UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI

Informatica

SCUOLA DI DOTTORATO DI

Scienze Naturali ed Ingegneristiche

DOTTORATO DI RICERCA IN

Informatica

Con il contributo di (ENTE FINANZIATORE)

Università degli Studi di Verona

CICLO /ANNO (1° anno d'Iscrizione) XXXVI/2020

TITOLO DELLA TESI DI DOTTORATO

Advancing Precision Medicine: Assessing Genetic Diversity Impact on Regulatory Elements and Genome Editing Outcomes

S.S.D. INF/01
(indicare il settore scientifico disciplinare di riferimento della tesi dato obbligatorio)*

Coordinatore: Prof./ssa Ferdinando Cicalese

Firma _____

Tutor: Prof./ssa Rosalba Giugno

Firma _____

Dottorando: Dott./ssa Manuel Tognon

Firma _____

* Per l'elenco dei Settori Scientifico-Disciplinari (SSD) si veda il D.M. del 4 Ottobre 2000, Allegato A *"Elenco dei Settori Scientifico –Disciplinari"* reperibile sul sito del Ministero dell'Università e della Ricerca al seguente indirizzo: http://www.miur.it/atti/2000/alladm001004_01.htm

Abstract

Over the past decade, the landscape of precision medicine has experienced a notable transformation, driven by the intersection of omics sciences and computational methodologies. This convergence has empowered researchers to explore individual-specific biological markers more deeply, offering the potential for tailored treatments. Central to this shift are genetic variants, which can profoundly influence the genomic environment by modifying regulatory elements. However, fully leveraging the potential of genetic variants in precision medicine necessitates accurate computational tools capable of deciphering their impact on cellular mechanisms. In response to this need, this PhD thesis introduces two innovative computational methods: GRAFIMO and MotifRaptor. GRAFIMO utilizes genome graph data structures to identify potential transcription factor binding sites while considering individual- and population-specific genetic variants. By accounting for various genetic events, from single nucleotide variants to complex structural variants, GRAFIMO provides a comprehensive analysis of transcription factor binding. Similarly, MotifRaptor adopts a Transcription Factor-centric approach to annotate the potential functional impact of non-coding variants, enhancing our understanding of regulatory mechanisms underlying complex traits and diseases. CRISPR genome editing holds promise for programmable editing of genomic sequences in therapeutic settings. However, its application hinges on two critical factors: accurately quantifying editing outcomes, and predicting and mitigating off-target effects. Addressing the former challenge, we benchmarked different methods used to assess genome editing outcomes across various cellular contexts, providing valuable insights into their efficacy and reliability. CRISPRme, on the other hand, represents a pioneering tool for variant- and haplotype-aware CRISPR off-target nomination. By considering genetic diversity across populations, CRISPRme facilitates thorough off-target analysis, potentially enabling safer and more precise genome editing interventions. These computational methods offer significant insights into exploiting genetic diversity for potential individual-oriented therapies, potentially contributing to the development of precision medicine-oriented approaches and therapies.

Acknowledgements

I am filled with immense gratitude towards those who have supported and guided during my PhD journey.

First and foremost, I would like to express my deepest gratitude to my Supervisors, Prof. Rosalba Giugno and Prof. Luca Pinello, for their unwavering support, guidance, and mentorship throughout this research journey. Your expertise, encouragement, and dedication have been invaluable in shaping my scientific endeavors and navigating the challenges of academic research.

I extend my heartfelt appreciation to my esteemed colleagues from both Infomics lab and Pinello lab. Your collaborative spirit, insightful discussions, and camaraderie have enriched my research experience and contributed significantly to the progress of this work. The collaborative environment fostered within the labs has been instrumental in my growth as a researcher.

To my beloved family, especially my parents, Franco and Lorella, and brother Luca, I owe an immeasurable debt of gratitude. Your boundless love, encouragement, and unwavering belief in my aspirations have been the cornerstone of my academic and personal growth throughout my PhD. Your sacrifices and enduring support have propelled me forward, and for that, I am profoundly grateful.

I would also like to extend my thanks to all the individuals with whom I had the privilege to collaborate on the papers presented in this PhD thesis, in particular to Dr. Erik Garrison and Dr. Vincenzo Bonnici. Your expertise, dedication, and collaboration have been instrumental in the successful completion of this work, and I am grateful for the opportunity to have worked alongside you.

Additionally, I am grateful for the unwavering support and camaraderie of my friends. Your encouragement, laughter, and companionship have provided solace during challenging times and made the journey more enjoyable. Your presence has been a constant source of motivation and joy throughout my journey.

Lastly, I extend my appreciation to all those who have supported and encouraged me along this journey, whether through discussions, feedback, or moral support. Your contributions have not gone unnoticed and have played a significant role in shaping this research endeavor.

Thank you all for your unwavering support and encouragement throughout my journey.

Contents

List of Figures

List of Tables

Introduction

During the last decade, omics sciences have swiftly emerged as fundamental tools for informing medical decisions, serving as the foundational pillars supporting precision medicine (Ginsburg and Willard, 2009). Precision medicine is a rapidly progressing healthcare approach using individual-specific clinical, genetic, genomic, environmental and social information (Ginsburg and Willard, 2009) to develop individual-tailored treatments. Although precision medicine encompasses a multidisciplinary approach, it largely relies on omics sciences to guide and enhance medical strategies and therapies. This technological progress had a profound impact, drastically reducing the costs associated with analyses involving omics data. Furthermore it empowered the accumulation of large datasets, even for individual patients (Voelkerding *et al.*, 2009). Indeed, understanding the genetic basis of diseases is generally expected to lead to better characterized and tailored therapies (Ashley, 2016). The evolution and enhancements in sequencing technologies brought about a profound transformation, substantially improving both the quantity and quality of available omics data. This facilitated the discovery of previously unknown causative genes (Ng *et al.*, 2009) and wider usage of omics data in medical decision-making processes (Ashley *et al.*, 2010; Worthey *et al.*, 2011). Collecting individual-specific omics data serves as a priceless asset for capturing the distinctive biological attributes, or *biomarkers*, that define individuals and even offer insights into potential medical conditions, including diseases. The discovery of individual-specific biomarkers potentially causing medical conditions is the cornerstone of precision medicine. Among these biomarkers, genetic variants hold a pivotal position (Raphael *et al.*, 2014). Genetic variants are differences in the DNA sequence when compared to a standard or reference sequence. They can occur naturally as mutations or alterations in the DNA sequence. They arise due to errors during different biological processes, such as DNA replication, *mitosis* or *meiosis*, or due to other DNA damages. The resulting mutations may undergo error-prone repair processes, particularly microhomology-mediated end joining (Sinha *et al.*, 2017; Seol *et al.*, 2018), leading to errors during other repair or replication mechanisms (Rodgers and McVey, 2016). Moreover, mutations in DNA sequences can even occur through the insertion or deletion of short DNA segments (indels), or large fragments consisting of thousands of nucleotides (structural variants). Genetic variants are common and fundamental aspects characterizing genetic diversity among individuals and populations (Siva, 2008). Some variants may have no noticeable effect on an individual's traits or health, while others can contribute to differences in susceptibility to diseases, response to medications, or various physical characteristics (Bodmer and Bonilla, 2008; Ingelman-Sundberg *et al.*, 2018; Mitchell-Olds *et al.*, 2007). Genetic variants may be located either within genes (coding regions), or in other genomic regions (non-coding regions). Historically, before the molecular causes were known, human genetics and clinical genetics studies focused on analyzing family pedigrees and inheritance descriptions to understand the occurrence of certain traits between single or groups of individuals. The advent of sequencing technologies allowed researchers to study and discover the molecular mechanisms underlying the occurrence of traits or diseases in single individuals or populations. Furthermore, the recent advances in sequencing technologies significantly increased the throughput of sequencing experiments, providing a huge amount of data. In this context, many international efforts, such as the 1000 Genomes Project (Siva, 2008; Consortium *et al.*, 2015) and the Human Genetic Diversity Project (Cavalli-Sforza, 2005; Bergström *et al.*, 2020), emerged to better study and understand the genetic diversity between human subjects and populations. Similarly, other consortia (Sherry *et al.*, 2001; Landrum *et al.*, 2020) collected variants data to discover potential associations between variants and traits or diseases. Simultaneously, several computational methods, such as variant calling (McKenna *et al.*, 2010) pipelines or genome-wide association studies (GWAS) (Uffelmann *et al.*, 2021) to analyze and interpret variants-related data were proposed within the scientific community. However, the huge amount of data proved itself difficult to be fully exploited by using the common reference genome sequences. Reference genomes are commonly represented as strings, reporting the DNA sequence reconstructed from raw sequencing data. However, such representation cannot include the genetic variability and diversity present in genomes. Recently, the in-

troduction of advanced data structures known as genome graphs (Paten *et al.*, 2017; Garrison *et al.*, 2018) has revolutionized the representation of genomes and genetic diversity across individuals and populations. Genome graphs, essentially, are graph-based data structures, that employ nodes to represent DNA sequences and edges to represent links between successive sequences. Paths within genome graphs, often labeled (e.g., in reference genomes), denote haplotypes associated with different genomes (Sirén *et al.*, 2020). This data structure not only integrates reference genome sequences with variants but also overcomes limitations inherent in traditional methods working on *linear* reference genomes. In fact, genome graphs can accommodate both simple variant events, such as SNVs and indels, across multiple individuals, and complex events, such as SNVs-indels combinations and large structural variants. Despite the several methods and frameworks for constructing and representing genome graphs (Andreace *et al.*, 2023), these structures have proven immensely beneficial in enhancing various genetic data analysis pipelines. For instance, they improve read mapping accuracy (Sibbesen *et al.*, 2023), enable more precise variant calling incorporating SNPs, indels, and structural variants (Garrison *et al.*, 2018; Ebler *et al.*, 2022), and facilitate the capture of genome-wide epigenetic marks within genomic regulatory elements (GREs) (Groza *et al.*, 2020; Liao *et al.*, 2023). Of particular note is the recent release of the first draft of the human pangenome (Liao *et al.*, 2023), which holds great promise for the widespread adoption of genome graph frameworks to explore and represent the intricacies of genetic variability. Consequently, the development of additional software and pipelines to exploit the potential of genome graphs in genetic data analysis offers significant promise in advancing precision medicine-oriented approaches (Yu and Chen, 2023). As mentioned, genome graphs have emerged as powerful tools for analyzing epigenetic marks within GREs, offering a comprehensive representation of genetic diversity introduced by non-coding variants. Genetic variants within non-coding regions, notably within GREs, are key players in shaping gene expression and are linked to various traits and disease susceptibilities (Weinhold *et al.*, 2014; Wienert *et al.*, 2015). These variants have the potential to perturb the intricate networks governing gene expression by altering GRE sequences and functions. Gene expression is finely regulated by sequence-specific protein interactions, mediated by histones and transcription factors (TFs) (Lambert *et al.*, 2018), which modulate chromatin accessibility and compaction. Histone modifications and DNA methylation further regulate gene expression by influencing DNA-histone interactions and recruiting chromatin remodeling complexes. TFs specifically target cis-regulatory sequences within GREs, such as enhancers and silencers, exerting their effects on gene expression even when distal from gene promoters (Lambert *et al.*, 2018). Notably, enhancers collaborate closely with TFs to fine-tune gene expression. Understanding how genetic variants impact these genomic elements, in particular TF binding sites within GREs, is crucial for deciphering the implications of individual and population-specific genetic diversity in cellular environments. CRISPR genome editing (Cong *et al.*, 2013) offers promising avenues for the development of innovative therapeutics by precisely modifying genetic or epigenetic elements, including GREs, within specific genomic regions. In particular, CRISPR genome editing holds great promises precision medicine applications, through a precise and accurate personalized editing of potentially malignant genome sequence segments. Accurate design of guide RNAs (gRNAs) to target intended sequences (referred to as *on-targets*) is crucial to prevent unintended and potentially harmful edits in non-targeted sequences, known as *off-targets* (Pattanayak *et al.*, 2013; Cho *et al.*, 2014). Off-target prediction is essential for guiding CRISPR experiment design, considering various sequence characteristics and the influence of genetic diversity on potential off-target site creation (Scott and Zhang, 2017). In fact, genetic variants can enhance the affinity of previously unbound sequences by the gRNA and may introduce novel Protospacer Adjacent Motifs (PAMs), potentially posing risks to the cellular environment. While computational methods have been extensively developed to predict off-target sites during gRNA design (Hanna and Doench, 2020), most operate exclusively on the reference genome. Alternative approaches have been proposed to analyze off-target potential while accounting for genetic diversity. However, these methods often face limitations, such as inefficient handling of large datasets or accommodating alternative haplotypes and indels. Additionally, they may require advanced computational skills, restricting their accessibility to a broader audience. The adoption of precision medicine-oriented approaches in experimental and clinical settings faces significant hurdles, particularly in accurately predicting and accommodating the impact of individual- and population-specific genetic diversity. In this PhD thesis, we address two specific challenges critical to realizing the shift toward precision medicine-oriented approaches (**Figure 1.1**): (i) understanding the influence of genetic diversity on transcription factor binding landscapes, and (ii) assessing the effects of genetic variants on CRISPR genome editing outcomes (**Table 1.1**). To tackle the first challenge, GRAFIMO (Tognon *et al.*, 2021) and MotifRaptor (Yao *et al.*, 2021) offer innovative solutions (**Figure 1.1 (B-C)**). GRAFIMO, a novel Transcription Factor Binding Site scanning tool, leverages genome graph data structures to identify potential binding sites with remarkable accuracy (**Figure 1.1 (C)**).

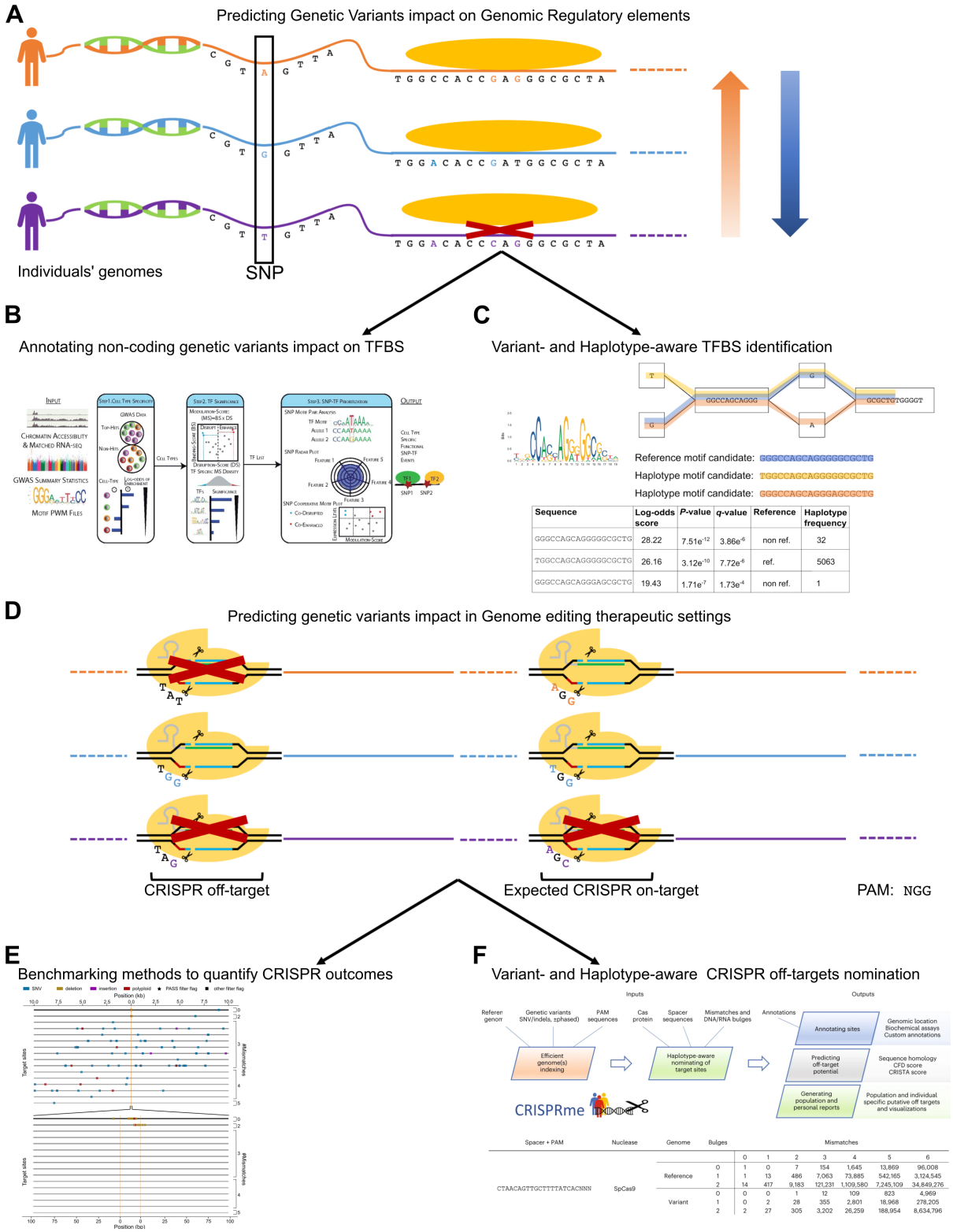


Figure 1.1. Overview of challenges and topics discussed throughout the PhD thesis. **(A)** Genetic variants may alter transcription factor binding affinity, necessitating consideration of individual haplotypes for accurate assessment. **(B)** MotifRaptor (**Chapter 6**) introduces a novel Transcription Factor-centric approach for annotating non-coding variants from GWAS analyses. It evaluates tissue-specific effects on transcription factor binding, providing insights into regulatory mechanisms. **(C)** GRAFIMO (**Chapter 5**) utilizes genome graph data structures to identify Transcription Factor Binding Sites, accounting for simple and complex genetic events and offering a comprehensive analysis of binding. **(D)** Genetic variants can also impact CRISPR genome editing outcomes by altering on-targets and introducing novel off-targets. **(E)** Benchmarking widely-used genome editing tools (**Chapter 8**) aims to assess their accuracy across different cellular contexts. **(F)** CRISPRme (**Chapter 9**) introduces a novel variant- and haplotype-aware, and user-friendly tool for evaluating CRISPR off-target potential, considering genetic diversity across populations.

Challenge	Proposal	Aim
Predict genetic variants impact on genomic regulatory elements to guide experimental analyses	GRAFIMO	A novel Transcription Factor Binding Site scanning tool that takes advantage of genome graph data structures. GRAFIMO effectively identifies potential binding sites while accounting for both simple (SNVs and indels) and complex (structural variants) genetic events. By considering individual and population-specific genetic landscapes, it provides a more comprehensive analysis of transcription factor binding.
	MotifRaptor	A novel Transcription Factor-centric approach designed to annotate non-coding variants identified through GWAS analyses. This approach evaluates the tissue- and cell type-specific effects of genetic diversity on the landscape of transcription factor binding. By scrutinizing whether genetic variants enhance, disrupt, or exhibit dual effects on the binding potential of investigated transcription factors, it provides a nuanced understanding of regulatory mechanisms underlying complex traits and diseases.
Predict the impact of genetic variants in genome editing therapeutic settings	WGS Editing quantification	Assessement and quantification of genome editing outcomes comparing four extensively used tools: CRISPResso, Mutect2, Strelka2, and Varscan2. This comparison utilizes three guide RNAs with potential future therapeutic applications across various cell lines. The aim is to determine whether these tools may over- or under-estimate editing outcomes. By examining these tools across different cellular contexts, we gain valuable insights into their efficacy and reliability in predicting and analyzing genome editing results.
	CRISPRme	A novel variant- and haplotype-aware CRISPR off-target nomination tool. Designed to be user-friendly, CRISPRme offers a comprehensive framework for evaluating the off-target potential of user-defined guide RNAs (gRNAs). By considering genetic diversity across individuals and populations, CRISPRme ensures thorough off-target analysis across the genome.

Table 1.1. Novel tools for variant- and haplotype-aware genomic regulatory element analysis and genome editing assessment. Overview of proposed tools addressing challenges in genomic regulatory element analysis and genome editing assessment. GRAFIMO and MotifRaptor provide solutions for predicting genetic variants' impact on transcription factor binding sites, while WGS Editing Quantification and CRISPRme offer approaches for evaluating genome editing outcomes and off-target effects, respectively.

By accounting for both simple (SNVs and indels) and complex (structural variants) genetic events and considering individual and population-specific genetic landscapes, GRAFIMO provides a comprehensive analysis of transcription factor binding. Similarly, MotifRaptor introduces a Transcription Factor-centric approach, enabling the annotation of non-coding variants identified through GWAS analyses (**Figure 1.1 (B)**). By assessing tissue- and cell type-specific effects of genetic diversity on transcription factor binding, MotifRaptor enhances our understanding of regulatory mechanisms underlying complex traits and diseases. To tackle the second challenge, we propose two studies aimed at evaluating widely used methods to quantify CRISPR genome editing outcomes on whole genome sequencing data, and to introduce a novel haplotype- and variant-aware computational method to nominate potential CRISPR off-targets (**Figure 1.1 (E-F)**). Our benchmark on methods to quantify CRISPR genome editing outcomes compares four widely adopted tools on high-depth whole genome sequencing data: CRISPResso (Clement *et al.*, 2019), Mutect2 (McKenna *et al.*, 2010), Strelka2 (Kim *et al.*, 2018), and Varscan2 (Koboldt *et al.*, 2012) (**Figure 1.1 (E)**). This comparison, conducted across various cell lines using three guide RNAs with potential therapeutic applications, aims to determine whether these tools may under- or over-estimate editing outcome with a genome-wide focus. By examining these tools' performance across different cellular contexts, valuable insights are gained into their efficacy and reliability in predicting and analyzing genome editing results. Notably, the correct quantification and estimation of editing outcome is a key aspect to enable potential applications of CRISPR genome editing in precision medicine contexts. Furthermore, CRISPRme (Cancellieri *et al.*, 2023) proposes a novel variant- and haplotype-aware CRISPR off-target nomination tool (**Figure 1.1 (F)**). Designed to be user-friendly, CRISPRme offers a comprehensive framework for evaluating the off-target potential of user-defined guide RNAs. By considering genetic diversity across individuals and populations, CRISPRme ensures thorough off-target analysis across the genome, thus facilitating safer and more precise genome editing interventions. The subsequent chapters offer a comprehensive view of the biological challenges we addressed and our proposals to solve them. In **Chapter 2**, we delve into the fundamental concept of genetic variants, elucidating their molecular mechanisms and nomenclature. **Chapter 3** expands upon the genome graph data structure, showcasing its pivotal role in representing population-wide genetic diversities. Moving forward to **Chapter 4**, we shine a spotlight on genomic regulatory elements (GREs), particularly focusing on Transcription Factors and their binding mechanisms. Additionally, this chapter discusses existing methods for analyzing transcription factor-related data. **Chapter 5** introduces GRAFIMO, a powerful tool designed to analyze transcription factor binding sites, providing insights into its functionality and significance. Building upon this, **Chapter 6** unveils MotifRaptor, offering a deeper dive into its algorithms, recent applications in studies, and potential future enhancements. **Chapter 7** shifts gears to explore the molecular principles underlying CRISPR genome editing, along with the current computational analyses applicable to CRISPR genome

editing data. In **Chapter 8**, we provide insights into the WGS editing quantification project, outlining ongoing analyses and preliminary results. **Chapter 9** brings forth CRISPRme, a groundbreaking tool characterized by its computational innovations and experimental findings derived from its application. Lastly, **Chapter 10** ties everything together, offering conclusions and discussions regarding the results and tools presented throughout the PhD thesis.

Genetic Variants

Genetic variants are modifications occurring in the DNA sequence within the genome of an organism. Mutations may result from errors during DNA replication processes, *mitosis*, *meiosis*, or other forms of DNA damage (such as the formation of pyrimidine dimers induced by exposure to ultraviolet radiation). In this discussion, we treat genetic variants and mutations as synonymous terms. Mutations may undergo error-prone repair processes, such as microhomology-mediated end joining (MEMJ) (Sinha *et al.*, 2017; Seol *et al.*, 2018), resulting in errors during repair and replication of the original DNA sequence (Rodgers and McVey, 2016). Mutations may consist in substitutions of single nucleotide in the sequence, as well as in insertions or deletions of DNA segments. The different versions of the same variant are referred to as *alleles*. We define the variant observed in the reference (original) DNA sequence *reference allele*, and any other distinct version from the reference is called *alternative allele*. The impact of mutations on the phenotype depends on several factors, ranging from the organism's genomic environment to the social environment in which the organism is located. While some genetic variants may not impact the phenotype, others may introduce differences in the organism's susceptibility to disease, response to drugs and treatments, or physical characteristics (Bodmer and Bonilla, 2008; Ingelman-Sundberg *et al.*, 2018; Mitchell-Olds *et al.*, 2007). Furthermore, mutations furnish the raw material upon which evolutionary forces, such as natural selection, can act (Akey *et al.*, 2004; Teotónio *et al.*, 2009). Genetic variants may occur either within gene sequences (coding regions), or outside genes (non-coding regions). Mutations occurring in coding regions may have either advantageous or deleterious consequences on the organism's phenotype, by altering the gene product. Although variants in non-coding regions do not alter gene sequences, they still may impact gene products and the organism's phenotype (Zhang and Lupski, 2015).

2.1 Genetic Variants Genesis

DNA may undergo naturally occurring or artificially induced modifications resulting in mutations. Mutations can occur spontaneously in nature, as a result of exposure to chemical or physical agents increasing the mutation rate in DNA sequences called *mutagens*, or induced experimentally using laboratory protocols (e.g. CRISPR genome editing (Cong *et al.*, 2013)). Variants can be categorized into four classes depending on the process that induced the emergence of the mutations: (i) spontaneous mutations, (ii) error-prone translesion synthesis, (iii) errors introduced during DNA repair processes, and (iv) induced mutations resulting from exposure to mutagens. Spontaneous mutations occur randomly across the genome sequence even in healthy cells. Spontaneous mutations are caused by different random events, such as incorrect base pairing during DNA replication caused by tautomerism (Podolyan *et al.*, 2003), slipped-strand mispairing during replication (Levinson and Gutman, 1987), deamination (Duncan and Miller, 1980), or depurination (Kunkel, 1984). Some spontaneous mutations are induced by error-prone replication mechanisms (translesion synthesis), acting on damaged DNA. While naturally occurring damages to the DNA sequence, such as double-strand breaks, exhibit low frequency, their repair processes often introduce mutations. A prominent pathway for repairing double-strand breaks is non-homologous end joining (NHEJ) (Weterings and Chen, 2008). In NHEJ, a few nucleotides are removed to facilitate the imprecise alignment alignment of the broken ends for rejoining, and subsequently, additional nucleotides are added to fill the remaining gaps. However, as consequence NHEJ introduces insertions or deletions into the original DNA sequence (Weterings and Chen, 2008). Finally, mutations may be induced by the exposure of DNA sequences to mutagens, such as chemical (nitrous acid, hydroxylamine, alkylating agents, etc.) or radiation (UV light, ionizing radiation) agents. Importantly, some mutagens are commonly found in the environment, and organisms are normally exposed to their action.

2.2 Classifying Genetic Variants

Over the years, various methods have been proposed to classify genetic variants based on their inheritance patterns, the modifications they induce in the DNA structure, their location within the genome, and their impact on gene products.

2.2.1 Classifying Genetic Variants by their inheritance

In multicellular organisms, variants can be categorized in *germline* mutations, which can be inherited by descendants, and *somatic* mutations, which generally are not transmitted to descendants. Diploid organisms, such as human, possess two copies of each chromosome, one inherited from the father, the other from the mother. Depending on the occurrence of mutations on each chromosome, variants can be classified in three types: (i) heterozygous mutations, when only one copy of the chromosomes carries the alternative allele, (ii) homozygous mutations, when both copies carry the same alternative allele, and (iii) compound mutations, when the two copies carry different alternative alleles. Germline mutations occurring in the reproductive cells of an individual result in constitutional mutations in the offspring, spread in every organism's cell. Importantly, germline mutations have the potential to be transmitted through successive generations. A newly occurring germline mutation, not inherited from parents is defined *de novo* mutation. Somatic mutations do not affect the germline and are generally not transmitted to descendants. However, they are inherited by all the progeny derived from a mutated cell within the same organism during mitosis. Somatic mutations are often linked to deleterious consequences for the cell, such as cancer (Martincorena and Campbell, 2015).

2.2.2 Classifying Genetic Variants by their impact on DNA sequence structure

Genetic variants can be classified into three classes based on the structural modification they induce to the DNA sequence: (i) single nucleotide variants (SNVs), (ii) insertions and deletions (indels), and (iii) structural variants.

Single Nucleotide Variants

A single nucleotide variant (SNV) denotes a germline substitution involving a single base at a precise position within the genome. For example a `G` occurring at a specific genome location may be replaced by an `A`. In this example the two alleles (nucleotide variations) are `G` (reference allele) and `A` (alternative allele). SNVs exhibits an allele frequency $< 1\%$. When single nucleotide substitutions show a frequency $\geq 1\%$ in the general population they are defined Single Nucleotide Polymorphisms (SNPs). The human genome within the global population has revealed the presence of over 600 million SNPs, which occur more frequently in non-coding regions than in coding. Human populations exhibit variations, leading to the prevalence of certain SNP alleles in specific geographic or ethnic groups while being comparatively rare in others. However, this variation is not frequent on a global scale (Bergström *et al.*, 2020). Within a population, SNPs can be characterized by their minor allele frequency (MAF). MAF is defined as the lowest frequency of an allele at a locus observed in a particular population. This frequency represents the lesser of the two allele frequencies for single-nucleotide polymorphisms.

Indels

Indel denote the insertion or deletion of nucleotides within a genome. Indels with length ≥ 50 bp are often referred to as structural variants. Indels are used as genetic markers in natural populations, particularly in phylogenetic studies (Väli *et al.*, 2008). Between 16% and 25% of all sequence polymorphisms in humans are likely to be represented by indels (Mills *et al.*, 2006). In most deeply studied genomes, as the human genome, indels frequency is notably lower compared to that of SNPs, except in the vicinity of highly repetitive regions, such as homopolymers and microsatellites.

Structural Variants

Structural variants refer to diverse alterations in the chromosome structure of an organism. Historically, structural variants were defined to affect sequences of length between $\sim$1 Kb to 3 Mb (Feuk *et al.*, 2006). However, recently their range has been expanded to consider events affecting sequences > 50 bp

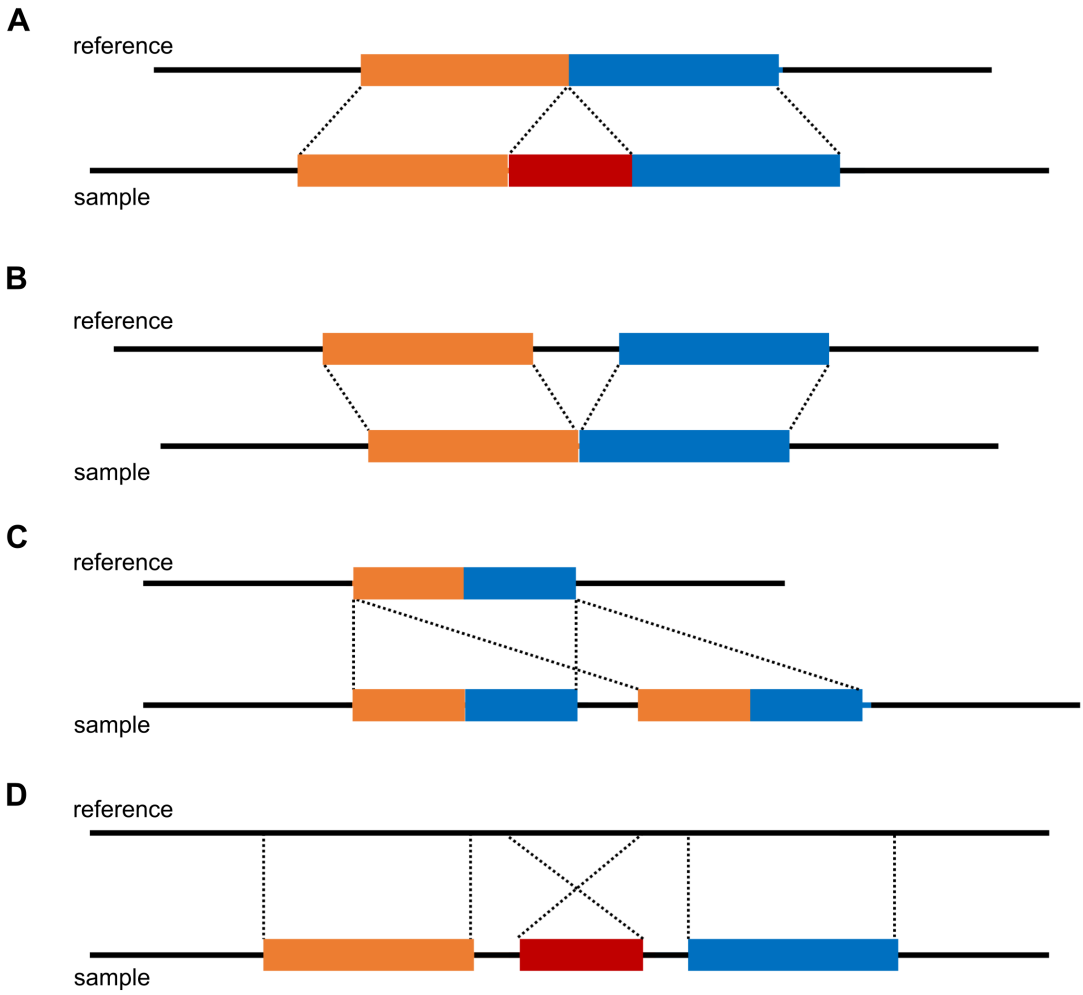


Figure 2.2. Structural variants: large alterations in the chromosomal structure. (A) Insertions insert large DNA segments within the genome sequence. **(B)** Large deletions remove large genomic segments from the DNA. **(C)** Duplications duplicate large portions of DNA sequences, potentially duplicating genes and regulatory elements. **(D)** Inversions flip large segments of DNA, potentially flipping or even break genes or regulatory elements sequences.

(Alkan *et al.*, 2011). Approximately 13% of the human genome is structurally variant in the normal population (Sudmant *et al.*, 2015). Furthermore, there are several evidences suggesting that structural variants can encompass millions of nucleotides of heterogeneity within each genome, potentially playing a pivotal role in human diversity and susceptibility to diseases. Many structural variants are linked to diseases (Lakich *et al.*, 1993). Structural variants encompass several types of variations, such as inversions, copy-number variants (CNVs), translocations, deletions, insertions, and duplications (**Figure 2.2**). Insertions, deletions, and duplications are particular cases of CNVs (**Figure 2.2 (A-C)**). Inversions are chromosomal rearrangements wherein a chromosomal segment undergoes an inversion, flipping its original position (**Figure 2.2 (D)**). Inversions occur when a chromosome experiences two breaks along its arm, and the DNA segment between the breakpoints inserts itself in the opposite direction, within the same chromosome arm. Generally, breakpoints occur in regions rich of repetitive nucleotides. Importantly, inversions may flip or even break genes or regulatory elements sequences (Puig *et al.*, 2015). Inversions can occur through different mechanisms, such as chromosomal breakage and repair, or non-homologous end joining. There are two types of inversions: paracentric inversions, which do not involve chromosome's centromere, and pericentric, which span the centromere and have breakpoints in each chromosome's arm. In CNVs sections of an organism's genome are duplicated or deleted, resulting in variations in the number of repeated nucleotide sequences among individuals (McCarroll and Altshuler, 2007; Alkan *et al.*, 2011). Therefore, this structural variant involves insertions or deletions >50 bp in the genome. CNVs represent a fundamental aspect of genomic diversity, encompassing $4.8 - 9.5\%$ of the genome in human (Zarrei *et al.*, 2015). Translocations, instead, are structural variants leading to atypical rearrangements of chromosomes. There exist two types of translocations: reciprocal and Robertsonian translocations. The former occurs when non-homologous chromosomes exchange DNA fragments between each other.

Robertsonian translocations, instead, occur when two non-homologous chromosomes become attached. Importantly, translocations can result in the fusion of two genes(*gene fusion*) or regulatory elements, otherwise separate. Translocations may be balanced, when the exchange of genetic material is even, or unbalanced. The latter may result in extra or missing genes or regulatory elements on the donor/receiver chromosome.

2.2.3 Classifying Genetic Variants by their location

Another method to distinguish genetic variants is based on their location across the genome. *Coding* genetic variants occur within the coding sequence of genes (exons), while *non-coding* variants occur in non-coding sequences of genes (introns) or intergenic regions. While coding variants may directly impact the gene product, non-coding variants may still affect gene splicing, gene expression levels, or the gene regulatory mechanisms perturbing the cell environment Li *et al.* (2016); Maurano *et al.* (2015). In particular, non-coding variants occurring within large and small regulatory elements, such as enhancers or transcription factor binding sites, have been shown to significantly impact gene expression levels and regulation (De Gobbi *et al.*, 2006; Wienert *et al.*, 2015), and have been linked to increased susceptibility to some diseases, such as cancer (Weinhold *et al.*, 2014). Coding variants may affect the gene product, potentially modifying the sequence of the resulting protein. The impact that such mutations may have on the gene product provides another method to classify genetic variants.

2.2.4 Classifying Genetic Variants by their impact on gene product

Coding variants may or may not impact the gene product. Considering SNVs, in the former case variants are defined *synonymous substitutions*, in the latter *nonsynonymous substitutions*. Despite the different nucleotide, synonymous substitutions do not change the amino acid encoded by the affected codon. Nonsynonymous substitutions, instead, alter the gene product either changing the amino acid encoded by the affected codon (*missense*), or inserting a premature stop codon (*nonsense*). Indels, in contrast, if their length is a multiple of 3, they can either eliminate or add one or more amino acids to the gene product. However, if their length is not a multiple of 3, they may induce a reading frame shift, potentially leading to the disruption of the gene product or the introduction of a premature stop. Structural variants are excluded from this classification.

2.2.5 Combining variants: the concept of Haplotype

In organisms that inherit genetic material from both parents, the genetic material is typically organized into two sets of paired chromosomes, with one set coming from each parent. *Haplotypes* are clusters of alleles inherited together from a single parent. A diploid set encompasses pairs of chromosomes, whereas a haploid set comprises only one half of each pair. The haploid genotype focuses on individual chromosomes rather than pairs, encompassing either all chromosomes from one parent or even a specific segment. The term haplotype has multiple applications. It refers to a collection of specific alleles clustered within tightly linked genes on a chromosome, likely to be inherited together and preserved across generations. It also indicates sets of linked SNP alleles statistically associated and likely to occur together. Identifying these associations, along with a few alleles of a specific haplotype sequence, is often used to locate other nearby polymorphic sites on the chromosome. Importantly, haplotypes are crucial to investigate the genetics of common diseases and traits, and their occurrence across different human groups or populations. The crucial importance of haplotypes is highlighted by international efforts such as the International HapMap Project (Gibbs *et al.*, 2003), whose aim was to develop a haplotype map of the human genome, describing common patterns of human genetic variation and diversity.

2.3 Interpreting and understanding Genetic Variants impact

Advancements in sequencing technologies have revolutionized the accumulation of extensive data on genetic variants and genomic diversity, with a primary focus on the human genome. Notably, international initiatives like the 1000 Genomes Project (Siva, 2008; Consortium *et al.*, 2015) and the Human Genome Diversity Project (Bergström *et al.*, 2020) have diligently worked towards capturing the spectrum of genetic diversity across individuals and populations. A wealth of experimentally validated data is stored in various databases, such as dbSNP (Sherry *et al.*, 2001), ClinVar (Landrum *et al.*, 2020) or COSMIC (Bamford *et al.*, 2004), linking known variants to genetic traits or diseases. These resources are

fundamental tools to understand and interpret the potential impact of genetic variants on the genetic landscape. Given the intricate nature of sequencing data, several computational pipelines have emerged to sift through this vast information, identifying variants from raw sequencing data. Variant calling tools, such as the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010), Strelka (Kim *et al.*, 2018), Varscan (Koboldt *et al.*, 2012) or FreeBayes (Garrison and Marth, 2012), analyze sequencing data to unveil the variants present within an individual's genome, for example. Simultaneously, methodologies and pipelines have been developed to explore potential associations between specific variants and diseases, such as genome-wide association studies (GWAS) (Uffelmann *et al.*, 2021), or to predict their functional impact on the molecular mechanisms governing the cell (McLaren *et al.*, 2016). These studies and achievements unraveled the intricate complexity and diversity inherent in the human genome. However, representing such intricacy using the conventional reference genome sequences becomes challenging. To address this limitation, complex data structures known as genome graphs (Paten *et al.*, 2017) have been recently introduced. These graphs offer a comprehensive framework to embed the genetic diversity within individuals and populations, presenting a novel, graph-based data structure. This innovative approach is poised to provide a more nuanced understanding of the intricate tapestry of the human genome. Genome graphs also provide a comprehensive and efficient framework to consider individual- and population specific mutations, supporting the shift towards precision medicine-oriented approaches. Moreover, the recent release of the first draft of the human pangenome reference (Liao *et al.*, 2023) testifies how the shift towards the use of frameworks embedding genetic variation are more and more supported by the scientific community.

2.3.1 Identifying Genetic Variants

Identifying and interpreting the consequences of variants occurrence is crucial to unveil the molecular mechanisms underlying several common and rare traits and diseases. Historically, variants and in particular SNVs were identified employing techniques such as mass spectrometry, or single-strand conformation polymorphism (Orita *et al.*, 1989). However, this methods allowed to analyze a few variants. DNA sequencing technologies almost completely replaced these protocols. In fact, sequencing technologies allow to recover and study the entire variation of genomes at lower costs. To analyze sequencing data and *call* variants, during the last decades have been developed several *variant calling* pipelines and methods. Variant calling pipelines encompass three main steps: (i) genome sequencing, (ii) sequence alignment to a reference genome, and (iii) variants identification. During genome sequencing the DNA is sequenced by sequencer machines, which reconstruct the genomic sequences splitting it in short or long overlapping and redundant sequences, called short or long reads, respectively. The resulting reads are stored in FASTQ files (**Appendix A.2**). The generated reads are then mapped to a reference genome. Aligning reads helps identify the genomic locations of these reads within the genome, allowing to reconstruct the original genomic sequence and enables the identification of genetic variations, such as SNPs, indels, and structural variations, during the variant identification. The accurate alignment of reads is essential for downstream analyses like variant calling and understanding the functional implications of genetic variations in the context of the reference genome. Aligned reads are stored in BAM or CRAM files (**Appendix A.4** and **A.5**). Various algorithms and tools, like Bowtie (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), and HISAT (Kim *et al.*, 2019), are commonly used for efficient and accurate sequence alignment. The last step focuses on identifying and characterizing genetic variants in the sequenced genome. Once the reads are aligned, variant calling algorithms are employed to detect differences between the sample and the reference. These algorithms analyze the aligned reads to identify positions where the sample's sequence differs from the reference genome. Variant calling algorithms may apply filters on the identified variants to remove false positives and improve the accuracy of variant calls. Moreover, they can annotate the identified variants, adding information such as their genomic location, functional impact, and population frequency. Variant callers can also call determines the specific genetic makeup (*genotype*) of an individual at each variant position. The identified variants are generally stored in VCF (Variant Calling Format) files. Variant calling is crucial for understanding the genetic differences between individuals, populations, or disease states. Multiple tools, including GATK (McKenna *et al.*, 2010), SAMtools (Li *et al.*, 2009a), Strelka (Kim *et al.*, 2018) and VarScan (Koboldt *et al.*, 2012), are commonly used for variant calling. Accurate and reliable variant calling is essential for studies focused on identifying disease-causing mutations, understanding genetic diversity, and uncovering associations between genetic variants and specific traits or diseases. Furthermore, variant calling constitutes a fundamental tool to recover individual and population-specific genetic diversity, playing a pivotal role in defining precision medicine oriented computational methods.

The Variant Calling Format (VCF)

The Variant Call Format (VCF) is a standard file format used for storing information about genetic variants identified through variant calling analyses. VCFs are plain text file providing several and different information organized in fields regarding the reported variants. Due to their dimensions, VCFs are often compressed. VCFs have two main components: (i) a header and (ii) a body. The header contains metadata and information about the reference genome, describes the sample information, file creation date, the tools used for variant calling, and specifies the version of the VCF format. The body comprises rows of variant records, each representing a genetic variant. Variants records consist of eight mandatory tab-separated fields, along with an unlimited number of optional columns (**Table 2.2**). The fields in a variant record provide detailed information about the variant. The VCF file format is versatile and supports the representation of a wide range of genetic variants, including SNPs, indels, and structural variants. It has become a standard for exchanging variant information across different bioinformatics tools and databases, facilitating the sharing and integration of genomic data in research and clinical settings.

Field	Field name	Description	Type
1	CHROM	The identifier of the sequence on which the variant has been called, commonly denoting the chromosome name	Mandatory
2	POS	The variant's position relative to the reference sequence	Mandatory
3	ID	Variant's identifier	Mandatory
4	REF	Reference allele	Mandatory
5	ALT	Altrenative allele	Mandatory
6	QUAL	The assigned quality score for the inference of the called allele	Mandatory
7	FILTER	A flag value indicating whether the variant has successfully passed the filtering steps	Mandatory
8	INFO	An expandable list of key-value pairs describing the variant, following the syntax <key>=<data>[,data]	Mandatory
9	FORMAT	Sample's related fields	Optional
+	SAMPLE	Values are provided for the FORMAT fields for each described sample	Optional

Table 2.2. Variant Calling Format (VCF) fields. The table outlines the eight mandatory fields of a VCF file along with two frequently used optional fields (FORMAT and SAMPLE)

2.3.2 Genetic Variants Databases

The recent strides in sequencing technologies have substantially augmented the throughput of sequencing experiments, generating vast datasets. In response to this, numerous global initiatives, exemplified by the International HapMap Project (Gibbs *et al.*, 2003; Consortium, 2005), the 1000 Genomes Project (1KGP) (Siva, 2008; Consortium *et al.*, 2015) and the Human Genetic Diversity Project (HGDP) (Cavalli-Sforza, 2005), have surfaced. These endeavors aim to delve deeper into the genetic diversity among human individuals and populations. Similarly, consortia like dbSNP (Sherry *et al.*, 2001) and ClinVar (Landrum *et al.*, 2020) have aggregated variant data to explore potential associations between genetic variants and various traits or diseases. The HapMap project was among the first initiative to perform large scale experiments to uncover human genetic diversity. Their goal was to construct a haplotype map of the human genome, elucidating potential prevalent variants patterns among populations. In fact, while generally shared across populations, haplotypes can exhibit significant variations in frequency. The primary objective of 1KGP was to identify common genetic variants with frequencies of at least 1% within the studied populations (SNPs). Leveraging advancements in sequencing technology, the project pioneered large-scale sequencing of genomes from a diverse group of individuals, creating a comprehensive database on human genetic variation (Fairley *et al.*, 2020). While the cost of deep sequencing for the extensive number of samples was still prohibitive, the genome's specific regions typically harbor a limited number of haplotypes. By aggregating data across samples, the project efficiently detected the majority of variants in a region. The initial plan aimed to achieve 4x genomic coverage for each sample, recognizing that this depth might not unveil all variants in each sample but could detect most variants with frequencies

as low as 1%. In the project's final phase, data from 2,504 samples from 26 populations were amalgamated, enabling accurate genotyping at all discovered variant sites. The multi-sample strategy, coupled with genotype imputation (Li *et al.*, 2009b), facilitated determining samples' genotype, even for variants not covered by sequencing reads. However, 1KGP dataset does not cover all human populations. To fill this gap the 1KGP data have been extended performing further experiments as well as including data produced by other consortia, such as HGDP (Bergström *et al.*, 2020). Similarly to 1KGP, HGDP aim was to identify and map the genetic variation and diversity among human populations. 1KGP and HGDP could be considered complementary datasets. In the meantime, other consortia focused on collecting the variants experimentally identified in databases. dbSNP (Sherry *et al.*, 2001) is a publicly accessible repository providing comprehensive information on genetic variation. It functions as a unified database containing all identified genetic variations, facilitating the investigation of a broad range of genetically influenced natural phenomena. Despite its name suggests a focus solely on SNPs, dbSNP encompasses a broader spectrum of molecular variations (Sherry *et al.*, 1999), including SNPs and short indels for example. dbSNP plays a crucial role in guiding applied research in fields such as exploring variants-trait associations (Kitts and Sherry, 2002). While dbSNP reports broader information about genetic variants, ClinVar (Landrum *et al.*, 2020) details variants-traits association focusing on the clinical significance of such associtions. ClinVar main purpose is to streamline access to information and facilitate communication regarding the relationships posited between human genetic variations and observed health conditions, along with the historical context of these interpretations. These repositories provide valuable resources providing several data to analyse the potential impact of genetic variation on the cell environment. By providing raw data and functional details, these databases are significant resources, fundamental to develop precision medicine-oriented methods that consider genetic variants.

2.3.3 Discovering Variants - Trait associations

Once identified, a key question is to discover whether our variants are linked to a particular trait, or to predict their potential impact on the cellular environment. Genome-wide association studies (GWAS) (Uffelmann *et al.*, 2021) provide answers to the first question. GWAS is an observational investigation of a comprehensive set of genetic variants across different individuals to identify potential associations with a particular trait or disease. These studies predominantly focus on relationships between SNPs and major human diseases, although they are equally applicable to other genetic variants, such as indels for example. GWAS typically involve large cohorts of individuals, comparing the genomes of those with a particular trait or disease to those without. Case-control designs are common, where individuals with a specific trait or disease (cases) are compared to those without (controls). To collect GWAS data high-throughput genotyping technologies, such as SNP arrays, are employed to assay hundreds of thousands to millions of SNPs across the genome. To identify between genetic variants and traits, GWAS employ statistical methods. Correction for multiple testing is crucial to minimize false-positive results. Linkage Disequilibrium analysis helps identify blocks of correlated genetic variants, providing insights into the genomic regions associated with the trait of interest. Importantly, population stratification must be carefully addressed to avoid spurious associations due to differences in genetic ancestry. Different computational tools, such as PLINK (Purcell *et al.*, 2007) or SNPTEST (Marchini *et al.*, 2007), have been developed to carry out GWAS. While other methods focus on a limited number of predefined genetic regions, GWAS explores the entire genome. Consequently, GWAS adopts a non-candidate-driven approach, in contrast to gene-specific candidate-driven studies. However, while GWAS identifies SNPs and other DNA variants associated with a disease, it does not independently determine which impact the variants have on the cell environment and which gene/genomic element is causative of the investigated trait/disease. Functional annotation annotates and interprets the functional significance of variants identified during GWAS. Several databases and tools have been developed for this task. For example PROVEAN (Choi and Chan, 2015) and VEP (McLaren *et al.*, 2016) predict the impact of variants on gene products, while RegulomeDB (Boyle *et al.*, 2012) investigates the impact of genetic variants on genomic regulatory elements.

2.3.4 Limitations of current analysis methods

As described throughout this chapter, various methods and computational tools have been proposed over the last few decades to analyze, assess, and predict the impact of genetic variants on the cellular environment. However, these methods may have limitations and may not fully exploit the information of the analyzed variants data. Methods employing variants to enrich the reference genome sequence may lose or not consider simultaneously SNVs and indels, because the latter may break the sequence

coordinate system. Moreover, these methods generally are not able to consider large variant events, such as structural variants, and complex SNVs-indels combinations in their analyses. Most methods ignores haplotypes and may introduce recombinant sequences not observed in the data. Furthermore, most methods are not scalable when analyzing large set of individual-specific data and do not perform population-based analyses. These limitations may impact the shift towards the development of precision medicine-oriented methods. The recent advent of genome graphs (Paten *et al.*, 2017) presents a powerful framework for representing genomes alongside their complete genetic variation and diversity. Importantly, genome graphs may enhance precision medicine-oriented analyses by encapsulating the genetic diversity of numerous samples, ranging from several to thousands, within a unified data structure (Yu and Chen, 2023). The subsequent chapter delves into the details of the genome graphs data structure, shedding light on recent advancements and exploring the prospective enhancements this data structure could bring to the forefront of developing novel methods for analyzing the impact of genetic variants on cellular processes.

Genome Graphs

The use of the human reference genome as a basis for studying genetic variations in other human genomes introduces a significant challenge known as reference allele bias. This bias, stemming from mapping errors during sequencing experiments, tends to underreport data that deviates from the reference allele (Degner *et al.*, 2009; Brandt *et al.*, 2015). Structural variations pose a particular challenge, requiring distinct algorithms for their detection, due to their absence from the reference (Sudmant *et al.*, 2015). To address these limitations, the field of genome inference is moving towards unbiased *de novo* assembly for individual samples and developing richer reference structures that more comprehensively represent population variations, such as graph-based reference structures. This transition towards graph-based reference structures is particularly evident in human genomics, signaling a shift in the approach to variant calling and genome analysis. Genome graphs (Paten *et al.*, 2017) offer a revolutionary approach to address the limitations of traditional linear genome representations (**Figure 3.3**). Stemming from the concept of sequence graphs (Paten *et al.*, 2017), genome graphs provide a versatile framework to encapsulate the genetic diversity inherent in multiple individual genomes within a unified data structure. A sequence graph (**Figure 3.3 (A)**) is essentially a bidirectional graph where nodes correspond to DNA fragments, and edges signify the adjacency between different segments in a sequence. These edges connect the terminus of one DNA segment to the inception of another, bearing a label denoting a string of DNA (which can be empty). Consequently, traversing the graph through adjacent nodes linked by edges allows for the retrieval of genetic sequences. Originally conceived to succinctly represent multiple sequences sharing variations, sequence graphs have found extensive applications. Mainly, sequence graphs serve as efficient data structures for storing diverse genetic information and have been instrumental in tasks such as the succint representation of multiple sequence alignments (Lee *et al.*, 2002; Paten *et al.*, 2014). Therefore, genome graphs can be perceived as an evolution of bidirectional sequence graphs. Notably, genome graphs offer a comprehensive representation enriched with a collection of paths encapsulating genome sequences across a population of individuals (**Figure 3.3 (B)**). These graphs serve as a compact yet robust depiction of genetic variation within a population, encompassing structural variants, such as inversions and duplications (Paten *et al.*, 2017). Formally, a genome graph is defined as $G = (N, E, P)$, where $N = \{n_1, n_2, \ldots, n_k\}$ denotes the set of nodes representing DNA segments, $E = e_1, e_2, \ldots, e_w$ denotes the set of edges connecting nodes, and $P = p_1, p_2, \ldots, p_q$ comprises the set of paths embedded in G. Each path $p \in P$ represents one of the embedded sequences or genomes within the genome graph. Each node in the graph corresponds to a DNA segment constructed from the alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{N}\}$. Therefore, by incorporating a multitude of paths, genome graphs offer a holistic view of genomic diversity within a population of individuals, enabling the representation of complex genetic structures and facilitating comprehensive genetic analyses. Traversal through the graph is facilitated in both forward and reverse directions, presenting two distinct "entry points" for each node $n \in N$ (**Figure 3.4 (A)**). When traversing a node in the reverse direction, the DNA fragment associated with the node is read as its reverse complement (**Figure 3.4 (B)**). Edges within the genome graph signify adjacency between the DNA fragments of the nodes they connect. Consequently, long sequences are implicitly encoded by concatenating the fragments of adjacent nodes. Similar to nodes, edges can be traversed in the reverse direction, enabling comprehensive exploration of genetic sequences. Moreover, to increase graph's compactness, genome graphs are capable of accommodating cycles to represent repetitive segments across the embedded sequences (**Figure 3.4 (C)**). Genome graphs usually accommodate different types of cycles: regular cycles where a node n_i can be reached from n_i, reversing cycles, where n_i^r is reachable from n_i, and instances of non-cyclic reversal (**Figure 3.4 (D)**), where both n_i and n_i^r are accessible from another node n_j. In the context of representing genomes for a population of individuals, paths within genome graphs correspond to the genomic sequences of each subject. Variants manifest as bubbles in the graph, creating diverging paths anchored by a common start and end sequence on the reference (Paten *et al.*, 2018) or a shared DNA segment. Genome graphs provide an efficient and compact framework to

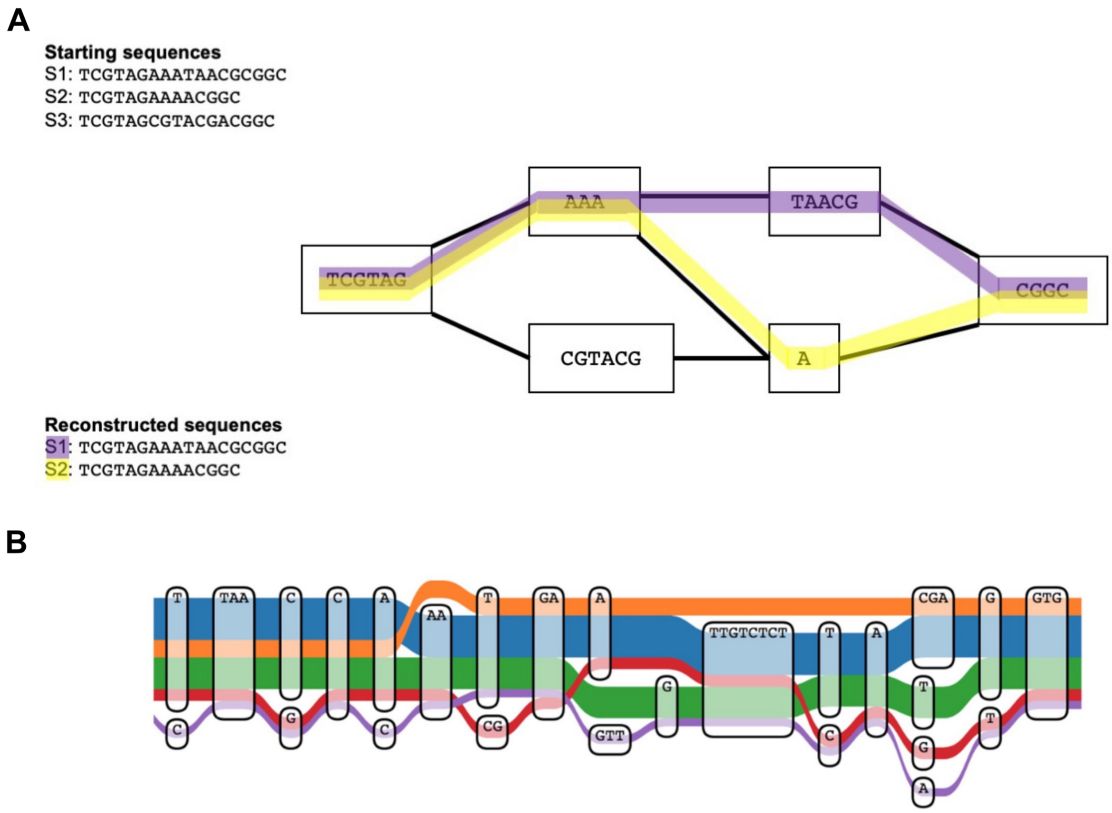


Figure 3.3. The genome graph and sequence graph data structures. **(A)** A sequence graph, denoted as S, is constructed from three sequences: s_1 = TCGTAGAAATAACGCGGC, s_2 = TCGTAGAAAACGGC, and s_3 = TCGTAGCGTACGACGGC. By tracing the purple path within the graph, the original sequence s_1 can be reconstructed, while following the yellow path yields the reconstructed sequence s_2. It's noteworthy that in sequence graph S, the edges are not labeled, although in general, they could carry labels for additional information. **(B)** Visualization of Genome Graphs data structure, with each color representing a distinct path within the graph. Each path delineates the genomic sequence of an individual genome encoded within the Genome Graph structure. This comprehensive representation facilitates the simultaneous depiction of multiple genetic sequences, enabling the exploration of genetic diversity and complexity across individuals within a population.

represent and encode the genetic diversity observed within populations of individuals. Notably, genome graphs offer a groundbreaking framework that expands classical omics analysis to encompass not only simple variations like SNPs and indels, but also more complex genetic alterations such as large indels, SNP-indel combinations, and structural variants. This versatility makes genome graphs a powerful tool for capturing the intricate genetic landscape present in diverse populations. In recent years, there has been a noticeable shift towards adopting genome graph-based representations of genomes over traditional genome assemblies for genetic data analysis (Hurgobin and Edwards, 2017). This transition has been further accelerated by the emergence of several consortia dedicated to constructing and releasing population-based (Gao *et al.*, 2023) or comprehensive pangenomes (Wang *et al.*, 2022c), with these pangenomes being represented as genome graphs. Notably, a significant milestone was reached with the release of the first draft of a genome-graph-based human reference genome, known as the pangenome reference (Liao *et al.*, 2023). This landmark achievement underscores the growing recognition and adoption of genome graph technology in genomic research, promising a more comprehensive and accurate understanding of genetic variation and its implications. Furthermore, genome graphs hold great promise in facilitating the shift towards the adoption of precision medicine-oriented approaches to analyse genetic data accounting for genetic diversity across individuals and populations (Yu and Chen, 2023). However, scaling pangenomes' genome graph data structures to accommodate hundreds of genomes poses a significant computational challenge. In recent years, numerous software tools have been developed to construct and analyze genome graph-based pangenomes (Andreace *et al.*, 2023). While construction procedures and implementations may vary, they collectively underscore how genome graphs significantly enhance a wide array of genomic analyses. These improvements are evident in various areas, including high-quality short read mapping (Sirén *et al.*, 2021), accurate genotyping of SNPs, indels, and structural variants (SVs) (Ebler *et al.*, 2022), RNA-seq mapping (Liao *et al.*, 2023), and de novo variant calling (Garrison *et al.*, 2018). The following sections delve into some of the prominent methods for constructing genome graph-based pangenomes and elucidate their potential to enhance genetic data analysis.

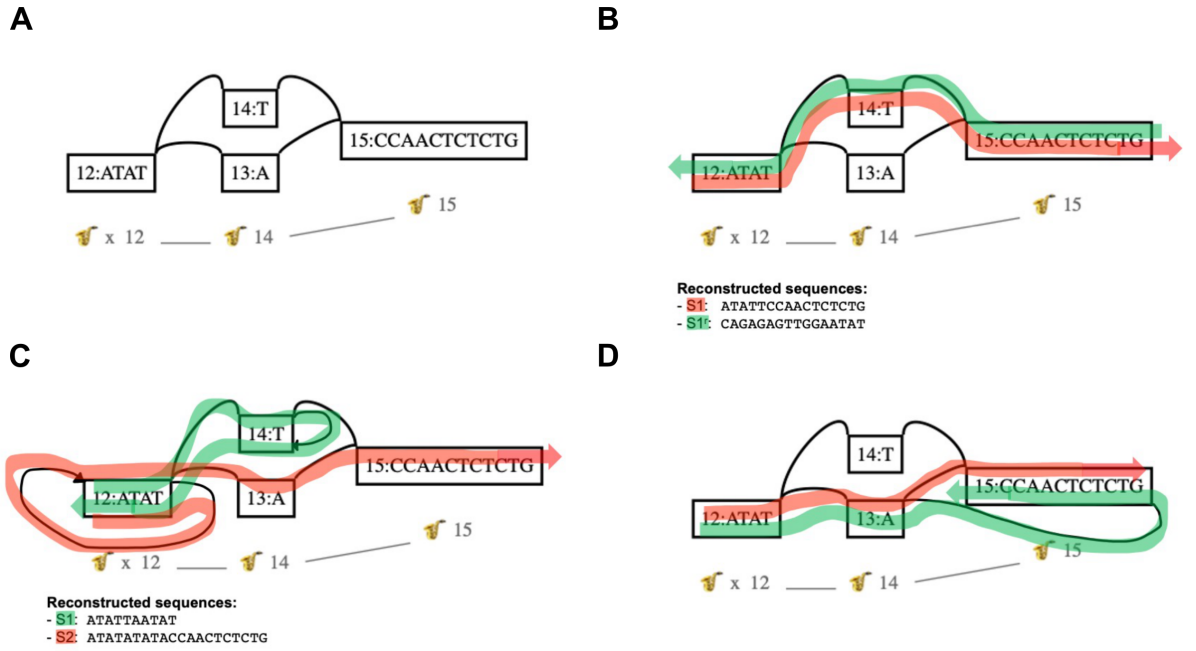


Figure 3.4. Traversing the genome graph along the embedded paths. **(A)** A genome graph provides multiple traversal possibilities based on its nodes, edges, and paths. **(B)** By following the red path, the forward sequence `ATATTCCAACTCTCTG` can be reconstructed, while tracing the green path yields its reverse complement. **(C)** Genome graphs accommodate ordinary cycles and reversing cycles. The red path reveals an ordinary cycle at node 12, while the green path highlights a reversing cycle at node 14, where an edge exits and re-enters the node, necessitating consideration of the DNA label as its reverse complement. Such intricate graph structures efficiently represent complex genomic rearrangements, such as sequence duplications (as seen in the red path case) or inversions (as observed in the green path case). **(D)** Genome graphs may also exhibit non-cyclic instances of reversal; for instance, node 13 can access both the forward DNA fragment of node 15 (via the red path) and its reverse complement (via the green path).

3.1 Methods to Construct Genome Graphs

In recent years, we have witnessed a dynamic landscape of research endeavors aimed at developing methods and data structures tailored to construct and represent pangenomes as genome graphs (Consortium, 2018; Andreace *et al.*, 2023). This surge in activity reflects the growing recognition of genome graphs as powerful tools for capturing the complexity and diversity inherent in genomic data. Among the diverse array of approaches currently available, several notable methods have garnered attention for their efficacy in handling large datasets and facilitating efficient representation of genome graphs. The Variation Graph Toolkit (Garrison *et al.*, 2018), for instance, has emerged as a foundational tool in the field, offering robust capabilities for constructing genome graphs that encompass intricate genetic variations. Notably, the Variation Graph Toolkit provides a comprehensive suite of tools for analyzing genomic sequence data utilizing the genome graph data structure. The resulting variation graphs (VGs) can be constructed either from user-provided data or from precomputed graphs in a variety of supported graph formats. With the Variation Graph Toolkit, users can manipulate and visualize VGs, enabling detailed exploration and analysis of genomic variations. The introduction of VGs represents a significant advancement in genomic data representation. By encapsulating the complexity introduced by genetic variations in a compact and efficient manner, VGs provide a powerful framework for understanding genomic diversity. Unlike traditional reference genomes, VGs can succinctly summarize the complexity of genetic information within a population of individuals in a single structure. This capability enhances the accuracy of describing genetic changes and variations, offering insights that linear reference genomes cannot convey. Pantools (Sheikhizadeh *et al.*, 2016) and Bifrost (Holley and Melsted, 2020) have also gained prominence for their adeptness in generating pangenomes from extensive collections of genomes, particularly in the context of phylogenetics and bacterial genomics. However, Pantools have also been used to build a human pangenome encompassing seven genomes (Sheikhizadeh *et al.*, 2016). Bifrost specializes in constructing dynamic, colored compacted de Bruijn Graphs (Muggli *et al.*, 2017). Its approach involves first generating a standard de Bruijn Graph using an efficient variant of Bloom Filters (Tarkoma *et al.*, 2011), followed by the computation of the compacted de Bruijn Graph. In this process, colors—identifiers represent the sample of origin of each k-mer and are incorporated into the data structure by storing an array for each k-mer. A noteworthy characteristic of the colored de Bruijn Graph constructed by Bifrost is its single

large connected component when representing human genomes. This occurrence arises from the sharing of common k-mers among chromosomes. Consequently, this pangenome representation encapsulates all the variants present in the input sequences, providing a comprehensive view of genetic diversity across individuals. Minigraph (Li *et al.*, 2020), Minigraph-Cactus, and PGGB (Garrison *et al.*, 2023) stand out for their versatility and efficiency in constructing genome graphs capable of representing complex genetic structures. These methods leverage sophisticated data structures and algorithms to handle large-scale genomic datasets while providing detailed insights into genetic variations and structural complexities. Notably, these three approaches have been employed to assemble and release the first draft of the human pangenome reference (Liao *et al.*, 2023). Minigraph constructs a VG in a directed, bidirected, and acyclic manner through an iterative process. It achieves this by iteratively mapping new haplotypes combining minimap2 (Li, 2018) and a graph wavefront alignment algorithm (Li *et al.*, 2020). Minigraph focuses on capturing variations longer than 50 base pairs, making it particularly suited to handle larger-scale genetic alterations. However, it does not consider isolated SNPs and small indels, as its resolution is not at the base level. While Minigraph provides base-level alignment for contigs, the resulting graphs offer a higher-level resolution. Furthermore, the resulting graph is partitioned into connected components, each corresponding to the chromosomes present in the initial input genome. This organization facilitates the analysis and interpretation of genomic data within the context of chromosome-level structures. The Minigraph-Cactus pipeline is a sophisticated tool for constructing VGs, which integrates the capabilities of Minigraph (Li *et al.*, 2020) and the Cactus (Armstrong *et al.*, 2020) base aligner. This pipeline is designed to generate a structural VG using Minigraph and then leverage Cactus to produce base-level pangenome graphs from a set of input assemblies, incorporating haplotypes as paths within the graph. Importantly, Cactus ensures the resulting VG remains acyclic by considering the reference sequence employed by Minigraph. Additionally, the resulting pangenome graph undergoes further processing using GFAffix. This versatile pipeline offers the flexibility to generate multiple graphs, each corresponding to a specific chromosome, or to produce a single graph that encompasses inter-chromosomal variants. By combining Minigraph and Cactus, the Minigraph-Cactus pipeline provides a robust framework for constructing comprehensive variation graphs suitable for diverse genomic analyses. PGGB, instead, encompasses different tools to create a pipeline dedicated to constructing directed acyclic VGs. Firstly, it conducts pairwise base-level alignment of haplotypes using wfmash (Garrison and Guarracino, 2023). Subsequently, it constructs the graph from these alignments using seqwish (Garrison and Guarracino, 2023). Lastly, the pipeline employs two more tools, smoothxg and GFAffix, to sort and normalize the graph. The resulting variation graph comprehensively represents variations of all lengths present in the input sequences. This approach ensures that the constructed graph encapsulates the full spectrum of genetic variations, making it a powerful resource for genomic analyses. However, many of these tools and pipelines for constructing genome graphs come with certain limitations. They often demand computational expertise from users and may require substantial computational resources to operate effectively. As the field progresses, there is a pressing need to enhance the scalability, accuracy, and user-friendliness of genome graph-based approaches. Future research endeavors are poised to concentrate on refining existing methods, investigating innovative algorithms, and developing more user-friendly tools to facilitate wider adoption of genome graphs in genomic research and analysis. Nevertheless, genome graphs have already found applications in various biological analyses, offering notable enhancements compared to analyses conducted using classical linear reference genomes.

3.2 Improving genome analysis pipelines

One immediate application of genome graphs lies in enhancing reference-based genetic data analysis workflows. In such workflows, genome graphs seamlessly replace existing linear references subsequent processing analyses. For instance, the consortium that released the human pangenome draft demonstrated as pangenome improves sequence mapping workflows. By projecting reads mapping from the pangenome space back to the linear reference, they improved variant calling workflows combining Giraffe mapper (Sirén *et al.*, 2021) and the DeepVariant variant caller (Poplin *et al.*, 2018). In the Giraffe–DeepVariant workflow, DeepVariant remains unaware of the intricacies of the pangenome, yet the workflow gains from a mapping step that accommodates sequences absent from the linear reference. Notably, transitioning to pangenome mapping does not entail significantly higher computational costs and has yielded a significant reduction in false-positive and false-negative calls compared to standard reference-based methods. Moreover, the authors showed that these improvements are particularly pronounced at complex loci. Pangenomes not only enhance variant calling accuracy but have also been shown to improve transcript mapping (Sibbesen *et al.*, 2023) and facilitate detection of ChIP-seq peaks (Groza *et al.*, 2020). For

example, Liao *et al.* (2023) demonstrated how pangenome-based transcriptome mapping pipeline using vg mpmap (Sibbesen *et al.*, 2023) significantly decreased false mapping rates compared to a linear reference-based pipeline employing either vg mpmap or STAR (Dobin *et al.*, 2013). Furthermore, the pangenome-based pipeline exhibited reduced allelic bias and enhanced mapped coverage on heterozygous variants. Additionally, they observed an increase in peak called analyzing ChIP-seq (**Section 4.2.1**) targeting two hystones (H3K4me1 and H3K27ac) and ATAC-seq data (**Section 4.2.1**), compared to the linear reference-based analyses. Moreover, they identified regulaory features specific to structural variant alleles not detected in the linear reference genome. Therefore, genome graphs serve as invaluable frameworks, enabling the comprehensive capture and representation of genetic diversity among individuals. Additionally, they offer a robust framework for effectively capturing and representing genomic regulatory elements, including their variability across subjects and populations.

