Università degli Studi di Verona

DIPARTIMENTO DI INFORMATICA
Ph.D. in Computer Science

Third Year Report

Predicting genetic variants impact on genomic regulatory elements and CRISPR genome editing

Student:
Manuel Tognon
Matricola VR456869

Supervisor:
Prof. Rosalba Giugno

Cosupervisor:
Prof. Luca Pinello

Anno Accademico 2022–2023

Contents

List of Figures

List of Tables

Introduction

Omics sciences have swiftly emerged as fundamental tools for informing medical decisions. They serve as the foundational pillars supporting precision or personalized medicine (Ginsburg and Willard, 2009). The evolution and enhancements in sequencing technologies have brought about a profound transformation, substantially improving both the quantity and quality of available omics data. This technological progress has had a profound impact, drastically reducing the costs associated with analyses involving omics data. Furthermore it has empowered the accumulation of large dataset, even for individual patients (Voelkerding *et al.*, 2009). Individual-specific omics data serve 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. Among these biomarkers, genetic variants hold a pivotal position, operating at the level of DNA sequences. Genetic variants can be located within genes (coding regions), or in other genomic regions (non-coding regions). in the complex landscape of cellular regulation, non-coding genetic variants frequently play important roles in the epigenetic machinery that governs the cellular environment (Maurano *et al.*, 2012). While interpreting the functional impact of genetic variants mapped within coding regions on the cellular environment is often straightforward and widely explored in the existing literature, non-coding variants functional interpretation presents challenges. We refer to the genomic regions governing the cellular environment as genomic regulatory elements (GREs). Currently, several computational tools designed to predict the effects of genetic variants within gene sequences are available to the scientific community. However, while these tools are valuable in assessing genetic variants impacts at broader level, they often fall short when it comes to accomodating the requirements of precision genomics. In the evolving landscape of precision medicine, ther is a growing need for computational resources that can seamlessly integrate individual-oriented aspects into their analyses. Moreover, since genomics plays an increasingly central role in healthcare decision-making, the development of more individual centric computational tools becomes fundamental. We addressed these challenges by proposing innovative algorithms and computational tools tailored for the analysis of genetic variant impact on GREs and CRISPR genome editing, at broad and individual-specific levels. Our primary focus has been on creating haplotype- and individual-aware methods, aligning with the growing demand for individual-centric genomics applications. Our research focused on developing novel algorithms designed to discover and find transcription factor binding sites (TFBSs) within DNA sequences, accounting for individual- and cell type-specific genetic variants. We propose a novel computational model to represent TFBSs called MotifGraphs, which employs graph data structures to efficiently represent genetic diversity between binding sites from different individuals or cell types, without sacrificing interpretability. To address the effects of genetic variants on TFBSs, we developed two novel algorithms GRAFIMO (Tognon *et al.*, 2021) and MotifRaptor (Yao *et al.*, 2021). GRAFIMO searches for occurrences of known TFBSs genome graphs (Paten *et al.*, 2017), while accounting for for individual haplotypes and genetic variants. MotifRaptor predicts and annotates genetic variants impact on TFBSs integrating different omics data, such as chromatin accessibility, gene expression and GWAS summary statistics. Additionaly we introduced CRISPRme (Cancellieri *et al.*, 2023), a tool designed to CRISPR off-targets and avaluate the potential impact of individual genetic variants on target specificity. Importantly, CRISPRme considers single-nucleotide variants, accomadate bona fide haplotypes and handles spacer:protospacer mismatches and bulges. This tool is specifically designed to perform individuals-oriented analyses, considering the wide genetic diversity present in different populations.

Transcription Factors

Transcription factors (TFs) (**Fig.2.1**) are fundamental regulatory proteins playing a key role in regulating the transcriptional state, cellular differentiation and developmental state of cells (Lambert *et al.*, 2018; Reimold *et al.*, 2001; Whyte *et al.*, 2013). In human, approximately 1600 proteins are recognized as TFs (Babu *et al.*, 2004). This number accounts for roughly 8% of all human genes, highlighting the critical role played by TFs orchestrating genetic regulation. TFs exhibit their regulatory prowess by often collaborating in a coordinated manner to influence gene expression. This collaborative orchestration is vital for fine-tuning and precisely controlling cellular process. Moreover, TFs display a remarkable versatility as they govern the activity of multiple genes across different cell types (Lambert *et al.*, 2018). TFs exhibit a modular structure, that is divided into three distinct domains (Latchman, 1997). (i) The DNA binding domain directs the TF to its precise target site on the genome. Through a specific recognition of DNA sequences, the DNA binding domain enables the TF to dock onto regulatory regions located across the genome. (ii) the activation domain facilitates interactions between the TF and other transcriptional regulators. By engaging with different co-factors and regulatory proteins, the activation domain plays a crucial role modulating gene expression, often acting as a bridge between the factor and the transcriptional machinery. (iii) The signal sensing domain captures external signals and transmits them to the broader transcriptional complex. These signals can originate from different sources, including cellular cues and environmental stimuli, and are essential for fine-tuning the TF regulatory actions in response to changing conditions. The interplay between these three domains allows TFs to function as highly versatile and adaptable components of the gene regulation machinery, responding to both internal and external cues to precisely control gene expression in a dynamic and context-dependent manner. TFs exert their function through different strategies. (i) TFs can either facilitate the recruitment of RNA polymerase to gene promoter regions, thus promoting transcription initiation, or block RNA polymerase access (Fuda *et al.*, 2009). (ii) TFs play a crucial role in shaping chromatin landscape by weakening DNA-histone interactions, increasing DNA accessibility and consequently facilitating gene expression. (iii) Some TFs catalyze histone deacetylation (Liu *et al.*, 2016), by removing acetyl groups from histones, thus promoting a more compact chromatin structure and consequently reducing gene transcription. (iv) Other TFs enhance DNA-histone interactions, leading to a more tightly packed chromatin structure and consequently repressing gene expression. Therefore, by binding short DNA sequences (∼6-20 nucleotides (Stewart *et al.*, 2012)), known as transcription factor binding sites (TFBSs), they finely regulate gene expression in a cell-specific manner. TFBSs are located within gene promoters (Whitfield *et al.*, 2012) or in more distant regulatory elements such as enhancers, silencers, or insulators (Gotea *et al.*, 2010; Lemon and Tjian, 2000; Nolis *et al.*, 2009). While TFBS often exhibit recurring sequence patterns, referred to as *motifs*, TFs display a remarkable ability to bind to similar but not identical sequences, often differing by just a few nucleotides. The precise configuration of TFBS, coupled with the local chromatin structure, plays a pivotal role in fine-tuning TFs' regulatory functions within cells (Mendenhall *et al.*, 2013; Maurano *et al.*, 2015). During the process of DNA binding, Transcription Factors harness a combination of electrostatic and Van der Waals forces. Although TFs exhibit high specificity in binding to their target sequences, not every nucleotide within the binding site directly interacts with the TF. These interactions vary in strength, resulting in TFs binding not to a single specific sequence but to a closely related subset of targets. However, the sequence composition of the TFBS decisively dictates the strength of the TF-DNA interaction, known as binding affinity. Numerous studies have established links between different diseases and cancer types and genetic variants occurring within TFBS (Docquier *et al.*, 2005; Katainen *et al.*, 2015; Yu *et al.*, 2019). Furthermore, variants within TFBS can disrupt the precise regulation of gene expression by TFs, potentially affecting the entire cellular environment and even propagating effects to neighboring cells. Moreover, the misregulation of gene expression governed by TFs caused by variants occurring in TFBS could affect the entire cell environment and be propagated to neighboring cells. Therefore, identifying such regulatory motifs would provide fundamental insights

Figure 2.1. A human transcription factor (CTCF) binding its DNA target sequence.

on the complex mechanisms governing gene expression and the cell environment.

2.1 Discovering Transcription Factor Binding Site motifs

Several experimental assays have been developed to determine the binding site sequences of TFs in living cells or organisms (*in vivo*), or in test-tubes using synthetic or purified components (*in vitro*) (Jolma and Taipale, 2011). Early methods, like electrophoretic mobility shift assay (EMSA) (Garner and Revzin, 1981) or footprinting (Hampshire *et al.*, 2007), generally analyze a relatively small number of target sequences to find TFBS. As a result, they return small datasets of bound sequences. *In vitro* and *in vivo* high-throughput protocols such as PBM, SELEX or ChIP methods (Berger *et al.*, 2006; Jolma *et al.*, 2010; Collas and Dahl, 2008), facilitated the analysis of most target sites for factors of interest. As a result, large datasets of bound sequences have been generated, presenting an unprecedented opportunity to study and determine the TF binding landscapes. Experimental assays can recover the sequences bound by TFs along with their relative or absolute binding affinity. However, such datasets can incorrectly report unbound sequences as binding sites. In addition, the assays usually capture extra nucleotides in target sites, reducing data resolution and making manual analysis challenging. Motif discovery algorithms provide a computational framework to analyze these large datasets generated by experimental assays, discovering the 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). Given a sequence dataset, these algorithms typically recover sets of short and similar sequence elements. The prioritized sequence elements are later used to construct a motif model, summarizing the diverse binding site configurations observed among the prioritized sequences, and encoding their recurrent patterns and similarities (**Fig.2.2 (A)**). Several methods and models have been proposed to discover and represent TFBS motifs. Position weight matrices (PWMs) (Stormo, 2000) are the most popular models. PWMs are simple yet powerful and interpretable models, encoding the probability of observing a given nucleotide in each TFBS position. However, PWMs have some limitations, like the assumption of independence among

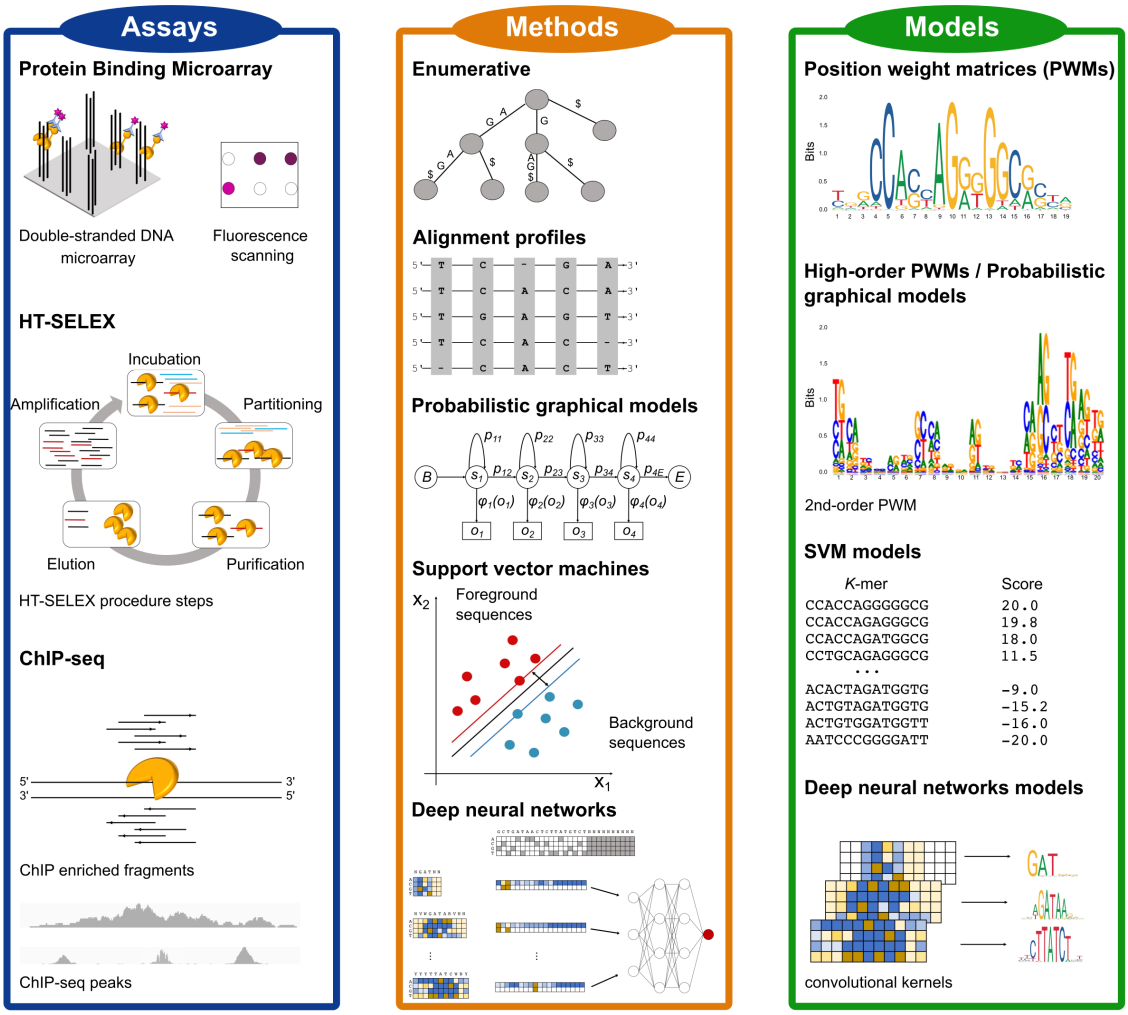


Figure 2.2. Experimental and computational methods to discover TFBS and popular models to represent binding site motifs. Protein binding microarray (PBM), HT-SELEX and ChIP-seq have become the most popular assays to determine TF binding preferences and identify their target sites (TFBS) in recent years. 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.

the binding site positions. Therefore, several alternative motif models have been proposed (Siddharthan, 2010; Gorkin *et al.*, 2012; He *et al.*, 2021) , as described below. The derived motif models can be employed in many downstream analyses, like searching potential binding site occurrences in regulatory genomic sequences, predicting the sets of genes regulated by the investigated TFs or assessing how genetic variants could affect their binding landscape.

2.1.1 Experimental methods to discover Transcription Factor Binding Sites

During the last decades, several techniques have been introduced to experimentally identify and assess TF binding sites and binding preferences (Jolma and Taipale, 2011) (**Fig.2.2** and **Table 2.1**). Early studies on TF binding focused their analysis on gene promoters (Stormo, 2000) and employed in vitro methods, such as Electro-Mobility Shift Assay (EMSA) (Garner and Revzin, 1981) or DNase footprinting (Galas and Schmitz, 1978). EMSA exploits non-denatured polyacrylamide gel properties to separate bound and unbound DNA sequences. DNase footprinting combines EMSA with DNase I cleavage, identifying uncut regions (footprints) due to the protection of the bound TF. Generally, these assays produce datasets of a few hundred of bound sequences, exploring a limited spectrum of TFs binding landscape. Moreover, EMSA and DNase footprinting may be subject to technical constraints that could lead to inaccuracies in the reported sequences and binding preferences (Jolma and Taipale, 2011). The introduction of NGS technologies revolutionized the study of TFBS identification by encouraging researchers to develop methods that exploit the power of massively parallel sequencing (**Fig.2.2**). These methods have two major advan-

tages: (i) they do not require any prior knowledge on the binding site sequence (Jolma and Taipale, 2011; Zia and Moses, 2012) and (ii) produce datasets of thousands of bound sequences allowing a better 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 bp) and measure TF binding preferences in vitro. In PBMs, a tagged TF is released on a glass slide containing thousands of spots filled with short, immobilized DNA sequences. The tagged TFs are then incubated with fluorescent antibodies against the tag and subsequently washed to remove weakly bound factors. The fluorescence and DNA sequence enrichment are then used to quantify the TF–DNA binding strength and capture the bound sequences. Generally, the recovered sequences do not contain nucleotides flanking the investigated binding sites, producing high-resolution datasets. However, since the number of possible sequences grows as a function of the target length, PBMs can assess only a limited number of target sequences (Jolma and Taipale, 2011; Zia and Moses, 2012). PBM analysis is usually constrained to binding sites ∼10–12 bp long. HT-SELEX (Jolma and Taipale, 2011; Jolma *et al.*, 2010) is a widely used *in vitro* method, coupling SELEX with high-throughput sequencing. A TF is released on a pool of randomized DNA sequences to allow the factor to select its target sites. The resulting TF–DNA complexes are separated from unbound sequences using affinity capture, and subsequently amplified through polymerase chain reaction (PCR) and sequenced. The resulting DNA library is enriched in binding sites for the studied TF and is used as the starting pool for another SELEX run (Jolma and Taipale, 2011; Jolma *et al.*, 2010). SELEX does not require any prior knowledge on the target sites of the investigated factor (Jolma *et al.*, 2013). Since SELEX reaction is typically performed in liquid phase and consequently does not suffer from physical constraints, the sequence space covered by HT-SELEX is often larger than that of PBMs. Moreover, by coupling sequencing with DNA barcode indexing, HT-SELEX allows to analyze hundreds of TFs in parallel. HT-SELEX produces datasets of thousands of high-resolution bound sequences, which include only a few nucleotides flanking the binding sites. However, since the starting DNA library is constituted by randomized sequences, HT-SELEX cannot recover the genomic binding locations for the investigated factor. The introduction of chromatin immunoprecipitation (ChIP) technologies (Collas and Dahl, 2008). radically changed the study of TFBS binding, enabling the genome-wide identification of regions bound by TFs in vivo. In ChIP, the TF–DNA complexes are cross-linked using formaldehyde. The DNA is then fragmented in ∼100–1000 bp long fragments and subsequently immunoprecipitated with antibodies specific for the investigated TF. To recover the bound sequences, the cross-links are reverted. Then, the resulting fragments are amplified through 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 the binding regions, the recovered DNA fragments are mapped onto the genome. After ChIP-seq reads mapping, peak calling algorithms (Thomas *et al.*, 2017; Guo *et al.*, 2012; Zhang *et al.*, 2008) are employed to predict the genomic binding locations for the investigated factor. Peak calling algorithms identify the genomic regions showing greater enrichment in mapped DNA probes with respect to a control experiment and mark those regions as binding locations, or peaks (Pepke *et al.*, 2009). ChIP methods produce large datasets of thousands of genomic regions, whose length ranges from few hundreds to thousands of nucleotides, from which we can identify the likely TFBS for the investigated factor. Although ChIP technologies, and particularly ChIP-seq, are currently considered the current 'golden standard', they have some limitations. (i) ChIP can detect indirect binding, identifying other TFBS not belonging to the investigated factor (Worsley Hunt and Wasserman, 2014). (ii) ChIP-seq peaks may be false positives, recovered because of poor antibody quality (Pickrell *et al.*, 2011). (iii) ChIP-seq returns low-resolution datasets, whose sequences include several nucleotides flanking the target TFBS. ChIP-exo (Rhee and Pugh, 2011) addresses the latter issue, employing a lambda exonuclease to trim ChIP sequences, removing some of the nucleotides flanking the target sites. Alternatively, since most TFs bind their target sequences in open chromatin regions, experimental assays targeting open chromatin like ATAC-seq or DNase-seq (Buenrostro *et al.*, 2013; John *et al.*, 2011) can be employed to recover in vivo genomic locations likely to contain TFBS. ATAC-seq and DNase-seq are generally employed when the factors binding the target regions are not known.In summary, the current high-throughput in vivo and in vitro assays generate datasets of thousands of sequences potentially containing several possible binding configurations of TFBS, thereby enabling better characterizations of TFs binding landscapes.

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	Yes. Limited to short motifs (∼12bp)	*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 2.1. Generally, EMSA and DNase footprinting are used to validate known TFBS, while currently PBMs, HT-SELEX and ChIP-based methods are preferred to discover novel binding sites. ChIP-based assays are the only methods that recover the TF genomic binding locations. The throughput column refers to the number of samples that can be processed in parallel by each method (high: hundreds of samples; low:a few samples).

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

The TFBS motif discovery problem can be formalized as follows. Given a set of positive DNA sequences S, obtained from an experimental assay targeting a certain TF, and a set of negative sequences B the goal is to find one or more recurrent, short and similar subsequences in S that maximize the discriminatory power between S and B. Such subsequences are called patterns or motifs and are likely bound by the investigated TF. The negative set B can contain randomly generated or selected genomic sequences, with similar nucleotide content and length of those in S. The retrieved patterns are used to construct and train a computational model M (motif model), representing the discovered motif. These models can then be used to identify new potential binding sites, given a new set of sequences, and to predict the strength of the TF–DNA binding. Motif discovery can be considered a classification or a regression problem, depending on the type of data used to train M (Tognon *et al.*, 2023). The datasets derived by experimental assays like ChIP-seq or HT-SELEX provide hundreds or thousands of sequences containing binding sites. In this setting, motif discovery becomes a classification problem. In fact, the goal is to discriminate between bound and unbound sites in the input sequences and train the motif model with the identified binding sites. The datasets produced by other experimental technologies like PBMs provide the relative binding strength for large sets of sequences of equal length. Therefore, rather than discriminating between bound and unbound sequences, in this setting M learns the relative binding affinities associated to each target site in the input dataset, transforming motif discovery into a regression problem. In both settings, the final goal is to derive a computational model M, describing the recovered TFBS and capable of predicting new binding events, along with their affinity, in sequences not used during model training. Motif discovery algorithms can be classified in enumerative, alignment-based, probabilistic graphical models, support vector machine (SVM)-based and deep neural network-based methods (**Fig.2.2**).Other approaches to discover TFBS motifs in genomic sequences use phylogenetic footprinting (McCue *et al.*, 2001; Blanchette and Tompa, 2002). The core principle of phylogenetic footprinting is 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. Although phylogenetic footprinting was one of the first techniques proposed for identifying TFBS, it is still widely used to examine 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 utilizes phylogenetic footprinting to discover TFBS. Before describing the algorithms, we briefly review the models to describe TFBS motifs.The most common models to represent TFBS are 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 summarize the discovered TFBS by denoting the most frequently observed nucleotide at each motif position in a prioritized sequence set. Although TFBS have conserved positions not tolerant to mutations (Li and Ovcharenko, 2015), other binding site locations admit alternative nucleotides. Degenerate consensus accommodates ambiguous motif positions employing IUPAC symbols. However, consensus sequences cannot encode the contribution to TF–DNA binding of each nucleotide at each motif position. PWMs address this limitation, providing an additive model with the contribution of each motif position to the binding site. PWMs construct an ungapped alignment between motif candidate sequences and count the frequency of each nucleotide at each position. The statistical significance of PWMs is often measured employing relative entropy (RE) (Stormo, 1998). RE quantifies the difference between computed nucleotide frequencies and those obtained 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 wide success, PWMs still assume independence between motif positions.Probabilistic graphical models address this limitation by modeling dependency between motif nucleotides. These models include high-order PWMs like 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 often visualized as logos with q-mers replacing the single nucleotides, where q is the dependency order between neighboring nucleotides. Importantly, probabilistic graphical models can account for variable spacing between half-sites of two box motifs. However, the number of model's parameters and its complexity grow exponentially with q, often resulting in the model overfitting the

input dataset. SVM-based models train a SVM kernel learning the binding site structure from the input sequence dataset. TFBS are represented by either 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 account for variable spacing between the half-sites of two box motifs, like probabilistic graphical models. Importantly, k-mers indirectly capture k-th order dependencies between neighboring nucleotides. However, simple SVM-based models are limited to consider short k ($\sim$10 bp) and cannot represent longer motifs. Gapped k-mers (Ghandi *et al.*, 2014b) addressed this limitation, handling longer TFBS and sequence degeneration in non-informative motif positions. To visualize the discovered motifs, SVM-based models are often reduced to PWMs computed 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. Although DNN-based models are accurate and powerful, their 'black box' nature is a major limitation (Park *et al.*, 2020). Many frameworks visualize the discovered motifs as PWMs, computed 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. Therefore, motifs learned by single kernels and the resulting PWMs are often redundant with each other. DeepLIFT (Shrikumar *et al.*, 2017) proposed a method to assign importance scores to the kernels. Comparing the activation of each neuron to a reference value, DeepLIFT selects which kernels contribute most to the TFBS definition, 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, computing interpretable models without losing some information learned by the DNN is still an open challenge.

