from Chimpanzee and archaic humans. Human evolutionary success remains poorly understood, while many researchers point towards modern human extraordinary brain capacities [145,158]. The role of 3D genome architecture, and especially human-specific CTCF binding sites, on human brain development and evolution remains elusive. With this computational approach, I show that public datasets can shed light on human-specific evolution.

By prioritizing variants that show differential CTCF binding profiles and showing proximity to 3D genomic boundary regions as well as coding sequence, I obtain 4 variant lists of human evolutionary relevant CTCF binding sites containing a total of 762 CTCF binding site gain (217) and loss (545) mutations.

I show that using publicly available datasets is valuable, even if datasets are not specifically focused on brain evolution. Experimental results from human NPCs as well as organoids show that prioritized lists of gain and loss candidate CTCF binding sites are enriched in brain-specific gained or lost CTCF binding sites. These specific experiments are able to further filter candidate lists, resulting in 8 experimental validated gain variants, and 22 experimental validated loss variants. I recommend prioritizing computationally prefiltered variant lists further into CTCF gain variants showing complete absence in all ape experiments and presence in all human experiments while filtering loss variants for complete absence in human read outs and presence in the apes.

4.4.1 Conservation

Candidates show increased conservation for all four variant lists. This observation is especially striking in the recent evolutionary candidate sets, potentially because numbers for these sets are small. Conservation hints towards functional genomic regions, as variants disrupting sequence function are being removed by purifying selection. This observation is not surprising though as filtering requires presence of CTCF binding in apes or humans, prefiltering for functional DNA stretches.

4.4.2 Candidate genes

Candidate lists contain promising targets for further analysis. Enrichment analyses show that especially genes expressed in the developing brain are encountered. These lists, however, are incomplete, as various additional genes show brain expression while not being part of specific expressed “brain development” genes. For instance, a human specific A to G CTCF gain variant upstream of AKAP6 arose after the modern human split from archaic humans. AKAP6 is highly expressed in the brain and is associated to brain relevant disease phenotypes [159] such as schizophrenia [160]. Rewired 3D genome structure mediated by an additional CTCF binding site could potentially alter gene expression of AKAP6 in time or space.

Three G to A variants leading to disruption of a CTCF binding motif in ACLS6 arose since the evolutionary split from chimpanzees. ACLS6 is crucial for acyl-CoA creation from fatty acids in the brain [161,162]. Finally, a small insertion on chromosome 4 position XX (G to GAGCCAACTGCAG) in the transcript of APBB2 creates a novel CTCF binding site. Here, experimental CTCF binding is completely absent in the great apes while being present in human NPCs and organoids. APBB2 is potentially relevant for brain development as it is associated with abnormally large and dysfunctional axons [163] as well as Alzheimer’s disease [164].

4.4.3 Experimental validation

Validating the computational filtering approach using NPC and organoid CTCF binding sites as well as individual examples mentioned above shows the potential of the prioritized candidate lists. Few variants remain that fulfill all criteria when taking CUT&Tag experimental results into account (8 human gain, 21 human loss, 1 human recent loss variant). Criteria are strict and result in a conservative set of highly prioritized variants. However, as experimental validation is cost and labor intensive, few candidates can be tested *in vitro*. Functional validation is needed to further understand the role of CTCF and therefore 3D genome architecture on human brain evolution. Proposed candidate lists can be used to experimentally validate individual targets by introducing gain or loss variants into chimpanzee cell lines, organoids or mouse models using CRISPR systems. Experimental readouts such as RNA-seq can be used to infer expression changes due to the altered CTCF binding or Hi-C and other assays to identify alterations to the genomic compartments. In addition computational approaches like deepC[125] can be used to infer the validity computationally. Phenotypic readouts such as axon length or mouse behavior can be measured in brain organoids or mouse models.