Genomic Regulatory Elements

Genetic variants located outside coding regions can significantly influence cellular environments. Specifically, non-coding variants within genomic regulatory elements (GREs) have been identified as key players in altering gene expression (De Gobbi *et al.*, 2006; Wienert *et al.*, 2015), linking them to the emergence of certain traits and increased disease susceptibility (Weinhold *et al.*, 2014). Mutations within GREs have the potential to disrupt the intricate regulatory networks governing gene expression by modifying GRE sequences and functions. Gene expression unfolds during RNA biosynthesis, and its orchestration relies on sequence-specific protein binding, such as histones or transcription factors (TFs). Histones, forming nucleosomes (DNA-histone complexes), govern gene expression by controlling chromatin accessibility and compactness. Alterations in chromatin compactness directly impact gene expression by, for instance, reducing DNA accessibility to the RNA polymerase complex. Histone modifications, including acetylation, methylation, and phosphorylation, can either promote or inhibit gene expression by modulating DNA-histone interactions. Moreover, histones contribute to the three-dimensional organization of chromatin, shaping higher-order structures and influencing the physical arrangement of genomic elements. TFs regulate gene expression by binding to short target sequences known as transcription factor binding sites, often located within GREs called cis-regulatory sequences. These sequences, encompassing enhancers, silencers, and insulators (Kolovos *et al.*, 2012), can significantly impact gene expression rates, even when distant from gene promoters (Wittkopp and Kalay, 2012). Enhancers, in particular, play a pivotal role in gene expression regulation by collaborating with TFs (Spitz and Furlong, 2012). Furthermore, DNA methylation, another critical mechanism, can enhance or inhibit gene expression by adding methyl groups to GRE sequences, such as gene promoters. Interacting with methyl-binding domain (MBD) proteins, DNA methylation influences the recruitment of protein complexes with chromatin remodeling and histone modifying activity to methylated CpG islands (Du *et al.*, 2015). MBD proteins, with a strong affinity for highly methylated CpG islands, primarily function to repress local chromatin through mechanisms like inducing repressive histone marks or establishing an overall repressive chromatin environment via nucleosome remodeling and chromatin reorganization (Du *et al.*, 2015). Understanding how genetic variants impact these fundamental genomic elements is crucial for interpreting the consequences of individual- and population-specific genetic variations on the cellular environment. The next sections will delve into enhancers, silencers, insulators, and the essential proteins regulating gene expression by interacting with GREs: transcription factors.

4.1 Enhancers, Silencers and Insulators: a molecular regulatory symphony

The genome harbors a diverse array of genes, each encoding one or more products. The precise regulation of gene expression is crucial. The most efficient regulatory mechanism occurs at the transcriptional level, where cis-regulatory elements (CREs) play a pivotal role in controlling transcription (Wittkopp and Kalay, 2012). Enhancers and promoters are well-characterized CREs regulating gene expression. Cis-regulatory modules (CRMs), a subset of CREs, serve as functional regulatory elements by harboring transcription factor binding sites. Briefly, CRMs are DNA sequences with clustered transcription factor binding sites, encompassing diverse modular structures like locus control regions, promoters, enhancers, silencers, and boundary control elements. These modules are classified into four main classes: (i) promoters, (ii) enhancers, (iii) silencers, and (iv) insulators. Promoters are relatively short sequences encompassing the site where transcription initiates and the surrounding region ∼35 base pairs upstream or downstream from the initiation site (Butler and Kadonaga, 2002). Importantly, a single gene may harbor multiple promoter sites. To start the expression of the downstream gene, a series of TFs bind sequentially the promoter. Only upon the binding of the appropriate set of TFs in the correct order, the RNA polymerase attach

to the promoter and begins the transcriptional process. Enhancers positively regulate (enhance) gene expression. They can be situated upstream, downstream, within introns, or far from the controlled gene. Coordination among multiple enhancers is common in the regulation of gene transcription. Insulators indirectly interact with neighboring CRMs. Silencers, instead, possess the ability to turn off (silence) gene transcription through the binding of transcriptional regulatory proteins, known as repressors.

4.1.1 Enhancers

Enhancers are short DNA region, typically ranging from 50 to 1500 bp, where activators proteins, known as transcription factors, can bind. TFs binding increases the likelihood of transcription initiation for the controlled gene or genes set. In the human genome there exist >100,000 enhancers (Pennacchio *et al.*, 2013). Enhancers exhibit a remarkable ability to influence transcription even when located at considerable distances from the transcription initiation site, with some identified enhancers are located >100 Kb upstream or downstream of the start site (Pennacchio *et al.*, 2013). Enhancers exert their control on gene expression through the binding of activator proteins rather than directly on the promoter sequence. In fact, the activator proteins engage with the mediator complex, recruiting RNA polymerase and TFs, which subsequently start the transcription process. Enhancers interact with far controlled genes by forming loops that reduce the physical distance from the targets (Schoenfelder and Fraser, 2019). Enhancers are not confined to specific locations within the gene structure. They can be located within introns, and their orientation or even their physical location in the chromosome can be altered without compromising their functionality. Furthermore, enhancers may be located within exonic region of unrelated genes and they may act on genes situated on different chromosomes (Spilianakis *et al.*, 2005).

4.1.2 Silencers

Silencers are specific sequences with the capacity to bind transcriptional regulatory factors, known as repressors. However, unlike enhancers, when repressor proteins interact with silencers, they block the RNA polymerase, preventing and "silencing" gene expression (Pang *et al.*, 2023). Similarly to enhancers, silencers may be located Kbs either upstream or downstream of the target gene start site. There can be identified two categories of silencers: classical silencer and non-classical negative regulatory element (NRE). Classical silencers actively suppress gene expression by directly impeding the assembly of TFs, predominantly interfering with their proper functioning (Ogbourne and Antalis, 1998). On the other hand, NREs exert passive repression on genes, generally by hindering elements upstream of the gene. In classical silencers repressors target the TFs assembly initiating gene transcription, which are typically located upstream of the gene and can exhibit varying distances, ranging from short to long. In the case of long-range silencers, they form DNA loops to facilitate the physical proximity of silencer-promoter pairs (Maston *et al.*, 2006). Additionally, NREs may induce bends in the promoter region, obstructing TF-RNA polymerase interactions. If silencers are located within introns, two types of repressions may occur: physical blockage of a splice site and a bent DNA structure inhibiting RNA processing.

4.1.3 Insulators

Unlike enhancers and silencers, insulators act only as a long-range regulatory element, by controlling target genes located at considerable distances. Typically spanning 300-2000 bp, insulators feature clustered binding sites for sequence-specific DNA-binding proteins, facilitating intra- and inter-chromosomal interactions (Valenzuela and Kamakaka, 2006). The functional roles of insulators encompass acting as enhancer-blockers, barriers, or a combination of both. These functions are orchestrated through mechanisms involving loop formation and nucleosome modifications (Gaszner and Felsenfeld, 2006). Various examples of insulators exist, such as the CTCF insulator (Ishihara *et al.*, 2006). Insulators with an enhancer-blocking function form chromatin loop domains that physically separate the enhancer from the promoter of the target gene. The creation of loop domains occurs through interactions between enhancer-blocking elements, either by engaging with each other or by anchoring the chromatin to structural elements, such as nucleosome (Gaszner and Felsenfeld, 2006). The effectiveness of these insulators relies on their location between the promoter of the target gene and the adjacent enhancer. If enhancers directly engage with their target promoters through looping (direct-contact model), insulators disrupt this interaction by establishing a loop domain spatially dividing the enhancer and promoter sites, thus preventing the formation of the promoter-enhancer loop (Gaszner and Felsenfeld, 2006). Alternatively, when an enhancer influences a promoter through a signal, following the tracking model of enhancer action, insulators can impede this signal by targeting a nucleoprotein complex at the base of the loop formation (Gaszner

and Felsenfeld, 2006). On the other hand, barrier insulators alter the nucleosomal substrate in the reaction cycle crucial to heterochromatin formation (Gaszner and Felsenfeld, 2006). Different mechanisms are employed for these modifications, encompassing nucleosome removal and silencing (chromatin-mediated silencing). Additionally, modification can occur through the recruitment of histone acetyltransferase(s) and ATP-dependent nucleosome remodeling complexes (Gaszner and Felsenfeld, 2006).

4.2 Transcription Factors

Transcription factors (TFs) (**Figure 4.5 (A)**) are essential regulatory proteins that play a pivotal role in governing the transcriptional state, cellular differentiation, and developmental status of cells (Lambert *et al.*, 2018; Reimold *et al.*, 2001; Whyte *et al.*, 2013). In humans, around 1,600 proteins are identified as TFs (Babu *et al.*, 2004), constituting approximately 8% of all human genes. This highlights the crucial function of TFs in orchestrating genetic regulation. TFs demonstrate their regulatory capabilities by frequently collaborating in a coordinated manner, exerting a collective influence on gene expression. This collaborative orchestration is essential for finely tuning and precisely controlling cellular processes. This collective coordination is indispensable for finely adjusting and accurately regulating cellular processes. Additionally, transcription factors (TFs) demonstrate remarkable versatility by overseeing the activity of numerous genes across different cell types (Lambert *et al.*, 2018). TFs exhibit a modular structure, which is categorized into three distinct domains (Latchman, 1997) (**Figure 4.5 (B)**): (i) the DNA binding domain, (ii) the activation domain, and (iii) the signal sensing domain. The DNA binding domain serves as the navigational guide for the transcription factor (TF), directing it to its specific target site on the genome. By precisely recognizing DNA sequences, this domain empowers the TF to anchor onto regulatory regions dispersed throughout the genome. This targeted interaction allows the TF to exert its regulatory influence on gene expression within those specific genomic locations. The activation domain of a transcription factor serves a pivotal role in facilitating interactions between the TF and various transcriptional regulators. Through engagements with diverse co-factors and regulatory proteins, the activation domain acts as a key modulator of gene expression. Its function often involves serving as a bridge between the transcription factor and the intricate machinery involved in transcription, thereby influencing the regulatory processes that govern gene activity. The signal sensing domain of a transcription factor captures external signals and relays them to the broader transcriptional complex. These signals originate from different sources, encompassing cellular cues and environmental stimuli. The signal sensing domain acts as a sophisticated sensing mechanism that allows TFs to finely adjust their regulatory activity in response to dynamic and changing conditions within the cellular and environmental milieu. This adaptive responsiveness ensures a nuanced and context-specific modulation of gene expression. The intricate interplay between these three domains empowers TFs to operate as highly versatile and adaptable components within the gene regulation machinery. By effectively coordinating the DNA binding domain, activation domain, and signal sensing domain, TFs can respond to a spectrum of internal and external cues. This enables them to exert precise control over gene expression in a dynamic, context-dependent manner. TFs exert their function through different strategies (**Figure 4.5 (B)**): (i) promoting or blocking RNA polymerase recruitment, (ii) shaping chromatin landscape, (iii) catalyzing histone deacetylation, and (iv) enhancing DNA-histone interactions. Transcription factors (TFs) wield the ability to either enhance the recruitment of RNA polymerase to gene promoter regions, thereby promoting the initiation of transcription, or act as inhibitors by impeding RNA polymerase access (Fuda *et al.*, 2009). TFs shape the chromatin landscape by actively modulating DNA-histone interactions. This regulatory action includes the capacity to weaken these interactions, thereby enhancing DNA accessibility. Certain TFs actively participate in histone deacetylation processes (Liu *et al.*, 2016). This involvement manifests as the removal of acetyl groups from histones, resulting in the promotion of a more compact chromatin structure. This compacted chromatin configuration contributes to the downregulation of gene transcription and acts as a mechanism suppressing gene expression. TFs exert their functions through the recognition and binding of short DNA sequences, typically spanning ∼6-20 nucleotides (Stewart *et al.*, 2012). These specific binding sites, referred to as transcription factor binding sites (TFBSs), serve as the molecular targets where TFs interact with the genome to regulate gene expression. The precise binding of TFs to these sites is a crucial step in initiating the regulatory processes that govern various cellular activities and responses. TFBSs are situated in crucial genomic regions, including gene promoters (Whitfield *et al.*, 2012), as well as more distal regulatory elements such as enhancers, silencers, or insulators (Gotea *et al.*, 2010; Lemon and Tjian, 2000; Nolis *et al.*, 2009). Although TFBSs frequently feature recurring sequence patterns, known as *motifs*, TFs exhibit a remarkable capacity to bind to sequences that are similar but not identical. This ability allows TFs to recognize and interact with a range of target sequences, often

differing by just a few nucleotides. The precise configuration of TFBS, along with the local chromatin structure, plays a pivotal role in finely tuning the regulatory functions of TFs within cells (Mendenhall *et al.*, 2013; Maurano *et al.*, 2015). During the DNA binding process, TFs utilize a combination of electrostatic and Van der Waals forces. While TFs demonstrate high specificity in binding to their target sequences, not every nucleotide within the binding site directly interacts with the TF. These interactions exhibit varying strengths, leading TFs to bind not to a single specific sequence but to a closely related subset of targets. However, the sequence composition of the TFBS decisively influences the strength of the TF-DNA interaction, known as *binding affinity*. Numerous studies have established connections between various diseases, cancer types, and genetic variants occurring within TFBS (Docquier *et al.*, 2005; Katainen *et al.*, 2015; Yu *et al.*, 2019). Additionally, variants within TFBS have the potential to disrupt the precise regulation of gene expression by TFs, exerting an impact on the entire cellular environment and possibly propagating effects to neighboring cells. The misregulation of gene expression governed by TFs, resulting from variants occurring in TFBS, could influence the broader cell environment and extend its effects to neighboring cells. Therefore, the identification of such regulatory motifs holds the key to fundamental insights into the intricate mechanisms that govern gene expression and the cellular environment. Several experimental techniques have been devised to identify TF binding site sequences, either within living cells or organisms (*in vivo*) or in controlled environments like test tubes using synthetic or purified components (*in vitro*) (Jolma and Taipale, 2011) (**Figure 4.6**). Early methods, such as electrophoretic mobility shift assay (EMSA) (Garner and Revzin, 1981) or footprinting (Hampshire *et al.*, 2007), typically focus on a limited number of target sequences for TFBS identification, resulting in small datasets of bound sequences. *In vitro* and *in vivo* high-throughput protocols like PBM, SELEX, or ChIP methods (Berger *et al.*, 2006; Jolma *et al.*, 2010; Collas and Dahl, 2008) have revolutionized the analysis of target sites for specific factors, yielding extensive datasets of bound sequences. This abundance of data offers an unprecedented opportunity to explore and characterize TF binding landscapes. Despite their capability to recover TF-bound sequences and their relative or absolute binding affinities, these experimental assays may mistakenly identify unbound sequences as binding sites. Moreover, they often capture additional nucleotides within target sites, compromising data resolution and presenting challenges for manual analysis. Computational frameworks, such as motif discovery algorithms, offer a means to analyze extensive datasets produced by experimental assays, unveiling sequences potentially bound by TFs and predicting their affinities (Pavesi *et al.*, 2004a; Tompa *et al.*, 2005; D'haeseleer, 2006; Das and Dai, 2007; Zambelli *et al.*, 2013; Tognon *et al.*, 2023). When provided with a sequence dataset, these algorithms usually identify sets of short and similar sequence elements. Subsequently, these prioritized sequence elements are employed to build a *motif model*, which summarizes the different configurations of binding sites observed in the prioritized sequences. The motif model encodes recurring patterns and similarities, summarizing the diversity of binding site configurations (**Figure 4.6**). Several methods and models have been proposed for the discovery and representation of TFBS motifs, with Position Weight Matrices (PWMs) (Stormo, 2000) emerging as the most prevalent models. PWMs, recognized for their simplicity, effectiveness, and interpretability, encode the likelihood of encountering a specific nucleotide at each TFBS position. Nevertheless, PWMs come with limitations, such as assuming independence among binding site positions. Hence, alternative motif models have been proposed (Siddharthan, 2010; Gorkin *et al.*, 2012; He *et al.*, 2021). The derived motif models find application in diverse downstream analyses, including the exploration of potential binding site occurrences in regulatory genomic sequences, forecasting the sets of genes regulated by the scrutinized TFs, or evaluating how genetic variants might impact their binding landscape. In the subsequent sections, we provide an overview of the current landscape in motif discovery, encompassing both traditional and contemporary experimental and computational approaches for the discovery and representation of TFBS motifs within DNA sequences. Our exploration delves into the innovations introduced by each algorithm and model, elucidating their respective merits and limitations. We also examine how researchers have addressed these limitations over time. We also delve into common downstream analyses utilizing motif models. Finally, we discuss current challenges and potential avenues for future research in the realm of developing innovative motif discovery algorithms and downstream analysis methods. In particular we focus on the future perspective of applicability in the precision medicine domain, highlighting our contributions in this context.

4.2.1 Experimental methods to discover Transcription Factor Binding Sites

Over the past few decades, different techniques have been introduced to experimentally identify and evaluate TF binding sites along with their binding preferences (Jolma and Taipale, 2011) (**Figure 4.6** and **Table 4.3**). Initial investigations into TF binding primarily concentrated on gene promoters (Stormo,

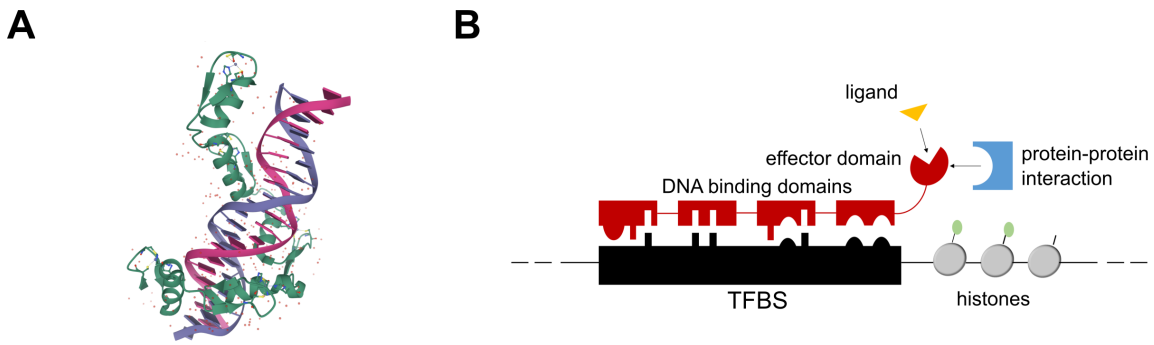


Figure 4.5. The human transcription Factors. (A) Graphical representation of a human transcription factor (CTCF) binding its target sequence across the genome. **(B)** The human transcription factor domains and the interactions with other elements in the within the cell environment (adapted from Lambert *et al.* (2018))

2000) and employed *in vitro* methods, including the Electro-Mobility Shift Assay (EMSA) (Garner and Revzin, 1981) and DNase footprinting (Galas and Schmitz, 1978). EMSA leverages the properties of non-denatured polyacrylamide gel to separate DNA sequences into bound and unbound fractions. DNase footprinting integrates EMSA with DNase I cleavage, identifying protected regions (footprints) due to the presence of the bound TF. Typically, these assays generate datasets comprising a few hundred bound sequences, providing a constrained view of the TFs' binding landscape. Additionally, technical limitations in EMSA and DNase footprinting may introduce inaccuracies in the reported sequences and binding preferences (Jolma and Taipale, 2011). The advent of Next-Generation Sequencing (NGS) technologies has transformed the exploration of TFBS identification, prompting researchers to devise approaches that harness the capabilities of massively parallel sequencing (**Figure 4.6**). These methods offer two major advantages: (i) they operate without requiring prior knowledge of the binding site sequence (Jolma and Taipale, 2011; Zia and Moses, 2012) and (ii) generate datasets comprising thousands of bound sequences, enabling more comprehensive characterization of TF binding preferences (Stormo and Zhao, 2010). Protein binding microarrays (PBMs) (Berger *et al.*, 2006; Berger and Bulyk, 2009) recover short TFBS sequences (∼10 base pairs) and assess TF binding preferences in vitro. In PBMs, a labeled TF is applied to a glass slide containing numerous spots, each filled with short, immobilized DNA sequences. Subsequently, the labeled TFs are incubated with fluorescent antibodies targeting the label, followed by washing to eliminate weakly bound factors. The fluorescence and DNA sequence enrichment are then utilized to quantify TF-DNA binding strength and capture the bound sequences. Generally, the recovered sequences lack nucleotides flanking the investigated binding sites, yielding high-resolution datasets. However, as the number of possible sequences grows with the target length, PBMs can assess only a limited number of target sequences (Jolma and Taipale, 2011; Zia and Moses, 2012). PBM analysis is typically restricted to binding sites ∼10–12 bp long. HT-SELEX, a widely employed *in vitro* technique, seamlessly integrates SELEX with high-throughput sequencing (Jolma and Taipale, 2011; Jolma *et al.*, 2010). This method involves releasing a TF onto a pool of randomized DNA sequences, allowing the factor to selectively pick its target sites. Subsequently, TF–DNA complexes are isolated from unbound sequences through affinity capture, followed by polymerase chain reaction (PCR) amplification and sequencing. The resulting DNA library, enriched in binding sites for the studied TF, serves as the initial pool for another SELEX run (Jolma and Taipale, 2011; Jolma *et al.*, 2010). Notably, SELEX does not necessitate prior knowledge of the target sites for the investigated factor (Jolma *et al.*, 2013). As the SELEX reaction is typically conducted in a liquid phase, free from physical constraints, the sequence space covered by HT-SELEX often surpasses that of PBMs. Additionally, through the integration of sequencing with DNA barcode indexing, HT-SELEX enables the parallel analysis of hundreds of TFs. It yields datasets comprising thousands of high-resolution bound sequences, letting only a few nucleotides surrounding the binding sites. However, since the starting DNA library consists of randomized sequences, HT-SELEX cannot pinpoint the genomic binding locations of the investigated factor. The advent of chromatin immunoprecipitation (ChIP) technologies (Collas and Dahl, 2008) has revolutionized the study of TFBS binding, allowing for the genome-wide identification of regions bound by TFs *in vivo*. In the ChIP method, TF–DNA complexes undergo cross-linking using formaldehyde. The DNA is then fragmented into ∼100–1000 bp long fragments and subsequently immunoprecipitated using antibodies specific to the investigated TF. To recover the bound sequences, the cross-links are reversed. The resulting fragments

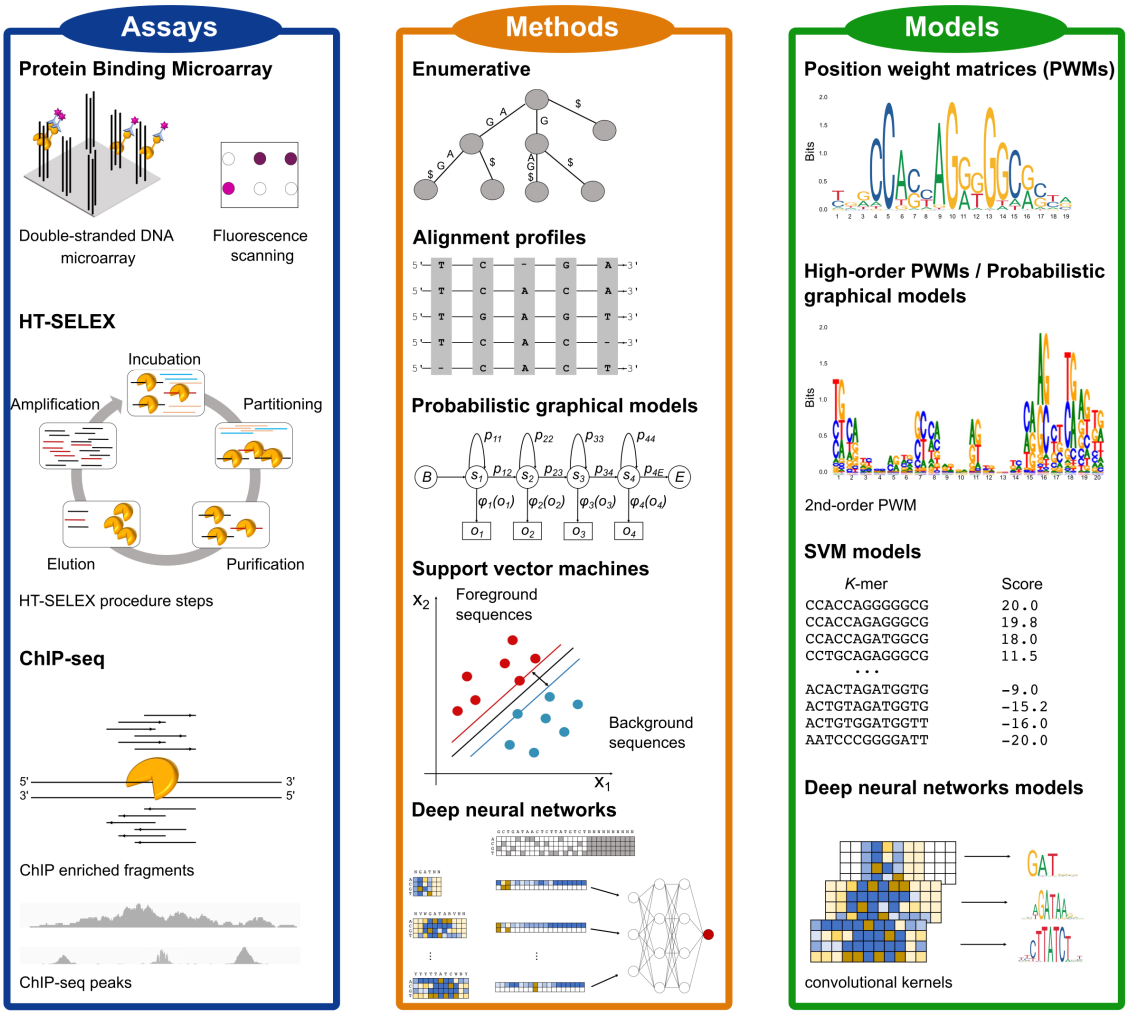


Figure 4.6. Experimental and computational methods to discover TFBS and popular models to represent binding site motifs. Protein binding microarray (PBM), HT-SELEX and ChIP-seq have become the most popular assays to determine TF binding preferences and identify their target sites (TFBS) in recent years. Computational motif discovery methods can be grouped into five classes, based on the algorithms employed to discover TFBS: enumerative, alignment-based, probabilistic graphical model-based, SVM-based and DNN-based methods. TFBS sequences prioritized by motif discovery algorithms are encoded in computational models representing the binding preferences of the investigated TFs.

can be amplified through either microarray hybridization (ChIP-on-Chip (Collas and Dahl, 2008; Pillai and Chellappan, 2015)) or sequencing (ChIP-seq (Johnson *et al.*, 2007; Mardis, 2007)). To locate binding regions, the recovered DNA fragments are mapped onto the genome. After mapping ChIP-seq reads, peak calling algorithms (Thomas *et al.*, 2017; Guo *et al.*, 2012; Zhang *et al.*, 2008) are employed to predict the genomic binding locations of the investigated factor. These algorithms identify genomic regions with higher enrichment in mapped DNA probes compared to a control experiment and designate those regions as binding locations or peaks (Pepke *et al.*, 2009). ChIP methods yield large datasets containing thousands of genomic regions, ranging from a few hundred to thousands of nucleotides, allowing the identification of potential TFBS for the studied factor. Despite ChIP technologies, especially ChIP-seq, being considered the current 'golden standard,' they have some limitations. ChIP can detect indirect binding, identifying other TFBS not belonging to the investigated factor (Worsley Hunt and Wasserman, 2014). ChIP-seq peaks may be false positives, arising from poor antibody quality (Pickrell *et al.*, 2011). Finally, ChIP-seq returns low-resolution datasets, including several nucleotides flanking the target TFBS. To address the latter issue, ChIP-exo (Rhee and Pugh, 2011) utilizes lambda exonuclease to trim ChIP sequences, removing some of the nucleotides flanking the target sites. Alternatively, experimental assays targeting open chromatin regions, such as ATAC-seq (Buenrostro *et al.*, 2013) or DNase-seq (John *et al.*, 2011), can be employed to recover *in vivo* genomic locations likely to contain TFBS, as most TFs bind their target sequences in open chromatin regions. ATAC-seq and DNase-seq are generally used when the factors binding the target regions are not known. In summary, the current high-throughput *in vivo* and *in vitro* assays generate datasets comprising thousands of sequences potentially containing various

Experimental assay	Description	Output	*De novo* motif discovery capability	Type	Identification of genomic binding locations	Throughput
Competition EMSA	Bound DNA sequences are identified by observing changes in the electrophoretic migration of DNA sequences through non-denatured polyacrylamide gel	Bound DNA sequences	No. Used to validate known binding sites	*in vitro*	No	Low
DNase footprinting	Pools of DNA sequences are incubated with the TF of interest; then, the DNA is degraded using DNase I. The unbound fragments are cut in all positions, while the bound DNA is protected by the TF	Bound DNA sequences	No. Used to validate known binding sites	*in vitro*	No	Low
Protein Binding Microarrays	Arrays of ∼40 000 spots with short, immobilized DNA sequences are incubated with a tagged TF, and then washed to remove weakly bound proteins. The bound sequences are identified through fluorescence-based detection & Continuous values describing fluorescence intensity on each array spot	Continuous values describing fluorescence intensity on each array spot	Yes. Limited to short motifs	*in vitro*	No	High
HT-SELEX	The TF is added to a pool of randomized DNA fragments. The bound sequences are selected and constitute the starting pool for the next experimental round. The procedure is repeated for several rounds. Sequencing is employed to recover the sequence of the bound DNA fragments	DNA sequences	Yes	*in vitro*	No	High
ChIP-based technologies	TF–DNA complexes are cross-linked with formaldehyde and immunoprecipitated employing TF-specific antibodies. The bound sequences are then prioritized employing qPCR microarrays (ChIP-on-Chip) or through sequencing (ChIP-seq). ChIP-exo integrates exonuclease treatment to enhance sequence resolution	Genomic binding location coordinates	Yes. Limited by the inability to distinguish direct and indirect binding	*in vivo*	Yes	Low

Table 4.3. *In vivo* and *in vitro* experimental assays to identify and validate transcription factor binding sites. Traditional techniques like EMSA or DNase footprinting are employed for validating established TFBS, whereas modern methods such as PBMs, HT-SELEX, and ChIP-based approaches are favored for uncovering new binding sites. Among these, ChIP-based assays stand out as the sole methods capable of retrieving the genomic locations where TFs bind. The "throughput" designation denotes the number of samples each method can process simultaneously (high: hundreds of samples; low: a few samples) (Adapted from Tognon *et al.* (2023)).

possible binding configurations of TFBS, thereby enabling more comprehensive characterizations of TFs binding landscapes.

4.2.2 Computational methods and models to discover and represent Transcription Factor Binding Sites

As described in the previous section, experimental assays recover large binding site sequence datasets. However, the reported sequences often include unbound sequences and contain additional nucleotides flanking the target sites. These limitations add a layer of complexity and make manual analyses and curations challenging and often unfeasible. Motif discovery algorithms address this challenges, by providing a computational framework to analyse the datasets produced by experimental assays. Motif discovery algorithms discover potential binding sites in the analyzed sequence datasets and construct computational models representing TFBS. The problem of TFBS motif discovery can be formalized as follows. Given a set of positive DNA sequences S, obtained from an experimental assay targeting a specific TF, and a set of negative sequences B, the objective is to identify one or more recurrent, short, and similar subsequences in S that maximize the discriminatory power between S and B (Tognon *et al.*, 2023). These identified subsequences are referred to as patterns or motifs and are likely to be bound by the investigated TF. The negative set B may consist of randomly generated or carefully selected genomic sequences with similar nucleotide content and length to those in S. The patterns identified are then employed to construct and train a computational model M (motif model) that represents the discovered motif. These models can then be employed to identify potential new binding sites and predict the strength of the TF–DNA binding when presented with a new set of sequences. Motif discovery can be viewed as either a classification or a regression problem, depending on the type of data used for training (Tognon *et al.*, 2023). Datasets from experimental assays, such as ChIP-seq or HT-SELEX, offer hundreds or thousands of sequences containing binding sites, transforming motif discovery into a classification problem. In this scenario, the objective is to differentiate between bound and unbound sites in the input sequences and train the motif model with the identified binding sites. On the other hand, datasets from other experimental technologies like PBMs provide relative binding strength for large sets of sequences of equal length. In this case, rather than distinguishing between bound and unbound sequences, M learns the relative binding affinities associated with each target site in the input dataset, making motif discovery a regression problem. In both settings, the ultimate goal is to derive a computational model M that characterizes the recovered TFBS and can predict new binding events, along with their affinity, in sequences not used during model training. Motif discovery algorithms can be classified into enumerative, alignment-based, probabilistic graphical models, support vector machine (SVM)-based, and deep neural network-based

methods (**Figure 4.6** and **Appendix B.2**). Alternative approaches for identifying TFBS motifs in genomic sequences involve phylogenetic footprinting (McCue *et al.*, 2001; Blanchette and Tompa, 2002). The fundamental principle of phylogenetic footprinting posits that functional elements, such as TFBS, are more likely to be conserved across evolutionarily related species, while non-functional elements are more susceptible to mutations. Despite being one of the earliest techniques suggested for TFBS identification, phylogenetic footprinting remains widely used for assessing TFBS conservation across different organisms (Balazadeh *et al.*, 2011; Xu *et al.*, 2012; Katara *et al.*, 2012). In a recent study (Glenwinkel *et al.*, 2014), the authors proposed a novel method that employs phylogenetic footprinting to discover TFBS. Before delving into the algorithms, we provide an overview about the models used to describe TFBS motifs. The prevalent models for representing TFBS include consensus sequences (Day and McMorris, 1992), PWMs (Stormo, 2000, 2013), high-order PWMs (Siddharthan, 2010; Korhonen *et al.*, 2017), SVM-based (Gorkin *et al.*, 2012), and deep neural network-based (He *et al.*, 2021) models. Consensus sequences succinctly characterize identified TFBS by indicating the most frequently observed nucleotide at each motif position within a prioritized sequence set. While TFBS possess conserved positions intolerant to mutations (Li and Ovcharenko, 2015), other TFBS locations admit alternative nucleotides. Degenerate consensus accommodates ambiguous motif positions using IUPAC symbols. However, consensus sequences are unable to capture the contribution of each nucleotide at every motif position to TF–DNA binding. PWMs address this limitation by offering an additive model that considers the contribution of each motif position to the binding site. PWMs construct an ungapped alignment between candidate motif sequences and tally the frequency of each nucleotide at each position. The statistical significance of PWMs is often assessed using relative entropy (RE) (Stormo, 1998). RE measures the difference between computed nucleotide frequencies and those derived from aligning random sequences. PWMs are visualized as logos (Schneider and Stephens, 1990), where the height of each nucleotide is proportional to its RE. Despite their widespread success, PWMs still assume independence between motif positions. To address this limitation, probabilistic graphical models model dependencies between motif nucleotides. These models encompass high-order PWMs such as dinucleotide weight matrices (DWMs), Bayesian networks (BNs), Markov models (MMs), or hidden Markov models (HMMs) (Siddharthan, 2010; Korhonen *et al.*, 2017; Barash *et al.*, 2003; Siebert and Söding, 2016). DWMs and high-order PWMs are commonly visualized as logos with q-mers replacing single nucleotides, where q represents the dependency order between adjacent nucleotides. Notably, probabilistic graphical models can accommodate variable spacing between half-sites of two box motifs (Mathelier and Wasserman, 2013). However, the number of model parameters and their complexity increase exponentially with q, often leading to models overfitting the input dataset. SVM-based models use an SVM kernel to learn the binding site structure from the input sequence dataset. TFBS is represented either by a list of k-mers with associated weights or support vectors used to discriminate between bound and unbound sequences, depending on the employed kernel (Boeva, 2016). In the former case, the weights reflect the k-mer contribution to the motif sequence. SVM-based models can accommodate variable spacing between the half-sites of two box motifs, similar to probabilistic graphical models. Importantly, k-mers indirectly capture k-th order dependencies between neighboring nucleotides. However, simple SVM-based models are restricted to considering short ($\sim$10 bp) and cannot effectively represent longer motifs. Gapped k-mers (Ghandi *et al.*, 2014b) addressed this limitation by handling longer TFBS and accounting for sequence degeneration in non-informative motif positions. To visualize the discovered motifs, SVM-based models are often simplified to PWMs by aligning the informative k-mers. Deep neural network (DNN)-based models integrate the diverse, complex, and hierarchical patterns governing TF–DNA binding events in input nucleotide sequences. Despite their accuracy and power, a major limitation of DNN-based models is their 'black box' nature (Park *et al.*, 2020). Many frameworks visualize the discovered motifs as PWMs, computed by aligning the sequences activating the convolutional kernels of the DNN (Koo and Ploenzke, 2020). However, DNNs often learn distributed representations where multiple neurons cooperate to describe single patterns. Consequently, motifs learned by single kernels and the resulting PWMs tend to be redundant with each other. DeepLIFT (Shrikumar *et al.*, 2017) introduced a method to assign importance scores to the kernels. By comparing the activation of each neuron to a reference value, DeepLIFT identifies which kernels contribute most to defining the TFBS, thereby reducing motif redundancy. TF-MoDISco (Avsec *et al.*, 2021a) extended this idea by clustering and aggregating the discovered motifs, using the importance scores assigned to the kernels. However, generating interpretable models without sacrificing some information learned by the DNN remains an open challenge.

Enumerative methods

Enumerative algorithms for motif discovery (**Figure 4.6**) assume that motifs are patterns overrepresented in the input dataset S, compared to a background set of genomic sequences B. Enumerative algorithms typically assume that the motif length $|M|$ is known *a priori*. The fundamental idea involves collecting the approximate occurrences of all potential 4^k k-mers within the sequences of S and evaluating the statistical significance by comparing the observed matches in S with those expected from a background model. Subsequently, a PWM is derived by constructing an ungapped alignment from the statistically significant k-mers. However, the exhaustive search for approximate occurrences of all k-mers quickly becomes impractical, particularly for larger S. Early methods incorporated heuristics to address this challenge, such as limiting the search to patterns occurring at least once in each sequence $s \in S$ (Li *et al.*, 1999) or restricting mismatching locations to specific motif positions (Califano, 2000). Nonetheless, allowing mismatches at any motif position is crucial. To address this, Weeder (Pavesi *et al.*, 2001, 2004b) and SMILE (Marsan and Sagot, 2000) proposed using suffix trees (STs) (Weiner, 1973) to efficiently navigate the entire motif search space. Leveraging the indexing capabilities of STs enables the algorithms to conduct approximate pattern matching without imposing restrictions on mismatching positions. This approach ensures high accuracy in motif discovery while mitigating computational costs. For assessing the statistical significance of motif candidates, SMILE and Weeder compare motif frequencies in S with those in a set of randomly generated genomic sequences or the promoters of the same organism, respectively (**Appendix B.1.1**). However, these strategies often prove computationally demanding and lack scalability when applied to the extensive datasets produced by PBMs, HT-SELEX, or ChIP assays (Liu *et al.*, 2018). As a result, more efficient methods tailored for large datasets have been proposed. MDscan (Liu *et al.*, 2002) and Amadeus (Linhart *et al.*, 2008) employ word enumeration to identify potential motifs in sequence datasets (**Appendix B.1.1**). MDscan leverages the shape of ChIP peaks to identify non-redundant patterns that are abundant in the most enriched sequences, employing a third-order Markov background model to gauge the statistical significance of motifs. Amadeus assesses all k-mers in S and organizes similar patterns into lists, which are then grouped into motifs and statistically evaluated using a hypergeometric test. However, word enumeration may still be computationally demanding. In response to this, DREME (Bailey, 2011) introduced the use of regular expressions to approximate motif frequencies in S and B. To evaluate motifs' statistical significance, DREME employs Fisher's exact test, comparing the number of motif occurrences detected in S and B. Nevertheless, regular expressions can be computationally expensive for large S, potentially leading to false positives or missing motifs. Trawler (Ettwiller *et al.*, 2007), HOMER (Heinz *et al.*, 2010), and STREME (Bailey, 2021) reintroduced STs and proposed different optimizations to enhance scalability for large datasets (**Appendix B.1.1**). Trawler and HOMER improved the statistical significance assessment step using z-scores derived from the normal approximation to the binomial distribution and the hypergeometric distribution, respectively. In contrast, STREME reduces the motif search space by initially identifying overrepresented seed words of different lengths on the ST. Subsequently, STREME counts the number of approximate matches for the most significant words on the ST. By identifying seeds of varying lengths, STREME can discover motifs of different lengths in a single tree visit.

Alignment-based methods

Alignment-based motif discovery algorithms generate alignment profiles to characterize motifs' binding preferences (**Figure 4.6**), avoiding the exhaustive enumeration of k-mers. In this approach, an alignment is constructed by selecting motif candidate sequences from the input dataset S and evaluating the resulting profile using various metrics, such as nucleotide conservation, information content, or the statistical significance of the profile. The motif's statistical significance is assessed by computing the probability of obtaining the same alignment from a background dataset B or random sequences. Typically, alignment-based motif discovery algorithms assume that the motif length $|M|$ is known in advance. Formally, for alignment-based algorithms, motif discovery can be viewed as a combinatorial problem. Given $|M| = k$, the goal is to identify the optimal alignment profile by combining k-mers from S, according to a scoring criterion. The best alignments are then used to generate the corresponding PWMs. Most alignment-based algorithms assume that each sequence in S contains either zero or one binding site, resulting in $(\sum_{s \in S} |s| - |M| + 1)^{|S|}$ possible profiles built combining k-mers in all possible ways. However, enumerating all potential solutions is computationally impractical, even for small datasets. As a solution, alignment-based algorithms employ heuristics such as greedy strategies (Hertz and Stormo, 1999), expectation–maximization (EM) (Bailey *et al.*, 1994), stochastic methods (e.g., Gibbs sampling) (Lawrence *et al.*, 1993), or genetic algorithms (Lee *et al.*, 2018) (**Appendix B.1.2**). CONSENSUS

(Hertz and Stormo, 1999) introduced a greedy strategy to incrementally construct alignment profiles. It initially solves the problem for two sequences and gradually expands by incorporating the remaining sequences one by one. CONSENSUS stores the best partial alignments with the hope of identifying the highest-scoring profiles. However, in cases where motifs lack conservation, CONSENSUS might discard potentially highest-scoring solutions. The MEME algorithm (Bailey *et al.*, 1994; Bailey and Elkan, 1995; Bailey *et al.*, 2006) adopts a distinct strategy based on Expectation-Maximization (EM). It iteratively refines an initial profile by substituting some k-mers in the profile with others more likely to yield improved solutions. MEME evaluates the fit of each k-mer in the dataset to the current profile, bypassing the need for a background model. MEME identifies motifs occurring multiple times in each sequence and calculates their statistical significance without relying on TFBS conservation. However, the algorithm may prematurely converge to local maxima, and convergence heavily relies on the initial conditions of the algorithm. In contrast to MEME, Gibbs sampling (Lawrence and Reilly, 1990) employs a stochastic approach to add k-mers to the alignment rather than a deterministic one based on profile fit. Gibbs sampling replaces k-mers in the profile with others selected with a probability proportional to their likelihood score (**Appendix B.1.2**). The stochastic nature of the Gibbs sampling algorithm reduces the likelihood of converging to local maxima, but achieving reliable results may necessitate multiple runs. Despite this, several methods employing Gibbs sampling and its extensions have been proposed (Neuwald *et al.*, 1995; Hughes *et al.*, 2000; Workman and Stormo, 1999; Liu *et al.*, 2000; Thijs *et al.*, 2001; Frith *et al.*, 2004b, 2008) (**Appendix B.1.2**). Genetic algorithms offer an alternative approach to address the limitations of EM and stochastic methods. GADEM (Li, 2009) combines EM local search with genetic algorithms to refine profiles, preventing convergence to local maxima and mitigating the stochastic nature of Gibbs sampling. However, due to their computational complexity, genetic algorithms pose challenges when analyzing datasets containing thousands of sequences. Since the solution space for algorithms using alignment profiles grows exponentially with the size of S, and even with the use of heuristics, analyzing thousands of sequences becomes computationally impractical (Zambelli *et al.*, 2013). Therefore, researchers have directed their efforts toward developing algorithms specifically tailored to analyze the large datasets produced by high-throughput assays (**Appendix B.1.2**). MEME-ChIP (Machanick and Bailey, 2011) and STEME (Reid and Wernisch, 2011) enhance the MEME algorithm for ChIP datasets. While MEME-ChIP focuses the analysis on a random subset of sequences, STEME accelerates EM steps by indexing the sequences in a suffix tree. However, using random subsets of S may lead to missing critical motif instances, and constructing a suffix tree from thousands of sequences may be computationally demanding. ChIPMunk (Kulakovskiy *et al.*, 2010) introduces a greedy profile optimization, similar to EM, designed to discover motifs in large ChIP-seq datasets, while accounting for ChIP peak shapes. XXmotif (Hartmann *et al.*, 2013) and ProSampler (Li *et al.*, 2019b) propose methods that combine enumerative motif discovery with iterative and stochastic profile refinement, respectively.

Probabilistic graphical model-based methods

Including dependencies between nucleotides within TFBS has been a topic of discussion within the community (Tomovic and Oakeley, 2007; Morris *et al.*, 2011; Zhao and Stormo, 2011). Some studies have demonstrated the existence of dependencies between both neighboring and non-neighboring nucleotides in TFBS (Slattery *et al.*, 2014; Rohs *et al.*, 2010). Enumerative and alignment-based algorithms typically represent motifs as PWMs, which do not inherently incorporate dependencies between binding site positions. Although PWMs can be extended to account for di- or trinucleotide frequencies (high-order PWMs), such as DWMs (Siddharthan, 2010), methods like Dimont (Grau *et al.*, 2013) and diChIPMunk (Kulakovskiy *et al.*, 2013a) have proposed extensions to alignment-based approaches, representing motifs as DWMs to capture dependencies, albeit primarily focusing on neighboring nucleotides (**Appendix B.1.3**). Probabilistic graphical models (**Figure 4.6**), including Bayesian Networks (BNs), Markov Models (MMs), or Hidden Markov Models (HMMs), offer robust frameworks for capturing dependencies between nucleotides within TFBS. A study (Barash *et al.*, 2003) suggested using BNs trained through Expectation-Maximization (EM) to model TFBS, capturing dependencies between both neighboring and non-neighboring positions. However, this method assumes a consistent order of dependence throughout the entire motif. Another approach, Variable Order Bayesian Networks (VOBNs) (Ben-Gal *et al.*, 2005), utilizes BNs considering variable orders of dependencies between positions. Nonetheless, training BNs can be computationally challenging when dealing with thousands of sequences, and these models may be prone to overfitting even when trained on hundreds of sequences. MMs and HMMs offer more efficient and scalable frameworks compared to BNs for incorporating dependencies between motif positions. Consequently, recent algorithmic developments have leaned towards models like TFFMs (Mathelier and Wasserman, 2013) and Discrover (Maaskola and Rajewsky, 2014), proposing HMM-based models learning

dinucleotide dependencies between neighboring motif positions in extensive sequence datasets. TFFMs, for instance, also learn characteristics of the sequences flanking the TFBS. MMs have been extended to capture various orders of dependencies between neighboring nucleotides, as demonstrated in a study (Eggeling *et al.*, 2014) focusing on discovering CTCF (Bell *et al.*, 1999) motifs using variable-order MMs. Similarly, Slim (Keilwagen and Grau, 2015) has extended MMs to capture dependencies between non-neighboring nucleotides. However, MMs and HMMs typically only capture low-order dependencies. In contrast, BaMMotif (Siebert and Söding, 2016; Ge *et al.*, 2021) has proposed a motif discovery algorithm using a Bayesian approach to efficiently train Markov models with dependencies up to the fifth order, even on large datasets comprising thousands of sequences.

SVM-based methods

Support Vector Machines (SVMs) (Boser *et al.*, 1992) have proven successful in various computational biology problems, including the discovery of transcription factor binding site (TFBS) motifs (**Figure 4.6**). This is achieved by decomposing bound sequences (foreground dataset S) and unbound sequences (background dataset B) into k-mers, using their frequencies as features to train a sequence similarity kernel (Ben-Hur *et al.*, 2008). Typically, each k-mer is assigned a weight proportional to its contribution to defining positive or negative training sets or its likelihood of being a motif candidate. While earlier methods were designed for protein sequence homology (Leslie *et al.*, 2001; Eskin *et al.*, 2002; Kuang *et al.*, 2005), recent SVM-based algorithms are designed specifically for TFBS motif discovery. SVMs demonstrate efficiency in analyzing datasets comprising thousands of sequences. Kmer-SVM (Lee *et al.*, 2011; Fletez-Brant *et al.*, 2013) proposed a method for discovering TFBS motifs in sequence datasets employing the spectrum kernel. By counting exact matches for all contiguous k-mers in S and B, Kmer-SVM constructs the k-mers feature space (**Appendix B.1.4**). The introduction of mismatch (Kuang *et al.*, 2005) and wildcard (Leslie and Kuang, 2003) kernels allowed counting k-mer frequencies while allowing a fixed number of mismatching positions for each k-mer. This concept was later expanded to enable less restrictive k-mer frequency estimation, providing flexibility in motif structure without compromising scalability on large datasets. Agius and colleagues (Agius *et al.*, 2010) further extended this concept by developing the di-mismatch kernel, a first-order Markov mismatch kernel based on the dinucleotide alphabet. This kernel addresses sequence variability and considers dependencies between neighboring nucleotides (**Appendix B.1.4**). To ensure scalability on large datasets, a small k ($k = 4$) is employed, which enables the discovery of short motifs. However, TFBS lengths typically range between 6 to 20 base pairs, making it challenging to fully characterize longer motifs with short k-mers. Additionally, increasing k often leads to sparse feature vectors overfitting to the training dataset. Gapped k-mers (Ghandi *et al.*, 2014b) proposed representing longer motifs as k-mers with gaps in non-informative or degenerate TFBS positions, accounting for motif variability in sequence and length. Gkm-SVM (Ghandi *et al.*, 2014a, 2016) extends Kmer-SVM to train SVM kernels employing gapped k-mers as features. The algorithm considers larger k, preventing model overfitting and reducing dependency on parameter choice. LS-GKM (Lee, 2016) optimizes the algorithm for scalable SVM training with gapped-k-mers on large-scale sequence datasets and provides various kernels for SVM training (**Appendix B.1.4**).

Deep Neural Networks-based methods

DNNs have gained significant popularity in computational biology (Talukder *et al.*, 2021; Zeng *et al.*, 2020b; Singh *et al.*, 2016, 2019; Zeng *et al.*, 2018; Kelley *et al.*, 2018; Li *et al.*, 2019a; Yin *et al.*, 2019; Manzanarez-Ozuna *et al.*, 2018) owing to their ability to discern complex patterns (Park and Kellis, 2015) from large omics datasets (Zhang *et al.*, 2019). Originally designed for image classification (LeCun *et al.*, 2015; Sainath *et al.*, 2013; Vu *et al.*, 2017), Convolutional Neural Networks (CNNs) (LeCun *et al.*, 2015) have proven successful in analyzing *in vivo* TF–DNA interactions (Alipanahi *et al.*, 2015; Zhou and Troyanskaya, 2015; Kelley *et al.*, 2016; Zeng *et al.*, 2016a) (**Figure 4.6**). By applying non-linear transformations to input data, CNNs learn and represent complex patterns in a high-dimensional space (Bengio *et al.*, 2013), simplifying classification tasks and enabling accurate prediction of TFBS in genomic sequences. Genomic sequences are often represented as 1D or 2D images with four associated channels (A, C, G, T) (Zeng *et al.*, 2016a), transforming TFBS classification into a two-class image classification problem. Typically, CNN architectures for motif discovery and classification consist of sets of four layers: the convolutional layer, max-pooling layer, fully connected NN layer, and the output layer (Zeng *et al.*, 2016a) (**Appendix B.1.5**). DeepBind (Alipanahi *et al.*, 2015) and Basset (Kelley *et al.*, 2016) introduced two CNN architectures for motif discovery across different datasets, including ChIP-seq, HT-SELEX, PBM, and DNase-seq (**Appendix B.1.5**). The motifs discovered by DeepBind and Basset

are visualized as PWMs, computed by aligning and grouping sequences that activate the convolutional layer. While DeepBind and Basset have shown promising results, their performance may be constrained by the quality of training data and the significant computational resources and time required for model training. These challenges have spurred the development of novel methods like BPNet (Avsec *et al.*, 2021a), which address some of these issues by incorporating additional features into the model and using more efficient training processes. BPNet proposed a dilated CNN architecture that enables the model to learn and integrate diverse complex features without sacrificing the spatial and base resolution of the input data (**Appendix B.1.5**). However, TF–DNA interactions involve not only direct binding but also interactions between multiple binding subregions (long-term interactions) and nucleotides with high-order structures of TFs (short-term interactions). Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and Bidirectional LSTMs (BLSTMs) efficiently capture long-term and short-term dependencies in sequential signals. Given that genomic sequences can be viewed as sequential signals with long-term and short-term dependencies, LSTMs and BLSTMs are well-suited for modeling TF–DNA interactions (**Appendix B.1.5**). DeeperBind (Hassanzadeh and Wang, 2016) introduced a hybrid CNN–LSTM architecture by removing the pooling layer to maintain positional information of potential motif instances. Similarly, DanQ (Quang and Xie, 2016) proposed a hybrid CNN–BLSTM architecture to capture the positional dynamics of genomic sequences for TFBS motif discovery. The BLSTM replaces the fully connected NN. Factornet (Quang and Xie, 2019) extended DanQ's framework by incorporating additional features and employing a Siamese BLSTM architecture to enhance model training.

4.2.3 Transcription Factor Databases

With the recent advancements in experimental technologies, an extensive amount of transcription factor-related data has been generated and has been stored in different databases (**Table 4.4**). The ENCODE project (Consortium *et al.*, 2012) stands out as a valuable resource, offering a multitude of data on functional elements within the human genome, gathered across different tissues and cell types. ENCODE stores several TF-related genomic data encompassing ChIP-seq data targeting various TFs and DNase-seq information. Similarly, Cistrome (Zheng *et al.*, 2019) and GTRD (Kolmykov *et al.*, 2021) contribute to the landscape by providing TF-related genomic data from different organisms and spanning various species, cell types, and tissues. GTRD, in particular, provides extensive collections of curated ChIP-seq, ChIP-exo, and ChIP-nexus datasets. HOCOMOCO (Kulakovskiy *et al.*, 2013b, 2018) and JASPAR (Sandelin *et al.*, 2004; Fornes *et al.*, 2020) provide large collections of curated and experimentally derived TFBS motifs. These databases store Position Weight Matrices (PWMs) and Dinucleotide Weight Matrices (DWMs) obtained through the analysis of ChIP-seq and SELEX datasets. Notably, HOCOMOCO models integrate sequence datasets with evolutionary conservation and DNA shape. Similarly, Cis-BP (Weirauch *et al.*, 2014) incorporates experimentally derived and computationally predicted PWMs from various sources, including published literature, other databases, and experimental datasets. TRANSFAC (Wingender *et al.*, 1996, 2000) emerges as a comprehensive repository, collecting experimentally validated and manually curated PWMs for different TFs from diverse eukaryotic organisms. It includes data on TF-associated proteins, DNA binding domains, and regulatory elements. FactorBook (Pratt *et al.*, 2022) provides computationally predicted PWMs generated by analyzing ENCODE data, coupled with TF expression data across tissues and cell types. Unibind (Puig *et al.*, 2015) collects experimentally validated and curated PWMs from different organisms. It provides insights on the structural properties and conformation of TF–DNA complexes, along with their genomic binding locations across various cell types and tissues. UniPROBE (Newburger and Bulyk, 2009) stores curated PWMs for several eukaryotic TFs, generated by analyzing Protein Binding Microarray datasets. HTRIdb (Bovolenta *et al.*, 2012) focuses on data pertaining to TF–target gene interactions in humans. This information, derived from published literature and other databases, spans different cell types, experimental methods, and disease states. The database also provides functional annotations for the target genes. Similarly, DoRothEA (Garcia-Alonso *et al.*, 2019) offers a comprehensive resource of regulatory TF–target interactions by integrating diverse datasets and assigning confidence scores based on experimental evidence. DoRothEA includes TF–target interactions from literature-curated databases, top interactions identified from ChIP-seq experiments, motif scans, and regulons inferred from both normal tissues and cancer types. The normal tissue collection comprises 1 million interactions between 1,402 TFs and 26,984 targets, while the pancancer collection includes 636,753 interactions between 1,412 TFs and 26,939 targets. DoRothEA assigns confidence scores to the inferred TF–target interactions, with the most reliable interactions being supported by multiple sources. TFcancer (Huang *et al.*, 2021b) collects TF–gene interactions across 33

Type	Name	Reference	Data type	Model Organisms	TFs
Sequence database	ENCODE	Consortium *et al.* (2012)	ChIP-seq DNase-seq ATAC-seq	*C. elegans* *D. melanogaster* *H. sapiens* *M. musculus*	>1,500
	Cistrome	Zheng *et al.* (2019)	ChIP-seq DNase-seq	*H. sapiens* *M. musculus*	1,773 (ChIP-seq)
	GTRD	Kolmykov *et al.* (2021)	ChIP-seq ChIP-exo ChIP-nexus DNase-seq	*A. thaliana* *C. elegans* *D. rerio* *D. melanogaster* *H. sapiens* *M. musculus* *R. norvegicus* *S. cerevisiae* *S. pombe*	3,988 (ChIP-seq) 1,708 (ChIP-exo + ChIP-nexus)
Motif models database	HOCOMOCO	Kulakovskiy *et al.* (2013b, 2018)	PWMs DWMs	*H.sapiens* *M. musculus*	680 (human) 453 (mouse)
	JASPAR	Sandelin *et al.* (2004); Fornes *et al.* (2020)	PWMs DWMs	53 species	>1,500
	Cis-BP	Weirauch *et al.* (2014)	PWMs	*A. thaliana* *C. elegans* *D. rerio* *D. melanogaster* *H. sapiens* *M. musculus* *N. crassa* *R. norvegicus* *S. cerevisiae* *X. tropicalis*	>5,000
	TRANSFAC	Wingender *et al.* (1996, 2000)	PWMs	>300 species	>10,000
	FactorBook	Pratt *et al.* (2022)	PWMs	*H. sapiens* *M. musculus*	881 (human) 49 (mouse)
	UniPROBE	Newburger and Bulyk (2009)	PWMs	*C. elegans* *C. parvum* *H. sapiens* *M. musculus* *P. falciparum* *S. cerevisiae* *V. harveyi*	726
	Unibind	Puig *et al.* (2015)	PWMs	*A. thaliana* *C. elegans* *D. rerio* *D. melanogaster* *H. sapiens* *M. musculus* *R. norvegicus* *S. cerevisiae* *S. pombe*	841
TF-target gene interaction database	HTRIdb	Bovolenta *et al.* (2012)	TF-gene interaction networks	*H. sapiens*	284
	DoRothEA	Garcia-Alonso *et al.* (2019)	TF-gene interaction networks	*H.sapiens* *M. musculus*	1,402
TF-disease association database	TFcancer	Huang *et al.* (2021b)	TF-cancer associations	*H. sapiens*	364

Table 4.4. Transcription Factor Databases. The table provides a condensed overview of the TF-related databases discussed in **Section 4.2.3**. It outlines key aspects for each database, including its primary purpose (**Type**), the variety of available data types (**Data type**), the model organisms covered by the database (**Model organisms**), and the total number of included TFs (**TFs**).

cancer types. The database offers tools to identify TF expression alterations and elucidate their roles in biological processes and signaling pathways associated with cancer.

4.2.4 Downstream analysis using Transcription Factor Binding Site motifs

The discovered motifs serve various purposes in downstream analyses, including motif comparison, motif scanning, motif enrichment analysis, and evaluating the effects of genetic variants on TF–DNA binding affinity (Tognon *et al.*, 2023). Motif comparison measures the similarity between the discovered motifs and annotated TFBS, facilitating the connection of known TFs to newly identified motifs (Gupta *et al.*, 2007). Tools such as Tomtom (Gupta *et al.*, 2007), STAMP (Mahony and Benos, 2007), MACRO-APE (Vorontsov *et al.*, 2013), or MoSBAT (Lambert *et al.*, 2016) are commonly used for this task, searching annotated databases for motifs matching the input consensus sequence or inferred motif matrix. Additionally, motif comparison tools have been developed to interpret and annotate potential motifs encoded in the convolutional filters of CNN models. Motif scanning involves searching sets of genomic regions for potential occurrences of the input motif, aiming to recover potential binding locations for the investigated factor. Motif scanning algorithms assign scores to sequences using the input model (e.g., a

PWM). A challenge in motif scanning is the choice of a reliable cutoff for discriminating between true and false binding events (Boeva, 2016). Several motif scanning tools are available, such as (Korhonen *et al.*, 2009), FIMO (Grant *et al.*, 2011), PWMscan (Ambrosini *et al.*, 2018), and the HOMER suite (Heinz *et al.*, 2010). Recently, MOODS has been extended to search instances of motifs modeled as high-order PWMs (Korhonen *et al.*, 2017). However, motif scanning algorithms generally do not consider individual- and population-specific genetic variants and haplotypes while searching for potential motif occurrences. This may result in several potential TFBS missed and missing information which could facilitate the interpretation of genetic variants impact within the cell. Motif enrichment analysis (MEA) investigates over- and underrepresented motifs in gene regulatory regions, linking investigated TFs to their functions within the cell environment. MEA involves scanning regulatory regions for motif occurrences and statistically testing motif enrichment. There are many MEA tools available, such as Clover (Frith *et al.*, 2004a), Pscan (Zambelli *et al.*, 2009), AME (McLeay and Bailey, 2010), or oPOSSUM-3 (Kwon *et al.*, 2012) are commonly used for MEA. The HOMER suite provides MEA functionality. Haystack (Pinello *et al.*, 2018) proposed an integrated MEA strategy, exploring motif enrichment in cell type-specific regions and incorporating gene expression data to assess the transcriptional activity of studied factors and their impact on regulated genes. ChEA (Lachmann *et al.*, 2010; Kou *et al.*, 2013; Keenan *et al.*, 2019) is a web-based MEA tool that provides high-fidelity predictions of TF-gene interactions. To enhance prediction quality, ChEA incorporates benchmarked TF-target gene set libraries derived from ChIP-seq experiments, RNA-seq co-expression data, TF co-occurrence, and single TF perturbation studies.

4.2.5 Evaluating genetic variants impact on Transcription Factor Binding Sites

Genetic variants have been demonstrated to influence TF–DNA binding events (De Gobbi *et al.*, 2006; Wienert *et al.*, 2015; Weinhold *et al.*, 2014), with implications for common diseases through their association with regulatory elements (Maurano *et al.*, 2012). These variants have the potential to alter the transcriptional state of the cell (Deplancke *et al.*, 2016), leading to an increasing interest in developing tools to predict the impact of variants on TFBS (**Table 4.5**). TRAP (Thomas-Chollier *et al.*, 2011) and CATO (Maurano *et al.*, 2015) employ PWMs to predict variant impact on TFBS by comparing binding affinity scores between reference and alternative sequences. TRAP repeats this process across a collection of TFBS, reporting the motif with the most significant score change. CATO, on the other hand, generates a ranked list of disrupted motifs using a logistic model trained with information content differences, TF occupancy, and phylogenetic conservation between reference and alternative sequences. However, these methods face scalability issues when analyzing thousands of SNPs. atSNP (Zuo *et al.*, 2015) introduces a scalable strategy to assess the impact of thousands of SNPs on TFBS by computing statistical significance for the affinity score differences between reference and alternative sequences using PWMs. DeltaSVM (Lee *et al.*, 2015) and GkmExplain (Shrikumar *et al.*, 2019) use SVM-based motif models to evaluate variant impact. DeltaSVM scans DNA positions overlapping each SNP, computing the difference between reference and alternative sequence scores using a pretrained list of k-mers with associated weights. However, it assesses the impact of individual variants without accounting for relationships between variants. GkmExplain overcomes this limitation by considering the impact of variants not at individual positions but on sequence features, such as entire k-mers. DeepBind (Alipanahi *et al.*, 2015) and DeepSEA (Zhou and Troyanskaya, 2015) employ DNN-based models to predict variant impact on TFBS. DeepBind uses mutation maps to assess the effect of variants on binding affinities, considering the importance of each motif position within the model. DeepSEA uses in silico saturated mutagenesis to predict the impact of individual variants on the whole sequence context and features like TFBS. Similarly, Basset (Kelley *et al.*, 2016) uses *in silico* saturated mutagenesis by learning critical nucleotides governing chromatin accessibility, assigning importance scores to each position in the input sequences and attempting to map the variants' impact to the TFBS. Basenji (Kelley *et al.*, 2018) extends Basset's workflow by providing functional annotations to SNPs affecting sequence features and TFBS, and returning potential changes in gene expression patterns. However, Basenji is limited to predicting SNP effects on distal regulatory elements within a 20 kb range. Enformer (Avsec *et al.*, 2021b) overcomes this limitation by employing transformer architectures to extend the range up to 200 kb, providing a more comprehensive and accurate assessment of the functional effects of variants on sequence elements and gene expression.

Motif model	Software	Reference	Original input data type	Output	Year
PWM	TRAP	Thomas-Chollier *et al.* (2011)	ChIP-seq	Allele-specific score	2011
	CATO score	Maurano *et al.* (2015)	DHS sites	Ranked list of TFBS affected by SNPs	2015
	atSNP	Zuo *et al.* (2015)	Sequences overlapping input SNPs	Allele-specific score	2015
SVM-based	DeltaSVM	Lee *et al.* (2015)	DNase-seq	Allele-specific score	2015
	GkmExplain	Shrikumar *et al.* (2019)	DNase-seq	SNP impact on whole TFBS	2019
DNN-based	DeepBind	Alipanahi *et al.* (2015)	ChIP-seq HT-SELEX	Single SNP impact	2015
	DeepSEA	Zhou and Troyanskaya (2015)	ATAC-seq DNase-seq	Single SNP impact	2015
	Basset	Kelley *et al.* (2016)	ChIP-seq DNase-seq	Single SNP impact	2016
	Basenji	Kelley *et al.* (2018)	ChIP-seq DNase-seq	Single SNP impact	2018
	Enformer	Avsec *et al.* (2021b)	DNA sequences	SNP functional impact	2021

Table 4.5. Software to assess genetic variants impact on Transcription Factor Binding Sites The table provides a comprehensive overview of tools for predicting the impact of variants on TFBS. For each tool, the table presents details such as the utilized TFBS model (**Motif model**), the original data type used for testing in their original publication (**Original input data type**), the output type (**Output**), the year of publication (**Year**), and the corresponding reference (**Reference**).

4.2.6 Limitations on current Transcription Factor Binding Site Motif analysis

The exploration of TFBS motifs in DNA sequences has been an extensively studied domain over the past few decades. In this section we delved into diverse algorithms and computational models designed to discover and represent motifs. Moreover, we described the main downstream analysis using TFBS motifs. However, several unresolved issues and potential avenues for further research persist in this field. The debate around the selection between simple and complex motif discovery algorithms is still ongoing (Tognon *et al.*, 2023). Enumerative and alignment-based methods, despite their simplicity, have exhibited comparable performance in terms of scalability and accuracy when juxtaposed with complex methods (Weirauch *et al.*, 2013). Notably, these methods offer user-friendly interfaces and typically they do not require computational expertise. Additionally, they can be applied to any sequence dataset without necessitating additional information beyond the sequences themselves. Well-established tools like MEME (Bailey *et al.*, 1994; Bailey and Elkan, 1995; Bailey *et al.*, 2006), HOMER (Heinz *et al.*, 2010), and the newer STREME (Bailey, 2021) continue to be popular due to their user community support and continued maintenance. In contrast, probabilistic graphical model-based algorithms often struggle with scalability issues, particularly when analyzing thousands of sequences due to the intricate nature of model training involving dependencies. SVM-based methods have demonstrated scalability and accuracy in discovering TFBS motifs across different sequence dataset and predicting PBM binding affinities. One of the pivotal advantages of SVM-based algorithms lies in their ability to learn features across the entire sequence context. However, the performance of SVM-based methods is heavily influenced by the quality of the background dataset, which needs careful design based on various sequence characteristics, such as repeats and GC content. DNN-based motif discovery algorithms showcased high accuracy compared to

other methods, albeit with increased complexity requiring expertise and fine parameter tuning. While scalable in terms of dataset size, DNN-based methods often demand substantial computational resources and dedicated hardware components (e.g., GPUs) for effective model training. Despite these challenges, DNN-based methods are gaining popularity within the community. The debate revolving around the best method is still open. While several papers benchmarked motif discovery algorithms on diverse datasets, comprehensive benchmarks considering a wide range of datasets and methods from different algorithm classes are imperative for robust evaluations (Tognon *et al.*, 2023). The necessity and efficacy of motif models that capture dependencies between positions within a TFBS have been extensively discussed over the past few decades (Weirauch *et al.*, 2013; Bulyk *et al.*, 2002; Benos *et al.*, 2002; Siggers and Gordân, 2014). While probabilistic models are expected to outperform simpler models, numerous studies showed that simpler models like PWMs perform comparably well on both *in vitro* PBM and *in vivo* ChIP-seq data for the majority of transcription factors (Zhao and Stormo, 2011; Weirauch *et al.*, 2013). Weirauch *et al.* (2013) suggested that these results could be attributed to the observed degeneracy in eukaryotic TFBS. Many TFs exhibit binding to sequences with variations relative to the motif consensus, albeit with lower affinity. Since PWMs can accommodate variations in the motif consensus, they possess the capability to capture a broader spectrum of target sites, encompassing even those with weaker binding. However, this advantage comes at the cost of heightened susceptibility to noise, potentially leading to the recovery of several false positives. Probabilistic graphical models, by encoding dependencies between TFBS positions, are expected to offer more robust models. However, due to the multiple parameters learned by these models, there is a risk of overfitting the training data if not appropriately trained. SVM-based motif models have demonstrated generally superior performance compared to PWMs when predicting potential TFBS (Ghandi *et al.*, 2014a). Nonetheless, these models are often reduced to PWMs for visualization and interpretation, leading to a loss of the learned information. Recent studies have observed that DNN-based models better capture the sequence specificities underlying TF–DNA interactions, yielding improved predictions compared to other models (Trabelsi *et al.*, 2019). However, for the purpose of visualization and interpretation of the discovered motifs, DNN models are typically reduced to PWMs computed using the sequences activating the convolutional kernels. Consequently, while complex motif models provide powerful frameworks, they sacrifice interpretability, whereas simpler models are more susceptible to noise but remain interpretable. The trade-off between model accuracy and interpretability remains an open challenge in the context of motif analysis. The choice of TFBS model influences motif downstream analyses, in particular the evaluation of genetic variants impact. As outlined in **Section 4.2.5**, several methods using diverse models have been proposed to evaluate genetic variants impact on transcription factor binding sites. Mutations affecting TFBSs can persist within population-specific haplotypes (Kasowski *et al.*, 2010), potentially generating novel TFBS specific to a population of individuals. Similarly, genetic variation across cell types can generate cell type-specific binding sites. Current tools partially address these challenges, by evaluating the impact of individual genetic variants on TFs' binding landscape. This poses important limitation on the usage of these methods in precision medicine contexts. GRAFIMO (Tognon *et al.*, 2021) addresses these challenges by introducing a variant- and haplotype-aware motif scanning tool to search potential occurrences of known TF motifs on genome graphs (Paten *et al.*, 2017). By searching motifs (given as PWMs) on genome graphs rather than on linear reference genomes, GRAFIMO finds potential binding site occurrences accounting for the effect of genetic variants from populations of even thousands of individuals. On the other hand, tools predicting genetic variants impact often ignore other fundamental information, such as TFs' expression level or chromatin accessibility. MotifRaptor (Yao *et al.*, 2021) interpolates different omics data, such as cell type-specific transcriptomic data and chromatin accessibility, while predicting genetic variants impact on TFBSs. MotifRaptor is designed to annotate non-coding variants predicting their potential functional impact. The next two chapters describe these novel approaches.