Enumerative methods

Enumerative motif discovery algorithms (**Fig.2.2**) assume that motifs are overrepresented patterns in the input dataset S, with respect to a set of background genomic sequences B. Enumerative algorithms may assume that the motif length $|M|$ is known a priori. Given $|M| = k$, the general idea is to collect the approximate occurrences of all potential 4^k k-mers in the sequences of S and assess if the difference between the number of matches found in S and B or the expected number of matches from a background model is statistically significant. Then, a PWM is obtained building an ungapped alignment from the statistically significant k-mers. Searching the approximate occurrences of all 4^k k-mers quickly becomes impractical, even for small k. Early proposals introduced the usage of heuristics to reduce the search space, for example, searching only 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). However, mismatches can occur at any motif position. Weeder (Pavesi *et al.*, 2001, 2004b) and SMILE (Marsan and Sagot, 2000) proposed using suffix trees (STs) (Weiner, 1973) to efficiently explore the entire motif search space. They leverage the indexing capabilities of STs to perform approximate pattern matching, without restrictions on mismatching positions. This enabled achieving high accuracy in motif discovery, while reducing computational costs. To determine the statistical significance of motif candidates, SMILE and Weeder compare the motifs frequencies in S with those in a set of random genomic sequences or the promoters of the same organism, respectively. However, these approaches can be computationally intensive and are not scalable on the large datasets generated by PBMs, HT-SELEX or ChIP assays (Liu *et al.*, 2018). Therefore, more efficient approaches specifically tailored to work on large datasets were proposed. MDscan (Liu *et al.*, 2002) and Amadeus (Linhart *et al.*, 2008) use word enumeration to discover motif candidates in sequence datasets. MDscan employs ChIP peaks shape to identify non-redundant patterns abundant in the most enriched sequences and uses a third-order Markov background model to assess motif statistical significance. Amadeus evaluates all k-mers in S and groups similar patterns in list. Each list is grouped into motifs, statistically evaluated using a hypergeometric test. However, word enumeration can be still computationally demanding. To address this challenge, DREME (Bailey, 2011) proposed using regular expressions to count approximate frequencies of motifs in S and B. To evaluate the motifs' statistical significance, DREME employs Fisher's exact test, comparing the number of sequences in S and B in which the motifs occur. However, regular expressions can be computationally expensive when analyzing large S, and may detect false positives or miss motifs. Trawler, HOMER and STREME (Ettwiller *et al.*, 2007; Heinz *et al.*, 2010; Bailey, 2021) reintroduced STs, proposing different optimizations to make the methods scalable on large datasets. Trawler and HOMER optimized the statistical assessment step using z-scores derived from the normal approximation to the binomial distribution and the hypergeometric distribution, respectively. Instead of improving the statistical assessment, STREME reduces the motif search space by first identifying overrepresented seed words of different lengths on the ST. Then, STREME counts the

number of approximate matches of the most significant words on the ST. By identifying seeds of different lengths, STREME discover motifs of different lengths in one single tree visit.

Alignment-based methods

Alignment-based motif discovery algorithms compute alignment profiles to describe motifs binding preferences (**Fig.2.2**), avoiding exhaustive k-mer enumeration. This approach involves constructing an alignment by selecting motif candidate sequences from the input dataset S and evaluating the resulting profile using various measures, like nucleotide conservation, information content or profile statistical significance. Motif statistical significance is determined by computing the probability of obtaining the same alignment from either a background dataset B or random sequences. Alignment-based motif discovery algorithms typically assume that the motif length $|M|$is known a priori. For alignment-based algorithms, motif discovery can be formalized as a combinatorial problem. Given $|M| = k$ the goal is to find the best 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 zero or one binding site. Therefore, there exist $(\sum_{s \in S} |s| - |M| + 1)^{|S|}$ possible profiles, built by combining k-mers in all possible ways. Since enumerating all possible solutions is computationally impractical even for small datasets, alignment-based algorithms employ heuristics, such as greedy (Hertz and Stormo, 1999), expectation-maximization (EM) (Bailey *et al.*, 1994), stochastic (e.g. Gibbs sampling) (Lawrence *et al.*, 1993), or genetic algorithms (Lee *et al.*, 2018). CONSENSUS (Hertz and Stormo, 1999) proposed a greedy approach to construct alignment profiles incrementally. It solves the problem initially on two sequences and progressively solves it by adding the remaining sequences $s \in S$ one by one. CONSENSUS stores the best partial alignments hoping to find the highest-scoring profiles. However, if motifs are not conserved, CONSENSUS may potentially discard the highest-scoring solutions. The MEME algorithm (Bailey *et al.*, 1994; Bailey and Elkan, 1995; Bailey *et al.*, 2006) proposed a different strategy based on EM. It iteratively refines an initial profile by substituting some k-mers in the profile, with others more likely to produce better solutions. MEME evaluates the fit of each k-mer in $s \in S$ to the current profile, rather than a background model. MEME identifies motifs occurring more than once in each sequence and computes their statistical significance, and the method does not rely on TFBS conservation. However, the algorithm may converge prematurely to local maxima and convergence heavily depends on the algorithm starting conditions. In contrast to MEME, Gibbs sampling (Lawrence and Reilly, 1990) employs a stochastic approach to add k-mers to the alignment instead of a deterministic one based on the profile fit. Gibbs sampling replaces k-mers in the profile with others selected with probability proportional to its likelihood score. The algorithm's stochastic nature reduces its likelihood to converge to local maxima, but it may require multiple runs to achieve reliable results. However, several methods using 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). Genetic algorithms are an alternative approach overcoming the limitations of EM and stochastic methods. GADEM (Li, 2009) combined EM local search with genetic algorithms to refine profiles, avoid convergence to local maxima and overcome Gibbs sampling stochastic nature. However, due to their computational complexity, genetic algorithms are computationally demanding when analyzing thousands of sequences. Using alignment profiles, the solution space grows exponentially with the size of S and even with employing heuristics analyzing thousands of sequences is computationally impractical (Zambelli *et al.*, 2013). MEME-ChIP (Machanick and Bailey, 2011) and STEME (Reid and Wernisch, 2011) improved the MEME algorithm to analyze ChIP datasets. While MEME-ChIP focuses the analysis on a random subset of sequences, STEME speeds up EM steps indexing the sequences in a suffix tree. However, using random subsets of S may cause missing critical motif instances and constructing ST from thousands of sequences may be computationally demanding. ChIPMunk (Kulakovskiy *et al.*, 2010) proposed a greedy profile optimization like EM developed to discover motifs in large ChIP-seq datasets, while accounting for ChIP peaks shape. XXmotif (Hartmann *et al.*, 2013) and ProSampler (Li *et al.*, 2019b) proposed methods combining enumerative motif discovery with iterative and stochastic profile refinement, respectively.

Probabilistic graphical model-based algorithms

The inclusion of dependencies between nucleotides in TFBS has been subject of debate (Tomovic and Oakeley, 2007; Morris *et al.*, 2011; Zhao and Stormo, 2011). Some studies have shown that dependencies exist between neighboring and non-neighboring nucleotides in TFBS (Slattery *et al.*, 2014; Rohs *et al.*, 2010). Enumerative and alignment-based algorithms represent motifs as PWMs, which do not account

for dependencies between the binding site positions. PWMs can be extended to account for the frequency of di- or trinucleotides (high-order PWMs), like DWMs (Siddharthan, 2010). Dimont (Grau *et al.*, 2013) and diChIPMunk (Kulakovskiy *et al.*, 2013a) proposed extensions to alignment-based methods to discover and represent motifs as DWMs. However, these methods capture dependencies only between neighboring nucleotides. Probabilistic graphical models (**Fig.2.2**) such as BNs, MMs or HMMs provide powerful frameworks for capturing dependencies between TFBS nucleotides. In (Barash *et al.*, 2003) the authors proposed using BNs trained via EM to model TFBS. The proposed approach captures dependencies between neighboring and non-neighboring positions but assumes the same order of dependence throughout the entire motif. Similarly, in (Ben-Gal *et al.*, 2005), the authors introduced VOBN models. VOBNs use BNs accounting for variable orders of dependencies between positions. However, training BNs is not computationally scalable when analyzing thousands of sequences and these models are prone to overfitting when trained on hundreds of sequences. MMs and HMMs provide more efficient and scalable frameworks than BNs to include dependencies between motif positions. Therefore, researchers focused on developing algorithms using these models to learn dependencies in large sequence datasets produced by NGS assays. TFFMs (Mathelier and Wasserman, 2013) and Discrover (Maaskola and Rajewsky, 2014) proposed HMM-based models learning the dinucleotide dependencies between neighboring motif positions in large sequence datasets. In addition, TFFMs learn the properties of the sequences flanking the TFBS. MMs can be extended to capture different orders of dependencies between neighboring nucleotides, as demonstrated in (Eggeling *et al.*, 2014), where the authors proposed a method to discover CTCF citepbell1999protein motifs using variable-order MMs. Similarly, MMs can also be extended to capture dependencies between non-neighboring nucleotides as proposed in Slim (Keilwagen and Grau, 2015). However, MMs and HMMs typically only capture low-order dependencies. BaMMotif (Siebert and Söding, 2016; Ge *et al.*, 2021) proposed a motif discovery algorithm employing a Bayesian approach to efficiently train Markov models up to fifth-order dependencies on thousands of sequences.

SVM-based methods

SVMs (Boser *et al.*, 1992) have been successfully applied to different problems in computational biology (Ben-Hur *et al.*, 2008), including TFBS motif discovery (**Fig.2.2**). This is achieved by decomposing bound (foreground dataset S) and unbound sequences (background dataset B) in k-mers and using their frequencies as features to train a sequence similarity kernel (Ben-Hur *et al.*, 2008). Generally, to each k-mer is assigned a weight proportional to its contribution to the definition of the positive or negative training sets, or to its likelihood of being a motif candidate. While earlier methods (Leslie *et al.*, 2001; Eskin *et al.*, 2002; Kuang *et al.*, 2005) were designed for protein sequence homology, recent SVM-based algorithms have been developed to discover TFBS motifs. Furthermore, SVMs can efficiently analyze datasets of thousands of sequences. Kmer-SVM (Lee *et al.*, 2011; Fletez-Brant *et al.*, 2013) proposed a method to discover TFBS motifs in sequence datasets, using the spectrum kernel (Leslie *et al.*, 2001) Kmer-SVM counts the exact matches for all contiguous k-mers in S and B, building the k-mers feature space. The mismatch and wildcard kernels (Kuang *et al.*, 2005; Leslie and Kuang, 2003) were introduced to count k-mer frequencies while allowing a fixed number of mismatching positions for each k-mer. This approach was later extended to allow for less restrictive k-mer frequency estimation, offering flexibility in the motif structure without affecting scalability on large datasets. Agius and coworkers (Agius *et al.*, 2010) extended the concept of mismatch kernels by developing the di-mismatch kernel. The di-mismatch kernel is a first-order Markov mismatch kernel based on the dinucleotide alphabet, which handles sequence variability and accounts for dependencies between neighboring nucleotides. To maintain scalability on large datasets small k ($\sim$10) is used, discovering short motifs. However, TFBS lengths range between 6 and 20 bp, making it challenging to fully characterize longer motifs with short k-mers. In addition, increasing k often results in sparse feature vectors overfitting the training dataset. Gapped k-mers (Ghandi *et al.*, 2014b) proposed to represent 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 the method's dependency on parameters' choice. LS-GKM (Lee, 2016) optimizes the algorithm for scalable SVM training with gapped k-mers on large-scale sequence datasets. LS-GKM also provides other kernels for SVM training.

DNN-based methods

DNNs have become increasingly popular 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). due to their ability to learn complex patterns (Park and Kellis, 2015) from large omics datasets (Zhang *et al.*, 2019). Convolutional neural networks (CNNs) (LeCun *et al.*, 2015) , originally developed for image classification (LeCun *et al.*, 2015; Sainath *et al.*, 2013; Vu *et al.*, 2017), have been successfully applied to analyze in vivo TF–DNA interactions (Alipanahi *et al.*, 2015; Zhou and Troyanskaya, 2015; Kelley *et al.*, 2016; Zeng *et al.*, 2016a) (**Fig.2.2**). CNNs apply non-linear transformation to input data, learning and representing complex patterns in a high-dimensional space (Bengio *et al.*, 2013) This simplifies classification tasks and enables accurate prediction of TFBS in genomic sequences. CNNs represent genomic sequences as 1D or 2D images with four associated channels (A, C, G, T) (Zeng *et al.*, 2016a). Therefore, classifying TFBS in genomic sequences becomes a two-class image classification problem. Typically, CNN architectures designed for motif discovery and classification consist of one or more sets of four layers: the convolutional layer, the max-pooling layer, the fully connected NN layer and the output layer (Zeng *et al.*, 2016a). DeepBind (Alipanahi *et al.*, 2015) and Basset (Kelley *et al.*, 2016) proposed two CNN architectures to discover motifs in different datasets, such as ChIP-seq, HT-SELEX, PBM and DNase-seq. The discovered motifs in DeepBind and Basset are visualized as PWMs. The PWMs are computed by aligning and grouping the sequences that activate the convolutional layer. While DeepBind and Basset have demonstrated promising results in predicting TFBS, their performance may be limited by the quality of training data and the significant computational resources and time required for model training. These limitations have led to the development of novel methods, such as BPNet (Avsec *et al.*, 2021a), which address some of these issues by incorporating additional features in the model and using more efficient training processes. BPNet proposed a dilated CNN architecture, allowing the model to learn and integrate diverse complex features without sacrificing the spatial and base resolution of the input data. However, TF–DNA interactions involve not only the direct binding between TF and DNA but also the interactions between multiple binding subregions (long-term interactions) and the nucleotides with high-order structures of TFs (short-term interactions). Long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and bi-directional LSTMs (BLSTMs) can efficiently capture long-term and short-term dependencies of sequential signals. LSTMs and BLSTMs are well suited for modeling TF–DNA interactions as genomic sequences can be viewed as sequential signals with long-term and short-term dependencies. DeeperBind (Hassanzadeh and Wang, 2016) introduced a hybrid CNN–LSTM architecture removing the pooling layer to maintain the 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 the DanQ approach by incorporating additional features in the model and using a Siamese BLSTM architecture to improve model training.

2.2 Transcription factor binding sites databases

With the recent advancement in experimental technologies, a vast amount of TF-related data have been generated and stored in databases (**Table 2.2**). The ENCODE project (Consortium *et al.*, 2012) provides multiple data on functional elements in the human genome collected across different tissues and cell types. ENCODE stores TF-related genomic data such as ChIP-seq targeting several TFs and DNase-seq. Similarly, Cistrome (Zheng *et al.*, 2019) and GTRD (Kolmykov *et al.*, 2021) provide TF-related genomic data from different organisms and across different species, cell types and tissues, respectively. Furthermore, GTRD stores large 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, experimentally derived and computationally predicted TFBS motifs for several TFs from different species. They store PWMs and DWMs obtained by analyzing ChIP-seq and SELEX datasets. In addition, HOCOMOCO models were generated integrating sequence datasets with evolutionary conservation and DNA shape. Similarly, Cis-BP (Weirauch *et al.*, 2014) stores experimentally derived and computationally predicted PWMs, obtained integrating multiple sources, including published literature, other databases and experimental datasets. TRANSFAC (Wingender *et al.*, 1996, 2000) collects experimentally validated and manually curated PWMs for various TFs from different eukaryotic organisms, and includes data on TF-associated proteins, DNA binding domains and, regulatory elements. FactorBook (Pratt *et al.*, 2022) provides computationally predicted PWMs generated analyzing ENCODE data and includes TF expression data across tissues and cell types. Unibind (Puig *et al.*, 2021) collects experimentally validated and curated PWMs from different organisms, providing information on structural properties and conformation of TF–DNA complexes and their genomic binding locations across different cell types and tissues. UniPROBE (Newburger and Bulyk, 2009) stores curated PWMs

for several eukaryotic TFs, generated analyzing PBM datasets. HTRIdb (Bovolenta *et al.*, 2012) stores data on TF–target genes interactions in human, collected from published literature and other databases, in different cell types, experimental methods and disease state, also providing functional annotations for the target genes. TFcancer (Huang *et al.*, 2021b) collects TF–gene interactions across 33 cancer types, providing tools to identify TF expression alterations and their roles in biological processes and signaling pathways in cancer.

Type	**Name**	**Reference**	**Data type**	**Model organism**	**TFs**
Sequence database	ENCODE	(Consortium *et al.*, 2012)	ChIP-seq DNase-seq ATAC-seq	*Caenorhabditis elegans* *Drosophila melanogaster* *H. sapiens* *M. musculus*	> 1,500
Sequence database	Cistrome	(Zheng *et al.*, 2019)	ChIP-seq DNase-seq	*H. sapiens* *M. musculus*	1,773 (ChIP-seq)
Sequence database	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)
Motif models database	JASPAR	(Sandelin *et al.*, 2004; Fornes *et al.*, 2020)	PWMs DWMs	53 species	> 1,500
Motif models database	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
Motif models database	TRANSFAC	(Wingender *et al.*, 1996, 2000)	PWMs	> 300 species	> 10,000
Motif models database	FactorBook	(Pratt *et al.*, 2022)	PWMs	*H. sapiens* *M. musculus*	> 881 (human) 49 (mouse)
Motif models database	Unibind	(Puig *et al.*, 2021)	PWMs	*A. thaliana* *C. elegans* *D. rerio* *D. melanogaster* *H. sapiens* *M. musculus* *R. norvegicus* *S. cerevisiae* *S. pombe*	841

Motif models database	UniPROBE	(Newburger and Bulyk, 2009)	PWMs	*C. elegans* *C. parvum* *H. sapiens* *M. musculus* *P.falciparum* *S. cerevisiae* *V. harveyi*	726
TF–target gene interaction database	HTRIdb	(Bovolenta *et al.*, 2012)	TF–gene interaction networks	*H. sapiens*	284
TF–disease association database	TFcancer	(Huang *et al.*, 2021b)	TF–cancer associations	*H. sapiens*	364

Table 2.2. The table presents a summary of the TF-related databases discussed in Section 4. For each database, the table reports the database main purpose (Type), the available type of data (Data type), the model organisms for which data are provided (Model organisms), the number of TFs (TFs) and the database website (Website).

2.3 Downstream analysis on transcription factor binding sites

The discovered motifs can be employed in several downstream analyses: motif comparison, motif scanning, motif enrichment analysis and assessing genetic variants effects on TF–DNA binding affinity (Tognon *et al.*, 2023). Motif comparison measures the similarity between the discovered motifs and annotated TFBS. Motif comparison allows for linking known TFs to the newly discovered motifs (Gupta *et al.*, 2007) and inferring the relationship between the input sequences and function of the annotated TF (Weirauch *et al.*, 2014). For this task, several tools have been developed such as Tomtom, STAMP, MACRO-APE or MoSBAT (Gupta *et al.*, 2007; Mahony and Benos, 2007; Vorontsov *et al.*, 2013; Lambert *et al.*, 2016). These tools search annotated database for motifs matching the input consensus sequence or inferred motif matrix. Moreover, motif comparison tools have been developed to interpret and annotate the potential motifs encoded in the convolutional filters of a CNN model. Motif scanning scans sets of genomic regions searching for potential occurrences of the input motif. The goal is to recover sets of potential binding locations for the investigated factor. Given a motif model (e.g. a PWM) and a set of sequences, motif scanning algorithms assign a score to each sequence using the input model. A common challenge is to determine a reliable cutoff on the scores assigned to the sequences to discriminate between true and false binding events (Boeva, 2016). Several motif scanning tools are currently available such as MOODS, FIMO or PWMscan (Korhonen *et al.*, 2009; Grant *et al.*, 2011; Ambrosini *et al.*, 2018). The HOMER suite (Heinz *et al.*, 2010) also provides a motif scanning functionality. Recently, MOODS was extended to search instances of motifs modeled as high-order PWMs (Korhonen *et al.*, 2017). GRAFIMO (Tognon *et al.*, 2021) extended classical motif scanning to panels of thousands of genomes encoded in genome graphs (Paten *et al.*, 2017), considering individual genetic variants and haplotypes while searching for potential motif occurrences. Motif enrichment analysis (MEA) searches for over- and underrepresented motifs in gene regulatory regions. Analyzing the TFBS enrichment in regulatory regions governing sets of genes, researchers can link the investigated TFs to their function within the cell environment. MEA consists of two steps: (i) scanning regulatory regions for motif occurrences and (ii) statistical testing of motif enrichment. TFs whose motifs are significantly overrepresented (enriched) in the scanned regulatory regions are marked as transcriptional regulators for the target gene set. There are many MEA tools available to the community, such as Clover, Pscan, AME or oPOSSUM-3 (Frith *et al.*, 2004a; Zambelli *et al.*, 2009; McLeay and Bailey, 2010; Kwon *et al.*, 2012). HOMER (Heinz *et al.*, 2010) provides a functionality to perform MEA. Haystack (Pinello *et al.*, 2018) proposed an integrated MEA strategy, investigating motif enrichment in cell-type-specific regions and incorporating gene expression data to assess the transcriptional activity of the studied factors and their impact on the regulated genes. Genetic variants have been shown to impact TF–DNA binding events (De Gobbi *et al.*, 2006; Wienert *et al.*, 2015; Weinhold *et al.*, 2014) including variants associated with common diseases in regulatory elements (Maurano *et al.*, 2012) potentially altering the transcriptional state of the cell (Deplancke *et al.*, 2016). As a result, there has been a growing interest in developing tools to predict the impact of variants on TFBS (**Table 2.3**). TRAP (Thomas-Chollier *et al.*, 2011) and CATO (Maurano *et al.*, 2015) use PWMs to predict the impact of variants on TFBS by comparing the binding affinity scores of reference and alternative sequences. TRAP repeats the procedure on a collection of TFBS, reporting the motif showing

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

Table 2.3. Software to assess genetic variants impact on transcription factor binding sites. The table provides an overview of the tools for predicting the impact of variants on TFBS as discussed in Section 5. For each tool, the table report the employed TFBS model (Motif model), the original data type used to test each method in their original publication (Original input data type), the output type (Output), the year (Year) and the associated publication (Reference)

the largest score change. CATO, instead, provides a ranked list of disrupted motifs, obtained using a logistic model trained with the information content difference between reference and alternative sequences, TF occupancy and phylogenetic conservation. However, these methods are not scalable when analyzing thousands of single nucleotide polymorphisms (SNPs). atSNP (Zuo *et al.*, 2015) proposed a scalable strategy to assess the impact of thousands of SNPs on TFBS by computing the statistical significance of the computed affinity scores, in addition to the difference between the reference and alternative sequence binding scores using PWMs. GRAFIMO (Tognon *et al.*, 2021) extended the scalability to millions of SNPs by scanning collections of PWMs on genome graphs, while accounting for haplotypes. MotifRaptor (Yao *et al.*, 2021) integrates chromatin accessibility, gene expression and GWAS summary statistics, to predict and annotate functional effects for large non-coding variant datasets, using PWMs. DeltaSVM (Lee *et al.*, 2015) and GkmExplain (Shrikumar *et al.*, 2019) se SVM-based motif models to assess variant impact. DeltaSVM scans DNA positions overlapping each SNP in the input dataset using a pretrained list of k-mers with associated weights and computing the difference between the reference and alternative sequence scores. However, it assesses the impact of individual variants, not accounting for relationships between variants. GkmExplain overcomes this limitation by considering the impact of variants not in individual positions, but on sequence features, like 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 variant effect on binding affinities by 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) employs in silico saturated mutagenesis by learning critical nucleotides governing chromatin accessibility. Basset assigns importance scores to each position in the input sequences and attempts to map the variants' impact to the TFBS in the input sequences. Basenji (Kelley *et al.*, 2018) extends Basset's workflow by providing functional annotations to SNPs affecting sequence features like TFBS and returning potential changes in gene expression patterns. However, Basenji is limited to predict 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 more comprehensive and accurate functional effects of variants on sequence elements and gene expression.