4.4.4 Computational validation and limitations

Computational validation of these candidate lists is limited as comprehensive datasets naturally correlate with the input data and filtering steps. Independent CTCF datasets, such as computationally predicted CTCF binding sites, will be enriched in the candidates by design. Further limitations include the limitation of public human and great ape brain tissue derived datasets such as CTCF peaks or open chromatin. Even though validation using NPCs and organoid derived CTCF peaks show high overlap with existing datasets. More refined projects will require higher quality and more specific datasets to be used in filtering. The workflow of this pipeline can be easily extended using different datasets or by focusing on novel evolutionarily relevant mechanisms.

4.4.5 Conclusion

In conclusion, in this project I prioritize genetic variants for their impact on 3D genome architecture mediated by CTCF. I show that variants can be filtered comprehensively using public datasets. I report four lists of human-specific gained or lost CTCF binding sites using great ape and human datasets. Candidates were validated using experimental CTCF peaks from chimp, orang and human NPCs and organoids. Further validation can be applied using various experimental readouts, including directed editing of alleles in human and chimpanzee cells using CRISPR technologies.

5 Discussion

5.1 Contributions to Variant Interpretation

In this thesis, I outline three approaches to help understand human genetic variants. I look at targeted sequencing datasets and provide a pipeline for clinicians as well as researchers that helps interpret variants from data generated using molecular inversion probes (MIPs) [3]. I provide a comprehensive framework to estimate the deleterious effects of structural variants in the human genome [4]. And finally, I link variant interpretation to human specific brain development, focusing on CTCF mediated changes in the recent human evolutionary past. All projects accomplish a better understanding of human genomic variants and contribute novel insights to the genomics research community.

5.1.1 hemoMIPs

While individual pipelines exist to analyze MIP-derived next generation sequencing datasets, none is open source and published. Providing a pipeline that is version controlled, uses state-of-the-art software management, is well described in a publication and is free to use and to extend with additional features, makes MIP data generation more accessible to the genetics community. In addition, MIP-targeted datasets can be compared to one another using standardized pipelines. MIP applications for other cohorts can be easily designed and clinically validated by non-bioinformatic researchers with the fully automated processing pipeline. Variants of interest for hemophilia can be further extended as well as individually designed and added into the pipeline for other disease cohorts. Therefore, hemoMIPs provides a framework to easily analyze and compare targeted sequencing datasets in a fully automated manner that has not been available before.

5.1.2 CADD-SV

Computational frameworks to estimate the functional impact of structural variants have been developed over recent years [119–121,138]. Many pipelines have been published recently or are currently in development to prioritize SVs. However, most frameworks suffer from certain ascertainment biases intrinsic to their training datasets. For instance, AnnotSV is a powerful tool that is easy to install and fast and efficient in

annotating and prioritizing variants. However, many of its annotation features rely on clinical interpretation such as presence in clinical databases such as ClinVar. Therefore, rare disease phenotypes or poorly studied mechanisms might be underestimated. CADD-SV on the other hand was designed to be as free as possible from known ascertainment biases. It does not make use of clinical, hand-picked annotations or training data. Instead, it uses evolutionary fixed SVs in chimps and humans which are presumably free of biases intrinsic to human annotated SV sets. Therefore CADD-SV is the first framework that identifies functional SVs while not suffering from common ascertainment biases. Additionally, CADD-SV is the first framework designed to capture the diversity of potential SV effects by providing users the full annotation set in a normalized and easy to interpret manner. All SV annotations are Z-transformed using SVs of the same type from a healthy cohort. This novel approach enables users to directly observe outliers in individual categories or annotations beyond an aggregated score value.

5.1.3 CTCF evolution

Finally, non-coding DNA still strongly lacks interpretability compared to coding DNA. Genes within regulatory domains are often co-expressed and in proximity. Variants influencing these domains (for instance in enhancer sequence or sites influencing 3D genome architecture) can impact expression patterns and therefore functionally affect individuals. One comparatively well understood mechanism is changes in binding affinity due to variants in transcription factor binding motifs [165]. However, no comprehensive framework exists to prioritize variants in this manner. Further, no framework exists that specifically looks at CTCF occupancy changes, mediated by heritable (germline) variants in the human lineage. This novel approach therefore combines public datasets with human variant lists to specifically prioritize non-coding variants that affect CTCF mediated 3D genome architecture for the first time.