GRAFIMO: Variant- and Haplotype-aware Transcription Factor Binding sites identification on Genome Graphs

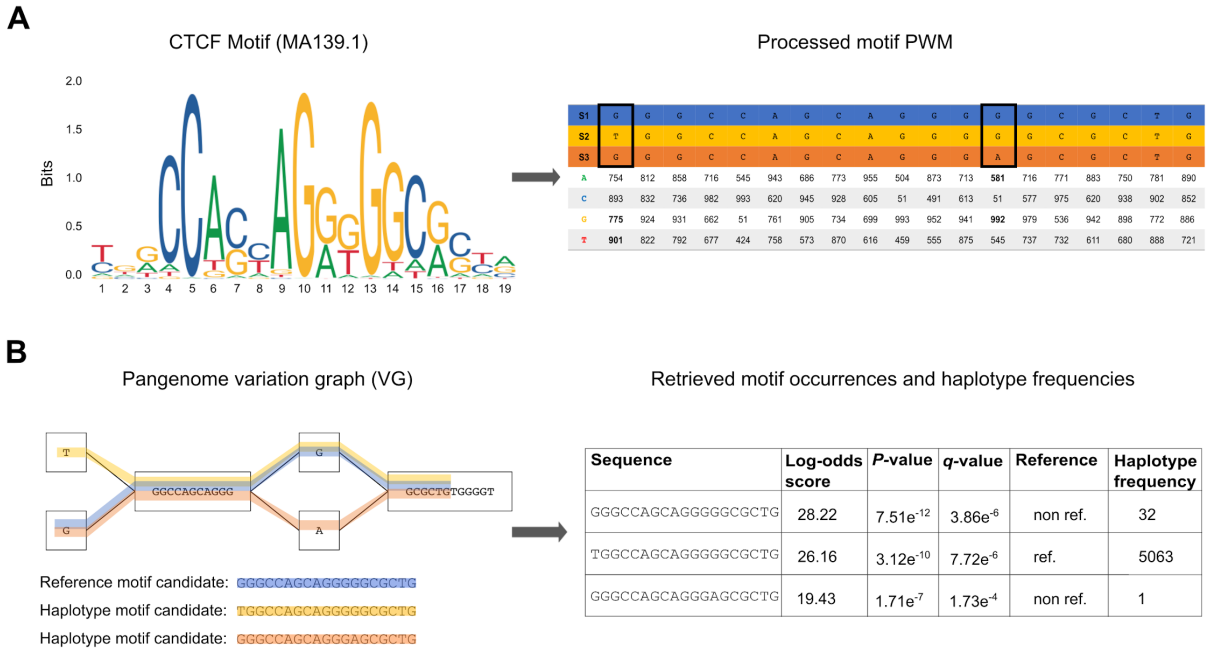
Over the past decade, significant strides have been made in the development of methods for searching known TFBS on linear reference genomes (Tognon *et al.*, 2023; Boeva, 2016). Notable examples include FIMO (Grant *et al.*, 2011) and MOODS (Korhonen *et al.*, 2009). They scan a set of genomic sequences searching for potential occurrences of known TFBS represented as PWMs. Other tools, such as is-rSNP, TRAP, and atSNP (Macintyre *et al.*, 2010; Thomas-Chollier *et al.*, 2011; Zuo *et al.*, 2015), were introduced to accommodate SNPs and short indels in the sequences to scan. However, these tools fall short in considering individual haplotypes and providing summaries of the frequency of these events in a population, limiting their applicability to precision medicine contexts. To address these challenges, we developed GRAFIMO (GRAph-based Finding of Infividual Motif Occurrences) (Tognon *et al.*, 2021), a novel tool performing variant- and haplotype aware identification of known TFBS in genome graphs. By leveraging genome graphs, GRAFIMO accounts for population-wide and individual-specific genetic variation while searching TFBSs. We demonstrate the utility of GRAFIMO by searching TFBS on a genome graph encoding the haplotypes from all individuals sequenced by the 1000 Genomes Project (1KGP) (Siva, 2008; Zheng-Bradley *et al.*, 2017). This innovative approach provides a more nuanced understanding of TF-DNA interaction dynamics in the context of population-wide genetic diversity. GRAFIMO overcomes the limitations of competitor motif scanning tools, providing a computational framework to predict the impact of genetic variation on TFs' binding landscape in precision medicine contexts.

5.1 Design and implementation

GRAFIMO is a command-line tool enabling an efficient variant- and haplotype-aware search of known TFBS, within a population of individuals encoded in a genome graph. GRAFIMO provides two main functionalities: genome graph construction from user data, and the search of one or more TFBS motifs on input precomputed graphs. Briefly, given a TF motif PWM and a set of genomic regions, GRAFIMO leverages genome graphs to efficiently scan and report all the TFBS candidates and their frequencies in a single pass. Along with the motif candidates, GRAFIMO reports the predicted changes in binding affinity mediated by the genetic variants embedded in the graph. GRAFIMO is written in Python3 and Cython and it has been designed to interface with *vg* toolkit to handle genome graphs.

5.1.1 Genome variation graph construction

GRAFIMO offers an intuitive command-line interface for constructing genome graphs from user data, when needed. Given a reference genome (FASTA format) and a set of genomic variants (VCF format), GRAFIMO seamlessly interfaces with the *vg* toolkit to build the genome graphs (`VG` format) and its indexes (`XG` (Garrison *et al.*, 2018) and `GBWT` (Sirén *et al.*, 2020; Novak *et al.*, 2017) formats). The two indexes are required to perform efficient graph traversal and track the haplotypes embedded in the data structure. To minimize the footprint of the genome graph files, GRAFIMO constructs a graph per chromosome. Moreover, this graph construction design speeds-up the TFBS motif search by allowing to



Sequence	Log-odds score	*P*-value	*q*-value	Reference	Haplotype frequency
GGGCCAGCAGGGGGCGCTG	28.22	$7.51e^{-12}$	$3.86e^{-6}$	non ref.	32
TGGCCAGCAGGGGGCGCTG	26.16	$3.12e^{-10}$	$7.72e^{-6}$	ref.	5063
GGGCCAGCAGGGAGCGCTG	19.43	$1.71e^{-7}$	$1.73e^{-4}$	non ref.	1

Figure 5.7. GRAFIMO TF motif search workflow. (A) The motif PWM (in MEME or JASPAR format) is processed and its values are scaled in the range [0, 1000]. The resulting score matrix is used to assign a score and a corresponding P-value to each motif occurrence candidate. In the final report GRAFIMO returns the corresponding log-odds scores, which are retrieved from the scaled values. **(B)** GRAFIMO slides a window of length k, where k is the motif width, along the haplotypes (paths in the graph) of the genomes used to build the VG. The resulting sequences are scored using the motif scoring matrix and are statistically tested assigning them the corresponding P-value and q-value. Moreover, for each entry is assigned a flag value stating if it belongs to the reference genome sequence ("ref") or contains genomic variants ("non.ref") and is computed the number of haplotypes in which the sequence appears.

scan different chromosomes in parallel. Alternatively, for systems with limited RAM, the search can be conducted one chromosome at a time, providing flexibility in resource usage.

5.1.2 Transcription factor binding site motif search

The TFBS motif search in GRAFIMO requires a set of genomes embedded in a genome graph (`XG` format), a database of known TF motif PWMs (MEME or JASPAR formats), and a set of genomic regions (BED format). GRAFIMO reports all the potential TFBS motif occurrences found in the input regions and their associated statistical significance (**Fig.5.7**). To identify potential TFBS, GRAFIMO employs a sliding window approach with a window length of k, where k is the query motif width. The window traverses the paths in the graphs that correspond to the genome sequences encoded in it (**Fig.5.7(B)**). This process is accomplished by an extension to the `vg find` functionality, using the graph's `GBWT` index to explore the k-mer space, while considering the embedded haplotypes (Sirén *et al.*, 2020). By default, GRAFIMO focuses its search on paths corresponding to observed haplotypes, but it also provides an option to consider all possible recombinants, even if absent in any embedded sample. The significance (log-likelihood) of each TFBS candidate is determined by evaluating the nucleotide preferences encoded in the motif PWM, akin to FIMO (Grant *et al.*, 2011). Specifically, the PWM undergoes processing into a Position Specific Scoring Matrix (PSSM) (**Fig.5.7(A)**). The resulting PSSM log-likelihood values are scaled within the range $[0, 1000]$. The scaling allows efficient statistical significance (P-value) computations via dynamic programming (Grant *et al.*, 2011). Subsequently P-values undergo conversion to q-values through the Benjamini-Hochberg procedure to correct for multiple hypothesis testing. In this procedure, all P-values correspond to all the k-mer-paths extracted within the scanned graph's regions. Additionally, GRAFIMO provides insights into the number of haplotypes in which a significant motif is observed, along with its presence in the reference genome and/or alternative genomes (**Fig.5.7(B)**). The subsequent sections provide an in-depth exploration of the algorithms embedded within GRAFIMO.

Processing of Position Weight Matrix to Position Specific Scoring Matrix

GRAFIMO utilizes Position Specific Scoring Matrices (PSSMs) to assign scores to potential motif occurrences. These PSSMs are generated through the processing of input PWMs provided in JASPAR (**Appendix A.8**) or MEME (**Appendix A.7**) format. Let us assume the input motif PWM M is provided in JASPAR format and $|M| = m$, then GRAFIMO starts by deriving the corresponding position probability matrix M_p:

$$M_p = \frac{M[n,i]}{\sum_{n\in\Sigma} M[n,i]}$$

where $\Sigma = \{A, C, G, T\}$ and $1 \leq i \leq m$. On the other hand, if the input binding site is provided in MEME format, GRAFIMO skips this step since the motif is already represented as a probability matrix. After computing M_p, GRAFIMO normalizes each probability returning a normalized position probability matrix M'_p:

$$M'_p = \frac{M_p[n,i]}{\sum_{n\in\Sigma} M_p[n,i]}$$

where $\Sigma = \{A, C, G, T\}$ and $1 \leq i \leq m$. To account for the frequency of emission of each symbol in the alphabet $\Sigma = \{A, C, G, T\}$ (genomic frequency of each nucleotide), users can provide GRAFIMO with a 0-th order background model B. In the absence of a user-defined background distribution, GRAFIMO assumes a uniform distribution. To prevent divisions by zero in subsequent processing steps and accommodate the potential lack of information, each normalized probability in M'_p is incremented by a pseudocount. Typically, pseudocounts are small values, ranging between 0.1 and $1e^{-4}$. Next, each value in M'_p is adjusted by adding the pseudocount and weighting the resulting values with the background model probability distribution:

$$M''_p = \frac{M'_p[n,i] \cdot \sum_{n\in\Sigma} M[n,i] + p \cdot B_n}{\sum_{n\in\Sigma} M[n,i] + p}$$

where $\Sigma = \{A, C, G, T\}$ and $1 \leq i \leq m$. Following these steps, the probabilities are transformed into log-likelihood scores, often referred to as *log-odds*. Each value in M''_p, undergoes this transformation to obtain the final PSSM:

$$\text{PSSM}[n,i] = \log_2\left(\frac{M''_p[n,i]}{B_n}\right)$$

The resulting PSSM is then utilized to calculate a log-likelihood or binding affinity score for each candidate motif occurrence.

Evaluating motif statistical significance via dynamic programming

To evaluate the significance of the score assigned to a candidate binding site occurrence, is essential to consider the likelihood of such a log-odds value occurring by chance. This probability can be estimated by establishing a null hypothesis model, representing the random occurrence of m nucleotides matching the motif. Recall that $|M| = m$, where M is the input motif. A naïve approach to derive a null model involves randomly shuffling nucleotides $n \in \Sigma$, obtaining random sequences s, where $|s| = m$, and score each s using the PSSM. The resulting scores under this null model would be purely chance-driven. The probability that a score, obtained under the null hypothesis, is at least as large as the one observed in the real model is defined P-value in motif scanning context (Noble, 2009). Essentially, the P-value for an observed score represents the area under the score score(s) distribution obtained after the random shuffle of nucleotides, where $x \geq \text{score}(s)$. Although shuffling nucleotides to obtain would result in accurate P-values, this procedure is not computationally feasible. To tackle this challenge efficiently, GRAFIMO implements a dynamic programming (DP) algorithm (Griffin *et al.*, 1994). This algorithm assumes that the scored sequence is randomly generated with a specified frequency associated to each nucleotide. To facilitate the implementation of the DP-algorithm, GRAFIMO scales the values stored in the log-odds matrix PSSM, ensuring they fall within the range $0 \leq \text{PSSM}(n,i) \leq 1000$:

$$\text{PSSM}_{\text{scaled}}[n,i] = \left\lfloor \left(\text{PSSM}[n,i] - \lfloor \min \text{PSSM} \rfloor \cdot \left\lfloor \frac{1000}{\max \text{PSSM} - \min \text{PSSM}} \right\rfloor \right) \right\rceil$$

where $\Sigma = \{A, C, G, T\}$ and $1 \leq i \leq m$. The result is a new matrix $\text{PSSM}_{\text{scaled}}$ with integer values ranging from 0 to 1000. Using integers instead of floatong-point values, enables fast computations. Subsequently,

a null model distribution is computed for each sequence s_r randomly obtained, with $1 \leq |s_r| \leq m$. This is done following the probabilities defined in the background distribution B. A $m \times (1000 \cdot m + 1)$ matrix P is built, with entries filled using **Algorithm 1**. The rows of P represent sequences length, while columns represent a score that can be obtained for a motif occurrence using $\text{PSSM}_{\text{scaled}}$. At the end of the procedure, each entry $P[i, j]$ contains the number of sequences $|S_r|$, where $|s_r| = i$ and $\text{score}(s_r) = j$, for each $s_r \in S_r$ with a score j. Sequences s_r, represent randomly generated strings following the background distribution. The time complexity of this algorithm is $O(n \cdot m)$, where n is the number of rows and m is the number of columns of P. The space needed by the algorithm is $O(n)$, as once the current row is completely computed, the preceding one is removed and is no longer needed in subsequent steps of the algorithm.

Algorithm 1: Dynamic programming algorithm computing null model in GRAFIMO (Griffin *et al.*, 1994)

Data: $\text{PSSM}_{\text{scaled}}$, M, B, Σ
Result: P

```
m ← |M| ;                                          /* M is the input motif */
c ← 1000 · m + 1;
for i in 1 to m do
    for n ∈ Σ do
        if i = 0 then
            P[0, M[n, i]] ← P[0, M[n, i]] + 1 · B_n;
        else
            for j in 1 to c do
                if P[i − 1, j] ≠ 0 then
                    P[i, M[n, i] + j] ← P[i, M[n, i] + j] + P[i − 1, j] · B_n;
return P
```

Searching motif occurrences on genome graphs

After computing the PSSM and its corresponding P-value matrix P from the input motif M, GRAFIMO begins a scanning process across the input genome graph within the input genomic regions (**Algorithm 2**). The traversal of graph nodes is achieved performing a Breadth-First Search (BFS)-like algorithm operating in parallel. The scanning progresses by employing a sliding window of width m, where $|M| = m$. While scanning the graph, GRAFIMO recover all consecutive overlapping segments of length m. When a node with two or more outgoing edges is encountered, indicating the presence of one or more alternative alleles at that position, GRAFIMO visits only the haplotypes belonging to the allowed paths in the graph. Subsequently, each sequence extracted from the genome graph is scored using the PSSM (**Algorithm 3**). The score for each sequence s is the sum of the values of entries $\text{PSSM}_{\text{scaled}}[n, i]$, where $n = s_i$ and $1 \leq i \leq m$. To speed-up the score computation, GRAFIMO utilizes the PSSM with scaled values instead of the one with log-likelihood scores, as operations on integers are faster than those on floating-point values. The resulting score for each sequence $\text{score}(s)$ is then transformed back into a log-likelihood score $LL(s)$:

$$LL(s) = \left(\frac{\text{score}(s)}{1000/\max \text{PSSM} - \min \text{PSSM}} \right) + m \cdot \lfloor \min \text{PSSM} \rfloor$$

The complexity of the scoring procedure is $O(n \cdot m)$, where n is the total number of sequences to score, and m is their length. In scenarios where numerous scores are reported, such as scanning a genome for TFBS motif occurrences, relying solely on a P-value is insufficient to assess statistical significance. In fact, P-values are statistically valid only when computing a single score (Noble, 2009). To address this limitation, GRAFIMO corrects P-values for multiple hypothesis tests using false discovery rate estimation (FDR), providing a corresponding q-value. FDRs are computed directly from P-values using the Benjamini-Hochberg procedure.

Algorithm 2: GRAFIMO motif scanning algorithm

Data: G, M
Result: scores, pvalues
PSSM $\leftarrow$ computePSSM(M) ; `/* see previouse sections for details */`
$P \leftarrow$ computePvalueMatrix(PSSM) ; `/* see Algorithm 1 for details */`
$S \leftarrow$ recoverSequencesBFS(G) ; `/* each` $s \in S$ `satisfies` $|s| = |M|$ `*/`
scores $\leftarrow \emptyset$;
pvalues $\leftarrow \emptyset$;
for $s \in S$ **do**
    score(s), pvalue(s) $\leftarrow$ computeScorePvalue(s, PSSM, P) ; `/* details in Algorithm 3 */`
    scores[i] $\leftarrow$ score(s);
    pvalues[i] $\leftarrow$ pvalue(s);
return *scores, pvalues*

Algorithm 3: Compute sequence score and P-value

Data: s, PSSM, P
Result: score, pvalue
$\Sigma \leftarrow \{A, C, G, T\}$;
score $\leftarrow 0$;
for i ***in*** 1 ***to*** $|s|$ **do**
    $n \leftarrow s(i)$;
    if $n \notin \Sigma$ **then**
        score $\leftarrow 0$; `/* The sequence contains N, assign the lowest possible score */`
        break;
    score $\leftarrow$ score + PSSM(n, i);
pvalue $\leftarrow \sum_{j=\text{score}}^{|P|} \frac{P(j)}{\sum_{k=1}^{|P|} P(k)}$;
return *score, pvalue*

5.1.3 Report generation

GRAFIMO interface has been designed with inspiration drawn from FIMO, allowing seamless integration into pipelines and workflows originally built on FIMO. Similar to FIMO, GRAFIMO generates three distinct reports: a tab-delimited file (TSV), an HTML report, and a GFF3 file compatible with the UCSC Genome Browser (Lee *et al.*, 2020). The TSV report (**Fig.5.8**) lists each found candidate TFBS along with its score, genomic location (start, stop, and strand), and statistical significance (P-value and q-value). Moreover, it reports the number of haplotypes in which each motif candidate has been observed, with a flag value highlighting whether the occurrence was found only on alternative haplotypes or also in the reference sequence. GRAFIMO also provides an HTML version of the TSV report (**Fig.5.9**), that can be easily explored with any web browser. The GFF3 report is compatible with the UCSC Genome Browser, allowing users to load it as a custom track (**Fig.5.10**). This enables the visualization and exploration of the identified TFBS candidates alongside additional annotations available in the Genome Browser, such as nearby genes, enhancers, promoters, or pathogenic variants sourced from the ClinVar database (Landrum *et al.*, 2020).

5.2 Searching motif occurrences with GRAFIMO

GRAFIMO's main aim is to investigate the potential impact of genetic variants on the binding affinity of putative TFBS across a set of individuals. By leveraging genome graphs, GRAFIMO may recover additional sites that might be missed when considering linear reference genomes exclusively, without accounting for genetic variants. To showcase its utility, we constructed a genome graph based on 2,548 individuals from 1KGP phase 3 (hg38 human genome assembly) (Zheng-Bradley *et al.*, 2017; Lowy-Gallego *et al.*, 2019). The resulting graph encoded their genetic variants (∼78 millions, SNPs and indels (**Table 5.6**)) and phased haplotypes (total of 5,096 haplotypes). Subsequently, we searched the graph for putative TFBS of three TF motifs retrieved from the JASPAR database (Sandelin *et al.*, 2004; Fornes *et al.*, 2020):

	motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence	haplotype_frequency	reference
1	MA0139.1	CTCF	chr2:231612642-231612803	231612720	231612739	+	29.114754098360663	5.988799754224379e-13	6.040843464356013e-07	TGGCCACCAGGGGGCGCCG	5096	ref
2	MA0139.1	CTCF	chr3:98023327-98023504	98023430	98023411	-	29.04918032786884	8.440969121103595e-13	6.040843464356013e-07	CGGCCACCAGGGGGCGCCA	5096	ref
3	MA0139.1	CTCF	chr7:98021604-98021781	98021693	98021712	+	28.983606557377072	1.0141213123241087e-12	6.040843464356013e-07	CGGCCACCAGGGGGCGCCG	5096	ref
4	MA0139.1	CTCF	chr12:121499983-121500138	121500062	121500081	+	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5096	ref
5	MA0139.1	CTCF	chr16:67993308-67993508	67993402	67993421	+	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5094	ref
6	MA0139.1	CTCF	chr12:108618200-108618398	108618301	108618282	-	28.42622950819674	4.0897345623855974e-12	1.218071546367531e-06	CGGCCACCAGGGGGCGCCC	5096	ref
7	MA0139.1	CTCF	chr8:142782577-142782754	142782661	142782680	+	28.327868852459005	5.273880274923456e-12	1.3463598544475949e-06	TGGCCACCAGGGGGCGCTC	2835	non.ref
8	MA0139.1	CTCF	chr1:156090811-156090991	156090877	156090896	+	28.032786885245912	1.1988060111075724e-11	2.3803199115082577e-06	TGGCCACCAGGTGGCGCCG	5096	ref
9	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	28.032786885245912	1.1988060111075724e-11	2.3803199115082577e-06	TGGCCACCAGGTGGCGCCG	5088	ref
10	MA0139.1	CTCF	chr19:35067213-35067372	35067283	35067302	+	27.91803278688525	1.594209932358396e-11	2.559578506660857e-06	TGGCCACCAGGGGGCACTG	5096	ref
11	MA0139.1	CTCF	chr15:34210116-34210292	34210206	34210187	-	27.918032786885525	1.594209932358396e-11	2.559578506660857e-06	TGGCCACCAGGGGGCACTG	787	non.ref
12	MA0139.1	CTCF	chr9:130461251-130461422	130461337	130461356	+	27.885245901639337	1.7187819081805712e-11	2.559578506660857e-06	TGGCCACCAGAGGGCGCCA	5096	ref
13	MA0139.1	CTCF	chr1:7647328-7647487	7647424	7647405	-	27.803278688524586	2.1113539694844316e-11	2.6950196770286664e-06	TGGCCACCAGGTGGCGCTG	5096	ref
14	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	27.803278688524586	2.1113539694844316e-11	2.695019677028664e-06	TGGCCACCAGGTGGCGCTG	8	non.ref
15	MA0139.1	CTCF	chr3:120558991-120559170	120559086	120559067	-	27.672131147540995	2.6237024511211056e-11	3.0606872796384374e-06	TGGCCAGCAGGGGGCGCTC	5052	ref
16	MA0139.1	CTCF	chr19:4809674-4809833	4809756	4809775	+	27.590163934426243	2.9455541904678063e-11	3.0606872796384374e-06	TGGCCACCAGAGGGCGCTG	5096	ref
17	MA0139.1	CTCF	chr20:44685269-44685461	44685354	44685373	+	27.57377049180326	3.0829219981831114e-11	3.0606872796384374e-06	TGGCCAGCAGGGGGCGGTG	5091	ref
18	MA0139.1	CTCF	chr19:47863504-47863679	47863606	47863587	-	27.57377049180326	3.0829219981831114e-11	3.0606872796384374e-06	TGGCCACTAGGGGGCGGCA	5095	ref
19	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.491803278688508	3.5727124816933935e-11	3.3144635195783872e-06	TGGCCACCAGGGGGCAGCG	2	non.ref
20	MA0139.1	CTCF	chr5:177501044-177501224	177501155	177501136	-	27.459016393442596	3.7094909167992564e-11	3.3144635195783872e-06	CGGCCACCAGAGGGCGCTG	5096	ref
21	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.360655737704917	4.440796897956725e-11	3.7789447576156337e-06	CGGCCACCAGGGGGCAGCG	5086	ref
22	MA0139.1	CTCF	chr20:49979066-49979252	49979155	49979174	+	27.229508196721326	5.8822012428205414e-11	4.5090853829073590e-06	TGGCCAGCAGAGGGCGCCA	4789	ref
23	MA0139.1	CTCF	chr12:111397009-111397212	111397090	111397109	+	27.213114754098342	6.217717166529091e-11	4.5090853829073590e-06	CAGCCACCAGGGGGCGCCA	5096	ref
24	MA0139.1	CTCF	chr20:63973975-63974156	63974071	63974052	-	27.13114754098359	7.240291664795245e-11	4.5090853829073590e-06	CGGCCACCAGGGGGCAGTG	5094	ref
25	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.13114754098359	7.240291664795245e-11	4.5090853829073590e-06	CGGCCACCAGGGGGCAGTG	2	non.ref

Figure 5.8. GRAFIMO TSV summary report. The tab-delimited report shows the first 25 potential CTCF occurrences retrieved by GRAFIMO, searching CTCF motif in ChIP-seq peak regions on A549 cell line (ENCODE experiment ENCFF816XLT).

	motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence	haplotype_frequency	reference
1	MA0139.1	CTCF	chr2:231612642-231612803	231612720	231612739	+	29.114754	5.988800e-13	6.040843e-07	TGGCCACCAGGGGGCGCCG	5096	ref
2	MA0139.1	CTCF	chr3:98023327-98023504	98023430	98023411	-	29.049180	8.440969e-13	6.040843e-07	CGGCCACCAGGGGGCGCCA	5096	ref
3	MA0139.1	CTCF	chr7:98021604-98021781	98021693	98021712	+	28.983607	1.014121e-12	6.040843e-07	CGGCCACCAGGGGGCGCCG	5096	ref
4	MA0139.1	CTCF	chr12:121499983-121500138	121500062	121500081	+	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5096	ref
5	MA0139.1	CTCF	chr16:67993308-67993508	67993402	67993421	+	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5094	ref
6	MA0139.1	CTCF	chr12:108618200-108618398	108618301	108618282	-	28.426230	4.089735e-12	1.218072e-06	CGGCCACCAGGGGGCGCCC	5096	ref
7	MA0139.1	CTCF	chr8:142782577-142782754	142782661	142782680	+	28.327869	5.273880e-12	1.346360e-06	TGGCCACCAGGGGGCGCTC	2835	non.ref
8	MA0139.1	CTCF	chr1:156090811-156090991	156090877	156090896	+	28.032787	1.198806e-11	2.380320e-06	TGGCCACCAGGTGGCGCCG	5096	ref
9	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	28.032787	1.198806e-11	2.380320e-06	TGGCCACCAGGTGGCGCCG	5088	ref
10	MA0139.1	CTCF	chr19:35067213-35067372	35067283	35067302	+	27.918033	1.594210e-11	2.559579e-06	TGGCCACCAGGGGGCACTG	5096	ref
11	MA0139.1	CTCF	chr15:34210116-34210292	34210206	34210187	-	27.918033	1.594210e-11	2.559579e-06	TGGCCACCAGGGGGCACTG	787	non.ref
12	MA0139.1	CTCF	chr9:130461251-130461422	130461337	130461356	+	27.885246	1.718782e-11	2.559579e-06	TGGCCACCAGAGGGCGCCA	5096	ref
13	MA0139.1	CTCF	chr1:7647328-7647487	7647424	7647405	-	27.803279	2.111354e-11	2.695020e-06	TGGCCACCAGGTGGCGCTG	5096	ref
14	MA0139.1	CTCF	chr16:66608629-66608805	66608709	66608728	+	27.803279	2.111354e-11	2.695020e-06	TGGCCACCAGGTGGCGCTG	8	non.ref
15	MA0139.1	CTCF	chr3:120558991-120559170	120559086	120559067	-	27.672131	2.623702e-11	3.060687e-06	TGGCCAGCAGGGGGCGCTC	5052	ref
16	MA0139.1	CTCF	chr19:4809674-4809833	4809756	4809775	+	27.590164	2.945542e-11	3.060687e-06	TGGCCACCAGAGGGCGCTG	5096	ref
17	MA0139.1	CTCF	chr20:44685269-44685461	44685354	44685373	+	27.573770	3.082922e-11	3.060687e-06	TGGCCAGCAGGGGGCGGTG	5091	ref
18	MA0139.1	CTCF	chr19:47863504-47863679	47863606	47863587	-	27.573770	3.082922e-11	3.060687e-06	TGGCCACTAGGGGGCGGCA	5095	ref
19	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.491803	3.572712e-11	3.314464e-06	TGGCCACCAGGGGGCAGCG	2	non.ref
20	MA0139.1	CTCF	chr5:177501044-177501224	177501155	177501136	-	27.459016	3.709491e-11	3.314464e-06	CGGCCACCAGAGGGCGCTG	5096	ref
21	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.360656	4.440797e-11	3.778945e-06	CGGCCACCAGGGGGCAGCG	5086	ref
22	MA0139.1	CTCF	chr20:49979066-49979252	49979155	49979174	+	27.229508	5.882201e-11	4.509085e-06	TGGCCAGCAGAGGGCGCCA	4789	ref
23	MA0139.1	CTCF	chr12:111397009-111397212	111397090	111397109	+	27.213115	6.217717e-11	4.509085e-06	CAGCCACCAGGGGGCGCCA	5096	ref
24	MA0139.1	CTCF	chr20:63973975-63974156	63974071	63974052	-	27.131148	7.240292e-11	4.509085e-06	CGGCCACCAGGGGGCAGTG	5094	ref
25	MA0139.1	CTCF	chr10:123994704-123994883	123994783	123994802	+	27.131148	7.240292e-11	4.509085e-06	CGGCCACCAGGGGGCAGTG	2	non.ref

Figure 5.9. GRAFIMO HTML summary report. The HTML report displays the first 25 potential CTCF occurrences retrieved by searching CTCF motif occurrences with GRAFIMO on ChIP-seq regions on A549 cell line (ENCODE experiment ENCFF816XLT).

CTCF (JASPAR ID MA0139.1), ATF3 (JASPAR ID MA0605.2), and GATA1 (JASPAR ID MA0035.4) (**Fig.5.11**). The three motifs exhibit diversity in length (from 11 to 19 bp), information content, and evolutionary conservation. To investigate regions with likely true binding events, we focused our analysis on ChIP-seq peak regions in six different cell lines (A549, GM12878, H1, HepG2, K562, MCF-7) retrieved from the ENCODE Project database (Consortium *et al.*, 2012; Davis *et al.*, 2018) (**Table 5.7**). For each TF, we systematically acquired the optimal IDR thresholded peaks from ENCODE (bigBED format). Subsequently, we employed UCSC's `bigBedToBed` tool (Kent *et al.*, 2010) to convert each bigBED file into its corresponding BED. To enhance data quality, we filtered the resulting BED files, excluding features mapped to non-canonical chromosomes. The filtered BEDs were then subjected to sorting based on both q-values and peak signals. This sorting allowed us to prioritize the identification of the most informative regions, specifically selecting the top 3,000 peaks for each experiment. Then, we performed a comprehensive scan on the prioritized regions using GRAFIMO. In our subsequent downstream analyses, we retained those sites that exhibited a P-value $< 1e^{-4}$. This stringent threshold ensured a focus on statistically significant potential motif occurrences. Finally, we considered the sites meeting these criteria as potential binding sites for the respective TFs under investigation. Based on the retrieved sites, we

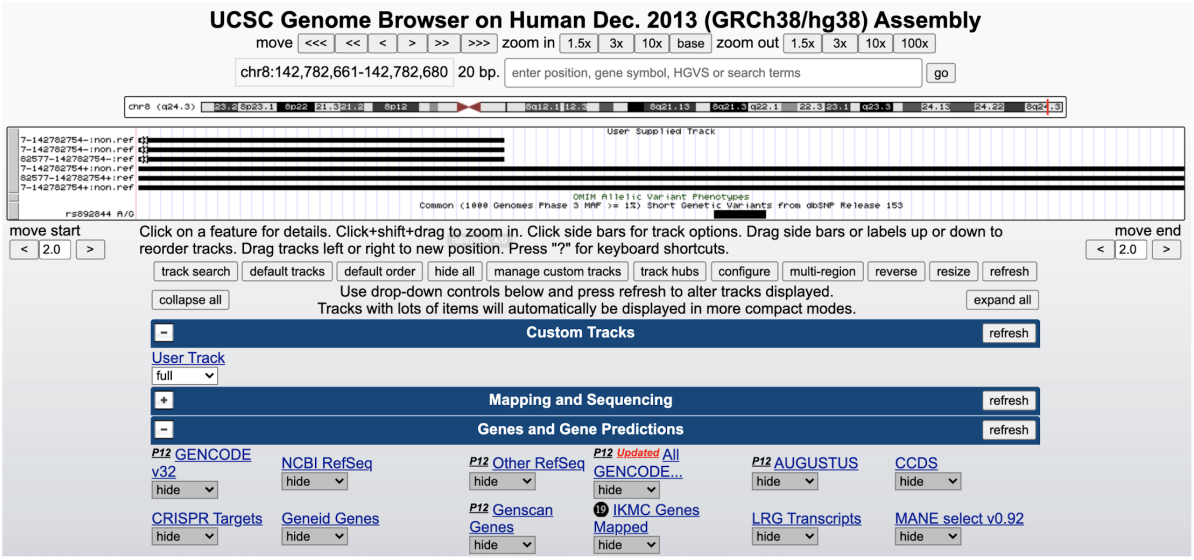


Figure 5.10. GFF3 track returned by GRAFIMO and loaded on the UCSC Genome Browser. GRAFIMO produces a GFF3 report which can be loaded on the UCSC Genome Browser. The custom track loaded in the example shows three potential CTCF occurrences (region chr8:142,782,661-142,782,680) recovered by GRAFIMO, which overlap a dbSNP annotated variant (rs892844).

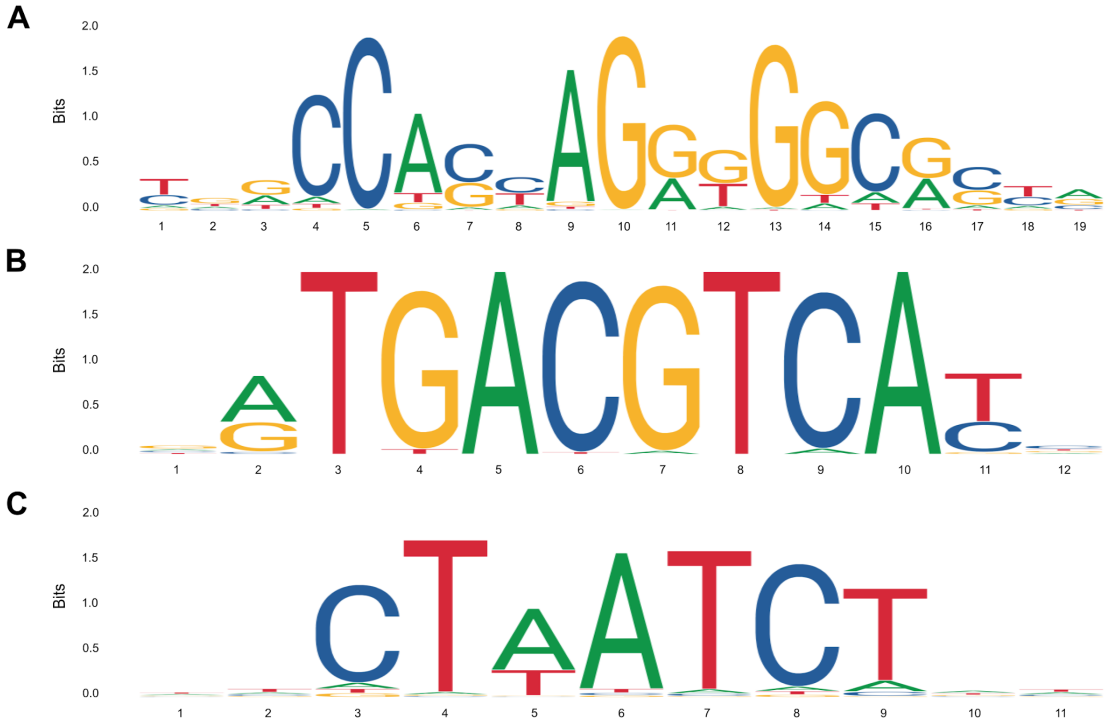


Figure 5.11. Transcription factor motifs used to test GRAFIMO. Transcription factor binding site motifs of **(A)** CTCF, **(B)** ATF3, and **(C)** GATA1.

consistently observed across the three investigated TFs that genetic variants can significantly affect the estimated binding affinity.

5.2.1 Searching for CTCF occurrences

CTCF is a zinc-finger transcription factor involved in transcriptional regulation, playing a pivotal role in epigenetic control (Ishihara *et al.*, 2006), and functioning as a tumor suppressor (Fiorentino and Giordano, 2012). In our experiments, we performed a targeted search for CTCF motif (JASPAR ID MA0139.1) (**Fig.5.11 (A)**) occurrences within the 1KGP genome graph. Interestingly, we found a substantial number of CTCF motif occurrences exclusive to non-reference haplotypes, indicating that a significant pool of

Chromosome	Number of genetic variants
Chr1	6,191,833
Chr2	6,790,551
Chr3	5,641,493
Chr4	5,477,810
Chr5	5,115,036
Chr6	4,863,337
Chr7	4,511,408
Chr8	4,425,449
Chr9	3,384,360
Chr10	3,874,259
Chr11	3,881,791
Chr12	3,745,465
Chr13	2,760,845
Chr14	2,548,903
Chr15	2,301,453
Chr16	2,548,920
Chr17	2,209,149
Chr18	2,189,529
Chr19	1,738,824
Chr20	1,817,492
Chr21	1,045,269
Chr22	1,059,079
ChrX	106,963

Table 5.6. Genetic variants in the 1KGP genome graphs. Number of genetic variants (SNPs and indels) in the genome graph constructed using 1KGP phase 3 on hg38 data. The variants belong to 2,548 individuals from 26 populations. In total the genome graphs encoded ∼78 millions variants

.

Motif	A549	GM12878	H1	HepG2	K562	MCF-7
CTCF	ENCFF816XLT	ENCFF267NYF		ENCFF015OJG	ENCFF895HAG	ENCFF088JWU
ATF3			ENCFF207AVV	ENCFF753WNT	ENCFF787GVU	
GATA1					ENCFF811YFQ ENCFF939ODZ	

Table 5.7. ENCODE ChIP-seq experiments. To test our software we searched potential occurrences of three transcription factor motifs (CTCF, ATF3, and GATA1) in a hg38 genome graph enriched with genetic variants and haplotypes of 2,548 individuals from 1000 Genomes Project phase 3.

potential TFBS is overlooked when scanning the genome without considering genetic variants (**Fig.5.12 (A)**). Furthermore, our analysis identified several highly significant CTCF occurrences within rare haplotypes (**Fig.5.12 (B)**), that may influence gene expression in individuals showing these haplotypes. We investigated the genomic locations of significant motif occurrences to assess how individual binding sites might be affected by genetic diversity–whether disrupted, created, or modulated. Interestingly, our experiments revealed that 6.13% of potential CTCF binding sites were exclusive to non-reference haplotypes, 5.94% were disrupted by variants in non-reference haplotypes, and approximately 30% retained significance in non-reference haplotypes, but with different binding scores (**Fig.5.12 (C)**). Notably, a considerable portion of putative binding sites recovered solely on individual haplotypes exhibited population specificity. For instance, 24.66%, 6.74%, 5.68%, 13.01%, and 12.52% of potential CTCF binding sites retrieved exclusively from individual haplotypes were specific to AFR, EUR, AMR, SAS, and EAS populations, respectively (**Fig.5.12 (D)**). Among the unique motif occurrences identified exclusively in non-reference haplotypes in CTCF ChIP-seq peaks, we uncovered a TFBS (chr19:506,910-506,929) that illustrates the pitfalls of relying solely on reference genomes for motif scanning. In this genomic locus, we identified a heterozygous SNP aligning with position 10 of the CTCF matrix, that significantly modulates the binding affinity of the corresponding binding site. By inspecting ChIP-seq reads (experiment ENCSR000DZN on GM12878), an allelic imbalance emerged towards the alternative allele. The alternative allele G exhibited a clear prevalence (70.59% of reads), while the reference allele A lagged behind at 29.41% of reads. This allelic imbalance is not observed in the control reads (experiment ENCSR000EYX) (**Fig.5.13**).

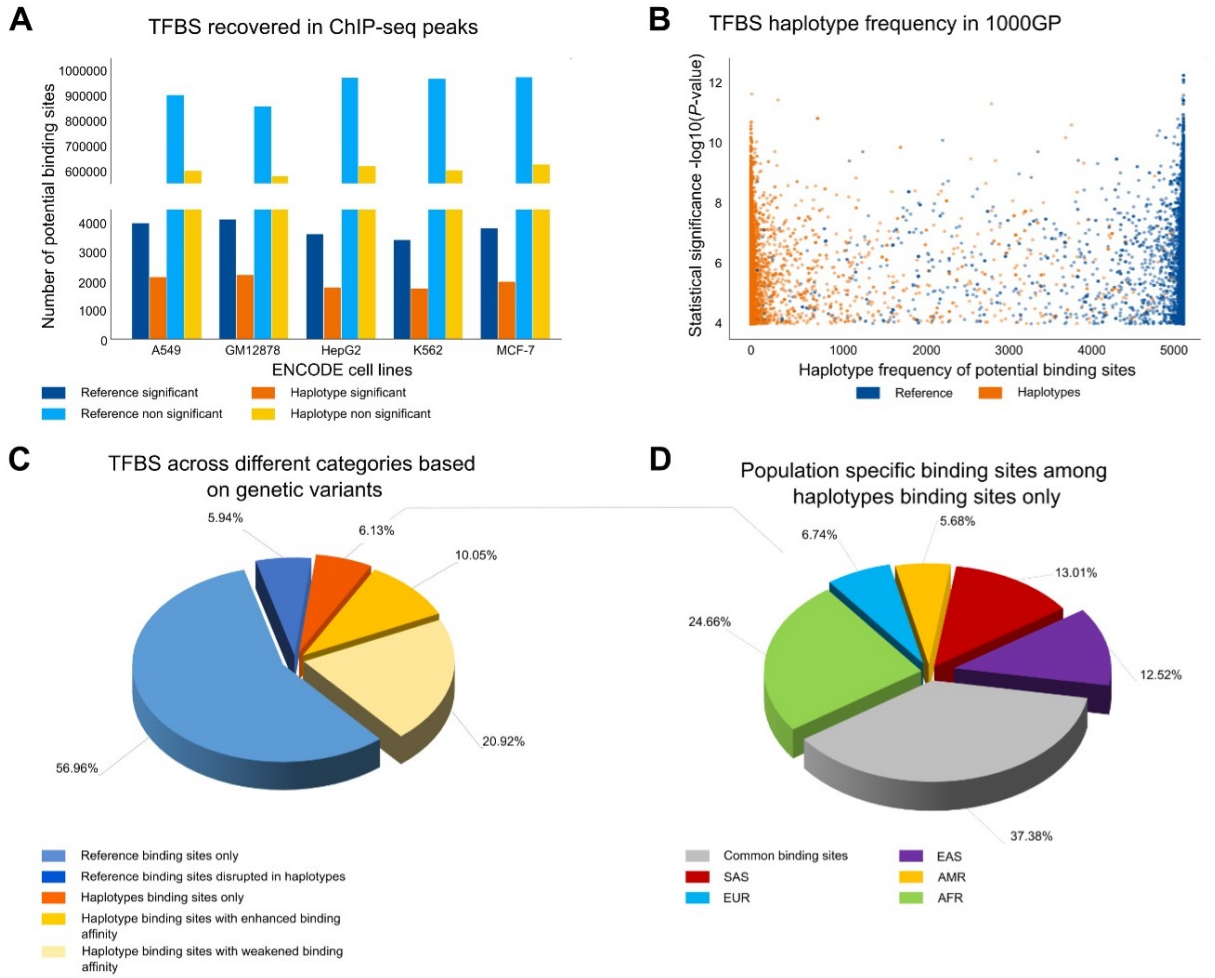
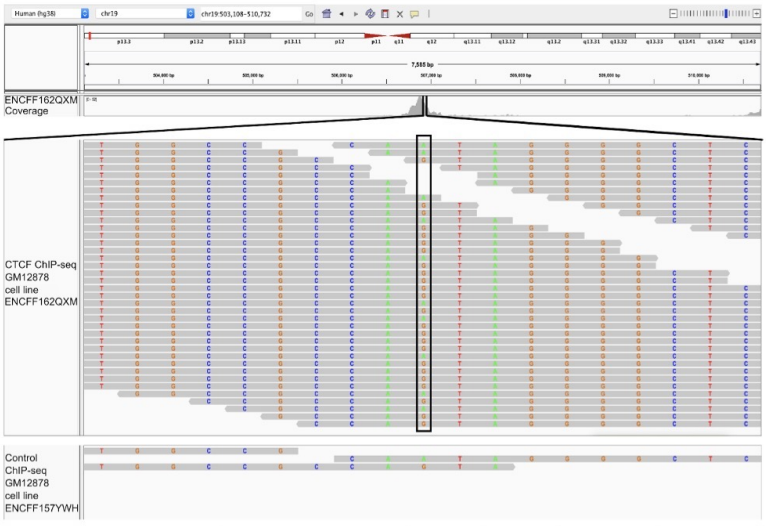


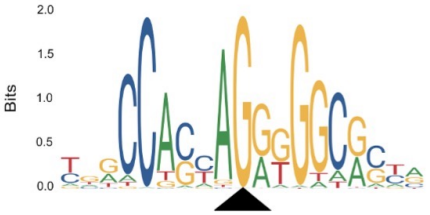
Figure 5.12. Searching CTCF motif on genome graphs using GRAFIMO provides insights on the impact of genetic diversity on putative binding sites. **(A)** Statistically significant (P-value $< 1e^{-4}$) and non-significant potential CTCF occurrences found in the reference and haplotype sequences found using GRAFIMO on 1KGP genome graph. **(B)** Statistical significance of the identified CTCF occurrences and their frequency within the haplotypes embedded in the genome graph. **(C)** Percentage of statistically significant potential CTCF binding sites found only in the reference genome, only in alternative haplotypes, and with their binding scores modulated by 1KGP genetic variants. **(D)** Percentage of population-specific and common (shared by two or more populations) CTCF binding sites found in individual haplotypes.

5.2.2 Searching for ATF3 occurrences

Activating Transcription Factor 3 (ATF3) is a member of the cAMP responsive element-binding family, exhibits increased activity in response to physiological stress across diverse tissues (Chen *et al.*, 1996). Moreover, ATF3 has versatile roles in immunity and cancer (Thompson *et al.*, 2009). ATF3 binds to short conserved DNA sequences. In our investigations, we searched ATF3 motif (JASPAR ID MA0605.2) (**Fig.5.11 (B)**) occurrences within the 1KGP genome graph. For subsequent analyses we considered ATF3 occurrences with P-value $< 1e^{-4}$ as potential binding sites. Our results unveiled several potential motif occurrences that would be lost scanning only the reference genome sequences (**Fig.5.14 (A)**). Additionally, we observed that several ATF3 motif occurrences with high statistical significance were identified in alternative haplotypes (**Fig.5.14 (B)**). Furthermore, we observed that 7.03% of potential ATF3 binding sites are exclusively identified in non-reference haplotype sequences, 11.28% are disrupted by genomic variants, and approximately 13% of ATF3 TFBS maintain significance in non-reference haplotypes but exhibit different binding scores (**Fig.5.14 (C)**). Importantly, a substantial fraction of putative ATF3 binding sites displayed population specificity: 19.81%, 3.77%, 7.55%, 19.81%, and 19.81% of potential binding sites retrieved in individual haplotypes were specific to AFR, EUR, AMR, SAS, EAS populations, respectively (**Fig.5.14 (D)**).



Reads	#A in REF	#G in ALT
ChIP-seq	10 (29.41%)	24 (70.59%)
Control	1 (50%)	1 (50%)



Sequence	Genomic location	Strand	*P*-value	Reference	Haplotype frequency
TGGCCGCCAGTAGGGGCTC	chr19:506,910-506,529	+	$7.64e^{-5}$	non ref.	3529
TGGCCGCCAATAGGGGCTC	chr19:506,910-506,529	+	$1.31e^{-3}$	ref.	1567

Figure 5.13. Considering genomic diversity, GRAFIMO captures additional binding events. GRAFIMO reports a potential CTCF binding site at chr19:506,910-506,929 exclusively present in haplotype sequences. The binding site was identified by scanning CTCF ChIP-seq peaks (experiment ENCSR000DZN on GM12878). The analysis of ChIP-seq reads (ENCFF162QXM) unveils an allelic imbalance at position 10 of the motif, towards the alternative allele (G). GRAFIMO accurately captures this imbalance by reporting sequences carrying a G at position 10 (as found in the alternative haplotypes), while the potential TFBS on the reference, which carries an A, is not reported as statistically significant. This discrepancy aligns with CTCF motif logo, which illustrates G as the dominant nucleotide at position 10.

5.2.3 Searching for GATA1 occurrences

GATA1 is a zinc-finger transcription factor playing a pivotal role in the development of hematopoietic cell lineages (Calligaris *et al.*, 1995). GATA1 binds short (11 bp) highly conserved DNA sequences. In our experiments, we searched GATA1 motif (JASPAR ID MA0035.4) (**Fig.5.11**) occurrences within the 1KGP genome graph. In our downstream analyses we considered GATA1 occurrences with *P*-value $< 1e^{-4}$ as potential binding sites. In our experiments we observed that several GATA1 occurrences are lost when not considering genetic diversity while scanning genomic sequences (**Fig.5.15 (A)**). Moreover, several potential motif occurrences with high statistical significance were found scanning non-reference haplotypes (**Fig.5.15 (B)**). Further investigations unveiled that 9.78% of potential GATA1 binding sites are exclusive to non-reference haplotype sequences, 12.58% are disrupted by genetic variants, and ∼4% maintain significance in non-reference haplotypes but exhibit different binding scores (**Fig.5.15 (C)**). We also identified population specific GATA1 binding sites among those retrieved only in individual haplotypes, with 25.97% specific to AFR, 3.90% to EUR, 9.09% to AMR, 19.48% to SAS, and 11.69% to EAS populations (**Fig.5.15(D)**).

5.3 Comparing GRAFIMO and FIMO

To validate GRAFIMO accuracy, we conducted a comparative analysis with FIMO, running the latter on the same ChIP-seq regions used to test the former. FIMO and GRAFIMO runs were performed on a Linux-based machine with an Intel(R) Core (TM) i7-5960X 3.00GHz CPU (16 cores) and 64 GB of memory (RAM). Since FIMO scans sets of sequences given as FASTA files, for each investigated TF, we recovered the reference genome sequences representing the ChIP-seq optimal IDR thresholded peaks using BEDTools (Quinlan and Hall, 2010). For each studied TF, we observed that GRAFIMO reports all potential motif occurrences identified by FIMO. Therefore, GRAFIMO successfully identifies additional motif candidates found in individual haplotypes embedded in the genome graph without sacrificing potential motif occurrences in the reference genome sequence. We further benchmarked GRAFIMO against FIMO in terms of running time and memory usage. Using CTCF motif (19 bp) as an example, we searched for potential motif occurrences on forward and reverse strands in 1000 genomic regions of human chr22 with increasing length (1 to 9 Mb). To run FIMO, for each set of genomic regions we created the corresponding

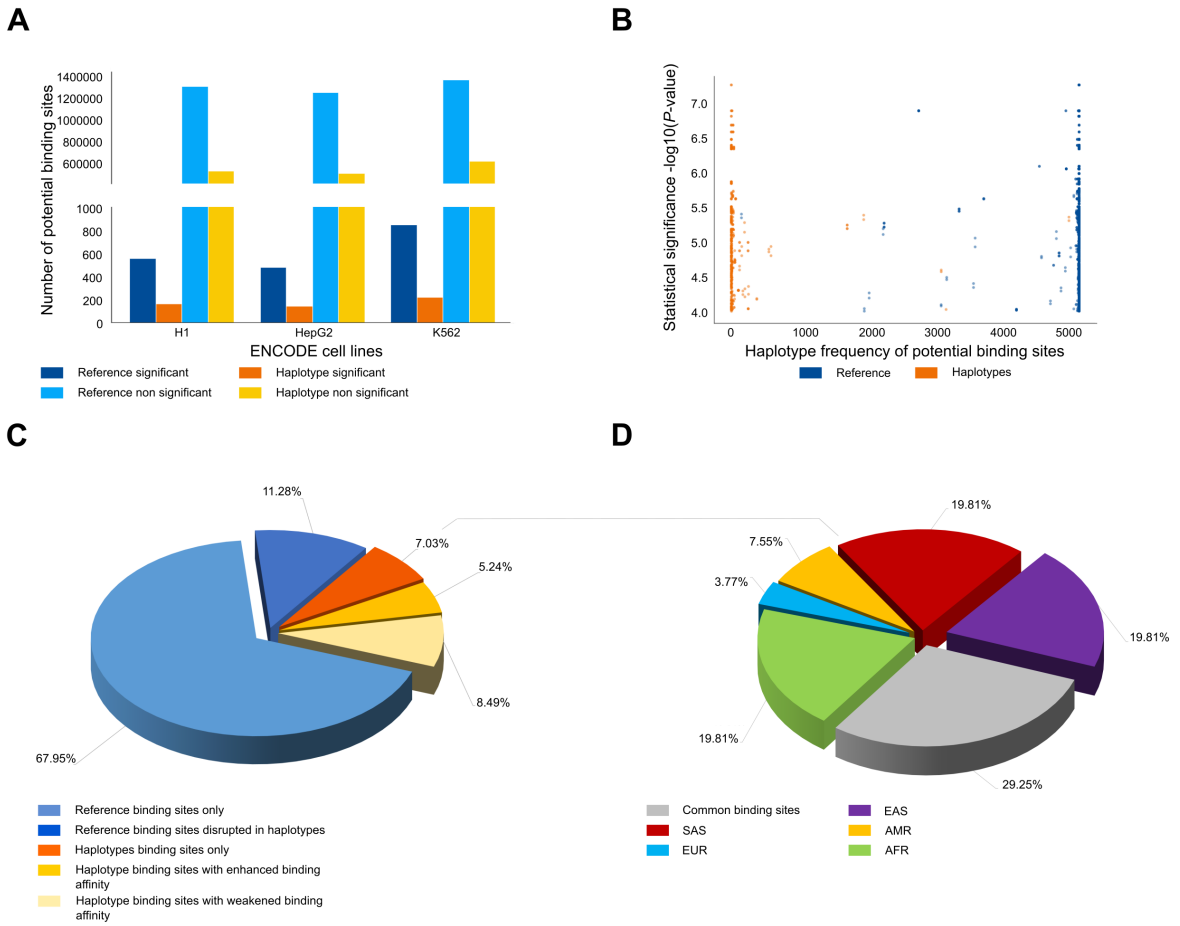


Figure 5.14. Searching ATF3 motif on genome graphs using GRAFIMO provides insights on the impact of genetic diversity on putative binding sites. (A) Statistically significant (P-value $< 1e^{-4}$) and non-significant potential ATF3 occurrences found in the reference and haplotype sequences found using GRAFIMO on 1KGP genome graph. **(B)** Statistical significance of the identified ATF3 occurrences and their frequency within the haplotypes embedded in the genome graph. **(C)** Percentage of statistically significant potential ATF3 binding sites found only in the reference genome, only in alternative haplotypes, and with their binding scores modulated by 1KGP genetic variants. **(D)** Percentage of population-specific and common (shared by two or more populations) ATF3 binding sites found in individual haplotypes.

FASTA file. To assess GRAFIMO performance we scanned chr22 genome graph without encoded variants, on the previously computed set of regions. Since FIMO does not provide a parallel implementation we run GRAFIMO using a single thread. Although FIMO exhibits superior speed and lower memory requirements in a single-thread scenario (**Fig.5.16 (A)** and **(B)**), GRAFIMO surpasses FIMO in speed when scanning regions accounting for genetic variants of 2,548 individuals (**Fig.5.16 (C)**). We excluded from the measurements for FIMO the overhead introduced to compute the FASTA files from the original BED files. These results were expected since FIMO is a highly optimized tool scanning linear reference genomic sequences, while GRAFIMO efficiency shines while scanning panels of individuals accounting for their genetic diversity. We further evaluated GRAFIMO performance using 1, 4, 8, and 16 threads to scan the chr22 genome graph embedding the 2,548 individuals and their genetic variants from 1KGP (**Fig.5.17**). As expected, using multiple threads significantly reduces running time, although memory usage remains consistent with the number of threads. For the analyses presented in the previous section each scan on average took ∼15 minutes and consumed around 24 GB of memory.

5.4 Discussion and limitations

By leveraging genome graphs, GRAFIMO introduces an effective approach for investigating the impact of genetic variation on the binding landscape of transcription factors (TFs) across diverse populations. Notably, our findings revealed the presence of numerous potential and exclusive TF binding sites (TFBS) in individual haplotype sequences. Additionally, genomic variants exert a substantial influence on the binding affinity of several motif occurrence candidates identified in the reference genome sequence. This

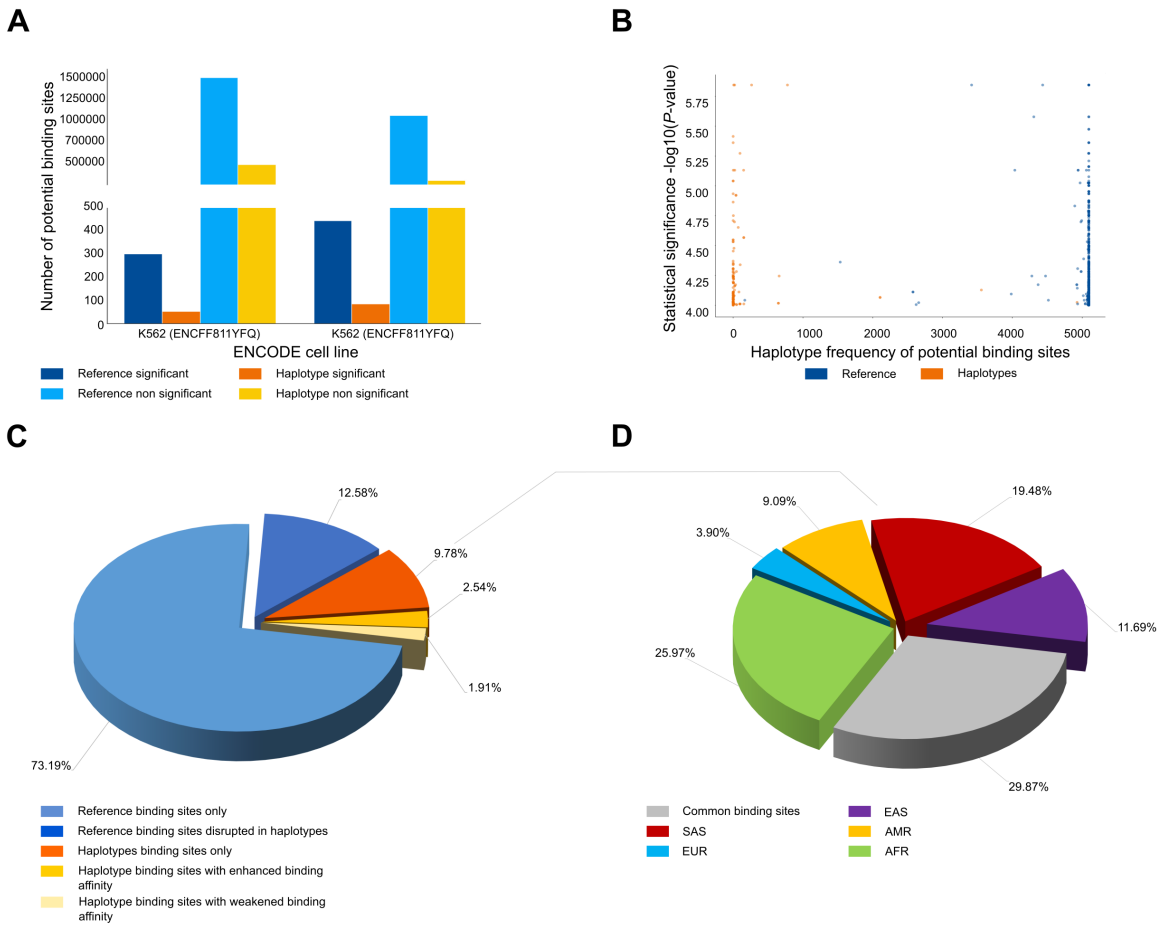


Figure 5.15. Searching GATA1 motif on genome graphs using GRAFIMO provides insights on the impact of genetic diversity on putative binding sites. **(A)** Statistically significant (P-value $< 1e^{-4}$) and non-significant potential GATA1 occurrences found in the reference and haplotype sequences found using GRAFIMO on 1KGP genome graph. **(B)** Statistical significance of the identified GATA1 occurrences and their frequency within the haplotypes embedded in the genome graph. **(C)** Percentage of statistically significant potential GATA1 binding sites found only in the reference genome, only in alternative haplotypes, and with their binding scores modulated by 1KGP genetic variants. **(D)** Percentage of population-specific and common (shared by two or more populations) GATA1 binding sites found in individual haplotypes.

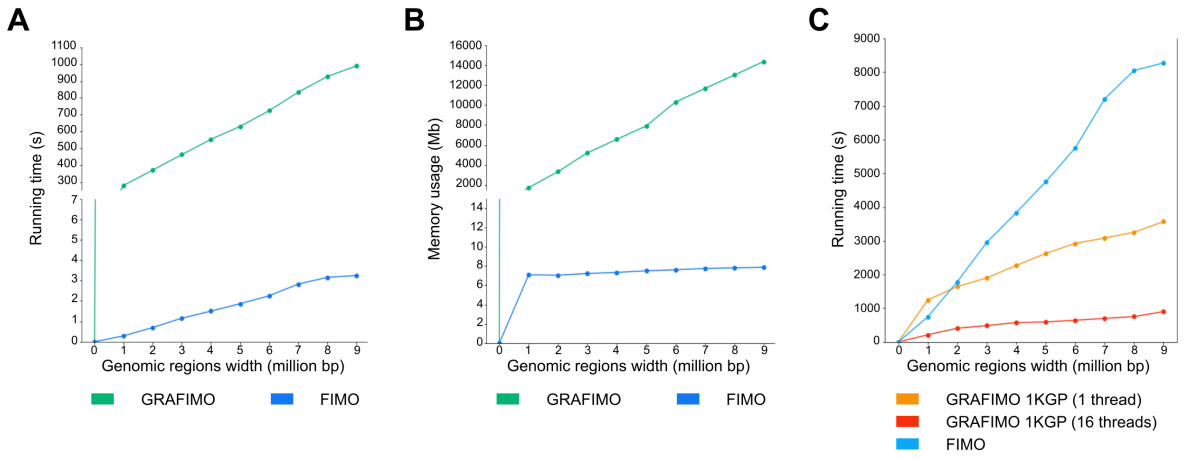


Figure 5.16. Comparing GRAFIMO and FIMO performance. **(A)** FIMO is faster than GRAFIMO (using 1 thread) when searching CTCF motif (JASPAR ID MA0139.1) on human chr22 regions (total width ranging from 1 to 9 Mb) and without accounting for genetic variants. **(B)** FIMO uses less memory than GRAFIMO, however it only scan reference sequences. **(C)** GRAFIMO is generally faster than FIMO while searching CTCF occurrences when considering genetic diversity in large panels of individuals (e.g. 1KGP phase 3), even with single thread.

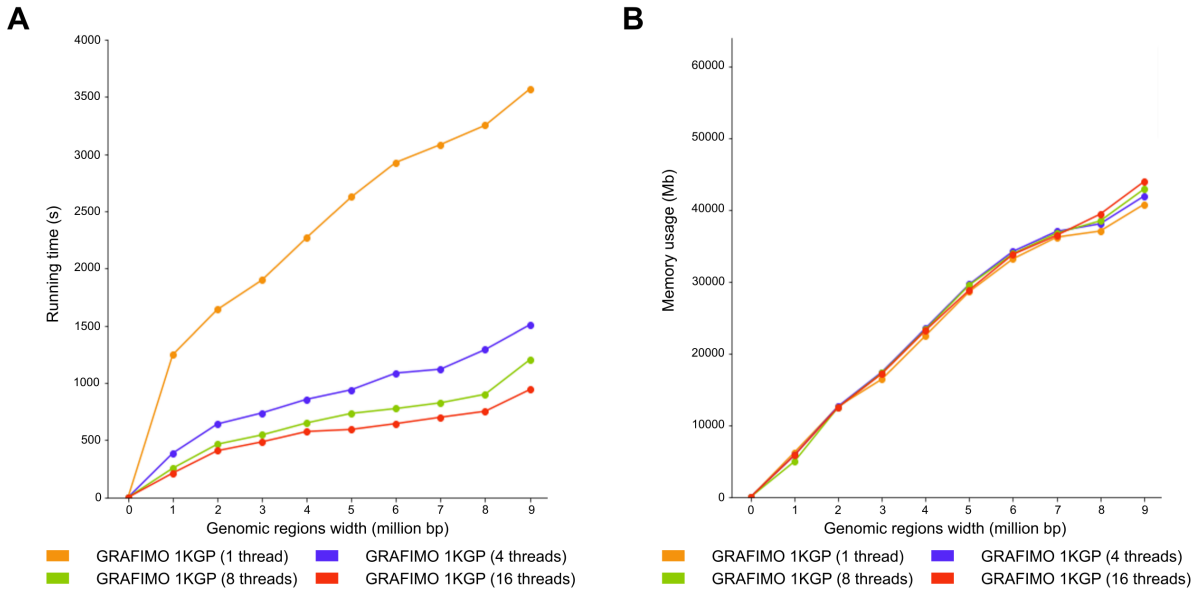


Figure 5.17. GRAFIMO running time efficiently scales with the number of threads. By running GRAFIMO with multiple threads **(A)** the running time significantly decreases, while **(B)** memory usage remains similar.

tool offers a valuable resource for prioritizing regions that could underlie individual-specific alterations in gene expression, a facet often overlooked when relying solely on reference genomes.

MotifRaptor: a Transcription Factor-centric method to evaluate non-coding genetic variants impact

Several studies highlighted the impact of genetic variants on the modulation of transcription factor binding affinity (De Gobbi *et al.*, 2006; Weinhold *et al.*, 2014; Wienert *et al.*, 2015). Genome-wide association studies (GWASs) revealed numerous SNPs associated with complex traits or human diseases (Buniello *et al.*, 2019). However, despite these extensive endeavors, functional studies aimed at prioritizing potential causal variants have faced challenges, leading to a constrained understanding of the pathophysiological mechanisms linking variants to phenotypes (Gallagher and Chen-Plotkin, 2018). This poses challenges when interpreting the functional impact of individual- or population-specific variants on TF binding landscape. Missense variants may modulate the function of a TF by influencing its coding sequence, altering the protein structure, and consequently impacting its DNA binding capability. This scenario is particularly relevant in Mendelian diseases (Barrera *et al.*, 2016). Conversely, for common diseases and complex traits, most associated variants (over 90%) are located in non-coding regions, predominantly within DNase I hypersensitive sites. Often non-coding variants are located within genomic regulatory elements, such as enhancers, silencers, or promoters (Maurano *et al.*, 2012). Mutations that alter TF binding activities may serve as mediators for chromatin state alterations and gene deregulation. Therefore, genetic variants within non-coding domains may alter TF recognition sequences, enhancing or disrupting TF-DNA binding events, consequently inducing changes in downstream gene expression programs (Deplancke *et al.*, 2016). While individual non-coding variants may exert only moderate effects on binding sites and lack the power to comprehensively elucidate gene expression programs, the statistical analysis of a set of SNPs that modulate common TF binding sites holds the promise of revealing convergent regulatory mechanisms underlying complex traits. Despite the existence of various approaches aimed at investigating the impact of genetic variants on TF binding sites, challenges persist in achieving a comprehensive understanding (Tognon *et al.*, 2023). MotifRaptor (Yao *et al.*, 2021) addresses these challenges by providing a unique framework. The rationale behind its development stems from the limitations of existing tools, such as MMARGE (Link *et al.*, 2018), GERV (Zeng *et al.*, 2016b), DeepSEA (Zhou and Troyanskaya, 2015), Basset (Kelley *et al.*, 2016), or IMPACT (Amariuta *et al.*, 2019), which heavily rely on the availability of genome-wide maps of TF occupancy and chromatin marks in specific cellular contexts. MotifRaptor, instead, accommodates scenarios where only motif PWM models and gene expression data are available. Existing models based on ChIP-seq or PWM data lack a systematic approach to globally rank and assess the significance of TFs based on all trait-associated variants. While tools like CADD (Maurano *et al.*, 2015), CENTIPEDE (Moyerbrailean *et al.*, 2016), or atSNP (Zuo *et al.*, 2015) offer valuable insights, they are limited in providing a formal testing procedure for the global effect of a set of GWAS variants on the overlapping TF binding sites. MotifRaptor overcomes this limitation by introducing a novel genome-wide statistic. This genome-wide statistic prioritizes putative causal TFs based on the entire set of binding sites and overlapping variants, offering a more comprehensive approach. Furthermore, linkage disequilibrium (LD) is often overlooked in current methods, leading to potential confounding in the analysis. MotifRaptor addresses this issue by implementing strategies that reduce the space of variants in each LD block and explicitly account for local LD structure. By focusing on cell type-specific chromatin accessibility regions and sampling the background set of chromatin accessibility regions in close proximity, Motif-Raptor mitigates the problem of false positives, offering specificity in identifying variants within regions that are cell type-specific. In comparison to available tools, MotifRaptor overcomes these challenges and provides a genome-wide significance score for each TF, taking into account the directional modulation of TF binding sites by variants. While tools like SLDP (Reshef *et al.*,

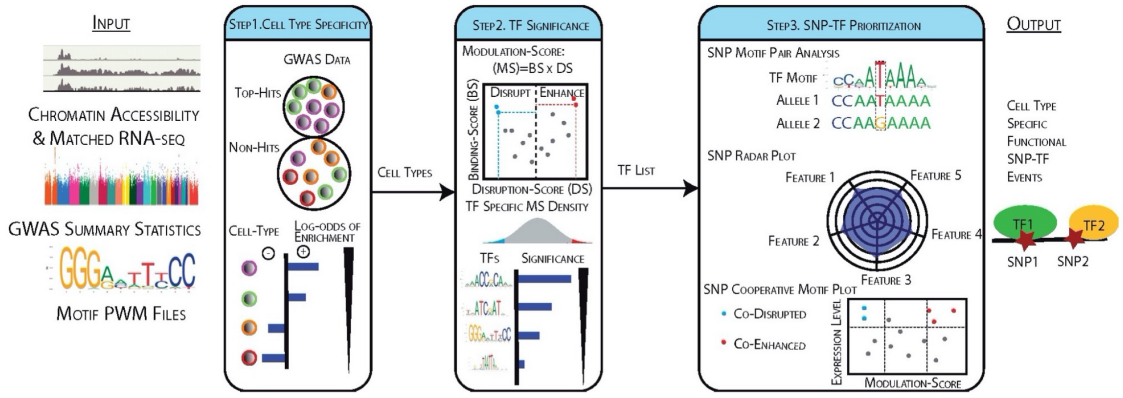


Figure 6.18. MotifRaptor analysis pipeline. MotifRaptor executes a three-step process to enhance its functionality: (i) characterization of pertinent cell types, determined by the enrichment of trait-associated SNPs within chromatin accessible sites; (ii) identification of TFs with binding sites significantly influenced by genetic variants in these cell types; and (iii) pinpointing and visually representing individual TF-SNP regulatory events (taken from Yao *et al.* (2021)).