MotifGraph

Several studies have demonstrated that transcription factors exhibit binding sites that can vary across populations (Kasowski *et al.*, 2010), cell types (Arvey *et al.*, 2012; Gertz *et al.*, 2013), and even among individuals (Tognon *et al.*, 2021). Employing algorithms that solely rely on the reference genome may yield generalized models, leading to erroneous predictions of TFBS when applied to personal genomics settings. A more robust approach involves the development of motif models incorporating data from diverse populations, cell types, or individual genomes, ultimately enhancung the accuracy of predictions regarding the occurrence of TFBS (**Fig.3.3**). Furthermore, these models offer the potential for more accurate predictions on the imoact of non-coding variants on TF-DNA binding events and allow for the encoding of variable orders of nucleotide dependencies. Notably, graph data structures often provide interpretable and intuitive frameworks. With these goals in mind, we have developed the MotifGraph framework, which seamlessly combines the strengths of probabilistic graphical-based and SVM-based motif discovery algorithms and models. The MotifGraph model G is defined by a set of vertices V, edges E, and paths P. Each vertex $v \in V$ in the graph is assigned a label representing a nucleotide, denoted mathematically as label$(v) \in A, C, G, T$. Meanwhile, each edge $e \in E$ denotes the allowed connections

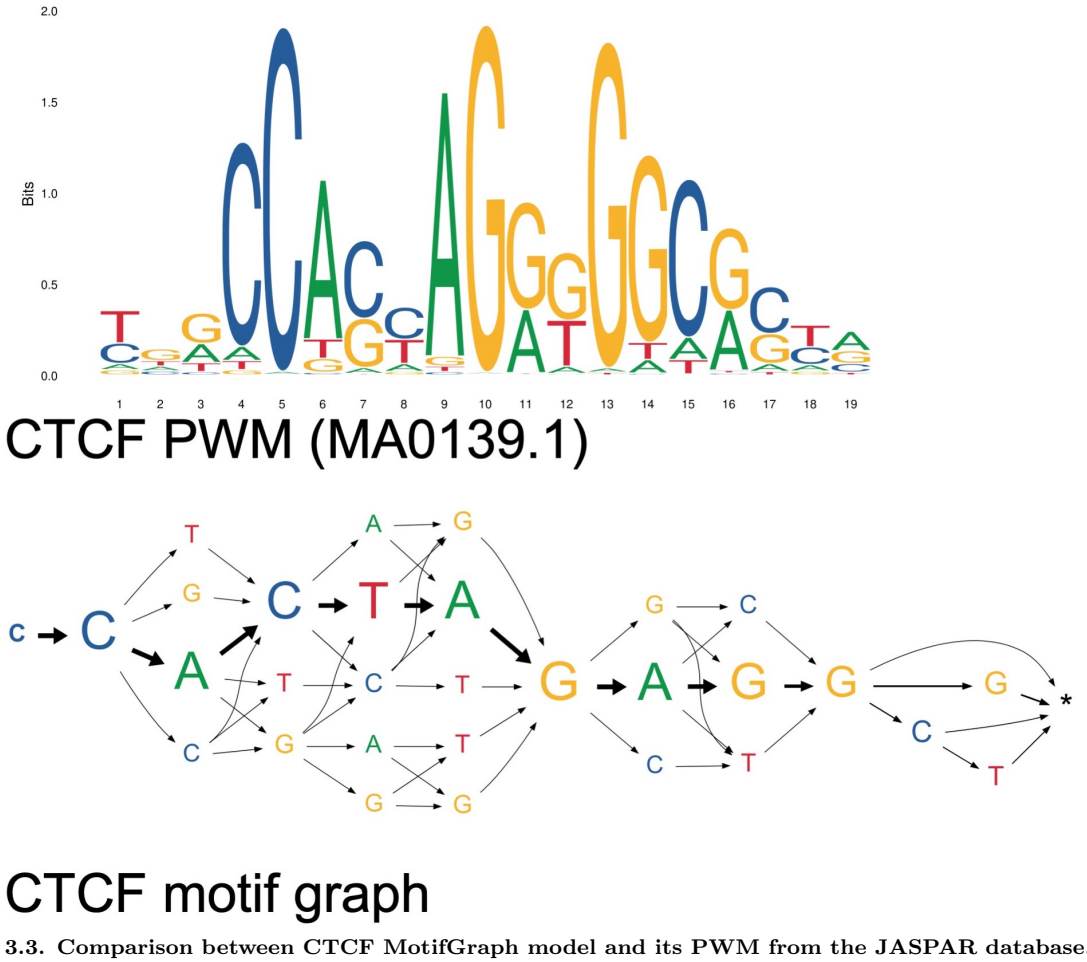


Figure 3.3. Comparison between CTCF MotifGraph model and its PWM from the JASPAR database. On top the CTCF motif available on JASPAR database (MA0139.1). On the bottom the Motif Graph model trained using 100 k-mers obtained from a ChIP-seq experiment trageting CTCF binding site on HepG2 cell line.

between consecutive nucleotides within the TFBS motif. Further, each path $p \in P$ corresponds to a haplotype embedded in G that represents an individual sequence used to train the MotifGraph model. Currently the model is primarily designed to encode first-order dependencies between consecutive nucleotides. However, it retains the flexibility to accomodate motifs of variable lengths, all while storing essential information about the training sequences.

3.1 Motif Graph model construction

MotifGraph motif discovery process (**Algorithm 1**) encompasses two primary phases: the prioritization of k-mers and the training and construction of the graph model. During the k-mer prioritization step, MotifGraph leverages the k-mer based motif discovery method proposed in LS-GKM (Ghandi *et al.*, 2014a; Lee, 2016). Within this framework, for each k-mer of length k present in the positive sequence dataset S and the background dataset B, the algorithm tallies the number of matches in both S and B, accommodating mismatching positions. Subsequently, a cosine similarity kernel is trained using the frequencies of the recovered k-mers, with the trained kernel assigning a weight to each k-mer in S and B. The k-mers weights reflect their influence in defining either the foreground or background dataset. The algorithm then sorts the k-mers based on their weight scores. The Motif Graph model undergoes an iterative training process employing a greedy approach, systematically incorporating the top-ranked k-mers into G, one by one (**Algorithm 2**.) Each k-mer is aligned with the existing MotifGraph model, aiming to maximize the alignment of nucleotides with the current model. During this alignment process, the algorithm adjusts the input k-mer both to the right and left, within a specified offset number of nucleotides (a value set at 3 in our experiments). After constructing the model, we create a scoring matrix analogous to the well-known Position Specific Scoring Matrix (*PSSM*), taking into account first-order nucleotide dependencies, akin to DWMs. This scoring matrix assigns a likelihood score and categorizes new sequences as potentially bound or not by the investigated factor. Essentially, the score indicates the likelihood of the scanned sequence containing a potential binding site. To score a sequence, we slide the scoring matrix along the string, employing a method akin to classical PWM scanning tools like FIMO (Grant *et al.*, 2011).

Algorithm 1: Motif Graph motif discovery.

Input: S, B, k
Output: G
frequencies ← countFrequencies(S, B, k)
kernel ← trainKernel(frequencies)
kmers, weights ← extractWeights(kernel)
rankedKmer ← sort(kmers, weights)
$G \leftarrow \emptyset$
for *kmer in rankedKmers* **do**
  $G \leftarrow$ addKmers(G, kmer)
return G

Algorithm 2: Motif Graph model training.

Input: G, kmer, i
Output: G
if $i = 1$ **then**
  return G
for j *in* 1 *to* 3 **do**
  matchesLeftOffset, alignmnetLeft ← countMatchesLeftOffset(G, kmer, j)
for j *in* 1 *to* 3 **do**
  matchesRightOffset, alignmnetRight ← countMatchesRightOffset(G, kmer, j)
alignment ← getBestAlignment(matchesLeftOffset, alignmentLeft, matchesRightOffset, alignmentRight)
$G \leftarrow$ insertKmer(G, alignment)
return G

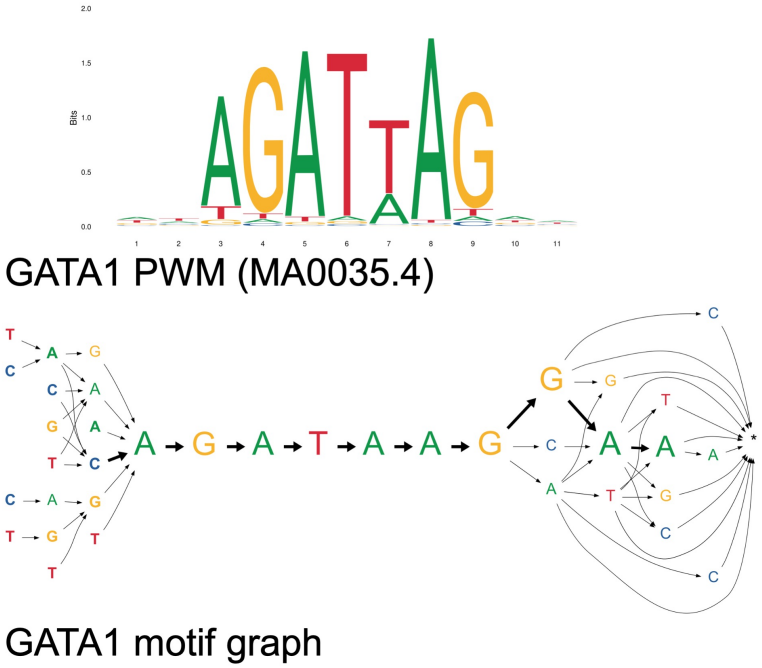


Figure 3.4. Comparison between GATA1 MotifGraph model and its PWM from the JASPAR database. On top the GATA1 motif available on JASPAR database (MA0035.4). On the bottom the Motif Graph model trained using 100 k-mers obtained from a ChIP-seq experiment trageting GATA1 binding site on K562 cell line.

Training k-mers	10	50	100	200	350	500	750
AUC	0.69	0.69	0.72	0.63	0.60	0.55	0.55
F1-score	0.62	0.69	0.71	0.65	0.62	0.57	0.56

Table 3.4. CTCF Motif Graph model AUC and F1-scores values with different number of training k-mers.

3.2 Testing the MotifGraph model

3.2.1 A preliminary training approach

To test our motif discovery algorithm we obtained 10,000 ChIP-seq peak sequences from the ENCODE Project database (Consortium *et al.*, 2012) for CTCF and GATA1 obtained on the HepG2 and K562 cell lines, respectively. We sorted the ChIP-seq peaks according to their enrichment score, hence we tested our algorithm on reliable peaks. Interestingly, the trained MotifGraph models were closed to the motif PWMs on the JASPAR database (Sandelin *et al.*, 2004) for both TFs (**Fig.3.3** and **Fig.3.4**), capturing their consensus sequence. In the models the main motif sequences are highlighted by edge thickness, which is proportional to the number of training k-mers supporting each $p \in P$. We tested the discriminative power of CTCF and GATA1 MotifGraph, training the models with different numbers of k-mers (**Fig.3.5**), to establish the optimal number of training sequences. We assessed models' discriminative power via cross-validation splitting S and B (75% training and 25%, testing). We trained the models using 10, 50, 100, 200, 350, 500, and 750 k-mers. For CTCF the best prformance in terms of both AUC (0.72) and F1-score (0.71) were obtained using 100 k-mers (**Table 3.4**). For GATA1 the model returned the best AUC using 200 k-mers (0.76), while the best F1-score (0.70) was obtained training the model with 100 sequences (**Table 3.5**). We also compared MotifGraph models discriminative power against the corresponding PWMs and DWMs downloaded from JASPAR (Sandelin *et al.*, 2004) and HOCOMOCO (Kulakovskiy *et al.*, 2016), respectively. For CTCF, both PWM and DWM showed better performance than our model (**Fig.3.6 (A)**). On the other hand, on GATA1 data our model behaved better than PWMs, although it still performed worse than DWMs (**Fig.3.6 (B)**).

Training k-mers	10	50	100	200	350	500	750
AUC	0.73	0.75	0.75	0.76	0.74	0.74	0.71
F1-score	0.68	0.68	0.70	0.69	0.70	0.69	0.67

Table 3.5. GATA1 Motif Graph model AUC and F1-scores values with different number of training k-mers.

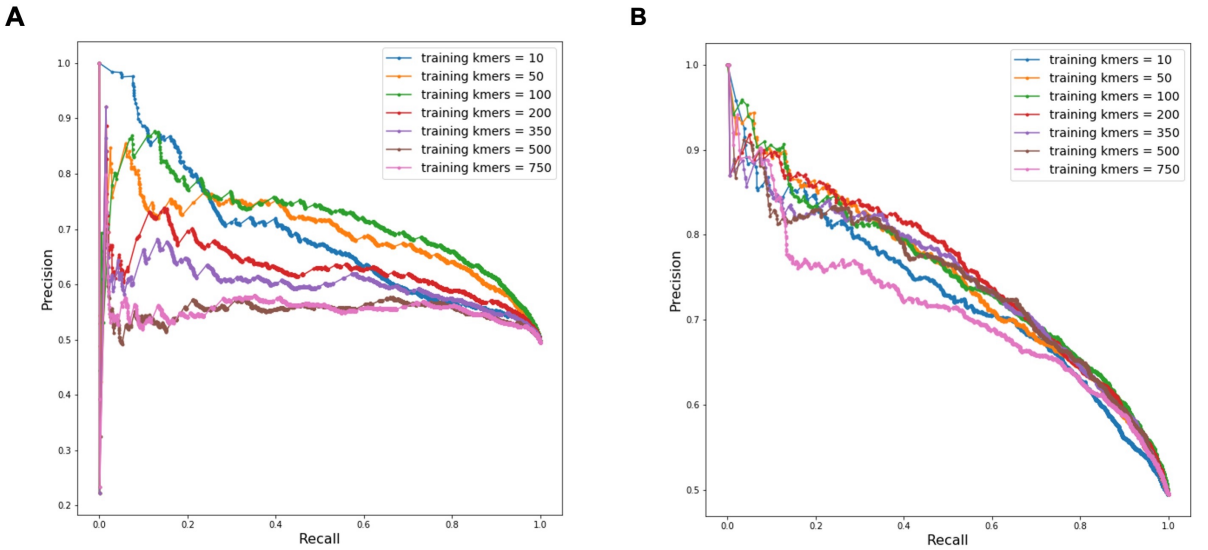


Figure 3.5. Precision-Recall curves obtained varying the number of k-mers used to train the Motif Graph models. To establish the number of training k-mers returning the best discriminative performance we computed the Precision-Recall curves of different Motif Graph models trained using 10, 50, 100, 200, 350, 500, and 750 k-mers on **(A)** CTCF and **(B)** GATA1

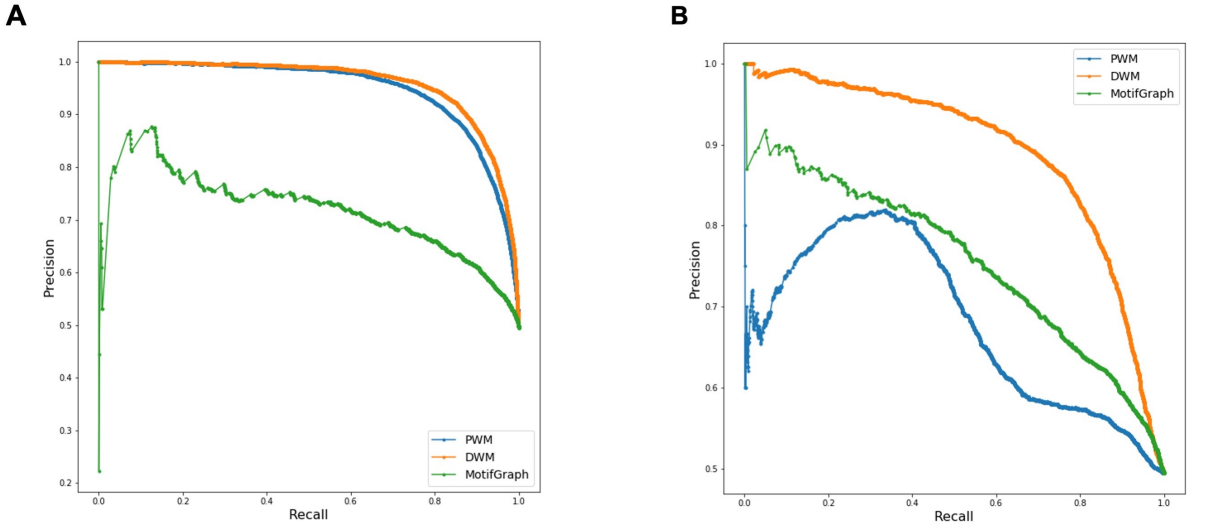


Figure 3.6. Comparing Motif Graph, PWM, and DWM Precision-Recall curves. We compared the discriminative power of the Motif Graph models against that of PWMs and DWMs for both **(A)** CTCF and **(B)** GATA1.

3.2.2 Improving k-mers prioritazion

To improve the performance of our models we focused on improving k-mers selection procedure. To this goal in mind, we developed a k-mers prioritazition procedure similar to LS-GKM algorithm. Given S and B, we compute gapped k-mers frequencies for each sequence in the input datasets. Using the gapped k-mers frequencies we compute a cosine similarity matrix measuring cosine similarity between each sequence pair in S and B. We then identify potential sequence clusters running the Leiden algorithm (Traag *et al.*, 2019) on a graph G_L constructed from the cosine similarity matrix. G_L is defined by a set of vertices V and edges E, where each $v \in V$ is a sequence in S and B, $e \in E$ is a term of comparison between two sequences, and weight(e) is the cosine similarity value between the two sequences linked by e. We then train a MotifGraph G with ech sequence cluster separately. To test our procedure we used CTCF ChIP-seq peak sequences on K562 cell lines for S, while B was obtained by shuffling the sequences in S, maintaining first-order dependencies.By plotting the cosine similarity matrix using UMAP we observe two clear clusters separating foreground (bound) and background (unbound) sequences (**Fig.3.7 (A)**). Furthermore, the clusters correlates with the binding affinity scores computed running FIMO (Grant *et al.*, 2011) on the sequences in S and B (**Fig.3.7 (B)**).

By running Leiden algorithm on G_L, it identifies five different clusters. While it keeps the background sequences in a single cluster, the foreground sequences are divided in four clusters (**Fig.3.8 (A)**). To

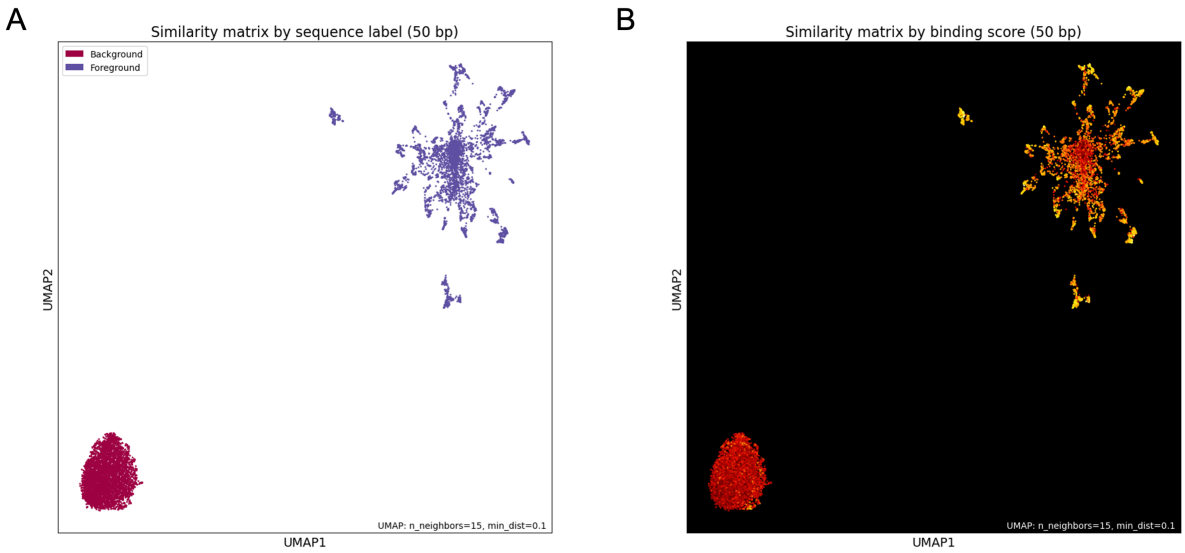


Figure 3.7. UMAP plot of foreground and background sequence clusters. **(A)** Plotting the cosine similarity matrix with UMAP computed on CTCF ChIP-seq data (cell line K562) the foreground and background sequence clusters are identified. **(B)** The clusters correlates with the binding affinity scores computed on the sequences running FIMO.

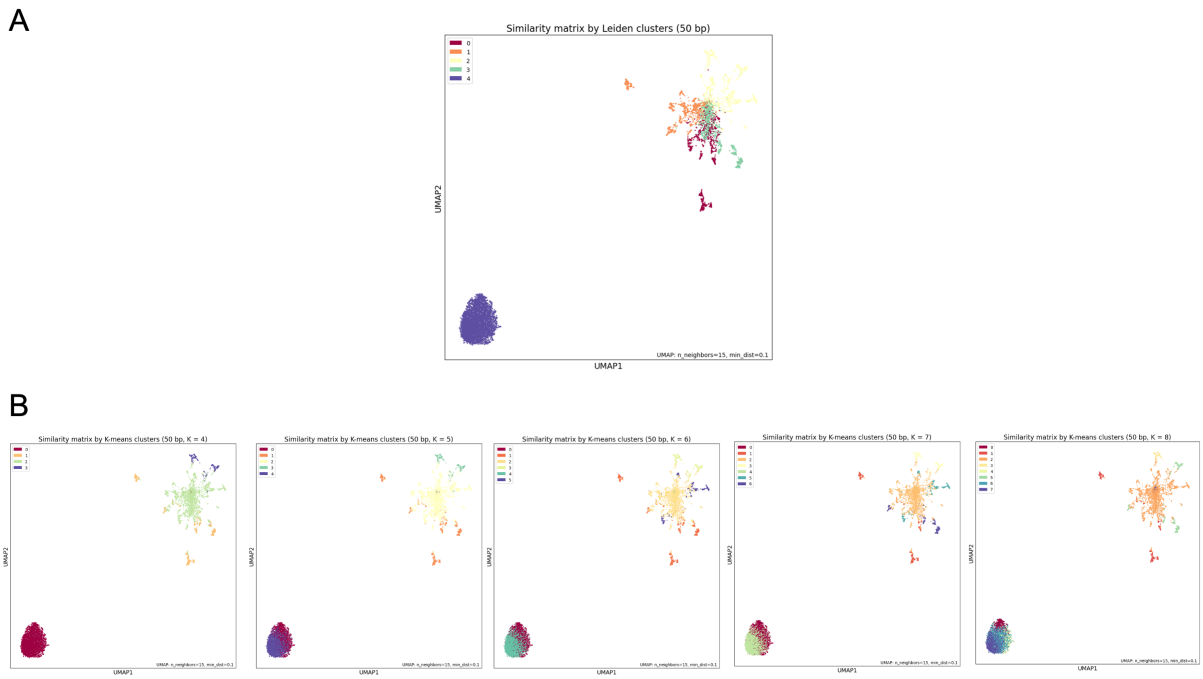


Figure 3.8. Leiden algorithm identifies five clusters. **(A)** Running the Leiden algorithm on G_L identifies five clusters. While it keeps background sequences in a single cluster, it splits foreground sequences in four different clusters. **(B)** K-means clustering identifies different sequence clusters and often splits even the background sequences cluster.

assess Leiden's results, we run k-means on the cosine similarity matrix using $k \in 4, 5, 6, 7, 8$. However, increasing k the algorithm split even the background sequences in different clusters, suggesting poor clustering performance **Fig.3.8 (B)**).

3.3 Future directions

In the next future we plan to construct the MotifGraph using the sequences prioritized by the Leiden algorithm in a cluster-wise fashion. , We expect that different cluster will return different motifs, corresponding to different motif grammars recovered from different genomic regions. That enforces our initial

aim of providing a comprehensive model learning different features in a single framework.

Predicting genetic variants impact on transcription factor binding sites

Many have highlighted the significant influence of genetic variations on TF-DNA binding events (De Gobbi *et al.*, 2006; Weinhold *et al.*, 2014; Guo *et al.*, 2018). Genome-wide association studies (GWASs) have revealed thousands of genetic variants, known as single nucleotide polymorphisms (SNPs), that are associated with complex human traits. Notably, these SNPs are typically situated within noncoding regions of the genome, many of which function as regulatory elements such as enhancers (Maurano *et al.*, 2012) Consequently, SNPs have the potential to influence gene expression by modulating TF-DNA interactions. These genetic variants can disrupt TF-DNA binding sequences, potentially leading to alterations in downstream gene expression patterns (Deplancke *et al.*, 2016). Crucially, mutations that affect transcription factor binding sites (TFBS) can persist within specific haplotypes found in populations (Kasowski *et al.*, 2010),, contributing to the emergence of population-specific TFBS. Similarly, genetic variability between cell types can give rise to cell-type-specific TF target sequences. Therefore, it becomes imperative to develop software tools capable of predicting the potential impact of genetic variations on TFBS specificity, while also considering haplotype and cell-type-specific mutations. In pursuit of this goal, we have introduced two tools designed for predicting the effects of mutations on TFBS within haplotypes and across different cell types. GRAFIMO (Tognon *et al.*, 2021) is a variant- and haplotype-aware motif scanning tool searching potential occurrences of known TF motifs on genome graphs (Paten *et al.*, 2017). Briefly, genome graphs are graph-based data structures, where nodes correspond to DNA sequences and edges describe allowed links between successive sequences. Paths through the graph, which may be labelled (such as in the case of a reference genome), correspond to haplotypes belonging to different genomes (Sirén *et al.*, 2020) (**Fig.5.9**). MotifRaptor (Yao *et al.*, 2021) investigates the imapct of genetic variants on TFBS interpolating different omics data, such as cell type-specific transcriptomic data and chromatin accessibility. In particular, MotifRaptor is designed to support researchers while annotating variants located within non-coding regions, and provides a potential functional annotation for these mutations. In our research we extend the current MotifRaptor to employ SVM-based motif models (**section 2.1.2**), which often provide better predictions on variants impact than PWM models (Tognon *et al.*, 2023).