5.2 The future of the developed approaches

As genetic research never sleeps and continuously develops, all approaches have been designed to be extendable for potential future usage.

5.2.1 hemoMIPs

hemoMIPs, as the name suggests, was designed using data generated for a hemophilia cohort conducted by the "My Life Our Future" Project consortium. However, targeted sequencing using MIPs is very powerful and can be easily extended for further disease cohorts. The pipeline is limited to sequencing data generated from Illumina machines using molecular inversion probes as it was tuned to handle specific characteristics of this type of datasets, such as strong coverage imbalance and adaptor trimming. However, as hemoMIPs is public and open source, it can be easily used and extended by users. For example, tailoring hemoMIPs for other disease cohorts such as BRCA1 sequencing for breast cancer screenings is feasible [65,166]. MIP targets can be designed for new diseases as well as extended for hemophilia to additional loci/genes inferred from clinically described variants. To conclude, the hemoMIPs pipeline is well equipped to be used in future projects due to its usage of workflow management tools as well as its open-source approach.

5.2.2 CADD-SV

The CADD-SV framework is not just easy to use due to its documentation, open-source approach and web-based scoring feature, the tool can easily be extended to future aspects of SV scoring. The framework can be used to annotate new training datasets as they emerge. The CADD-SV annotation rules are not limited to evolutionarily motivated training datasets but can also be used for other types of training sets. For instance, a model trained on gnomAD rare versus common SVs could be explored using the CADD-SV annotation pipeline. For instance, additional models for inversions can be handled with minor additions to the code, once a training dataset of sufficient size is available. One future improvement of CADD-SV could be its current limitation to insertion site annotations. CADD-SV does not make use of the inserted sequence itself but only infers deleteriousness from the insertion site its proximity to certain annotations. However, sequence models exist to estimate the functionality and the direct impact of the inserted sequence itself [167]. Incorporation of such information into a set of CADD-SV models could be a useful addition to the current framework. Further, CADD-SV scores could be ranked depending on certain feature groups. For instance, variants could be ranked with SVs of similar size, allele frequency or gene density. The employed approach does not do that. Instead, SVs are ranked relative to the biggest and least biased population SV call set currently available (gnomAD-SV). Exploring relative ranking based on all or individual features mentioned above could be included in future CADD-SV versions.

Further, the set of annotations used in CADD-SV can be easily extended. CADD-SV is based on a snakefile containing individual rules that annotate each SV using datasets containing features and their respective genome coordinates. It is already designed to be able to handle various data formats, such as BED files or GFF files. Further feature sets and transformations can be easily added to the annotation pipeline and included in future models. CADD-SV is limited by the limitations of SV detection itself. This is especially the case for SV breakpoints as they tend to be in highly repetitive regions and are sometimes unreliable, with intervals rather than exact location. CADD-SV uses 100bp up- and downstream of the SV of interest. More refined methods to detect SVs might benefit the current approach. Alternative definitions of the up- and downstream regions could be explored for training, potentially optimizing the exact number of base pairs used.

To conclude, the approach does hold limitations for scoring breakpoints and is unable to score inversions as well as making use of inserted sequence. Due to the open-source approach, the aforementioned limitations can be addressed in future updates, using extended training datasets as well as additional or updated models for inversions as well as inserted sequence.

5.2.3 CTCF-pipeline

The pipeline to identify and prioritize human specific CTCF gain, or loss mutations can be easily extended to additional datasets, species comparisons or transcription factors. The approach uses experimental CTCF binding sites to identify evolutionary differences in transcription factor occupancy mediated by germline variants. Currently the experimental data was generated in ape fibroblast cell lines. However, as this project focuses specifically on 3D architecture changes in brain development, CTCF CUT&Tag datasets from more relevant cell-types or organoids could be used in the future.

This approach can be further extended to all kinds of transcription factors and evolutionary relationships. Other TFs with known binding sites can be incorporated into the framework, identifying gain and loss sites. Human lineage derived variant lists would be untouched. Additional datasets such as information on open chromatin regions as well as distance to coding sequence would be identical if differences are to be identified for human.