2018) and GREGOR (Schmidt *et al.*, 2015) consider LD structure, MotifRaptor offers a distinctive approach with its TF-centric analysis, statistics, and visualization functionalities. This tool is poised to enhance the discovery and interpretation of the impact of non-coding variants on crucial regulators of complex traits, providing a scalable framework to shift towards precision medicine-oriented methods.

6.1 Design and implementation

MotifRaptor analysis requires GWAS summary statistics, TF PWM models (PFM format (**Appendix A.9**)), chromatin accessibility (BED format (**Appendix A.6**)), and transcriptomic data as input. MotifRaptor leverages and interpolates the input data to generate a prioritized list of putative trait-associated TFs. This list contains TFs whose binding sites may be modulated by genetic variants in different cell types. Furthermore, the tool offers intuitive visualizations for exploring external annotations and assessing potential co-factor involvement for each variant. MotifRaptor analysis involves three key steps: (i) characterizing relevant cell and tissue types through the enrichment of the trait-associated SNPs in open chromatin regions, (ii) identifying significant TFs in these cell types, and (iii) pinpointing and annotating crucial TF-SNP regulatory events (**Figure 6.18**). Developed as an open-source command-line utility, MotifRaptor is implemented with Python and Cython.

6.1.1 Quantifying genetic variants impact on Transcription Factor Binding Sites

To evaluate the impact of genetic variants on TFBS, MotifRaptor implements an efficient and comprehensive genome-wide scoring procedure. MotifRaptor scoring procedure scans all binding sites overlapping the input variants. Although the scanning procedure is similar to the algorithm implemented by other motif scanning algorithms, such as MOODS (Korhonen *et al.*, 2009), FIMO (Grant *et al.*, 2011) or GRAFIMO (Tognon *et al.*, 2021), MotifRaptor proposes a threshold-free scanning. Threshold-free motif scanning enables to efficiently compute modulation scores and null models, used to test scores statistical significance. Given a genomic sequence s, where $|s| = m$, and a PWM, the score $M(i, S_i)$ denotes the likelihood of observing the nucleotide $S_i \in \{A, C, G, T\}$ at position i, where $1 \leq i \leq m$. We derive the binding score BS from $M(i, S_i)$ as a log-likelihood score computed over the entire binding site and corrected to account for genome-wide nucleotide frequencies, or background frequencies, $B(s_i)$:

$$\mathrm{BS} = \log \prod_{i=1}^{m} \frac{M(i, s_i)}{B(s_i)} = \sum_{i=1}^{m} (\log M(i, s_i)) - \log B(s_i))$$

Based on this formulation, MotifRaptor computes a disruption score DS, which reflects the potential impact of genetic variants on a given binding site. Given a target variant, MotifRaptor assumes two haplotypes, denoted as s_{ref} and s_{alt} for the reference and alternative alleles, respectively. MotifRaptor

assigns a BS to each allele, resulting in binding scores for the reference and alternative alleles (i.e. $\mathrm{BS}(s_{ref})$ and $\mathrm{BS}(s_{alt})$). To ensure computational scalability, MotifRaptor restricts the calculations to a region spanning 61 bp, centered around the target variant. For each region, MotifRaptor considers the best putative binding position K for both s_{ref} and s_{alt}:

$$K = \underset{1 \leq k \leq m}{\arg\max}(\mathrm{BS}(s_{ref,k:k+m-1}, M), \mathrm{BS}(s_{alt,k:k+m-1}, M))$$

Therefore, the disruption score DS at position K is defined as:

$$\mathrm{DS} = \Delta\mathrm{BS}$$

The magnitude and sign of DS are valuable indicators, conveying information about the direction and strength of the target variant's impact on TF binding. Specifically, positive DS suggest an increased binding affinity, while negative values imply a reduced binding potential. Since TFBS motifs exhibit varying length and specificity, MotifRaptor scales both BS and DS to the ranges $[0, 1]$ and $[-1, 1]$, respectively:

$$\mathrm{BS}_{\max}(M) = \max_{s_i \in \{A,C,G,T\}} \sum_{i=1}^{m} \log M(i, s_i) - \log B(s_i)$$

$$\mathrm{DS}_{\max}(M) = \max_{s \neq w \in \{A,C,G,T\}} |\log M(i, s_i) - \log M(i, w_i)|$$

Given the scaled scores, MotifRaptor establishes a Binding-Disruption (BD) space. The BD-space offers a comprehensive framework to visualize and summarize TF-SNP modulation events globally and across different factors. Within the BD-space, events located near the distal corners from the origin (coordinates $(1, 1)$ or $(1, 1)$) signify robust binding and substantial modulation induced by genetic variants. Furthermore, MotifRaptor combines BS and DS into a composite score, called modulation score (MS):

$$\mathrm{MS}(s, M) = \left(\frac{\mathrm{BS}(s, M)}{\mathrm{BS}_{\max}(M)}\right) \times \left(\frac{\mathrm{DS}(s, M)}{\mathrm{DS}_{\max}(M)}\right)$$

MS encapsulates the rectangular area spanned by the scaled DS and BS. Larger absolute MS values, correspond to events close to the distal corners, and indicate meaningful modulation. Since no parametric distributions adequately fit the observed data, MotifRaptor proposes to determine the statistical significance of SNP-TF modulation events, by estimating a null model based on the central limit theorem (CLT). However, this strategy involves a complete enumeration of all potential binding sites across the genome. MotifRaptor overcomes this issue by leveraging efficient data structures, that efficiently compute genome-wide BS and DS for all input SNPs and TFs. Importantly, MotifRaptor avoids to rely on pre-determined scores, P-value cutoffs, or computationally intensive shuffling procedures. Moreover, MotifRaptor generates a ranked list of transcription factors built defining a TF score, that combines its cell type-specific expression and MS in chromatin accessible regions. Since TFs may share similar motifs but their expression and function is different across cellular contexts, the TF score, given a motif M and a cell type C, is defined as:

$$\text{TF-score}(M, C) = \mathrm{EP}(M, C) \times (1 - \min \mathrm{FDR}(M, C))$$

where the TF-score is constrained in $[0, 1]$, $\mathrm{EP}(M, C)$ is the expression percentile, and $\min \mathrm{FDR}(M, C)$ is the corrected P-value for MS significance. $\min \mathrm{FDR}(M, C)$ compares MSs' distribution in cell type-specific chromatin accessible regions.

6.1.2 Fast genome-wide motif scanning

The construction of a null model to statistically evaluate the BD-space and MS requires the exhaustive enumeration of all potential TFBS across the genome coupled with an assessment of their potential modulation by SNPs unrelated to any phenotype. While tools like FIMO or MOODS offer fast binding site occurrences enumeration, they are not optimized to efficiently compute the modulation of binding affinity induced by a set of SNPs (Zuo *et al.*, 2015). GRAFIMO partially overcomes this limitation, by considering genetic variation while searching motif occurrences (**Chapter 5**). However, some downstream analysis is required to recover the binding affinity modulation induced by genetic variants. MOODS, FIMO or GRAFIMO rely on a pre-defined threshold on statistical significance (P-value or q-value) to filter potential false positives or enhance computational efficiency. However, this strategy may exclude

weak but genuine binding events (Boeva, 2016; Tognon *et al.*, 2023). In MotifRaptor context, if employed, this strategy may bias the estimation of the complete empirical distribution of MSs. Moreover, establishing a uniform threshold for different TFBS motifs is neither practical nor reasonable (Tognon *et al.*, 2023). In fact, different TFBS display varying lengths and complexities. atSNP (Zuo *et al.*, 2015) addresses the first challenge by proposing an efficient P-value estimation procedure. However, it lacks an effective threshold-free scanning technique. On the other hand, MotifRaptor overcomes these limitation by avoiding redundant calculations. To avoid redundant calculations, MotifRaptor leverages two text-indexing data structures: a suffix array (SA) (Puglisi *et al.*, 2007) and a longest common prefix (LCP) array (Landau *et al.*, 2001). SA and LCP enables bypassing of repeated portions of the genome, significantly reducing redundancy in calculations. MotifRaptor algorithm performing fast genome-wide motif scanning encompasses different steps (**Algorithm 4**). MotifRaptor starts by recovering all input SNPs and their flanking genomic sequences. Let us assume N input variants and that the length of the flanking sequences surrounding each SNP is $2m-1$, where $m = |M|$. MotifRaptor enrich each reference sequence with its linked variant returning two sequences: s_{ref} and s_{alt}. Subsequently, s_{ref} and s_{alt} are concatenated constructing a single sequence, and the original sequences' genome positions are recorded. This results in a unique extended pseudo-genome (P) of length $2 \times (2m-1) \times N$. MotifRaptor then computes the SA and LCP arrays on P in linear time and space ($O(mN)$), using the divsufsort algorithm (Fischer and Kurpicz, 2017). By leveraging SA and LCP data structures, MotifRaptor scan sequences in lexicographical order. Let us assume the algorithm is at iteration j and it inductively computed an auxiliary array BS$[1:m]$, storing binding scores between M and the first m nucleotides in P starting at position SA$[j-1]$. Then, BS$[m] = BS(P[\text{SA}[j-1]])$ against entire M. At iteration $j = 0$, MotifRaptor fills BS with 0s. At iteration $j = 1$, MotifRaptor inductively computes BS$[1:m]$ for the first m nucleotides of suffix SA$[j]$. The existing values in BS$[1:m]$ referring to SA$[j-1]$ and the LCP$[j]$ value are exploited for efficient computation. If the common prefix's length between SA$[j-1]$ and SA$[j]$ is $p = \text{LCP}[j]$, the update of binding scores can start from position $p+1$, since the binding scores in BS$[1:p]$ remain unchanged, given that SA$[j-1]$ and SA$[j]$ share the initial p characters for which binding scores have already been computed, BS$[1:p]$. Explicitly, for $p+1 \leq x \leq m$:

$$\text{BS}[x] = \text{BS}[p] + \sum_{i=p+1}^{x} \log M(i, s_i) - \log B(s_i)$$

To map the binding score BS$[m]$ relative to an substring s, where $s \in P$ and $|s| = m$, aligned against the motif M, back to the SNP site, MotifRaptor determines the SNP position and the binding position according to:

$$\text{SNP} = \left\lfloor \frac{\text{SA}[j]}{2m-1} \right\rfloor$$

$$\text{binding\_site} = \text{SA}[j] \mod (2m-1)$$

It is fundamental to note that not all prefixes w, where $|w| = m$, of P's suffixes are substrings of the original genome. Recall that P is constructed by concatenating sequences blocks centered around SNP positions. MotifRaptor further improves the running time by skipping cross-block sequences scanning. By exploiting these data structures and score design, MotifRaptor efficiently compute genome-wide binding and disruption scores for each TF, enabling complete null models estimation factor.

Algorithm 4: MotifRaptor motif scanning algorithm

Data: SNPs, G, M
Result: BS
$P \leftarrow$ computePseudoGenome(SNPs, G, M) ; `/* see algorithm 5 */`
SA $\leftarrow$ computeSuffixArray(P) ; `/* use divsufsort algorithm */`
LCP $\leftarrow$ computeLCP(P, SA);
BS $\leftarrow$ initializeArray($0, |P|$);
for $i \in |BS|$ **do**
  BS$[i] \leftarrow$ scoreSequence($P[i:i+m], M, \text{SA}[i-1], \text{LCP}[i]$);
return BS

Algorithm 5: Algorithm computing Pseudo-genome P

Data: SNPs, G, M
Result: P

$m \leftarrow |M|$;
$P \leftarrow \emptyset$;
for *snp* $\in$ *SNPs* **do**
    $s_{\text{ref}} \leftarrow G[\text{snp}_{\text{pos}} - m : \text{snp}_{\text{pos}} + m]$;
    $s_{\text{alt}} \leftarrow \text{enrichSequence}(s_{\text{ref}}, \text{snp})$;
    $s \leftarrow \text{concatenate}(s_{\text{ref}}, s_{\text{alt}})$;
    $P \leftarrow \text{concatenate}(P, s)$;
return P

6.1.3 Evaluating TF-SNP modulations significance

To assess the statistical significance of TF-SNP modulation scores (MSs), MotifRaptor employs a method leveraging the central limit theorem (CLT). This involves estimating a complete null model through the enumeration of all potential binding sites across the genome. To ensure scalability, MotifRaptor leverages suffix and LCP arrays, that enable exhaustive enumeration while maintaining method's scalability. For a given set of target variants V, MotifRaptor evaluates whether they significantly modulate TF binding events by testing if MS distribution $D_{\text{MS}}(V)$ significantly differs from the MS distribution obtained from a set of background (genome-wide) SNPs W ($D_{\text{MS}}(W)$). To assess the distinction between the distributions, MotifRaptor employs a non-parametric test with a null hypothesis:

$$E(D_{\text{MS}}(V)) = E(D_{\text{MS}}(W))$$

According to CLT, the distribution of the sample mean of W, converges to a normal distribution, with its mean equal to $E(D_{\text{MS}}(W))$, and variance equal to $\frac{\text{Var}(D_{\text{MS}}(W))}{N_{\text{sample}}}$, regardless of the underlying MS distribution. Consequently, MotifRaptor assesses the significance of enhanced scores $E(D_{\text{MS}}(V))$, disrupted scores $E(D_{\text{MS}}(W))$, or both. This robust procedure, combined with efficient data structures, enables the identification of significant genome-wide shifts in TF-SNP binding modulations without relying on predetermined scores or cutoffs on P-values.

6.2 MotifRaptor helps prioritizing variants affecting Transcription Factor Binding Sites involved in LDL-C uptake

We demonstrate the capabilities of MotifRaptor by utilizing the tool to prioritize potential TFs modulated by a set of lipoprotein cholesterol (LDL-C)-associated GWAS variants and low-density lipoprotein receptor (LDLR) coding variants on LDL-C uptake in HepG2 hepatocellular carcinoma cells. Genetic differences in LDL-C levels significantly impact the risk of coronary artery disease. This is supported by quantitative serum LDL-C measurements, which are consistently recorded across various biobanks, making them one of the most reliable human phenotypic datasets available. Through a trans-ancestry GWAS meta-analysis conducted by the Global Lipids Genetics Consortium (GLGC), over 900 genome-wide significant loci associated with blood lipid levels have been identified, with more than 400 loci specifically linked to LDL-C (Graham *et al.*, 2021). Notably, these LDL-C GWAS loci exhibit a significant overlap with gene expression patterns in the liver, suggesting the liver's pivotal role in modulating the effects of LDL-C variants (Wang *et al.*, 2022a; Finucane *et al.*, 2018). Despite this correlation, the specific causal variants and mechanisms underlying many of these loci's influence on LDL-C levels remain elusive. LDL-C levels are profoundly influenced by rare coding variants, with the most severe cases linked to Familial Hypercholesterolemia (FH), a disease characterized by markedly elevated LDL-C levels and premature cardiovascular disease (Bouhairie and Goldberg, 2015). FH arises from inherited monogenic variants in several genes, primarily affecting LDLR, a cell surface receptor pivotal for LDL uptake, thereby regulating its circulation levels (Brown and Goldstein, 1984). Despite the efficacy of lipid-lowering therapies, individuals with FH still face a significantly heightened risk of coronary events compared to the general population (Mundal *et al.*, 2018). Given that elevated LDL-C levels elevate cardiovascular disease risk across the lifespan, early identification of at-risk individuals holds paramount clinical significance (Bouhairie and Goldberg, 2015). However, many LDLR variants currently lack clear clinical interpretation. For instance, among the 1,427

Variant	Chrom	Position (hg38)	Major Allele	Minor Allele	HepG2 Editable REF Allele	HepG2 Edited Allele	MAF
rs4149309	chr9	104,827,299	A	T	A	G	0,11951
rs12042481	chr1	45,499,094	T	C	T	C	0,169
rs186701924	chr19	56,948,710	A	C	A	G	0,00125
rs4457349	chr8	125,482,075	A	G	A	G	0,0259
rs429358	chr19	44,908,684	T	C	T	C	0,156
rs11149612	chr16	83,947,360	C	T	T	C	0,461
rs4719853	chr7	26,029,463	T	A	T	C	0,237
rs76895963	chr12	4,275,678	T	G	T	C	0,02104
rs35081008	chr19	58,150,868	C	T	T	C	0,147
rs58198139	chr5	156,972,028	C	T	T	C	0,364
rs1250259	chr2	215,435,759	T	A	A	G	0,265
rs704	chr17	28,367,840	G	A	A	G	0,474
rs77542162	chr17	69,085,137	A	G	A	G	0,0229
rs1434282	chr1	199,041,592	C	T	T	C	0,274
rs8126001	chr20	64,080,106	C	T	T	C	0,48964
rs10107182	chr8	58,480,178	C	T	T	C	0,336
rs62084210	chr17	67,832,594	A	G	A	G	0,29613
rs771555783	chr17	80,025,096	CACCAT	C			0,224
rs115421711	chr5	52,787,392	A	G	A	G	0,0413
rs3767844	chr1	214,007,378	A	G	A	G	0,28746
rs2618566	chr20	17,864,040	G	T	T	C	0,339
rs4390169	chr1	155,133,578	A	G	A	G	0,481
rs402348	chr18	63,367,187	T	G	T	C	0,211
rs116734477	chr5	52,799,190	C	T	T	C	0,0412

Table 6.8. LDL-C uptake associated variants prioritized through BEAN and tested for Transcription Factor binding modulation. The table provides details of genetic variants linked to LDL-C uptake, which were prioritized using BEAN (Ryu *et al.*, 2024). These variants underwent testing for modulation of Transcription Factor binding using MotifRaptor (Yao *et al.*, 2021). It includes information such as the genomic location (hg38 genome assembly), major and minor alleles on both the reference genome and HepG2 cells, as well as the minor allele frequency (MAF) associated with each variant. MAF values are sourced from the UK Biobank cohort (Bycroft *et al.*, 2018).

LDLR missense variants cataloged in the ClinVar database (Landrum *et al.*, 2020), half are designated as variants of unknown significance (VUS) or exhibit conflicting interpretations of pathogenicity, thereby complicating FH diagnosis. Similarly, within the UK Biobank cohort (Bycroft *et al.*, 2018), 69% of the 758 unique LDLR missense variants carried by sequenced individuals lack reporting or have uncertain annotations in ClinVar. A more comprehensive understanding of the impacts of LDLR variants would facilitate the earlier diagnosis and treatment of numerous individuals at risk for FH and associated cardiovascular complications. To assess the effects of both common GWAS-associated and rare coding variants in the LDLR gene, we developed a novel approach, called BEAN (Ryu *et al.*, 2024), leveraging base editing techniques (**Section 7.2.1**) followed by the cellular uptake of fluorescent LDL-C in HepG2 cells (**Figure 6.19**). BEAN offers a scalable flow cytometric assay to measure a critical determinant of serum LDL-C levels (Hamilton *et al.*, 2023), as the liver plays a predominant role in clearing LDL-C from circulation (Spady, 1992). Our experimental-computational pipeline enables us to scrutinize LDL uptake-altering GWAS-associated variants and elucidate their downstream effects on chromatin accessibility, transcription factor binding, and gene expression, ultimately impacting LDL uptake. Intriguingly, we identify causal variants that influence LDL-C uptake by affecting the genes OPRL1, VTN, and ZNF329, which were not previously linked to LDL-C levels.

6.2.1 Evaluating candidate variants impact on Transcription Factor Binding potential

To investigate the potential effects of LDL-C uptake-associated variants, prioritized using BEAN (**Table 6.8**), on transcription factor binding sites and their regulatory impact on genes involved in LDL-C uptake, we adapted the pipeline of MotifRaptor for our analysis. Each variant's genomic context was investigated by retrieving sequences spanning 61 bp centered around the SNP location. To recover the genomic context we used the GRCh38 genome assembly as reference. For each variant, we generated both a reference and an alternative sequence by substituting the major allele with the minor allele at the SNP position. Subsequently, we employed all human TF PWMs from the CIS-BP database (Weirauch *et al.*, 2014) to scan each pair of reference and alternative sequences, generating binding scores for each major and minor allele. The comparison of these scores for each TF across the reference and alternative

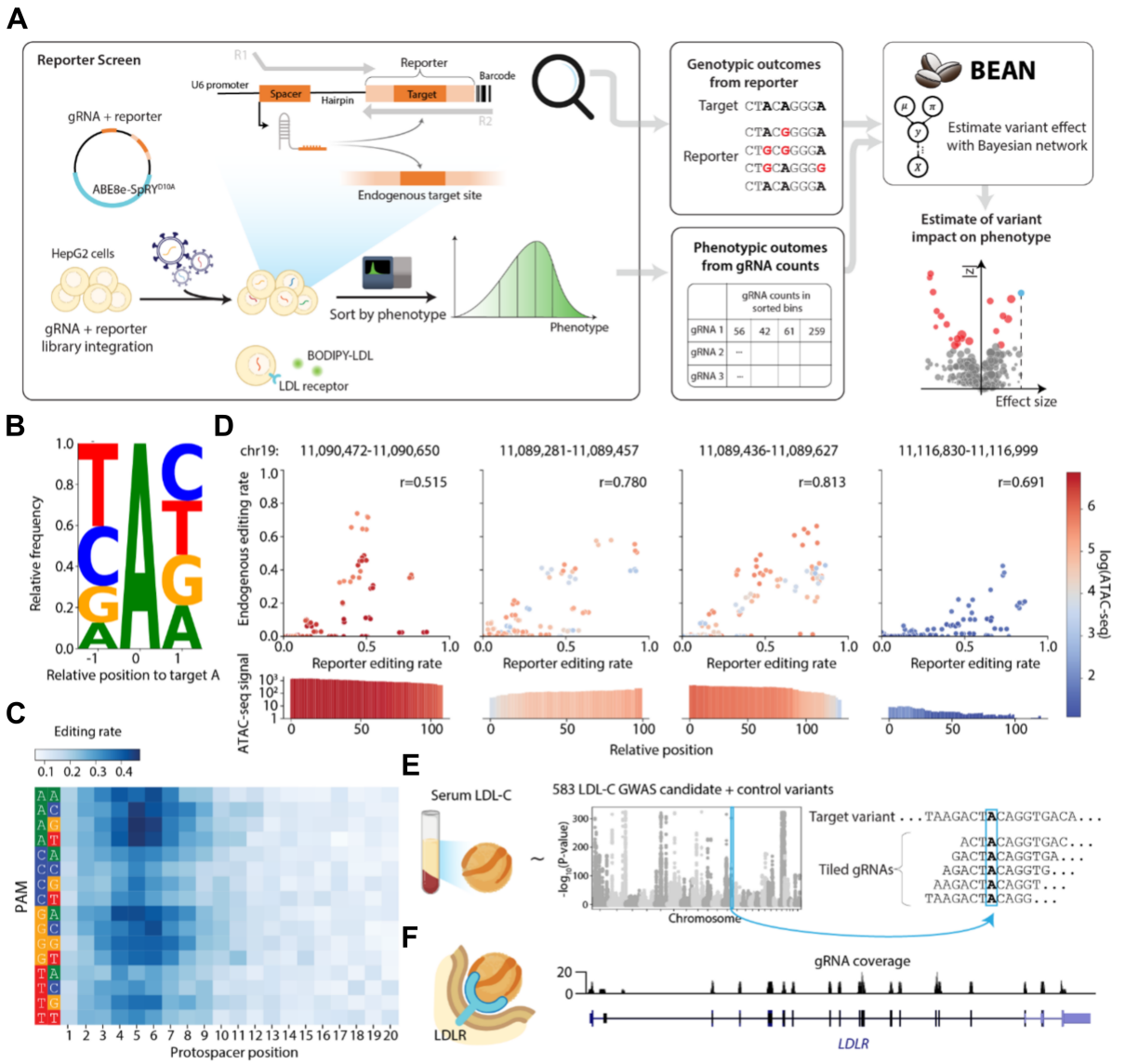


Figure 6.19. BEAN computational-experimental base editing screening pipeline. (A) The plot illustrates the process and analysis workflow for activity-normalized base editing screening, followed by analysis using BEAN. Initially, a library of gRNAs, each coupled with a reporter sequence containing its genomic target, is inserted into a lentiviral base editor expression vector. Lentiviral transduction is conducted in HepG2 cells, succeeded by flow cytometric sorting of four distinct populations based on fluorescent LDL-cholesterol (BODIPY-LDL) uptake levels. Paired-end NGS is then employed to decode gRNA counts and reporter editing outcomes within each flow cytometric bin. Subsequently, BEAN is utilized to model the editing frequency of the reporter and allelic outcomes, along with gRNA enrichments across flow cytometric bins. This enables BEAN to estimate variant phenotypic effect sizes. **(B)** The sequence logo illustrates the adjacent nucleotide specificity of ABE8e-SpRY editing, derived from 7,320 gRNAs. In the logo, the height of each base symbolizes the relative frequency of observing that particular base given an edit at position 0. **(C)** Average editing efficiency of ABE8e-SpRY by protospacer position and PAM sequence. **(D)** The scatterplots illustrate the comparison of nucleotide-level editing efficiency between the reporter and endogenous target sites across three experimental replicates for a total of 49 gRNAs spanning four loci. The top panel displays the accessibility of the four loci as measured by ATAC-seq signal in HepG2 cells, with scatterplot markers colored according to the accessibility of each nucleotide. Pearson correlation coefficients (r) are provided to indicate the strength of correlation between the editing efficiency of reporter and endogenous target sites. **(E)** The schematic depicts the design of the LDL-C variant library gRNA, targeting selected GWAS candidate variants. The Manhattan plot illustrates variant P-values obtained from a recent GWAS study (Klimentidis *et al.*, 2020). The gRNAs are strategically designed to tile the variant at five positions, maximizing editing efficiency at protospacer positions 4–8. **(F)** gRNA coverage of the LDLR tiling library across LDLR coding sequence along with 5' and 3' UTRs and several regulatory regions.

alleles within every sequence pair was crucial for assessing a variant's impact on TF binding. Higher binding scores for the alternative sequence indicated an increase in TF binding potential, while lower scores suggested a decrease. To quantify these changes, we computed disruption scores (DS) (**Section 6.1.1**). Recall, that DS captures the directional change induced by each variant, where a negative value

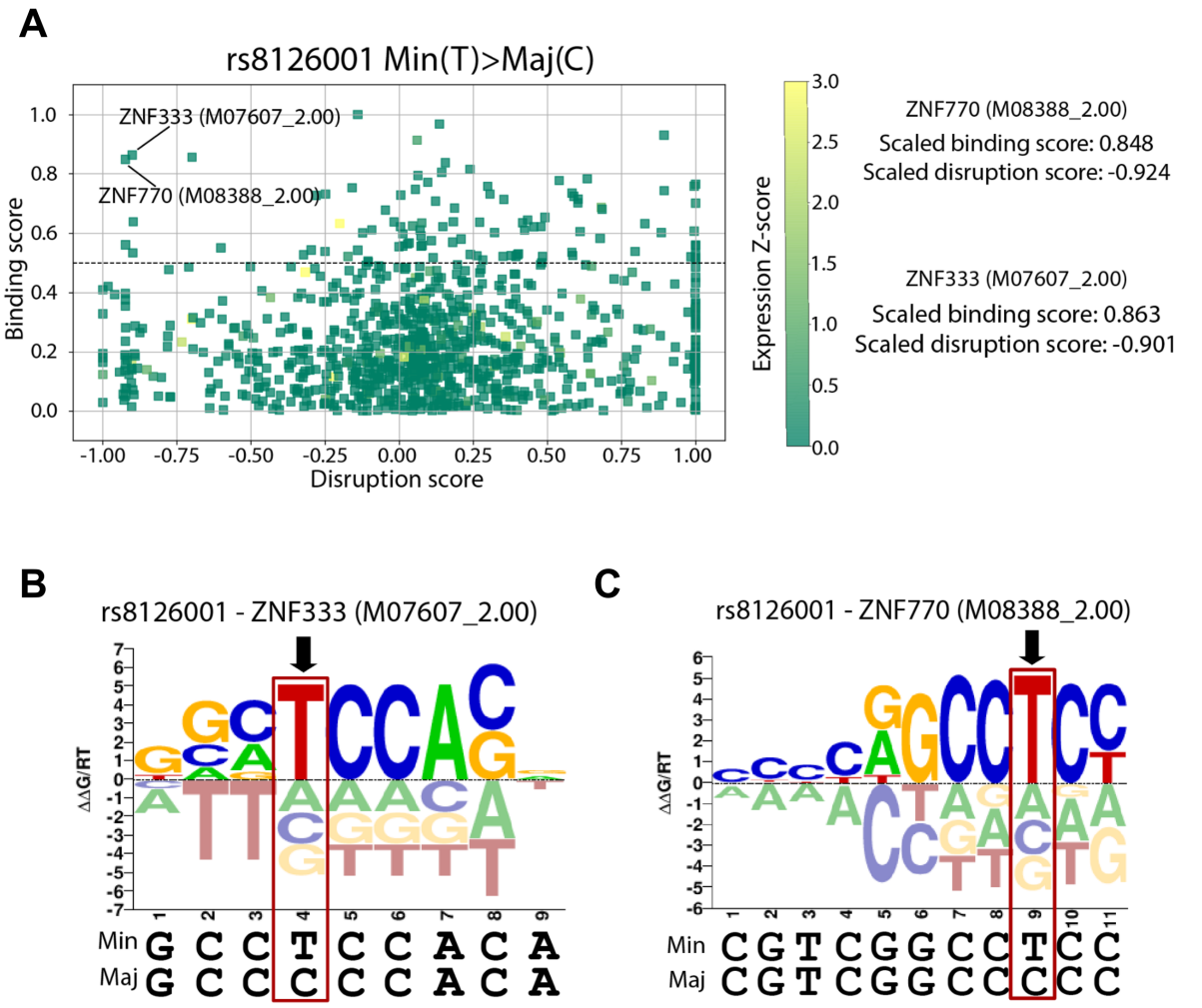


Figure 6.20. MotifRaptor analysis reveals insights into potential disruption of transcription factor binding by candidate variants. (A) The disruption is graphically represented against the binding scores of motifs, specifically examining the transition from the minor to major allele of rs8126001. **(B-C)** The identified motifs of ZNF333 and ZNF770 are showcased, aligned with the loci of rs8126001, featuring both minor and major alleles.

signifies reduced TF binding potential, and a positive value indicates an increase. For each variant, we ranked TFs exhibiting heightened binding potential and susceptibility to disruption by the variant. In the case of rs8126001, our methodology highlighted two zinc finger TFs, ZNF333 and ZNF770, showcasing enhanced binding site sequences attributed to the presence of the heterozygous minor allele in HepG2 cells (**Figure 6.20**). HepG2 ChIP-seq data provides supportive evidence for the binding of these TFs at the locus, even though the variant lies at the periphery of the peaks (**Figure 6.21**). While drawing definitive conclusions necessitates additional experimental validation, our findings are consistent with prior research (Farh *et al.*, 2015), which implies that a minority of causal variants directly modify canonical TF binding sequences. Instead, the majority tend to influence non-canonical sequences, emphasizing the intricacies involved in understanding the impact of genetic variants on transcription factor interactions.

6.3 Discussion and future directions

MotifRaptor is a computational toolkit designed to investigate the impact of genetic variants on TFBS within non-coding regions. MotifRaptor analysis encompasses three key steps that enable users to identify relevant cell types, cell type-specific genomic regions, and TFs, and to analyze and annotate SNPs based on their effects on TF binding landscape. Importantly, MotifRaptor is independent from ChIP-seq data, which may not be readily accessible for many cell type-TF combinations. Instead, MotifRaptor employs TF PWMs models, which are more widely available. Its efficient algorithmic design enables the computation of genome-wide null models for each TF, facilitating a comprehensive exploration of potential relationship between SNP-trait focusing on TF binding sites. Moreover, MotifRaptor incorporates gene

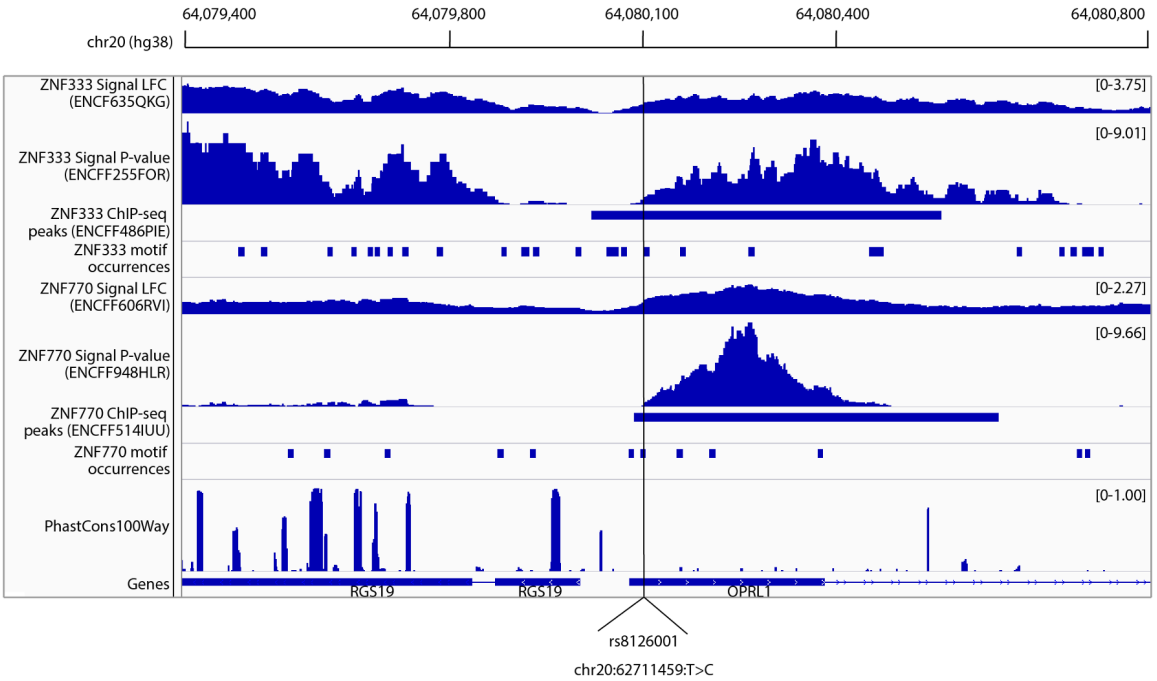


Figure 6.21. An overview of ChIP-seq signal Log2 Fold Change (LFC), signal *P*-values, peaks, and motif occurrences is provided for ZNF333 and ZNF770 in proximity to rs8126001. PhastCons100way conservation scores and gene annotations are co-displayed. ENCODE accession numbers are encapsulated in parentheses. The ENCODE signal LFC and *P*-values are derived through the ENCODE processing pipeline, utilizing MACS2 (Zhang *et al.*, 2008) for the generation of signal *P*-values.

expression data to filter out false positives, enhancing the accuracy of its predictions. Additionally, the MotifRaptor integrates established annotations to evaluate individual SNPs based on their conservation and potential deleterious effects, providing users with valuable insights into the functional significance of genomic variants. However, further extensions or downstream analyses are necessary to evaluate the significance of individual SNPs and their potential target genes. Additionally, it is important to note that MotifRaptor does not currently incorporate any preprocessing or filtering steps for the summary statistic files provided as input. Therefore, while it assumes the completeness and accuracy of these files, the biological insights and prioritized factors that can be derived may be influenced by the quality of the input data. MotifRaptor may be extended to consider other TFBS models, such as Support Vector Machine (SVM)-based or Deep Neural Network (DNN)-based. Since these models have been demonstrated to outperform PWMs in many predictive tasks (Tognon *et al.*, 2023), they may boost the accuracy of MotifRaptor's predictions. However, while several complete collections of PWMs are available, comprehensive libraries of SVM-based or DNN-based models are still not available to the community.

CRISPR genome editing

The discovery of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas system has revolutionized genetic engineering, offering a precise, versatile and programmable framework to modify the genomes of living organisms (Cong *et al.*, 2013). Originally identified in bacterial immune system (Garneau *et al.*, 2010), CRISPR-Cas system has become a transformative gene-editing tool with profound implications for molecular biology and the potential for innovative therapeutic interventions (Jinek *et al.*, 2012; Hsu *et al.*, 2014; Kantor *et al.*, 2020). CRISPR-Cas systems encompass an ensemble of diverse components each employing distinct mechanisms of action. Despite different Cas proteins (**Table 7.9**) may be employed depending on the editing goal (Makarova *et al.*, 2020), all systems rely on CRISPR RNA (crRNA) for precise target specificity. Notably, some variants introduce an additional requirement for trans-activating RNA (tracrRNA), forming a crucial scaffold structure (Kantor *et al.*, 2020). The aforementioned CRISPR RNAs are joined into a single small guide RNA (sgRNA) (Ran *et al.*, 2013). Using sgRNAs rather than crRNAs and tracrRNAs significantly simplifies the editing process design. Protospacer adjacent motif (PAM) sequences are other key components of most CRISPR-Cas systems. PAM sequences are short motifs adjacent to the target sequence, restricting Cas proteins action to locations presenting the target sequence followed by the appropriate PAM. The simplicity of the 5'-`NGG`-3' PAM sequence requirement has led to the predominant use of Streptococcus pyogenes' Cas9 (SpCas9) protein in different applications (**Figure 7.22 (A)**). However, spCas9's PAM sequence significantly restricts its editing landscape. Consequently, recent studies focused on proposing novel engineered Cas proteins enhancing Cas9 editing scope (Anders *et al.*, 2016; Miller *et al.*, 2020). Notably, the SpG and SpRY Cas9 variants significantly expanded the editing landscape (Walton *et al.*, 2020), targeting less stringent PAMs. SpG is designed to target an expanded set of `NGN` PAMs. SpRY is a nearly PAM-less variant (PAM `NRN`) enabling to edit regions previously considered "un-targetable". By delivering the Cas9 nuclease complex coupled with a synthetic sgRNA (Cas9:sgRNA) CRISPR provides a simple and programmable means to modify genomic sequences at specific locations (**Figure 7.22 (B)**). To perform its editing action, the Cas9:sgRNA complex engages in a randomized exploration of the cellular DNA, initially seeking the appropriate PAM. Upon locating the PAM sequence, Cas9 unwinds the DNA to facilitate the hybridization of the Cas9-associated sgRNA with the exposed DNA strand, known as the protospacer. When the DNA sequence aligns with the sgRNA target sequence, the catalytic domains of the endonuclease execute the cleave the DNA double strand, resulting in a double-strand break (Jinek *et al.*, 2012). Subsequently, the host cell undertakes the repair of this break through either non-homologous end joining (NHEJ) or homology-directed repair (HDR) mechanisms (**Figure 7.22 (C-D)**). Genome editing mediated by other Cas proteins works similarly. While NHEJ leads to insertions or deletions at the repair site, HDR can be employed to insert a defined DNA template, to precisely repair the edited DNA segment. However, HDR pathway exhibits limited efficiency and high rates of unintended indels mutations, potentially compromising the mutation repairing (Mao *et al.*, 2008; Song and Stieger, 2017). CRISPR-Cas-mediated single-base-pair editing (DNA base-editors) systems have been developed to address these limitations (Nishida *et al.*, 2016; Kantor *et al.*, 2020). There are two classes of base-editors (BEs): cytosine base-editors (CBEs) and adenine base-editors (ABEs). CBEs and ABEs can introduce all four nucleotide transitions. The introduction of prime-editors (PEs) broadened the spectrum of donor-free precise DNA editing. In fact, PEs not only cover all nucleotide transitions and transversions but also extend their capability to handle small insertion and deletion mutations (Anzalone *et al.*, 2019). CRISPR-Cas systems empower precise yet potentially random sequence editing, the experiment success relies on metculous design of guide sgRNAs to direct the Cas complex to the intended target sequences (*on-targets*). Achieving this precision is crucial to avoid unintended and potentially risky alterations in non-targeted sequences (*off-targets*) (Pattanayak *et al.*, 2013; Cho *et al.*, 2014). Additionally, genetic variants may introduce novel unexpected target sites and PAMs, potentially influencing experiment outcome (Scott and Zhang, 2017). As a result, while designing sgRNA designs it becomes imperative to consider genetic diversity,

Nuclease	PAM
spCas9	NGG
SpCas9-VQR	NGAG
spCas9-VRER	NGCG
spCas9-NG	NGN
spCas9-xCas9	NG/GAA/GAT
spCas9-Sc++	NNG
spCas9-SpG	NGN
spCas9-SpRY	NRN
SaCas9	NNGRRT
SaCas9-KKH	NNNRRT
AsCas12a	TTTV
AsCas12a-RR	TYCV
AsCas12a-RVR	TATV
Cas12j	TTN
Cas12e	TTCN
Un1Cas12f1	TTTN
CnCas12f1	CCN

Table 7.9. Cas nucleases and their PAM sequences. List of different Cas nucleases and their Protospacer Adjacent Motif (PAM) sequences. Underlined Cas proteins denote "near PAM-less" nucleases

especially in clinical applications, to prevent potential undesired and hazardous consequences within the host genome. In aiding the design of suitable sgRNAs, various computational approaches have emerged. Guide-design tools primarily focus on computationally predicting potential guide RNAs that can target a specified input sequence, such as a gene or regulatory element, given a PAM. In parallel, off-target nomination tools offer insight into the potential sequences edited using a specific sgRNA and PAM sequence within a host genome. These tools furnish a list of both the intended target sequence and potentially unintended edited sequences. Furthermore, additional tools have been developed to assist researchers in evaluating and quantifying the success of their CRISPR genome editing experiments by analyzing the observed editing events in their data. Altogether, CRISPR genome editing has the potential to revolutionize the precision medicine paradigm. Its ability to make targeted modifications to the genome presents a powerful framework for addressing diseases linked to individual-specific molecular reasons (Sgro and Blancafort, 2020; Ho *et al.*, 2020). Beyond disease correction, CRISPR-based interventions hold promise in advancing our understanding of intricate genetic and epigenetic mechanisms underlying various health conditions (Syding *et al.*, 2020). Moreover, the adaptability of CRISPR technologies allows for the exploration of therapeutic avenues in the context of genetic disorders, neurodegenerative diseases (Kolli *et al.*, 2018), and even certain types of cancer (Wang *et al.*, 2022b), paving the way for innovative and personalized treatment strategies. Additionally, the continuous refinement of CRISPR tools and the exploration of new delivery methods contribute to the ongoing evolution of this revolutionary technology, expanding its potential applications in diverse biomedical and therapeutic realms. The subsequent sections delve into the fundamental components of the Cas9 complex (with similar principles applicable to systems employing different Cas proteins), explore Base Editors (BEs) and Prime Editors (PEs) along with their distinctions from other CRISPR-Cas systems, address the off-target problem, and present an overview of computational methods available for the analysis of CRISPR genome editing experiment data.

7.1 Essential elements in CRISPR-Cas9 Genome Editing

SpCas9 popularity in several CRISPR genome editing applications is mainly attributed to the simplicity of its 5'-NGG-3' PAM sequence and its widespred availability (**Figure 7.22**). The precision of CRISPR-Cas9 editing systems hinges on two crucial factors (Ran *et al.*, 2013): (i) the target sequence, a 20 bp string within each CRISPR locus of the sgRNA array, which typically has multiple unique targets, and (ii) the PAM sequence, which Cas9 identifies to initiate editing. Once identified its target sequences, Cas9 nuclease opens both DNA strands to introduce novel and precise modifications. Cas9 operates primarily through two mechanisms: (i) knock-in and (ii) knock-out mutations. In knock-in, HDR employs DNA sequences that resemble the target to mend the Cas9-induced breaks in the genome, using exogenous DNA as a template for repair. Conversely, in knock-out, the mutations in the DNA introduced by Cas9 result in the repair of breaks via NHEJ. Repair through NHEJ often leads to random indels in the target sequence, which can disrupt, enhance, or alter the function of the targeted site.

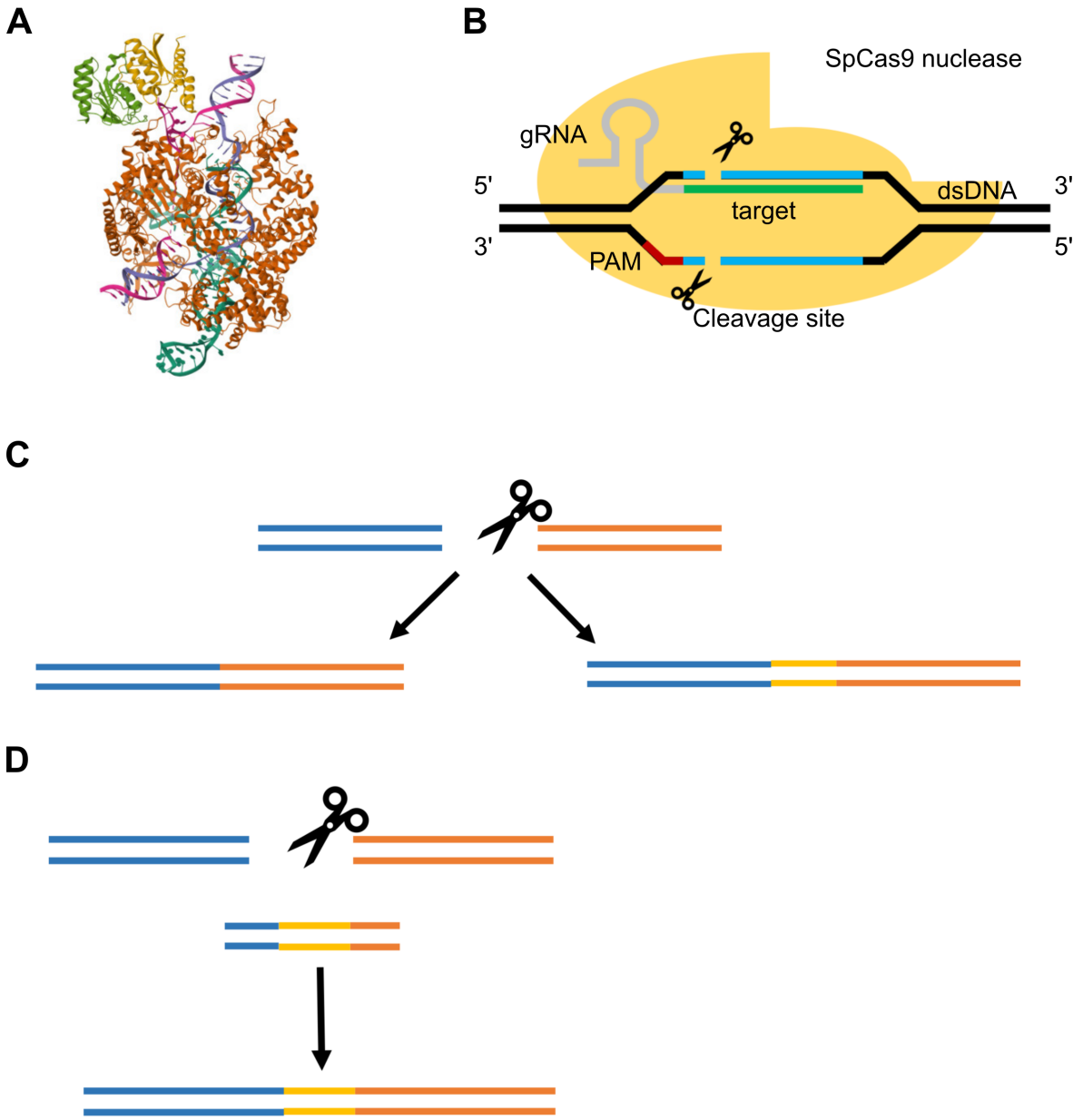


Figure 7.22. CRISPR-Cas9 genome editing. **(A)** X-Ray resolved structure of SpCas9 nuclease (obtained from the PDB database, accession number 6VPC (Lapinaite *et al.*, 2020)). **(B)** sgRNA:Cas9 complex. The sgRNA guides Cas9 to its target sequence. Once the sgRNA is paired with the target sequence, Cas9 opens the double-stranded DNA (dsDNA) near the Protospacer Adjacent Motif (PAM) sequence. After unwinding the DNA, Cas9 cleaves the dsDNA 3-4 bp upstream with respect to the PAM sequence, inducing a DNA breakage. The DNA breaks are repaired through either: **(C)** Non-homologous End Joining (NHEJ), which introduces insertions or deletions (indels) in the edited sequences, or **(D)** Homologous Directed Repair (HDR), which inserts a template DNA within the edited segment.

7.1.1 The Cas9 nuclease

Cas9 is a protein that holds a crucial role in the immune defense mechanisms of certain bacteria against DNA viruses and plasmids. Widely utilized in genetic engineering applications, its primary function revolves around cleaving DNA, enabling modifications to a cell's DNA sequence. To survive in diverse and challenging environments populated with bacteriophages, certain bacteria have evolved mechanisms to resist and repel predatory viruses. The CRISPR system serves as a fundamental element in these defense strategies. Functionally, CRISPR-Cas systems act as self-programmable restriction enzymes. CRISPR loci consist of short, palindromic repeats occurring at regular intervals, comprising alternating CRISPR repeats and variable CRISPR spacers ranging from 24 to 48 bp in length. Typically, CRISPR loci are accompanied by adjacent CRISPR-associated (cas) genes, which encode DNA restricting enzymes, such as SpCas9 in Streptococcus pyogenes (**Figure 7.22 (A)**). The discovery that spacer regions within CRISPR

loci exhibit homology to foreign DNA elements, including plasmids and viruses, marked the first biological evidence suggesting that CRISPRs might function as an immune system, providing bacteria with a versatile and adaptive defense mechanism against viral intruders. This discovery laid the foundation for applying the CRISPR-Cas system to edit genome sequences (Mali *et al.*, 2013). Cas9, a dual RNA-guided DNA endonuclease enzyme associated with the CRISPR adaptive immune system in Streptococcus pyogenes, plays a pivotal role in this system (**Figure 7.22 (B)**). In Streptococcus pyogenes, CRISPR memorizes foreign DNA, while Cas9 functions as an executor, interrogating and cleaving intruding DNA (Garneau *et al.*, 2010; Jinek *et al.*, 2012; Heler *et al.*, 2015). Cas9's interrogation involves unwinding foreign DNA and scrutinizing it for sites that align with the 20 bp spacer region of the guide RNA (gRNA). Upon identifying complementarity, Cas9 executes the cleavage of the invading DNA. Beyond its role in bacterial immunity, Cas9 has emerged as a robust genome engineering tool, inducing site-directed double-strand breaks in DNA. These breaks can lead to the inactivation of genomic functional elements or the introduction of foreign segments through non-homologous end joining (NHEJ) or homology-directed repair (HDR), respectively (**Figure 7.22 (C-D)**). Cas9's popularity has surged due to its capability to cleave nearly any sequence complementary to the guide RNA (Jinek *et al.*, 2012). Importantly, its target specificity relies on guide RNA:DNA complementarity rather than modifications to the protein itself. This unique feature makes Cas9 directly applicable to target DNA sequences (Mali *et al.*, 2013). The original bacterial Cas9 requires a guide RNA composed of two distinct RNAs: the CRISPR RNA (crRNA) and the trans-activating crRNA (tracrRNA) (Deltcheva *et al.*, 2011). To simplify Cas9 DNA targeting, chimeric single guide RNAs (sgRNAs) have been engineered. These advancements underscore the versatility and potential of Cas9 as a powerful tool in genetic engineering and genome editing (Ran *et al.*, 2013).

7.1.2 sgRNA: guiding Cas9 to target sequences

Guide RNAs (sgRNAs) are RNA segments that serves as a guiding tool for enzymes, such as Cas nucleases, targeting RNA or DNA sequences (Ran *et al.*, 2013). By forming complexes with Cas enzymes, sgRNAs guide the sgRNA:Cas complex to the intended target sequences to edit (**Figure 7.22 (B)**). While guide RNAs have natural occurrences and play crucial roles in various biological functions, sgRNAs are synthetically designed in genome editing applications to achieve precise and targeted modifications. sgRNAs achieve targeting through straightforward Watson-Crick base pairing with complementary sequences. Through complementary base-pairing sgRNAs direct Cas enzymes to the specific target region. sgRNAs are synthetically designed by joining two RNAs: (i) the trans-activating RNA (tracrRNA), and (ii) the CRISPR RNA (crRNA) (Deltcheva *et al.*, 2011). The tracrRNA facilitates Cas9 endonuclease activity, while the crRNA binds to the target-specific DNA segments. tracrRNA possesses a stem-loop structure, serving as the base-pairing partner for the endonuclease enzyme. crRNA contains a spacer-flanked region resulting from repeat sequences, comprising 18-20 bp. crRNA recognizes the specific complementary target region, initiating cleavage by Cas9 upon binding with the crRNA and tracrRNA effector complex. Modifications in the crRNA sequences enable changes in the target regions, rendering it a user-defined programmable tool. sgRNAs design is crucial for the precision of CRISPR-mediated genome editing experiments. For instance, in CRISPR-Cas9 regions targeting is dictated by the 20 bp sequence at the 5' end of sgRNA. However, to ensure cleavage the target sequence must be followed by the Protospacer Adjacent Motif (PAM). PAMs are short DNA segments located immediately after the target region. PAMs are crucial for Cas nucleases to cleave targets and are generally located 3-4 bp downstream from the cleavage site.

7.1.3 Protospacer Adjacent Motifs

Protospacer Adjacent Motifs (PAMs) (**Figure 7.22 (B)**) constitute a 2–6 bp DNA sequence positioned immediately following the sequence targeted by Cas nucleases (Shah *et al.*, 2013). Notably, PAMs are absent in the bacterial host genome, despite their integration into the invading virus or plasmid. The effective binding and cleavage of the target DNA sequence by Cas proteins hinge on the presence of the PAM sequence following it (Jinek *et al.*, 2012; Shah *et al.*, 2013). In fact, Cas9 refrains from cleaving the target sequence unless there is an immediately adjacent PAM sequence. Serving as an indispensable targeting component, PAM plays a pivotal role in discerning bacterial self from non-self DNA, thereby shielding the CRISPR locus from being targeted and disrupted by the CRISPR-associated nuclease. In CRISPR-Cas9 genome editing applications, sgRNAs recognize target sequences with a `NGG` PAM sequence at the 3'-end, guiding the nuclease to a specific sequence for cleavage (Cong *et al.*, 2013). After sgRNA-target pairing, Cas9 starts breaking the DNA double strand 3-4 bp upstream the PAM location. Crucially,

most Cas proteins start editing exclusively at sites where they recognize a PAM. Numerous endeavors have been undertaken to engineer SpCas9 and other Cas nucleases capable of recognizing various PAMs, with the goal of enhancing the versatility of CRISPR genome editing at specified locations within the genome (Kleinstiver *et al.*, 2015) (**Table 7.9**).

7.2 Precision Genome Editing: Maintaining DNA Double Strand Structure with Base-Editing and Prime-Editing

While Homologous Directed Repair (HDR) can insert DNA templates within edited sites to accurately restore genome sequences, this pathway is hindered by restricted efficiency and elevated rates of undesirable indel mutations. These mutations often offset the potential advantages gained from the mutation repair process (Kantor *et al.*, 2020). Additionally, HDR-mediated editing is constrained to cell types that are actively dividing, as it heavily relies on homologous recombination processes (Bollen *et al.*, 2018). The introduction of base-editing (BE) and prime-editing (PE) systems addressed these issues, introducing frameworks to precisely edit the genome sequence without breaking double strand DNA (Komor *et al.*, 2017; Anzalone *et al.*, 2019). Importantly, BEs and PEs have potential to rectify disease-causing mutations within the human genome. Targeting the four transition mutations enables the correction of over 25% of human pathogenic SNPs. By inserting or deleting short DNA segments beside targeting the four nucleotide transitions, PEs may address up to 89% of known genetic variants linked to human diseases (Komor *et al.*, 2017; Anzalone *et al.*, 2019; Kantor *et al.*, 2020). Therefore, BEs and PEs hold great promises in expanding the therapeutical significance of genome editing technologies. Notably, BEs and PEs may become fundamental tools in precision medicine settings and bring closer the realization of the paradigm (Porto *et al.*, 2020). The next sections describe the main concepts behind BE and PE.

7.2.1 Base-editing

BE systems present similar components to CRISPR-Cas. However, beside encompassing a Cas enzyme, BEs include a single-stranded DNA modifying enzyme to precisely modify nucleotides. BEs are divided in two distinct classes based on the single-stranded DNA modifying enzyme: cytosine base-editors (CBEs) and adenine base-editors (ABEs). Collectively, CBEs and ABEs can perform all four nucleotide transitions (`C` $\rightarrow$ `T`, `T` $\rightarrow$ `C`, `A` $\rightarrow$ `G`, `G` $\rightarrow$ `A`). Furthermore, the introduction of dual base-editor systems designed for combinatorial editing within human cells, expanded BEs' scope incorporating transversions (Kurt *et al.*, 2021; Grünewald *et al.*, 2020; Sakata *et al.*, 2020; Zhang *et al.*, 2020). Such advances may facilitate targeting more intricate compound edits significantly expanding DNA BE landscape.

Cytosine base-editors

Cytosine base-editors (CBEs) deaminate cytosine to uracil (`C` $\rightarrow$ `U`), which is recognized as a thymine during cell replication. This process results in `C` $\rightarrow$ `G` and `T` $\rightarrow$ `A` transitions fixed in the DNA sequence after cell replication (Komor *et al.*, 2016, 2017). The introduction of the engineered nCas9 protein dramatically improved CBE efficiency (Komor *et al.*, 2016). This Cas9 variant induces a nick in the strand containing the `G` in the `U-G` intermediate. The nick on the non-edited DNA strand favors cellular repair towards a `U-A` outcome. The latter is subsequently transformed into `T-A` during DNA replication. Employing nCas9 resulted in a six-fold increase in editing efficiency. Despite nCas9 exhibited 1.1% indel rate, it remains significantly lower compared to the rate induced by editing systems using HDR and breaking double strand DNA (Komor *et al.*, 2016). Further improvements have been introduced to reduce indel rate (Kim *et al.*, 2017; Kantor *et al.*, 2020). In parallel, researchers focused on broadening the range of targetable bases, by developing BEs that incorporate different CRISPR nucleases. In fact, CBEs using SpCas9 face limitations due to their dependence on its `G`/`C`-rich PAM. To address the challenge a range of engineered Cas9 variants with modified PAM sequences and enhanced cleavage specificity has been developed. These variants hold the potential to further expand the targeting scope (Kim *et al.*, 2020; Chatterjee *et al.*, 2020; Miller *et al.*, 2020; Zetsche *et al.*, 2015).

Adenine base-editors

Since their editing capabilities are constrained to install `C-G` to `T-A` mutation, CBEs present a limited spectrum of correction of disease-linked variants. Moreover, methylated cytosines are prone to spontaneous deamination (Alsøe *et al.*, 2017), potentially introducing pathogenic variants. Adenine base-editors

(ABEs) address these limitations by converting A-T to G-C. Importantly, this enable to potentially reverse ∼50% of pathogenic SNPs. This development significantly broadened the scope of BEs applications to treatment of diseases with genetic and molecular bases. ABEs employ an ABE-dCas9 complex binding to a target DNA sequence guided by gRNA. The complex's deaminase domain catalyzes the transition of adenine to inosine, that is interpreted as guanine during DNA replication (Gaudelli *et al.*, 2017). This facilitates the replacement of the original A-T with G-C at the targeted site. Generally, ABEs editing efficiency is high introducing almost no indels. While they worked well when paired with SpCas9, ABEs exhibited limited compatibility with Cas9 variants. ABE8e (Gaudelli *et al.*, 2017; Richter *et al.*, 2020) solved this issue, providing a base-editor with enhanced efficiency and compatibility with Cas9 variants other than SpCas9. This significantly expanded the editing landscape of ABEs, enabling the editing of a wide range of genomic regions.

7.2.2 Prime-editing

Originally, CBEs and ABEs limited their editing capability to four transition mutations. Prime-editing (PE) (Anzalone *et al.*, 2019) addressed the issue proposing a BE method operating without inducing double-strand breaks in the DNA, and employing an engineered reverse transcriptase (RT). The RT is fused to Cas9 nickase (nCas9) and a prime-editing guide RNA (pegRNA) (Anzalone *et al.*, 2019). pegRNAs have two major functions purpose: (i) directing nCas9 to the target site, and (ii) harboring an additional sequence specifying the desired sequence edits (Anzalone *et al.*, 2019). Briefly, the 5 end of the pegRNA binds to the primer binding site (PBS) region on the DNA. This bind induces the exposure of the non-complementary strand. Subsequently, Cas9 nicks the unbound DNA strand containing the PAM, creating a primer for the RT to start its retrotranscription. While elongating the nicked strand, RT uses the interior of the pegRNA as a template, editing the target region. This operation produces two redundant PAM DNA flaps: the edited 3 and unedited 5' flaps. To include the edit within the DNA, PE exploits cellular endonucleases actions. In fact, The cell endonucleases are abundant during lagging-strand DNA synthesis, and preferentially degrade the 5 flaps (Kantor *et al.*, 2020). The resulting heteroduplex containing the unedited and the edited strands is resolved and integrated into the host genome through cellular replication and repair processes.

7.3 Editing unwanted regions: addressing the Off-Targets issue and understanding the influence of Genetic Diversity

CRISPR genome editing offers signifcant opportunities for developing innovative therapeutics by precisely modifying genetic or epigenetic elements, such as regulatory elements, within specific genomic regions. The design of gRNAs to precisely target the intended sequences (on-targets) is of paramount importance. gRNAs precision is essential to prevent unintended and potentially deleterious edits in non-targeted sequences, called off-targets (Pattanayak *et al.*, 2013; Cho *et al.*, 2014). Off-target edits often occur at sites exhibiting sequence homology to the intended on-target. However, the precise rules governing off-target editing remain unclear (Tycko *et al.*, 2016). The presence of mismatches in sequences may result in either a decrease in cleavage activity or may have no discernible effect, contingent upon the specific nature of the base alteration and its relative position (Doench *et al.*, 2014, 2016). Furthermore, while off-target editing generally decreases with the number of mismatches (Cho *et al.*, 2014), several studies demonstrated that off-targets mutations occur in sites displaying up to 6 mismatches with respect to the on-target (Tsai *et al.*, 2015). Off-target prediction is a fundamental practice guiding the development and design of CRISPR experiment. Beyond the mere count of mismatches or bulges, various sequence characteristics, such as the position of the mismatch or bulge relative to the PAM, or specific base alterations, contribute to the likelihood of off-target effects. Notably, genetic diversity significantly influences the potential creation of new off-target sites (Scott and Zhang, 2017). Indeed, genetic variants can enhance the affinity of certain sequences previously unbound by the sgRNA. Furthermore, they may generate novel PAM sequences due to their simple structure. PAMs occurring in novel positions may induce Cas proteins to break double stranded DNA in previously uncut positions, with potential hazardous effects on the cell environment. Several methods are currently available to assess on- and off-target genome editing at individual loci (Clement *et al.*, 2020). These approaches encompass both sequencing and non-sequencing-based methods. While non-sequencing-based methods, generally, provide cost-effective solutions to detect sequence edits, sequecing-based technologies precisely quantify editing frequencies and define the mutation alleles induced by genome editing (Clement *et al.*, 2020). Among these methods, GUIDE-seq (Tsai *et al.*,

2015) and Circle-seq (Tsai *et al.*, 2017) have been succesfully employed to experimentally identify off-target edits *in cellula* and *in vivo*, respectively.

7.4 The computational aspects of CRISPR Genome Editing experiments

Designing effective gRNAs is crucial for precisely targeting specific sequences within the genome while minimizing off-target effects. Several computational tools are to design guides (guide-designer). Guide-designer provide a list of potential guides targeting a region. However, these tools must account for potential off-target sites during gRNA design to mitigate unintended genetic alterations. Off-target nomination tools predict and analyze off-target effects of input gRNAs. They provide a comprehensive landscape of the potential experimental outcomes. CRISPR edits can be quantified using several tailored computational methods, that measure editing events from sequencing data. Although, quantification methods detect and quantify editing events in common CRISPR genome editing settings, they fall short when measuring the phenotypic outcomes in more compelx experiments, like CRISPR screens. CRISPR screens evaluate the impact of genetic variants on phenotypic traits, providing insights into the functionality of specific genomic elements and their implications for disease and therapeutics, such as cancer.

7.4.1 Designing guide RNAs

Designing guide RNAs (gRNAs) to precisely target specific sequences within the genome is a critical aspect of planning CRISPR genome editing experiments. These gRNAs not only need to efficiently target the intended target sequence but also must minimize the risk of binding to unintended sites, thereby reducing the likelihood of off-target effects. To assist researchers in this endeavor, a variety of computational tools have emerged to facilitate the design of gRNAs with optimal characteristics (Li *et al.*, 2023). These tools, available both online and offline, leverage sophisticated algorithms to predict gRNA efficacy and specificity. Among the available tools, web-based platforms like CRISPOR (Haeussler *et al.*, 2016), CRISPR-SURF (Hsu *et al.*, 2018), and CHOPCHOP (Montague *et al.*, 2014) offer user-friendly interfaces for designing gRNAs. Additionally, specialized tools such as Prime-design (Hsu *et al.*, 2021) cater to specific applications, such as designing pegRNAs used in PE experiments. A crucial aspect of these gRNA design tools is their ability to assess the off-target potential associated with the predicted guides. Once gRNAs are designed, their efficacy and specificity are typically evaluated using off-target prediction tools. These tools aim to detect potential off-target sites that may be edited using the predicted gRNAs, providing researchers with valuable insights into the potential risks associated with their experimental designs.

7.4.2 Nominating potential CRISPR off-targets

Off-target genomic cleavage, occurring at sequences sharing high similarity with the intended on-target site, poses a significant challenge in CRISPR genome editing (Pattanayak *et al.*, 2013; Cho *et al.*, 2014). Computational methods nominating off-target sites during guide RNA design have been pursued extensively to address this issue (Hanna and Doench, 2020). Various methods have been developed to predict off-target sites based on factors like sequence homology, mismatches, RNA/DNA bulges, and the effects of mismatches on cleavage activities (Doench *et al.*, 2014, 2016; Sanson *et al.*, 2018). Traditional sequence aligners offer a rapid way to identify putative off-target sites by aligning the target sequence to the reference genome and reporting loci with specified mismatches. Notable examples are CRISPR-P (Lei *et al.*, 2014), Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), CHOPCHOP (Montague *et al.*, 2014), GT-scan (O'Brien and Bailey, 2014), CRISPOR (Haeussler *et al.*, 2016). However, their effectiveness is limited as they may miss some potential off-target sites. Newer search algorithms have been devised to overcome these limitations, incorporating considerations like PAM requirements (Xiao *et al.*, 2014; Zhu *et al.*, 2014), RNA and/or DNA bulges (Bae *et al.*, 2014; Cancellieri *et al.*, 2020), and enzymatic binding and cleaving processes (Klein *et al.*, 2018). Despite off-target editing generally decreasing with an increasing number of mismatches, experiments have shown mutations occurring at sites with numerous mismatches, highlighting the complexity of off-target effects. Therefore, determining the optimal number of tolerated mismatches during *in silico* off-target site nomination remains a challenge (Clement *et al.*, 2020). Insights from experimental datasets have been leveraged to train computational models using various techniques such as logistic (Allen *et al.*, 2019), regression (Listgarten *et al.*, 2018), or

random forest regression (CRISTA) (Abadi *et al.*, 2017) models. However, generating sufficient data for model training remains a hurdle. Additionally, these models were trained on CRISPR-CAs9 derived data and are not directly applicable on other Cas proteins or on base editors. Together, off-target nomination tools provides a comprehensive *in silico* prediction of genome editing potential outcomes.

7.4.3 Quantifying CRISPR Genome Editing experiments outcome

Genome editing outcomes can be quantified through different sequencing methods. Sanger sequencing trace decomposition tools like TIDE and ICE offer quick and cost-effective quantification of editing and allelic frequencies, particularly useful for screening edited cell clones (Brinkman *et al.*, 2014; Conant *et al.*, 2022). Alternatively, PCR amplicons can be cloned into plasmids and sequenced via Sanger methods to identify specific alleles (Canver *et al.*, 2014). Next-generation sequencing (NGS) is considered the gold standard for determining editing frequency and characterizing resulting alleles. Compared to traditional methods, NGS provides more accurate results, although it can be cost-prohibitive and time-consuming. However, as sequencing costs decrease, NGS analysis becomes more accessible, enabling the development of multiplexed readouts and more sensitive assays for detecting rare alleles. Numerous computational tools have been developed for analyzing NGS data from CRISPR-Cas nucleases (Park *et al.*, 2017; Pinello *et al.*, 2016) and base editors (Clement *et al.*, 2019; Hwang *et al.*, 2018). These tools employ various alignment and analysis techniques to distinguish true genome edits from sequencing errors. For instance, the "editing window" approach focuses on mutations overlapping the predicted target site, reducing false-positive results (Clement *et al.*, 2019). Such methods enhance the accuracy of genome editing analysis and pave the way for more reliable experimental outcomes. Despite quantification tools may be used to quantify the phenotypic outcomes of genome wide CRISPR experiments, they fall short when quantifying the impact on extended pooled screens.

7.4.4 Quantifying gene essentiality from genome-wide pooled CRISPR screens

Perturbing gene activity and evaluating the resulting phenotype is a fundamental approach for identifying and understanding a gene's biological functions. Historically, the ability to induce complete gene knockouts (KOs) on a genomic scale was confined to model organisms like yeast. In higher eukaryotes, including human cell lines, researchers relied on RNA interference (RNAi) or gene trapping methods in haploid human cells (Carette *et al.*, 2011). RNAi uses the endogenous RNA-induced silencing complex (RISC) machinery to target messenger RNA transcripts, whose abundance can vary significantly. This variability often leads to incomplete target knockdown and off-target effects of varying severity, resulting in diluted data (Kaelin Jr, 2012; Hart *et al.*, 2014). The adaptation of the CRISPR-Cas system to pooled library gene KO screens in mammalian cells has revolutionized the identification of essential genes, enabling high-throughput functional analysis of the genome (Koike-Yusa *et al.*, 2014; Shalem *et al.*, 2014; Meyers *et al.*, 2017). In CRISPR pooled library screens, multiple gRNAs target each gene, with each cell affected by a single gRNA clone, while each guide targets numerous cells. Cells without any perturbation, or those with KOs showing no growth phenotype, exhibit normal growth rates, akin to wildtype cells. Conversely, cells with a gRNA targeting a gene critical for fitness demonstrate reduced growth rates. To pinpoint genes responsible for fitness defects upon KO, the frequency distribution of gRNAs in the population is analyzed via deep sequencing and compared to an early control timepoint. Changes in the gRNA frequency distribution are quantified as log fold changes, with severe negative values indicating gRNAs associated with significant fitness defects. However, accurately estimating the gene-level effect by aggregating individual reagents poses a major challenge in the analysis of CRISPR pooled screens. BAGEL (Hart and Moffat, 2016) introduced a method employing a Bayesian classifier, utilizing training sets comprising both known essential and non-essential genes to construct models to predict the expected fold change distributions for each gene category. When assessing the effects on uncharacterized genes, BAGEL consolidates experimental data from reagents targeting the gene of interest, such as gRNAs or short hairpin RNAs, and evaluates the probability of these observations aligning with the essential or non-essential training sets. However, the accuracy of cell proliferation measurements in CRISPR–Cas9 loss-of-function screens can be compromised by genomic copy number variations (Wang *et al.*, 2015; Aguirre *et al.*, 2016). This influence is particularly pronounced when the gRNA–Cas9 complex targets regions with copy number gain, triggering a DNA-damage response. This sequence-independent effect complicates the assessment of gene deletion consequences on cell viability, even with low-level copy number alterations, posing challenges in cancer cell lines with numerous genomic amplifications. Current methods for addressing this issue involve filtering schemes that exclude data from amplified regions, potentially overlooking low-level alterations (Wang *et al.*, 2017). CERES (Meyers *et al.*, 2017) introduced

an approach for estimating gene dependency from essentiality screens while computationally correcting for copy number effects. CERES employs a nonlinear model to address bias stemming from DNA cutting toxicity by estimating the fitness impacts caused by multiple DNA cuts. CRISPRCleanR (Iorio *et al.*, 2018) offers an unsupervised method for genome-wide screens, organizing reagents based on their targeted genomic locations. Regions showing gRNA enrichment or depletion are adjusted to the global mean under the assumption that they reflect a copy number alteration. Chronos (Dempster *et al.*, 2021) addresses challenges related to gRNA efficacy, such as variable screen quality, cell growth rates, and diverse DNA cutting outcomes by incorporating a mechanistic model of the experiment. Notably, Chronos directly analyzes read-count level data using a robust negative binomial noise model (Anders and Huber, 2010), rather than modeling log-fold change values with a Gaussian distribution as other tools do. Additionally, Chronos employs an enhanced model to eliminate biases associated with copy number variations, accounting for multiple sources of bias.