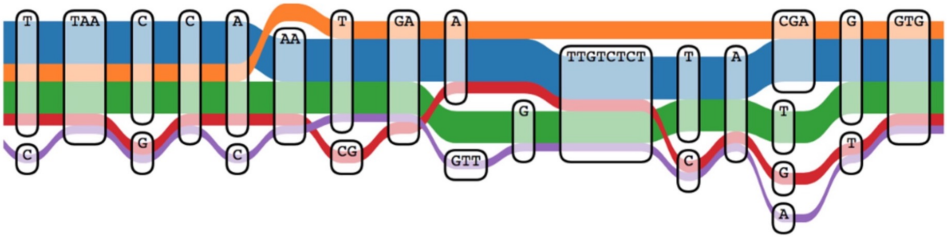


Figure 5.9. Genome Graphs data structure visualization. Each color corresponds to a path in the graph. Each path represents the genomic sequences of one of the individual genomes encoded in the Genome Graph structure.

5.1 GRAFIMO

In recent years, several tools have been proposed for scanning regulatory DNA regions, such as enhancers or promoters, with the goal of predicting which TF may bind these genomic locations. Importantly, it has been shown that regulatory motifs are under purifying selection (Li and Ovcharenko, 2015; Vorontsov *et al.*, 2016), and mutations occurring in these regions can lead to deleterious consequences on the transcriptional states of a cell (Guo *et al.*, 2018). In fact, mutations can weaken, disrupt or create new TFBS and therefore alter expression of nearby genes. Mutations altering TFBS can occur in haplotypes that are conserved within a population or private to even a single individual, and can correspond to different phenotypic behaviour (Kasowski *et al.*, 2010). For these reasons, population-level analysis of variability in TFBSs is of crucial importance to understand the effect of common or rare variants to gene regulation. Recently, a new class of methods and data structures based on genome graphs have enabled us to succinctly record and efficiently query thousands of genomes. Genome graphs optimally encode shared and individual haplotypes based on a population of individuals. An efficient and scalable implementation of this approach called variation graphs (VGs) has been recently proposed (Garrison *et al.*, 2018). VGs offer new opportunities to extend classic genome analyses originally designed for a single reference sequence to a panel of individuals. Moreover, by encoding individual haplotypes, VGs have been shown to be an effective framework to capture the potential effects of personal genetic variants on functional genomic regions profiled by ChIP-seq of histone marks (Groza *et al.*, 2020) During the last decade, several methods have been developed to search TFBS on linear reference genomes, such as FIMO (Grant *et al.*, 2011) and MOODS (Korhonen *et al.*, 2009) or to account for SNPs and short indels such as is-rSNP, TRAP and atSNP (Macintyre *et al.*, 2010; Thomas-Chollier *et al.*, 2011; Zuo *et al.*, 2015), however these tools do not account for individual haplotypes nor provide summary on the frequency of these events in a population. To solve these challenges, we have developed GRAFIMO (GRAph-based Finding of Individual Motif Occurrences) (Tognon *et al.*, 2021), a tool that offers a variation- and haplotype-aware identification of TFBS in VGs. Here, we show the utility of GRAFIMO by searching TFBS on a VG encoding the haplotypes from all the individuals sequenced by the 1000 Genomes Project (1000GP) (Siva, 2008; Zheng-Bradley *et al.*, 2017).

5.1.1 Desing and implementation

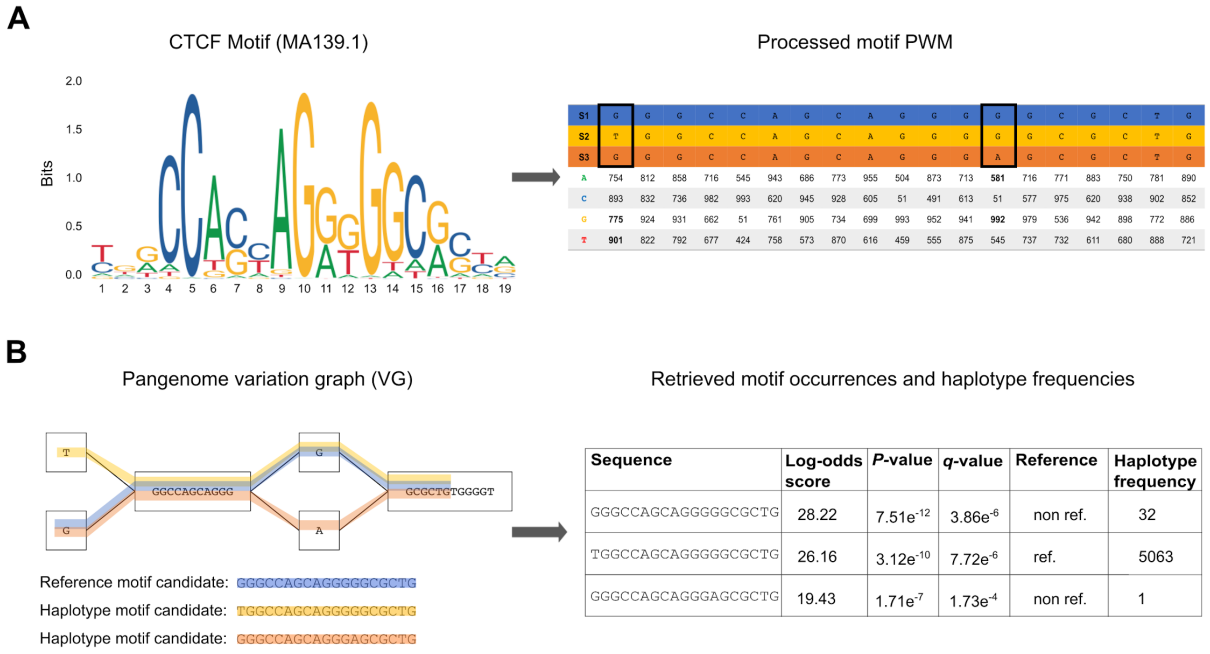
GRAFIMO is a command-line tool, which enables a variant- and haplotype- aware search of TFBS, within a population of individuals encoded in a VG. GRAFIMO offers two main functionalities: the construction of custom VGs, from user data, and the search of one or more TF motifs, in precomputed VGs. Briefly, given a TF model (PWM) and a set of genomic regions, GRAFIMO leverages the VG to efficiently scan and report all the TFBS candidates and their frequency in the different haplotypes in a single pass together with the predicted changes in binding affinity mediated by genetic variations. GRAFIMO is written in Python3 and Cython and it has been designed to easily interface with the *vg* software suite.

Genome variation graph construction

GRAFIMO provides a simple command-line interface to build custom genome variation graphs if necessary. Given a reference genome (FASTA format) and a set of genomic variants with respect to the reference (VCF format), GRAFIMO interfaces with the VG software suite to build the main VG data structure, the XG graph index (Garrison *et al.*, 2018) and the GBWT index (Novak *et al.*, 2017). used to track the haplotypes within the VG. To minimize the footprint of these files and speedup the computation, GRAFIMO constructs the genome variation graph by building a VG for each chromosome. This also speeds-up the search operation since we can scan different chromosomes in parallel. Alternatively, the search can be performed one chromosome at the time for machines with limited RAM.

Transcription factor motif search

The motif search operation takes as input a set of genomes encoded in a VG (.xg format), a database of known TF motifs and a set of genomic regions (BED format), and reports in output all the TFBS motifs occurrences in those regions and their estimated significance (**Fig.5.10**). To search for potential TFBS, GRAFIMO slides a window of length k (where k is the width of the query motif) along the paths of the VG corresponding to the genomic sequences encoded in it (**Fig.5.10(B)**). This is accomplished



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 5.10. 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.

by an extension to the vg find function, which uses the GBWT index of the graph to explore the k-mer space of the graph while accounting for the haplotypes embedded in it. By default, GRAFIMO considers only paths that correspond to observed haplotypes, however it is possible also to consider all possible recombinants even if they are not present in any individual. The significance (log-likelihood) of each potential binding site is calculated by considering the nucleotide preferences encoded in the PWM as in FIMO More precisely, the PWM is processed to a Position Specific Scoring Matrix (PSSM) (**Fig.5.10(A)**). and the resulting log-likelihood values are then scaled in the range [0, 1000] to efficiently calculate a statistical significance i.e. a P-value by dynamic programming as in FIMO (Grant *et al.*, 2011). *P*-values are then converted to q-values by using the Benjamini-Hochberg procedure to account for multiple hypothesis testing. For this procedure, we consider all the P-values corresponding to all the k-mer-paths extracted within the scanned regions on the VG. GRAFIMO computes also the number of haplotypes in which a significant motif is observed and if it is present in the reference genome and/or in alternative genomes (**Fig.5.10(B)**).

Report generation

We have designed the interface of GRAFIMO based on FIMO, so it can be used as in-drop replacement for tools built on top of FIMO. As in FIMO, the results are available in three files: a tab-delimited file (TSV), a HTML report and a GFF3 file compatible with the UCSC Genome Browser (Lee *et al.*, 2020). The TSV report contains for each candidate TFBS its score, genomic location (start, stop and strand), *P*-value, *q*-value, the number of haplotypes in which it is observed and a flag value to assess if it belongs to the reference or to the other genomes in VG. The HTML version of the TSV report can be viewed with any web browser. The GFF3 file can be loaded on the UCSC genome browser as a custom track, to visualize and explore the recovered TFBS with other annotations such as nearby genes, enhancers, promoters, or pathogenic variants from the ClinVar database (Landrum *et al.*, 2020).

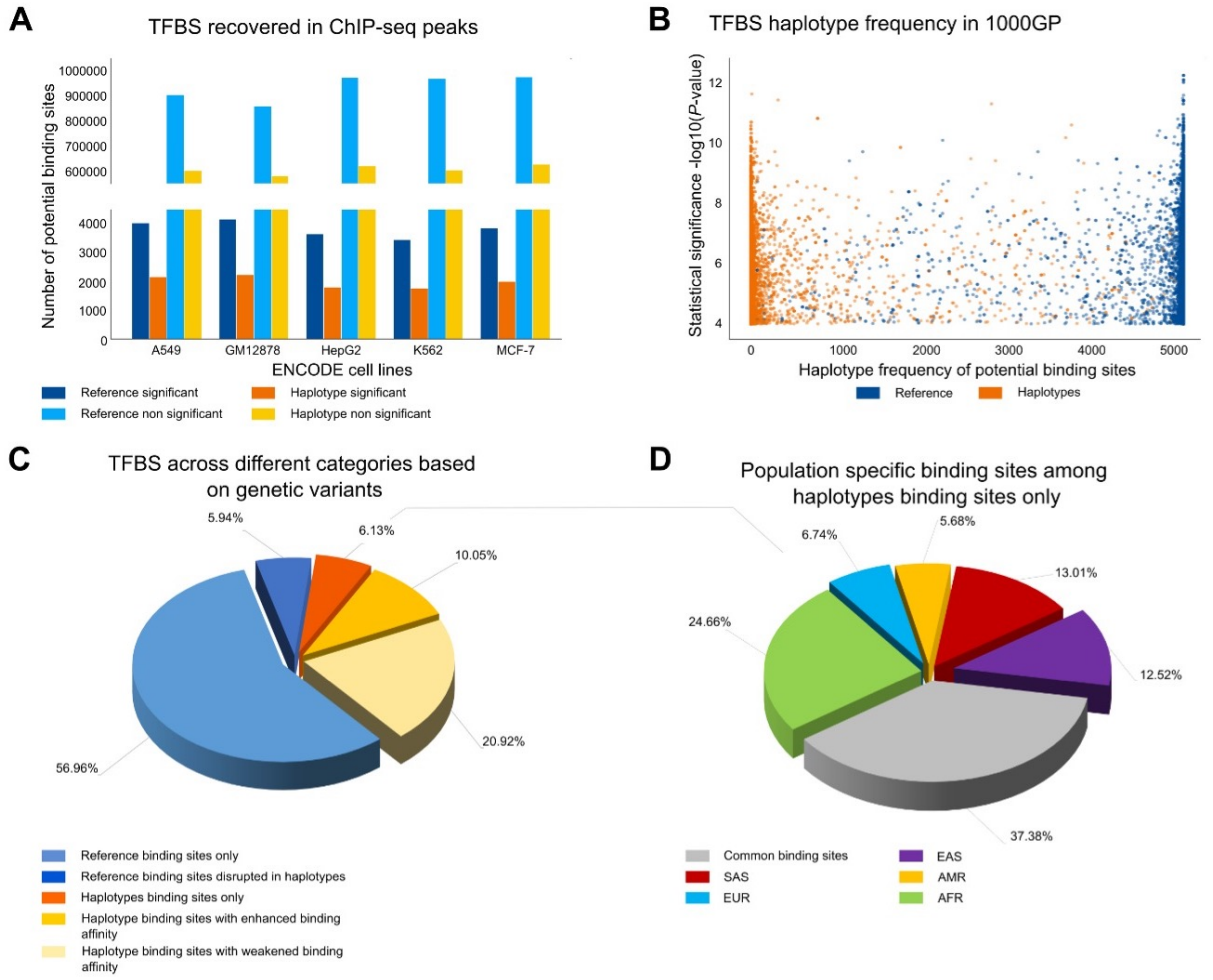
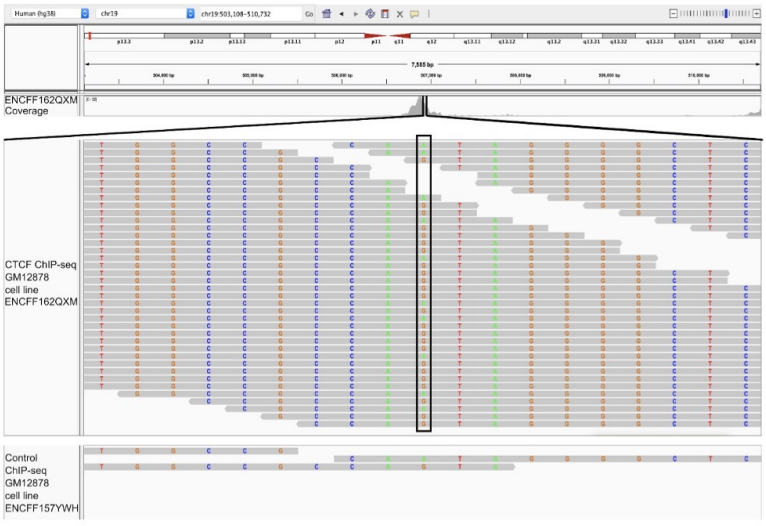


Figure 5.11. Searching CTCF motif on VG with GRAFIMO provides an insight on how genetic variation affects putative binding sites. **(A)** Potential CTCF occurrences statistically significant (P-value ¡ 1e-4) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh hg38 1000GP VG. (**B)** Statistical significance of retrieved potential CTCF motif occurrences and frequency of the corresponding haplotypes embedded in the VG. **(C)** Percentage of statistically significant CTCF potential binding sites found only in the reference genome or alternative haplotypes and with modulated binding scores based on 1000GP genetic variants **(D)** Percentage of population specific and common (shared by two or more populations) potential CTCF binding sites present on individual haplotypes.

5.1.2 Searching motif occurrences with GRAFIMO

GRAFIMO can be used to study how genetic variants may affect the binding affinity of potential TFBS within a set of individuals and may recover additional sites that are missed when considering only linear reference genomes without information about variants. To showcase its utility, we first constructed a VG based on 2548 individuals from the 1000GP phase 3 (hg38 human genome assembly) encoding their genomic variants and phased haplotypes. We then searched this VG for putative TFBS for three TF motifs with different lengths (from 11 to 19 bp), evolutionary conservation, and information content from the JASPAR database (Fornes *et al.*, 2020): CTCF (JASPAR ID MA0139.1), ATF3 (JASPAR ID MA0605.2) and GATA1 (JASPAR ID MA0035.4). To study regions with likely true binding events, for each factor we selected regions corresponding to peaks (top 3000 sorted by q-value) obtained by ChIP-seq experiments in 6 different cell types (A549, GM12878, H1, HepG2, K562, MCF-7) from the ENCODE project (Consortium *et al.*, 2012). We used GRAFIMO to scan these regions and selected for our downstream analyses only sites with a P-value ¡ 1e-4 and considered them as potential TFBS for these factors. Based on the recovered sites, we consistently observed across the 3 studied TFs that genetic variants can significantly affect estimated binding affinity. In fact, we found that thousands of CTCF motif occurrences are found only in non-reference haplotypes, suggesting that a considerable number of TFBS candidates are lost when scanning for TFBS the genome without accounting for genetic variants (**Fig.5.11(A)**). Similar results were obtained searching for ATF3 and GATA1. We also found several highly significant CTCF motif occurrences in rare haplotypes that may potentially modulate gene expression in these individuals (**Fig.5.11(B)**). Similar behaviours were observed for ATF3 and GATA1.



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 5.12. Considering genomic variation, GRAFIMO captures more potential binding events. GRAFIMO reports a potential CTCF binding site at chr19:506,910-506,929 found only in haplotype sequences, searching the motif in ChIP-seq peaks called on cell line GM12878 (experiment code ENCSR000DZN). The reads used to call for ChIP-seq peaks (ENCFF162QXM) show an allelic imbalance at position 10 of the motif sequence towards the alternative allele `G`, instead of the reference allele `A`. The imbalance is captured by GRAFIMO which reports the sequence presenting `G` at position 10 (found in the haplotypes), while the potential TFBS on the reference carrying an `A` is not reported as statistically significant (*P*-value ¿ 1e-4). CTCF motif logo shows that the `G` is the dominant nucleotide in position 10.

We also investigated the potential effects of the different length and type of mutations i.e. SNPs and indels on the CTCF, ATF3 and GATA1 binding sites. However, we did not observe a clear and general trend. By considering the genomic locations of the significant motif occurrences we next investigated how often individual TFBS may be disrupted, created or modulated. We observed that 6.13% of the potential CTCF binding sites can be found only on non-reference haplotype sequences, 5.94% are disrupted by variants in non-reference haplotypes and ∼30% are still significant in non-reference haplotypes but with different binding scores (**Fig.5.11(C)**). Similar results were observed for ATF3 and GATA1. Interestingly, we observed that a large fraction of putative binding sites recovered only on individual haplotypes are population specific. For CTCF we found that 24.66%, 6.74%, 5.68%, 13.01%, 12.52% of potential CTCF TFBS retrieved on individual haplotype sequences only are specific for AFR, EUR, AMR, SAS and EAS populations, respectively (**Fig.5.11(D)**).Among the unique CTCF motif occurrences found only on non-reference haplotypes in CTCF ChIP-seq peaks we uncovered one TFBS (chr19:506,910–506,929) that clearly illustrates the danger of only using reference genomes for motif scanning. Within this region we recovered a heterozygous SNP that overlaps (position 10 of the CTCF matrix) and significantly modulates the binding affinity of this TFBS. In fact, by inspecting the ChIP-seq reads (experiment ENCSR000DZN, GM12878 cell line), we observed a clear allelic imbalance towards the alternative allele G (70.59% of reads) with respect to the reference allele A (29.41% of reads). This allelic imbalance is not observed in the reads used as control (experiment code ENCSR000EYX) (**Fig.5.12**). Taken together these results highlight the importance of considering non-reference genomes when searching for potential TFBS or to characterize their potential activity in a population of individuals. We also compared the performance of GRAFIMO against FIMO. FIMO is faster and requires less memory, when scanning a single linear genome. However, when considering the 2548 individual genomes and their genetic variation, GRAFIMO proves to be generally faster than FIMO. Moreover, we benchmarked how GRAFIMO running time and memory usage change using an increasing number of threads. By increasing the number of threads, we observed a dramatical drop in running time, while memory usage remained similar.

5.2 MotifRaptor 2

Several studies have reported that genetic variants can enhance or disrupt TF-DNA binding affinity (De Gobbi *et al.*, 2006; Weinhold *et al.*, 2014; Wienert *et al.*, 2015). Genome-wide association studies (GWASs) have uncovered thousands of genetic variants (SNPs) associated with complex traits or human disease (Buniello *et al.*, 2019). Despite these efforts, functional studies to prioritize potential causal variants have lagged behind (Gallagher and Chen-Plotkin, 2018), resulting in a limited interpretation of the underlying pathophysiology mechanisms connecting variant to phenotype. A few missense SNPs can alter the function of a given TF by affecting its coding sequence, protein structure and therefore DNA binding capability, especially for Mendelian disease (Barrera *et al.*, 2016). For common diseases and complex traits, the great majority (>90%) of associated SNPs are in non-coding regions and mainly in DNase I hypersensitive sites. These SNPs correspond to functionally relevant non-coding regions such as enhancers and promoters (Maurano *et al.*, 2012). This observation suggests that chromatin state alterations and gene deregulation may be mediated by SNPs that modulate TF binding activities. In other words, genetic variants in these non-coding regions may perturb TF recognition sequences to enhance or disrupt TF-DNA binding events ultimately changing the downstream gene expression programs (Deplancke *et al.*, 2016). Even if single non-coding SNPs may only moderately alter binding sites and are underpowered to explain gene expression programs, statistics on a set of SNPs modulating common TF binding sites could be significant enough to reveal the convergent regulatory mechanism in complex traits. The method we present is based on this key idea.Despite the fact that several approaches have been proposed to explore how TF binding sites could be affected by genetic variants, challenges remain. The next paragraphs provide a short summary and the rationale behind the development of Motif-Raptor. First, current availability of ChIP-seq data unfortunately limit the utility of tools such as MMARGE (Link *et al.*, 2018), GERV (Zeng *et al.*, 2016b), DeepSEA (Zhou and Troyanskaya, 2015), Basset (Kelley *et al.*, 2016), IMPACT (Amariuta *et al.*, 2019), RegulomeDB (Boyle *et al.*, 2012), HaploReg4 (Ward and Kellis, 2012). In fact, these tools are extremely powerful and practical only when genome-wide maps of TF occupancy and/or chromatin marks in relevant cellular contexts are available. Therefore, in (Yao *et al.*, 2021), the authors found a unique value proposition in developing a framework to accommodate scenarios in which only PWM models and gene expression data are available. Second, available models based on ChIP-seq or PWM data do not systematically provide a global ranking and the significance of the TFs based on all trait-associated variants, rather a per SNP scoring. In fact, current methods based on PWM and/or DNase I-seq data, such as Combined Annotation Dependent Depletion (CADD, (Maurano *et al.*, 2015)), CENTIPEDE (Moyerbrailean *et al.*, 2016; Pique-Regi *et al.*, 2011), Affinity Testing for regulatory SNPs (atSNP, (Zuo *et al.*, 2015)). do not provide a procedure to formally test the global effect of a set of GWAS variants on the set of overlapping TF binding sites. To solve this limitation, MotifRaptor proposed a novel genome-wide statistic to prioritize putative causal TFs based on the entire set of binding sites and overlapping variants rather than single loci. Third, these methods do not consider linkage disequilibrium (LD) for the tagged loci by the GWAS-associated variants. This is important given that several non-causal SNPs have similar association scores as the true causal ones and that this potentially confound the analysis. In fact, these false positives can dilute our power of detecting the true mechanisms behind the causal variants. MotifRaptor's approach tries to account for this problem based on the two following strategies. By relying on cell type-specific chromatin accessibility regions, it is already reducing the space of variants in each LD block. To implicitly account for local LD structure, MotifRaptor samples the background set of chromatin accessibility regions in close proximity of the regions that are specific for each cell type. With these strategies, MotifRaptor mitigates the problem by specifically looking for variants within regions that are cell type specific. Among available tools only SLDP (Signed LD Profile) (Reshef *et al.*, 2018) overcomes this problem and offers a genome-wide significance score for each TF, based on the directional modulation of TF binding sites by SNPs. Another tool, GREGOR (Genomic Regulatory Elements and Gwas Overlap algoRithm) (Schmidt *et al.*, 2015), also explicitly accounts for LD structure to assess the enrichment on sentinel SNPs in arbitrary genomic regions (e.g. to prioritize cell types based on cell type specific annotations). MotifRaptor addresses the above limitations of current approaches, by providing a cell type-specific TF-centric analysis with associated statistics, comprehensive reporting and visualization functionalities. This tool can facilitate the discovery and interpretation of the action of non-coding variants on key regulators of complex traits.

5.2.1 Design and implementation

MotifRaptor pipeline consists of three main steps (**Fig.5.13**). (i) TCharacterize important cell-types through the enrichment of the phenotype associated SNPs in open chromatin regions. (ii) Search TFBS

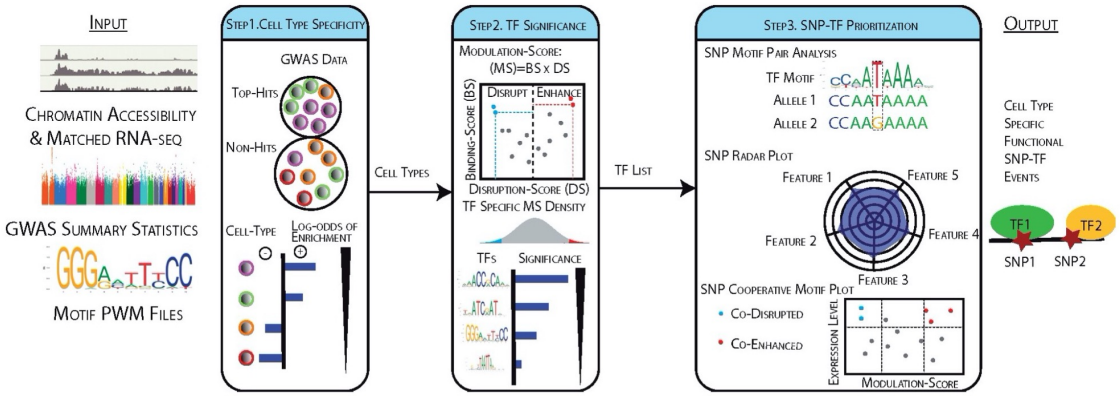


Figure 5.13. MotifRaptor analysis workflow. Three steps are performed: (1) characterize relevant cell types based on the enrichment of phenotype associated SNPs in chromatin accessible sites, (2) find TFs with binding sites that are significantly modulated by genetic variants in these cell types and (3) identify and visualize individual TF-SNP regulation events

whose binding potential is significantly modulated by the previously prioritized variants, in the investigated cell-types. (iii) Identify and visualize individual TF-SNP modulation events.