Variant lists for other species are available, or could be easily generated, using state of the art alignment tools. Therefore, CTCF or other transcription factor differences could be identified in other species

comparisons as well. Non-coding DNA is still a mystery to many clinicians [168] even though disease causing variants might be found in non-coding DNA stretches. This approach is well suited to shed light on variants impacting transcription factor binding.

On top of the computational approach, future work on human specific CTCF evolution must include experimental validation. Organoid or mouse models can be generated to validate the effect of variants prioritized with this framework.

The CTCF pipeline is currently not open source or publicly available. Releasing the approach on GitHub would be needed in the future to make the framework available to the community.

5.3 Future of variant interpretation

5.3.1 Diverse population sets

All three projects contribute novel insights into interpretation of human genetic variants. However, challenges to correctly identify and interpret variants beyond the power of a single PhD student remain. Diverse populations may improve fine mapping to overcome the challenges introduced by linked variants to better identify causative variants that can be prioritized. Currently, a majority of existing genetic databases rely on data from European ancestry [169]. Further increased population cohorts can increase the resolution for association studies and therefore help identify functional variants. Increased number of individuals also increases the resolution to fine map variants. It is assumed that among the nearly 8 billion human individuals any SNV compatible with life can be observed [170].

5.3.2 Improved Computational Frameworks

As outlined in this manuscript, computational approaches are key to handle complex and large datasets for variant interpretation. Recent advances in computational protein structure predictions show the potential of computational approaches to predict variant effects. Billions of known protein sequences cannot be experimentally assessed for their structure. AlphaFold introduces a state-of-the-art neural network-based model to predict structures to atomic accuracy that might excel experimental approaches [171]. Strategically employed training datasets combined with powerful models and advances in the field of

deep learning algorithms [172] have the potential to revolutionize research fields beyond protein structure predictions.

5.3.3 Functional Characterization

Experimental variant characterization has many limitations because functions act through a wide set of molecular and cellular processes. Often a limited set of variants can be tested in a limited number of cell types. High throughput assays pave the way forward. Genome perturbations become increasingly common with new technologies such as MPRAs [38], CRISPR/Cas9 [47], single cell technologies or new approaches to synthesize whole genomes [173] currently under development and being further improved. Ever increasing numbers of variants will become feasible to be tested in parallel for an ever-increasing number of cellular or organismal systems. Organoid models are already used to functionally characterize complex diseases such as Alzheimer's disease[174]. Many traits are difficult to assess in cell culture due to their phenotypic complexity. Methods to further improve functional characterization have the potential to revolutionize our understanding of genomic variants in the near future. This thesis provides computational approaches to prioritize variants for further functional characterizations.

5.4 Accessibility

The accessibility and availability of basic research tools is often overlooked. I find accessibility crucial for future developments, to continue to build upon the foundations laid down by previous researchers. Therefore CADD-SV and hemoMIPs are both open-source, use workflow and dependency management and are released on GitHub. CADD-SV is further available as an online scoring resource, enabling researchers or clinicians without bioinformatic knowledge to prioritize variants themselves. Access to resources especially for clinicians is still far from universal for software in variant interpretation. Some disease phenotypes might be well described in cell lines or mouse models but lack clinical applications and therapies that benefit human patients directly. Some clinical impact is limited due to availability of computational frameworks. A webserver for variant interpretation is therefore crucial to link basic research to clinicians for instance in university hospitals. The CADD-SV webserver further provides direct links to additional resources for SVs of interest. gnomAD-SV [19], USCS genome browser [123] tracks and region details available from ENSEMBL [175], are just one click away.

GitHub manuals for CADD-SV and hemoMIPs were thoroughly laid out and tested accordingly. All additional steps to run the software are described in detail on each tool's respective GitHub page. Additional steps to prepare the datasets to be ready for usage in the pipelines are also described.