7.4.5 Quantifying variants impact on phenotypic traits through CRISPR screens

CRISPR screens are also valuable for evaluating the impact of genetic variants on phenotypic traits and disease susceptibility at genome-wide scale (Doench, 2018). They have been instrumental in prioritizing potential cancer therapeutic targets on a genome-wide scale (Behan *et al.*, 2019). This approach has led to the establishment of consortia like the Cancer Dependency Map (DepMap) (Tsherniak *et al.*, 2017), which enables the research community to investigate cancer vulnerabilities by providing open access to key cancer dependencies, along with analytical and visualization tools. DepMap utilizes large-scale functional genomics profiling to identify genes essential for cell growth, conducting genome-wide RNAi and CRISPR loss-of-function screens in over 1,000 cancer cell lines. These extensive efforts are crucial for uncovering novel cancer targets and advancing precision medicine strategies. Sequencing of gRNAs for analysis often introduces overdispersion due to multiple PCR rounds in library construction, impacting accuracy (Clement *et al.*, 2020). To address this, analysis tools incorporate distributions like negative binomial or beta binomial for P-value calculation, while permutation-based non-parametric analysis avoids distribution assumptions (Spahn *et al.*, 2017; Jeong *et al.*, 2019; Jia *et al.*, 2017). The P-value indicates the impact of a gRNA on the observed phenotypes. For pooled screens with multiple gRNAs targeting each gene, gene-level statistics are derived by aggregating individual P-values, considering factors like allelic diversity and cleavage efficiency (Li *et al.*, 2014, 2015; Yu *et al.*, 2016). Tools like MAGeCK and hierarchical mixture-based models adjust for variable gRNA efficiencies and offer quality control and visualization capabilities (Li *et al.*, 2015; Wang *et al.*, 2019). In single-cell analysis, scRNA-seq readouts help measure perturbation effects on gene expression, with methods like scMAGeCK and MUSIC modeling gRNA effects using linear models or topic modeling (Yang *et al.*, 2020; Duan *et al.*, 2019). CRISPR base editing screens leverage base editors (BEs) for targeted introduction of transition variants, valuable for investigating coding variant effects and GWAS-associated variants. Notably, BE screens have been employed to dissect coding variant effects and to evaluate GWAS-associated variant functions (Kweon *et al.*, 2020; Després *et al.*, 2020; Hanna *et al.*, 2021; Cuella-Martin *et al.*, 2021; Cheng *et al.*, 2021; Huang *et al.*, 2021a; Sánchez-Rivera *et al.*, 2022; Kim *et al.*, 2022; Sangree *et al.*, 2022; Morris *et al.*, 2023; Martin-Rufino *et al.*, 2023; Pablo *et al.*, 2023; Coelho *et al.*, 2023; Lue *et al.*, 2023; Garcia *et al.*, 2023). BEAN (Ryu *et al.*, 2024) proposes using genotypic outcome data to normalize phenotypic scores of target variants and can analyze densely tiled coding sequence data, enhancing accuracy in evaluating phenotypic scores for each coding variant.

Assessing CRISPR genome editing outcomes: A comparative analysis of advanced quantification methods using high-depth Whole Genome Sequencing

CRISPR genome editing (Cong *et al.*, 2013) holds significant promise for various clinical applications, particularly in the realm of precision or personalized medicine. However, one of the major challenges hindering the clinical adoption of CRISPR genome editing technologies lies in the risk associated with unintended genetic modifications at unpredicted genomic loci (Fu *et al.*, 2013). To address this concern, both computational and experimental approaches have been devised to identify potential off-target editing sites (Clement *et al.*, 2020). These strategies typically generate a ranked list of candidate sites, with those ranking highest being prioritized for further validation using amplicon sequencing to assess the extent of genome editing (Akcakaya *et al.*, 2018). However, such nomination strategies often lack a clear cutoff point to determine which sites warrant experimental validation. As a result, the decision on the number of sites selected for validation may be subjective, influenced by factors such as budget constraints or arbitrary thresholds (e.g., selecting the top 100 sites). This approach leaves open the possibility of editing occurring at lower-ranked sites, while also raising concerns about the nomination strategies' ability to comprehensively capture the dynamics of CRISPR genome editing. Consequently, relying solely on nomination strategies may not be sufficient to mitigate the risk of unintended modifications at unpredicted genomic locations. Whole-genome sequencing (WGS) has emerged as a standard method to comprehensively analyze the genome to identify editing evidences. WGS offers numerous advantages, including the potential to assess editing activity at every nucleotide position, as well as the potential to detect complex editing outcomes such as translocations (Yin *et al.*, 2022). However, despite the increasing accessibility and declining costs of sequencing technologies, in the context of editing detection, WGS still presents challenges due to its significant expense. In fact, extensive sequencing coverage across the entire genome is essential to detect rare editing events. Moreover, distinguishing rare editing events from sequencing errors poses additional complexities. WGS has already been used to identify genome editing events in clonal cell lines (Smith *et al.*, 2014; Veres *et al.*, 2014). This approach leverages individual cells within a clone as replicates to mitigate the bias introduced by sequencing artifacts. However, while effective for detecting editing events that occur with relatively high frequency, this method is less suitable for identifying rare off-target events. The sensitivity of detection is constrained by the number of clones analyzed, with a larger number of clones required to reliably detect rare events (e.g., performing WGS on 100 clonal lines to detect an editing event occurring at a 1% rate). In this chapter, we explore the effectiveness of whole-genome sequencing (WGS) for detecting genome editing events across entire cell populations. By generating ultra-deep 1000x WGS datasets from genome-edited samples, we thoroughly analyze editing occurrences at both intended and unintended target sites. Additionally, we compare the detection efficiency of WGS with various experimental and computational methods for nomination and identification of editing events. This comprehensive approach, integrating models, datasets, and analytical techniques, offers valuable insights into the utility of WGS for detecting genome editing events.

Guide	Specificity	Maximum Mismatches	Potential off-targets
RNF2	High	4	7
EMX1	Moderate	4	293
HEKSite4	Moderate	4	832

Table 8.10. Editing gRNAs selected to benchmark editing detection rate. The table displays the three chosen guides (RNF2, EMX1, and HEKSite4) used to assess the detection rate of editing events across various computational methods. These three gRNAs were selected based on their effectiveness in editing when paired with the Cas9 protein, each exhibiting different levels of specificity. Our study examined potential off-target sites for each gRNA, considering matches of up to 4 mismatches. Note that DNA/RNA bulges were not taken into consideration in our analysis.

8.1 Establishing sequencing depth requirements to detect genome editing events

The genome editing detection problem can be dissected into two primary components: (i) sequencing genomic DNA with adequate depth to identify editing events at specific loci, and (ii), accurately interpreting the sequencing data to differentiate true CRISPR edits from technical artifacts. Addressing the first aspect, we propose a binomial model to assess sequencing depth requirements, and use simulated samples to evaluate the efficacy of existing mutation callers in identifying edited reads. Sequencing genomic DNA to detect editing events can be conducted in a heterogeneous population of edited cells, where the detection limit and power are directly proportional to the sequencing depth. This principle applies not only to amplicon sequencing targeting a single locus but also to WGS. Notably, while WGS can be conducted at low coverage in clonal cell lines to identify genome editing in individual cells, the limit of detection in this scenario is determined by the number of single-cell clones that can be sequenced. To detect genome editing events is essential to sequence samples at sufficient depth. Deeper sequencing is necessary to detect rarer editing occurrences. To formalize the ability to detect editing events across various frequencies, we employed a statistical model based on the binomial distribution as a framework (Petrackova *et al.*, 2019). With this model, we can determine the sequencing depth required to observe a specific number of reads, given a certain frequency of edited reads in a sample. For instance, to detect a single edited read in a sample with a mutation rate of 10%, 30 reads are required, while 300 reads are needed to detect a single edited read in a sample with a mutation rate of 1%, with a power of 95%. Additionally, the binomial distribution confidence interval may be used to estimate the range of true editing rates based on an observed editing rate and sequencing depth. For example, the 95% confidence interval for an edit observed in 15 out of 30 reads ranges from 31.3% to 68.7%.

8.2 Generating WGS datasets

Whole-genome sequencing is a fundamental tool to comprehensively identify mutations throughout the genome. Notably, WGS has found application in clinical settings for detecting disease-associated germline mutations (Thiffault *et al.*, 2019). Clinical WGS protocols typically target sequencing depths of 30x-50x coverage, with 40x coverage being recognized as adequate for detecting both heterozygous and homozygous single nucleotide polymorphisms (SNPs) (Sun *et al.*, 2021). According to our models (**Section 8.1**), sequencing at 30x coverage can identify a single edited read amidst approximately 5% genome editing at 80% confidence. However, for detecting rare editing events, deeper sequencing is required. Our goal is to construct WGS libraries at a depth of 1000x, enabling the detection of a single edited read with a 0.2% editing rate and 80% power. In our study, we edited K562 and GM12878-Cas9 cell lines (Ma *et al.*, 2017) using three guides exhibiting varying specificities, as determined by the number of sites in the genome with sequence homology up to 4 mismatches (**Table 8.10**). Specifically, we selected guides targeting RNF2 (high specificity, 7 off-by-4 sites), EMX1 (moderate specificity, 293 off-by-4 sites), HEKSite4 (moderate specificity, 832 off-by-4 sites). Additionally, a Cas-12a guide targeting DNMT1Site3 was utilized as a control guide. For each sample, PCR-free sequencing libraries were prepared. Genomic sequencing depth was assessed using Mosdepth (Pedersen and Quinlan, 2018), revealing a minimum of 89% genomic coverage at 1000x in GM12878 and at least 57% genomic coverage at 1000x in K562.

8.3 Measuring genome editing detection rates using existing tools

We leveraged our WGS datasets to evaluate the efficacy of established tools commonly utilized for variant detection in non-CRISPR contexts: Mutect2 (McKenna *et al.*, 2010), Varscan2 (Koboldt *et al.*, 2012), and Strelka2 (Kim *et al.*, 2018). While these tools have been developed to ignore technical artifacts, such as sequencing errors, and to identify rare alleles, none is specifically designed to detect CRISPR-edited reads or genomic sites affected by CRISPR editing. Nevertheless, they are frequently employed to identify rare editing events. We therefore tested the ability of these tools to detect CRISPR editing at and around sites nominated by experimental (GUIDE-seq (Tsai *et al.*, 2015)), and computational (Casoffinder (Bae *et al.*, 2014)) methods.

8.3.1 Detecting genome editing events in GUIDE-seq sites

For each site identified by GUIDE-seq, we defined 20 Kb windows to analyze the occurrence and characteristics of events detected by Mutect2, Strelka2, and Varscan2 in both K562 and GM12878-Cas9 cell lines. Each GUIDE-seq nominated site was expanded to include 10 Kb upstream and downstream of the target site. Within our analysis, edits situated within the predicted target site were considered true positives, while those located upstream or downstream were labeled as false positives. Upon inspecting Mutect2 editing events identified in GM12878 (using gRNA EMX1) (**Figure 8.23**) and K562 cells (**Figure 8.24**), a significant proportion of the called events were found outside (both upstream and downstream) of the GUIDE-seq target sites. This prevalence of false positives was also observed with Strelka2 and Varscan2, across cells treated with other gRNAs (**Figure 8.25**). However, in GM12878 cells targeted by RNF2, both Mutect2 and Strelka2 did not detect any editing events outside of the designated target site (**Figure 8.25 (C)**). Similarly, in K562 cells targeted with EMX1 gRNA, Strelka2 exclusively identified edits within the specified target region (**Figure 8.25 (D)**). Interestingly, we noticed a higher frequency of SNV calls compared to insertions or deletions (indels) in the samples, indicating that Mutect2, Strelka2, and Varscan2 are more inclined to detect the genetic diversity inherent within cell lines rather than authentic editing events. Although Mutect2 was able to detect true positive editing events lying within the target site, particularly within strong GUIDE-seq sites, most events were discarded upon applying filters, tagged as artifacts (**Figure 8.23 (B)** and **8.24 (B)**). However, such explicit filtering steps were not provided by Strelka2 and Varscan2. These observations are expected, as Mutect2, Strelka2, and Varscan2 were not specifically tailored to identify CRISPR editing events. Consequently, while they recovered several true positive events, they also reported numerous false positive events, potentially biasing the quantification and detection of editing events, especially those occurring outside the expected edited regions. Moreover, the noise in the signal returned by the variant callers could further skew the quantification of editing efficiency. It is noteworthy that the variant callers seem to capture true positive editing events mainly on the strongest sites. However, studies have demonstrated that editing may be observed on targets displaying up to 6 mismatches to the gRNA (Tsai *et al.*, 2015; Clement *et al.*, 2020). Given these considerations, we further explored how variant callers behave when focusing solely on computationally predicted target sites.

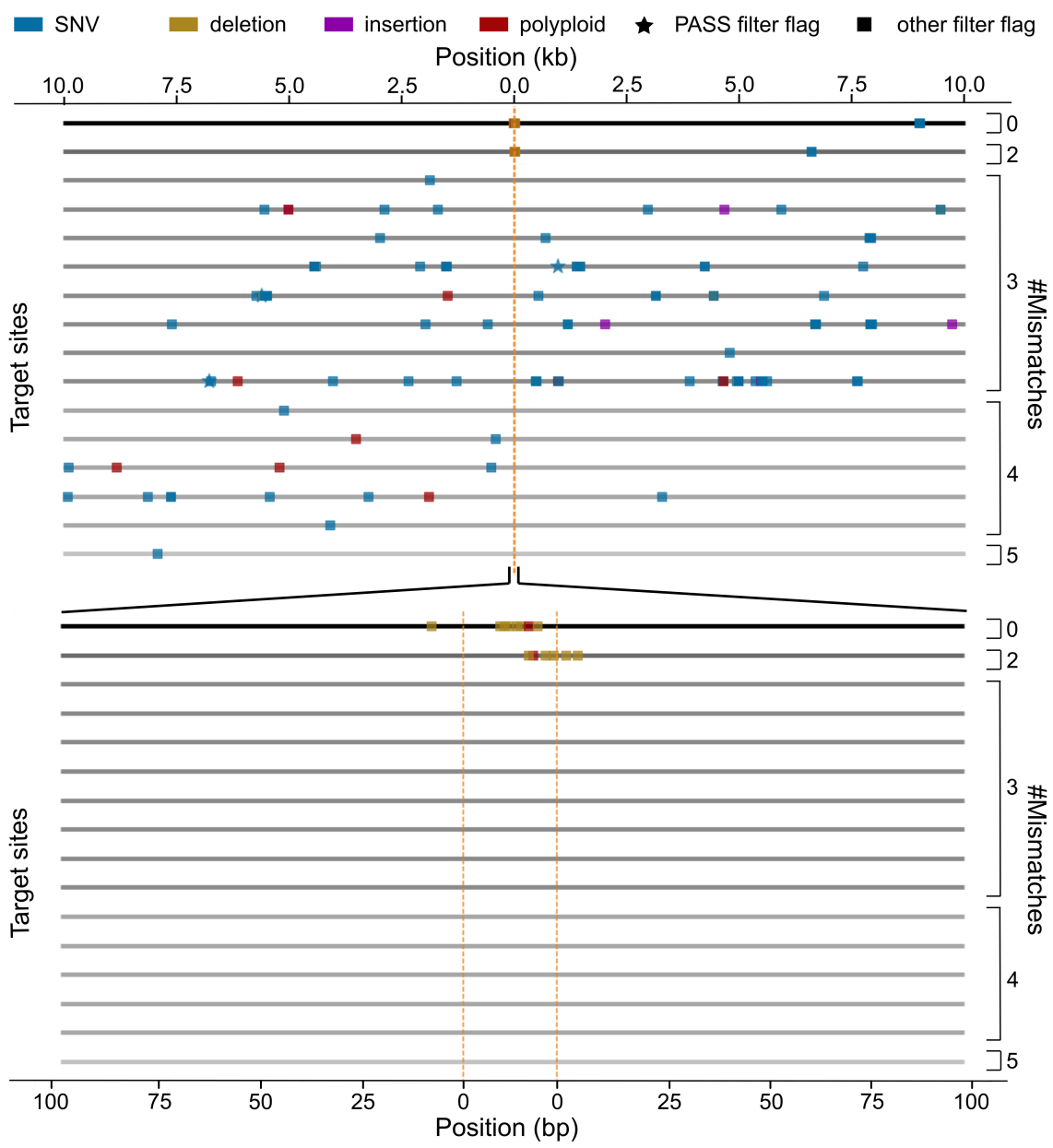
8.3.2 Detecting genome editing events in CasOffinder sites

We extended our investigation by focusing on target sites predicted computationally by CasOffinder (Bae *et al.*, 2014), comparing the performance of variant callers with that of CRISPResso2 (Clement *et al.*, 2019), a widely recognized computational tool for quantifying genome editing events in WGS experiments. For each CasOffinder site, we computed an editing rate by comparing the number of reads supporting the edited allele against those supporting the wild-type allele. The editing rate for a site s is calculated using the following formula:

$$\text{editing}_{\text{rate}}(s) = \sum_{a} \frac{\text{reads}_{\text{alt}}}{\text{reads}_{\text{ref}_a} + \text{reads}_{\text{alt}_a}}$$

where a represents each variant called within that site by the variant caller or CRISPResso. In this analysis, we focused on the strongest on-target and off-target sites predicted by CasOffinder. Generally, we observed that both variant callers and CRISPResso identified edits primarily in treated cells, predicting minimal editing in control cells (**Figure 8.26** and **8.27**). However, a notable exception was observed with Strelka2, which identifies some edits in control samples for gRNA EMX1 and HEKSite4 on GM12878

A

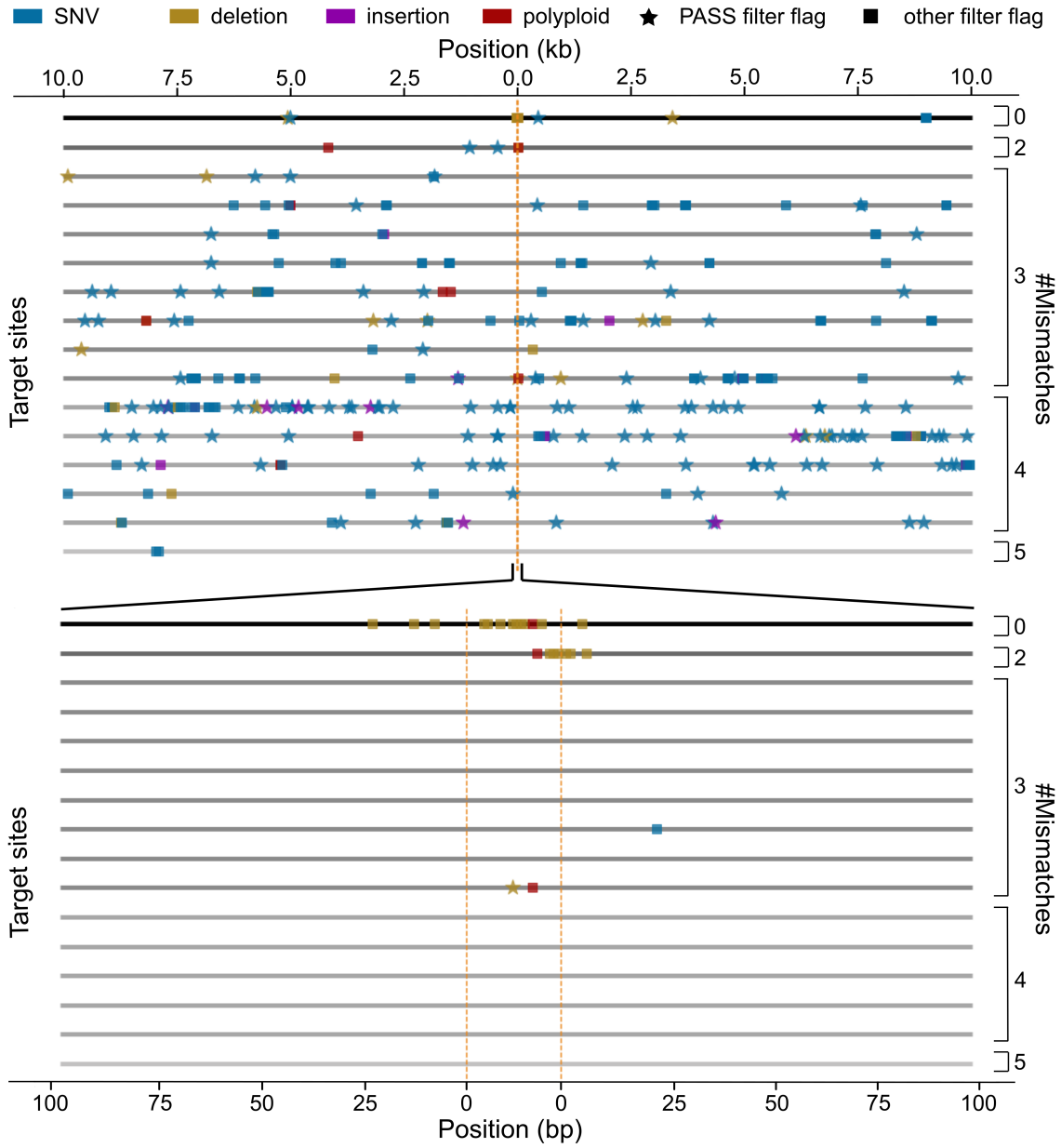


B

Variant type	TP-PASS	FP-PASS	TP-Other filter	FP-Other filter
SNV	0	3	0	99
deletion	0	0	12	8
insertion	0	0	0	7
polyploid	0	0	0	9

Figure 8.23. Editing events detected on GM12878 cell line (gRNA EMX1) using Mutect2. (A) Mutect2 detected editing events on the edited GM12878 (using gRNA EMX1) within a ∼20 Kb window around the GUIDE-seq nominated sites. Mutect2 identifies most events outside the intended target site, suggesting high false positives rate. When the window size was restricted to 100 bp, Mutect2 detects true editing events (indels and polyploid events) only at strong sites with up to 2 mismatches. **(B)** The table provides a summary of the true and false positive events detected by Mutect2. After applying filtering steps, Mutect2 discarded a significant number of true editing events, affecting its accuracy.

A



B

Variant type	TP-PASS	FP-PASS	TP-Other filter	FP-Other filter
SNV	0	122	0	148
deletion	1	12	16	18
insertion	0	8	0	11
polyploid	0	0	3	8

Figure 8.24. Editing events detected on K562 cell line (gRNA EMX1) using Mutect2. (A) Mutect2 detected editing events on the edited K562 (using gRNA EMX1) within a ∼20 Kb window around the GUIDE-seq nominated sites. As expected from K562 genetic make-up, Mutect2 recovers significantly more events compared to GM12878. However, Mutect2 still identifies most events outside the intended target site. Restricting the window size to 100 bp, we observe that Mutect2 detects true editing events (indels and polyploid events) at both strong and weak sites (up to 2 and 3 mismatches). **(B)** The table provides a summary of the true and false positive events detected by Mutect2. After applying filtering steps, Mutect2 discarded a significant number of true editing events. However, Mutect2 filter keeps a deletion detected on a target site (3 mismatches).

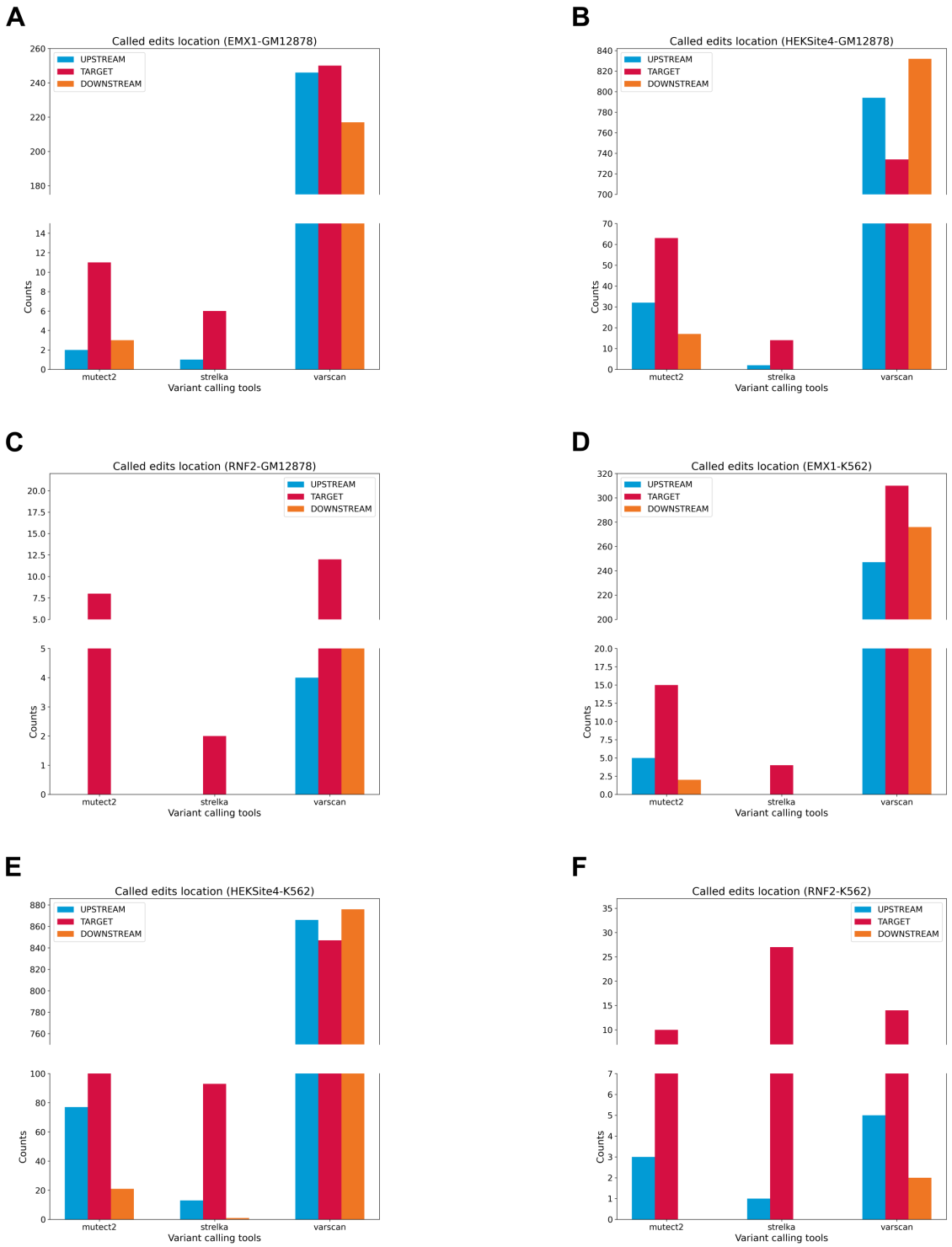


Figure 8.25. Detected editing events locations on GM12878 and K562 cells. (A-C) The locations of editing events identified by Mutect2, Strelka2, and Varscan2 in GM12878-Cas9 cells are analyzed. The majority of the editing events reported by the three tools are located outside the target sites predicted by GUIDE-seq. **(C)** Interestingly, in cells targeted by RNF2, both Mutect2 and Strelka2 do not detect any editing events outside of the designated target site. **(D-F** The analysis of editing events detected by Mutect2, Strelka2, and Varscan2 in K562 cells reveals that, similar to GM12878, most of the reported editing events occur outside the target sites predicted by GUIDE-seq. **(D)** Notably, when targeting EMX1, Strelka2 exclusively identifies edits within the specified target region.

cells (**Figure 8.26 (A)** and **(C)**), and Mutect2, which identifies some edits in control samples for gRNA RNF2 on GM12878 cells (**Figure 8.26 (A)**). As expected, editing rates are higher in K562 cells than GM12878, due to their genetic diversity. Interestingly, all variant callers and CRISPResso successfully identified and recovered editing events occurring in the on-target sites. However, variant callers appeared to generally underestimate editing events occurring in the best off-target site. This observation aligns with our previous findings, where we demonstrated how variant callers struggle to recover editing events in weaker sites.

8.4 Discussion, limitations and future directions

Whole-genome sequencing has emerged as a promising method for detecting genome editing due to its ability to capture sequence variations across the entire genome without the biases associated with amplicon sequencing. Unlike experimental nomination methods, WGS does not rely on biochemical properties and offers a comprehensive view of genomic alterations. Previous attempts at WGS, however, have been limited by insufficient power to detect rare mutations, particularly in heterogeneous populations. Despite its potential, using WGS for identifying CRISPR-Cas editing sites presents several challenges. The primary obstacle is the cost associated with next-generation sequencing, especially considering that the majority of sequenced reads do not correspond to potential editing sites. Moreover, existing computational tools for WGS analysis are not optimized to identify sequence alterations induced by CRISPR-Cas9 editing, and the computational resources required for analyzing ultra-deep WGS samples are often prohibitive. We investigated the effectiveness of ultra-deep WGS in detecting rare editing events resulting from CRISPR-Cas9 editing, benchmarking different computational methods commonly used to detect and quantify editing events. Our findings reveal that although the compared computational tools may capture some editing events, other are ignored or underestimated. Moreover, the development of methods to tailored to exactly quantify editing events is required to better capture and estimate editing at less deep sequencing rates. Furthermore, translating WGS findings to clinical applications poses additional challenges. WGS is a destructive process, meaning that cells analyzed by WGS may not accurately represent those used in therapy. There is also a risk that cells with rare editing events may be present in the therapeutic cell population, and biological conditions observed in WGS samples may not fully reflect those in the clinical setting. However, despite these challenges, using WGS at a specified depth offers a means to quantify the risk of failure in medical interventions, particularly the risk of editing at unintended targets. This approach provides a quantifiable measure of risk that can inform decision-making in precision medicine applications.

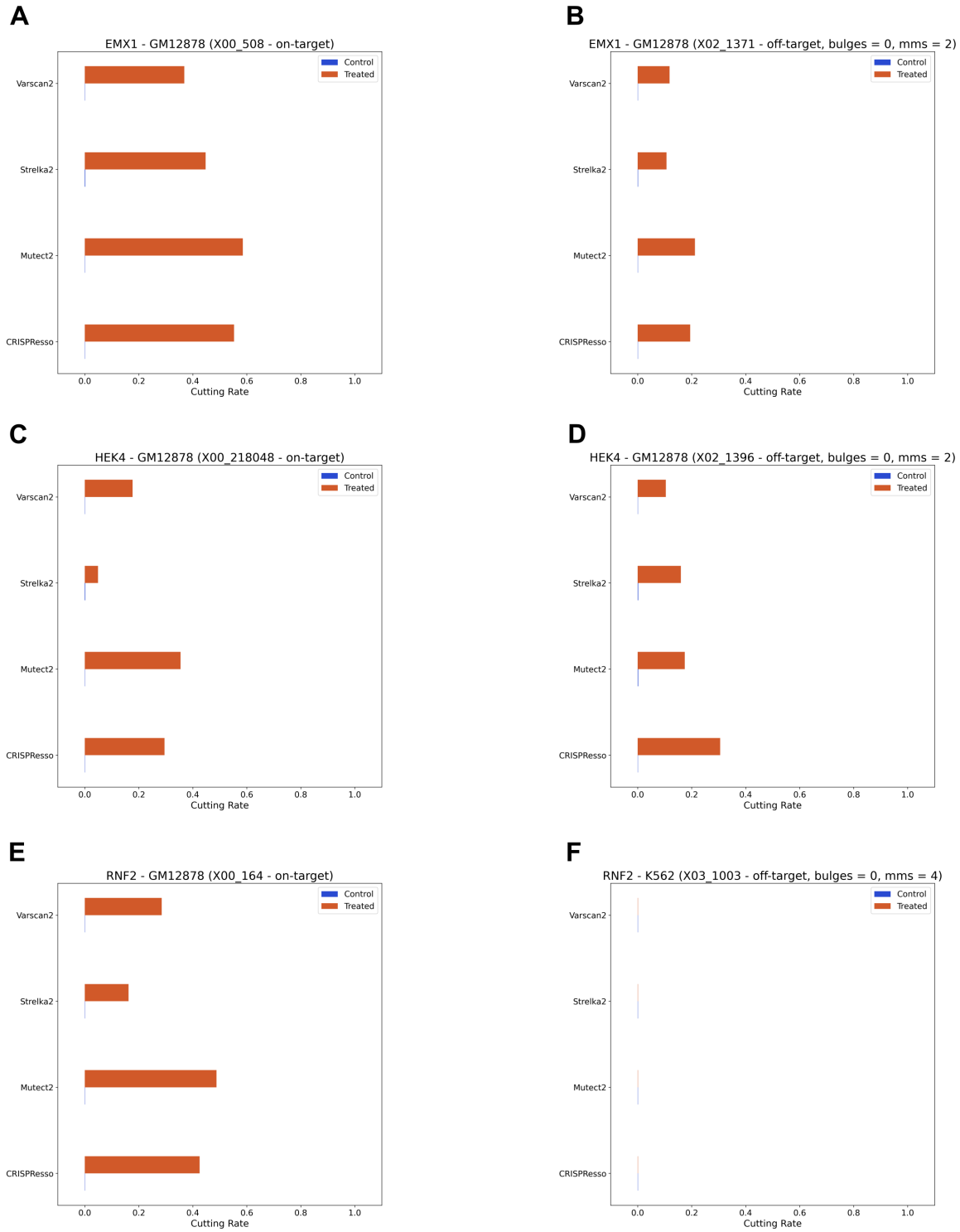


Figure 8.26. Editing rates on treated and control GM12878 cells. Editing rates computed analyzing treated and control GM12878 cells on computationally predicted on-target **(A)** and off-target **(B)** targeting EMX1, **(C-D)** HEKSite4, and RNF2 **(E-F)**.

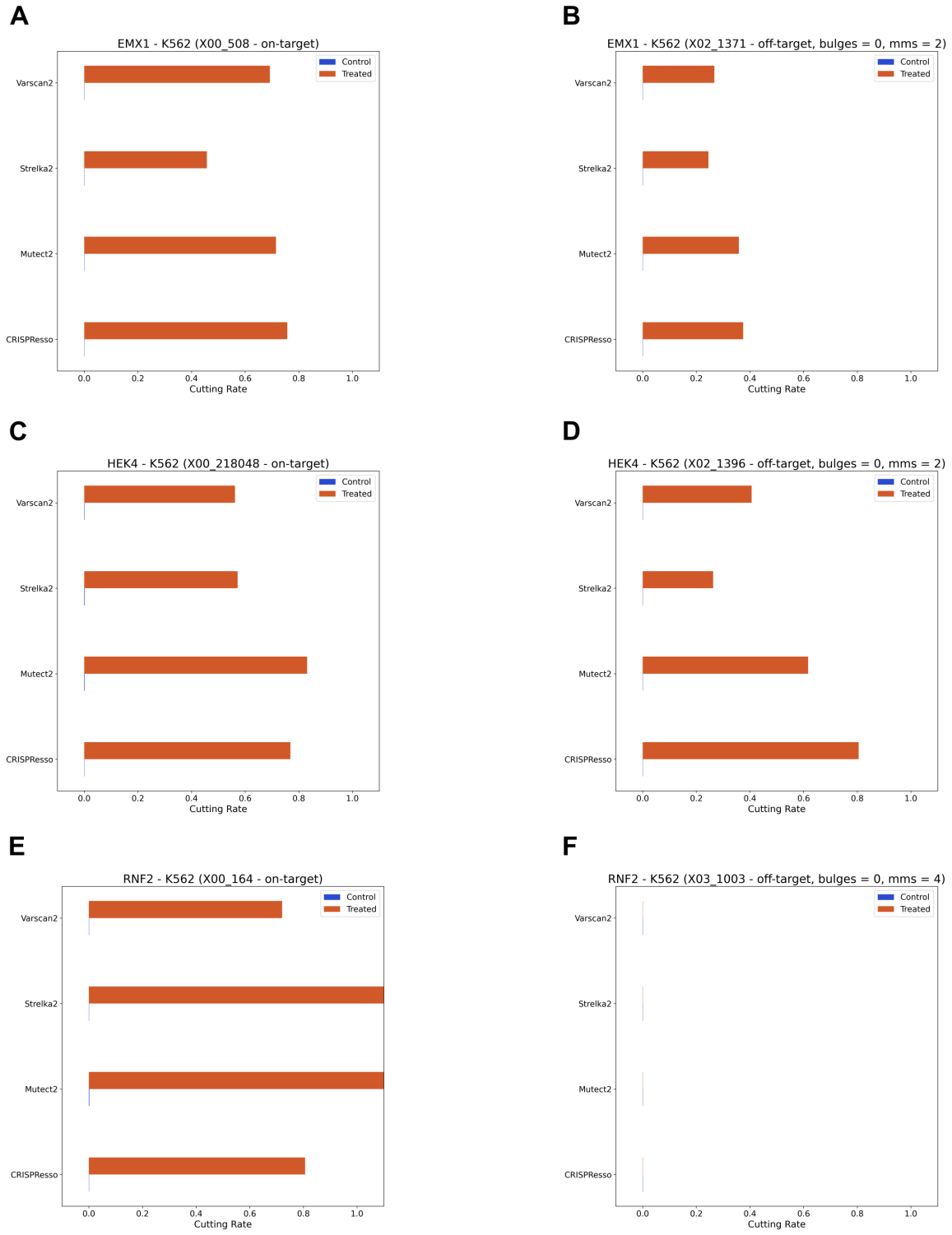


Figure 8.27. Editing rates on treated and control K562 cells. Editing rates computed analyzing treated and control K562 cells on computationally predicted on-target **(A)** and off-target **(B)** targeting EMX1, **(C-D)** HEKSite4, and RNF2 **(E-F)**.

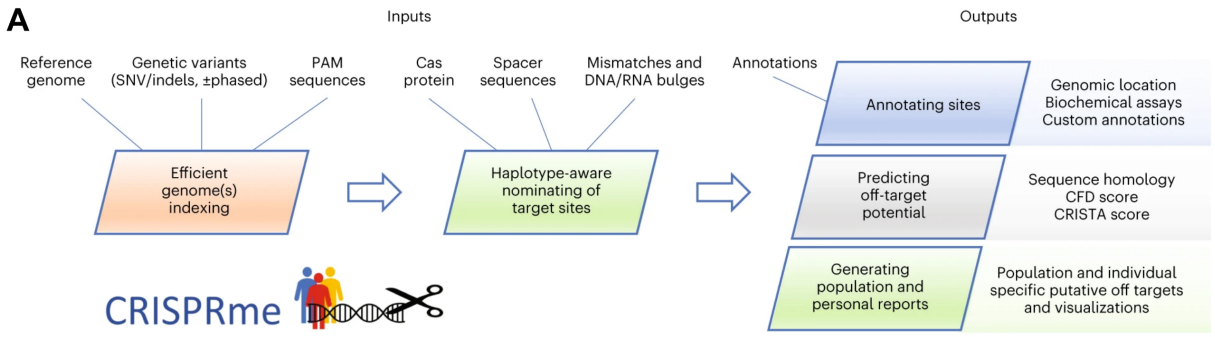
Genetic diversity alters potential therapeutic CRISPR genome editing off-targets outcomes

CRISPR genome editing stands as a groundbreaking approach for crafting innovative therapeutics through the precise manipulation of genetic or epigenetic elements within targeted genomic regions. However, the therapeutic potential of CRISPR-based systems is tempered by the risk of unintended off-target modifications, which could potentially lead to genotoxic effects. To mitigate this risk, a plethora of experimental assays and computational methodologies have been devised to uncover or predict these off-target sites (Clement *et al.*, 2020). Off-target site can be partially predicted based on their homology to the spacer and PAM sequence (**Section 7.3**). However, a multitude of sequence features beyond simply the number of mismatches or presence of bulges influence off-targets potential (Clement *et al.*, 2020; Bao *et al.*, 2021; Hsu *et al.*, 2013; Doench *et al.*, 2016). Computational models play a pivotal role in complementing experimental approaches to off-target nomination by triaging gRNAs before experiments (**Section 7.4.2**). These methods predict the number and cleavage potential of off-target sites, and prioritize target sites for experimental scrutiny. Moreover, genetic variants may significantly alter the landscape of both on-target and off-target editing. Strategies designed to specifically target patient mutations may increase the likelihood of editing mutant alleles, while variants that reduce homology to the intended target sequence may diminish the efficiency of the desired genetic modification. Although various experimental methods exist for empirically nominating off-target sites *in vitro* and *in cellula*, these methods typically rely on homology to the reference genome or a limited set of human donor genomes to evaluate off-target potential (Bao *et al.*, 2021; Chaudhari *et al.*, 2020). Therefore, computational methods offer a particularly valuable avenue for predicting the impact of off-target sequences not present in reference genomes. Previous studies have utilized population-based variant databases such as the 1000 Genomes Project (1KGP) and the Exome Aggregation Consortium (ExAC) to demonstrate how genetic variants can significantly alter the off-target landscape. Specifically, genetic variants may generate novel and personalized off-target sites not present in a single reference genome (Lessard *et al.*, 2017; Scott and Zhang, 2017). However, existing computational tools often have limitations, such as their inability to efficiently handle large variant datasets, consider higher numbers of mismatches, model RNA:DNA hybrid bulges, and accommodate alternative haplotypes and indels. Moreover, these methods often require advanced computational skills, limiting their potential audience. Several user-friendly websites have emerged to facilitate the design of gRNAs and evaluate their potential off-target effects (Concordet and Haeussler, 2018; Listgarten *et al.*, 2018; Labun *et al.*, 2019; Park *et al.*, 2015). However, while these scalable graphical user interface (GUI) tools are invaluable for gRNA design, they do not address genetic variants. These tools typically impose artificial constraints on the number of mismatches for the search and often lack support for DNA/RNA bulges. Consequently, using the available tools to design gRNAs for therapeutic purposes may overlook significant off-target sites, potentially leading to unintended genotoxicity. A comprehensive and exhaustive haplotype-aware off-targets search, encompassing an arbitrary number of mismatches, bulges, and genetic variants while being, poses a significant computational challenge requiring specialized and efficient data structures. CRISPRitz (Cancellieri *et al.*, 2020), a recently introduced command line tool, address this challenge. By employing optimized data structures, CRISPRitz efficiently searches off-target sites accounting for SNVs, mismatches and bulges. However, CRISPRitz has inherent limitations. CRISPRme substantially extends CRISPRitz by proposing a comprehensive tool designed to assist in gRNA design while incorporating support for haplotype-aware off-target enumeration, short indel variants, and flexible parameters for mismatches and bulges (Cancellieri *et al.*, 2023). CRISPRme is a unified, user-friendly web-based application providing several reports to prioritize putative off-target sites based on their risk

at population or individual level. Additionally, CRISPRme offers flexibility by allowing users to incorporate custom genomic annotations, including empirically identified off-target sites or cell type-specific chromatin features. Moreover, CRISPRme can integrate population genetic variants from phased individual variants datasets (e.g., from 1000 Genomes Project (Consortium *et al.*, 2015; Lowy-Gallego *et al.*, 2019)), unphased individual variants (e.g., from the Human Genome Diversity Project (Bergström *et al.*, 2020)), and population-level variants (e.g., from the Genome Aggregation Database (Karczewski *et al.*, 2020)). Furthermore, CRISPRme can analyze personal genomes from individual subjects to identify and prioritize private off-targets created by variants unique to a single individual. To demonstrate the tool's utility, we conduct an analysis of the off-target potential of a gRNA currently undergoing clinical trials for sickle cell disease (SCD) and β-thalassemia (Frangoul *et al.*, 2021; Canver *et al.*, 2015; Wu *et al.*, 2019). Our analysis reveals potential off-targets introduced by genetic variants within and extending beyond the 1KGP dataset. Notably, we identify a previously overlooked off-target site introduced by a variant prevalent in individuals of African ancestry (rs114518452, minor allele frequency (MAF) = 4.5%). Moreover, we provide experimental evidence of its off-target potential in gene-edited human CD34+ hematopoietic stem and progenitor cells (HSPCs). Additionally, we illustrate that allele-specific off-target potential is pervasive across various nucleic acid targeting therapeutic strategies.

9.1 CRISPRme: a computational tool for variant-aware off-target nomination

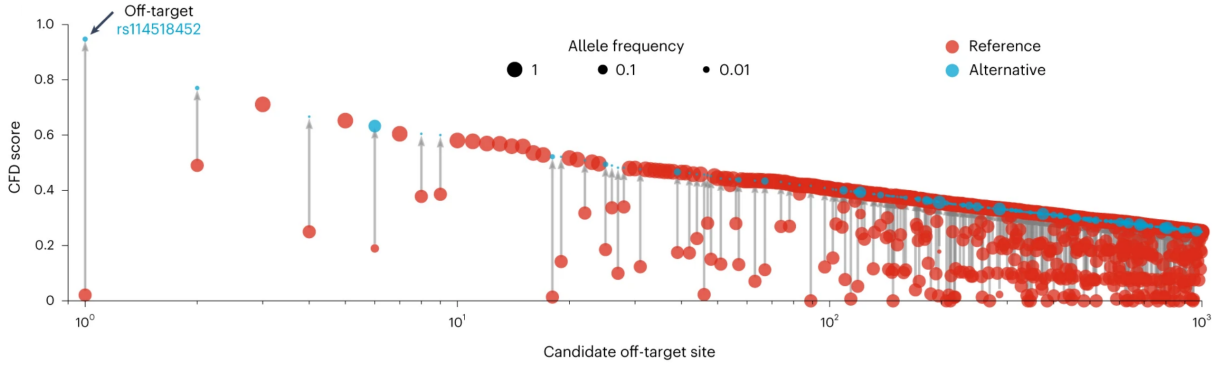
CRISPRme is a web-based tool designed to predict the off-target potential of CRISPR gene editing, that accounts for the impact of genetic variation. CRISPRme is available online at `http://crisprme.di.univr.it`. CRISPRme offers a versatile platform that can also be deployed to local, secure, and isolated environments, as both a web application or a command-line utility. Importantly, neither deployment option involves the transfer or storage of data online, ensuring compliance with genomic privacy and regulations. As input, CRISPRme requires a Cas protein, gRNA spacer sequence, and PAM, a genome build, sets of variants (provided as VCF files for populations or individuals), user-defined thresholds for mismatches and bulges, and optional genomic annotations specified by the user. These inputs collectively enable CRISPRme to generate comprehensive and personalized reports (**Figure 9.28 (A)**). CRISPRme is designed to accommodate diverse gene editors with flexible and relaxed PAM requirements (Walton *et al.*, 2020). Leveraging a PAM encoding method based on Aho-Corasick automata and an index based on a ternary search tree, CRISPRme efficiently performs genome-wide exhaustive searches. Even when tested with challenging conditions such as an `NNN` PAM, extensive mismatches (validated with up to 7 mismatches), and the presence of RNA:DNA bulges (validated with up to 2 bulges), CRISPRme demonstrates robust performance. Notably, a comprehensive search with up to 6 mismatches, 2 DNA/RNA bulges, and a fully nonrestrictive PAM (`NNN`), on a small computational cluster node equipped with 20 CPU cores and 128 GB RAM (Intel Xeon CPU E5-2609 v4 clocked at 2.2 GHz) took ∼34 hours of real-time processing and ∼152 hours of CPU time (including both user and system times). Importantly, all available 1KGP variants, including both SNVs and indels, can be included in the search alongside accompanying metadata for each individual, such as sex, superpopulation, and age. The search accounts for the haplotypes observed in the input variants datasets. To prevent the inflation of reported sites, off-target sites representing alternative alignments to a specific genomic region are merged. While several tools are available for off-target enumeration, only two command-line tools incorporate genetic variants in the search (Lessard *et al.*, 2017; Fennell *et al.*, 2021). However, these tools have several limitations concerning scalability for large searches, supported number of mismatches, bulges, haplotypes, and variant file formats, and notably lack an easy-to-use graphical user interface (GUI). CRISPRme generates different reports: (i) gRNA off-targets report, (ii) a comparison of gRNAs against customizable annotations, (iii) cumulative distribution of homologous sites based on the reference genome or superpopulations, and (iv) individual-focused reports. The gRNA off-targets report provides a summary for each gRNA, listing all potential off-targets identified in the reference or variant genomes based on mismatches and bulges (**Figure 9.28 (B)**). Additionally, CRISPRme generates a file containing comprehensive information on each identified candidate off-target. By default, CRISPRme annotates gRNAs by categorizing potential off-target sites using GENCODE (Frankish *et al.*, 2019) for genomic features and ENCODE (Consortium *et al.*, 2012) for candidate cis-regulatory elements (cCREs). Users can incorporate their own annotations in BED format (**Appendix A.6**), such as empirically derived off-target scores or cell type-specific chromatin features (**Figure 9.29**). By leveraging 1KGP (Consortium *et al.*, 2015) and/or HGDP (Bergström *et al.*, 2020) variants, CRISPRme presents the cumulative distribution of homologous sites based on the



B

Spacer + PAM	Nuclease	Genome	Bulges	Mismatches						
				0	1	2	3	4	5	6
CTAACAGTTGCTTTTATCACNNN	SpCas9	Reference	0	1	0	7	154	1,645	13,869	96,008
			1	1	13	486	7,063	73,885	542,165	3,124,545
			2	14	417	9,183	121,231	1,109,580	7,245,109	34,849,276
		Variant	0	0	0	1	12	109	823	4,969
			1	0	2	28	355	2,801	18,968	278,205
			2	2	27	305	3,202	26,259	188,954	8,634,796

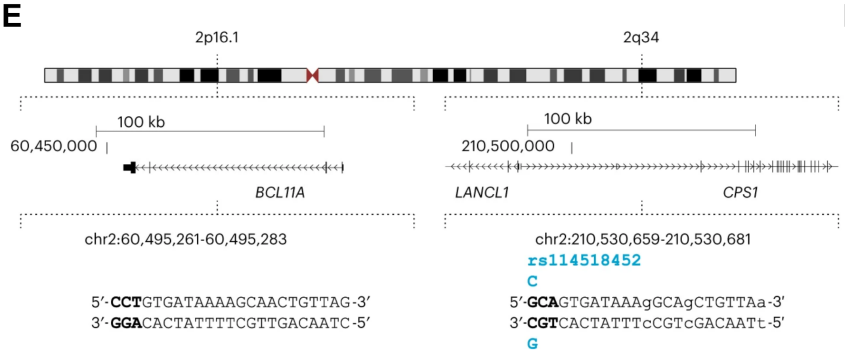
C



D

Sequence	Alignment	Chr	Position	Strand	Variant ID	CFD	MAF	Annotation
Spacer+PAM	CTAACAGTTGCTTTTATCACNNN							
Reference	tTAACAGcTGCcTTTATCACTGC	2	210530658	-		0.021		Intron:*CPS1*
Alternative	tTAACAGcTGCcTTTATCACTG**G**				rs114518452 2-210530659-G-C	0.947	0.02	

E



F

rs114518452 2-210530659-G-C

MAF (%)	Population
4.55	African/African American
0.91	Other
0.66	Latino/Admixed American
0.12	South Asian
0.01	European (non-Finnish)
0.00	Finnish/Amish/East Asian/ Middle Eastern/Ashkenazi Jewish

Figure 9.28. CRISPRme offers a web-based platform to analyze the off-target potential of CRISPR-Cas gene editing, taking into account population-level genetic diversity. **(A)** CRISPRme analyzes CRISPR-Cas gene editing off-target potential. It accepts inputs such as a reference genome, genetic variants, PAM sequence, Cas protein type, spacer sequence, homology threshold, and genomic annotations. Through a user-friendly interface, it provides comprehensive analyses tailored to specific targets, allowing both target-focused and individual-focused assessments of off-target potential. Available as an online web tool, CRISPRme can also be deployed locally or utilized offline as command-line software. **(B)** Analysis of off-target's potential of BCL11A-1617 spacer targeting the +58 erythroid enhancer with SpCas9, `NNN` PAM, considering 1KGP variants (up to 6 mismatches and 2 bulges). **(C)** The top 1,000 predicted off-target sites ranked by CFD score, provide insight into their potential off-target effects. The CFD score is displayed for both the reference and alternative allele, where applicable, with the size of the circle representing the allele frequency. **(D)** The off-target showing the highest CFD score corresponds to the minor allele of rs114518452 (coordinates the potential off-target on hg38 are 0-based and 1-based for the variant ID). The minor allele frequency (MAF) is based on 1KGP data. **(E)** The top predicted off-target site exhibits allele-specific off-targeting, with 3 mismatches to the BCL11A-1617 spacer. rs114518452-C minor allele creates a *de novo* `NGG` PAM sequence, as indicated by the PAM (in bold). Mismatches to BCL11A-1617 spacer are displayed in lowercase. Coordinates on hg38 are 1-based. **(F)** Allele frequencies of rs114518452, sourced from gnomAD v3.1.

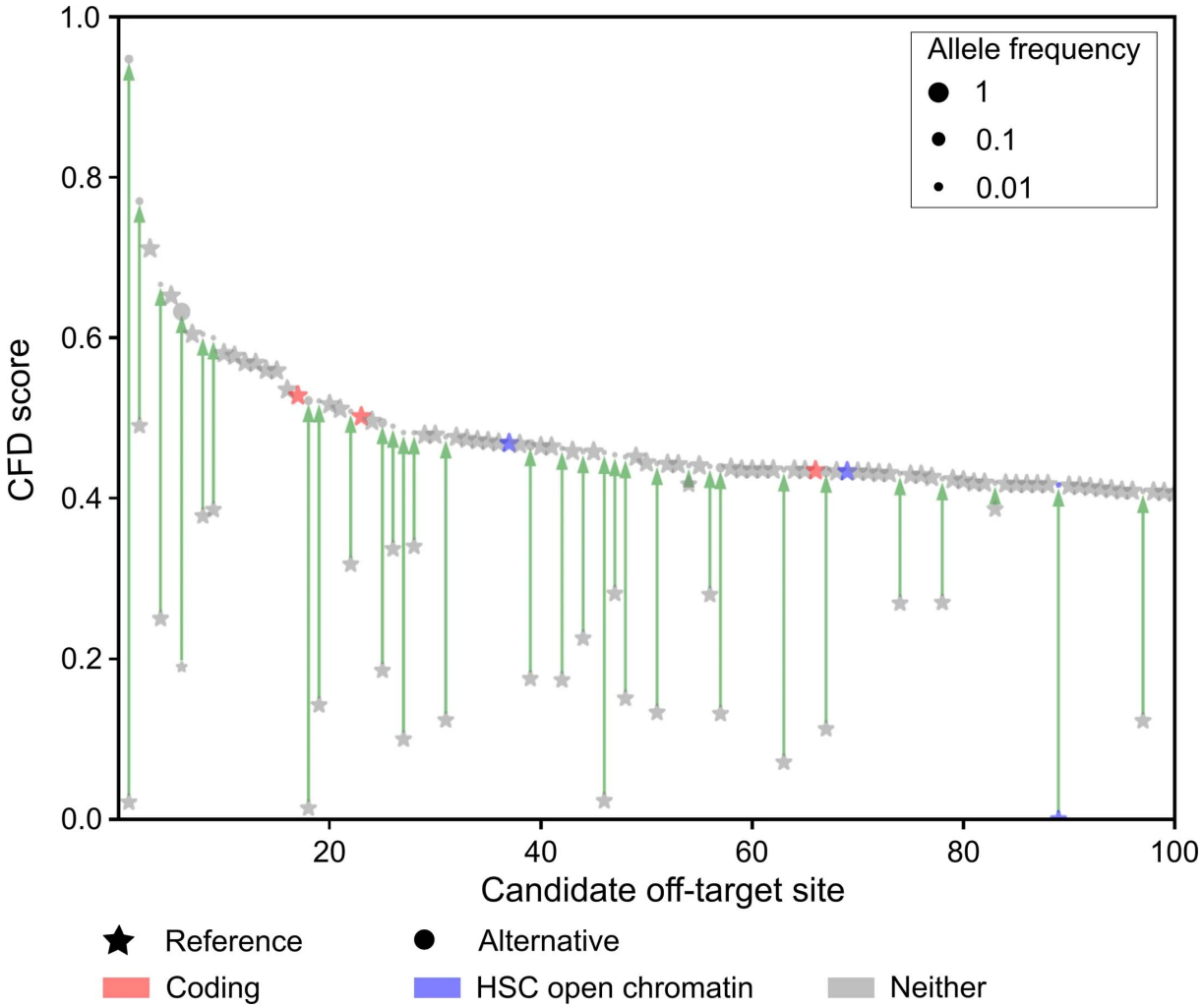


Figure 9.29. The top 100 predicted off-target sites for the BCL11A-1617 spacer ranked by their CFD scores. Results from CRISPRme's search depicted in Figure 9.28. Candidate off-target sites located within coding regions, as identified by GENCODE annotations, and within ATAC-seq peaks in hematopoietic stem cells (HSCs), based on user-provided annotations are highlighted (data sourced from Corces *et al.* (2016)).

reference genome or superpopulation. These global reports enable comparison of different gRNA sets based on the impact of genetic variation on their predicted on- and off-target cleavage potential using cutting frequency determination (CFD) (Doench *et al.*, 2016) or CRISPR Target Assessment (CRISTA) (Abadi *et al.*, 2017) scores (**Figure 9.30**). CRISPRme offers multiple scoring metrics and can readily incorporate new ones, including those tailored for various gene editors. CRISPRme also generates personalized genome-focused reports, called personal risk cards, which highlight private off-target sites attributed to unique genetic variants.

9.1.1 CRISPRme off-targets search

CRISPRme is available through an online web application hosted at `http://crisprme.di.univr.it/` (tested for compatibility with browsers like Google Chrome, Mozilla Firefox, and Apple Safari). Additionally, CRISPRme can be deployed offline as either a local web application or a standalone command-line package. To initiate a search, users need to input parameters such as gRNA spacer(s), Cas protein, PAM sequence, genome build, and thresholds for mismatches and DNA/RNA bulges. Optional inputs include genetic variant datasets (such as 1KGP, HGDP, and/or personal variants) and annotations. CRISPRme search is straightforward and involves just three simple steps facilitated by the user-friendly interface (**Figure 9.31 (A)**). Furthermore, the tool offers several customization options to tailor the search according to specific requirements.

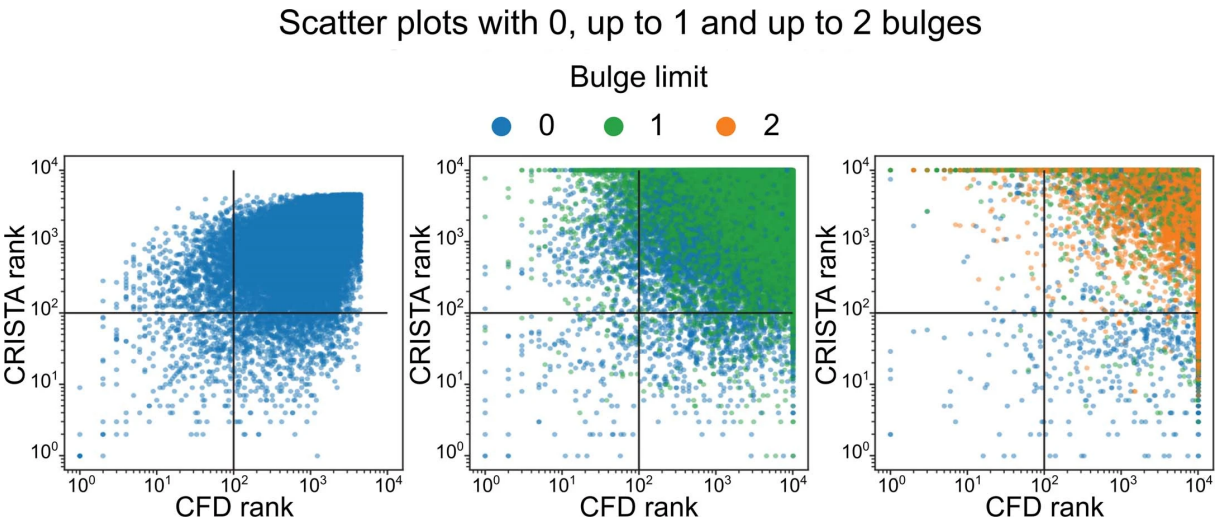


Figure 9.30. Plots depicting the rank-ordered correlation between CFD and CRISTA reported targets. The scatter plots illustrate the correlation between ranked targets, obtained by selecting the top 10,000 targets sorted by CFD and CRISTA scores, respectively. From left to right, the plots depict the correlation of targets with 0, 1, and 2 bulges. In the left plot, which represents targets with 0 bulges, Pearson's correlation is 0.57 ($P < 1e^{-10}$) and Spearman's correlation is 0.55 ($P < 1e^{-10}$). Moving to the center plot, which shows targets with 1 bulge, Pearson's correlation is -0.16 ($P < 1e^{-10}$), and Spearman's correlation is -0.33 ($P < 1e^{-10}$). Finally, in the right plot, representing targets with 2 bulges, Pearson's correlation is -0.55 ($P < 1e^{-10}$), and Spearman's correlation is -0.80 ($P < 1e^{-10}$). These correlation values and P-values (two-sided) were computed using standard functions from the Python scipy library. The colors in the plots indicate the lowest count of bulges for each target, as different alignments prioritized by the two scoring methods may result in varying numbers of mismatches and bulges for the same target.

Spacer, Cas protein, and PAM selection

The gRNA spacer sequence, typically 20 bp long, matches the genomic target protospacer sequence and directs Cas protein binding in the presence of a PAM (**Section 7.1.2**). In CRISPRme gRNAs are represented as DNA for easy comparison to the aligned protospacer sequence. CRISPRme allows the user to input a set of gRNA spacers each without a PAM (max 100 sequences in the online version). Alternatively, users can input a set of genomic coordinates in BED format (**Appendix A.6**) or DNA sequences in FASTA format (**Appendix A.1**). BED coordinates are treated as 0-based, and CRISPRme extracts the first 100 possible spacer sequences within these coordinates, starting from the positive strand. The protospacer adjacent motif (PAM) is essential for Cas protein binding to a specific DNA target (**Section 7.1.3**). CRISPRme supports various PAMs. In the online version, users can select one available PAM to perform the search. Both 3' (e.g., SpCas9) and 5' (e.g., Cas12a) PAM sequences are supported.

Genome selection and threshold configuration

CRISPRme supports genome builds based on FASTA files (**Appendix A.1**). The software accommodates reference sequences available in FASTA format such as transcriptomes, genomes from other organisms, and cancer genomes. The default hg38 genomic build, including mitochondrial DNA, is available in the online version, with options to incorporate variants from 1KGP and/or HGDP in the search. Personal variants can be added in the local offline and command-line versions. Users can specify the number of mismatches, DNA, and RNA bulges tolerated in enumerating potential off-targets. The online tool allows up to 6 mismatches and up to 2 DNA/RNA bulges (consecutive or interleaved). For local web and command-line versions, these thresholds are customizable based on available computational resources. Furthermore, CRISPRme enables users to specify the window for base editing susceptibility (**Section 7.2.1**) when selecting a base editor as the Cas protein (**Figure 9.31 (B)**). CRISPRme includes in the final report whether a candidate off-target displays base editing susceptibility or not.

Annotations and Email Notification

To assess off-target activity impact, CRISPRme provides functional annotations for coding and non-coding regions based on files from the Encyclopedia of DNA Elements (ENCODE) (Consortium *et al.*, 2012) and GENCODE (Frankish *et al.*, 2019). Users can add custom genome annotations (provided in BED format) in the offline version. Users may provide an email address in the online version to be notified upon job completion. Upon submitting the job, the search begins. Progress is shown on a new page, and

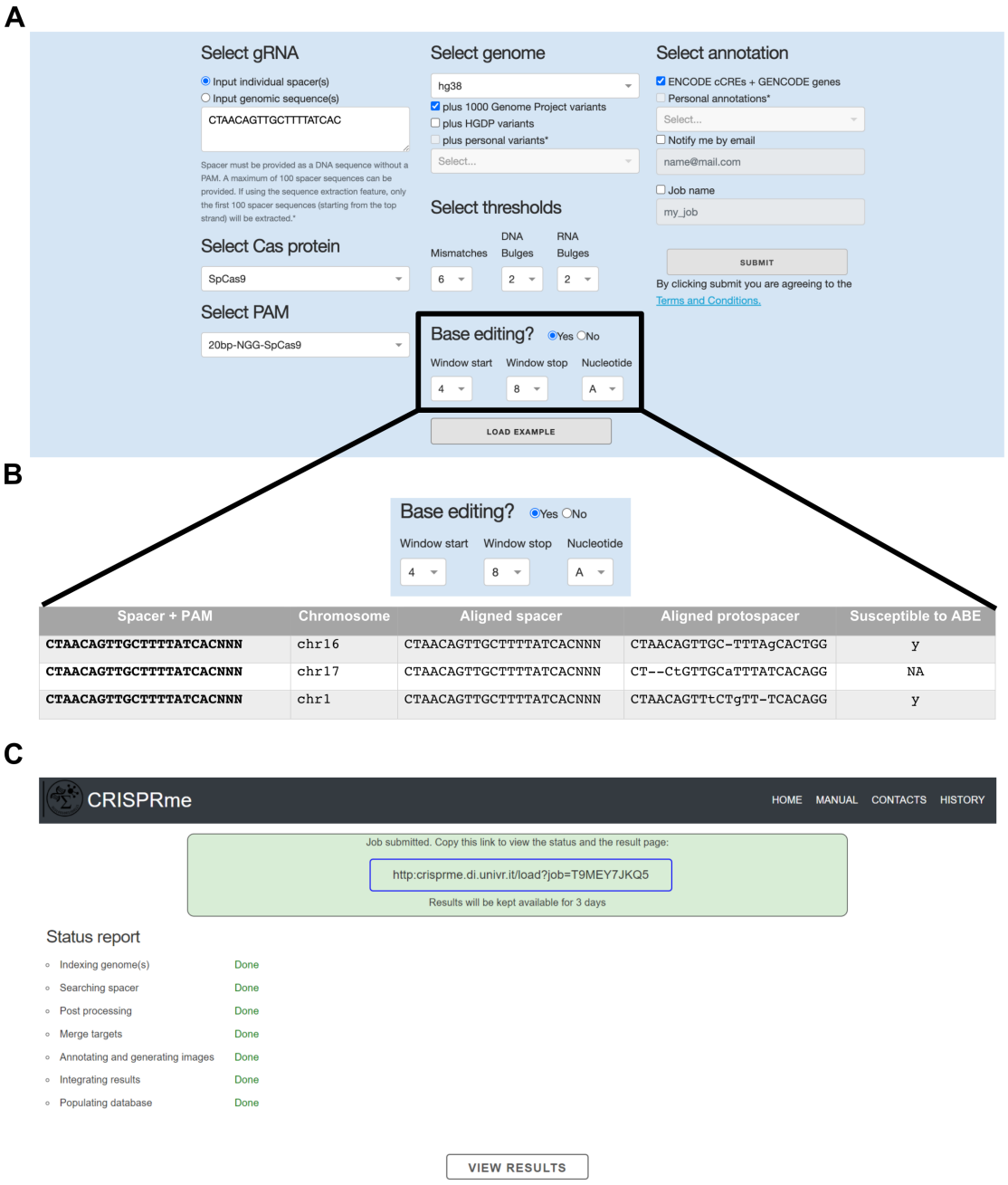


Figure 9.31. CRISPRme graphical user interface. (A) Main page layout of CRISPRme's graphical user interface. **(B)** Base editing options to flag potential off-target sites susceptible to base editing. The example illustrates two targets identified as susceptible to A base editing, while one target shows no susceptibility. A target is labeled as susceptible to <nt>BE if it contains the selected nucleotide within the specified window. Conversely, if the nucleotide is absent in the window, the target is labeled as non-susceptible. **(C)** CRISPRme's status report page.

upon completion, a "View Results" link appears at the bottom of the status report page (**Figure 9.31 (C)**).

9.1.2 CRISPRme output and graphical reports

CRISPRme summarizes the results in a table, emphasizing, for each gRNA, its CFD specificity score alongside the count of on-targets and off-targets detected in both the reference and variant genomes, grouped by the number of mismatches and bulges (**Figure 9.32 (A)**). Notably, the CFD specificity score was originally devised for searches encompassing up to 3 or 4 mismatches. However, as the number of mismatches increases, the specificity score diminishes non-linearly. Therefore, it's crucial to compare with caution these scores across searches with varying numbers of mismatches/bulges or different ge-

netic variant datasets. Furthermore, six other downloadable interactive reports are generated for each guide: *Custom Ranking*, *Summary by Mismatches/Bulges*, *Summary by Sample*, *Query Genomic Region*, *Graphical Reports*, and *Personal Risk Cards*.

CRISPRme targets tabular reports

Custom ranking reports enable users to filter and prioritize potential off-targets according to various criteria such as the number of mismatches and/or bulges, CFD/CRISTA score, Risk Score (indicating the impact of genetic variants), or a combination of these preferences (**Figure 9.32 (B)**). The report by mismatches/bulges presents a matrix grouping off-targets into subgroups according to their type, mismatch count, and bulge size. Targets labeled with "X" exclusively feature mismatches, while those labeled "DNA" include DNA bulges (potentially alongside mismatches), and "RNA" targets comprise RNA bulges (also potentially alongside mismatches) (**Figure 9.33 (A-B)**). The Summary by Sample displays data related to the samples available in the input VCFs, enabling users to extract and visualize specific targets associated with each sample (**Figure 9.33 (C-D)**). The Query Genomic Region page enables users to fetch off-targets that intersect with a particular genomic region. For instance, this capability allows rapid assessment of potential off-targets within a designated regulatory element or coding region (**Figure 9.33 (E)**).

CRISPRme graphical reports

The Graphical Reports page generates and displays different graphical reports for each input gRNA. Stem plots (**Figure 9.34 (A)**) illustrate how genetic variants impact the predicted off-target potential. The arrow connecting the red (reference allele off-target) and blue (alternative allele off-target) dots displays the change in predicted cleavage potential due to the variant. Bar plots display the distribution of candidate off-targets across super-populations based on the number of mismatches and bulges (**Figure 9.34 (B)**). A radar chart, derived from annotations from GENCODE and ENCODE, depicts a gRNA's potential off-targets falling within annotated regions. A larger area in the chart signifies more potential off-targets within these regions, which could indicate an undesirable outcome. A summary table provides the count and percentage of off-targets with a given annotation (**Figure 9.34 (B)**). A motif logo summarizes the frequency of mismatches and bulges among the predicted off-targets for each base of the protospacer + PAM (**Figure 9.34 (B)**).