Quantifying the genetic variants effects on TF binding sites

To assess the impact of genetic variants at a certain TF binding site,, MotifRaptor implements an efficient genome-wide and threshold-free scoring procedure, scanning all the binding sites overlapping the target variants. Given a genomic sequence s, where $|s| = m$, and a position weight matrix (PWM), the score $M(i, S_i)$ indicates 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, $B(s_i)$:

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

from this formulation, MotifRaptor derives a disruption score DS, capturing the potential impact of genetic variants on a given binding site. Given a target variant, MotifRaptor assumes two haplotypes, denoted by s_{ref} and s_{alt}, for the reference and alternative alleles, respectively. MotifRaptor scores each allele to obtain the binding score for the reference and alternative alleles (i.e. $BS(s_{ref})$ and $BS(s_{alt})$). For scalability reasons it limits the computations 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}:

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

Therefore, the disruption score at K is defined as:

$$DS = \Delta BS$$

The value and sign of the DS become informative indicators on the directionality and strength of the target variant impact on TF binding. In fact, positive DS indicate enhanced binding affinities, while negative values a reduced binding potential. Since different TF binding site motifs have diverse length and specificities, MotifRaptor rescales both the BS and DS in $[0, 1]$ and $[-1, 1]$, respectively. Given the rescaling of the two scores, MotifRaptor defined a Binding-Disruption (BD) space to visualize and summarize the TF-SNP modulation events globally and across different factors.

Assessing TF-SNP modulations significance

To quantify the significance of the predicted TF-SNP modulation events prioritized, MotifRaptor employed a method based on the central limit theorem (CLT). Therefore, estimated a complete null-model

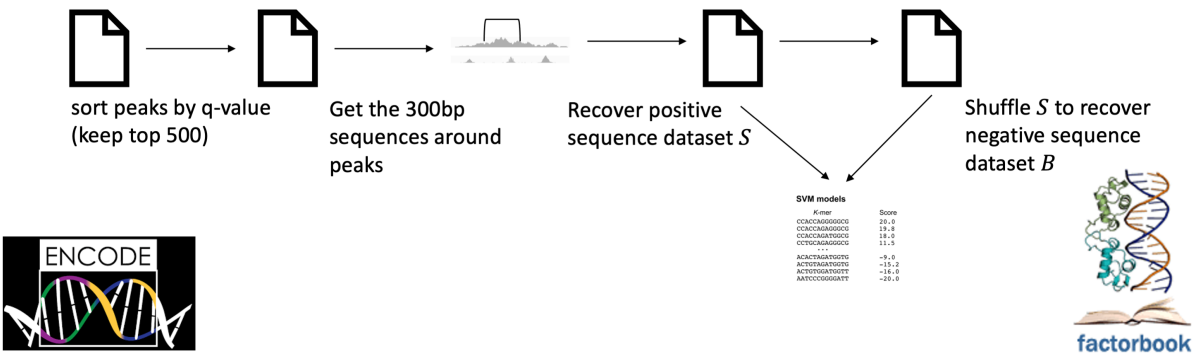


Figure 5.14. SVM-based motif models computation workflow. For the sake of reproducibility, we follow the model computation pipeline employed by FactorBook database, to derive PWM models from ENCODE data.

based on the enumeration of all potential binding sites on the genome. To maintain the scalability, MotifRaptor employed advanced sequence data structures like Suffix Arrays and Longest Common Prefix Arrays. These data structures allow the exhaustive enumeration keeping the method scalability. Given a set of target variants T, to assess if they significantly modulate TF binding events, MotifRaptor tests whether the B-S space significantly differed from the B-S space obtained from a set of background SNPs B. To assess the difference between the distributions MotifRaptor uses a non-parametric test with null hypothesis $E(D_{B-S_{space}}(T)) = E(D_{B-S_{space}}(B))$. Since based on the CLT, the distribution of the sample mean of B converges to a normal distribution, its mean and variance are equal to $E(D_{B-S_{space}}(B))$ and $\frac{\mathrm{Var}(D_{B-S_{space}}(B)}{N_{samples}}$, regardless of the underlying B-S space values distribution. Therefore, MotifRaptor tests the significance of both the enhanced binding ($(D_{B-S_{space}}(T)) > E(D_{B-S_{space}}(B))$), disrupted binding ($(D_{B-S_{space}}(T)) < E(D_{B-S_{space}}(B))$), or both. This procedure combined with efficient data structures allows to identify significant genome-wide shifts in TF-SNP binding modulations without using predetermined scores or cutoff on p-values.

5.2.2 Improving MotifRaptor by employing SVM-based motif models

In the case study presented in the original paper, MotifRaptor's predictions were supported by literature, suggesting that it leverages promising ideas. However, MotifRaptor 's main limitation is the employed computational models to represent TF binding motifs. Despite position weight matrices (Stormo, 2000) (PWMs) are simple, intuitive, and widely employed TF motif models, PWMs carry several drawbacks that often result in under- or over-estimation of non-coding variant impact on TFs binding landscape (Tognon *et al.*, 2023). TF SVM-based motif models (Tognon *et al.*, 2023; Boeva, 2016) have been demonstrated to overperform PWMs in several tasks, such as prediction of binding events and variants impact on TF binding sequences. However, while several publicly available resources provide complete and extensive collections of motif PWMs, to our knowledge there is no complete collection of SVM-based TF motifs. Therefore, we propose an extensive and curated collection of SVM-based motifs as well as a novel approach to annotate the functional impact of non-coding variants on TF binding landscape using SVM-based models, which interpolates different omics data, such as transcriptome, DNase-seq (John *et al.*, 2011), ATAC-seq (Buenrostro *et al.*, 2013), etc.

Building a library of transcription factor SVM-based models

SVM-based models have demonstrated superior performance compared to traditional PWMs (Tognon *et al.*, 2023). However, a significant drawback persists as no motif database currently offers these advanced SVM-based models. Moreover, the interpretability of these models remains a challenging issue, and there is a scarcity of motif analysis tools that harness the power of SVMs-based motif models. To address these limitations, we propose the creation of a comprehensive library comprising SVM-based motif models. To compute the SVM-based motif models we use ChIP-seq data from ENCODE (Consortium *et al.*, 2012). To enhance model reproducibility, we modified FactorBook's model computation pipeline (Pratt *et al.*, 2022) (**Fig.5.14**). To compute the models we selected ChIP-seq peaks datasets from ENCODE retrieved on different tissues and cell types. For each ChIP-seq dataset, we selected the top 500 peak sequences sorted by q-value. Then, we shrinked the surviving peaks to 300 bp centered around the peak signal center. After recovering the resulting genomic sequences, we constructed a background dataset by

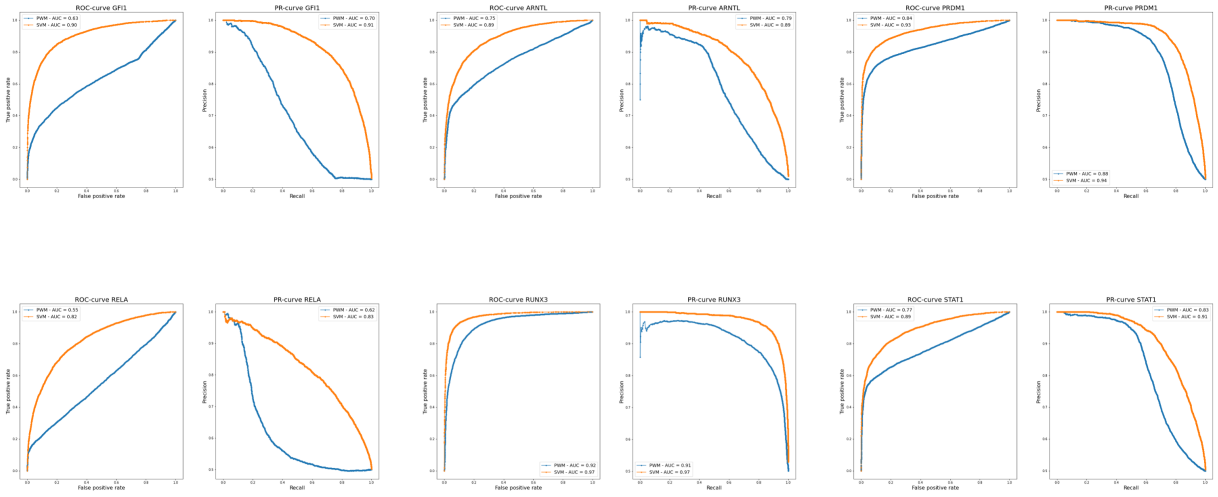


Figure 5.15. Comparing SVM-based and PWM motif models predictive power. We compared the predictive power of SVM-based and PWM motif models for six transcription factors: ARNTL, GFI1, PRDM1, RELA, RUNX3 and STAT1. Both models were trained on ENCODE ChIP-seq data (cell line GM12878). To train SVM-based models we followed the described pipeline, while to train PWM models we replaced LS-GKM with MEME. The predictive perfomance of the two models have been compared computing ROC- and PR-curves.

shuffling the original sequences. The two dataset are, then, used as input to compute the SVM models using LS-GKM (Lee, 2016).

Comparing SVM-based and PWM motif models predictive power

We demonstrate the improvements brought by SVM-based motif models over PWMs, by comparing the predictive power of the two models in terms of ROC- and PR-curves. For this analysis we selected the six TFs that have been previously identified in the original MotifRaptor publication to be affected by mutations linked to rheumatoid arthritis: ARNTL, GFI1, PRDM1, RELA, RUNX3, and STAT1. For the assessment we employed the original and the modified FactorBook pipeline, where the latter replaces MEME with LS-GKM to compute the SVM models. MEME (Bailey *et al.*, 1994) is a widely used software to compute PWM motif models from sequence data. To train and test the models we selected one ChIP-seq dataset for each TF (GM12878 cell line) from the ENCODE data portal. After training, we compared models' performance via a 4-fold cross-validation (75% training and 25% testing) on the top 500 peaks, selected following FactorBook's pipeline. By running our comparison, we observed that generally SVM-base motif models better predict binding events than PWMs, in both terms of ROC and PR-curves (**Fig.5.15**).

5.2.3 Future directions

In the next future we plan to complete our SVM-based motif models catalog extending the presented analysis to all the remaining TFs whose ChIP-seq data are available in ENCODE. Additionally, ENCODE datasets could be integrated employing datatasets from different databases, such as Cistrome (Zheng *et al.*, 2019) or GTRD (Kolmykov *et al.*, 2021). Moreover, we plan to release to the community the complete SVM motif models catalog. We also plan to train models on SELEX data rather than ChIP-seq and compare the resulting models. Importantly, since SELEX is a method less susceptible to noise than ChIP-seq the resulting models could potentially provide an even larger boost of models' predictive performance. We plan to modify MotifRaptor scanning algorithm for both PWM and SVM motif models to consider haplotypes, and indels. In fact, currently MotifRaptor consider only SNP and treats them as independent events, however it is known that complex SNP-SNP, SNP-indel, indel-indel combination can happen (Tognon *et al.*, 2021), often with signifcant effects on TF binding potential.

CRISPR genome editing

CRISPR gene editing (Cong *et al.*, 2013) enabled the genetic engineering of the genomes of living organisms. CRISPR provides a simple and programmable platform coupling the binding to genomic target sequences with diverse effector proteins restricted by protospacer adjacent motif (PAM) sequences. By delivering the Cas9 nuclease complexed with a synthetic guide RNA (gRNA) into a cell, CRISPR provides a simple and programmable platform to modify the genomic sequence at a desired location, potentially allowing the removal or addition of genes *in vivo*. Importantly, CRISPR offers unprecedented opportunities to develop novel therapies by introducing targeted genetic or epigenetic modifications to the genomic regions of interest. CRISPR-Cas9 offers high fidelity and simple construction and its specificity depends on two factors: (i) the target sequence and (ii) the PAM sequence. The target sequence is 20 bp long as part of each CRISPR locus in the gRNA array (Ran *et al.*, 2013). Typically, crRNA has multiple unique targets. Cas9 selects the genomic location by pairing the gRNA with its complementary sequence on the host DNA. Since the gRNA sequence is not part of the Cas9 complex, it can be designed independently to target specific genomic locations (Bialk *et al.*, 2015). To exploit the exonucleasic function Cas9 recognize its PAM sequences. PAMs are very short nonspecific sequences, occurring in several locations along the genome (Ran *et al.*, 2013). Once assembled the required sequences, Cas9 finds the targets on the genome, guided by the gRNA. The Cas9 nuclease opens both genomic strands of the target sequence to introduce novel modifications in it. Cas9 works in two main methods: (i) knock-in, and (ii) knock-out mutations (**Fig.6.16**). In knock-in, homology directed repair (HDR) employs DNA sequences similar to the targets to repair the breaks in the genome caused by Cas9 exonucleasic actions, using exogenous DNA as repairing template. Importantly, this method relies on periodic and isolated damaged spots in the target sites to start the DNA repair operations. In knock-out, mutations in the DNA inserted by Cas9 result in the repair of breaks through nonhomologous end joining (NHEJ). DNA repair via NHEJ often results in random insertions and deletions in the target sequence, which may disrupt, enhance, or alter the function of the target site. Since CRISPR-Cas9 enables a targeted random gene disruption, designing gRNA finely guiding Cas9 to the desired target sequence (*on-targets*) is fundamental to avoid unexpected and dangerous outcomes on undesired targeted sequences (*off-targets*). Most importantly, genetic variants may alter protospacer and PAM sequences and may influence both on-target and off-target potential. Therefore, it is fundamental to consider genetic variability when designing gRNA, to avoid potential undesired and dangerous outcomes on the host genome, in particular in clinical settings.

6.1 Benchmarking whole genome sequencing to detect CRISPR genome editing events

One of the major hurtles for the clinical adoption of CRISPR genome editing technologies is the risk of genome editing at unpredicted genomic sites (Fu *et al.*, 2013). Computational and experimental strategies have been developed to nominate off-target editing sites (Clement *et al.*, 2020).). In practice, these nomination strategies produce a ranked list of sites, of which the top hits are prioritized for validation using amplicon sequencing to quantify evidence of genome editing(Akcakaya *et al.*, 2018). However, nomination strategies have an arbitrary cutoff or no cutoff to distinguish which sites should be experimentally validated, and the number of sites chosen for validation are sometimes selected subjectively based on budget or choosing nice round numbers (e.g. the top 100). Unfortunately, the possibility of editing at nominated sites lower down the prioritized list-combined with the possibility that the nomination strategies may not comprehensively capture the genome editing dynamics of actual CRISPR genome editing–have led many to suggest that reliance on nomination strategies may be insufficient to exclude editing at unpredicted sites. Whole-genome sequencing (WGS) has been proposed as a standard to comprehensively examine the genome for evidence of genome editing. WGS has several benefits, including the ability to potentially

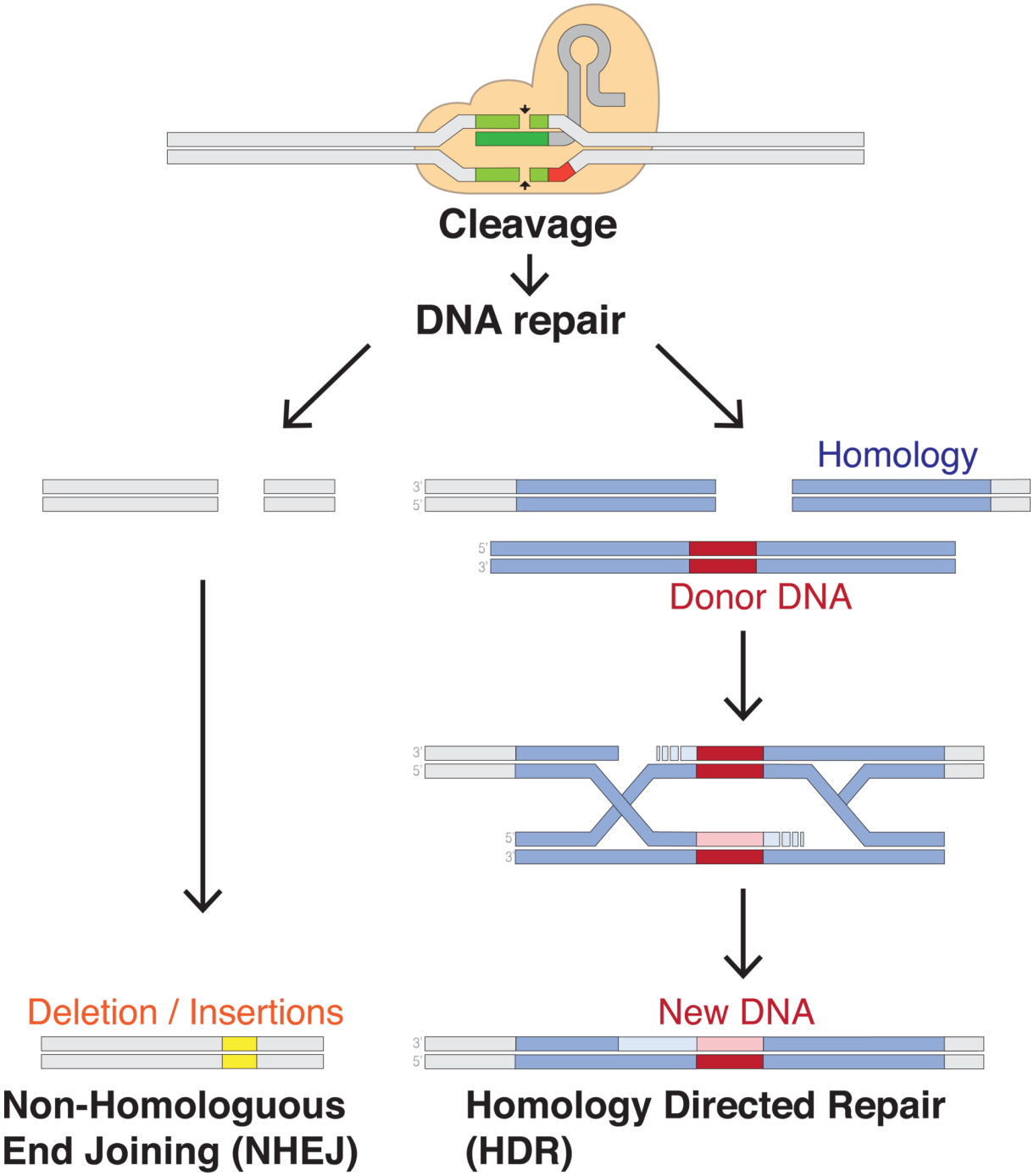


Figure 6.16. CRISPR gene editing via HDR and NHEJ.

measure genome editing activity at every base in the genome, and to measure complex genome editing outcomes such as translocations. However, despite increased availability and reduced cost of genome sequencing, WGS remains an expensive proposition as significant sequencing of the entire genome must be performed to detect rare editing events, and it is difficult to distinguish rare editing events from sequencing errors. WGS has been used to detect genome editing events in clonal cell lines (Smith *et al.*, 2014; Veres *et al.*, 2014) where individual cells within a clone serve as replicates to reduce bias of sequencing artifacts. However, this method is inadequate for use in detecting rare off-targets because the limit of detection is based on the number of clones (e.g. WGS of 100 clonal lines would need to be performed to detect an editing event at a 1/100 (1%) rate). To investigate the utility of WGS as a method to detect genome editing events genome-wide in bulk populations, we have created models for the sequencing depth required for detection of editing events. We have produced ultra-deep 1000x WGS datasets of genome-edited samples and used them to characterize editing at on- and off-target sites, and compare editing detection in WGS to several experimental nomination and computational detection methods. We

then propose an unbiased genome editing detection method which can identify the genome editing from WGS data in cases when the CRISPR guide sequence is not known. Together, these models, datasets, and methods provide an assessment of the utility for whole genome sequencing in the detection of genome editing.

6.1.1 Modeling sequencing depth requirements to detect genome editing

The actual problem of detecting genome editing can be broken into two parts: (i) sequencing genomic DNA with sufficient depth to detect editing at a given site, and (ii) correctly interpreting the sequencing data (e.g. read alignment and variant calling) to discriminate true CRISPR edits from technical noise. We address the first question regarding sequencing depth using a binomial model, and use simulated samples to measure the ability of existing mutation callers to detect edited genome-edited reads as such. DNA sequencing for the detection of editing events can be performed in a heterogenous population of edited cells with a detection limit and detection power proportional to the sequencing depth. This approach to determining sequencing depth requirements applies to amplicon sequencing of a single locus as well as to whole-genome sequencing. We note that WGS can be performed at low coverage on clonal lines to detect genome editing in single cells, but in this case the limit of detection is defined based on the number of single cell clones that can be sequenced. To detect genome editing – especially rare editing events – in cell populations, samples must be sequenced with sufficient depth, meaning that the sequence of sufficiently many DNA molecules must be read in order to find one containing an editing event. To measure rarer editing events, deeper sequencing must be performed. We used a statistical model to formalize the power to recover editing events at varying frequencies based on the binomial distribution (Petrackova *et al.*, 2019). Using this model we can calculate the sequencing depth required to observe a certain number of reads given a frequency of edited reads in a sample. For example, 30 reads are required to detect a single edited read in a sample with a mutation rate of 10% and 300 reads are required to detect a single edited read in a sample with a mutation rate of 1% at 95% power, or that in 95%. We can also use the binomial distribution confidence interval to calculate the range of true editing rates given an observed editing rate and sequencing depth. For example, the 95% confidence interval for an edit in 15/30 reads is between 31.3% and 68.7%. In order to assess the ability of mutation callers to detect genome editing events, we created simulated samples by inserting mutations found in reads from amplicon sequencing of 14 on- and off-targets from Fu *et al.* (Fu *et al.*, 2014) into whole-genome sequence from the Genome in a Bottle Consortium (Zook *et al.*, 2016). We plan to measure the ability of three commonly-used mutation callers to recover simulated reads resulting from genome editing (McKenna *et al.*, 2010; Kim *et al.*, 2018; Koboldt *et al.*, 2012).

6.1.2 Generation of WGS datasets

Whole-genome sequencing is commonly performed to identify mutations genome-wide, and has been applied in clinical settings to detect disease-associated germline mutations (Thiffault *et al.*, 2019). Clinical whole-genome sequencing aims for 30x-50x coverage of samples, and 40x coverage has been reported to provide sufficient depth to detect heterozygous and homozygous SNPs (Sun *et al.*, 2021). According to our model introduced above, sequencing at 30x can detect one edited read in the context of approximately 5% genome editing at 80%. In order to detect rarer editing events, deeper sequencing must be performed. We aimed to create whole genome sequencing libraries at 1000x, which should be able to detect one edited read with a 0.2% editing rate with 80% power. We performed genome editing in K562 and GM12878-Cas9 cell lines (Ma *et al.*, 2017) with four guides with varying specificities as measured by the number of sites in the genome with sequence homology with up to 4 mismatches. We selected RNF2 (precise guide, 7 off-by-4 sites), EMX1 (mid specificity, 293 off-by-4 sites), HEKSite4 (mid specificity, 832 off-by-4 sites), and VEGFASite3 (promiscuous, 6509 off-by-4 sites). Additionally we used a Cas-12a guide targeting DNMT1Site3 as a control guide. For each sample we prepared PCR-free sequencing libraries. Genomic sequencing depth was measured using Mosdepth (Pedersen and Quinlan, 2018), and showed a minimum of 89% genomic coverage at 1000x in GM12878 and at least 57% genomic coverage at 1000x in K562.