5.5 Conclusion

I present three approaches to shed light on the function of human genomic variants. I aid variant interpretation by providing a workflow implementation of software to analyze targeted clinical sequencing datasets. I introduce CADD-SV, a framework to predict the deleterious effects of structural variants in the human genome. And finally, I interpret non-coding variants in a novel approach, prioritizing differential CTCF binding, prioritizing regions of evolutionary derived expression changes in human brain development genes for experimental follow-up.

Using Snakemake workflows [50], hemoMIPs performs sample demultiplexing, overlap paired-end merging, alignment using BWA, MIP-arm trimming, variant calling using GATK, coverage analysis and HTML report generation for single molecule and paired end sequencing datasets. While hemoMIPs was developed to analyze targeted sequencing data of the MLOF Initiative, it can be applied to a broad set of MIP sequencing data sets for direct visualization of clinically relevant disease variants.

CADD-SV integrates rich sets of annotations in predictive models of SV effects. CADD-SV is built from machine learning models with an unbiased training using evolutionary-derived and putative benign variants that underwent millions of years of purifying selection. Clinical interpretation of structural variants is improved by CADD-SV due to its comprehensive score design, online scoring as well as provisions for clinicians with the respective annotations to identify outlier SVs.

Finally, in the CTCF evolution section I prioritize genetic variants for their impact on 3D genome architecture mediated by CTCF. I show that variants can be filtered comprehensively using public datasets. I report four lists of human-specific, gained or lost CTCF binding sites using great ape and human datasets. Candidates were validated using experimental CTCF peaks from chimp, orang and human NPCs and organoids. While having no direct functional link yet, putative CTCF-mediated expression changes may play a role in Alzheimer's disease, schizophrenia and many more diseases.

Hence , my thesis projects improve our understanding of non-coding variants, the identification of disease variants as well as structural variants and their contribution to human evolution and health.

6 References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–424 (2015).
3. Kleinert, P., Martin, B. & Kircher, M. HemoMIPs—Automated analysis and result reporting pipeline for targeted sequencing data. *PLOS Computational Biology* **16**, e1007956 (2020).
4. Kleinert, P. & Kircher, M. A framework to score the effects of structural variants in health and disease. *Genome Res* gr.275995.121 (2022) doi:10.1101/gr.275995.121.
5. *Molecular biology of the cell*. (Garland Science, 2008).
6. Peyvandi, F., Garagiola, I. & Young, G. The past and future of haemophilia: diagnosis, treatments, and its complications. *Lancet* **388**, 187–197 (2016).
7. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769–1776 (2011).
8. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
9. van Berkum, N. L. *et al.* Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *J Vis Exp* 1869 (2010) doi:10.3791/1869.
10. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21**, 207–226 (2020).
11. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* **20**, 535–550 (2019).
12. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* **17**, 743–755 (2016).
13. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
14. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437–455 (2019).
15. Benfey, P. N. & Mitchell-Olds, T. From Genotype to Phenotype: Systems Biology Meets Natural Variation. *Science* **320**, 495–497 (2008).
16. A framework for exhaustively mapping functional missense variants. *Molecular Systems Biology* **13**, 957 (2017).
17. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**, 327–335 (1996).
18. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
19. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
20. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nature Reviews Genetics* 1 (2018) doi:10.1038/s41576-018-0007-0.
21. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).
22. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

23. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1932–1941 (2014).
24. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
25. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol Mol Biol Rev* **79**, 347–372 (2015).
26. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
27. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).
28. Darwin, C. *On the Origin of Species, 1859.* (Taylor and Francis, 2012).
29. Gregory, T. R. Understanding Natural Selection: Essential Concepts and Common Misconceptions. *Evo Edu Outreach* **2**, 156–175 (2009).
30. Masel, J. Genetic drift. *Current Biology* **21**, R837–R838 (2011).
31. Isenbarger, T. A. *et al.* The most conserved genome segments for life detection on Earth and other planets. *Orig Life Evol Biosph* **38**, 517–533 (2008).
32. Bräuer, G., Yokoyama, Y., Falguères, C. & Mbua, E. Modern human origins backdated. *Nature* **386**, 337–338 (1997).
33. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
34. Chen, F.-C. & Li, W.-H. Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. *The American Journal of Human Genetics* **68**, 444–456 (2001).
35. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).
36. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
37. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
38. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**, 1–15 (2019).
39. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).
40. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
41. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–913 (2005).
42. Lewontin, R. C. The Apportionment of Human Diversity. in *Evolutionary Biology: Volume 6* (eds. Dobzhansky, T., Hecht, M. K. & Steere, W. C.) 381–398 (Springer US, 1972). doi:10.1007/978-1-4684-9063-3_14.
43. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls | Nature. https://www.nature.com/articles/nature05911.
44. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
45. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).