CRISPRme Personal Risk cards

CRISPRme provides a dedicated feature known as Personal Risk Cards, designed to generate reports summarizing potential off-target editing by a specific gRNA in an individual, influenced by its genetic variants. This feature is particularly important to retrieve and investigate private off-targets. The report encompasses two dynamically generated plots showcasing all candidate variant off-targets for the sample, including non-unique and individual-unique targets (**Figure 9.35 (A)**). These plots underscore how the introduction of genetic variants can alter predicted off-target cleavage potential, emphasizing the importance of CRISPRme's variant-aware off-target assessment. Furthermore, the report includes two tables (**Figure 9.35 (B)**): (i) a summary table at the top, and (ii) detailed information on each extracted candidate off-target at the bottom. The summary table features several informative columns. The Personal column counts all candidate variant off-targets for the selected sample, encompassing both variants unique and non-unique to the individual. Similarly, the PAM creation column counts all instances where a genetic variant in the sample introduces a new PAM, with the PAM used in the search not found in the reference genome at the same position. Finally, the Private column counts all candidate variant off-targets uniquely identified in the selected sample. CRISPRme offers the opportunity to download the personal card table file, as a separate instance.

9.1.3 Computational details on CRISPRme implementation

The CRISPRme web version and frontend were developed using Dash, a Python framework renowned for building responsive and interactive web applications (`https://plotly.com/dash/`). Delving into the backend, which encompasses graphical report generation and data analysis, Python and bash scripts form the core components. Furthermore, to ensure optimal performance and stability, the search engine powering CRISPRme is developed in C++ and extends the original CRISPRitz off-target nomination tool (Cancellieri *et al.*, 2020). Leveraging the inherent speed and robustness of C++, the search engine

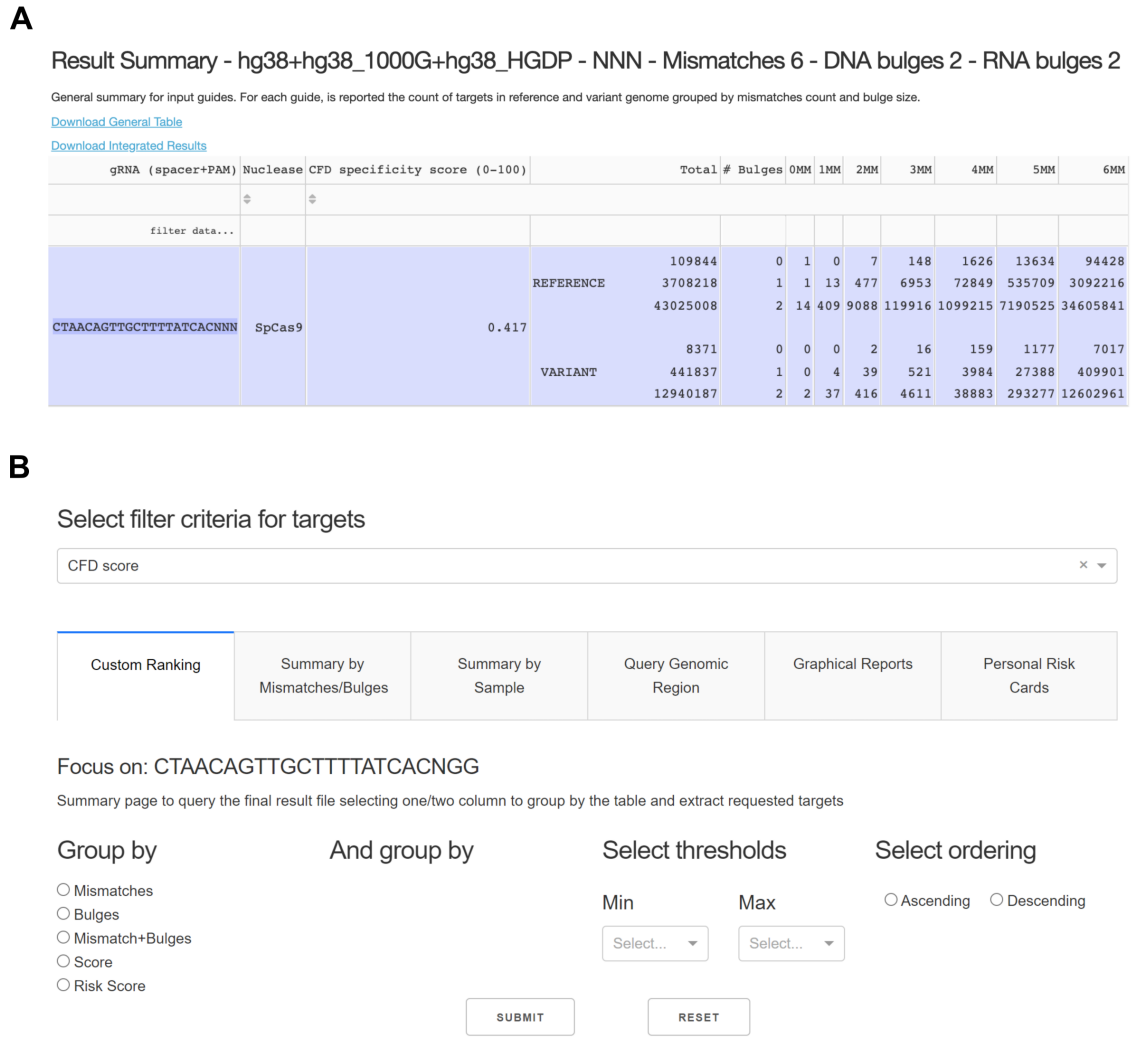


Figure 9.32. CRISPRme targets summary report. (A) The table provides a comprehensive overview of the search results performed with sg1617, utilizing an `NNN` PAM, allowing for up to 6 mismatches and 2 DNA or RNA bulges. The search was performed on the human reference genome, augmented with the 1KGP dataset comprising 5 super-populations, as well as the HGDP dataset featuring 7 super-populations. It includes columns detailing the nuclease used, the CFD specificity score, and the count of targets categorized based on the number of mismatches and bulges. Additionally, situated in the top left corner are two buttons: "Download General Table," enabling the download of the table as a text file, and "Download Integrated Results," facilitating the retrieval of the full results. **(B)** Users can specify filters, orders, and group-by operations to efficiently retrieve results based on custom logic tailored to their specific needs. Here, we display the first seven columns of the table, which include sequence and positional information.

fully harnesses parallel computation and compiler optimizations, thereby enhancing efficiency and scalability. This multi-faceted approach underscores CRISPRme's commitment to delivering a seamless user experience while conducting comprehensive off-targets nomination analyses. To perform variant- and haplotype-aware off-target nomination CRISPRme employs a tailored procedure enriching the reference genome sequence with variants (SNPs and short indels) defined in the input VCFs.

Enriching reference genomes with genetic variants

CRISPRme conducts off-target searches based on reference genomic sequences stored in FASTA files. To enhance the accuracy of this search, reference genomes can be "enriched" with genetic variants sourced from VCF files obtained from databases such as 1KGP (Consortium *et al.*, 2015), HGDP (Bergström *et al.*, 2020), or personal datasets. The enrichment process involves encoding SNPs and indels using IUPAC notation (**Algorithm 6**). IUPAC symbols allow to represent alternative alleles via ambiguous DNA characters. For instance, at a specific position where the reference allele is `G` and the alternative allele is `A`, both alleles are represented by the ambiguous character `R`, corresponding to the IUPAC symbol encoding both `G` or `A`. This enrichment technique ensures that the resulting genome encapsulates

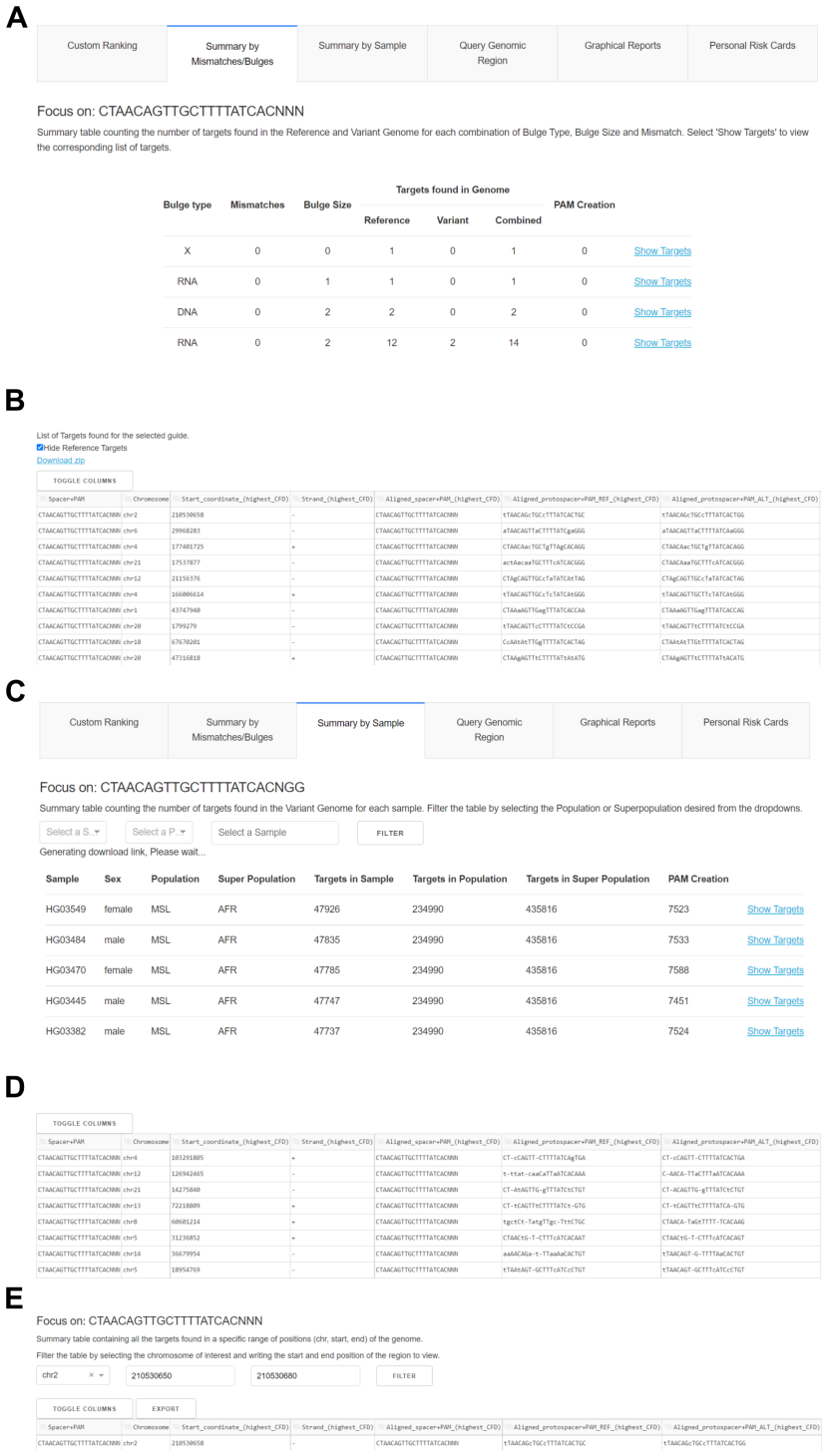


Figure 9.33. Summary report by Mismatches/Bulges, by Sample, and by Region. (A) The Mismatches/Bulges summary table displays the first 4 out of 33 rows for a search allowing up to 6 mismatches and 2 DNA or RNA bulges. The combined column aggregates the counts of off-targets found in both the reference and variant genomes. **(B)** View of "Show Targets" with 3 mismatches and no bulges. Users can toggle between different columns using the "Toggle Columns" button located at the top of the table. The first seven columns, which include sequence and positional information, are detailed in **Table**. **(C)** A tabulated list displays samples along with their corresponding gender, population, and super-population information, along the variant off-target counts specific to each sample, its population, and super-population, as well as the number of PAM creation events observed for each sample. **(D)** A glimpse into the "Show Targets" view for subject HGDP01211 is provided. Users can choose which columns to display by utilizing the "Toggle Columns" button located at the top of the table. Presented here are the initial seven columns, furnishing sequence and positional insights. **(E)** The table displays the candidate off-target(s) within the specified region. Here, the table's initial seven columns are displayed, providing sequence and positional details.

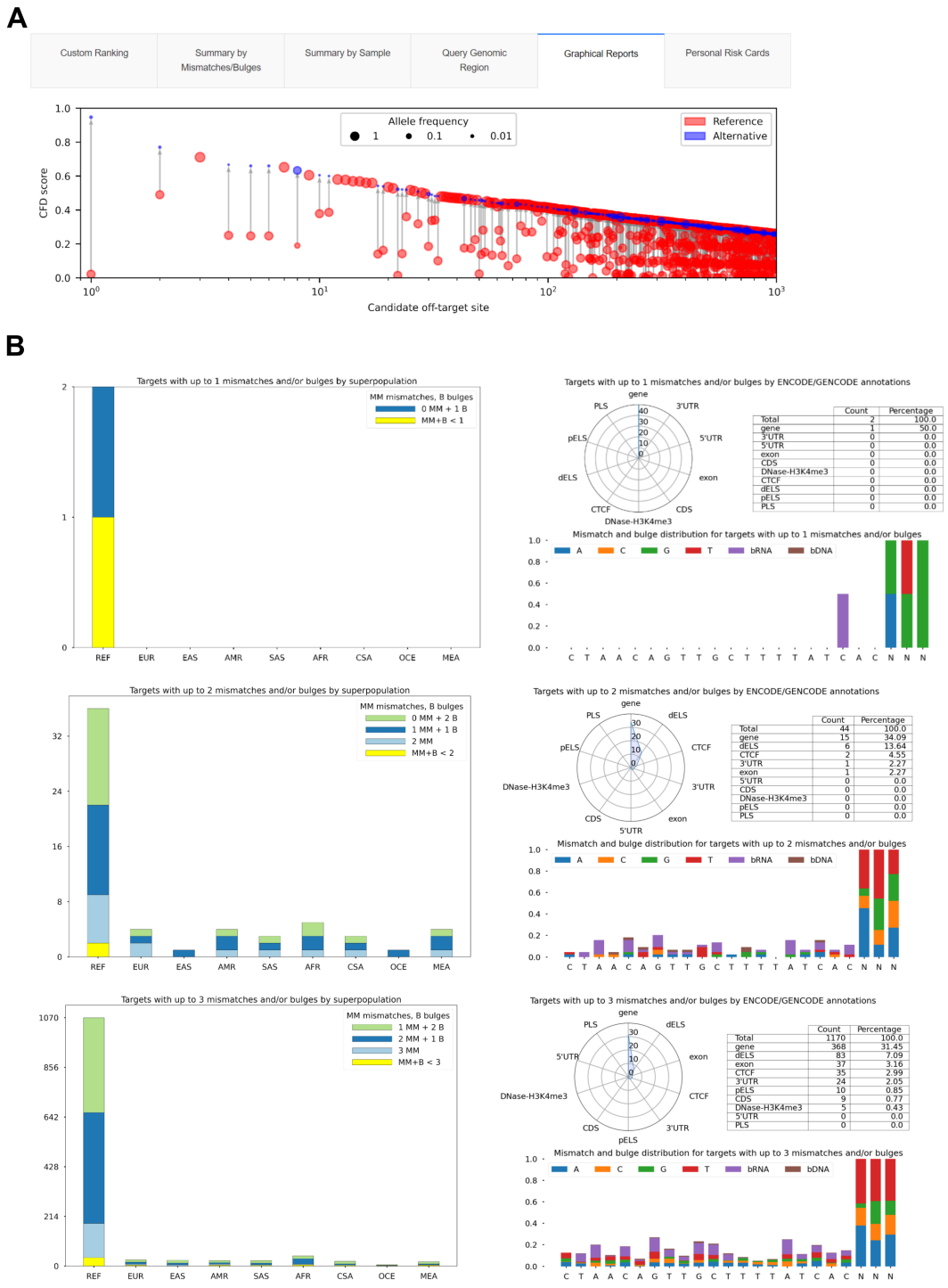
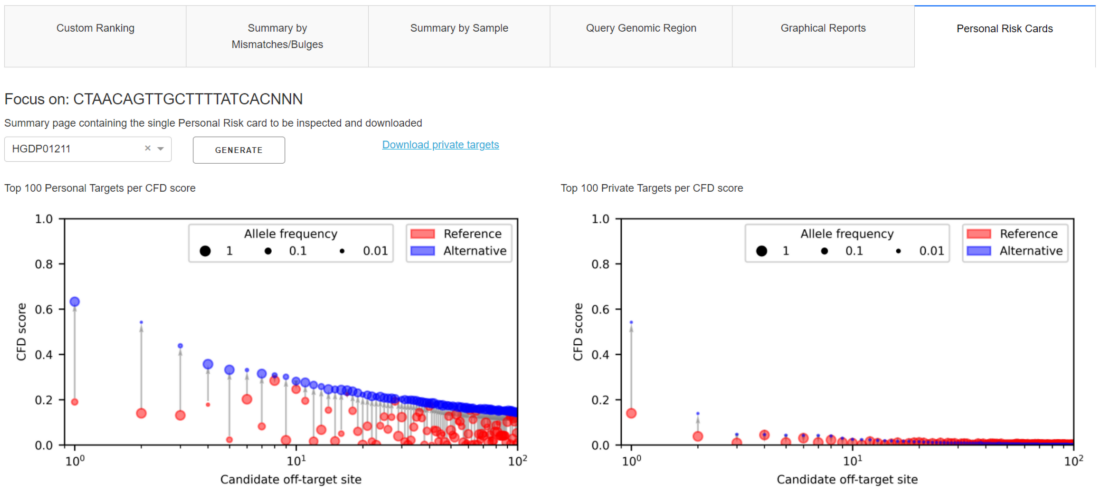


Figure 9.34. CRISPRme Graphical Reports. (A) CRISPRme comparison using sg1617, `NNN` PAM, 6 mismatches, and 2 DNA/RNA bulges, assessed on hg38 reference genome along with 1KGP and HGDP variants. A stem plot displays the CFD score distribution for candidate off-targets, sorted by descending CFD score. When a genetic variant enhances the CFD score, both alternative (blue) and reference (red) allele CFD scores at the same locus are shown, with circle size proportional to allele frequency. **(B)** On the left, stacked bar plots summarize the number of candidate off-targets for each mismatch + bulge category across super-populations in the input variant data. On the right, a radar chart illustrates the percentage of off-targets categorized by specific genomic annotations relative to the total count, accompanied by a table detailing the precise number of off-targets falling into each category (note: an off-target can belong to multiple categories). Additionally, a motif plot exhibits the distribution of mismatches and bulges concerning the spacer+PAM sequence.

all SNP variants across different samples, including multiallelic sites with three or more observed alleles. Indels, however, are handled differently. For each indel, CRISPRme generates a simulated chromosome containing the variant DNA sequence along with 50 surrounding nucleotides (25 on each side). The association between samples and variants is facilitated through a hash table, enabling efficient querying

A



B

Personal	PAM Creation	Private
374323	0	2636

TOGGLE COLUMNS

Spacer+PAM	Chromosome	Start_coordinate_(highest_CFD)	Strand_(highest_CFD)	Aligned_spacer+PAM_(highest_CFD)	Aligned_protospacer+PAM_REF_(highest_CFD)	Aligned_protospacer+PAM_ALT_(highest_CFD)
CTAACAGTTGCTTTTATCACNNN	chr3	99137592	+	CTAACAGTTGCTTTTATCACNNN	aT-ACAGcTtaTTTTATCACCAG	aT-ACAGcTtaTTTTATCACCGG
CTAACAGTTGCTTTTATCACNNN	chr3	55688983	-	CTAACAGTTGCTTTTATCACNNN	g-AAgttTcGCTTTcAcCACTGG	g-AAgttTcGCTTTTAcCACTGG
CTAACAGTTGCTTTTATCACNNN	chr6	87185387	+	CTAACAGTTGCTTTTATCACNNN	g-AgCA-cTGCTTTcgTtACCGG	g-AgCA-cTGCTTTcATtACCGG
CTAACAGTTGCTTTTATCACNNN	chr15	31234527	+	CTA-ACAGTTGCTTTTATCAC-NNN	CaACACAcaTGtaTaTATCACATAG	CaACACAcaTGCaTaTATCACATAG
CTAACAGTTGCTTTTATCACNNN	chr6	142941722	-	CTAACAGTTGCTTTTATCACNNN	tT-AaAGT-GaTaTTATaAaAAG	tT-AaAGT-GaTaTTATaAaAGG
CTAACAGTTGCTTTTATCACNNN	chr5	4322940	-	CTAACAGTTGCTTTTATCACNNN	CT-tCAGTTaaaTTTATtcCCAG	CT-tCAGTTaaaTTTATtcCCGG
CTAACAGTTGCTTTTATCACNNN	chr4	127257387	+	CTA-ACAGTTGCTTTTATC-ACNNN	taACAtgGcaGCTTTTATCCAaTGG	taACAtAGcaGCTTTTATCCAaTGG
CTAACAGTTGCTTTTATCACNNN	chr4	8252994	+	CTAACAGTTGCTTTTATCACNNN	CT-cCtaT-GCTcTTATCcCTGG	CT-tCtaT-GCTcTTATCcCTGG

Figure 9.35. CRISPRme Personal Risk Card. The example shows data for subject HGDP01211. **(A)** On the left: Plot illustrating potential variant off-targets specifically associated with the chosen sample. On the right: Plot displaying potential off-targets exclusive to the selected sample. **(B)** The top table provides a summary of personal variant off-targets, instances of PAM creation, and private off-targets. The bottom table enumerates all private targets identified for the selected sample. Here, the table's initial seven columns are displayed, comprising sequence and positional information.

to identify which samples contain a given SNP/indel. Notably, this approach removes the need for users to manually generate multiple variant-containing genomes, simplifying the operation and automating the association of targets with specific haplotypes. The haplotype-aware search implemented in CRISPRme ensures that only off-targets present in real genomes are reported, thereby eliminating the chance of reporting spurious off-targets. This operation considers the potential presence of multiple variants in any given off-target, generating and analyzing various combinations efficiently. An ad-hoc algorithm has been developed to process candidate off-targets in polynomial time, grouping variants by sets of individuals sharing them to avoid redundant computations. Additionally, the algorithm accommodates both phased and unphased VCFs as input. When phased VCFs are utilized, CRISPRme generates haplotypes for the positive and negative strands, considering the position of the analyzed variant, even in cases where multiple variants are present in the same target. This optimization significantly reduces computation time and ensures the exclusion of artificial off-targets from the final report.

CRISPRme off-target nomination: indexing, search, and analysis procedures

To perform efficient target searches, CRISPRme employs an indexed reference or enriched genome and bit-wise matching operations (**Figure 9.36**). This index employs a tree-based data structure encoding all potential candidate off-targets. The tree data structure allows for swift retrieval of both reference and variant-enriched sites. However, regions of the genome containing ambiguous bases (e.g., poorly assembled or repetitive regions represented by `Ns`) are excluded from the search operation for accuracy. For this task CRISPRme employs CRISPRitz's search engine (Cancellieri *et al.*, 2020). To identify all genome regions matching the input PAM, the algorithm constructs a compact deterministic automaton (**Figure 9.36 (A)** and **Algorithm 7**). This machine scans and enumerates all potential PAM matches in linear time ($O(|\text{PAM}|)$). The automaton allows to reduce time complexity of targets search to the

Algorithm 6: CRISPRme reference genome enrichment algorithm

Data: G, VCFS
Result: E, I, hapMap

$E \leftarrow G$; /* Genome enriched with SNPs */
$I \leftarrow G$; /* Genome enriched with indels */
hapMap $\leftarrow \emptyset$;
for *vcf* ***in*** *VCFS* **do**
  $\text{pos}_{\text{prev}} \leftarrow 1$;
  for v ***in*** *vcf* **do**
    ref, alt, id, pos, samples $\leftarrow$ parseVariant(v) ; /* Recover variants VCF fields */
    if v *is SNP* **then**
      $E_{\text{pos}_{\text{prev}}:\text{pos}} \leftarrow$ enrichSequence($G_{\text{pos}_{\text{prev}}:\text{pos}}, v, \text{pos}$) ; /* Encode ref and alt alleles using IUPAC symbols */
    else
      $I_{\text{pos}_{\text{prev}}:\text{pos}} \leftarrow$ enrichSequence($G_{\text{pos}_{\text{prev}}:\text{pos}}, v, \text{pos}$) ; /* Encode ref and alt alleles using IUPAC symbols */
    hapMap[id] $\leftarrow$ samples;
return *E, I, hapMap*

minimum number of comparisons, with a complexity of $O(N)$, where $N = |G|$ and G is the input genome. Notably, this algorithm scans each nucleotide in the genome only once. In CRISPRitz, the automaton is structured as a rooted directed tree, often referred to as a *trie* (Bodon and Rónyai, 2003), enriched with additional connections among nodes. Each path from the root to a leaf corresponds to a PAM sequence (**Figure 9.36 (A)**). Therefore, PAM search is executed by reading one nucleotide at a time and matching it with a node of the graph. When traversing trie's nodes, the algorithm advances along the corresponding path by moving one node forward in the tree and one nucleotide in the genome. This traversal iteratively continues as long as the current genome nucleotide matches the current node. If it is not possible to match the next available node, the algorithm backtracks to the parent node of the graph by following special failure edges (**Figure 9.36 (A)**), and the search resumes from the parent node. Importantly, the paths following the failure edges reduce the number of comparisons by allowing the search to restart from the longest common substring that could be matched at that stage. Every time the graph visit reaches a leaf node, the algorithm records the current index (i.e., the nucleotide position in the reference genome) as it may represent the starting position of a candidate target site. The result is the complete list of indices of candidate off-target sites. The indices are saved in two arrays that differentiate between positive and negative strands. For each identified PAM, the upstream or downstream sequences, depending on the Cas protein and strand orientation, adjacent to the PAM occurrence are extracted. The candidate target sequence length equal the width of the input gRNA. All identified candidate off-target sequences are gathered, sorted in lexicographical order, and encoded using four-bit notation (**Figure 9.36 (B)**). To efficiently consider bulges during candidate targets search, CRISPRitz harnesses a ternary search tree (TST) data structure (Bentley and Sedgewick, 1998) to index the genome. In a TST, each node represents a nucleotide and can have at most three children: left, center, and right. The TST construction involves inserting one candidate target sequence at a time, with insertion carried out through a recursive search. Beginning with the first nucleotide of the candidate target sequence and the TST root, the algorithm compares if the two nucleotides match. If they do, the comparison proceeds to the next nucleotide on the candidate target sequence, descending in the TST through the center child node without adding a new node to the TST. Conversely, if the nucleotide pair between the center child node and the nucleotide from the candidate off-target sequence do not match, the algorithm verifies the lexicographical order of the nucleotide on the candidate target. If it is smaller than the current nucleotide in the TST node, the search recursively continues to the left child. Conversely, if the candidate target nucleotide is lexicographically greater, the search recursively continues to the right child. Subsequently, if the node in the chosen direction does not exist, a new node is inserted into the TST with the current nucleotide of the candidate target sequence. Each genomic sequence inserted in the TST is two characters longer than the PAM sequence to accommodate searches with up to two DNA bulges (**Figure 9.36 (C)**). Genome-wide target search is a computationally intensive task (**Figure 9.36 (B)**). Nonetheless, CRISPRitz addresses this challenge by employing a 4-bit encoding to

Algorithm 7: CRISPRme PAM indexing algorithm

```
Data: E, PAM
Result: P
P ← ∅ ;                                                    /* Enriched genome */
A ← initializeAutomaton(PAM) ;        /* initialize PAM matching automaton (refer to
 figure 9.35 (A)) */
count ← 0;
for i in |E| do
    nuc ← E[i];
    s ← recoverState(nuc, A);
    if s = success then
        count + +;
        if count = |PAM| then
            P ← addPosition(P, i);
            count = 0 ;                                    /* Reset automaton states */
    else
        if count > 0 then
            count − −;
return P
```

represent each IUPAC nucleotide, enabling efficient bitwise operations to handle potentail mismatches between the gRNA and candidate target sequences (**Figure 9.36 (B)**). Therefore, if a gRNA aligns perfectly to a candidate sequence without mismatches, its position indicates the starting position of an on-target site. Conversely, if mismatches are present and their number is lower than the user-defined threshold, the position is reported as a candidate off-target. However, CRISPRitz employs a different strategy to handle potential bulges in the candidate target sequences. This algorithm leverages the TST index significantly reducing the computational complexity to $O(k)$ operations required to search for a sequence s, where $|s| = k$, on a TST containing n candidate off-targets. The search process is executed through a function that recursively traverses the TST. Beginning at the root, the algorithm explores all branches of the TST, determining whether nucleotide comparisons result in a match, mismatch, or bulge (DNA/RNA). Notably, when bulges are permitted, CRISPRitz may yield duplicate results, representing the same target with varying numbers of mismatches and/or bulges, or at different positions. In the mismatch-bulge search type, traversal along a branch may terminate at a leaf if the bulge threshold allows for the exploration of all nodes within the branch. This occurrence is feasible because DNA bulges are treated as supplementary characters that can align with supplementary characters on the candidate off-target site. The mismatch-bulge search type ceases when a branch of the TST is fully explored (i.e., reaching a leaf), when a partial exploration of the TST exceeds the specified threshold for mismatches or bulges, or when the traversal encounters a number of nodes equal to the length of the guide (**Figure 9.36 (C)**). Importantly, CRISPRme incorporates individual-specific analysis by extracting haplotype-specific off-targets and their corresponding samples from the IUPAC-encoded enriched genome. For each site, CRISPRme provides off-target potential scores. In cases where multiple alignments exist at a given site (e.g., due to RNA/DNA bulges), CRISPRme reports the alignment with the highest score, ensuring accuracy in off-target prediction. Currently, CRISPRme employs two widely used scoring functions, CFD (Doench *et al.*, 2016) and CRISTA (Abadi *et al.*, 2017) scores. These scoring functions are chosen for their efficiency in computation across thousands of sites, their ability to handle bulges, and their proven effectiveness in predicting off-targets, as validated by deep sequencing studies (Bao *et al.*, 2021). While CRISPRme currently supports these scoring functions, it has the potential to extend its support to include other predictive off-target scores in the future. Along individual CFD scores, CRISPRme provides a global score (CFD specificity) for each input gRNA:

$$\text{CFD}_{\text{specificity score}} = 100 \cdot \frac{100}{100 + \sum_{t_i} \text{CFD}(t_i)}$$

where t_i is one of the enumerated off-targets returned during the search. Its value falls within $(0, 100]$, where a gRNA that doesn't yield any predicted off-targets given the search parameters receives a perfect CFD specificity score of 100. It's important to note that the CFD score is specifically designed for SpCas9,

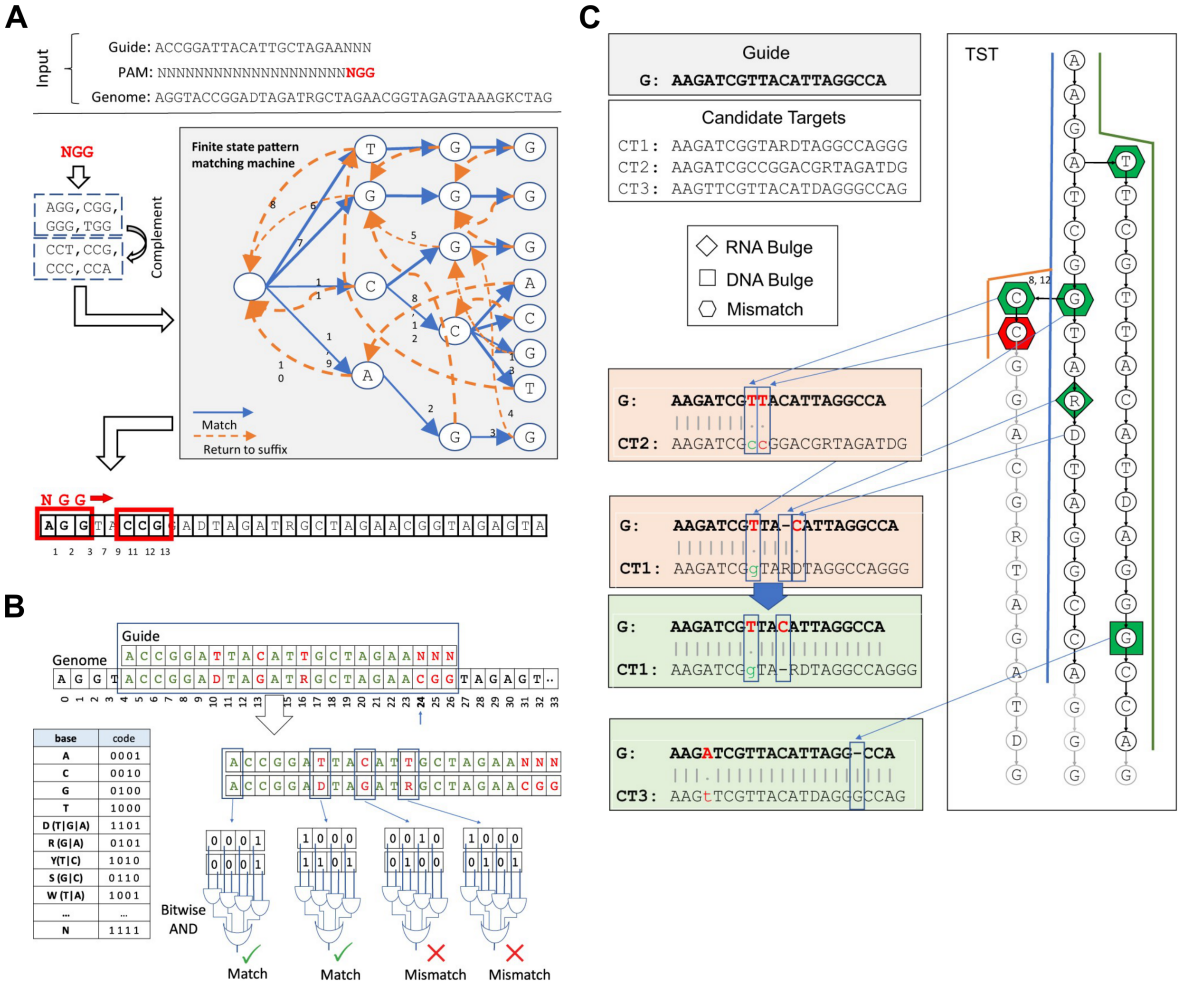


Figure 9.36. CRISPRme Off-target nomination indexing and search engine. **(A)** The PAM search process involves seeking the PAM sequence in the genome (`NGG` in the example). In the illustrated example, the process starts by matching of the base `A` at position 0 in the genome G with the root children `T`, `G`, `C`, and `A` of the pattern-matching machine. To illustrate, let's examine the initial 11 transitions of the automaton, which correspond to the identification of three candidate targets (at positions 0, 5, and 6). **(B)** The mismatch-aware guide matching strategy involves encoding characters using 4-bit notation. At a specific genomic position, the algorithm compares the gRNA's characters with those of G employing bitwise operations. In the illustrated example, gRNA matching starts from index 24. **(C)** Considering an example of gRNA matching where up to 1 mismatch and 1 RNA/DNA bulge are permitted. The search starts by exploring the left-most path of the tree, corresponding to the candidate sequence `CT2`. Upon encountering the second mismatch (`T` versus `C`), the algorithm confirms that no bulges are permissible and halts the traversal of the `CT2` path. Subsequently, the algorithm backtracks to the previous branch, `G`, indicating the first mismatch. It proceeds along the `CT1` path, where it determines that while the second mismatch cannot be treated as a DNA bulge, it can be regarded as an RNA bulge. This traversal concludes with the identification of `CT1` as an off-target with one mismatch and one RNA bulge. Similarly, the algorithm retreats to the previous branch, `A`, and follows the `CT3` path, thereby identifying `CT3` as an off-target with one mismatch and one DNA bulge (adapted from Cancellieri *et al.* (2020)).

and therefore, CRISPRme does not compute CFD scores for editors different from SpCas9. In such cases, a value of -1 is assigned to each off-target. Similarly, the CRISTA specificity score is calculated as:

$$\text{CRISTA}_{\text{specificity score}} = 100 \cdot \frac{100}{100 + \sum_{t_i} \text{CRISTA}(t_i)}$$

Crucially, CRISPRme employs a constructed hash table that maps samples to variants. This enables CRISPRme to filter results post-initial search, ensuring that only targets matching haplotypes present in the populations and corresponding individuals are reported. Since bulges may create multiple alignments for a genomic region, CRISPRme generates two lists of candidate off-target sites. To address this issue, CRISPRme merges off-target sites within a 3 bp range into a single entry per genomic region. The surviving off-targets are selected and sorted based on the user-defined criterion. For completeness, CRISPRme reports also the excluded targets in a different optional report.

CFD scoring function in CRISPRme

To evaluate targets using Cutting Frequency Determination (CFD) score (Doench *et al.*, 2016), CRISPRme employs a matrix based on empirical data developed by the authors. We recall that CFD score was specifically developed to evaluate spCas9 cutting potential. This matrix encompasses all possible pairs of mismatches between a RNA and DNA sequence of 20 nucleotides in length. Each entry in the matrix file signifies a pairing between a RNA and DNA nucleotide. For instance, the entry "rA:dG,20', F0.22" denotes that when the RNA nucleotide `A` pairs with the DNA nucleotide `G` at position 20, it receives a score of 0.227. This score is then multiplied with values corresponding to any additional mismatches present, yielding the overall CFD score for the off-target sequence. If a sequence has only one mismatch, its final score will simply be the score of that mismatch. Thus, in the aforementioned example, the CFD score of an off-target with a single mismatch (20`A`>`G`) would be 0.227. Furthermore, the matrix encompasses scores for bulges, denoted by "-". An example entry in the matrix representing a bulge is "sS'r-:dA,2', F0.692," which signifies that a RNA bulge pairing with the DNA nucleotide `A` at position 2 has a score of 0.692. Notably, bulges are not permitted in the first position of the RNA or DNA sequence. When dealing with targets featuring DNA bulges, computationally, they are reported with a longer sequence compared to the original spacer. To ensure consistency in calculating the CFD score for off-targets with DNA bulges, we compute the score based solely on the last 20 nucleotides of the protospacer, as intended by the original CFD scoring method. Moreover, the scoring process extends to include the PAM nucleotides. As the CFD score was derived from SpCas9 data, the matrix exclusively contains scores for `NGG` and all possible combinations of the last two positions of the PAM. Let's delve into an example of an off-target with a DNA bulge:

Spacer: `CTAACAGTTGCTT-TTATCACNNN`

Protospacer: `CTAACAGcTGCTTCTTATCACCTC`

In this instance, the off-target incorporates one mismatch (in lowercase) and one DNA bulge (aligned with the gap in the spacer sequence). When computing its CFD score, we exclude the first nucleotide of the protospacer since the protospacer is one nucleotide longer than the spacer. Each mismatch and bulge encountered is scored according to the matrix, and the values of each pair are multiplied to derive the final CFD score. For instance, the first mismatch is rT:dC at position 7, yielding a score of 0.588. Then, we come across the bulge r-:dC at position 13, which scores 0. The process continues until reaching the PAM. In this case, the nucleotide pair `TC` has a score of 0. Subsequently, we multiply all the values accumulated during the process. Hence, the final score is calculated as 0.588 * 0 * 0, resulting in an off-target CFD score of 0 for this example. Scoring off-targets with RNA bulges is relatively straightforward since the spacer sequences remain unchanged despite the presence of bulges. For instance:

Spacer: `CTAACAGTTGCTTTTATCACNNN`

Protospacer: `CTAACAGTTGCTTTTAT-ACGTG`

In this scenario, the off-target solely contains a RNA bulge (in the protospacer gap). Upon scanning the sequence, we identify the bulge at position 18, utilizing matrix entry "rC:d-,18" with a score of 0. Consequently, the final CFD score for the off-target will be 0.

9.1.4 Comparing CRISPRme with available off-target nomination tools

While several tools exist for enumerating CRISPR-Cas off-targets, only three previous studies have reported computational strategies to assess off-target potential in the presence of genetic variants (Lessard *et al.*, 2017; Scott and Zhang, 2017; Fennell *et al.*, 2021). Notably, among these, only crispRtool (Lessard *et al.*, 2017) and CALITAS (Fennell *et al.*, 2021) provide general command line software. However, upon testing, CALITAS (v1.0) was found to be unable to use phasing information and does not directly support publicly available 1KGP VCFs as input. This limitation significantly restricts the tool's utility for conducting comprehensive and general variant-aware analyses. Hence, we opted to narrow down our comparison to CRISPRme (v1.7.7) and crisprRtool (v2.0.5). To ensure a fair assessment, we evaluated both tools based on available features and running times using identical hardware specifications (AMD Ryzen Threadripper 3970X 32-Core Processor clocked at 2.2 GHz with 124 GB RAM). Our tests were conducted using the 1617 sgRNA (see **Section 9.2**), `NGG` PAM, variants sourced from 1KGP, and a variable number of mismatches and bulges. In summary, crisprRtool initially incorporates variants (limited to SNPs) into the reference genome using IUPAC notation, similarly to CRISPRme. It then proceeds

to scan the input gRNA(s) on the variant-enriched genome, generating a list of potential on- and off-targets with IUPAC nucleotides. The tool does offer an option to search each VCF file individually to resolve haplotypes (including SNPs and indels) of the identified off-targets. However, this requires manual editing and execution of a script for each VCF file, adding complexity to the process. Moreover, the search operation in crisprRtool is constrained by certain limitations, including a maximum of 5 allowed mismatches, lack of consideration for bulges, and restricted flexibility regarding the PAM location relative to the protospacer (limited to the 3' end). With the specified settings, crispRtool required ∼9 hours to execute a non-haplotype-resolved search with 5 allowed mismatches. Furthermore, a haplotype-resolved search restricted to chromosome 1, using 1KGP variants (6 million SNPs and indels), required ∼37 hours. Extrapolating this duration to encompass all other chromosomes, the entire search operation would exceed 300 hours. However, it's noteworthy that the results obtained from crispRtool lack the comprehensive graphical reports and textual summaries offered by CRISPRme, which would encompass results from all chromosomes. In contrast, CRISPRme capitalizes on an efficient genome index and auxiliary data structures, constructed just once during installation, which takes ∼4 hours for `NGG` PAM and ∼12 hours for `NNN` PAM. Alternatively, users can download these structures directly from our complete test package. Leveraging these resources, CRISPRme can execute a haplotype-aware search for a given gRNA across the entire genome with 5 allowed mismatches in approximately 1 hour. Furthermore, a haplotype-resolved search encompassing the entire genome, with up to 6 mismatches and 2 DNA/RNA bulges, requires only 2 hours (excluding the guide-independent indexing operation described above) and provides a comprehensive summary report.

9.2 A common allele-specific off-target for a gRNA in the clinic

We tested CRISPRme using a specific gRNA (#1617) designed to target a GATA1 binding motif located at the +58 erythroid enhancer of BCL11A (Canver *et al.*, 2015; Wu *et al.*, 2019). Recent clinical observations described two patients, one with sickle cell disease (SCD) and another with β-thalassemia, both treated with autologous gene-modified hematopoietic stem/progenitor cells (HSPCs) edited with Cas9 and this gRNA. Remarkably, these patients exhibited sustained increases in fetal hemoglobin levels, achieved transfusion independence, and showed an absence of vaso-occlusive episodes (in the case of the patient with SCD) post-therapy. Notably, neither this study nor prior preclinical investigations utilizing the same gRNA (#1617) have revealed any evidence of off-target editing in treated cells. This conclusion is drawn from the absence of off-target sites identified through bioinformatic analysis of the human reference genome and empirical analysis of *in vitro* genomic cleavage potential (Frangoul *et al.*, 2021; Wu *et al.*, 2019; Demirci *et al.*, 2019). CRISPRme analysis revealed a notable predicted off-target site within an intronic sequence of CPS1 (**Figure 9.28 (C-D)**), a genomic locus known to exhibit common genetic variation (as evidenced by a SNP with a minor allele frequency (MAF) of $\geq 1\%$). The CFD scores employed in our analysis range from 0 to 1, where a perfect match to the on-target site is assigned a score of 1. Specifically, the alternative allele rs114518452-`C` introduces a `TGG` PAM sequence, which aligns optimally with SpCas9, thereby generating a potential off-target site with three mismatches and a CFD score ($CFD_{alt} = 0.95$) approaching that of the on-target site (**Figure 9.28 (E)**). In contrast, the reference allele rs114518452-`G` disrupts the PAM to `TGC`, resulting in a markedly reduced predicted cleavage potential ($CFD_{ref} = 0.02$). In gnomAD v3.1, the rs114518452-`C` allele exhibits an overall MAF of 1.33%, with variable frequencies across different populations, including 4.55% in African/African American, 0.91% in Other, 0.66% in Latino/Admixed American, 0.12% in South Asian, 0.01% in European (non-Finnish), and 0.00% in all remaining represented populations (**Figure 9.28 (F)** and **Table 9.11**). To comprehensively assess potential off-target effects beyond those predicted by 1KGP variants, we scrutinized HGDP variants derived from whole-genome sequences of 929 individuals spanning 54 diverse human populations. Our analysis identified 249 candidate off-targets for gRNA #1617, each exhibiting a CFD score of ≥ 0.2, with HGDP-derived CFD scores surpassing those observed for either the reference genome or 1KGP variants by at least 0.1 (**Figure 9.37 (A)** and **9.38**). These additional variant off-targets, not found in the 1KGP dataset, were observed across all superpopulations, with the highest frequency in the African superpopulation (**Figure 9.37 (B)**). Among these variant off-targets, 229 (92.0%) were unique to a superpopulation, with 172 (69.1%) being private to a single individual (**Figure 9.37 (C)**). Moreover, individual-focused searches, such as the analysis of HGDP01211, a member of the Oroqen population within the East Asian superpopulation, revealed that while most variant off-targets (with higher CFD scores than the reference) were due to variants also present in 1KGP (n = 32,369, 90.4%), a subset originated from variants shared with other HGDP individuals but not present in 1KGP (n = 3,177, 8.9%). A small fraction of off-targets were private to the individual (n = 234, 0.7%) (**Figure 9.37**

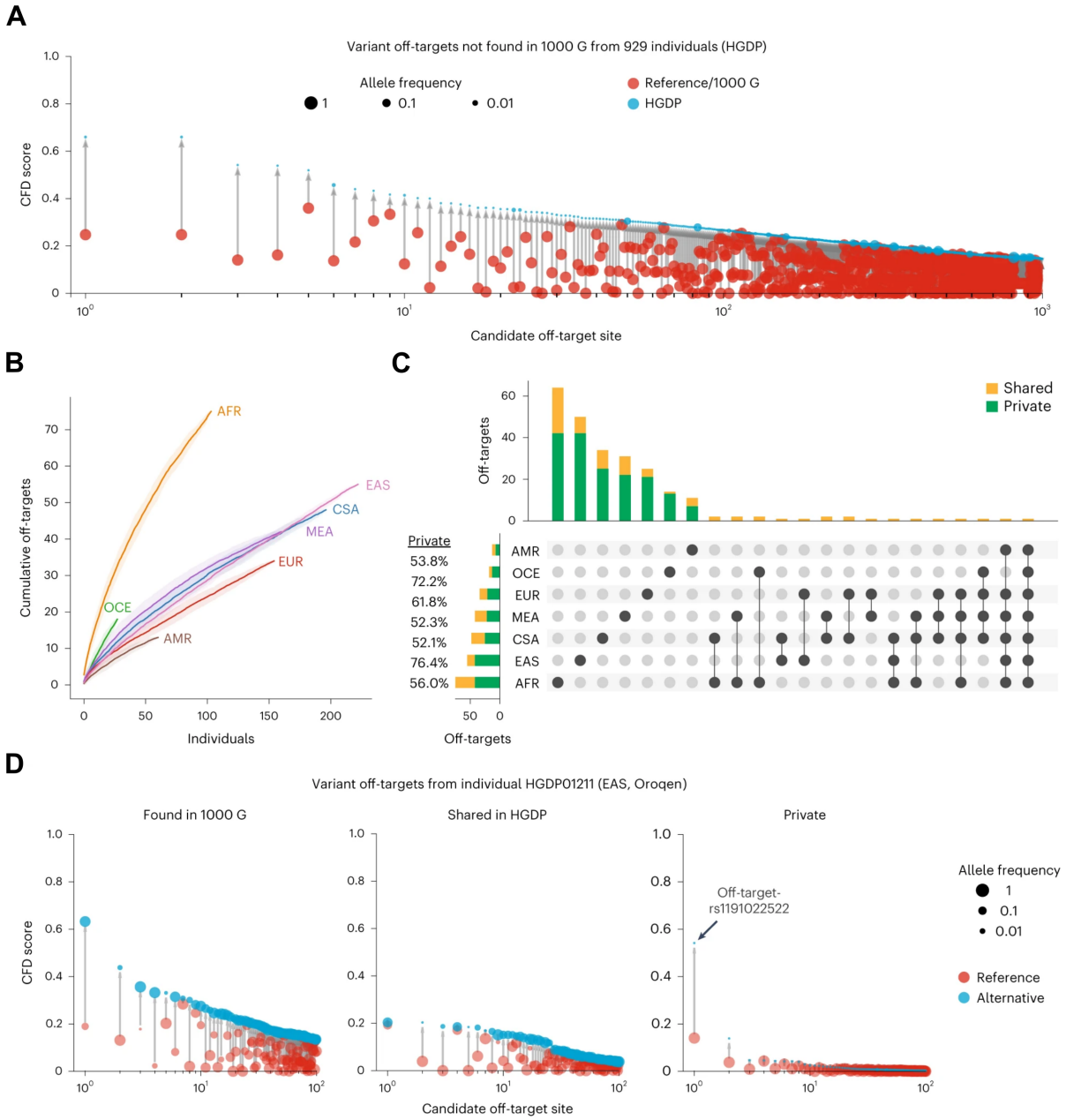
Population	Allele count	Allele number	Number of Homozygotes	Allele frequency
African/African-American	1,882	41,386	39	0.0455
Other	19	2,090	0	0.0091
Latino/Admixed American	100	15,246	0	0.0066
South Asian	6	4,830	0	0.0012
European (non-Finnish)	10	67,992	0	0.0001
European (Finnish)	0	10,612	0	0.0000
Amish	0	912	0	0.0000
East Asian	0	5,170	0	0.0000
Middle Eastern	0	316	0	0.0000
Ashkenazi Jewish	0	3,470	0	0.0000
XX	1,088	77,776	23	0.0140
XY	929	74,248	16	0.0125
Total	**2,017**	**152,024**	**39**	**0.0133**

Table 9.11. Complete population frequencies for rs114518452 from gnomAD v3.1.

(D)). Among these private off-targets was one generated by a variant (rs1191022522, 3-99137613-A-G, gnomAD v3.1 MAF 0.0053%), where the alternative allele produces a canonical NGG PAM that increases the CFD score from 0.14 to 0.54 (**Figure 9.37 (D-E)**). To experimentally validate the top predicted off-target identified by CRISPRme, we selected a CD34+ HSPC donor of African descent who was heterozygous for rs114518452-C, the variant predicted to induce the most significant increase in off-target cleavage potential (**Figure 9.28 (C-F)**). We performed ribonucleoprotein (RNP) electroporation using a gene editing protocol designed to preserve engrafting HSC function. Amplicon sequencing analysis revealed 92.0 ± 0.5% indels at the on-target site and 4.8 ± 0.5% indels at the off-target site. Notably, indels were exclusively detected at the alternative PAM-creating allele among reads spanning the variant position, with no indels observed at the reference allele (**Figure 9.39 (A-C)**). This observation suggests a 9.6 ± 1.0% off-target editing rate specifically targeting the alternative allele. In an additional analysis involving six HSPC donors homozygous for the reference allele rs114518452-G/G, no indels (0.00 ± 0.00%) were detected at the off-target site. This finding strongly indicates that off-target editing was strictly restricted to the alternative allele (**Figure 9.39 (D)**).The on-target BCL11A intronic enhancer site is located on chr2p, while the off-target-rs114518452 site resides on chr2q within an intron of a noncanonical transcript of CPS1. Utilizing inversion PCR, we identified inversion junctions indicative of ∼150 Mb pericentric inversions between BCL11A and the off-target site, specifically observed in edited HSPCs carrying the alternative allele (**Figure 9.40 (A-B)**). Subsequent deep sequencing of the inversion junction confirmed that inversions were solely present on the alternative allele in the heterozygous cells (**Figure 9.40 (C-D)**). Droplet digital PCR analysis revealed these inversions to occur at an allele frequency of 0.16 ± 0.04% (**Figure 9.40 (E)**). Different high-fidelity Cas9 variants hold promise for enhancing the specificity of gene editing, albeit potentially at the expense of reduced efficiency. Employing a HiFi variant, 3xNLS-SpCas9 (R691A), in cells heterozygous for the rs114518452-C variant yielded 82.3 ± 1.6% on-target indels, with only 0.1 ± 0.1% indels detected at the rs114518452-C off-target site. This represents a ∼48-fold reduction compared to editing with SpCas9 (**Figure 9.39 (C)**). Notably, inversions were not detected following editing with HiFi-3xNLS-SpCas9 (**Figure 9.40 (B)** and **(E)**).

9.3 Allele specific off-target potential of additional gRNAs

To explore the prevalence of off-target potential linked to alternative alleles, we assessed an additional set of 13 gRNAs commonly used in clinical development or SpCas9-based gene editing (Xu *et al.*, 2017, 2019a; Stadtmauer *et al.*, 2020; Gillmore *et al.*, 2021; DeWitt *et al.*, 2016; Xu *et al.*, 2019b; Métais *et al.*, 2019; Tsai *et al.*, 2015; Zeng *et al.*, 2020a; Musunuru *et al.*, 2021), along with 6 gRNAs for non-SpCas9-based editing, such as SaCas9 and Cas12a (Xu *et al.*, 2019b; Chu *et al.*, 2021; Newby *et al.*, 2021; Maeder *et al.*, 2019; De Dreuzy *et al.*, 2019). Our analysis, incorporating data from the 1KGP and HGDP genetic variant datasets, revealed that 18% (with a 95% confidence interval of 13–23%) of all identified off-target sites were attributed to alternative allele-specific off-targets. Interestingly, most of these alternative allele-specific off-targets were associated with rare variants (MAF < 1%), although candidate off-targets linked to common variants were also identified for each gRNA (**Figure 9.41 (A)**). Notably, none of these alternative allele-specific off-target sites were previously reported in the original manuscripts detailing the editing strategies and off-target analyses. CRISPRme generates visual representations specifically



Sequence	Alignment	Variant ID	CFD	MAF	Annotation
Spacer + PAM	CTAACAGTTGCTTTTATCACNNN				
Reference	aT-ACAGcTtaTTTTATCACCAG		0.140		intergenic:*DCBLD2*
Alternative	aT-ACAGcTtaTTTTATCACC**G**G	rs1191022522 3-99137613-A-G	0.542	0.00054	

Figure 9.37. CRISPRme offers comprehensive analysis of CRISPR-Cas gene editing's off-target potential, encompassing both population-wide and private genetic diversity. (A) CRISPRme analysis was performed using variants from HGDP comprising whole-genome sequencing of 929 individuals from 54 diverse human populations. Off-targets associated with HGDP variants exhibiting higher CFD scores compared to the reference genome or 1KGP were plotted and categorized by CFD score. HGDP variant off-targets are visualized in blue, while reference or 1KGP variant off-targets are depicted in red. **(B)** Cumulative distribution plot illustrating HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD by ≥ 0.1 per superpopulation. To generate robust statistics, individual samples from each of the seven superpopulations were randomly shuffled 100 times. calculating the mean and 95% confidence interval (shading around lines). **(C)** Intersection analysis of HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD of ≥ 0.1. Shared variants (black) are found in two or more HGDP samples whereas private variants (gray) are exclusive to a single sample. **(D)** CRISPRme analysis of a single individual (HGDP01211) showing the top 100 variant off-targets from each of the following three categories: shared with 1KGP variant off-targets (left panel), higher CFD score compared to reference genome and 1KGP but shared with other HGDP individuals (center panel) and higher CFD score compared to reference genome and 1KGP with variant not found in other HGDP individuals (right panel). In the center and right panels, the term "reference" denotes the CFD score derived from the reference genome or 1KGP variants. **(E)** The top predicted private off-target site associated with sample HGDP01211 is an allele-specific off-target. In this scenario, the rs1191022522-G minor allele replaces a noncanonical `NAG` PAM with a canonical `NGG` PAM sequence.For ease of visual alignment, the spacer is represented as a DNA sequence.

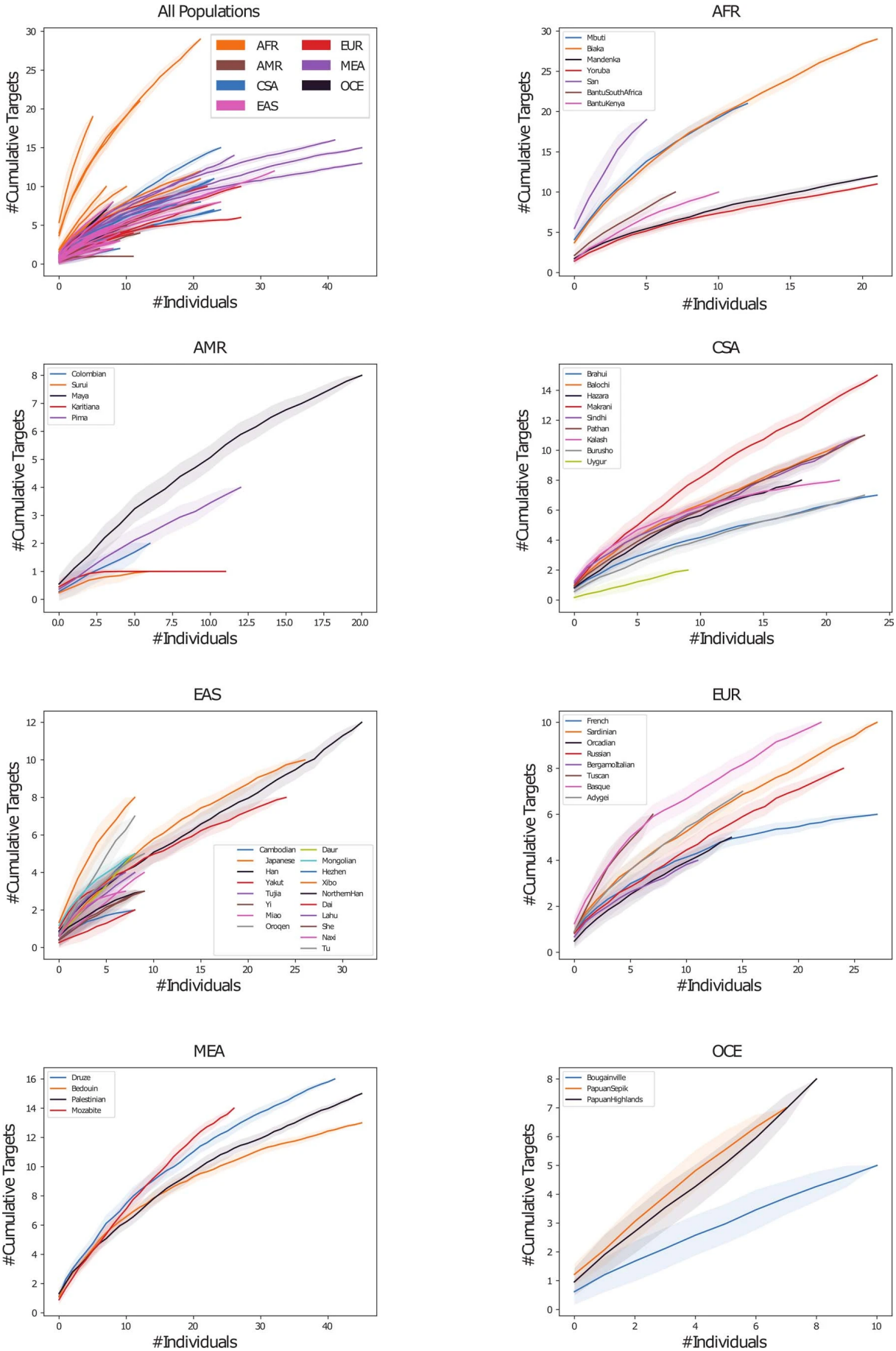


Figure 9.38. HGDP superpopulation distribution plots. HGDP Variant Off-Targets with CFD ≥ 0.2 and Increased CFD of ≥ 0.1. Individual samples from each of the seven superpopulations were randomly shuffled 100 times to compute the mean and 95% confidence interval. The first panel depicts the distribution across all 54 discrete populations, with colors indicating superpopulations. Additional seven panels display the distribution of discrete populations within each listed superpopulation.

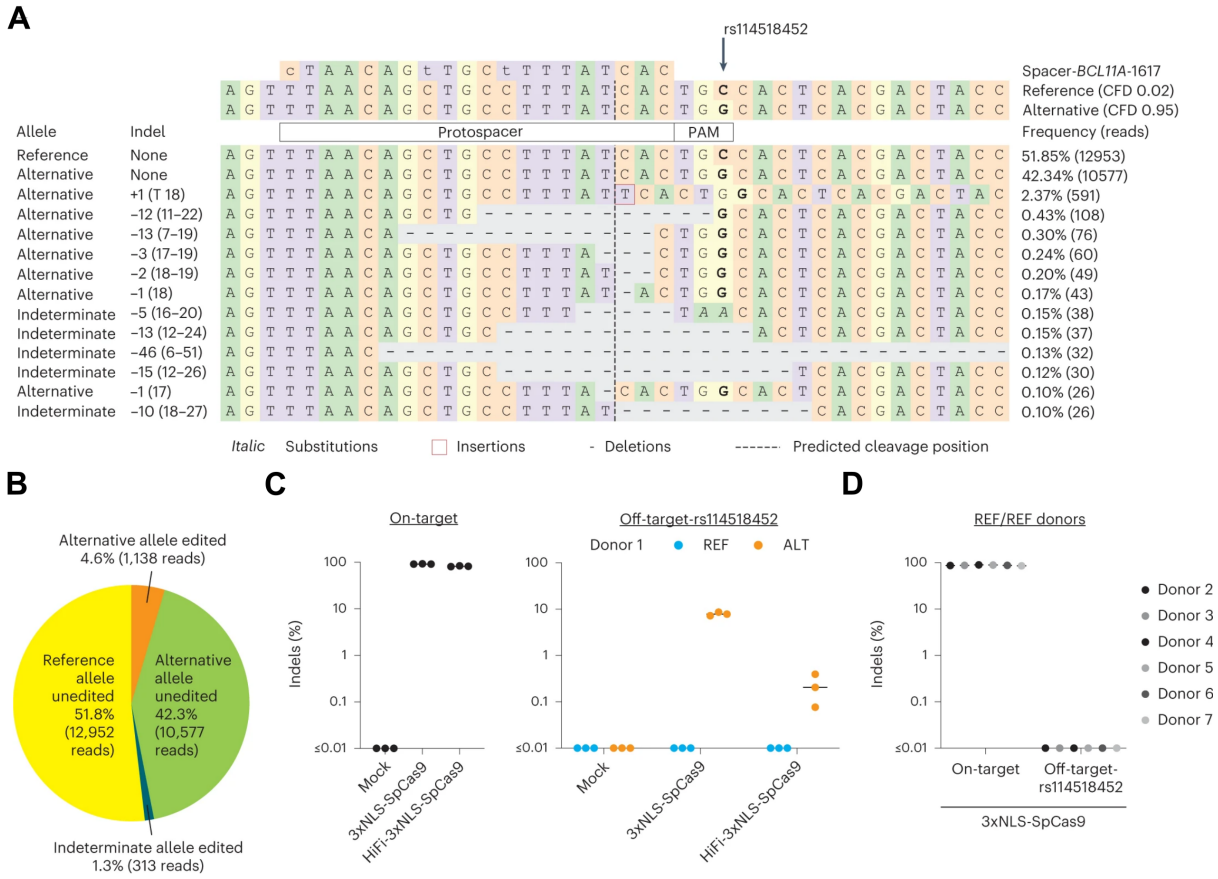


Figure 9.39. Allele-specific off-target editing by a BCL11A enhancer targeting gRNA in clinical trials associated with a common variant in African-ancestry populations. (A) Human CD34+ HSPCs obtained from a donor heterozygous for rs114518452-G/C (donor 1, REF/ALT) underwent 3xNLS-SpCas9:sg1617 RNP electroporation (NLS: nuclear localization signal), followed by amplicon sequencing of the off-target site located around chr2:210,530,659-210,530,681 (off-target-rs114518452 in 1-start hg38 coordinates). The CFD scores for both the reference and alternative alleles are provided, along with representative aligned reads. For ease of visual alignment, the spacer is depicted as a DNA sequence, with lowercase indicating mismatches and the rs114518452 position highlighted in bold. **(B)** Reads classified based on allele (indeterminate if rs114518452 position is deleted) and presence or absence of indels (edits). **(C)** Huma CD34+ HSPCs ffrom a donor heterozygous for rs114518452-G/C (donor 1) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation, HiFi-3xNLS-SpCas9:sg1617 RNP electroporation, or no electroporation (mock) followed by amplicon sequencing of the on-target and off-target-rs114518452 sites. Each dot represents an independent biological replicate (n = 3). Lines represent median values. Indel frequency was quantified for reads aligning to either the reference (REF) or alternative (ALT) allele. **(D)** Human CD34+ HSPCs from six donors homozygous for rs114518452-G/G (donors 2–7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation with one biological replicate per donor followed by amplicon sequencing of the on-target and off-target-rs114518452 sites.

designed to highlight alternative allele-specific candidate off-target sites that intersect with cCREs and protein-coding sequences, including putative tumor suppressor genes (Zhao *et al.*, 2016), and/or involve PAM creation events (**Figure 9.41 (B-C)**). For instance, among the top 20 candidate off-targets identified by CRISPRme for a SpCas9 gRNA targeting EMX1 (Tsai *et al.*, 2015), two sites are associated with genetic variants exhibiting high MAF (52% and 26%) and demonstrate substantial increases in CFD score from REF to ALT (+0.69 and +0.44). The first site involves an intronic PAM creation variant, while the second introduces two PAM-proximal matches to the gRNA (**Figure 9.41 (D)**). Remarkably, both of these candidate off-targets encompass indel variants, highlighting the capability of CRISPRme to consider variants beyond SNPs. In addition to visualizing candidate off-target sites based on predictive score ranks (such as CFD or CRISTA) for SpCas9-derived editors, CRISPRme offers the capability to visualize candidate off-targets according to the number of mismatches and bulges. This feature can be particularly valuable for Cas proteins with distinct PAMs, where predictive scores might not be readily available. For instance, SaCas9, a clinically relevant nuclease known for its small size and suitability for packaging into adeno-associated virus vectors. As an example, consider a SaCas9-associated gRNA targeting CEP290 (Maeder *et al.*, 2019), which is currently undergoing clinical trials for treating a form of congenital blindness (NCT03872479). CRISPRme identified two candidate off-targets linked to common SNPs (with MAFs of 7% and 5%) that reduce mismatches from five (REF) to four (ALT), and these sites

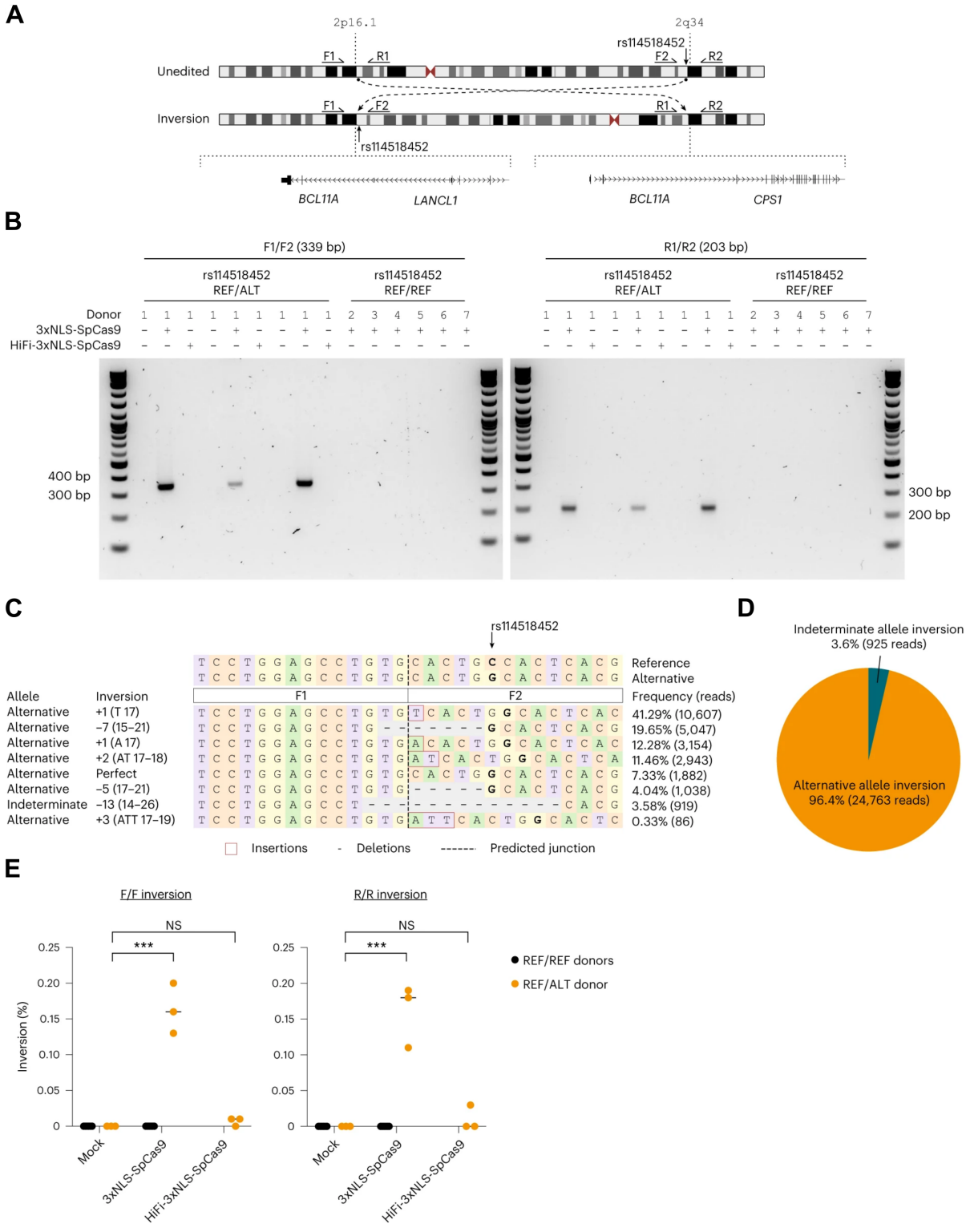


Figure 9.40. Allele-specific pericentric inversion following BCL11A enhancer editing due to off-target cleavage. (A) Concurrent cleavage of on- and off-target-rs114518452 sites may cause pericentric inversion of chr2. PCR primers F1, R1, F2 and R2 were designed to detect potential inversions. **(B)** Human CD43+ HSPCs from a donor heterozygous for rs114518452-G/C (donor 1) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation, HiFi-3xNLS-SpCas9:sg1617 RNP electroporation or no electroporation with three biological replicates. Human CD43+ HSPCs from six donors homozygous for rs114518452-G/G (donors 2–7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation with one biological replicate per donor. Gel electrophoresis for inversion PCR was performed with F1/F2 and R1/R2 primer pairs on left and right respectively with expected sizes of precise inversion PCR products indicated. **(C)** Reads from amplicon sequencing of the F1/F2 product (expected to include the rs114518452 position) from 3xNLS-SpCas9:sg1617 RNP treatment were aligned to reference and alternative inversion templates. rs114518452 position is shown in bold. **(D)** Reads classified based on allele (indeterminate if rs114518452 position deleted). **(E)** Inversion frequency by droplet digital PCR (ddPCR) from same samples as in panel b with three replicates from the single REF/ALT donor and one replicate each from the six REF/REF donors. F/F indicates forward and R/R reverse inversion junctions as depicted in panel a. NS, not significant.