6.1.3 Measurement of editing using existing tools

We used our WGS samples to test the ability to identify edited bases by existing tools which are used for variant calling in non-CRISPR settings. Although these tools have been developed to disregard technical artifacts such as sequencing errors, and some have been developed to detect rare alleles, none have been developed to identify CRISPR-edited reads or genomic sites of CRISPR editing. 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), CIRCLE-seq (Tsai *et al.*, 2017)), and computational (Casoffinder (Bae *et al.*, 2014)) methods. For each nominated region, we defined a region within 100bp of the predicted cutting location where we would expect to see CRISPR editing. We also defined 10Kb windows up- and downstream from the nominated region (not including the 100bp window) where we did not expect to see editing. We used Mutect (McKenna *et al.*, 2010), Varscan (Koboldt *et al.*, 2012), and Strelka (Kim *et al.*, 2018) were used to identify variants for each nominated region, and classified variants within the 100bp window as true-positives and variants in the 10kb flanking windows as false-positives.

6.1.4 Enhancement of off-target editing detection by predictive models

Our WGS dataset provides the benefit of being able to investigate editing across the genome without knowing beforehand the sites to be investigated – albeit at a depth of around 1000x coverage. This is in contrast to strategies such as amplicon sequencing where much greater depth can be achieved but the genomic sites to be amplified must be selected in advance, and editing cannot be investigated outside the selected regions. We have capitalized on this characteristic of our WGS dataset and identify nuclease effects at a variety of regions. We used Casoffinder to identify all putative cut sites with up to 4 mismatches and 1 bulge with respect to the guide sequence, and used CRISPRessoWGS (Clement *et al.*, 2019) to calculate the percentage of reads with indels in the treated and control sample.

6.1.5 Unbiased detection of genome editing targets

Next, we sought to identify sites of genome editing, but without relying on knowing the sgRNA sequence used to edit the sample. This unbiased approach is valuable because it allows detection of CRISPR edits in a sample where the sgRNA was not known, and it can also be used to find CRISPR edits that may occur at locations without sequence similarity to the sgRNA. We note that many experimental nomination methods constrain reported hits to sites with homology to the sgRNA sequence because naturally occurring double-strand breaks frequently occur at fragile or AT-rich genomic sites (Nobles *et al.*, 2019),), so an unbiased method using WGS could identify guide sequence-independent editing events. Unfortunately, WGS data contains noise from different sources that makes it difficult to identify genome editing. These sources of noise include sequencing errors, alignment errors, and allelic diversity. We developed an unbiased genome editing discovery algorithm that attempts to discriminate between sites of CRISPR editing and noise by comparing a treated sample to a control sample because we expect that these sources of noise will affect the control sample as well as the treated sample. In addition, we exploit two additional characteristics of genome editing: High rates of insertions/deletions at the on-target and diversity of editing at the on-target – we expect that more than one allele will be produced as the result of stochastic double-strand break repair after Cas9 editing. We ran our unbiased algorithm to detect ranked sites of editing in all WGS samples. The ranking of sites with putative editing correlates well between the GM12878 and K562 cell lines, suggesting that the majority of editing is reproducible across cell types. We observe slightly more sensitivity in the K562 line for detecting off-target sites nominated by CIRCLE-seq. This is likely be due to the higher editing rates in the K562 cell line. Importantly, our unbiased algorithm identifies the on-target edit site among the top hits in all guides and in both cell lines.

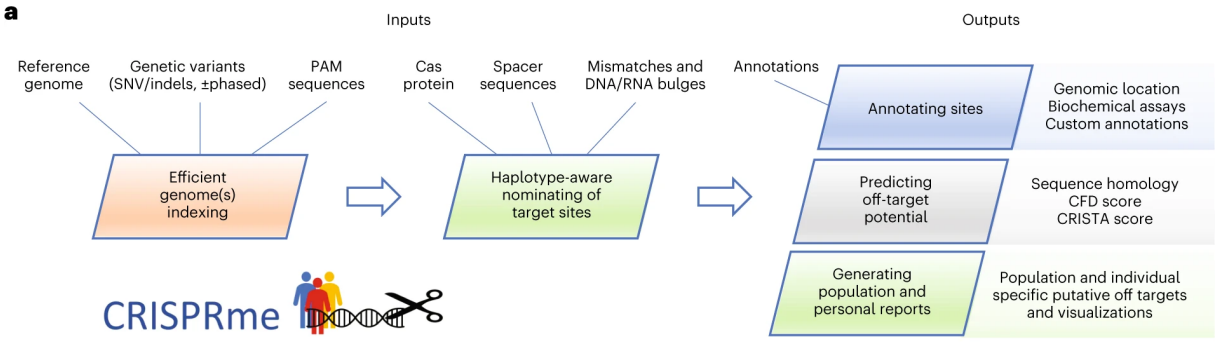
6.2 CRISPRme

CRISPR genome editing offers extraordinary opportunities to develop novel therapeutics by introducing targeted genetic or epigenetic modifications to genomic regions of interest. Briefly, CRISPR provides a simple and programmable platform that couples binding to a genomic target sequence of choice with diverse effector proteins through RNA:DNA (spacer:protospacer) complementary sequence interactions mediated by a gRNA spacer sequence matching a genomic protospacer sequence restricted by PAM sequences. Editing effectors may consist of nucleases to introduce targeted double strand breaks leading to short insertions/deletions (indels) and templated repairs (for example, Cas9), deaminases for precise substitutions (base editors) or chromatin regulators for transcriptional interference or activation (CRISPRi/a), among others, to achieve a range of desired biological outcomes (Anzalone *et al.*, 2020). CRISPR-based systems may create unintended off-target modifications posing potential genotoxicity for therapeutic use. Several experimental assays and computational methods are available to uncover or forecast these off-targets (Clement *et al.*, 2020). Off-target sites are partially predictable based on homology to the spacer and PAM sequence. Beyond the number of mismatches or bulges, a variety of sequence

features, like position of mismatch or bulge with respect to PAM or specific base changes, contribute to off-target potential (Clement *et al.*, 2020; Bao *et al.*, 2021; Hsu *et al.*, 2013; Doench *et al.*, 2016). Computational models can complement experimental approaches to off-target nomination in several respects: to triage gRNAs before experiments by predicting the number and cleavage potential of off-target sites and to prioritize target sites for experimental scrutiny. Genetic variants may alter protospacer and PAM sequences and therefore may influence both on-target and off-target potential. Gene editing strategies designed to specifically recognize patient mutations may increase the likelihood of editing mutant alleles, whereas variants that reduce homology to the anticipated target may decrease the efficiency of the desired genetic modification. Although a variety of in vitro and cell-based experimental methods can be used to empirically nominate off-target sites, these methods either use homology to the reference genome as a criterion to define the search space and/or use a limited set of human donor genomes to evaluate off-target potential (Bao *et al.*, 2021; Chaudhari *et al.*, 2020). Therefore, computational methods may be especially useful to predict the impact of off-target sequences not found in reference genomes. Prior studies considering gRNAs targeting therapeutically relevant genes using population-based variant databases like the 1000 Genomes Project (1000 G) and the Exome Aggregation Consortium have highlighted how genetic variants can substantially alter the off-target landscape by creating novel and personal off-target sites not present in a single reference genome (Lessard *et al.*, 2017; Scott and Zhang, 2017). Although these prior studies provide code to reproduce analyses, implementation choices make these tools not suitable to analyze large variant datasets and to consider higher numbers of mismatches. In addition, these methods ignore bulges between RNA:DNA hybrids, cannot efficiently model alternative haplotypes and indels, and require extensive computational skills to utilize. Several user-friendly websites have been developed to aid the design of gRNAs and to assess their potential off-targets (Concordet and Haeussler, 2018; Listgarten *et al.*, 2018; Labun *et al.*, 2019; Park *et al.*, 2015). Even though variant-aware prediction is an important problem for genome editing interventions, these scalable graphical user interface (GUI) based tools do not account for genetic variants. In addition, these tools artificially limit the number of mismatches for the search and/or do not support DNA/RNA bulges. Therefore, designing gRNAs for therapeutic intervention using current widely available tools could miss important off-target sites that may lead to unwanted genotoxicity. A complete and exhaustive off-target search with an arbitrary number of mismatches, bulges, and genetic variants that is haplotype-aware is a computationally challenging problem that requires specialized and efficient data structures.We have recently developed a command line tool that partially solves these challenges called CRISPRitz (Cancellieri *et al.*, 2020). This tool uses optimized data structures to efficiently account for single variants, mismatches and bulges but with substantial limitations Here, we substantially extend this work by developing CRISPRme, a tool to aid gRNA design with added support for haplotype-aware off-target enumeration, short indel variants and a flexible number of mismatches and bulges (Cancellieri *et al.*, 2023). CRISPRme is a unified, user-friendly web-based application that provides several reports to prioritize putative off-targets based on their risk in a population or individuals. CRISPRme is flexible to accept user-defined genomic annotations, which could include empirically identified off-target sites or cell-type-specific chromatin features. The tool can integrate population genetic variants from sets of phased individual variants (like those from 1000 G (Lowy-Gallego *et al.*, 2019)), unphased individual variants (like those from the Human Genome Diversity Project (HGDP) (Bergström *et al.*, 2020)) and population-level variants (like those from the Genome Aggregation Database (gnomAD) (Karczewski *et al.*, 2020)). Furthermore, it can accept personal genomes from individual subjects to identify and prioritize private off-targets due to variants specific to a single individual. Here, we demonstrate the utility of CRISPRme by analyzing the off-target potential of a gRNA currently being tested in clinical trials for sickle cell disease (SCD) and β-thalassemia (Frangoul *et al.*, 2021; Canver *et al.*, 2015; Wu *et al.*, 2019). We identify possible off-targets introduced by genetic variants included within and extending beyond 1000 G. We predict that the most likely off-target site, overlooked by prior analyses, is introduced by a variant common in African-ancestry individuals (rs114518452, minor allele frequency (MAF) = 4.5%) and provide experimental evidence of its off-target potential in gene edited human CD34+ HSPCs. Furthermore, we demonstrate that allele-specific off-target potential is widespread across various nucleic acid targeting therapeutic strategies.

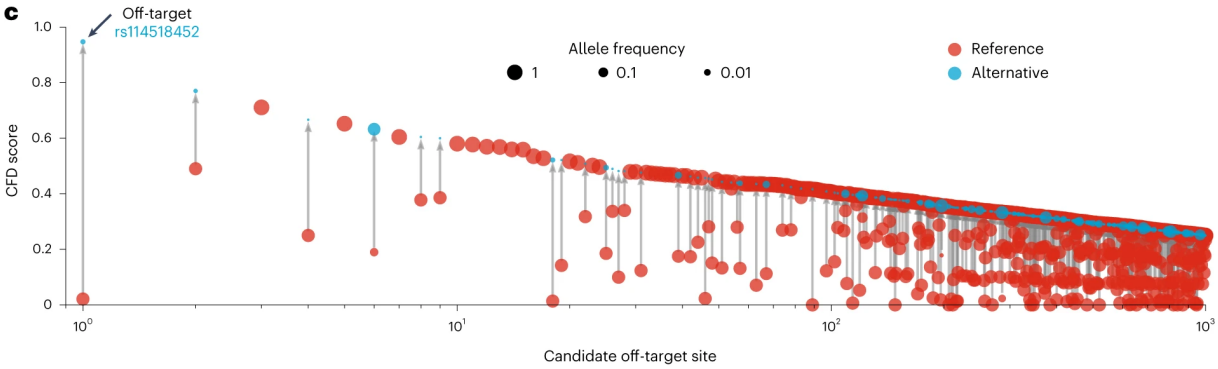
6.2.1 A computational tool for variant-aware off-target nomination

CRISPRme is a web-based tool to predict off-target potential of CRISPR gene editing that accounts for genetic variation. It is available online at `http://crisprme.di.univr.it`. CRISPRme can also be deployed to local, protected and isolated environments as a web app or command line utility, neither of which transfer or store data online, therefore respecting genomic privacy and regulations. CRISPRme



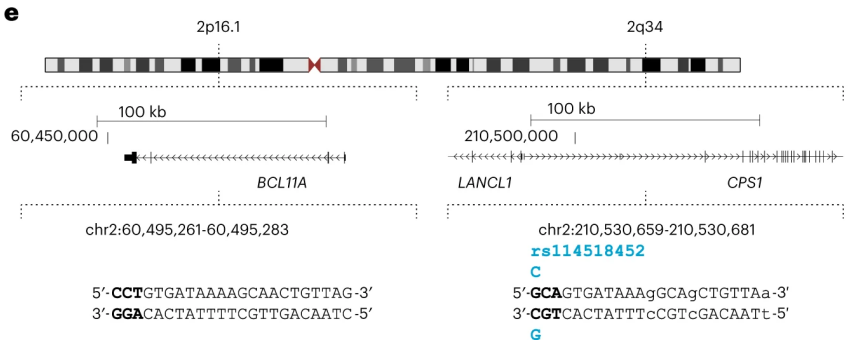
b

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



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	



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 6.17. CRISPRme provides web-based analysis of CRISPR-Cas gene editing off-target potential reflecting population genetic diversity. **(A)** CRISPRme software takes as input a reference genome, genetic variants, PAM sequence, Cas protein type, spacer sequence, homology threshold and genomic annotations and provides comprehensive, target-focused and individual-focused analyses of off-target potential. It is available as an online webtool and can be deployed locally or used offline as command-line software. **(B)** Analysis of the BCL11A-1617 spacer targeting the +58 erythroid enhancer with SpCas9, `NNN` PAM, 1000G variants, up to 6 mismatches and up to 2 bulges. **(C)** Top 1000 predicted off-target sites ranked by CFD score, indicating the CFD score of the reference and alternative allele if applicable, with allele frequency indicated by circle size. **(D)** The off-target site with the highest CFD score is created by the minor allele of rs114518452. Coordinates are for hg38 and 0-start for the potential off-target and 1-start for the variant-ID. MAF is based on 1000G. **(E)** The top predicted off-target site from CRISPRme is an allele-specific off-target with 3 mismatches to the BCL11A-1617 spacer sequence, where the rs114518452-C minor allele produces a de novo NGG PAM sequence. PAM sequence shown in bold and mismatches to BCL11A-1617 shown as lowercase. Coordinates are for hg38 and 1-start. **(F)** rs114518452 allele frequencies based on gnomAD v3.1. Coordinates are for hg38 and 1-start. Spacer shown as DNA sequence for ease of visual alignment.

takes as input a Cas protein, gRNA spacer sequence and PAM, genome build, sets of variants (VCF files for populations or individuals), user-defined thresholds of mismatches and bulges and optional user-defined genomic annotations to produce comprehensive and personalized reports (**Fig.6.17(A)**). We have designed CRISPRme to be flexible with support for new gene editors with variable and extremely relaxed PAM requirements (Walton *et al.*, 2020). Thanks to a PAM encoding based on Aho-Corasick automata and an index based on a ternary search tree, CRISPRme can perform genome-wide exhaustive searches efficiently even with an NNN PAM, extensive mismatches (tested with up to seven) and RNA:DNA bulges (tested with up to two). Notably, a comprehensive search performed with up to six mismatches, two DNA/RNA bulges and a fully nonrestrictive PAM (NNN) on a small computational cluster node using 20 CPUs and 128 GB RAM (Intel Xeon CPU E5-2609 v4 clocked at 2.2 GHz) takes ∼34 h of real time and ∼152 h of CPU time (including both user and system times). All the 1000 G variants, including both single-nucleotide variants and indels, can be included in the search together with all the available metadata for each individual (sex, superpopulation and age), and the search operation takes into account observed haplotypes. Importantly, off-target sites that represent alternative alignments to a given genomic region are merged to avoid inflating the number of reported sites. Although several tools exist to enumerate off-targets, to our knowledge, only two command line tools (Lessard *et al.*, 2017; Fennell *et al.*, 2021) incorporate genetic variants in the search. However, they have several limitations in terms of scalability to large searches, number of mismatches, bulges, haplotypes and variant file formats supported and do not provide an easy-to-use GUI. CRISPRme generates several reports. (i) it summarizes for each gRNA all the potential off-targets found in the reference or variant genomes based on their mismatches and bulges (**Fig.6.17(B)**) and generates a file with detailed information on each of these candidate off-targets. (ii) it compares gRNAs to customizable annotations. By default, it classifies possible off-target sites based on GENCODE (Frankish *et al.*, 2019) (genomic features) and ENCODE (Consortium *et al.*, 2012) (candidate cis-regulatory elements (cCREs)) annotations. It can also incorporate user-defined annotations in BED format, such as empiric off-target scores or cell-type-specific chromatin features (**Fig.6.18**). (iii) using 1000 G and/or HGDP variants, CRISPRme reports the cumulative distribution of homologous sites based on the reference genome or superpopulation. These global reports could be used to compare a set of gRNAs based on how genetic variation impacts their predicted on- and off-target cleavage potential using cutting frequency determination (CFD) or CRISPR Target Assessment (CRISTA) scores (Abadi *et al.*, 2017) (**Fig.6.19**). CRISPRme includes multiple scoring metrics and can be easily extended with new ones, including scores tailored for different editors. Finally, CRISPRme can generate personal genome focused reports called personal risk cards. These reports highlight private off-target sites due to unique genetic variants.

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

We tested CRISPRme with a gRNA (#1617) targeting a GATA1 binding motif at the +58 erythroid enhancer of BCL11A (Canver *et al.*, 2015; Wu *et al.*, 2019). A recent clinical report described two patients, one with SCD and one with β-thalassemia, each treated with autologous gene modified HSPCs edited with Cas9 and this gRNA, who showed sustained increases in fetal hemoglobin, transfusion independence and absence of vaso-occlusive episodes (in the patient with SCD) following therapy. This study, as well as prior preclinical studies with the same gRNA (#1617), did not reveal evidence of off-target editing in treated cells when considering off-target sites nominated by bioinformatic analysis of the human reference genome and empiric analysis of in vitro genomic cleavage potential (Frangoul *et al.*, 2021; Wu *et al.*, 2019; Demirci *et al.*, 2019). CRISPRme analysis found that the predicted off-target site with both the greatest CFD score and the greatest increase in CFD score from the reference to alternative allele was at an intronic sequence of CPS1 (**Fig.6.17(C)** and **(D)**), a genomic target subject to common genetic variation (modified by a SNP with MAF $\geq$ 1%). CFD scores range from 0 to 1, where the on-target site has a score of 1. The alternative allele rs114518452-C generates a TGG PAM sequence (that is, the optimal PAM for SpCas9) for a potential off-target site with three mismatches and a CFD score (CFD_{alt} 0.95) approaching that of the on-target site (**Fig.6.17(E)**). In contrast, the reference allele rs114518452-G disrupts the PAM to TGC, which markedly reduces predicted cleavage potential (CFD_{ref} 0.02). rs114518452-C has an overall MAF of 1.33% in gnomAD v3.1, with an MAF of 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 other populations represented in gnomAD (**Fig.6.17(F)**). To consider the off-target potential that could be introduced by personal genetic variation that would not be predicted by 1000 G variants, we analyzed HGDP variants identified from whole-genome sequences of 929 individuals from 54 diverse human populations. We observed 249 candidate off-targets

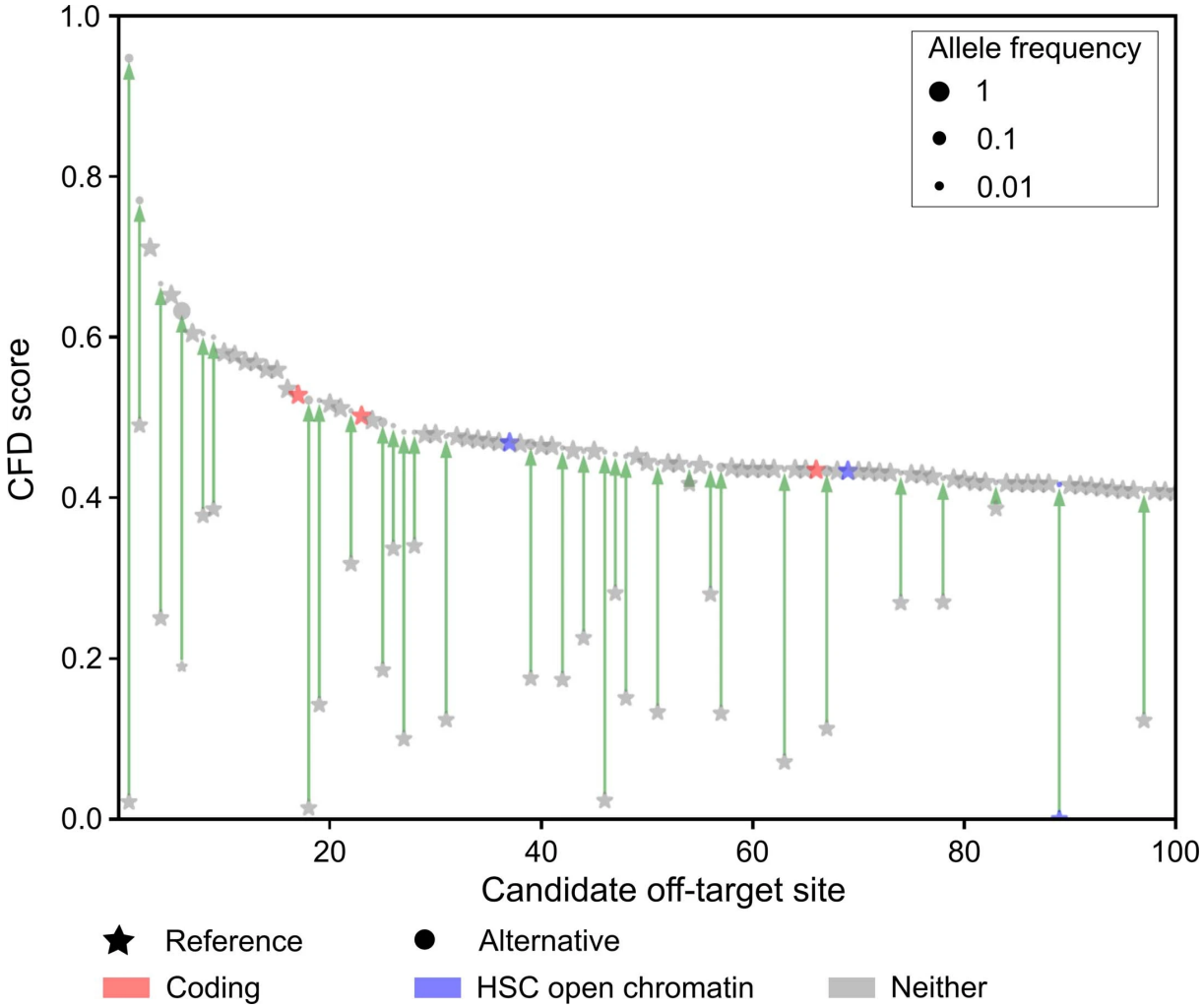


Figure 6.18. Top 100 predicted off-target sites for BCL11A-1617 spacer by CFD score. CRISPRme search as in **Fig.6.17**. Candidate off-target sites within coding regions based on GENCODE annotations and ATAC-seq peaks in HSCs based on user-provided annotations (data from (Corces *et al.*, 2016)) are highlighted.