46. Ioannidis, J. P. A., Trikalinos, T. A. & Khoury, M. J. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol* **164**, 609–614 (2006).
47. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
48. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
49. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15**, 475–476 (2018).
50. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
51. Cantsilieris, S., Stessman, H. A., Shendure, J. & Eichler, E. E. Targeted Capture and High-Throughput Sequencing Using Molecular Inversion Probes (MIPs). *Methods Mol Biol* **1492**, 95–106 (2017).
52. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).
53. O'Roak, B. J. *et al.* Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* **338**, 1619–1622 (2012).
54. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
55. Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**, 2670–2672 (2014).
56. Johnsen, J. M. *et al.* Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the My Life, Our Future initiative. *Blood Advances* **1**, 824–834 (2017).
57. Kircher, M. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol* **840**, 197–228 (2012).
58. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
59. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).
60. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
61. Rossetti, L. C., Radic, C. P., Larripa, I. B. & Brasi, C. D. D. Developing a new generation of tests for genotyping hemophilia-causative rearrangements involving int22h and int1h hotspots in the factor VIII gene. *Journal of Thrombosis and Haemostasis* **6**, 830–836 (2008).
62. Nijman, I. J. *et al.* Targeted next-generation sequencing: A novel diagnostic tool for primary immunodeficiencies. *Journal of Allergy and Clinical Immunology* **133**, 529-534.e1 (2014).
63. Weerakkody, R. A. *et al.* Targeted next-generation sequencing makes new molecular diagnoses and expands genotype–phenotype relationship in Ehlers–Danlos syndrome. *Genet Med* **18**, 1119–1127 (2016).
64. Zahari, M. *et al.* Mutational Profiles of F8 and F9 in a Cohort of Haemophilia A and Haemophilia B Patients in the Multi-ethnic Malaysian Population. *Mediterr J Hematol Infect Dis* **10**, e2018056 (2018).
65. Neveling, K. *et al.* BRCA Testing by Single-Molecule Molecular Inversion Probes. *Clinical Chemistry* **63**, 503–512 (2017).
66. Pedersen, B. S., Murphy, E., Yang, I. V. & Schwartz, D. A. Aligning sequence from molecular inversion probes. *bioRxiv* 007260 (2014) doi:10.1101/007260.

67. Aydemir, O. *et al.* Drug-Resistance and Population Structure of Plasmodium falciparum Across the Democratic Republic of Congo Using High-Throughput Molecular Inversion Probes. *J Infect Dis* **218**, 946–955 (2018).
68. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102**, 36–49 (2016).
69. Gong, T., Hayes, V. M. & Chan, E. K. F. Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics* **22**, bbaa056 (2021).
70. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
71. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
72. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
73. Courtens, W., Wuyts, W., Rooms, L., Pera, S. B. & Wauters, J. A subterminal deletion of the long arm of chromosome 10: A clinical report and review. *American Journal of Medical Genetics Part A* **140A**, 402–409 (2006).
74. Ibn-Salem, J. *et al.* Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biology* **15**, 423 (2014).
75. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125–138 (2013).
76. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).
77. Kehr, B., Melsted, P. & Halldórsson, B. V. PopIns: population-scale detection of novel sequence insertions. *Bioinformatics* **32**, 961–967 (2016).
78. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597–614 (2020).
79. Cardozo Gizzi, A. M. *et al.* Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Molecular Cell* **74**, 212-222.e5 (2019).
80. Chen, M. *et al.* Identification of a likely pathogenic structural variation in the LAMA1 gene by Bionano optical mapping. *npj Genom. Med.* **5**, 1–6 (2020).
81. Klopocki, E. *et al.* Deletions in PITX1 cause a spectrum of lower-limb malformations including mirror-image polydactyly. *Eur. J. Hum. Genet.* **20**, 705–708 (2012).
82. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput Biol* **13**, e1005708 (2017).
83. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310–315 (2014).
84. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
85. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology* **6**, e1001025 (2010).
86. Irimia, M. *et al.* Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
87. Hecker, N. & Hiller, M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *GigaScience* **9**, giz159 (2020).
88. Foster, K. A., Oster, C. G., Mayer, M. M., Avery, M. L. & Audus, K. L. Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. *Exp Cell Res* **243**, 359–366 (1998).

89. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
90. Nayan, V., Onteru, S. & Singh, D. Reproduction and nutriment–nurture crosstalk: epigenetic perspectives. *Journal of Reproductive Health and Medicine* **1**, 50–59 (2015).
91. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).
92. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Research* (2017) doi:10.1093/nar/gkx1092.
93. Piovesan, A. *et al.* On the length, weight and GC content of the human genome. *BMC Research Notes* **12**, 106 (2019).
94. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
95. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat Genet* **51**, 88–95 (2019).
96. Firth, H. V. & Wright, C. F. The Deciphering Developmental Disorders (DDD) study. *Developmental Medicine & Child Neurology* **53**, 702–703 (2011).
97. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7.20 (2013).
98. Samocha, K. E. *et al. Regional missense constraint improves variant deleteriousness prediction*. 148353 https://www.biorxiv.org/content/10.1101/148353v1 (2017) doi:10.1101/148353.
99. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
100. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* **22**, 1680–1688 (2012).
101. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods* **72**, 65–75 (2015).
102. Calandrelli, R., Wu, Q., Guan, J. & Zhong, S. GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data. *Genomics, Proteomics & Bioinformatics* **16**, 365–372 (2018).
103. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol.* **19**, 56 (2018).
104. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042–2059 (2016).
105. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**, baw093 (2016).
106. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electron Markets* **31**, 685–695 (2021).
107. Ying, X. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series* **1168**, 022022 (2019).
108. Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**, (2018).
109. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2021. (2021).
110. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
111. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
112. Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development* **22**, 191–203 (2012).

113. Gardner, E. J. *et al.* Contribution of retrotransposition to developmental disorders. *Nature Communications* **10**, 4630 (2019).
114. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
115. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* (2021) doi:10.1126/science.abf7117.
116. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
117. Qian, Y. *et al.* Identification of pathogenic retrotransposon insertions in cancer predisposition genes. *Cancer Genetics* **216–217**, 159–169 (2017).
118. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat Genet* **49**, 692–699 (2017).
119. Ganel, L., Abel, H. J. & Hall, I. M. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33**, 1083–1085 (2017).
120. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
121. Sharo, A. G., Hu, Z. & Brenner, S. E. StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants. *bioRxiv* 2020.05.15.097048 (2020) doi:10.1101/2020.05.15.097048.
122. Huynh, L. & Hormozdiari, F. TAD fusion score: discovery and ranking the contribution of deletions to genome structure. *Genome Biology* **20**, 60 (2019).
123. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144–161 (2013).
124. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
125. Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods* 1–7 (2020) doi:10.1038/s41592-020-0960-3.
126. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
127. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
128. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
129. Munroe, P. B. *et al.* Spectrum of mutations in the Batten disease gene, CLN3. *Am J Hum Genet* **61**, 310–316 (1997).
130. Kim, B. J. *et al.* Clarification of glycosylphosphatidylinositol anchorage of OTOANCORIN and human OTOA variants associated with deafness. *Hum Mutat* **40**, 525–531 (2019).
131. Chopra, M. *et al.* Heterozygous ANKRD17 loss-of-function variants cause a syndrome with intellectual disability, speech delay, and dysmorphism. *Am J Hum Genet* **108**, 1138–1150 (2021).
132. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
133. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics* 1–8 (2021) doi:10.1038/s41588-021-00865-4.
134. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications* **10**, 3240 (2019).
135. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data* **4**, 170107 (2017).
136. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci Rep* **8**, 1362 (2018).