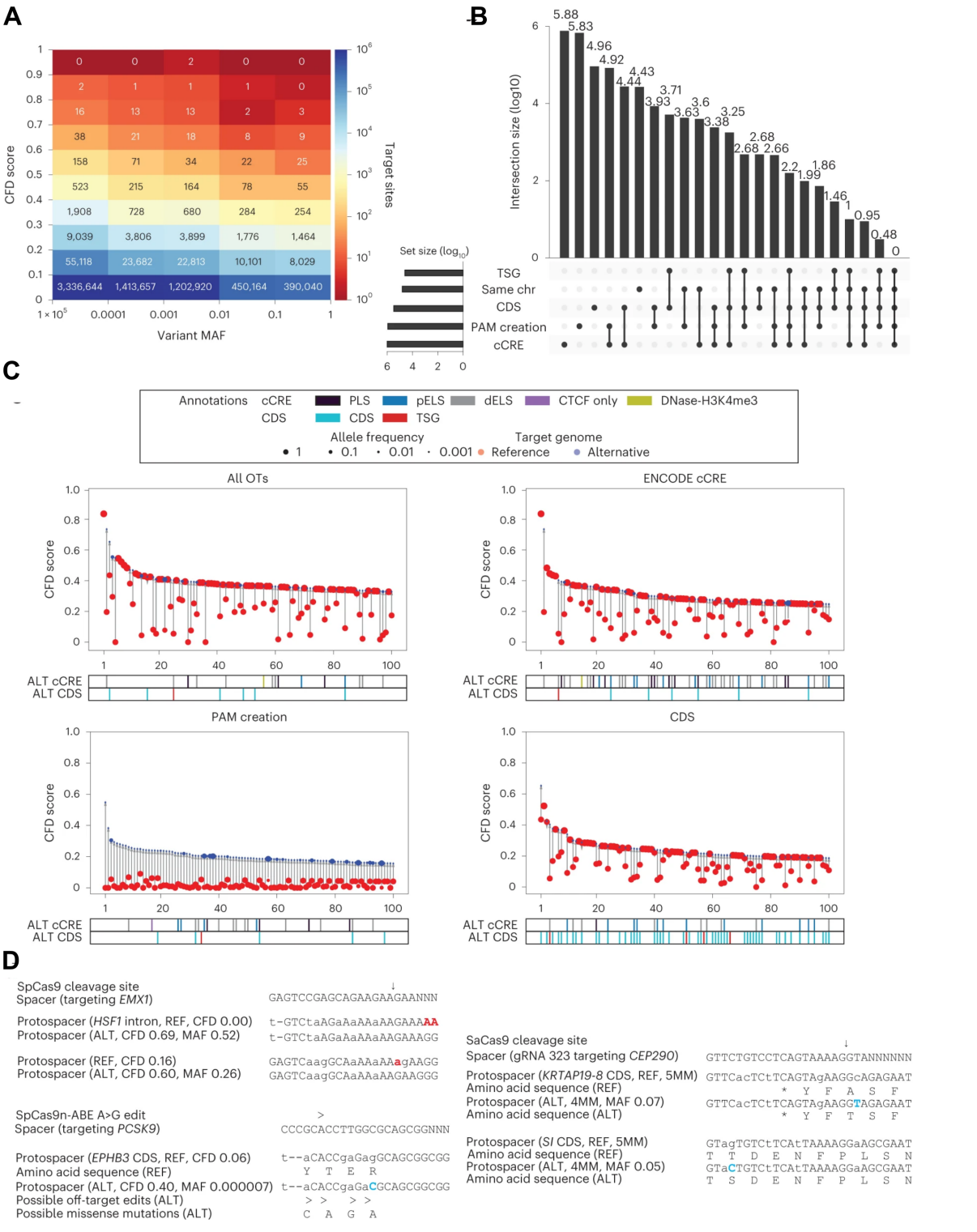


Figure 9.41. CRISPRme illustrates prevalent off-target potential due to genetic variation (A) Heatmap showing the distribution of alternative allele nominated off-targets for SpCas9 guides by CFD score and MAF. **(B)** UpSet plot showing overlapping annotation categories for candidate off-targets (tumor suppressor gene (TSG), candidate off-targets on the same chromosome (chr) as the on-target, CDS regions, cCRE from ENCODE and PAM creation events). **(C)** Top 100 predicted off-target sites ranked by CFD score for gRNA targeting PCSK9 with no filter, found in cCREs, corresponding to PAM creation events, and in CDS regions. **(D)** Candidate off-target sites with increased predicted cleavage potential introduced by common (MAF 52% and 26%) indel variants for a SpCas9 gRNA targeting EMX1 (top left). Candidate off-target cleavage sites within coding sequences with increased homology to a lead gRNA for SaCas9 targeting of CEP290 to treat congenital blindness in current clinical trials due to common SNPs (right). Potential missense mutations in the EPHB3 tumor suppressor resulting from candidate off-target A-to-G base editing by a preclinical lead gRNA targeting PCSK9 to reduce low-density lipoprotein cholesterol levels (bottom). MM denotes mismatches, deletions are shown in red, and SNPs are shown in blue.

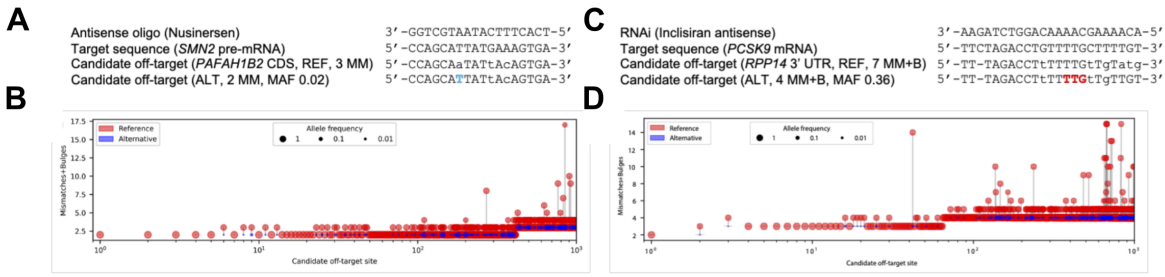


Figure 9.42. Candidate transcript off-targets introduced by common genetic variants for non-CRISPR sequence-based RNA-targeting therapeutic strategies. (A) A common SNP (in blue) introduces a candidate CDS off-target site with 2 mismatches for the FDA-approved antisense oligo Nusinersen. **(B)** Top 1000 candidate transcript off-targets ranked by mismatches and bulges for Nusinersen from a search performed with the 1000G and HGDP genetic variant datasets. **(C)** A common insertion variant (in red) introduces a candidate 3'UTR off-target site with 4 mismatches + bulges for the FDA-approved RNAi therapy Inclisiran. **(D)** Top 1000 candidate transcript off-targets ranked by mismatches and bulges for Inclisiran from a search performed with the 1000G and HGDP genetic variant datasets.

are predicted to induce cleavages within coding sequences (**Figure 9.41 (D)**). CRISPRme's functionality extends to the nomination of variant off-targets for base editors and the assessment of their susceptibility to base editing within a customizable editing window. Consider a gRNA targeting PCSK9 (Musunuru *et al.*, 2021), which has been utilized with SpCas9-nickase adenine base editor in vivo in preclinical studies aimed at reducing low-density lipoprotein cholesterol levels.Among the top five candidate off-target sites identified by CRISPRme, four involve alternative alleles. Notably, one off-target site exhibits a marked increase in CFD score, from 0.2 (REF) to 0.75 (ALT), and is situated within an ENCODE candidate enhancer element. Additionally, CRISPRme has pinpointed a candidate off-target linked to a rare variant (with a MAF of 0.0007%). This variant elevates the CFD score from 0.06 (REF) to 0.40 (ALT), potentially resulting in missense mutations in EPHB3, a putative tumor suppressor gene (**Fig.9.41 (D)**). CRISPRme tackles a fundamental computational hurdle that extends beyond CRISPR-based applications to encompass other technologies reliant on nucleic acid sequence recognition. For instance, CRISPRme has the capability to nominate off-targets for RNA-targeting strategies, including RNA-guided gene editors or oligonucleotide sequences used in RNA interference or antisense oligo therapies (**Figure 9.42**). We performed a variant-aware search (without PAM restriction) for the FDA-approved antisense oligonucleotide Nusinersen (Finkel *et al.*, 2017; Mercuri *et al.*, 2018), which targets SMN2 pre-mRNA to address spinal muscular atrophy. Employing CRISPRme, we pinpointed a potential off-target site within a coding region where a common SNP (MAF 2%) reduces the number of mismatches from three (REF) to two (ALT). Similarly, analysis of the FDA-approved RNA interference therapy Inclisiran (Raal *et al.*, 2020), which targets PCSK9 mRNA to combat hypercholesterolemia, unveiled an intriguing finding. The antisense strand of Inclisiran harbors a candidate off-target in the 3' untranslated region of the ribosomal gene RPP14. Notably, a common insertion variant (MAF 36%) decreases the number of mismatches and bulges from seven (REF) to four (ALT) at this site.

9.4 Analyzing candidate alternative allele-specific off-targets associated with therapeutic genome editing approaches with CRISPRme

In addition to the study involving sg1617, numerous other clinical trials of ex vivo SpCas9 gene editing have been reported. For instance, one trial utilized a gRNA targeting TRBC for gene-edited CAR-T cells (Stadtmauer *et al.*, 2020). In this case, 13 out of the top 20 ranked off-target sites, as determined by CFD score, are associated with alternative alleles. Notably, one off-target site, attributed to a rare variant, exhibited a CFD_{ref} of 0.05 and CFD_{alt} of 0.5, and is located within the coding sequences of TAS2R10, which encodes a G-protein coupled receptor expressed in T-cells. Moreover, a recent trial reported *in vivo* SpCas9 gene editing using lipid nanoparticle delivery to target the TTR gene in hepatocytes (Gillmore *et al.*, 2021). An off-target site for this gRNA displayed a CFD_{ref} of 0.01 and CFD_{alt} of 0.65, situated within an ENCODE candidate cis-regulatory element. CRISPRme analysis offers flexibility in accommodating PAM sequences beyond the conventional `NGG` motif. This capability is particularly valuable for base editors (BEs), where precise positioning of the desired edit within the editing window is crucial. For instance, various studies have explored SpCas9 variants with non-`NGG` PAM restrictions to optimize base editing outcomes. For example, a study detailed the conversion of the HBB sickle cell

mutation to a benign hemoglobin variant Makassar using A-to-G base editing (Chu *et al.*, 2021). This was achieved using a SpCas9 variant with a NGC PAM restriction and an IBE12 deaminase architecture, with the base editing window spanning from protospacer position 2-16, centered around position 9. During CRISPRme analysis, several alternative allele-nominated off-target sites were identified, some of which overlapped coding sequences. Notably, one such site exhibited four mismatches and bulges relative to the reference genome. However, in the presence of an alternative allele (with a MAF of 0.0007%), the mismatches and bulges reduced to three. This variant was located within ARHGAP26, where base edits positioned in the center of the editing window could potentially lead to missense mutations. ARHGAP26 is recognized as a tumor suppressor gene (Chen *et al.*, 2019), underscoring the importance of considering alternative allele-specific off-target effects in base editing strategies. In another study, the ABE8e-NRCH base editor was employed to convert the sickle cell allele to the Makassar variant. Upon CRISPRme analysis of alternative allele off-targets, a total of 1609 alternative allele off-targets were identified, each exhibiting four or fewer mismatches and bulges. Among these, 198 off-targets involved PAM creation events. One noteworthy alternative allele off-target site, with a MAF of 0.005%, was found to overlap the coding sequence for PLA2G4B. In this context, edits within the base editing window were anticipated to result in splice acceptor site disruption, potentially leading to loss-of-function alleles. PLA2G4B function has been associated with monocyte metabolic fitness (Saare *et al.*, 2020), suggesting the potential functional significance of off-target edits in this gene. CRISPRme extends its predictive capabilities to include off-target prediction for Cas12a-associated gRNAs featuring a 5' TTTV PAM sequence. For instance, in the case of a Cas12a gRNA directed at the HBG promoter for the treatment of sickle cell disease (SCD) in ongoing clinical trials (NCT04853576) (Morgan, 2020), CRISPRme identified an alternative allele-specific off-target site. In this scenario, the alternative allele introduces the TTTV PAM sequence, resulting in a protospacer with only 1 mismatch in the seed sequence. Notably, this off-target site is located within an ENCODE candidate cis-regulatory element. In summary, our study demonstrates that CRISPRme facilitates comprehensive off-target nomination for various guide RNAs under clinical development and across multiple Cas proteins (**Table 9.12**). Notably, none of these alternative allele off-target sites have been evaluated in the published descriptions of these gene editing approaches, highlighting the importance of tools like CRISPRme. Such tools can nominate candidate off-target sites that are absent from the human reference genome but present in specific populations or individual patients, thereby enabling a more thorough safety assessment.

9.5 Discussion and limitations

The described results underscore the impact of individual-specific genetic variation on off-target potential of sequence-based therapies, such as genome editing. Enhancing our ability to identify variant-associated off-target sites present in human populations would benefit from increased availability of haplotype-resolved genomes spanning diverse ancestries. Furthermore, these findings suggest how important is to consider genetic variation while developing individual-oriented genome editing therapies shifting towards precision medicine approaches. However, current tools, including CRISPRme, are limited in their capacity to enumerate potential off-targets based on structural variants or complex genetic events, such as combinations of indels and SNPs. Future extensions of CRISPRme, leveraging emerging data structures like genome graphs (Paten *et al.*, 2017; Garrison *et al.*, 2018), hold promise for handling these complex events and enhancing off-target nomination efficiency. It's crucial to assess the practical implications of allele-specific off-target editing on a case-by-case basis. For instance, in the context of BCL11A enhancer editing, approximately 10% of sickle cell disease (SCD) patients with African ancestry are expected to carry at least one rs114518452-C allele. Consequently, there could be up to 10% cleavage at an off-target site, which was not previously identified using currently available tools. These findings emphasize that allele-specific off-target editing potential is not uniformly distributed across all ancestral groups but is notably concentrated in populations of African ancestry, where genomic variation is most pronounced. Therefore, gene editing initiatives targeting conditions prevalent in populations of African ancestry, such as SCD, should pay particular attention to this issue. Gene editing efforts tailored to specific patient populations must account for genetic variants enriched in those populations during off-target assessment. However, our analysis reveals that variant off-targets can also be unique to individual patients, suggesting potential susceptibility across all human populations. Integrating off-target analysis and testing into therapeutic genome editing protocols is a multifaceted issue that extends beyond the scope of our study. Variant-aware off-target analysis is pivotal as it may uncover off-target potential overlooked by conventional methods. It's important to note that the mere presence of somatic genetic alterations does not necessarily imply functional consequences. While *ex vivo*-edited patient cells could theoretically undergo

Target gene	Cas protein	Status	Protospacer + PAM	Chrom	Start	PAM	Total Off-targets	ALT Off-targets	% ALT
CCR5	SpCas9	Clinical Trial	TCACTATGCTGC CGCCCAGTGGG	chr3	46,373,162	NGG	3,423,675	484,401	14
CCR5	SpCas9	Clinical Trial	CCCAGAAGGGGA CAGTAAGAAGG	chr3	46,373,138	NGG	3,365,086	559,834	17
TRAC	SpCas9	Clinical Trial	TGTGCTAGACAT GAGGTCTATGG	chr14	22,547,647	NGG	5,246,366	639,891	12
TRBC1/ TRBC2	SpCas9	Clinical Trial	GGAGAATGACGA GTGGACCCAGG	chr7	142,792,003/ 142,801,350	NGG	4,342,687	557,932	13
PDCD1	SpCas9	Clinical Trial	GGCGCCCTGGCC AGTCGTCTGGG	chr2	2.42E+08	NGG	2,423,918	249,404	10
TTR	SpCas9	Clinical Trial	AAAGGCTGCTGA TGACACCTGGG	chr18	31,592,987	NGG	4,169,153	497,569	12
HBB	SpCas9	Clinical Trial	CTTGCCCCACAG GGCAGTAACGG	chr11	5,226,967	NGG	2,384,266	340,029	14
HBB	SpCas9	Pre-Clinical	GGGTGGGAAAAT AGACTAATAGG*	chr11	5,226,803	NGG	2,839,539	328,619	12
HBG1/ HBG2	SpCas9	Pre-Clinical	CTTGTCAAGGCT ATTGGTCAAGG	chr11	5,249,955/ 5,254,879	NGG	2,943,866	316,718	11
EMX1	SpCas9	NA	GAGTCCGAGCAG AAGAAGAAGGG	chr2	72,933,852	NGG	4,887,079	902,572	18
FANCF	SpCas9	NA	GGAATCCCTTCT GCAGCACCTGG	chr11	22,625,785	NGG	3,328,636	338,978	10
BCL11A	A3A(N57Q)-BE3	Pre-Clinical	TTTATCACAGGC TCCAGGAAGGG	chr2	60,495,248	NGG	3,398,059	419,349	12
PCSK9	ABE8.8	Pre-Clinical	CCCGCACCTTGG CGCAGCGGTGG	chr1	55,040,028	NGG	3,335,810	386,504	12
HBB	IBE	Pre-Clinical	ACTTCTCCACAG GAGTCAGGTGC*	chr11	5,226,993	NGC	3,102,718	94,590	3
HBB	ABE8e-NRCH	Pre-Clinical	TTCTCCACAGGA GTCAGGTGCACC *	chr11	5,226,995	NRCH	5,822,863	1,995,125	34
CEP290	SaCas9	Clinical Trial	GTCAAAAGCTAC CGGTTACCTGAA GGGT	chr12	88,100,976	NNGRRN	373,955	150,445	40
CEP290	SaCas9	Clinical Trial	GTTCTGTCCTCA GTAAAAGGTATA GAGT	chr12	88,102,142	NNGRRN	906,516	318,522	35
HBG1/ HBG2	Cas12a	Clinical Trial	TTTGCCTTGTCA AGGCTATTGGTC	chr11	5,249,950/ 5,254,874	TTTV	1,258,797	415,379	33
HBB	Cas12a	Pre-Clinical	TTTATATGCAGA AATATTGCTATT ACC*	chr11	5,225,898	TTTV	453,213	138,953	31

Table 9.12. Additional gRNAs analyzed by CRISPRme representing a variety of target sequences, Cas proteins, and PAMs. The listed gRNAs are in clinical or preclinical development and/or widely tested in off-target studies. PAMs are indicated in bold. * indicates that the gRNA is specific to the sickle cell mutation (on-target site not found in hg38/REF). CRISPRme analysis was performed using the 1KGP and HGDP datasets with up to 6 mismatches and 2 bulges.

sequencing before infusion, the functional significance of off-target edits varies, ranging from potentially impactful to neutral. Thus, the detection of off-target editing in cell products may not inherently preclude their clinical utility, although such testing could deplete valuable material and delay therapy. To mitigate the risk of unintended allele-specific off-target effects during therapeutic genome editing, several steps are recommended. Firstly, prioritize the use of genome editing techniques that maximize specificity, such as high-fidelity editors and pulse delivery methods. Secondly, nominate off-targets in a variant-aware manner, with particular focus on genetic variants prevalent in relevant patient populations, utilizing tools like CRISPRme. Thirdly, employ off-target detection assays that are variant-aware to empirically assess the likelihood of off-target editing, recognizing that these assays may not perfectly reflect editing in a therapeutic context. When feasible, validate allele-specific off-target editing potential in primary cells of the relevant genotype through sequencing. However, obtaining such primary cells for biological validation in a clinically relevant context may pose challenges. Fourthly, conduct a risk assessment of variant off-target editing considering predicted genomic annotations, DNA repair mechanisms, delivery methods to target cells, and disease context. For instance, off-target edits within tumor suppressor loci may carry greater risk than those targeting unannotated non-coding sequences. Fifthly, if excess allele-specific genome editing risks are identified, consider incorporating genotype into subject inclusion/exclusion criteria. Finally, for therapeutic genome editing applications where feasible (e.g., hematopoietic cell targeting), proactively monitor somatic modifications in patient samples to gather data on the frequency and consequences of such events. This approach can help assess patient-specific risk and provide valuable insights into the frequency and in vivo dynamics of off-target edits if present. Overall, CRISPRme offers a user-friendly

framework to comprehensively evaluate off-target potential across diverse populations and within individuals. Notably, CRISPRme may significantly contribute to the development of more individual-focused genome editing therapies, helping the clinical shift towards precision medicine-oriented approaches.

Conclusions

Over the past decade, the landscape of medicine has undergone a profound transformation, signaling the waning relevance of the one-size-fits-all approach (Ginsburg and Willard, 2009). Precision medicine, characterized by its focus on tailoring treatments to the unique genetic profiles of individuals, stands poised to redefine the delivery of healthcare in the foreseeable future (Voelkerding *et al.*, 2009). This period has also marked a significant advancement in our comprehension of genetic variants, genomic diversity, and the intricate regulatory mechanisms governing gene expression. Crucially, genetic variants have emerged as pivotal biomarkers, illuminating the vast diversity among individuals and populations while holding substantial implications for human health (Raphael *et al.*, 2014). Within this PhD thesis, we have presented four studies that exemplify how leveraging the wealth of genetic variant data through computational methodologies can catalyze the transition toward precision medicine-oriented paradigms. The advent of genome graphs (Paten *et al.*, 2017) represents a pivotal breakthrough in genomics, offering a dynamic framework capable of capturing the intricate genetic diversity prevalent across populations and individuals. Unlike conventional linear reference genomes, genome graphs transcend limitations, providing a sophisticated lens through which to explore complex variant landscapes. Notably, these graphs offer an efficient means to unravel the genetic underpinnings of cellular environments and their regulatory mechanisms, particularly through the lens of genomic regulatory elements and transcription factors (Liao *et al.*, 2023). Within the cellular environment, genomic regulatory elements and transcription factors (TFs) orchestrate a symphony of molecular interactions crucial for maintaining regulatory balance (Lemon and Tjian, 2000; Lambert *et al.*, 2018). By delving into the intricate interplay between genetic variants and the binding landscapes of transcription factors, we unearth profound insights into the molecular foundations of complex traits and diseases. In this context, the introduction of GRAFIMO (Tognon *et al.*, 2021), with its novel variant- and haplotype-aware search for transcription factor binding site motifs on genome graphs, represents a significant leap forward. In our study, we demonstrated the capacity of GRAFIMO to uncover previously overlooked transcription factor binding sites (TFBSs), particularly those residing within population-specific and individual-specific genomic regions. This underscores the pivotal role of genetic variants in shaping the binding landscape of TFs. However, it's important to note that GRAFIMO's predictive performance may be constrained by the models employed to represent TFBS motifs (Tognon *et al.*, 2023). Using more sophisticated models holds the promise of enhancing the predictive capabilities of our tool, potentially unveiling further nuances in the landscape of TFBSs. MotifRaptor (Yao *et al.*, 2021) proposed a method to understand the functional implications of genetic variants, with a specific focus on their effects on TFBSs. Given that numerous disease-associated single nucleotide polymorphisms (SNPs) reside within non-coding regions of the genome, deciphering their functional significance poses a considerable challenge (Maurano *et al.*, 2012). However, MotifRaptor has introduced a novel TF-centric approach aimed at elucidating how these variants influence the binding landscape of transcription factors. Crucially, the methodology put forth by MotifRaptor integrates diverse omics datasets, including cell type-specific chromatin accessibility and transcriptomic profiles, enhancing its predictive capabilities. To illustrate its efficacy, we applied MotifRaptor to interpret the potential impact of two genetic variants associated with lipoprotein cholesterol (LDL-C) uptake. Notably, the predictions made by MotifRaptor were validated through experimental validation, underscoring its reliability and utility (Ryu *et al.*, 2024). Looking ahead, while MotifRaptor has demonstrated its prowess, there remains room for enhancement. Employing more intricate models beyond simple position weight matrices to represent TF motifs could further refine its predictive accuracy. Additionally, broadening the scope of analysis to encompass not only the binding sites themselves but also the genetic sequences surrounding these sites would provide a more comprehensive understanding of the impact of genetic variants on the genomic context guiding TF binding. Such advancements would elevate MotifRaptor's capacity to uncover the intricate interplay between genetic variation and transcriptional regulation, paving the way for deeper insights into the molecular mechanisms potentially underpinning complex traits and dis-

eases. The advent of CRISPR genome editing (Cong *et al.*, 2013) has empowered us to engineer precise modifications in the genome, offering unprecedented opportunities for therapeutic interventions. However, the successful translation of CRISPR-based therapies into clinical practice hinges upon our ability to accurately predict and quantify editing outcomes (Fu *et al.*, 2013). Whole genome sequencing has emerged as a cornerstone resource (Yin *et al.*, 2022), offering a comprehensive and unbiased means to evaluate and quantify editing outcomes without the limitations imposed by arbitrary site selection, as is often the case with amplicon sequencing (Akcakaya *et al.*, 2018). In this PhD thesis, we have undertaken a comprehensive study aimed at comparing and benchmarking three widely utilized methods (McKenna *et al.*, 2010; Kim *et al.*, 2018; Koboldt *et al.*, 2012) for calling genetic variants from sequencing data. Leveraging ultra-deep whole genome sequencing (1000x coverage), we meticulously examined control and treated cells subjected to guide RNAs targeting three genes with potential therapeutic relevance. Our findings underscore the critical importance of employing methods specifically tailored for genome editing quantification experiments. Notably, we observed that variant calling tools may exhibit tendencies to inaccurately estimate editing outcomes, potentially leading to under- or over-estimations. Furthermore, our analysis revealed limitations in the ability of these tools to detect editing events occurring at weak editing sites, despite their known susceptibility to editing (Tsai *et al.*, 2015). These insights highlight the pressing need for the development of specialized methodologies explicitly designed to recover and quantify editing outcomes across varying sequencing depths. Such tailored approaches are indispensable for realizing the full potential of CRISPR genome editing in clinical settings, where precision and accuracy are paramount. By addressing these methodological gaps, we can unlock new avenues for harnessing CRISPR technology in therapeutic applications, ultimately advancing the frontier of precision medicine. Moreover, the successful application of CRISPR genome editing in therapeutic contexts also hinges crucially on our ability to predict and mitigate off-target effects. Genetic variants play a pivotal role in potentially introducing novel off-targets or disrupting desired on-target sites (Lessard *et al.*, 2017; Scott and Zhang, 2017). In this context, the development of variant- and haplotype-aware computational tools, such as CRISPRme (Cancellieri *et al.*, 2023), marks a significant leap forward. This innovative tool offers an efficient, scalable, and user-friendly framework for predicting and evaluating off-target effects across the genome, while accounting for individual-specific genetic diversity. One notable strength of CRISPRme lies in its comprehensive approach to off-target prediction. It considers complex events such as the presence of potential DNA/RNA bulges and provides support for user-defined guide RNA and PAM sequences. Furthermore, CRISPRme generates several user-friendly reports and graphical summaries, facilitating the interpretation of the impact of genetic variants on CRISPR off-targets. In our study, we demonstrated the practical utility of CRISPRme by leveraging it to uncover an off-target site created by a genetic variant and specific to a subset of individuals. This off-target site had the potential to cause an inversion, disrupting a gene sequence and potentially leading to deleterious consequences in the patient's genomic environment. Looking ahead, CRISPRme stands to benefit from the integration of genome graphs into its predictive framework. By harnessing the power of genome graphs, CRISPRme could uncover and predict the impact of complex variants, such as SNP-indel combinations or structural variants, in the creation and modulation of novel off-target sites. This enhanced predictive capability would further bolster the utility of CRISPRme in guiding the safe and effective implementation of CRISPR-based therapies in clinical settings. In conclusion, our four studies vividly illustrate the transformative potential of omics sciences and computational tools in advancing precision medicine-oriented approaches. By centering our investigations on the pivotal role of genetic variants, fundamental biomarkers in precision medicine, we underscored the critical importance of accounting for individual- and population-specific genetic diversity. By harnessing the power of innovative computational methodologies, we elucidated significant events that have the potential to profoundly shape the cellular environment. These findings not only enhance our understanding of the genetic basis of diseases but also may help in developing tailored and effective potential precision medicine-oriented approaches.

Appendix A

Bioinformatics File Formats

This appendix describes the bioinformatics- and computational biology-related file formats encountered throughout the PhD thesis. The descriptions list the major details of each format, providing some context when possible.

A.1 FASTA file format

The FASTA file format is a widely used text-based format to represent nucleotide or protein sequences. It typically consists of two parts: a single-line description (header), beginning with ">", followed by the sequence data. The header line contains a brief description of the sequence and may include additional information, such as the sequence ID, source organism, or any relevant annotations. The sequence data comprises the nucleotide (`A`, `T`, `C`, `G`, `U`) or amino acid (protein) sequence. The sequence may be arranged as a single line or broken into multiple lines (each 80 bp long) for better readability. FASTA files may contain a single sequence or multiple sequences, with each sequence following the same two-part structure. There is no standard filename extension for FASTA files. Among the most common file extensions there are: `fasta` and `fa`.

A.2 FASTQ file format

The FASTQ format is a text-based representation designed to store both a biological sequence, typically a nucleotide sequence, and its corresponding quality scores. Each sequence letter and quality score is encoded using a single ASCII character to ensure brevity. Originally created at the Wellcome Trust Sanger Institute with the purpose of bundling a FASTA-formatted (**Appendix A.1**) sequence and its quality data, the FASTQ format has evolved into the prevailing standard for archiving the results produced by high-throughput sequencing instruments. There is no standard file extension for FASTQ files. However, `.fq` and `.fastq` are commonly used file extensions.

A.3 SAM file format

The Sequence Alignment Map (SAM) (Li *et al.*, 2009a) is a text-based format developed to store biological sequences aligned to a reference sequence. SAM files present a TAB-delimited structure. The SAM files comprise a header and an alignment section. The header section, when present, precedes the alignment section. Headers are denoted by "@". The alignment section is characterized by 11 mandatory fields and may include a variable number of optional fields (**Table A.1**). SAM files can be manipulated and examined using SAMtools (Li *et al.*, 2009a). The SAM format has become widely adopted for storing diverse data, including nucleotide sequences, generated by next-generation sequencing technologies. The standard has been expanded to accommodate unmapped sequences, and it supports both short and long reads (up to 128 Mbp). SAM is used to store and process mapped data by several tools, such as the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010) or strelka (Kim *et al.*, 2018), and organizations, such as the ENCODE project (Consortium *et al.*, 2012) or 1000 Genomes Project (Consortium *et al.*, 2015).

Field	Type	Description
QNAME	string	Query name
FLAG	int	bitwise FLAG
RNAME	string	Reference sequence name
POS	int	Mapping position (1-based)
MAPQ	int	Mapping quality
CIGAR	string	CIGAR string
RNEXT	string	Reference name of the mapped/next read
PNEXT	int	Position of the mapped/next read
TLEN	int	Template length
SEQ	string	Read sequence
QUAL	string	ASCII Phred-scaled base quality (+ 33)

Table A.1. SAM file format fields. Description of SAM format fields

A.4 BAM file format

BAM (Binary Alignment Map) (Barnett *et al.*, 2011) represents the compressed binary counterpart of the SAM format (**Appendix A.3**). BAM files provide a condensed and indexable representation of nucleotide sequence alignments. SAM/BAM are widely employed by numerous next-generation sequencing and analysis tools. In terms of custom track display, indexed BAM offers a notable advantage over other human-readable alignment formats. This approach enables the processing of alignments from large files that might otherwise cause issues when attempting to analyze the entire files.

A.5 CRAM file format

While BAM (**Appendix A.4**) files encompass all sequence data within a single file, CRAM files achieve a smaller size by leveraging an external "reference sequence" file, which is essential for both compression and decompression of the read information. Due to their higher compression density, many organizations and tools are transitioning to the CRAM format to optimize disk space utilization. The work properly, CRAM files necessitate an associated index file, similarly to BAM files.

A.6 BED file format

The BED (Browser Extensible Data) file format serves as a versatile method to display genomic features on genome browsers, such as the UCSC Genome Browser (Lee *et al.*, 2020). BED files are plain text documents consisting of twelve tab-separated fields, with the first three being mandatory. BED files can be analyzed and processed using various tools, including BEDtools (Quinlan and Hall, 2010). Currently, there exist several variations of the BED file format. Due to its adaptability, BED files have been extensively employed to summarize diverse genomic features, especially ChIP-seq peak regions. The ENCODE Project provides ChIP-seq peak data in the ENCODE BED `narrowPeak` format, deviating from the classical BED file format. The ENCODE `narrowPeak` consists of 10 fields (**Table A.2**). This format requires all ten fields to contain values. If data are not available for a certain field, the format defines null values to insert (**Table A.2**). Importantly BED `narrowPeak` files store statistical significance for the called peak, incorporating two fields to store both a P-value and a q-value.

A.7 MEME file format

The MEME file format (Bailey *et al.*, 2009) is a text representation for DNA or amino acid motifs, encoded as PWMs. Several motif discovery tools return motifs in MEME format, in particular those within the MEME suite (Bailey *et al.*, 2009). As the format is plain text (ASCII), it can be manually crafted using a simple text editor or word processor. A single MEME file can encompass one or more motifs and, if needed, specify the alphabet upon which the motifs are constructed, background frequencies of the alphabet letters, and strand information when dealing with DNA motifs. MEME files comprise five sections: (i) the version line, (ii) alphabet, (iii) strands (optional), (iv) background frequencies, and (v) motifs. The version line denotes the version of the tool (if part of the MEME suite) that produced the motif. The alphabet line includes the motif sequence alphabet, such as `ACGT` for DNA, `ACGU` for RNA, and `ACDEFGHIKLMNPQRSTVWY` for protein sequences. The strands line indicates whether the motif has been

Field	Description	Type
chrom	Name of the chromosome (or contig, scaffold, etc.)	mandatory
chromStart	Starting position of the feature in the chromosome. The first base in a chromosome is numbered 0	mandatory
chromEnd	Ending position of the feature in the chromosome. The chromEnd base is not displayed	mandatory
name	Feature name (preferably unique)	optional
score	Indicates feature's gray shade to be displayed in the genome browser (0-1000).	optional
strand	+/- denote feature's strand or orientation (whenever applicable)	optional
signalValue	Measurement of overall (usually, average) enrichment for the region	optional
pValue	Feature's statistical significance (-log10). Use -1 if no pValue is assigned	optional
qValue	Feature's statistical significance using false discovery rate (-log10). Use -1 if no qValue is assigned	optional
peak	Feature's peak summit position (0-based offset from chromStart). Use -1 if no peak summit called	optional

Table A.2. ENCODE BED narrowPeak file format fields. Description of ENCODE BED narrowPeak format fields.

discovered using both the forward and reverse strands. The background frequency line provides the letter frequencies of the motif alphabet in the input sequences. These frequencies must sum to 1. The motif section is further divided into three subsections: motif name, motif letter-probability matrix lines, and motif URL (optional). The motif name line contains the motif ID and its full name. This line marks the beginning of a new motif if the file contains more than one motif. The motif letter-probability matrix lines displays a probability table where rows represent motif positions and columns correspond to motif alphabet letters, sorted alphabetically. For a DNA motif, the columns contain A data in the first column, C data in the second, G data in the third, and T data in the fourth. Following the "letter-probability matrix" line, four optional values describe the motif alphabet length, motif width, number of source sites, and source *E*-value. The motif URL section includes a link to the source page for the current motif.

A.8 JASPAR file format

The JASPAR file format (Sandelin *et al.*, 2004) features a header line introduced by ">", similar to FASTA file headers, followed by the motif ID. In JASPAR motif files there are four lines, each one dedicated to a specific nucleotide (`A`, `C`,`G`, `T`). In JASPAR motif files columns represent motif positions, and rows represent the motif alphabet letters. Each line begins with the corresponding nucleotide, and the subsequent tab-separated raw count values are enclosed in square brackets "[ ]".

A.9 PFM file format

The PFM file format includes a header line introduced by ">" and followed by the unique motif identifier. In PFM files rows represent nucleotides, and columns motif positions. Each nucleotide's raw count values are outlined on a separate line, with the row letters ordered alphabetically.

Appendix B

Motif discovery algorithms

B.1 Algorithmic details of motif discovery software

B.1.1 Enumerative methods

Enumerative methods for motif discovery aim to identify overrepresented sequence patterns in the input dataset S compared to a background set of sequences B. These methods involve counting the approximate occurrences of all possible 4^k k-mers (assuming an alphabet $\Sigma = A, C, G, T$) in S and assessing the statistical significance of the difference between the observed matches in S and B or the expected match frequencies based on a background model (**Section 4.2.2**). The table in **Appendix B.2** provides a summary of the key characteristics of enumerative methods. Early approaches employed heuristics to explore the motif search space, allowing for mismatching positions. However, more recent methods such as Weeder (Pavesi *et al.*, 2001, 2004b) and SMILE (Marsan and Sagot, 2000) present more efficient and accurate strategies by utilizing suffix trees (STs) to index the input sequences. Both Weeder and SMILE leverage STs for efficient approximate pattern matching, searching all k-mers occurrences in S while allowing mismatches at any motif position. Although Weeder and SMILE share a similar data structure, they differ in their approaches to assess the significance of the identified motifs. SMILE evaluates motif significance by comparing the number of occurrences of a specific motif in S with its expected frequency in B. To estimate expected occurrences, SMILE either generates a random set of sequences using a Markov model or uses a negative set of sequences that do not contain instances of the target binding sites. On the other hand, Weeder compares the observed occurrences of each motif candidate with its expected frequencies derived from the regulatory regions of the same organism as the input sequences. It is important to note that both SMILE and Weeder may require several hours to complete on datasets consisting of thousands of sequences generated by high-throughput assays. To address this challenge, there has been a growing interest in the development of enumerative motif discovery methods capable of efficiently analyzing large datasets comprising thousands of sequences. MDscan (Liu *et al.*, 2002) uses word enumeration for motif discovery in sequence datasets. Initially designed for ChIP-on-Chip datasets, MDscan identifies non-redundant patterns that are abundant in the most enriched ChIP peak sequences, rather than enumerating all possible words from $s \in S$. To assess the statistical significance of motif candidates, MDscan employs a third-order Markov model as background. Amadeus (Linhart *et al.*, 2008) evaluates all k-mers in input sequence datasets. Amadeus has been originally designed to focus on datasets generated by ChIP-on-Chip assays. It groups similar sequence patterns into lists and combines these lists into motifs, which are statistically evaluated using a hypergeometric test to determine their significance. Recently, both MDscan and Amadeus have been extended to support other high-throughput sequencing technologies such as ChIP-seq and DNase-seq. Additionally, Amadeus has been adapted to analyze data from multiple organisms, aiming to identify conserved motifs across species. Enumerating words in datasets containing thousands of sequences can be computationally expensive in most scenarios. DREME (Bailey, 2011) presents an alternative approach using regular expressions to enhance motif discovery on large sequence datasets. DREME leverages regular expressions for counting approximate motif occurrences in both S and B, allowing a more efficient exploration of the motif search space across extensive sequence sets. DREME generates the background dataset from S using a Markov model or utilizes a user-provided negative set of sequences. To assess the statistical significance of the discovered motifs, DREME employs Fisher's exact test, comparing the occurrences of motifs in sequences from S and B. However, regular expressions for motif discovery hide certain drawbacks: they

(i) can be computationally demanding on large datasets, (ii) might identify patterns as motifs that are not biologically relevant, and (iii) may not capture all variations of a motif. While enumerative methods employing STs may address some of these challenges, they are computationally demanding and often lack scalability when analyzing thousands of sequences. Trawler, HOMER, and STREME (Ettwiller *et al.*, 2007; Heinz *et al.*, 2010; Bailey, 2021) propose different enhancements for discovering motifs using STs in the context of large-scale sequence datasets. Trawler (Ettwiller *et al.*, 2007) focuses on detecting transcription factor binding site (TFBS) motifs in extensive datasets like ChIP-seq and DNase-seq data. Employing suffix trees, Trawler enumerates all sequence patterns in both S and B, measuring their frequencies to identify overrepresented motifs in S. Trawler allows for mismatching positions by employing degenerate consensuses when matching motif sequences. Statistical evaluation of prioritized patterns in Trawler involves z-scores derived from the normal approximation to the binomial distribution. This approach facilitates fast computation without the need to correct for the effect of overlapping motifs. Trawler constructs motifs by clustering similar and significant patterns. HOMER (Heinz *et al.*, 2010), tailored to analyze ChIP-seq datasets, indexes both the input and background sequence datasets using STs. It searches for overrepresented patterns in the target sequences through approximate pattern matching on the ST. The identified overrepresented k-mers are clustered into motif candidates, which are statistically evaluated using either the hypergeometric or binomial test. By avoiding explicit counting of k-mer frequencies, HOMER reduces its running time. The significant motif candidates undergo further refinement by eliminating those lacking conserved positions or exhibiting low information content. While originally designed for ChIP-seq, HOMER demonstrates applicability to sequence datasets from diverse assays. STREME (Bailey, 2021) proposes a strategy to enhance motif discovery efficiency by employing approximate matching on suffix trees. STREME constructs an ST from S and identifies overrepresented seed words of different lengths through exact matching on the ST. Statistical significance of each seed word is assessed using either the Fisher's exact test or binomial distribution. STREME then counts the number of approximate matches for the most significant words on the ST, using a mismatch threshold. Subsequently, STREME prioritizes and groups the most significant seed words to derive the corresponding motifs. This approach enables STREME to identify motifs of different lengths in a single tree visit, substantially reducing the running time for motif discovery.

B.1.2 Alignment-based methods

Alignment-based motif discovery algorithms utilize employ alignment profiles to identify potential TFBS within the input sequence dataset S. These algorithms construct profiles by combining potential motif instances and then assess each profile's score using various statistical measures. However, the exhaustive enumeration of all possible alignments is impractical. As a result, efforts have been directed towards incorporating heuristics to efficiently generate high-quality alignments. The Table in **Appendix B.2** provides a summary of the key characteristics of the alignment-based methods discussed in this section. CONSENSUS (Hertz and Stormo, 1999) uses a greedy algorithm to identify TFBS alignment profiles. It assumes a single occurrence of the motif in each sequence of S. Assuming $|M| = k$, CONSENSUS starts by comparing all k-mers in two sequences, $s_1, s_2 \in S$, generating a $4 \times k$ profile for each k-mer pair. Each profile undergoes individual scoring based on information content, and the alignments with the highest scores are retained. Subsequently, CONSENSUS compares each profile to the k-mers of another sequence, $s_3 \in S/\{s_1, s_2\}$, producing a new set of profiles involving three sequences. These resulting profiles are scored using information content, and the alignments with the highest scores are preserved. CONSENSUS incrementally stores the best partial alignments with the hope of eventually finding the optimal profile. However, in cases where motifs lack conservation, the algorithm may store profiles corresponding to random k-mers, potentially excluding highest-scoring alignments. MEME (Bailey *et al.*, 1994; Bailey and Elkan, 1995; Bailey *et al.*, 2006) introduced an EM strategy to comprehensively explore the entire motif search space, distinguishing itself from CONSENSUS by not relying on partial solutions. Initially, MEME initializes a profile with one k-mer from each sequence, $s \in S$. This starting profile undergoes iterative refinement through an EM strategy involving two steps: the E-step and the M-step. In the E-step, MEME computes a likelihood score for each k-mer in S, using the current alignment profile. In the M-step, MEME assigns a weight to each k-mer in the current profile, proportional to the scores computed during the E-step, and updates the alignment by replacing low-scoring k-mers with others that better fit the current profile. MEME accommodates sequences with zero or more than one motif occurrence, provides background models of different orders, and assigns a P-value to each identified motif. This P-value indicates the probability of obtaining a profile with the same information content by chance, offering a measure of confidence for each discovered motif. Algorithms employing Gibbs sampling

address MEME's main limitation of potentially converging to local maxima, thus not always reporting the best alignment profile. The generic Gibbs sampling algorithm begins by initializing a profile with k-mers randomly chosen from each sequence in the input dataset S. In each iteration, the algorithm removes a k-mer from the profile originating from a specific sequence s and computes a likelihood score for each k-mer in s based on the modified profile. This score reflects how well each k-mer fits the profile, rather than a background model, akin to MEME. Subsequently, the algorithm replaces the removed k-mer with a new one chosen with a probability proportional to its likelihood score. This process repeats until a fixed number of iterations or until no further changes to the profile occur. The foundational Gibbs sampling algorithm (Lawrence and Reilly, 1990) assumes that each sequence in S contains precisely one binding site. Motif sampler (Neuwald *et al.*, 1995) extended the original algorithm to accommodate sequences containing one, multiple, or no binding sites. However, it necessitates the user to provide an estimate of the expected number of occurrences of the TFBS in the input sequence dataset. Further modifications to the Gibbs sampling procedure were introduced by AlignACE (Hughes *et al.*, 2000) and ANN-spec (Workman and Stormo, 1999), enabling the simultaneous exploration of TFBS on both strands. AlignACE also incorporates a sampling method that considers the relative position of each k-mer within a group in relation to gene transcription start sites (TSS). This ensures that functional motifs correspond to similar regions occurring at comparable distances from the TSS. ANN-spec couples Gibbs sampling with a perceptron artificial neural network, replacing the alignment profile. Bioprospector (Liu *et al.*, 2000) and MotifSampler (Thijs *et al.*, 2001) suggest using a third-order Markov model as a background, enhancing the predictive performance of the Gibbs sampling algorithm. Typically, alignment-based motif discovery methods assume that the motif length is known *a priori*. GLAM (Frith *et al.*, 2004b) modified the Gibbs sampling procedure to estimate the optimal alignment length and employed simulated annealing to optimize the motif profiles. GLAM was further extended to consider gapped motifs, introducing flexibility to motif variability (Frith *et al.*, 2008). While these methods employ heuristics for efficiently exploring the motif solution space, they were initially designed to analyze datasets comprising hundreds or a few thousand sequences. To address this limitation, developers of motif discovery algorithms have proposed novel alignment-based methods extending the original ideas to analyze the large datasets generated by NGS assays. MEME-ChIP (Machanick and Bailey, 2011) extends the original MEME algorithm to handle large ChIP datasets. Rather than analyzing the entire input dataset, MEME-ChIP runs the EM motif search on a random subset of sequences from S. Additionally, MEME-ChIP prioritizes motifs discovered around ChIP-seq peak summits. While originally designed for ChIP-seq and ChIP-on-Chip datasets, MEME-ChIP is versatile enough to be applied to datasets generated by different experimental assays. Similarly, STEME (Reid and Wernisch, 2011) enhances the MEME algorithm by indexing the input sequences using a suffix tree (ST). Employing a branch-and-bound strategy on the ST, STEME avoids evaluating the likelihood of all possible k-mers during MEME's E-step. ChIPMunk (Kulakovskiy *et al.*, 2010) introduces an efficient method for discovering motifs in thousands of genomic sequences. In contrast to traditional approaches using EM, ChIPMunk prioritizes a set of enriched k-mers initially and constructs a profile from them. Subsequently, the profile undergoes a scalable EM-like refinement to maximize the discrete information content (Kulakovskiy and Makeev, 2009). Moreover, ChIPMunk leverages ChIP peak shapes to weight the contribution of each sequence $s \in S$ to the motif definition, enhancing the accuracy of identified motifs. XXmotif (Hartmann *et al.*, 2013) integrates enumerative motif discovery with profile refinement. It iteratively selects sets of k-mers from S to maximize their fitness to the profile and enhances the motif's fitness to S. Similarly, ProSampler (Li *et al.*, 2019b) proposes a highly optimized and ultra-fast hybrid method for discovering motifs in ChIP-seq datasets. It combines motif enumeration with Gibbs sampling to refine motif profiles effectively.

B.1.3 Probabilistic graphical models-based methods

Whether including dependencies between neighboring and non-neighboring nucleotides in TFBS motif discovery and models has been a longstanding topic of discussion within the research community. The Table in **Appendix B.2** provides a concise overview of the key characteristics of probabilistic graphical models-based methods covered in this section. Motif discovery algorithms, such as Dimont (Grau *et al.*, 2013) and diChIPMunk (Kulakovskiy *et al.*, 2013a), have been developed to detect and represent TFBS motifs using DWMs. These methods, employing more sophisticated approaches than PWMs, demonstrate scalability to large datasets comprising thousands of sequences without compromising accuracy. In the case of Dimont (Grau *et al.*, 2013), the initial motif profile is initialized by combining the most overrepresented and non-redundant 7-mers in sequence set S, forming an alignment of length $|M|$. Dimont assigns higher weights to the positions at the center of the profile. To ensure scalability with large

datasets, Dimont uses a highly optimized supervised posterior probabilities estimation technique to refine the initial profile. The algorithm also filters redundant and reverse complement motifs that match the highest-scoring profiles, contributing to a reduction in runtime. Finally, the motif profiles undergo optimization using Kullback-Leibler divergence. Similarly, diChIPMunk (Kulakovskiy *et al.*, 2013a) extends the original ChIPMunk (Kulakovskiy *et al.*, 2010) procedure to construct DWMs, by considering motif dinucleotide frequencies. The algorithm employs an EM-like approach to optimize the starting motifs and weighs the contribution of sequences to the TFBS motif, accounting for the ChIP peak shape. Unlike ChIPMunk, which maximizes the discrete information content on individual nucleotides, diChIPMunk optimizes this measure based on dinucleotide frequencies. TFFMs (Mathelier and Wasserman, 2013) introduces a model based on hidden Markov models (HMMs) to capture dinucleotide dependencies between neighboring TFBS positions and learn the properties of sequences flanking the binding sites. By incorporating insertion and deletion states into the model, TFFMs can accommodate variable motif lengths, making it a flexible model. TFFMs leverage the MEME algorithm (Bailey *et al.*, 1994; Bailey and Elkan, 1995; Bailey *et al.*, 2006) to identify motif candidates, subsequently used to train the HMMs employing the Baum-Welch algorithm. Similarly, Discrover (Maaskola and Rajewsky, 2014) proposes a discriminative approach for motif discovery in ChIP-seq datasets, using HMMs. Initially, Discrover identifies overrepresented motifs in the input dataset S using regular expressions, similar to DREME (Bailey, 2011). The identified motif candidates serve as seeds to initialize an HMM. The algorithm prioritizes these candidates based on different objective functions (Maaskola and Rajewsky, 2014), such as likelihood or mutual information, to select the most promising seeds. Subsequently, the initial HMM undergoes refinement through iterative gradient optimization of the model likelihood. Markov Models (MMs) and HMMs can be extended to incorporate variable-order dependencies between neighboring and non-neighboring motif positions, akin to Bayesian Networks (BNs). Slim (Keilwagen and Grau, 2015) introduces a framework for learning dependencies of different orders between neighboring and non-neighboring nucleotides. The dependency orders are pruned in a data-driven manner, iteratively establishing the motif positions on which each nucleotide depends. The models are trained using motif candidates prioritized by running Dimont on S, maintaining scalability on large sequence datasets. However, to ensure scalability on large datasets, methods like HMMs and MMs often learn low-order dependencies from the input sequences. Addressing this, BaMMotif (Siebert and Söding, 2016; Ge *et al.*, 2021) presents an efficient motif discovery algorithm that learns high-order dependencies (up to the 5-th order) on thousands of sequences using an optimized Bayesian approach to train MMs. To prevent model overfitting, the algorithm uses the conditional probabilities of $(q-1)$-th order as priors for the q-th order conditional probabilities during training. The algorithm iteratively optimizes the model parameters until convergence through Expectation-Maximization (EM). During the E-step, BaMMotif estimates the probability that each position of each $s \in S$ is a motif starting position, using the current model. In the M-step, the algorithm refines the model using the motif candidates identified in the previous step. The algorithm dynamically adapts the model parameters during training to capture variable orders of dependencies.

B.1.4 SVM-based methods

Support Vector Machines (SVMs) offer an efficient and scalable approach for learning complex sequence features in different biological contexts, including TFBS. SVMs have demonstrated scalability, even when dealing with thousands of sequences typical of datasets generated by high-throughput protocols. Notably, SVMs not only capture the motif itself but also incorporate other sequence features surrounding the binding sites into the learned model. A summarized overview of the key features of SVM-based methods discussed in this section can be found in the Table in **Appendix B.2**. Kmer-SVM (Lee *et al.*, 2011; Fletez-Brant *et al.*, 2013) proposes a scalable framework using SVMs for the discovery and representation of TFBS as SVM models. The algorithm operates on a foreground dataset S and a background dataset B. It dissects the input sequences $s \in S$ into contiguous k-mers (with $k \sim 10$ base pairs) and computes their frequencies in both datasets. Employing a trie (Bodon and Rónyai, 2003) for indexing S and B, Kmer-SVM achieves linear runtime. The complete frequency profiles of k-mers within each sequence are stored as feature vectors, serving as input for training the SVM kernel. Kmer-SVM employs the spectrum kernel, optimizing parameters to effectively distinguish between positive and negative sequences (support vectors). Additionally, Kmer-SVM offers a weighted version of the spectrum kernel, where the contributions of k-mers are weighted based on their positional information within the sequence. However, this implementation lacks flexibility in k-mer frequency estimation and encounters scalability issues when analyzing datasets comprising tens of thousands of sequences containing degenerate and long (> 10 base pairs) motifs. Agius *et al.* (2010) present a more flexible framework through the training of Support

Vector Regression models (SVRs). The algorithm proposes the usage of a dinucleotide mismatch kernel (di-mismatch kernel). While the basic mismatch kernel permits up to m mismatches in each k-mer match, for small m values, the mismatch neighborhood of a given k-mer becomes excessively large. To address this issue, the authors introduced the di-mismatch kernel, which tallies mismatching nucleotides and favors k-mers with consecutive mismatches. Furthermore, the kernel is tailored for sets of unique k-mers occurring in training sequences, such as PBM probe sequences. Nevertheless, this method is constrained by considering short k, restricting the ability to discover longer motifs. To overcome this limitation, gapped-k-mers (Ghandi *et al.*, 2014b) introduce gaps in non-informative motif positions to accommodate longer motifs while preserving scalability on large datasets. A gapped-k-mer is defined by a pair of values (l, k) representing the full length and the number of non-gap positions, respectively. For example, `A*CG` is a gapped-k-mer with $l = 4$ and $k = 3$. Gkm-SVM (Ghandi *et al.*, 2014a, 2016) constructs feature vectors for each input sequence by tallying the occurrences of gapped-k-mers in both positive and negative datasets, as opposed to exact k-mer frequencies. These feature vectors are then used to train the gapped-k-mer kernel, which computes support vectors to distinguish between bound and unbound sequences. To determine the SVM hyperplane, the feature vectors for each sequence are mapped to the normalized counts of their distinct gapped-k-mers. However, this approach has been shown to be non-scalable when applied to large sequence datasets ($> 10,000$ sequences). LS-GKM (Lee, 2016) enhances Gkm-SVM to effectively handle large-scale datasets. Additionally, LS-GKM introduces a weighted version of the gapped-k-mer kernel, along with the **gkmrbf** kernel and its weighted counterpart, to capture non-linear relationships between sequence features.

B.1.5 Deep Neural Networks-based methods

Deep neural networks (DNNs) have demonstrated considerable success in addressing various challenges in computational biology, including the discovery and classification of transcription factor binding site motifs. The Table in **Appendix B.2** provides a concise overview of the key characteristics of DNN-based methods. Convolutional neural networks (CNNs) (LeCun *et al.*, 2015) have found widespread applications in motif discovery. Typically, CNNs designed for motif discovery consist of several fundamental layers (Zeng *et al.*, 2016a). In the first layer, known as the convolutional stage, input sequences from the dataset S undergo scanning with a set of convolutional filters (motif kernels). The resulting response values are then forwarded to the subsequent layer, called max-pooling layer. Within the max-pooling layer, maximal responses for each convolutional layer are selected and incorporated into feature vectors. The subsequent layer, the neural network stage, transforms these feature vectors into scalar scores, representing the likelihood of each site containing a TFBS. To mitigate overfitting, a dropout layer is frequently introduced at this stage, randomly masking portions of the neural network output. The output layer typically comprises two fully connected neurons, indicating the model predictions for "bound" or "unbound" sequences. Backpropagation and gradient descent are commonly employed to iteratively estimate and optimize the CNN parameters. One of the pioneering proposals utilizing CNN for motif discovery on ChIP-seq, HT-SELEX, and PBM is DeepBind (Alipanahi *et al.*, 2015). The DeepBind architecture introduces two variations to the basic CNN architecture: a rectification layer and the inclusion of averaging and maximization steps in the pooling layer. The rectification layer identifies sequences that align well with the convolutional kernel by shifting the response by some nucleotides and masking poorly matched positions. In the pooling layer, DeepBind calculates both maximum and average kernel responses, allowing for the identification of cumulative effects of short motifs and the determination of the locations of longer TFBS, respectively. To visualize the discovered motifs, DeepBind computes weighted ensembles of PWMs, obtained by aligning the sequences activating the kernels, centered around the position with the maximum response. Basset (Kelley *et al.*, 2016) enhances the basic CNN architecture by incorporating three additional convolutional layers after the pooling stage, followed by two fully connected neural networks. The model's parameters are initialized randomly and subsequently adjusted adaptively through backpropagation. Basset represents the identified motifs using PWMs. BPNet (Avsec *et al.*, 2021a) introduces further modifications to the CNN architecture to address the limitations observed in methods like DeepBind and Basset. By integrating nine dilated convolutional layers and eliminating the pooling stage, BPNet preserves the spatial and base resolution of the input data, capturing the intricacies of the discovered motifs. In its original implementation, BPNet utilizes ChIP-exo signal intensity values instead of binary labels during model training, enhancing the precision of predictions. BPNet also incorporates multitask learning, enabling the simultaneous discovery of multiple motifs. By integrating control experiment data, BPNet mitigates overfitting and enhances overall model performance. Considering that TF-DNA interactions involve complex long-term and short-term dependencies, recurrent neural networks

(RNNs) prove suitable for modeling such intricate relationships. RNNs were initially introduced to model sequential signals exhibiting stationary features over time. Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997), a variant of RNNs, adeptly capture both long-term and short-term dependencies by learning the positional dynamics of sequential signals. Bi-directional long/short-term memory (BLSTMs) networks represent variations of standard LSTMs, amalgamating the outputs of two RNNs—one analyzing sequential data from left to right and the other from right to left (corresponding to the forward and reverse strands in a genomic context). DeeperBind (Hassanzadeh and Wang, 2016) proposes a hybrid CNN-LSTMs architecture that effectively captures long-term and short-term positional dependencies. DeeperBind omits the pooling layer to prevent the loss of positional information concerning subregions activating the kernels in the convolutional stage. DanQ (Quang and Xie, 2016) introduces a hybrid architecture by incorporating CNNs and BLSTMs. The BLSTM layer replaces the fully connected neural network, followed by a dense layer of rectified linear units and a multi-task sigmoid output stage. Situating the BLSTM layer between the pooling and the dense rectified stages enables DanQ to capture the positional dynamics of the input sequences. The parameters of DanQ can be initialized with either random values or known motifs. The identified motifs are represented as PWMs, computed by aligning the sequences that activate the kernels. FactorNet (Quang and Xie, 2019) extends the DanQ framework by integrating additional features during model training, such as DNase-seq signals. To mitigate overfitting and reduce training complexity, FactorNet employs a Siamese architecture that accounts for the reverse complement. The Siamese architecture employs identical networks sharing weights for both the forward and reverse strands, ensuring consistent outputs and reducing the amount of required training data.

B.2 Catalogue of Motif discovery algorithms and software for discover Transcription Factor Binding sites in DNA sequences

Table B.2. Motif discovery algorithms and software for discover Transcription Factor Binding sites in DNA sequences. The table in this section lists the algorithms discussed in **Section 4.2.2**, and provides information for each algorithm, including associated publications (**Refs.**), the original data type used for development and testing in the original publication (**Original input data type**), the motif model returned (**Output**), theoretical advantages (**Pros**), drawbacks (**Cons**), the year of publication (**Year**), a link to the code or website (**Availability**), and details on how each software is distributed to the community (**Availability type**).

Motif discovery method	Algorithm	Refs	Original input data	Output	Pros	Cons	Year
Enumerative	SPLASH	Califano (2000)	DNA sequences (unspecified source)	Regular Expression	Simple output	Fixed mismatch positions	2000
	Weeder	Pavesi *et al.* (2001) Pavesi *et al.* (2004b)	DNA sequences (unspecified source)	PWM/ Consensus sequence	Simple output; use biological data as background dataset	Not scalable on large datasets	2001
	SMILE	Marsan and Sagot (2000)	DNA sequences (unspecified source)	Consensus sequence	Simple output	Not scalable on large datasets	2000
	MDScan	Liu *et al.* (2002)	ChIP-on-Chip	PWM	Use 3rd-order MMs as background	Based on word enumeration (computationally intensive)	2002
	Amadeus	Linhart *et al.* (2008)	ChIP-on-Chip	PWM	Use hypergeometric test to evaluate motifs	Based on word enumeration (computationally intensive)	2008
	DREME	Bailey (2011)	ChIP-seq	PWM/ consensus sequence	Detailed output; use Fisher's exact test to evaluate motifs	Based on regular expression	2011
	Trawler	Ettwiller *et al.* (2007)	ChIP-on-Chip	PWM	Use STs for fast motif search; use binomial distribution to evaluate motifs	Method performance rely on suitable background	2007
	HOMER	Heinz *et al.* (2010)	ChIP-seq	PWM/ consensus sequence	Detailed output; use ST for fast motif search; use hypergeometric distribution to evaluate motifs; well maintained	Motif statistical significance not reported	2010

Motif discovery method	Algorithm	Refs	Original input data	Output	Pros	Cons	Year
	STREME	Bailey (2021)	ChIP-seq	PWM/ consensus sequence	Detailed output; use ST and seed words evaluation for fast motif search	Assumes that each sequence contains one or zero motif occurrences; requires that more than one sequence contains the motif the report results	2021
Alignment-based	CONSENSUS	Hertz and Stormo (1999)	DNA sequences (unspecified source)	PWM	Simple approach	Method performance rely on motif conservation	1999
	MEME	Bailey *et al.* (1994) Bailey and Elkan (1995) Bailey *et al.* (2006)	DNA sequences (unspecified source)	PWM/ consensus sequence	Detailed output; fast motif search; well maintained	EM may converge prematurely	1994
	Motif Sampler	Neuwald *et al.* (1995)	DNA sequences (unspecified source)	Consensus sequence	Overcome EM limitation	Not scalable on large datasets; stochastic nature of Gibbs sampling	1995
	Align-ACE	Hughes *et al.* (2000)	DNA sequences (unspecified source)	PWM	Overcome EM limitation	Not scalable on large datasets; stochastic nature of Gibbs sampling	2000
	ANN-spec	Workman and Stormo (1999)	DNA sequences (unspecified source)	PWM	Overcome EM limitation	Not scalable on large datasets; use ANN replacing to replace alignment profiles (computationally intensive)	2000
	BioProspector	Liu *et al.* (2000)	ChIP-on-Chip	PWM	Use 3rd-order MMs as background	Not scalable on large datasets; stochastic nature of Gibbs sampling	2002
	MotifSampler	Thijs *et al.* (2001)	DNA sequences (unspecified source)	PWM	Use 3rd-order MMs as background	Not scalable on large datasets; stochastic nature of Gibbs sampling	2001
	GLAM2	Frith *et al.* (2004a) Frith *et al.* (2008)	DNA sequences (unspecified source)	PWM	Estimate optimal alignment length; allow gapped motifs; well maintained	Not scalable on large datasets; stochastic nature of Gibbs sampling	2008
	GADEM	Li (2009)	ChIP-on-Chip; ChIP-seq	PWM	Overcome EM limitation	Not scalable on large datasets; genetic algorithms are computationally intensive	2009
	MEME-ChIP	Machanick and Bailey (2011)	ChIP-seq	PWM	Scalable on large datasets; well maintained	Consider random subsamples of the input dataset	2011
	STEME	Reid and Wernisch (2011)	ChIP-on-Chip; ChIP-seq	PWM	Scalable on large datasets; faster MEME-like EM motif search	Computationally intensive ST construction	2011
	ChIPMunk	Kulakovskiy *et al.* (2010)	ChIP-seq	PWM	Scalable on large datasets; accounts for ChIP-seq peaks shape during motif search; well maintained	EM may converge prematurely	2010
	XXmotif	Hartmann *et al.* (2013)	ChIP-on-Chip; ChIP-seq	PWM	Integrate enumerativa and alignment-based motif search	May be computationally intensive on large datasets	2013
	ProSampler	Li *et al.* (2019b)	ChIP-seq	PWM	Scalable on large datasets; ultra-fast motif search	Stochastic nature of Gibbs sampling	2019
Probabilistic graphical models	Dimont	Grau *et al.* (2013)	ChIP-seq; ChIP-exo; PBM	Consensus sequence/ High-order PWM	Scalable on large datasets; include dependencies during motif search	Ignore non-neighboring dependencies	2013
	diChIPMunk	Kulakovskiy *et al.* (2013a)	ChIP-seq	High-order PWM	Include dependencies during motif search; well maintained	Not scalable on large datasets	2013
	TFFM	Mathelier and Wasserman (2013)	ChIP-seq	PWM/ High-order PWM	Scalable on large datasets; include dependencies during motif search; accomodate variable motif length	Depend on MEME motif discovery results	2013
	Discrover	Maaskola and Rajewsky (2014)	ChIP-seq	PWM	Scalable on large datasets; use HMMs	Ignore non-neighboring dependencies	2014
	Slim	Keilwagen and Grau (2015)	ChIP-seq; PBM	High-order PWM	Include neighboring and non-neighboring variable-order dependencies	Computationally intensive	2015
	BaMMotif	Siebert and Söding (2016) Ge *et al.* (2021)	ChIP-seq	High-order PWM	Include high-order dependencies; avoid model overfitting	Computationally intensive motif search; depends on XXmotif motif discovery results	2016

Motif discovery method	Algorithm	Refs	Original input data	Output	Pros	Cons	Year
Support Vector Machines	Kmer-SVM	Lee *et al.* (2011) Fletez-Brant *et al.* (2013)	ChIP-seq	PWM/ k-mer-based model	Scalable on large datasets	Limited to short motifs; Computationally intensive; rely on suitable background	2011
	Agius et al.	Agius *et al.* (2010)	ChIP-seq; PBM	PWM/ k-mer-based model	Scalable on large datasets	Limited to short motifs; Computationally intensive; rely on suitable background	2010
	Gkm-SVM	Ghandi *et al.* (2014a)	ChIP-seq	PWM/ k-mer-based model	Scalable on large datasets; consider long motifs using gapped k-mers	Computationally intensive; rely on suitable background	2014
	LS-GKM	Lee (2016)	ChIP-seq	PWM/ k-mer-based model	Improve Gkm-SVM perfomance on large datasets; introduce additional kernels	Rely on suitable background	2016
Deep Neural Networks	DeepBind	Alipanahi *et al.* (2015)	ChIP-seq; HT-SELEX; PBM	PWM/ DNN-based model	Scalable on large datasets; integrate different genomic data	Computationally intensive training; ignore kernel dependencies during motif reconstruction	2015
	Basset	Kelley *et al.* (2016)	ChIP-seq; DNase-seq	PWM/ DNN-based model	Scalable on large datasets; integrate different genomic data	Computationally intensive training; ignore kernel dependencies during motif reconstruction	2016
	BPNet	Avsec *et al.* (2021a)	ChIP-seq; ChIP-exo	PWM/ DNN-based model	Scalable on large datasets; integrate different genomic data; account for kernel dependencies during motif reconstruction; well maintained	Computationally intensive training	2021
	DeeperBind	Hassanzadeh and Wang (2016)	PBM	PWM/ DNN-based model	Scalable on large datasets; integrate short-range dependencies in sequences	Computationally intensive training; not maintained	2016
	DanQ	Quang and Xie (2016)	ChIP-seq; DNase-seq	PWM/ DNN-based model	Scalable on large datasets; integrate different genomic data; integrate long- and short-range dependencies in sequences	Computationally intensive training; weights differently sequences on positive and reverse strands	2016
	FactorNet	Quang and Xie (2019)	ChIP-seq; DNase-seq	PWM/ DNN-based model	Scalable on large datasets; integrate different genomic data; weights identically positive and reverse strands	Computationally intensive training	2019

References

Abadi, S., Yan, W. X., Amar, D., and Mayrose, I. (2017). A machine learning approach for predicting crispr-cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS computational biology*, **13**(10), e1005807.

Agius, P., Arvey, A., Chang, W., Noble, W. S., and Leslie, C. (2010). High resolution models of transcription factor-dna affinities improve in vitro and in vivo binding predictions. *PLoS computational biology*, **6**(9), e1000916.

Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W. F., Doshi, M. B., *et al.* (2016). Genomic copy number dictates a gene-independent cell response to crispr/cas9 targeting. *Cancer discovery*, **6**(8), 914–929.

Akcakaya, P., Bobbin, M. L., Guo, J. A., Malagon-Lopez, J., Clement, K., Garcia, S. P., Fellows, M. D., Porritt, M. J., Firth, M. A., Carreras, A., *et al.* (2018). In vivo crispr editing with no detectable genome-wide off-target mutations. *Nature*, **561**(7723), 416–419.

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS biology*, **2**(10), e286.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, **33**(8), 831–838.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature reviews genetics*, **12**(5), 363–376.

Allen, F., Crepaldi, L., Alsinet, C., Strong, A. J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., *et al.* (2019). Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nature biotechnology*, **37**(1), 64–72.

Alsøe, L., Sarno, A., Carracedo, S., Domanska, D., Dingler, F., Lirussi, L., SenGupta, T., Tekin, N. B., Jobert, L., Alexandrov, L. B., *et al.* (2017). Uracil accumulation and mutagenesis dominated by cytosine deamination in cpg dinucleotides in mice lacking ung and smug1. *Scientific reports*, **7**(1), 7199.

Amariuta, T., Luo, Y., Gazal, S., Davenport, E. E., van de Geijn, B., Ishigaki, K., Westra, H.-J., Teslovich, N., Okada, Y., Yamamoto, K., *et al.* (2019). Impact: genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors. *The American Journal of Human Genetics*, **104**(5), 879–895.

Ambrosini, G., Groux, R., and Bucher, P. (2018). Pwmscan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, **34**(14), 2483–2484.

Anders, C., Bargsten, K., and Jinek, M. (2016). Structural plasticity of pam recognition by engineered variants of the rna-guided endonuclease cas9. *Molecular cell*, **61**(6), 895–902.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.

Andreace, F., Lechat, P., Dufresne, Y., and Chikhi, R. (2023). Comparing methods for constructing and representing human pangenome graphs. *Genome Biology*, **24**(1), 274.

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., *et al.* (2019). Search-and-replace genome editing without double-strand breaks or donor dna. *Nature*, **576**(7785), 149–157.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., *et al.* (2020). Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**(7833), 246–251.

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, **17**(9), 507–522.

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A., *et al.* (2010). Clinical assessment incorporating a personal genome. *The Lancet*, **375**(9725), 1525–1535.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., *et al.* (2021a). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, **53**(3), 354–366.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021b). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, **18**(10), 1196–1203.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, **14**(3), 283–291.

Bae, S., Park, J., and Kim, J.-S. (2014). Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, **30**(10), 1473–1475.

Bailey, T. L. (2011). Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, **27**(12), 1653–1659.

Bailey, T. L. (2021). Streme: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**(18), 2834–2840.

Bailey, T. L. and Elkan, C. (1995). The value of prior knowledge in discovering motifs with meme. In *Ismb*, volume 3, pages 21–29.

Bailey, T. L., Elkan, C., *et al.* (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, **34**(suppl_2), W369–W373.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite: tools for motif discovery and searching. *Nucleic acids research*, **37**(suppl_2), W202–W208.