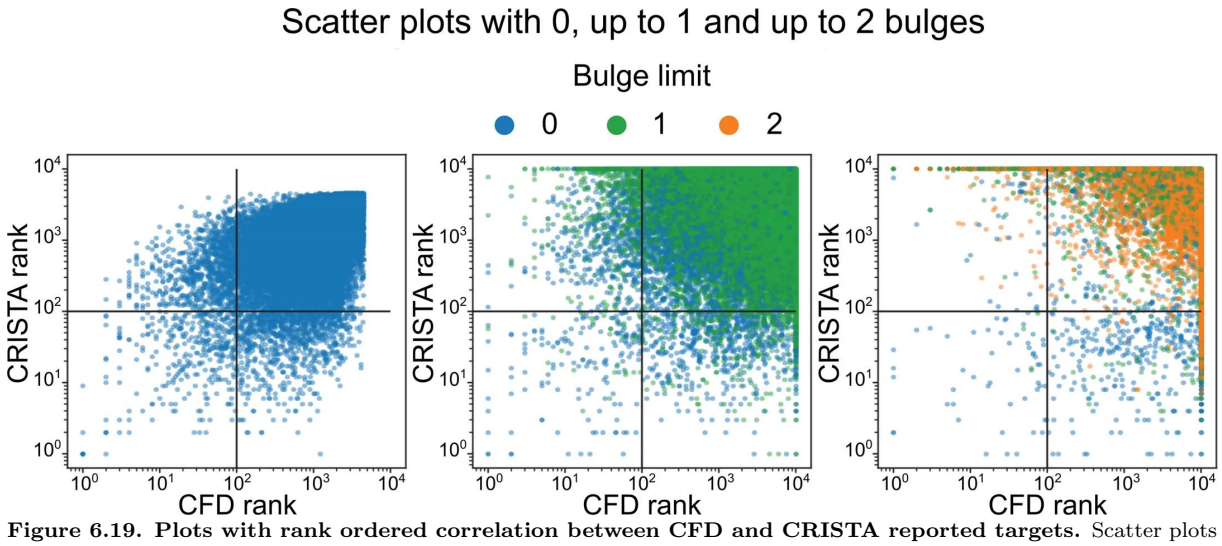
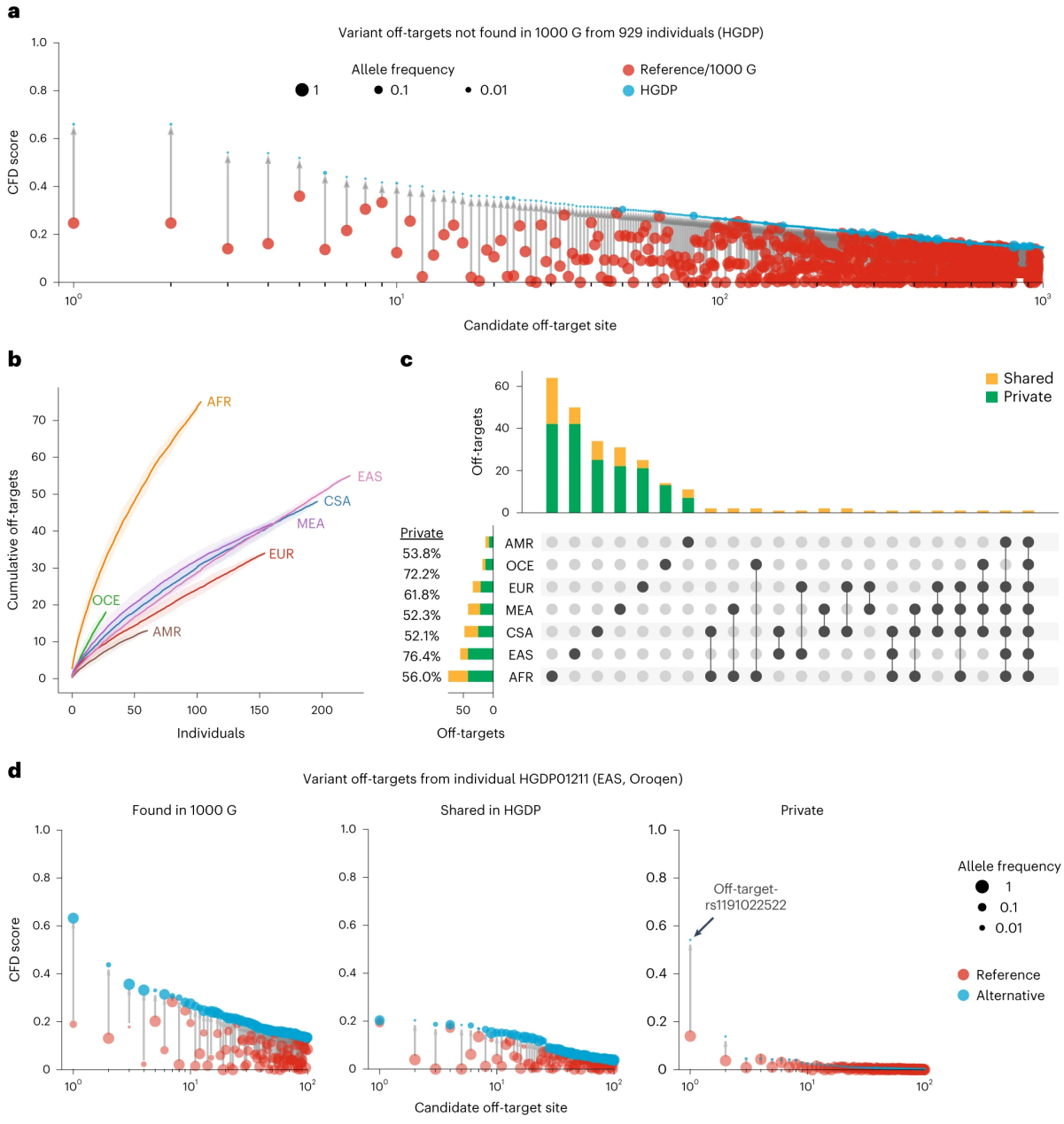


Figure 6.19. Plots with rank ordered correlation between CFD and CRISTA reported targets. Scatter plots show from left to right, the correlation of ranked targets, extracted by selecting top 10,000 targets sorted by CFD and CRISTA score, respectively. The left plot shows the rank correlation of targets with 0 bulges (Pearson's correlation: 0.57, $p < 1e^{-10}$, Spearman's correlation: 0.55, $p < 1e^{-10}$), the center plot shows the rank correlation of targets with 1 bulge (Pearson's correlation: -0.16, $p < 1e^{-10}$, Spearman's correlation: -0.33, $p < 1e^{-10}$) and the right plot shows the rank correlation of targets with 2 bulges (Pearson's correlation: -0.55, $p < 1e^{-10}$, Spearman's correlation: -0.80, $p < 1e^{-10}$). The correlation values and p-values (two-sided) were calculated using standard functions from the Python scipy library. The colors represent the lowest count of bulges for each target, because the two scoring methods may prioritize different alignments and thus different number of mismatches and bulges pf the same target.



e

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

Figure 6.20. CRISPRme provides analysis of off-target potential of CRISPR-Cas gene editing reflecting population and private genetic diversity. (A) CRISPRme analysis was conducted with variants from HGDP comprising whole-genome sequencing of 929 individuals from 54 diverse human populations. HGDP variant off-targets with greater CFD scores than the reference genome or 1000 G were plotted and sorted by CFD score, with HGDP variant off-targets shown in blue and reference or 1000 G variant off-targets shown in red. **(B)** Cumulative distribution plot of HGDP variant off-targets with CFD $\geq$ 0.2 and increase in CFD of $\geq$ 0.1 per superpopulation (AFR: Africa, AMR: Americas, CSA: Central & South Asia, EAS: East Asia, EUR: Europe, MEA: Middle East, OCE: Oceania). Individual samples from each of the seven superpopulations were shuffled 100 times to calculate the mean and 95% confidence interval (shading around lines). **(C)** Intersection analysis of HGDP variant off-targets with CFD $\geq$ 0.2 and increase in CFD of $\geq$ 0.1. Shared variants (black) were found in two or more HGDP samples whereas private variants (gray) were limited 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 1000G variant off-targets (left panel), higher CFD score compared to reference genome and 1000 G but shared with other HGDP individuals (center panel) and higher CFD score compared to reference genome and 1000 G with variant not found in other HGDP individuals (right panel). For the center and right panels, reference refers to CFD score from reference genome or 1000 G variants. **(E)** The top predicted private off-target site from HGDP01211 is an allele-specific off-target where the rs1191022522-G minor allele produces a canonical NGG PAM sequence in place of a noncanonical NAG PAM sequence. Spacer shown as DNA sequence for ease of visual alignment.

for gRNA #1617 with CFD $\geq$ 0.2 for which the CFD score in HGDP exceeded that found for either the reference genome or 1000 G variants by at least 0.1 (**Fig.6.20(A)** and **Fig.6.21**). hese additional variant off-targets not found from 1000 G were observed in each superpopulation, with the greatest frequency in the African superpopulation (**Fig.6.20(B)**); 229 (92.0%) of these variant off-targets were unique to a superpopulation, and 172 (69.1%) of these were private to just one individual (**Fig.6.20(C)**). Furthermore, single-individual-focused searches (for example, an analysis of HGDP01211, an individual of the Oroqen population within the East Asian superpopulation) showed that most variant off-targets (with higher CFD score than reference) were due to variants also found in 1000 G (n = 32369, 90.4%), a subset were due to variants shared with other individuals from HGDP but absent from 1000 G (n = 3177, 8.9%) and a small fraction were private to the individual (n = 234, 0.7%) (**Fig.6.20(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 (**Fig.6.20(D)** and **(E)**). o experimentally test the top predicted off-target from CRISPRme, we identified a CD34+ HSPC donor of African ancestry heterozygous for rs114518452-C, the variant predicted to introduce the greatest increase in off-target cleavage potential (**Fig.6.17(C)-(F)**). We performed ribonucleoprotein (RNP) electroporation using a gene editing protocol that preserves engrafting HSC function. Amplicon sequencing analysis showed 92.0 $\pm$ 0.5% indels at the on-target site and 4.8 $\pm$ 0.5% indels at the off-target site. For reads spanning the variant position, indels were strictly found at the alternative PAM-creation allele without indels observed at the reference allele (**Fig.6.22(A)-(C)**), suggesting 9.6 $\pm$ 1.0% off-target editing of the alternative allele. In an additional six HSPC donors homozygous for the reference allele rs114518452-G/G, 0.00 $\pm$ 0.00% indels were observed at the off-target site, suggesting strict restriction of off-target editing to the alternative allele (**Fig.6.22(D)**). The on-target BCL11A intronic enhancer site is on chr2p, whereas the off-target-rs114518452 site is on chr2q within an intron of a noncanonical transcript of CPS1. Inversion PCR demonstrated inversion junctions consistent with the presence of $\sim$150 Mb pericentric inversions between BCL11A and the off-target site only in edited HSPCs carrying the alternative allele (**Fig.6.23(A)** and **(B)**). Deep sequencing of the inversion junction showed that inversions were restricted to the alternative allele in the heterozygous cells (**Fig.6.23(C)** and **(D)**). Droplet digital PCR revealed these inversions to be present at 0.16 $\pm$ 0.04% allele frequency (**Fig.6.23(E)**). Various high-fidelity Cas9 variants may improve the specificity of gene editing, although at the possible cost of reduced efficiency (Schmid-Burgk *et al.*, 2020). Gene editing following the same electroporation protocol using a HiFi variant 3xNLS-SpCas9 (R691A) (Vakulskas *et al.*, 2018) in heterozygous cells revealed 82.3 $\pm$ 1.6% on-target indels with only 0.1 $\pm$ 0.1% indels at the rs114518452-C off-target site (that is, a $\sim$48-fold reduction compared to SpCas9) (**Fig.6.22(C)**). Inversions were not detected following HiFi-3xNLS-SpCAs9 editing (**Fig.6.23(B)** and **(E)**).

6.2.3 Allele specific off-target potential of additional gRNAs

To examine the pervasiveness of alternative allele off-target potential, we evaluated an additional 13 gRNAs in clinical development or otherwise widely used for SpCas9-based nuclease or base 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) and 6 gRNAs for non-SpCas9-based editing such as for SaCas9 and Cas12a (Xu *et al.*, 2019b; Chu *et al.*, 2021; Newby *et al.*, 2021; Maeder *et al.*, 2019; De Dreuzy *et al.*, 2019). CRISPRme analysis including the 1000 G and HGDP genetic variant datasets showed 18% (95% confidence interval 13-23%) of the total nominated off-targets were due to alternative allele-specific off-targets. Most alternative allele-specific off-targets were associated with rare variants (MAF < 1%), although candidate off-targets associated with common variants were identified for each gRNA (**Fig.6.24(A)**). None of these alternative allele-specific off-target sites were described in the original manuscripts reporting the editing strategies and off-target analyses. CRISPRme produces visualizations to specifically highlight alternative allele-specific candidate off-target sites overlapping cCREs and protein coding sequences (including putative tumor suppressor genes (Zhao *et al.*, 2016)) and/or that involve PAM creation events (**Fig.6.24(B)** and **(C)**). For example, within the top 20 candidate off-targets nominated by CRISPRme for a SpCas9 gRNA targeting EMX1 (Tsai *et al.*, 2015), two sites involve genetic variants with high MAF (52% and 26%) and are associated with substantial increases in CFD score from REF to ALT (+0.69 and +0.44). The first is an intronic PAM creation variant, whereas the second introduces two PAM-proximal matches to the gRNA (**Fig.6.24(D)**). Notably, both of these candidate off-targets involve indel variants, underscoring the utility of CRISPRme to account for variants beyond SNPs. In addition to visualizing candidate off-target sites by predictive score rank (such as CFD or CRISTA) for SpCas9-derived editors, CRISPRme can also visualize candidate

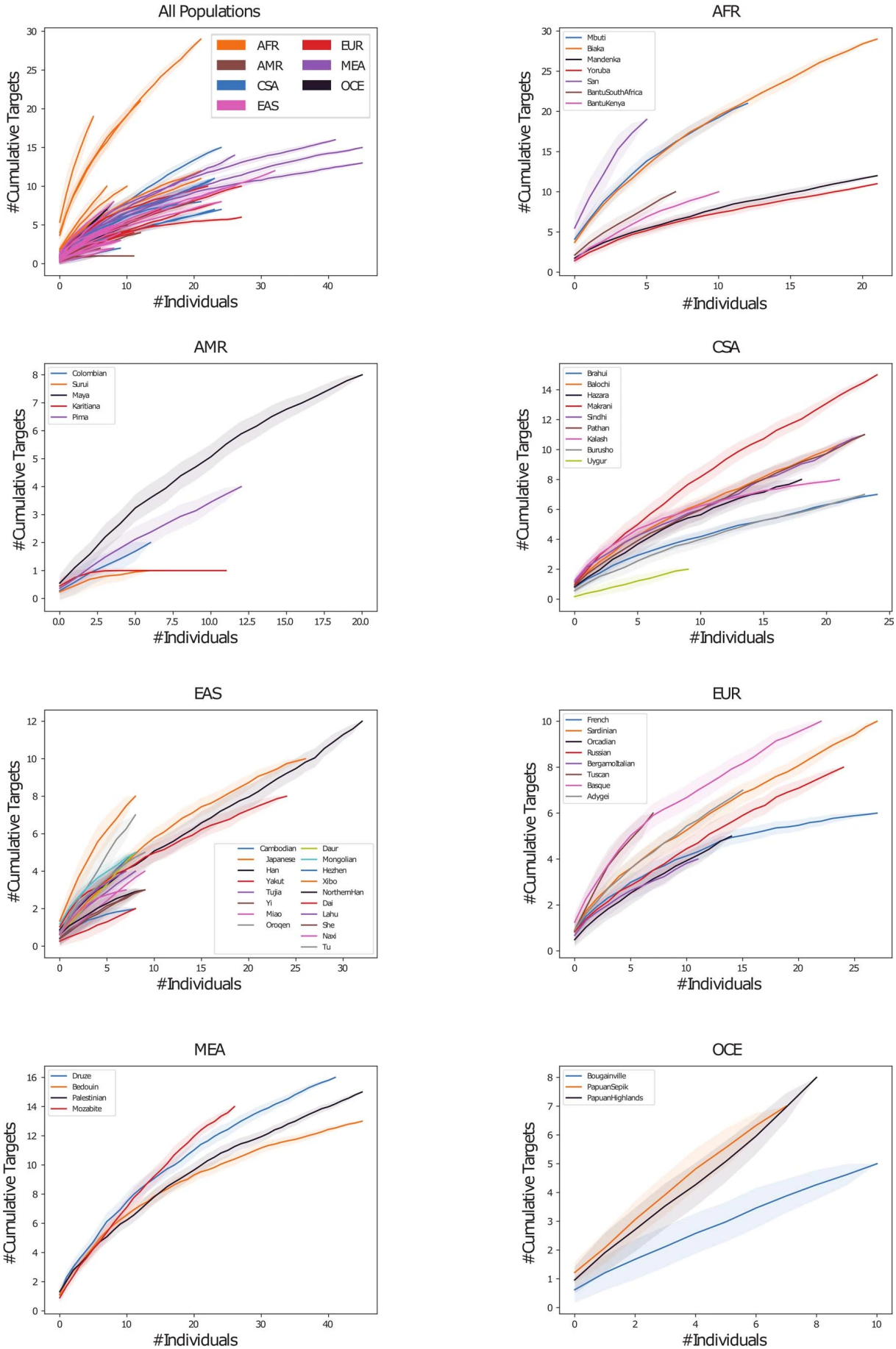


Figure 6.21. HGDP superpopulation distribution plots. HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD of ≥ 0.1. Individual samples from each of the seven superpopulations were shuffled 100 times to calculate the mean and 95% confidence interval. First panel shows distribution within all 54 discrete populations, colored by superpopulation. Additional seven panels show distribution of discrete populations within each listed superpopulation.

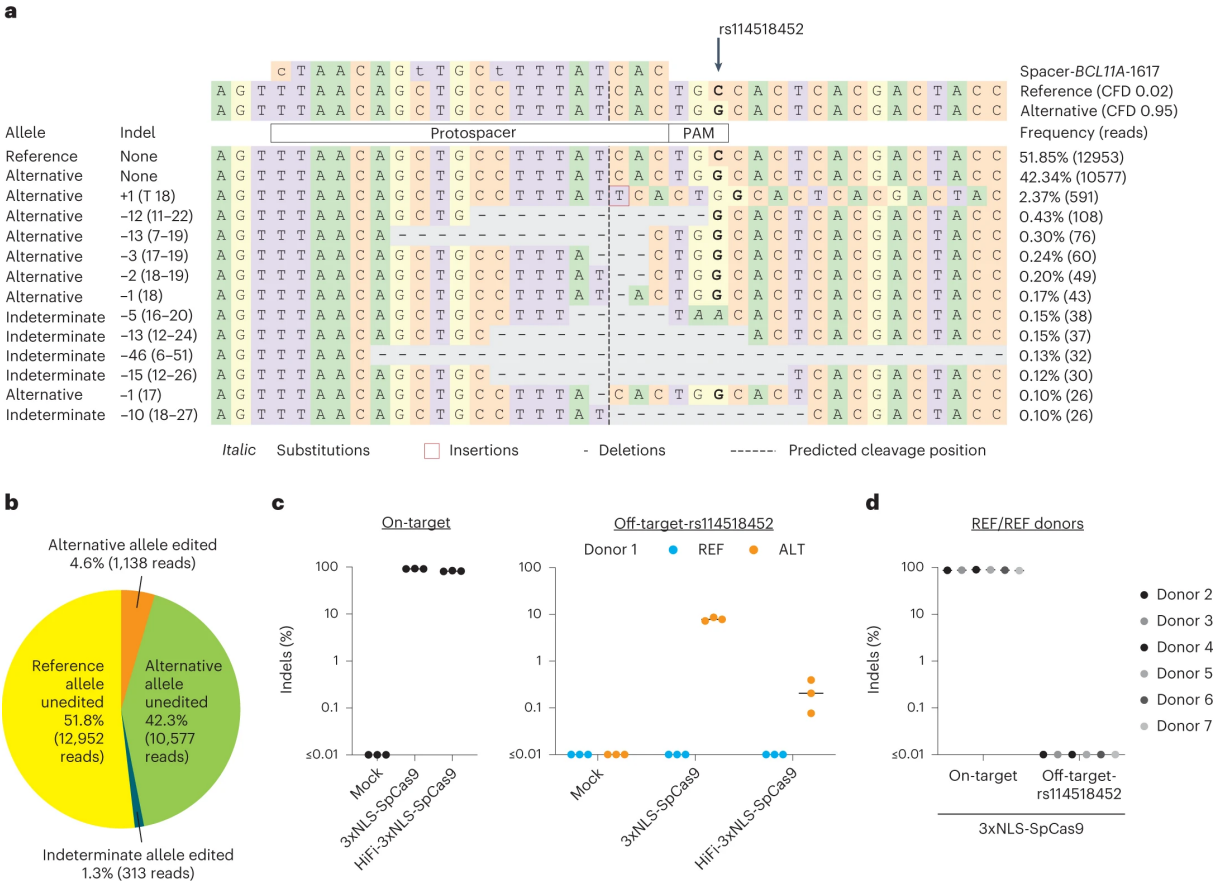


Figure 6.22. Allele-specific off-target editing by a BCL11A enhancer targeting gRNA in clinical trials associated with a common variant in African-ancestry populations. (A) Human CD34+ HSPCs from a donor heterozygous for rs114518452-G/C (donor 1, REF/ALT) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation (NLS: nuclear localization signal) followed by amplicon sequencing of the off-target site around chr2:210,530,659-210,530,681 (off-target-rs114518452 in 1-start hg38 coordinates). CFD scores for the reference and alternative alleles are indicated, and representative aligned reads are shown. Spacer shown as DNA sequence for ease of visual alignment, with mismatches indicated by lowercase and the rs114518452 position shown in bold. **(B)** Reads classified based on allele (indeterminate if the 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 independentbiological replicate (n = 3), and lines represent medians. 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.

off-targets by number of mismatches and bulges, which may be especially useful for Cas proteins with distinct PAMs for which predictive scores are not readily available. For example, SaCas9 is a clinically relevant nuclease whose small size favors packaging to adeno-associated virus. For a SaCas9-associated gRNA targeting CEP290 (Maeder *et al.*, 2019) currently being evaluated in clinical trials to treat a form of congenital blindness (NCT03872479), CRISPRme nominated two candidate off-targets associated with common SNPs (MAF 7% and 5%) that reduced mismatches from five (REF) to four (ALT) that are predicted to produce cleavages within coding sequences (**Fig.6.24(D)**). CRISPRme can nominate variant off-targets for base editors and evaluate their base editing susceptibility within a user-defined editing window. For a gRNA targeting PCSK9 (ref. 37) that has been used with SpCas9-nickase adenine base editor in vivo in preclinical studies to reduce low-density lipoprotein cholesterol levels, four of the top five candidate off-target sites involve alternative alleles, including one with CFD_{ref} 0.2 and CFD_{alt} 0.75 found in an ENCODE candidate enhancer element. CRISPRme nominated a candidate off-target associated with a rare variant (MAF 0.0007%) that increased the CFD score from 0.06 (REF) to 0.40 (ALT) that would be predicted to produce missense mutations in EPHB3, a putative tumor suppressor gene (**Fig.6.24(D)**). The underlying computational challenge that CRISPRme addresses extends beyond CRISPR-based applications to other technologies based on nucleic acid sequence recognition. For example, CRISPRme can nominate off-targets for RNA-targeting strategies, whether RNA-guided gene editors or even oligonucleotide sequences used as RNA interference or antisense oligo therapies (**Fig.6.25**). We

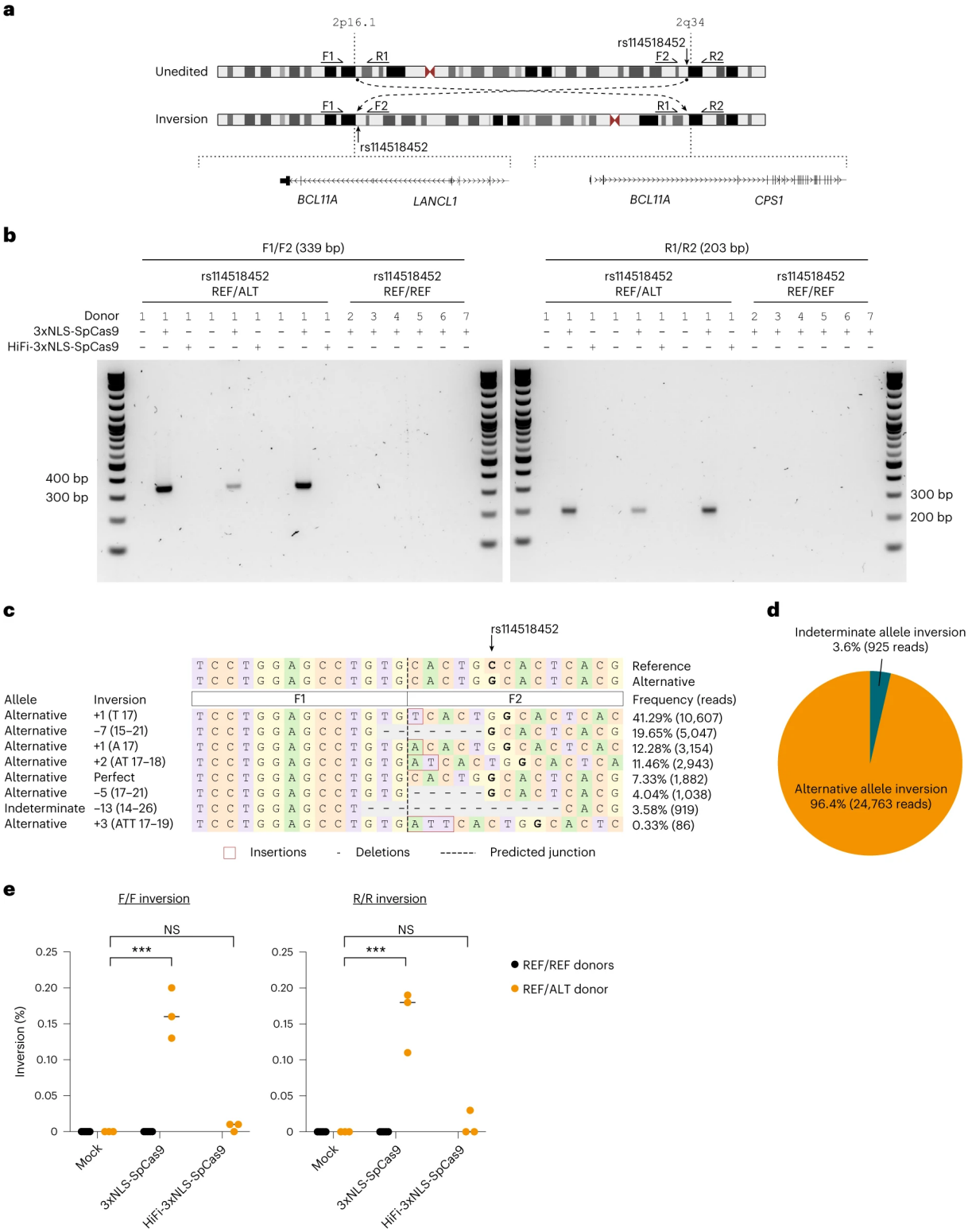


Figure 6.23. Allele-specific pericentric inversion following BCL11A enhancer editing due to off-target cleavage. (A) Concurrent cleavage of the on-target and off-target-rs114518452 sites could lead to pericentric inversion of chr2 as depicted. PCR primers F1, R1, F2 and R2 were designed to detect potential 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. The rs114518452 position is shown in bold. **(D)** Reads classified based on allele (indeterminate if the 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.