137. Hartley, T. *et al.* The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **178**, 458–463 (2018).
138. Kumar, S., Harmanci, A., Vytheeswaran, J. & Gerstein, M. B. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* **21**, 274 (2020).
139. Nichols, M. H. & Corces, V. G. A CTCF Code for 3D Genome Architecture. *Cell* **162**, 703–705 (2015).
140. Lobanenkov, V. V. *et al.* A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* **5**, 1743–1753 (1990).
141. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
142. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A* **115**, E6697–E6706 (2018).
143. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**, 3573–3599 (2017).
144. Banks, W. E. *et al.* Neanderthal Extinction by Competitive Exclusion. *PLoS One* **3**, e3972 (2008).
145. Ponce de León, M. S. *et al.* Neanderthal brain size at birth provides insights into the evolution of human life history. *Proceedings of the National Academy of Sciences* **105**, 13764–13768 (2008).
146. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* **119**, 1027–1040 (2004).
147. Trujillo, C. A. *et al.* Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science* **371**, eaax2537 (2021).
148. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLOS Genetics* **3**, e7 (2007).
149. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
150. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222–226 (2012).
151. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **33**, D29–D33 (2005).
152. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
153. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
154. Abugessaisa, I. *et al.* FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Research* **49**, D892–D898 (2021).
155. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J Neurosci* **34**, 1420–1431 (2014).
156. Schwartz, P. H. *et al.* Isolation and characterization of neural progenitor cells from post-mortem human cortex. *J Neurosci Res* **74**, 838–851 (2003).
157. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res* **38**, 4218–4230 (2010).
158. Hofman, M. Evolution of the human brain: when bigger is better. *Frontiers in Neuroanatomy* **8**, (2014).
159. Lee, S.-W. *et al.* AKAP6 inhibition impairs myoblast differentiation and muscle regeneration: Positive loop between AKAP6 and myogenin. *Sci Rep* **5**, 16523 (2015).

160. Smeland, O. B. *et al.* Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function. *JAMA Psychiatry* **74**, 1065–1075 (2017).
161. Ohkuni, A., Ohno, Y. & Kihara, A. Identification of acyl-CoA synthetases involved in the mammalian sphingosine 1-phosphate metabolic pathway. *Biochemical and Biophysical Research Communications* **442**, 195–201 (2013).
162. Nakahara, K. *et al.* The Sjögren-Larsson syndrome gene encodes a hexadecenal dehydrogenase of the sphingosine 1-phosphate degradation pathway. *Mol Cell* **46**, 461–471 (2012).
163. Brockington, A. *et al.* Downregulation of genes with a function in axon outgrowth and synapse formation in motor neurones of the VEGFδ/δ mouse model of amyotrophic lateral sclerosis. *BMC Genomics* **11**, 203 (2010).
164. Golanska, E. *et al.* Analysis of APBB2 gene polymorphisms in sporadic Alzheimer's disease. *Neurosci Lett* **447**, 164–166 (2008).
165. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).
166. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
167. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**, e1003711 (2014).
168. Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & Molecular Medicine* **50**, 1–8 (2018).
169. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
170. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-Specific Variants and the Limits of Human Genetics. *Trends Mol Med* **22**, 925–934 (2016).
171. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 (2021) doi:10.1038/s41586-021-03819-2.
172. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **8**, 53 (2021).
173. Boeke, J. D. *et al.* The Genome Project-Write. *Science* **353**, 126–127 (2016).
174. Papaspyropoulos, A., Tsolaki, M., Foroglou, N. & Pantazaki, A. A. Modeling and Targeting Alzheimer's Disease With Organoids. *Front Pharmacol* **11**, 396 (2020).
175. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* doi:10.1093/nar/gkaa942.