Balazadeh, S., Kwasniewski, M., Caldana, C., Mehrnia, M., Zanor, M. I., Xue, G.-P., and Mueller-Roeber, B. (2011). Ors1, an h2o2-responsive nac transcription factor, controls senescence in arabidopsis thaliana. *Molecular plant*, **4**(2), 346–360.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., *et al.* (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, **91**(2), 355–358.

Bao, X. R., Pan, Y., Lee, C. M., Davis, T. H., and Bao, G. (2021). Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nature protocols*, **16**(1), 10–26.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-dna binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 28–37.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T. (2011). Bamtools: a c++ api and toolkit for analyzing and managing bam files. *Bioinformatics*, **27**(12), 1691–1692.

Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., Gisselbrecht, S. S., Rossin, E. J., Woodard, J., Mariani, L., Kock, K. H., Inukai, S., *et al.* (2016). Survey of variation in human transcription factors reveals prevalent dna binding changes. *Science*, **351**(6280), 1450–1454.

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., *et al.* (2019). Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, **568**(7753), 511–516.

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein ctcf is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**(3), 387–396.

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., and Grosse, I. (2005). Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics*, **21**(11), 2657–2666.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, **4**(10), e1000173.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 1798–1828.

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein–dna interactions: how good an approximation is it? *Nucleic acids research*, **30**(20), 4442–4451.

Bentley, J. and Sedgewick, B. (1998). Ternary search trees. *Dr. Dobb's Journal*, **23**(4).

Berger, M. F. and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nature protocols*, **4**(3), 393–411.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, **24**(11), 1429–1435.

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., *et al.* (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, **367**(6484), eaay5012.

Blanchette, M. and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome research*, **12**(5), 739–748.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, **40**(6), 695–701.

Bodon, F. and Rónyai, L. (2003). Trie: an alternative data structure for data mining algorithms. *Mathematical and Computer Modelling*, **38**(7-9), 739–751.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in genetics*, **7**, 24.

Bollen, Y., Post, J., Koo, B.-K., and Snippert, H. J. (2018). How to create state-of-the-art genetic model systems: strategies for optimal crispr-mediated genome editing. *Nucleic acids research*, **46**(13), 6435–6454.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Bouhairie, V. E. and Goldberg, A. C. (2015). Familial hypercholesterolemia. *Cardiology clinics*, **33**(2), 169–179.

Bovolenta, L., Acencio, M., and Lemke, N. (2012). Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. *Nature Precedings*, pages 1–1.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., *et al.* (2012). Annotation of functional variation in personal genomes using regulomedb. *Genome research*, **22**(9), 1790–1797.

Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the hla genes in the 1000 genomes project phase i data. *G3: Genes, Genomes, Genetics*, **5**(5), 931–941.

Brinkman, E. K., Chen, T., Amendola, M., and Van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic acids research*, **42**(22), e168–e168.

Brown, M. S. and Goldstein, J. L. (1984). How ldl receptors influence cholesterol and atherosclerosis. *Scientific American*, **251**(5), 58–69.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, **10**(12), 1213–1218.

Bulyk, M. L., Johnson, P. L., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, **30**(5), 1255–1261.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.* (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, **47**(D1), D1005–D1012.

Butler, J. E. and Kadonaga, J. T. (2002). The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes & development*, **16**(20), 2583–2592.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Califano, A. (2000). Splash: structural pattern localization analysis by sequential histograms. *Bioinformatics*, **16**(4), 341–357.

Calligaris, R., Bottardi, S., Cogoi, S., Apezteguia, I., and Santoro, C. (1995). Alternative translation initiation site usage results in two functionally distinct forms of the gata-1 transcription factor. *Proceedings of the National Academy of Sciences*, **92**(25), 11598–11602.

Cancellieri, S., Canver, M. C., Bombieri, N., Giugno, R., and Pinello, L. (2020). Crispritz: rapid, high-throughput and variant-aware in silico off-target site identification for crispr genome editing. *Bioinformatics*, **36**(7), 2001–2008.

Cancellieri, S., Zeng, J., Lin, L. Y., Tognon, M., Nguyen, M. A., Lin, J., Bombieri, N., Maitland, S. A., Ciuculescu, M.-F., Katta, V., *et al.* (2023). Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nature Genetics*, **55**(1), 34–43.

Canver, M. C., Bauer, D. E., Dass, A., Yien, Y. Y., Chung, J., Masuda, T., Maeda, T., Paw, B. H., and Orkin, S. H. (2014). Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (crispr)/cas9 nuclease system in mammalian cells* . *Journal of Biological Chemistry*, **289**(31), 21312–21324.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., *et al.* (2015). Bcl11a enhancer dissection by cas9-mediated in situ saturating mutagenesis. *Nature*, **527**(7577), 192–197.

Carette, J. E., Guimaraes, C. P., Wuethrich, I., Blomen, V. A., Varadarajan, M., Sun, C., Bell, G., Yuan, B., Muellner, M. K., Nijman, S. M., *et al.* (2011). Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. *Nature biotechnology*, **29**(6), 542–546.

Cavalli-Sforza, L. L. (2005). The human genome diversity project: past, present and future. *Nature Reviews Genetics*, **6**(4), 333–340.

Chatterjee, P., Jakimo, N., Lee, J., Amrani, N., Rodriguez, T., Koseki, S. R., Tysinger, E., Qing, R., Hao, S., Sontheimer, E. J., *et al.* (2020). An engineered sccas9 with broad pam range and high specificity and activity. *Nature Biotechnology*, **38**(10), 1154–1158.

Chaudhari, H. G., Penterman, J., Whitton, H. J., Spencer, S. J., Flanagan, N., Lei Zhang, M. C., Huang, E., Khedkar, A. S., Toomey, J. M., Shearer, C. A., *et al.* (2020). Evaluation of homology-independent crispr-cas9 off-target assessment methods. *The CRISPR journal*, **3**(6), 440–453.

Chen, B., Wolfgang, C. D., and Hai, T. (1996). Analysis of atf3, a transcription factor induced by physiological stresses and modulated by gadd153/chop10. *Molecular and cellular biology*.

Chen, X., Chen, S., Li, Y., Gao, Y., Huang, S., Li, H., and Zhu, Y. (2019). Smurf1-mediated ubiquitination of arhgap26 promotes ovarian cancer cell invasion and migration. *Experimental & Molecular Medicine*, **51**(4), 1–12.

Cheng, L., Li, Y., Qi, Q., Xu, P., Feng, R., Palmer, L., Chen, J., Wu, R., Yee, T., Zhang, J., *et al.* (2021). Single-nucleotide-level mapping of dna regulatory elements that control fetal hemoglobin expression. *Nature genetics*, **53**(6), 869–880.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., and Kim, J.-S. (2014). Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickases. *Genome research*, **24**(1), 132–141.

Choi, Y. and Chan, A. P. (2015). Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**(16), 2745–2747.

Chu, S. H., Packer, M., Rees, H., Lam, D., Yu, Y., Marshall, J., Cheng, L.-I., Lam, D., Olins, J., Ran, F. A., *et al.* (2021). Rationally designed base editors for precise editing of the sickle cell disease mutation. *The CRISPR Journal*, **4**(2), 169–177.

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E., *et al.* (2019). Crispresso2 provides accurate and rapid genome editing sequence analysis. *Nature biotechnology*, **37**(3), 224–226.

Clement, K., Hsu, J. Y., Canver, M. C., Joung, J. K., and Pinello, L. (2020). Technologies and computational analysis strategies for crispr applications. *Molecular cell*, **79**(1), 11–29.

Coelho, M. A., Cooper, S., Strauss, M. E., Karakoc, E., Bhosle, S., Gonçalves, E., Picco, G., Burgold, T., Cattaneo, C. M., Veninga, V., *et al.* (2023). Base editing screens map mutations affecting interferon-γ signaling in cancer. *Cancer Cell*, **41**(2), 288–303.

Collas, P. and Dahl, J. A. (2008). Chop it, chip it, check it: the current status of chromatin immunoprecipitation. *Frontiers in Bioscience-Landmark*, **13**(3), 929–943.

Conant, D., Hsiau, T., Rossi, N., Oki, J., Maures, T., Waite, K., Yang, J., Joshi, S., Kelso, R., Holden, K., *et al.* (2022). Inference of crispr edits from sanger trace data. *The CRISPR journal*, **5**(1), 123–130.

Concordet, J.-P. and Haeussler, M. (2018). Crispor: intuitive guide selection for crispr/cas9 genome editing experiments and screens. *Nucleic acids research*, **46**(W1), W242–W245.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., *et al.* (2013). Multiplex genome engineering using crispr/cas systems. *Science*, **339**(6121), 819–823.

Consortium, . G. P. *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), 57.

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.

Consortium, T. C. P.-G. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, **19**(1), 118–135.

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., *et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, **48**(10), 1193–1203.

Cuella-Martin, R., Hayward, S. B., Fan, X., Chen, X., Huang, J.-W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C., *et al.* (2021). Functional interrogation of dna damage response variants with base editing screens. *Cell*, **184**(4), 1081–1097.

Das, M. K. and Dai, H.-K. (2007). A survey of dna motif finding algorithms. *BMC bioinformatics*, **8**(7), 1–13.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., *et al.* (2018). The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, **46**(D1), D794–D801.

Day, W. H. and McMorris, F. (1992). Critical comparison of consensus methods for molecular sequences. *Nucleic acids research*, **20**(5), 1093–1099.

De Dreuzy, E., Heath, J., Zuris, J. A., Sousa, P., Viswanathan, R., Scott, S., Da Silva, J., Ta, T., Capehart, S., Wang, T., *et al.* (2019). Edit-301: an experimental autologous cell therapy comprising cas12a-rnp modified mpb-cd34+ cells for the potential treatment of scd. *Blood*, **134**, 4636.

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., De Jong, P., *et al.* (2006). A regulatory snp causes a human genetic disease by creating a new transcriptional promoter. *Science*, **312**(5777), 1215–1217.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, **25**(24), 3207–3212.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., and Charpentier, E. (2011). Crispr rna maturation by trans-encoded small rna and host factor rnase iii. *Nature*, **471**(7340), 602–607.

Demirci, S., Zeng, J., Wu, Y., Uchida, N., Gamer, J., Yapundich, M., Drysdale, C., Bonifacino, A. C., Krouse, A. E., Linde, N. S., *et al.* (2019). Durable and robust fetal globin induction without anemia in rhesus monkeys following autologous hematopoietic stem cell transplant with bcl11a erythroid enhancer editing. *Blood*, **134**, 4632.

Dempster, J. M., Boyle, I., Vazquez, F., Root, D. E., Boehm, J. S., Hahn, W. C., Tsherniak, A., and McFarland, J. M. (2021). Chronos: a cell population dynamics model of crispr experiments that improves inference of gene fitness effects. *Genome biology*, **22**, 1–23.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor dna binding variation. *Cell*, **166**(3), 538–554.

Després, P. C., Dubé, A. K., Seki, M., Yachie, N., and Landry, C. R. (2020). Perturbing proteomes at single residue resolution using base editing. *Nature communications*, **11**(1), 1871.

DeWitt, M. A., Magis, W., Bray, N. L., Wang, T., Berman, J. R., Urbinati, F., Heo, S.-J., Mitros, T., Muñoz, D. P., Boffelli, D., *et al.* (2016). Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Science translational medicine*, **8**(360), 360ra134–360ra134.

D'haeseleer, P. (2006). How does dna sequence motif discovery work? *Nature biotechnology*, **24**(8), 959–961.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**(1), 15–21.

Docquier, F., Farrar, D., D'Arcy, V., Chernukhin, I., Robinson, A. F., Loukinov, D., Vatolin, S., Pack, S., Mackay, A., Harris, R. A., *et al.* (2005). Heightened expression of ctcf in breast cancer cells is associated with resistance to apoptosis. *Cancer research*, **65**(12), 5112–5122.

Doench, J. G. (2018). Am i ready for crispr? a user's guide to genetic screens. *Nature Reviews Genetics*, **19**(2), 67–80.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., and Root, D. E. (2014). Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nature biotechnology*, **32**(12), 1262–1267.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al.* (2016). Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, **34**(2), 184–191.

Du, Q., Luu, P.-L., Stirzaker, C., and Clark, S. J. (2015). Methyl-cpg-binding domain proteins: readers of the epigenome. *Epigenomics*, **7**(6), 1051–1073.

Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., *et al.* (2019). Model-based understanding of single-cell crispr screening. *Nature communications*, **10**(1), 2233.

Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in dna. *Nature*, **287**(5782), 560–561.

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., *et al.* (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature genetics*, **54**(4), 518–525.

Eggeling, R., Gohr, A., Keilwagen, J., Mohr, M., Posch, S., Smith, A. D., and Grosse, I. (2014). On the value of intra-motif dependencies of human insulator protein ctcf. *PLoS One*, **9**(1), e85629.

Eskin, E., Weston, J., Noble, W., and Leslie, C. (2002). Mismatch string kernels for svm protein classification. *Advances in neural information processing systems*, **15**.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature methods*, **4**(7), 563–565.

Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The international genome sample resource (igsr) collection of open human genomic variation resources. *Nucleic acids research*, **48**(D1), D941–D947.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., *et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**(7539), 337–343.

Fennell, T., Zhang, D., Isik, M., Wang, T., Gotta, G., Wilson, C. J., and Marco, E. (2021). Calitas: a crispr-cas-aware aligner for in silico off-target search. *The CRISPR Journal*, **4**(2), 264–274.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, **7**(2), 85–97.

Finkel, R. S., Mercuri, E., Darras, B. T., Connolly, A. M., Kuntz, N. L., Kirschner, J., Chiriboga, C. A., Saito, K., Servais, L., Tizzano, E., *et al.* (2017). Nusinersen versus sham control in infantile-onset spinal muscular atrophy. *N Engl J Med*, **377**, 1723–1732.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., *et al.* (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, **50**(4), 621–629.

Fiorentino, F. P. and Giordano, A. (2012). The tumor suppressor role of ctcf. *Journal of cellular physiology*, **227**(2), 479–492.

Fischer, J. and Kurpicz, F. (2017). Dismantling divsufsort. *arXiv preprint arXiv:1710.01896*.

Fletez-Brant, C., Lee, D., McCallion, A. S., and Beer, M. A. (2013). kmer-svm: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic acids research*, **41**(W1), W544–W556.

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., *et al.* (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, **48**(D1), D87–D92.

Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., *et al.* (2021). Crispr-cas9 gene editing for sickle cell disease and β-thalassemia. *New England Journal of Medicine*, **384**(3), 252–260.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., *et al.* (2019). Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, **47**(D1), D766–D773.

Frith, M. C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004a). Detection of functional dna motifs via statistical over-representation. *Nucleic acids research*, **32**(4), 1372–1381.

Frith, M. C., Hansen, U., Spouge, J. L., and Weng, Z. (2004b). Finding functional sequence elements by multiple local alignment. *Nucleic acids research*, **32**(1), 189–200.

Frith, M. C., Saunders, N. F., Kobe, B., and Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS computational biology*, **4**(5), e1000071.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., and Sander, J. D. (2013). High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, **31**(9), 822–826.

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, **461**(7261), 186–192.

Galas, D. J. and Schmitz, A. (1978). Dnaase footprinting a simple method for the detection of protein-dna binding specificity. *Nucleic acids research*, **5**(9), 3157–3170.

Gallagher, M. D. and Chen-Plotkin, A. S. (2018). The post-gwas era: from association to function. *The American Journal of Human Genetics*, **102**(5), 717–730.

Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., Wang, B., Kong, S., Li, S., Cui, Y., *et al.* (2023). A pangenome reference of 36 chinese populations. *Nature*, pages 1–10.

Garcia, E. M., Lue, N. Z., Liang, J. K., Lieberman, W. K., Hwang, D. D., Woods, J. C., and Liau, B. B. (2023). Base editor scanning reveals activating mutations of dnmt3a. *ACS Chemical Biology*.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, **29**(8), 1363–1375.

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H., and Moineau, S. (2010). The crispr/cas bacterial immune system cleaves bacteriophage and plasmid dna. *Nature*, **468**(7320), 67–71.

Garner, M. M. and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific dna regions: application to components of the escherichia coli lactose operon regulatory system. *Nucleic acids research*, **9**(13), 3047–3060.

Garrison, E. and Guarracino, A. (2023). Unbiased pangenome graphs. *Bioinformatics*, **39**(1), btac743.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., *et al.* (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, **36**(9), 875–879.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., *et al.* (2023). Building pangenome graphs. *bioRxiv*, pages 2023–04.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, **7**(9), 703–713.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., and Liu, D. R. (2017). Programmable base editing of a• t to g• c in genomic dna without dna cleavage. *Nature*, **551**(7681), 464–471.

Ge, W., Meier, M., Roth, C., and Söding, J. (2021). Bayesian markov models improve the prediction of binding motifs beyond first order. *NAR genomics and bioinformatics*, **3**(2), lqab026.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014a). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, **10**(7), e1003711.

Ghandi, M., Mohammad-Noori, M., and Beer, M. A. (2014b). Robust $$ k $$ k-mer frequency estimation using gapped $$ k $$ k-mers. *Journal of mathematical biology*, **69**(2), 469–500.

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmsvm: an r package for gapped-kmer svm. *Bioinformatics*, **32**(14), 2205–2207.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., *et al.* (2003). The international hapmap project. *Nature*.

Gillmore, J. D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M. L., Seitzer, J., O'Connell, D., Walsh, K. R., Wood, K., *et al.* (2021). Crispr-cas9 in vivo gene editing for transthyretin amyloidosis. *New England Journal of Medicine*, **385**(6), 493–502.

Ginsburg, G. S. and Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. *Translational research*, **154**(6), 277–287.

Glenwinkel, L., Wu, D., Minevich, G., and Hobert, O. (2014). Targetortho: a phylogenetic footprinting tool to identify transcription factor targets. *Genetics*, **197**(1), 61–76.

Gorkin, D. U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S. L., Loftus, S. K., Beer, M. A., Pavan, W. J., and McCallion, A. S. (2012). Integration of chip-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome research*, **22**(11), 2290–2301.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome research*, **20**(5), 565–577.

Graham, S. E., Clarke, S. L., Wu, K.-H. H., Kanoni, S., Zajac, G. J., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T. W., *et al.* (2021). The power of genetic diversity in genome-wide association studies of lipids. *Nature*, **600**(7890), 675–679.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7), 1017–1018.

Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, **41**(21), e197–e197.

Griffin, A. M., Griffin, H. G., and Staden, R. (1994). Staden: searching for motifs in nucleic acid sequences. *Computer Analysis of Sequence Data: Part II*, pages 93–102.

Groza, C., Kwan, T., Soranzo, N., Pastinen, T., and Bourque, G. (2020). Personalized and graph genomes reveal missing signal in epigenomic data. *Genome biology*, **21**(1), 1–22.

Grünewald, J., Zhou, R., Lareau, C. A., Garcia, S. P., Iyer, S., Miller, B. R., Langner, L. M., Hsu, J. Y., Aryee, M. J., and Joung, J. K. (2020). A dual-deaminase crispr base editor enables concurrent adenine and cytosine editing. *Nature biotechnology*, **38**(7), 861–864.

Guo, Y., Mahony, S., and Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS computational biology*.

Gupta, S., Stamatoyannopoulos, J., Bailey, T., and Stafford, W. (2007). Quantifying similarity between motifs. genome biology.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., *et al.* (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor. *Genome biology*, **17**, 1–12.

Hamilton, M. C., Fife, J. D., Akinci, E., Yu, T., Khowpinitchai, B., Cha, M., Barkal, S., Thi, T. T., Yeo, G. H., Barroso, J. P. R., *et al.* (2023). Systematic elucidation of genetic mechanisms underlying cholesterol uptake. *Cell Genomics*, **3**(5).

Hampshire, A. J., Rusling, D. A., Broughton-Head, V. J., and Fox, K. R. (2007). Footprinting: a method for determining the sequence selectivity, affinity and kinetics of dna-binding ligands. *Methods*, **42**(2), 128–140.

Hanna, R. E. and Doench, J. G. (2020). Design and analysis of crispr–cas experiments. *Nature biotechnology*, **38**(7), 813–823.

Hanna, R. E., Hegde, M., Fagre, C. R., DeWeirdt, P. C., Sangree, A. K., Szegletes, Z., Griffith, A., Feeley, M. N., Sanson, K. R., Baidi, Y., *et al.* (2021). Massively parallel assessment of human variants with base editor screens. *Cell*, **184**(4), 1064–1080.

Hart, T. and Moffat, J. (2016). Bagel: a computational framework for identifying essential genes from pooled library screens. *BMC bioinformatics*, **17**, 1–7.

Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology*, **10**(7), 733.

Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome research*, **23**(1), 181–194.

Hassanzadeh, H. R. and Wang, M. D. (2016). Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 178–183. IEEE.

He, Y., Shen, Z., Zhang, Q., Wang, S., and Huang, D.-S. (2021). A survey on deep learning in dna/rna motif mining. *Briefings in Bioinformatics*, **22**(4), bbaa229.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, **38**(4), 576–589.

Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D., and Marraffini, L. A. (2015). Cas9 specifies functional viral targets during crispr–cas adaptation. *Nature*, **519**(7542), 199–202.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, **15**(7), 563–577.

Ho, D., Quake, S. R., McCabe, E. R., Chng, W. J., Chow, E. K., Ding, X., Gelb, B. D., Ginsburg, G. S., Hassenstab, J., Ho, C.-M., *et al.* (2020). Enabling technologies for personalized and precision medicine. *Trends in biotechnology*, **38**(5), 497–518.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

Holley, G. and Melsted, P. (2020). Bifrost: highly parallel construction and indexing of colored and compacted de bruijn graphs. *Genome biology*, **21**(1), 1–20.

Hsu, J. Y., Fulco, C. P., Cole, M. A., Canver, M. C., Pellin, D., Sher, F., Farouni, R., Clement, K., Guo, J. A., Biasco, L., *et al.* (2018). Crispr-surf: discovering regulatory elements by deconvolution of crispr tiling screen data. *Nature methods*, **15**(12), 992–993.

Hsu, J. Y., Grünewald, J., Szalay, R., Shih, J., Anzalone, A. V., Lam, K. C., Shen, M. W., Petri, K., Liu, D. R., Joung, J. K., *et al.* (2021). Primedesign software for rapid and simplified design of prime editing guide rnas. *Nature Communications*, **12**(1), 1034.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., *et al.* (2013). Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, **31**(9), 827–832.

Hsu, P. D., Lander, E. S., and Zhang, F. (2014). Development and applications of crispr-cas9 for genome engineering. *Cell*, **157**(6), 1262–1278.

Huang, C., Li, G., Wu, J., Liang, J., and Wang, X. (2021a). Identification of pathogenic variants in cancer genes using base editing screens with editing efficiency correction. *Genome Biology*, **22**, 1–25.

Huang, Q., Tan, Z., Li, Y., Wang, W., Lang, M., Li, C., and Guo, Z. (2021b). Tfcancer: a manually curated database of transcription factors associated with human cancers. *Bioinformatics*, **37**(22), 4288–4290.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of molecular biology*, **296**(5), 1205–1214.

Hurgobin, B. and Edwards, D. (2017). Snp discovery using a pangenome: has the single reference approach become obsolete? *Biology*, **6**(1), 21.

Hwang, G.-H., Park, J., Lim, K., Kim, S., Yu, J., Yu, E., Kim, S.-T., Eils, R., Kim, J.-S., and Bae, S. (2018). Web-based design and analysis tools for crispr base editing. *BMC bioinformatics*, **19**, 1–7.

Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y., and Lauschke, V. M. (2018). Integrating rare genetic variants into pharmacogenetic drug response predictions. *Human genomics*, **12**, 1–12.

Iorio, F., Behan, F. M., Gonçalves, E., Bhosle, S. G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., *et al.* (2018). Unsupervised correction of gene-independent cell responses to crispr-cas9 targeting. *BMC genomics*, **19**, 1–16.

Ishihara, K., Oshimura, M., and Nakao, M. (2006). Ctcf-dependent chromatin insulator is linked to epigenetic remodeling. *Molecular cell*, **23**(5), 733–742.

Jeong, H.-H., Kim, S. Y., Rousseaux, M. W., Zoghbi, H. Y., and Liu, Z. (2019). Beta-binomial modeling of crispr pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome research*, **29**(6), 999–1008.

Jia, G., Wang, X., and Xiao, G. (2017). A permutation-based non-parametric analysis of crispr screen data. *BMC genomics*, **18**(1), 1–11.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, **337**(6096), 816–821.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, **43**(3), 264–268.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, **316**(5830), 1497–1502.

Jolma, A. and Taipale, J. (2011). Methods for analysis of transcription factor dna-binding specificity in vitro. *A Handbook of Transcription Factors*, pages 155–173.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., *et al.* (2010). Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, **20**(6), 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., *et al.* (2013). Dna-binding specificities of human transcription factors. *Cell*, **152**(1-2), 327–339.

Kaelin Jr, W. G. (2012). Use and abuse of rnai to study mammalian gene function. *Science*, **337**(6093), 421–422.

Kantor, A., McClements, M. E., and MacLaren, R. E. (2020). Crispr-cas9 dna base-editing and prime-editing. *International journal of molecular sciences*, **21**(17), 6240.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., *et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**(7809), 434–443.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., *et al.* (2010). Variation in transcription factor binding among humans. *science*, **328**(5975), 232–235.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., *et al.* (2015). Ctcf/cohesin-binding sites are frequently mutated in cancer. *Nature genetics*, **47**(7), 818–821.

Katara, P., Grover, A., and Sharma, V. (2012). Phylogenetic footprinting: a boost for microbial regulatory genomics. *Protoplasma*, **249**(4), 901–907.

Keenan, A. B., Torre, D., Lachmann, A., Leong, A. K., Wojciechowicz, M. L., Utti, V., Jagodnik, K. M., Kropiwnicki, E., Wang, Z., and Ma'ayan, A. (2019). Chea3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic acids research*, **47**(W1), W212–W224.

Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic acids research*, **43**(18), e119–e119.

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, **26**(7), 990–999.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, **28**(5), 739–750.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17), 2204–2207.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, **37**(8), 907–915.

Kim, H. K., Lee, S., Kim, Y., Park, J., Min, S., Choi, J. W., Huang, T. P., Yoon, S., Liu, D. R., and Kim, H. H. (2020). High-throughput analysis of the activities of xcas9, spcas9-ng and spcas9 at matched and mismatched target sequences in human cells. *Nature biomedical engineering*, **4**(1), 111–124.

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., *et al.* (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods*, **15**(8), 591–594.

Kim, Y., Lee, S., Cho, S., Park, J., Chae, D., Park, T., Minna, J. D., and Kim, H. H. (2022). High-throughput functional evaluation of human cancer-associated mutations using base editors. *Nature biotechnology*, **40**(6), 874–884.

Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T., and Liu, D. R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered cas9-cytidine deaminase fusions. *Nature biotechnology*, **35**(4), 371–376.

Kitts, A. and Sherry, S. (2002). The single nucleotide polymorphism database (dbsnp) of nucleotide sequence variation. *The NCBI handbook. McEntyre J, Ostell J, eds. Bethesda, MD: US national center for biotechnology information.*

Klein, M., Eslami-Mossallam, B., Arroyo, D. G., and Depken, M. (2018). Hybridization kinetics explains crispr-cas off-targeting rules. *Cell reports*, **22**(6), 1413–1423.

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P., Li, Z., Peterson, R. T., Yeh, J.-R. J., *et al.* (2015). Engineered crispr-cas9 nucleases with altered pam specificities. *Nature*, **523**(7561), 481–485.

Klimentidis, Y. C., Arora, A., Newell, M., Zhou, J., Ordovas, J. M., Renquist, B. J., and Wood, A. C. (2020). Phenotypic and genetic characterization of lower ldl cholesterol and increased type 2 diabetes risk in the uk biobank. *Diabetes*, **69**(10), 2194–2205.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**(3), 568–576.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral crispr-guide rna library. *Nature biotechnology*, **32**(3), 267–273.

Kolli, N., Lu, M., Maiti, P., Rossignol, J., and Dunbar, G. L. (2018). Application of the gene editing tool, crispr-cas9, for treating neurodegenerative diseases. *Neurochemistry international*, **112**, 187–196.

Kolmykov, S., Yevshin, I., Kulyashov, M., Sharipov, R., Kondrakhin, Y., Makeev, V. J., Kulakovskiy, I. V., Kel, A., and Kolpakov, F. (2021). Gtrd: an integrated view of transcription regulation. *Nucleic acids research*, **49**(D1), D104–D111.

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., and Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & chromatin*, **5**(1), 1–8.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature*, **533**(7603), 420–424.

Komor, A. C., Badran, A. H., and Liu, D. R. (2017). Crispr-based technologies for the manipulation of eukaryotic genomes. *Cell*, **168**(1), 20–36.

Koo, P. K. and Ploenzke, M. (2020). Deep learning for inferring transcription factor binding sites. *Current opinion in systems biology*, **19**, 16–23.

Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). Moods: fast search for position weight matrix matches in dna sequences. *Bioinformatics*, **25**(23), 3181–3182.

Korhonen, J. H., Palin, K., Taipale, J., and Ukkonen, E. (2017). Fast motif matching revisited: high-order pwms, snps and indels. *Bioinformatics*, **33**(4), 514–521.

Kou, Y., Chen, E. Y., Clark, N. R., Duan, Q., Tan, C. M., and Ma 'ayan, A. (2013). Chea2: gene-set libraries from chip-x experiments to decode the transcription regulome. In *Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013. Proceedings 8*, pages 416–430. Springer.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2005). Profile-based string kernels for remote homology detection and motif extraction. *Journal of bioinformatics and computational biology*, **3**(03), 527–550.

Kulakovskiy, I. and Makeev, V. (2009). Discovery of dna motifs recognized by transcription factors through integration of different experimental sources. *Biophysics*, **54**(6), 667–674.

Kulakovskiy, I., Levitsky, V., Oshchepkov, D., Bryzgalov, L., Vorontsov, I., and Makeev, V. (2013a). From binding motifs in chip-seq data to improved models of transcription factor binding sites. *Journal of bioinformatics and computational biology*, **11**(01), 1340004.

Kulakovskiy, I. V., Boeva, V., Favorov, A. V., and Makeev, V. J. (2010). Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, **26**(20), 2622–2623.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013b). Hocomoco: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, **41**(D1), D195–D202.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., *et al.* (2018). Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic acids research*, **46**(D1), D252–D259.

Kunkel, T. A. (1984). Mutational specificity of depurination. *Proceedings of the National Academy of Sciences*, **81**(5), 1494–1498.

Kurt, I. C., Zhou, R., Iyer, S., Garcia, S. P., Miller, B. R., Langner, L. M., Grünewald, J., and Joung, J. K. (2021). Crispr c-to-g base editors for inducing targeted dna transversions in human cells. *Nature biotechnology*, **39**(1), 41–46.

Kweon, J., Jang, A.-H., Shin, H. R., See, J.-E., Lee, W., Lee, J. W., Chang, S., Kim, K., and Kim, Y. (2020). A crispr-based base-editing screen for the functional assessment of brca1 variants. *Oncogene*, **39**(1), 30–35.

Kwon, A. T., Arenillas, D. J., Hunt, R. W., and Wasserman, W. W. (2012). opossum-3: advanced analysis of regulatory motif over-representation across genes or chip-seq datasets. *G3: Genes| Genomes| Genetics*, **2**(9), 987–1002.

Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., and Valen, E. (2019). Chopchop v3: expanding the crispr web toolbox beyond genome editing. *Nucleic acids research*, **47**(W1), W171–W174.

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., and Ma'ayan, A. (2010). Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments. *Bioinformatics*, **26**(19), 2438–2444.

Lakich, D., Kazazian Jr, H. H., Antonarakis, S. E., and Gitschier, J. (1993). Inversions disrupting the factor viii gene are a common cause of severe haemophilia a. *Nature genetics*, **5**(3), 236–241.

Lambert, S. A., Albu, M., Hughes, T. R., and Najafabadi, H. S. (2016). Motif comparison based on similarity of binding affinity profiles. *Bioinformatics*, **32**(22), 3504–3506.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, **172**(4), 650–665.

Landau, G. M., Kasai, T., Lee, G., Arimura, H., Arikawa, S., and Park, K. (2001). Linear-time longest-common-prefix computation in suffix arrays and its applications. In *Combinatorial Pattern Matching: 12th Annual Symposium, CPM 2001 Jerusalem, Israel, July 1–4, 2001 Proceedings 12*, pages 181–192. Springer.

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., *et al.* (2020). Clinvar: improvements to accessing data. *Nucleic acids research*, **48**(D1), D835–D844.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357–359.

Lapinaite, A., Knott, G. J., Palumbo, C. M., Lin-Shiao, E., Richter, M. F., Zhao, K. T., Beal, P. A., Liu, D. R., and Doudna, J. A. (2020). Dna capture by a crispr-cas9–guided adenine base editor. *Science*, **369**(6503), 566–571.

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, **29**(12), 1305–1312.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, **7**(1), 41–51.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, **262**(5131), 208–214.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, **521**(7553), 436–444.

Lee, C., Grasso, C., and Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**(3), 452–464.

Lee, C. M., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Nassar, L. R., Powell, C. C., *et al.* (2020). Ucsc genome browser enters 20th year. *Nucleic acids research*, **48**(D1), D756–D761.

Lee, D. (2016). Ls-gkm: a new gkm-svm for large-scale datasets. *Bioinformatics*, **32**(14), 2196–2198.

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from dna sequence. *Genome research*, **21**(12), 2167–2180.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. (2015). A method to predict the impact of regulatory variants from dna sequence. *Nature genetics*, **47**(8), 955–961.

Lee, N. K., Li, X., and Wang, D. (2018). A comprehensive survey on genetic algorithms for dna motif prediction. *Information Sciences*, **466**, 25–43.

Lei, Y., Lu, L., Liu, H.-Y., Li, S., Xing, F., and Chen, L.-L. (2014). Crispr-p: a web tool for synthetic single-guide rna design of crispr-system in plants. *Molecular plant*, **7**(9), 1494–1496.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes & development*, **14**(20), 2551–2569.

Leslie, C. and Kuang, R. (2003). Fast kernels for inexact string matching. In *Learning Theory and Kernel Machines*, pages 114–128. Springer.

Leslie, C., Eskin, E., and Noble, W. S. (2001). The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific.

Lessard, S., Francioli, L., Alfoldi, J., Tardif, J.-C., Ellinor, P. T., MacArthur, D. G., Lettre, G., Orkin, S. H., and Canver, M. C. (2017). Human genetic variation alters crispr-cas9 on-and off-targeting specificity at therapeutically implicated loci. *Proceedings of the National Academy of Sciences*, **114**(52), E11257–E11266.

Levinson, G. and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for dna sequence evolution. *Molecular biology and evolution*, **4**(3), 203–221.

Li, C., Chu, W., Gill, R. A., Sang, S., Shi, Y., Hu, X., Yang, Y., Zaman, Q. U., and Zhang, B. (2023). Computational tools and resources for crispr/cas genome editing. *Genomics, Proteomics and Bioinformatics*, **21**(1), 108–126.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, **25**(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009a). The sequence alignment/map format and samtools. *bioinformatics*, **25**(16), 2078–2079.

Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome biology*, **21**, 1–19.

Li, L. (2009). Gadem: a genetic algorithm guided formation of spaced dyads coupled with an em algorithm for motif discovery. *Journal of Computational Biology*, **16**(2), 317–329.

Li, M., Ma, B., and Wang, L. (1999). Finding similar regions in many strings. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 473–482.

Li, S. and Ovcharenko, I. (2015). Human enhancers are fragile and prone to deactivating mutations. *Molecular biology and evolution*, **32**(8), 2161–2180.

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., and Liu, X. S. (2014). Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome biology*, **15**(12), 1–12.

Li, W., Köster, J., Xu, H., Chen, C.-H., Xiao, T., Liu, J. S., Brown, M., and Liu, X. S. (2015). Quality control, modeling, and visualization of crispr screens with mageck-vispr. *Genome biology*, **16**, 1–13.

Li, W., Wong, W. H., and Jiang, R. (2019a). Deeptact: predicting 3d chromatin contacts via bootstrapping deep learning. *Nucleic acids research*, **47**(10), e60–e60.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009b). Genotype imputation. *Annual review of genomics and human genetics*, **10**, 387–406.

Li, Y., Ni, P., Zhang, S., Li, G., and Su, Z. (2019b). Prosampler: an ultrafast and accurate motif finder in large chip-seq datasets for combinatory motif discovery. *Bioinformatics*, **35**(22), 4632–4639.

Li, Y. I., Van De Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016). Rna splicing is a primary link between genetic variation and disease. *Science*, **352**(6285), 600–604.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., *et al.* (2023). A draft human pangenome reference. *Nature*, **617**(7960), 312–324.

Linhart, C., Halperin, Y., and Shamir, R. (2008). Transcription factor and microrna motif discovery: the amadeus platform and a compendium of metazoan target sets. *Genome research*, **18**(7), 1180–1189.

Link, V. M., Romanoski, C. E., Metzler, D., and Glass, C. K. (2018). Mmarge: motif mutation analysis for regulatory genomic elements. *Nucleic acids research*, **46**(14), 7006–7021.

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G., *et al.* (2018). Prediction of off-target activities for the end-to-end design of crispr guide rnas. *Nature biomedical engineering*, **2**(1), 38–47.

Liu, B., Yang, J., Li, Y., McDermaid, A., and Ma, Q. (2018). An algorithmic perspective of de novo cis-regulatory motif finding based on chip-seq data. *Briefings in bioinformatics*, **19**(5), 1069–1081.

Liu, F., Wang, L., Perna, F., and Nimer, S. D. (2016). Beyond transcription factors: how oncogenic signalling reshapes the epigenetic landscape. *Nature Reviews Cancer*, **16**(6), 359.

Liu, X., Brutlag, D. L., and Liu, J. S. (2000). Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Biocomputing 2001*, pages 127–138. World Scientific.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein–dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, **20**(8), 835–839.

Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P., Consortium, . G. P., *et al.* (2019). Variant calling on the grch38 assembly with the data from phase three of the 1000 genomes project. *Wellcome Open Research*, **4**.

Lue, N. Z., Garcia, E. M., Ngan, K. C., Lee, C., Doench, J. G., and Liau, B. B. (2023). Base editor scanning charts the dnmt3a activity landscape. *Nature Chemical Biology*, **19**(2), 176–186.

Ma, Y., Walsh, M. J., Bernhardt, K., Ashbaugh, C. W., Trudeau, S. J., Ashbaugh, I. Y., Jiang, S., Jiang, C., Zhao, B., Root, D. E., *et al.* (2017). Crispr/cas9 screens reveal epstein-barr virus-transformed b cell host dependency factors. *Cell host & microbe*, **21**(5), 580–591.

Maaskola, J. and Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden markov models. *Nucleic acids research*, **42**(21), 12995–13011.

Machanick, P. and Bailey, T. L. (2011). Meme-chip: motif analysis of large dna datasets. *Bioinformatics*, **27**(12), 1696–1697.

Macintyre, G., Bailey, J., Haviv, I., and Kowalczyk, A. (2010). is-rsnp: a novel technique for in silico regulatory snp detection. *Bioinformatics*, **26**(18), i524–i530.

Maeder, M. L., Stefanidakis, M., Wilson, C. J., Baral, R., Barrera, L. A., Bounoutas, G. S., Bumcrot, D., Chao, H., Ciulla, D. M., DaSilva, J. A., *et al.* (2019). Development of a gene-editing approach to restore vision loss in leber congenital amaurosis type 10. *Nature medicine*, **25**(2), 229–233.

Mahony, S. and Benos, P. V. (2007). Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic acids research*, **35**(suppl_2), W253–W258.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., *et al.* (2020). Evolutionary classification of crispr–cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*, **18**(2), 67–83.

Mali, P., Esvelt, K. M., and Church, G. M. (2013). Cas9 as a versatile tool for engineering biology. *Nature methods*, **10**(10), 957–963.

Manzanarez-Ozuna, E., Flores, D.-L., Gutiérrez-López, E., Cervantes, D., and Juárez, P. (2018). Model based on ga and dnn for prediction of mrna-smad7 expression regulated by mirnas in breast cancer. *Theoretical Biology and Medical Modelling*, **15**(1), 1–12.

Mao, Z., Bozzella, M., Seluanov, A., and Gorbunova, V. (2008). Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA repair*, **7**(10), 1765–1771.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, **39**(7), 906–913.

Mardis, E. R. (2007). Chip-seq: welcome to the new frontier. *Nature methods*, **4**(8), 613–614.

Marsan, L. and Sagot, M.-F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of computational biology*, **7**(3-4), 345–362.

Martin-Rufino, J. D., Castano, N., Pang, M., Grody, E. I., Joubran, S., Caulier, A., Wahlster, L., Li, T., Qiu, X., Riera-Escandell, A. M., *et al.* (2023). Massively parallel base editing to map variant effects in human hematopoiesis. *Cell*, **186**(11), 2456–2474.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, **349**(6255), 1483–1489.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS computational biology*, **9**(9), e1003214.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, **337**(6099), 1190–1195.

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J. A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature genetics*, **47**(12), 1393–1401.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, **39**(Suppl 7), S37–S42.

McCue, L. A., Thompson, W., Carmack, C. S., Ryan, M. P., Liu, J. S., Derbyshire, V., and Lawrence, C. E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic acids research*, **29**(3), 774–782.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, **20**(9), 1297–1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, **17**(1), 1–14.

McLeay, R. C. and Bailey, T. L. (2010). Motif enrichment analysis: a unified framework and an evaluation on chip data. *BMC bioinformatics*, **11**(1), 1–11.

Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K., and Bernstein, B. E. (2013). Locus-specific editing of histone modifications at endogenous enhancers. *Nature biotechnology*, **31**(12), 1133–1136.

Mercuri, E., Darras, B. T., Chiriboga, C. A., Day, J. W., Campbell, C., Connolly, A. M., Iannaccone, S. T., Kirschner, J., Kuntz, N. L., Saito, K., *et al.* (2018). Nusinersen versus sham control in later-onset spinal muscular atrophy. *New England Journal of Medicine*, **378**(7), 625–635.

Métais, J.-Y., Doerfler, P. A., Mayuranathan, T., Bauer, D. E., Fowler, S. C., Hsieh, M. M., Katta, V., Keriwala, S., Lazzarotto, C. R., Luk, K., *et al.* (2019). Genome editing of hbg1 and hbg2 to induce fetal hemoglobin. *Blood advances*, **3**(21), 3379–3392.

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., *et al.* (2017). Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. *Nature genetics*, **49**(12), 1779–1784.

Miller, S. M., Wang, T., Randolph, P. B., Arbab, M., Shen, M. W., Huang, T. P., Matuszek, Z., Newby, G. A., Rees, H. A., and Liu, D. R. (2020). Continuous evolution of spcas9 variants compatible with non-g pams. *Nature biotechnology*, **38**(4), 471–481.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome research*, **16**(9), 1182–1190.

Mitchell-Olds, T., Willis, J. H., and Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, **8**(11), 845–856.

Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., and Valen, E. (2014). Chopchop: a crispr/cas9 and talen web tool for genome editing. *Nucleic acids research*, **42**(W1), W401–W407.

Morgan, R. A. (2020). Ex vivo applications of cas12a.

Morris, J. A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D. A., Hao, S., *et al.* (2023). Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. *Science*, **380**(6646), eadh7699.

Morris, Q., Bulyk, M. L., and Hughes, T. R. (2011). Jury remains out on simple models of transcription factor specificity. *Nature biotechnology*, **29**(6), 483–484.

Moyerbrailean, G. A., Kalita, C. A., Harvey, C. T., Wen, X., Luca, F., and Pique-Regi, R. (2016). Which genetics variants in dnase-seq footprints are more likely to alter binding? *PLoS genetics*, **12**(2), e1005875.

Muggli, M. D., Bowe, A., Noyes, N. R., Morley, P. S., Belk, K. E., Raymond, R., Gagie, T., Puglisi, S. J., and Boucher, C. (2017). Succinct colored de bruijn graphs. *Bioinformatics*, **33**(20), 3181–3187.

Mundal, L. J., Igland, J., Veierød, M. B., Holven, K. B., Ose, L., Selmer, R. M., Wisloff, T., Kristiansen, I. S., Tell, G. S., Leren, T. P., *et al.* (2018). Impact of age on excess risk of coronary heart disease in patients with familial hypercholesterolaemia. *Heart*, **104**(19), 1600–1607.

Musunuru, K., Chadwick, A. C., Mizoguchi, T., Garcia, S. P., DeNizio, J. E., Reiss, C. W., Wang, K., Iyer, S., Dutta, C., Clendaniel, V., *et al.* (2021). In vivo crispr base editing of pcsk9 durably lowers cholesterol in primates. *Nature*, **593**(7859), 429–434.

Neuwald, A. F., Liu, J. S., and Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein science*, **4**(8), 1618–1632.

Newburger, D. E. and Bulyk, M. L. (2009). Uniprobe: an online database of protein binding microarray data on protein–dna interactions. *Nucleic acids research*, **37**(suppl_1), D77–D82.

Newby, G. A., Yen, J. S., Woodard, K. J., Mayuranathan, T., Lazzarotto, C. R., Li, Y., Sheppard-Tillman, H., Porter, S. N., Yao, Y., Mayberry, K., *et al.* (2021). Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature*, **595**(7866), 295–302.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., *et al.* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**(7261), 272–276.

Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K. Y., *et al.* (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science*, **353**(6305), aaf8729.

Noble, W. S. (2009). How does multiple testing correction work? *Nature biotechnology*, **27**(12), 1135–1137.

Nolis, I. K., McKay, D. J., Mantouvalou, E., Lomvardas, S., Merika, M., and Thanos, D. (2009). Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences*, **106**(48), 20222–20227.

Novak, A. M., Garrison, E., and Paten, B. (2017). A graph extension of the positional burrows–wheeler transform and its applications. *Algorithms for Molecular Biology*, **12**(1), 1–12.

Ogbourne, S. and Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, **331**(1), 1–14.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., and Sekiya, T. (1989). Detection of polymorphisms of human dna by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences*, **86**(8), 2766–2770.

O'Brien, A. and Bailey, T. L. (2014). Gt-scan: identifying unique genomic targets. *Bioinformatics*, **30**(18), 2673–2675.

Pablo, J. L. B., Cornett, S. L., Wang, L. A., Jo, S., Brünger, T., Budnik, N., Hegde, M., DeKeyser, J.-M., Thompson, C. H., Doench, J. G., *et al.* (2023). Scanning mutagenesis of the voltage-gated sodium channel nav1. 2 using base editing. *Cell Reports*, **42**(6).

Pang, B., van Weerd, J. H., Hamoen, F. L., and Snyder, M. P. (2023). Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology*, **24**(6), 383–395.

Park, J., Bae, S., and Kim, J.-S. (2015). Cas-designer: a web-based tool for choice of crispr-cas9 target sites. *Bioinformatics*, **31**(24), 4014–4016.

Park, J., Lim, K., Kim, J.-S., and Bae, S. (2017). Cas-analyzer: an online tool for assessing genome editing results using ngs data. *Bioinformatics*, **33**(2), 286–288.

Park, S., Koh, Y., Jeon, H., Kim, H., Yeo, Y., and Kang, J. (2020). Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Scientific reports*, **10**(1), 1–10.

Park, Y. and Kellis, M. (2015). Deep learning for regulatory genomics. *Nature biotechnology*, **33**(8), 825–826.

Paten, B., Zerbino, D. R., Hickey, G., and Haussler, D. (2014). A unifying model of genome evolution under parsimony. *BMC bioinformatics*, **15**, 1–31.

Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome research*, **27**(5), 665–676.

Paten, B., Eizenga, J. M., Rosen, Y. M., Novak, A. M., Garrison, E., and Hickey, G. (2018). Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, **25**(7), 649–663.

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., and Liu, D. R. (2013). High-throughput profiling of off-target dna cleavage reveals rna-programmed cas9 nuclease specificity. *Nature biotechnology*, **31**(9), 839–843.

Pavesi, G., Mauri, G., and Pesole, G. (2001). An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, **17**(suppl_1), S207–S214.

Pavesi, G., Mauri, G., and Pesole, G. (2004a). In silico representation and discovery of transcription factor binding sites. *Briefings in bioinformatics*, **5**(3), 217–236.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004b). Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, **32**(suppl_2), W199–W203.

Pedersen, B. S. and Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, **34**(5), 867–868.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics*, **14**(4), 288–295.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nature methods*, **6**(11), S22–S32.

Petrackova, A., Vasinek, M., Sedlarikova, L., Dyskova, T., Schneiderova, P., Novosad, T., Papajik, T., and Kriegova, E. (2019). Standardization of sequencing coverage depth in ngs: recommendation for detection of clonal and subclonal mutations in cancer diagnostics. *Frontiers in oncology*, **9**, 851.

Pickrell, J. K., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). False positive peaks in chip-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, **27**(15), 2144–2146.

Pillai, S. and Chellappan, S. P. (2015). Chip on chip and chip-seq assays: genome-wide analysis of transcription factor binding and histone modifications. In *Chromatin Protocols*, pages 447–472. Springer.

Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E., and Yuan, G.-C. (2016). Analyzing crispr genome-editing experiments with crispresso. *Nature biotechnology*, **34**(7), 695–697.

Pinello, L., Farouni, R., and Yuan, G.-C. (2018). Haystack: systematic analysis of the variation of epigenetic states and cell-type specific regulatory elements. *Bioinformatics*, **34**(11), 1930–1933.

Podolyan, Y., Gorb, L., and Leszczynski, J. (2003). Ab initio study of the prototropic tautomerism of cytosine and guanine and their contribution to spontaneous point mutations.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., *et al.* (2018). A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, **36**(10), 983–987.

Porto, E. M., Komor, A. C., Slaymaker, I. M., and Yeo, G. W. (2020). Base editing: advances and therapeutic opportunities. *Nature Reviews Drug Discovery*, **19**(12), 839–859.

Pratt, H. E., Andrews, G. R., Phalke, N., Huey, J. D., Purcaro, M. J., van der Velde, A., Moore, J. E., and Weng, Z. (2022). Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Research*, **50**(D1), D141–D149.

Puglisi, S. J., Smyth, W. F., and Turpin, A. H. (2007). A taxonomy of suffix array construction algorithms. *acm Computing Surveys (CSUR)*, **39**(2), 4–es.

Puig, M., Casillas, S., Villatoro, S., and Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in functional genomics*, **14**(5), 369–379.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *et al.* (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, **81**(3), 559–575.

Quang, D. and Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, **44**(11), e107–e107.

Quang, D. and Xie, X. (2019). Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, **166**, 40–47.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Raal, F. J., Kallend, D., Ray, K. K., Turner, T., Koenig, W., Wright, R. S., Wijngaard, P. L., Curcio, D., Jaros, M. J., Leiter, L. A., *et al.* (2020). Inclisiran for the treatment of heterozygous familial hypercholesterolemia. *New England Journal of Medicine*, **382**(16), 1520–1530.

Ran, F., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., and Zhang, F. (2013). Genome engineering using the crispr-cas9 system. *Nature protocols*, **8**(11), 2281–2308.

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine*, **6**(1), 1–17.

Reid, J. E. and Wernisch, L. (2011). Steme: efficient em to find motifs in large data sets. *Nucleic acids research*, **39**(18), e126–e126.

Reimold, A. M., Iwakoshi, N. N., Manis, J., Vallabhajosyula, P., Szomolanyi-Tsuda, E., Gravallese, E. M., Friend, D., Grusby, M. J., Alt, F., and Glimcher, L. H. (2001). Plasma cell differentiation requires the transcription factor xbp-1. *Nature*, **412**(6844), 300–307.

Reshef, Y. A., Finucane, H. K., Kelley, D. R., Gusev, A., Kotliar, D., Ulirsch, J. C., Hormozdiari, F., Nasser, J., O'Connor, L., Van De Geijn, B., *et al.* (2018). Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nature genetics*, **50**(10), 1483–1493.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, **147**(6), 1408–1419.

Richter, M. F., Zhao, K. T., Eton, E., Lapinaite, A., Newby, G. A., Thuronyi, B. W., Wilson, C., Koblan, L. W., Zeng, J., Bauer, D. E., *et al.* (2020). Phage-assisted evolution of an adenine base editor with improved cas domain compatibility and activity. *Nature biotechnology*, **38**(7), 883–891.

Rodgers, K. and McVey, M. (2016). Error-prone repair of dna double-strand breaks. *Journal of cellular physiology*, **231**(1), 15–24.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-dna recognition. *Annual review of biochemistry*, **79**, 233.

Ryu, J., Barkal, S., Yu, T., Jankowiak, M., Zhou, Y., Francoeur, M., Phan, Q. V., Li, Z., Tognon, M., Brown, L., *et al.* (2024). Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification. *Nature Genetics*, pages 1–13.

Saare, M., Tserel, L., Haljasmägi, L., Taalberg, E., Peet, N., Eimre, M., Vetik, R., Kingo, K., Saks, K., Tamm, R., *et al.* (2020). Monocytes present age-related changes in phospholipid concentration and decreased energy metabolism. *Aging Cell*, **19**(4), e13127.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y., and Ramabhadran, B. (2013). Improvements to deep convolutional neural networks for lvcsr. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 315–320. IEEE.

Sakata, R. C., Ishiguro, S., Mori, H., Tanaka, M., Tatsuno, K., Ueda, H., Yamamoto, S., Seki, M., Masuyama, N., Nishida, K., *et al.* (2020). Base editors for simultaneous introduction of c-to-t and a-to-g mutations. *Nature biotechnology*, **38**(7), 865–869.

Sánchez-Rivera, F. J., Diaz, B. J., Kastenhuber, E. R., Schmidt, H., Katti, A., Kennedy, M., Tem, V., Ho, Y.-J., Leibold, J., Paffenholz, S. V., *et al.* (2022). Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants. *Nature biotechnology*, **40**(6), 862–873.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, **32**(suppl_1), D91–D94.

Sangree, A. K., Griffith, A. L., Szegletes, Z. M., Roy, P., DeWeirdt, P. C., Hegde, M., McGee, A. V., Hanna, R. E., and Doench, J. G. (2022). Benchmarking of spcas9 variants enables deeper base editor screens of brca1 and bcl2. *Nature Communications*, **13**(1), 1318.

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., *et al.* (2018). Optimized libraries for crispr-cas9 genetic screens with multiple modalities. *Nature communications*, **9**(1), 5416.

Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E., and Willer, C. J. (2015). Gregor: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, **31**(16), 2601–2606.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, **18**(20), 6097–6100.

Schoenfelder, S. and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, **20**(8), 437–455.

Scott, D. A. and Zhang, F. (2017). Implications of human genetic variation in crispr-based therapeutic genome editing. *Nature medicine*, **23**(9), 1095–1101.

Seol, J.-H., Shim, E. Y., and Lee, S. E. (2018). Microhomology-mediated end joining: Good, bad and ugly. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **809**, 81–87.

Sgro, A. and Blancafort, P. (2020). Epigenome engineering: new technologies for precision medicine. *Nucleic acids research*, **48**(22), 12453–12482.

Shah, S. A., Erdmann, S., Mojica, F. J., and Garrett, R. A. (2013). Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, **10**(5), 891–899.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., *et al.* (2014). Genome-scale crispr-cas9 knockout screening in human cells. *Science*, **343**(6166), 84–87.

Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D., and Smit, S. (2016). Pantools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**(17), i487–i493.

Sherry, S. T., Ward, M., and Sirotkin, K. (1999). dbsnp—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research*, **9**(8), 677–679.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, **29**(1), 308–311.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Shrikumar, A., Prakash, E., and Kundaje, A. (2019). Gkmexplain: fast and accurate interpretation of nonlinear gapped k-mer svms. *Bioinformatics*, **35**(14), i173–i182.

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Sirén, J., Chang, X., Garrison, E., and Paten, B. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nature Methods*, **20**(2), 239–247.

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS one*, **5**(3), e9722.

Siebert, M. and Söding, J. (2016). Bayesian markov models consistently outperform pwms at predicting motifs in nucleotide sequences. *Nucleic acids research*, **44**(13), 6055–6069.

Siggers, T. and Gordân, R. (2014). Protein–dna binding: complexities and multi-protein codes. *Nucleic acids research*, **42**(4), 2099–2111.

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**(17), i639–i648.

Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, **7**(2), 122–137.

Sinha, S., Li, F., Villarreal, D., Shim, J. H., Yoon, S., Myung, K., Shim, E. Y., and Lee, S. E. (2017). Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. *PLoS genetics*, **13**(4), e1006714.

Sirén, J., Garrison, E., Novak, A. M., Paten, B., and Durbin, R. (2020). Haplotype-aware graph indexes. *Bioinformatics*, **36**(2), 400–407.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., *et al.* (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**(6574), abg8871.

Siva, N. (2008). 1000 genomes project. *Nature biotechnology*, **26**(3), 256–257.

Slattery, M., Zhou, T., Yang, L., Machado, A. C. D., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, **39**(9), 381–399.

Smith, C., Gore, A., Yan, W., Abalde-Atristain, L., Li, Z., He, C., Wang, Y., Brodsky, R. A., Zhang, K., Cheng, L., *et al.* (2014). Whole-genome sequencing analysis reveals high specificity of crispr/cas9 and talen-based genome editing in human ipscs. *Cell stem cell*, **15**(1), 12–13.

Song, F. and Stieger, K. (2017). Optimizing the dna donor template for homology-directed repair of double-strand breaks. *Molecular Therapy-Nucleic Acids*, **7**, 53–60.

Spady, D. K. (1992). Hepatic clearance of plasma low density lipoproteins. In *Seminars in liver disease*, volume 12, pages 373–385.

Spahn, P. N., Bath, T., Weiss, R. J., Kim, J., Esko, J. D., Lewis, N. E., and Harismendy, O. (2017). Pinapl-py: A comprehensive web-application for the analysis of crispr/cas9 screens. *Scientific reports*, **7**(1), 15854.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. *Nature*, **435**(7042), 637–645.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, **13**(9), 613–626.

Stadtmauer, E. A., Fraietta, J. A., Davis, M. M., Cohen, A. D., Weber, K. L., Lancaster, E., Mangan, P. A., Kulikovskaya, I., Gupta, M., Chen, F., *et al.* (2020). Crispr-engineered t cells in patients with refractory cancer. *Science*, **367**(6481), eaba7365.

Stewart, A. J., Hannenhalli, S., and Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, **192**(3), 973–985.

Stormo, G. D. (1998). Information content and free energy in dna-protein interactions [1]. *Journal of theoretical biology*, **195**(1), 135–137.

Stormo, G. D. (2000). Dna binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Stormo, G. D. (2013). Modeling the specificity of protein-dna interactions. *Quantitative biology*, **1**(2), 115–130.

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein–dna interactions. *Nature Reviews Genetics*, **11**(11), 751–760.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

Sun, Y., Liu, F., Fan, C., Wang, Y., Song, L., Fang, Z., Han, R., Wang, Z., Wang, X., Yang, Z., *et al.* (2021). Characterizing sensitivity and coverage of clinical wgs as a diagnostic test for genetic disorders. *BMC medical genomics*, **14**, 1–13.

Syding, L. A., Nickl, P., Kasparek, P., and Sedlacek, R. (2020). Crispr/cas9 epigenome editing potential for rare imprinting diseases: a review. *Cells*, **9**(4), 993.

Talukder, A., Barham, C., Li, X., and Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, **22**(3), bbaa177.

Tarkoma, S., Rothenberg, C. E., and Lagerspetz, E. (2011). Theory and practice of bloom filters for distributed systems. *IEEE Communications Surveys & Tutorials*, **14**(1), 131–155.

Teotónio, H., Chelo, I. M., Bradić, M., Rose, M. R., and Long, A. D. (2009). Experimental evolution reveals natural selection on standing genetic variation. *Nature genetics*, **41**(2), 251–257.

Thiffault, I., Farrow, E., Zellmer, L., Berrios, C., Miller, N., Gibson, M., Caylor, R., Jenkins, J., Faller, D., Soden, S., *et al.* (2019). Clinical genome sequencing in an unbiased pediatric cohort. *Genetics in Medicine*, **21**(2), 303–310.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, **17**(12), 1113–1122.

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that define the best chip-seq peak calling algorithms. *Briefings in bioinformatics*, **18**(3), 441–450.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'keeffe, S., Masri, N. E., Roider, H. G., Manke, T., and Vingron, M. (2011). Transcription factor binding predictions using trap for the analysis of chip-seq data and regulatory snps. *Nature protocols*, **6**(12), 1860–1869.

Thompson, M. R., Xu, D., and Williams, B. R. (2009). Atf3 transcription factor and its emerging roles in immunity and cancer. *Journal of molecular medicine*, **87**, 1053–1060.

Tognon, M., Bonnici, V., Garrison, E., Giugno, R., and Pinello, L. (2021). Grafimo: variant and haplotype aware motif scanning on pangenome graphs. *PLoS computational biology*, **17**(9), e1009444.

Tognon, M., Giugno, R., and Pinello, L. (2023). A survey on algorithms to characterize transcription factor binding sites. *Briefings in Bioinformatics*, page bbad156.

Tomovic, A. and Oakeley, E. J. (2007). Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**(8), 933–941.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., *et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, **23**(1), 137–144.

Trabelsi, A., Chaabane, M., and Ben-Hur, A. (2019). Comprehensive evaluation of deep learning architectures for prediction of dna/rna sequence binding specificities. *Bioinformatics*, **35**(14), i269–i277.

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., *et al.* (2015). Guide-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. *Nature biotechnology*, **33**(2), 187–197.

Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., and Joung, J. K. (2017). Circle-seq: a highly sensitive in vitro screen for genome-wide crispr–cas9 nuclease off-targets. *Nature methods*, **14**(6), 607–614.

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., *et al.* (2017). Defining a cancer dependency map. *Cell*, **170**(3), 564–576.

Tycko, J., Myer, V. E., and Hsu, P. D. (2016). Methods for optimizing crispr-cas9 genome editing specificity. *Molecular cell*, **63**(3), 355–370.

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, **1**(1), 59.

Valenzuela, L. and Kamakaka, R. T. (2006). Chromatin insulators. *Annu. Rev. Genet.*, **40**, 107–138.

Väli, Ü., Brandström, M., Johansson, M., and Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC genetics*, **9**(1), 1–8.

Veres, A., Gosis, B. S., Ding, Q., Collins, R., Ragavendran, A., Brand, H., Erdin, S., Cowan, C. A., Talkowski, M. E., and Musunuru, K. (2014). Low incidence of off-target mutations in individual crispr-cas9 and talen targeted human stem cell clones detected by whole-genome sequencing. *Cell stem cell*, **15**(1), 27–30.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, **55**(4), 641–658.

Vorontsov, I. E., Kulakovskiy, I. V., and Makeev, V. J. (2013). Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for Molecular Biology*, **8**(1), 1–11.

Vu, H., Cheng, E., Wilkinson, R., and Lech, M. (2017). On the use of convolutional neural networks for graphical model-based human pose estimation. In *2017 International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, pages 88–93. IEEE.

Walton, R. T., Christie, K. A., Whittaker, M. N., and Kleinstiver, B. P. (2020). Unconstrained genome targeting with near-pamless engineered crispr-cas9 variants. *Science*, **368**(6488), 290–296.

Wang, B., Wang, M., Zhang, W., Xiao, T., Chen, C.-H., Wu, A., Wu, F., Traugh, N., Wang, X., Li, Z., *et al.* (2019). Integrative analysis of pooled crispr genetic screens using mageckflute. *Nature protocols*, **14**(3), 756–780.

Wang, R., Lin, D.-Y., and Jiang, Y. (2022a). Epic: Inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell rna sequencing. *PLoS genetics*, **18**(6), e1010251.

Wang, S.-W., Gao, C., Zheng, Y.-M., Yi, L., Lu, J.-C., Huang, X.-Y., Cai, J.-B., Zhang, P.-F., Cui, Y.-H., and Ke, A.-W. (2022b). Current applications and future perspective of crispr/cas9 gene editing in cancer. *Molecular cancer*, **21**(1), 1–27.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, **350**(6264), 1096–1101.

Wang, T., Yu, H., Hughes, N. W., Liu, B., Kendirli, A., Klein, K., Chen, W. W., Lander, E. S., and Sabatini, D. M. (2017). Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras. *Cell*, **168**(5), 890–903.

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J., *et al.* (2022c). The human pangenome project: a global resource to map genomic diversity. *Nature*, **604**(7906), 437–446.

Weiner, P. (1973). Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, pages 1–11. IEEE.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, **46**(11), 1160–1165.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., *et al.* (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, **31**(2), 126–134.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., *et al.* (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**(6), 1431–1443.

Weterings, E. and Chen, D. J. (2008). The endless tale of non-homologous end-joining. *Cell research*, **18**(1), 114–124.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome biology*, **13**(9), 1–16.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**(2), 307–319.

Wienert, B., Funnell, A. P., Norton, L. J., Pearson, R. C., Wilkinson-White, L. E., Lester, K., Vadolas, J., Porteus, M. H., Matthews, J. M., Quinlan, K. G., *et al.* (2015). Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nature communications*, **6**(1), 7085.

Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, **24**(1), 238–241.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. (2000). Transfac: an integrated system for gene expression regulation. *Nucleic acids research*, **28**(1), 316–319.

Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, **13**(1), 59–69.

Workman, C. T. and Stormo, G. D. (1999). Ann-spec: a method for discovering transcription factor binding sites with improved specificity. In *Biocomputing 2000*, pages 467–478. World Scientific.

Worsley Hunt, R. and Wasserman, W. W. (2014). Non-targeted transcription factors motifs are a systemic component of chip-seq datasets. *Genome biology*, **15**(7), 1–16.

Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., Serpe, J. M., Dasu, T., Tschannen, M. R., Veith, R. L., *et al.* (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, **13**(3), 255–262.

Wu, Y., Zeng, J., Roscoe, B. P., Liu, P., Yao, Q., Lazzarotto, C. R., Clement, K., Cole, M. A., Luk, K., Baricordi, C., *et al.* (2019). Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nature medicine*, **25**(5), 776–783.

Xiao, A., Cheng, Z., Kong, L., Zhu, Z., Lin, S., Gao, G., and Zhang, B. (2014). Casot: a genome-wide cas9/grna off-target searching tool. *Bioinformatics*, **30**(8), 1180–1182.

Xu, F., Park, M.-R., Kitazumi, A., Herath, V., Mohanty, B., Yun, S. J., and de los Reyes, B. G. (2012). Cis-regulatory signatures of orthologous stress-associated bzip transcription factors from rice, sorghum and arabidopsis based on phylogenetic footprints. *BMC genomics*, **13**(1), 1–15.

Xu, L., Yang, H., Gao, Y., Chen, Z., Xie, L., Liu, Y., Liu, Y., Wang, X., Li, H., Lai, W., *et al.* (2017). Crispr/cas9-mediated ccr5 ablation in human hematopoietic stem/progenitor cells confers hiv-1 resistance in vivo. *Molecular Therapy*, **25**(8), 1782–1789.

Xu, L., Wang, J., Liu, Y., Xie, L., Su, B., Mou, D., Wang, L., Liu, T., Wang, X., Zhang, B., *et al.* (2019a). Crispr-edited stem cells in a patient with hiv and acute lymphocytic leukemia. *New England Journal of Medicine*, **381**(13), 1240–1247.

Xu, S., Luk, K., Yao, Q., Shen, A. H., Zeng, J., Wu, Y., Luo, H.-Y., Brendel, C., Pinello, L., Chui, D. H., *et al.* (2019b). Editing aberrant splice sites efficiently restores β-globin expression in β-thalassemia. *Blood, The Journal of the American Society of Hematology*, **133**(21), 2255–2262.

Yang, L., Zhu, Y., Yu, H., Cheng, X., Chen, S., Chu, Y., Huang, H., Zhang, J., and Li, W. (2020). scmageck links genotypes with multiple phenotypes in single-cell crispr screens. *Genome biology*, **21**(1), 1–14.

Yao, Q., Ferragina, P., Reshef, Y., Lettre, G., Bauer, D. E., and Pinello, L. (2021). Motif-raptor: a cell type-specific and transcription factor centric approach for post-gwas prioritization of causal regulators. *Bioinformatics*, **37**(15), 2103–2111.

Yin, J., Lu, R., Xin, C., Wang, Y., Ling, X., Li, D., Zhang, W., Liu, M., Xie, W., Kong, L., *et al.* (2022). Cas9 exo-endonuclease eliminates chromosomal translocations during genome editing. *Nature communications*, **13**(1), 1204.

Yin, Q., Wu, M., Liu, Q., Lv, H., and Jiang, R. (2019). Deephistone: a deep learning approach to predicting histone modifications. *BMC genomics*, **20**(2), 11–23.

Yu, J., Silva, J., and Califano, A. (2016). Screenbeam: a novel meta-analysis algorithm for functional genomics screens via bayesian hierarchical modeling. *Bioinformatics*, **32**(2), 260–267.

Yu, J., Liu, M., Liu, H., and Zhou, L. (2019). Gata1 promotes colorectal cancer cell proliferation, migration and invasion via activating akt signaling pathway. *Molecular and cellular biochemistry*, **457**(1-2), 191–199.

Yu, Y. and Chen, H. (2023). Human pangenome: far-reaching implications in precision medicine. *Frontiers of Medicine*, pages 1–7.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic acids research*, **37**(suppl_2), W247–W252.

Zambelli, F., Pesole, G., and Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, **14**(2), 225–237.

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature reviews genetics*, **16**(3), 172–183.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016a). Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, **32**(12), i121–i127.

Zeng, H., Hashimoto, T., Kang, D. D., and Gifford, D. K. (2016b). Gerv: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, **32**(4), 490–496.

Zeng, J., Wu, Y., Ren, C., Bonanno, J., Shen, A. H., Shea, D., Gehrke, J. M., Clement, K., Luk, K., Yao, Q., *et al.* (2020a). Therapeutic base editing of human hematopoietic stem cells. *Nature Medicine*, **26**(4), 535–541.

Zeng, W., Wu, M., and Jiang, R. (2018). Prediction of enhancer-promoter interactions via natural language processing. *BMC genomics*, **19**(2), 13–22.

Zeng, W., Wang, Y., and Jiang, R. (2020b). Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics*, **36**(2), 496–503.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., Van Der Oost, J., Regev, A., *et al.* (2015). Cpf1 is a single rna-guided endonuclease of a class 2 crispr-cas system. *Cell*, **163**(3), 759–771.

Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics*, **24**(R1), R102–R110.

Zhang, X., Zhu, B., Chen, L., Xie, L., Yu, W., Wang, Y., Li, L., Yin, S., Yang, L., Hu, H., *et al.* (2020). Dual base editor catalyzes both cytosine and adenine base conversions in human cells. *Nature biotechnology*, **38**(7), 856–860.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of chip-seq (macs). *Genome biology*, **9**(9), 1–9.

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., and Peng, S. (2019). Deep learning in omics: a survey and guideline. *Briefings in functional genomics*, **18**(1), 41–57.

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). Tsgene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic acids research*, **44**(D1), D1023–D1031.

Zhao, Y. and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology*, **29**(6), 480–483.

Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C. A., *et al.* (2019). Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic acids research*, **47**(D1), D729–D735.

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., and Consortium, . G. P. (2017). Alignment of 1000 genomes project reads to reference assembly grch38. *Gigascience*, **6**(7), gix038.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, **12**(10), 931–934.

Zhu, X., Xu, Y., Yu, S., Lu, L., Ding, M., Cheng, J., Song, G., Gao, X., Yao, L., Fan, D., *et al.* (2014). An efficient genotyping method for genome-modified animals and human cells generated with crispr/cas9 system. *Scientific reports*, **4**(1), 6420.

Zia, A. and Moses, A. M. (2012). Towards a theoretical understanding of false positives in dna motif finding. *BMC bioinformatics*, **13**(1), 1–9.

Zuo, C., Shin, S., and Keleş, S. (2015). atsnp: transcription factor binding affinity testing for regulatory snp detection. *Bioinformatics*, **31**(20), 3353–3355.