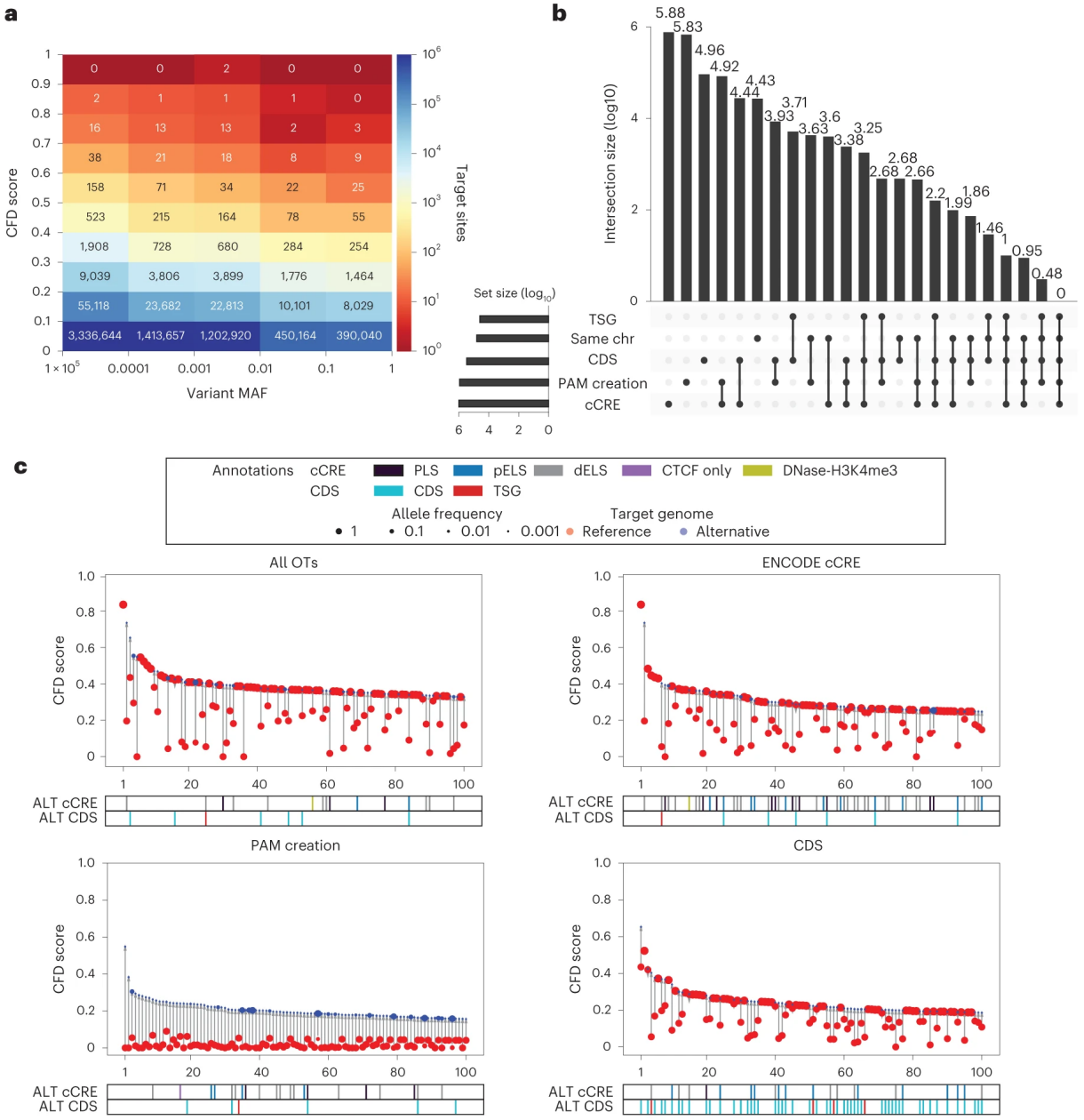


Figure 6.24. CRISPRme illustrates prevalent off-target potential due to genetic variation (A) Heatmap showing the distribution of alternative allele nominated off-targets for SpCas9 guides by CFD score and MAF. **(B)** UpSet plot showing overlapping annotation categories for candidate off-targets (tumor 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 the 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.

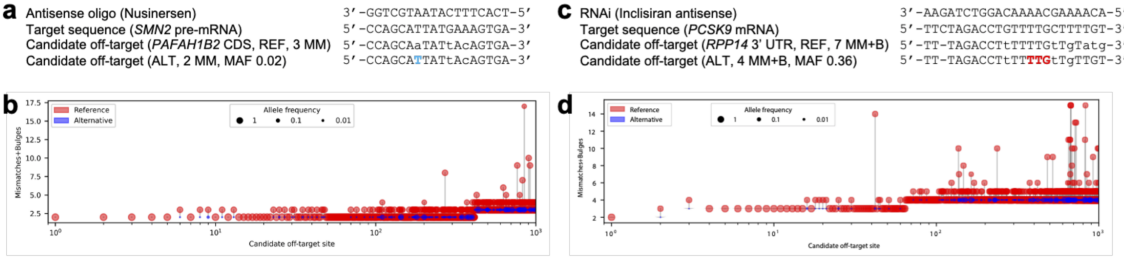


Figure 6.25. Candidate transcript off-targets introduced by common genetic variants for non-CRISPR sequence-based RNA-targeting therapeutic strategies. (A) A common SNP (in blue) introduces a candidate CDS off-target site with 2 mismatches for the FDA-approved antisense oligo Nusinersen. **(B)** Top 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.

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 treat spinal muscular atrophy. Using CRISPRme, we identified a potential off-target site within a coding region wherein 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 treat hypercholesterolemia, revealed that its antisense strand has a candidate off-target in the 3' untranslated region of the ribosomal gene RPP14 for which a common insertion variant (MAF 36%) reduces the number of mismatches and bulges from seven (REF) to four (ALT).

6.2.4 Limitations and Discussion

These results demonstrate how personal genetic variation may influence the off-target potential of sequence-based therapies like genome editing. Increased availability of haplotype-resolved genomes of diverse ancestry would enhance ability to nominate variant-associated off-target sites present in human populations. A limitation of current tools including CRISPRme is that potential off-targets cannot be enumerated based on structural variants or other complex genetic events such as combinations of indels and SNPs (Cancellieri *et al.*, 2023). Future extensions of CRISPRme based on new data structures such as graph genomes (Paten *et al.*, 2017; Garrison *et al.*, 2018) could enable these complex searches and improve their efficiency. The practical implications of allele-specific off-target editing need to be considered on a case-by-case basis. In the case of BCL11A enhancer editing, up to ∼10% of SCD patients with African ancestry would be expected to carry at least one rs114518452-C allele, leading to ∼10% cleavage at an off-target site that was not identified in prior studies of this gRNA using currently available tools. Our results highlight that allele-specific off-target editing potential is not equally distributed across all ancestral groups but is especially concentrated in those of African ancestry where genomic variation is most pronounced. Therefore, gene editing efforts that include subjects of African ancestry (like those targeting SCD) might pay particular attention to this issue. Gene editing efforts that focus on a specific patient population should consider genetic variants enriched in that population during off-target evaluation. However, our analysis also shows that variant off-targets may be private to a given individual, so all humans could potentially be susceptible to such an effect. Implementing off-target analysis and testing into therapeutic genome editing protocols in practice is an important issue that is broader in scope than our report. Fundamentally, variant-aware off-target analysis may identify off-target potential that would be overlooked by conventional analysis. Of note, as is true for off-target genetic changes in general, the mere possibility of somatic genetic alteration does not imply functional consequence. Although in principle, ex vivo-edited patient cells could be tested by sequencing before infusion, the functional importance of off-target edits may range from likely functional to likely neutral, so the mere presence of off-target editing in a cell product may not necessarily preclude its clinical use, and this testing could deplete precious material and delay therapy. We recommend several steps to minimize risk of unintended allele-specific off-target effects during therapeutic genome editing, consistent with regulatory guidance to consider effects of genetic variation. First, prioritize use of genome editing methods that maximize specificity, such as high-fidelity editors and pulse delivery. Second, nominate off-targets in a variant-aware manner, with particular attention toward genetic variants found in relevant patient populations, using a tool like CRISPRme (Cancellieri *et al.*, 2023). Third, use off-target detection assays that are variant-aware to empirically evaluate the likelihood of off-target editing, although these may imperfectly

reflect editing in a therapeutic context (Supplementary Note 7). When possible, allele-specific off-target editing potential should be validated in primary cells of relevant genotype by sequencing. However, it may be difficult to obtain such primary cells to perform biological validation in a relevant therapeutic context. Fourth, perform a risk assessment of variant off-target editing given predicted genomic annotations, mechanisms of DNA repair, delivery to target cells and disease context. For example, off-target edits within tumor suppressor loci might carry greater risk than those targeting unannotated noncoding sequences. Fifth, if excess allele-specific genome editing risks are identified, consider including genotype among the subject inclusion/exclusion criteria. Finally, for therapeutic genome editing indications in which it is feasible (such as hematopoietic cell targeting), prospectively monitor somatic modifications in patient samples to gather information about the frequency and consequence of such events to help assess patient-specific risk and provide valuable information for the broader field as to the frequency and in vivo dynamics of off-target edits if present.CRISPRme offers a simple-to-use tool to comprehensively evaluate off-target potential across diverse populations and within individuals. CRISPRme is available at `http://crisprme.di.univr.it` and may also be deployed locally to preserve privacy.

6.3 Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification

Genetic variation contributes substantially to complex disease risk. While well-powered genome-wide association studies (GWAS) (Tam *et al.*, 2019)and rare variant analyses from cohort studies such as the UK Biobank (UKB) (Bycroft *et al.*, 2018) have associated thousands of loci and genes with clinical phenotypes, these observational approaches are often insufficient to identify causal variants. Perturbation-based methods enable evaluation of the impact of an individual variant in a common genetic background, isolated from genetically linked variants, and such testing can be performed in high throughput through multiplex assays of variant effect (MAVEs) (Gasperini *et al.*, 2016). Numerous types of MAVEs have been developed, including deep mutational scanning (DMS) (Araya and Fowler, 2011), saturation mutagenesis (Myers *et al.*, 1986), massively parallel reporter assays (MPRA) (Inoue and Ahituv, 2015), and CRISPR-based screens (Bock *et al.*, 2022; Shalem *et al.*, 2014; Wang *et al.*, 2014). CRISPR base editing screens have emerged as a uniquely powerful method to study variants in their endogenous genomic context. Base editors, fusions of Cas9-nickase and single-stranded cytosine or adenine deaminase enzymes (Komor *et al.*, 2016; Gaudelli *et al.*, 2017),enable site-specific installation of transition variants. As the majority of disease-associated variants are single-nucleotide transitions citeprees2018base, base editors enable the installation of functionally relevant variants in a precise and scalable way. Base editing screens have been employed to dissect coding variant effects as well as to evaluate GWAS-associated variant functions (Hanna *et al.*, 2021; Morris *et al.*, 2023; Martin-Rufino *et al.*, 2023; Cuella-Martin *et al.*, 2021; Pablo *et al.*, 2023; Coelho *et al.*, 2023; Cheng *et al.*, 2021; Sánchez-Rivera *et al.*, 2022; Kim *et al.*, 2022; Kweon *et al.*, 2020; Huang *et al.*, 2021a; Sangree *et al.*, 2022; Lue *et al.*, 2023; Després *et al.*, 2020; Garcia *et al.*, 2023; Lue *et al.*, 2023). However, base editing efficiency varies substantially depending on the local sequence context surrounding the target base, the specific Cas9 variant and deaminase used, and the cellular context (Arbab *et al.*, 2020). Moreover, base edits can occur at multiple positions within the single-stranded DNA bubble created by the guide RNA (gRNA)-DNA binding on the opposite strand, therefore a single gRNA can install a variety of alleles, each with distinct efficiencies. While there have been efforts to predict editing outcomes using massively parallel base editor reporter assay data (Arbab *et al.*, 2020), these predictions do not generalize well to unprofiled base editors and cellular contexts (Sánchez-Rivera *et al.*, 2022). In previous base editing screens, analysis of phenotypic outcomes is confounded by variable editing efficiencies and outcomes. Phenotypic effects of gRNAs with robust editing are exaggerated, and effects of variants that are not installed as efficiently are underestimated. Such confounding is especially pernicious when the target elements are coding variants, as a single gRNA may install distinct coding variants with different frequencies, and current analysis methods are unable to deconvolve such data. Existing base editing screens have dealt with the heterogeneity in gRNA efficiency and genotypic outcomes in several ways. One approach that has been employed is to assume all editable nucleotides within the editing window are edited with uniform efficiency (Hanna *et al.*, 2021). Two recent studies have profiled the gRNAs used in phenotypic base editing screening using a base editor reporter (or sensor) assay (Sánchez-Rivera *et al.*, 2022; Kim *et al.*, 2022) to filter gRNAs with low editing efficiency when analyzing their phenotypic data. Despite these initial efforts, the computational analyses of these screens have not yet been formalized, often relying on existing tools that were not designed specifically for base editor data with or without the target site reporter. Here, we design an experimental-

computational pipeline to improve the accuracy of variant effect estimation in base editing screens. By incorporating a target site reporter sequence into the gRNA construct, we simultaneously measure the editing efficiency of a gRNA and its phenotypic impact. We develop a computational pipeline, BEAN, that normalizes the phenotypic scores of target variants using genotypic outcome information collected from the target site reporter. Moreover, we extend BEAN to analyze densely tiled coding sequence base editing screen data, sharing information among neighboring gRNAs to obtain accurate phenotypic scores for each coding variant. BEAN provides a first-in-class integrated solution to experimental assessment of variant effects through base editing screens. We systematically benchmark BEAN against current state-of-the-art methods for the analyses of pooled CRISPR screens and show substantially improved performance of BEAN. To leverage activity-normalized base editing screening, we have conducted screens assessing the impact of low-density 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 contribute substantially to coronary artery disease risk. Serum LDL-C measurements are quantitative and nearly uniformly measured in most biobanks, and thus they provide among the highest quality human phenotypic data for any trait. A trans-ancestry GWAS meta-analysis from the Global Lipids Genetics Consortium (GLGC) has identified >900 genome-wide significant loci associated with blood lipid levels, including >400 loci associated with LDL-C (Graham *et al.*, 2021). LDL-C GWAS loci overlap strongly with liver-enriched gene expression, nominating liver as the primary tissue driving LDL-C variant effects (Wang *et al.*, 2022; Finucane *et al.*, 2018). Yet, the causal variants and mechanisms by which many of these loci modulate LDL-C levels remain unknown. LDL-C levels are also impacted by rare coding variants. In the most severe instances, inherited monogenic variants in several genes cause Familial Hypercholesterolemia (FH), a disease associated with extremely elevated LDL-C levels and premature cardiovascular disease (Bouhairie and Goldberg, 2015). The majority of genetic mutations known to cause FH occur in LDLR, a cell surface receptor that uptakes LDL, thus removing it from circulation (Brown and Goldstein, 1984) Despite the effectiveness of lipid lowering therapies, FH patients are still 2-4-fold more likely to have coronary events than the general population (Mundal *et al.*, 2018). Elevated LDL-C levels increase cardiovascular disease risk throughout life, so the early identification of at-risk individuals would have immense clinical utility (Bouhairie and Goldberg, 2015). However, many LDLR variants currently lack clinical interpretation. Of the 1,427 LDLR missense variants in the ClinVar databasen (Landrum *et al.*, 2020), 50% are classified as variants of unknown significance (VUS) or to have conflicting interpretations of pathogenicity ("conflicting"), thus impeding FH diagnosis. Likewise, of the 758 unique LDLR missense variants carried by sequenced individuals in the UKB cohort, 69% are either unreported or have an uncertain annotation in ClinVar. Altogether, improved understanding of LDLR variant impacts would enable earlier diagnosis and treatment for a large number of at-risk individuals. We have modeled the impacts of both common GWAS-associated and rare LDLR coding variants through base editing installation followed by cellular uptake of fluorescent LDL-C in HepG2 cells, which provides a scalable flow cytometric assay to measure a key contributing factor of serum LDL-C levels (Hamilton *et al.*, 2023) given the majority of serum LDL-C is cleared in liver (Spady, 1992). By applying our experimental-computational pipeline to this screen model, we identify LDL uptake-altering GWAS-associated variants and characterize their downstream impact on chromatin accessibility, transcription factor binding, and gene expression that leads to differential LDL uptake. We nominate causal variants that alter LDL-C uptake through impacting the genes OPRL1, VTN, and ZNF329, which have not previously been connected with LDL-C levels. Through saturation tiled base editing of LDLR, not only do we accurately distinguish known pathogenic vs. benign variants, we find strong correlation between missense variant functional scores and the LDL-C levels of patients in the UKB who carry these variants. We combine functional scores with structural modeling to mechanistically classify deleterious variant impacts, revealing a key, conserved tyrosine residue in each LDLR class B repeat that interacts with the neighboring repeat to maintain structural integrity. Altogether, BEAN provides a widely applicable tool to characterize single-nucleotide variant functions.

6.3.1 A base editing reporter profiles endogenous editing outcomes

To enable accurate interrogation of variant effects at scale, we built a platform to perform dense, high-coverage base editing screens that accounts for variable editing efficiency and genotypic outcomes. To maximize coverage of variants in base editing screens, we built lentiviral adenine (ABE8e) (Gaudelli *et al.*, 2017; Richter *et al.*, 2020) and cytosine (AID-BE5) (Arbab *et al.*, 2020) deaminase base editor (BE) constructs using the near-PAM-less SpCas9 variant, SpRY (Walton *et al.*, 2020). Both BEs showed native genomic editing activity, as measured in HepG2 cells by ASGR1 splice site editing followed by

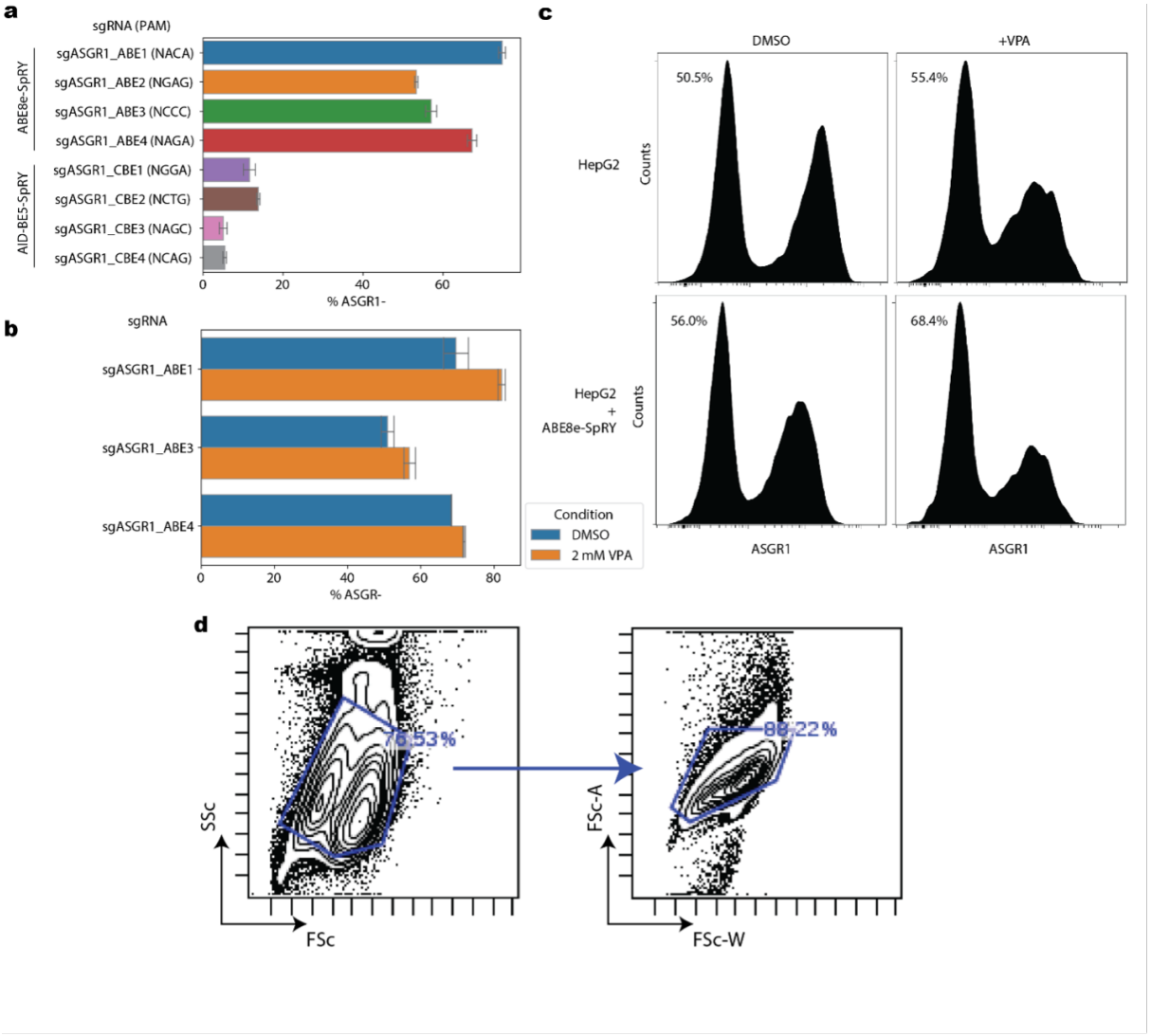


Figure 6.26. Optimization of SpRY base editing. (A)) gRNA editing efficiency in HepG2 with ABE8e-SpRY and AID-BE5-SpRY for ASGR1 splice site-targeted gRNAs, measured by the fraction of ASGR1 negative cell counts (% ASGR -) quantified by flow cytometry from two experimental replicates. **(B)** gRNA editing efficiency with and without valproic acid (VPA) treatment from 4 experimental replicates. **(C)** Flow cytometry signal with and without stable ABE8e-SpRY integration and valproic acid (VPA) treatment. Fraction of ASGR negative cell counts in each condition is labeled as percentage in the panel. **(D)** Example flow cytometric gates used in all analysis and sorting experiments to filter for single cells.

flow cytometric anti-ASGR1 antibody staining, with ABE8e-SpRY showing considerably more robust maximal activity (**Fig.6.26(A)**). Editing efficiency was increased by 5-10% by prior lentiviral integration of constitutively expressed BEs and by transient dosing of cells with the histone deacetylase valproic acid immediately after BE and gRNA transduction (**Fig.6.26(B)-(C)**), and thus these treatments were implemented in all screens.Base editing efficiency is known to vary depending on Cas9 binding efficiency as well as the local sequence and chromatin context surrounding the target base (Arbab *et al.*, 2020; Shin *et al.*, 2021; Yang *et al.*, 2023), and thus we expected gRNAs to vary substantially in editing efficiency across target sites. To account for this variability, we synthesized and cloned each gRNA paired with a 32-nt reporter sequence comprising the genomic target sequence of that gRNA into lentiviral base editor vectors (**Fig. 6.27(A)**), akin to previously published CRISPR mutational outcome reporter constructs (Sánchez-Rivera *et al.*, 2022; Kim *et al.*, 2022). When introduced into cells, the gRNA can edit both its native genomic target site and the adjacent target site (reporter) in the lentiviral vector, which can be read out using next-generation sequencing (NGS). We designed two gRNA libraries using this approach to improve understanding of the genetics of LDL-C levels. The first library (LDL-C GWAS library) targets 583 variants associated with LDL-C levels from the UK Biobank GWAS cohort. We included fine-mapped variants with posterior inclusion probability (PIP) >0.25 from either the SUSIE or Polyfun fine-mapping pipelines (Wang *et al.*, 2020; Weissbrod *et al.*, 2020), and also variants with PIP > 0.1 within 250 kb of any of 490 genes found to significantly alter LDL-C uptake from recent CRISPR-Cas9

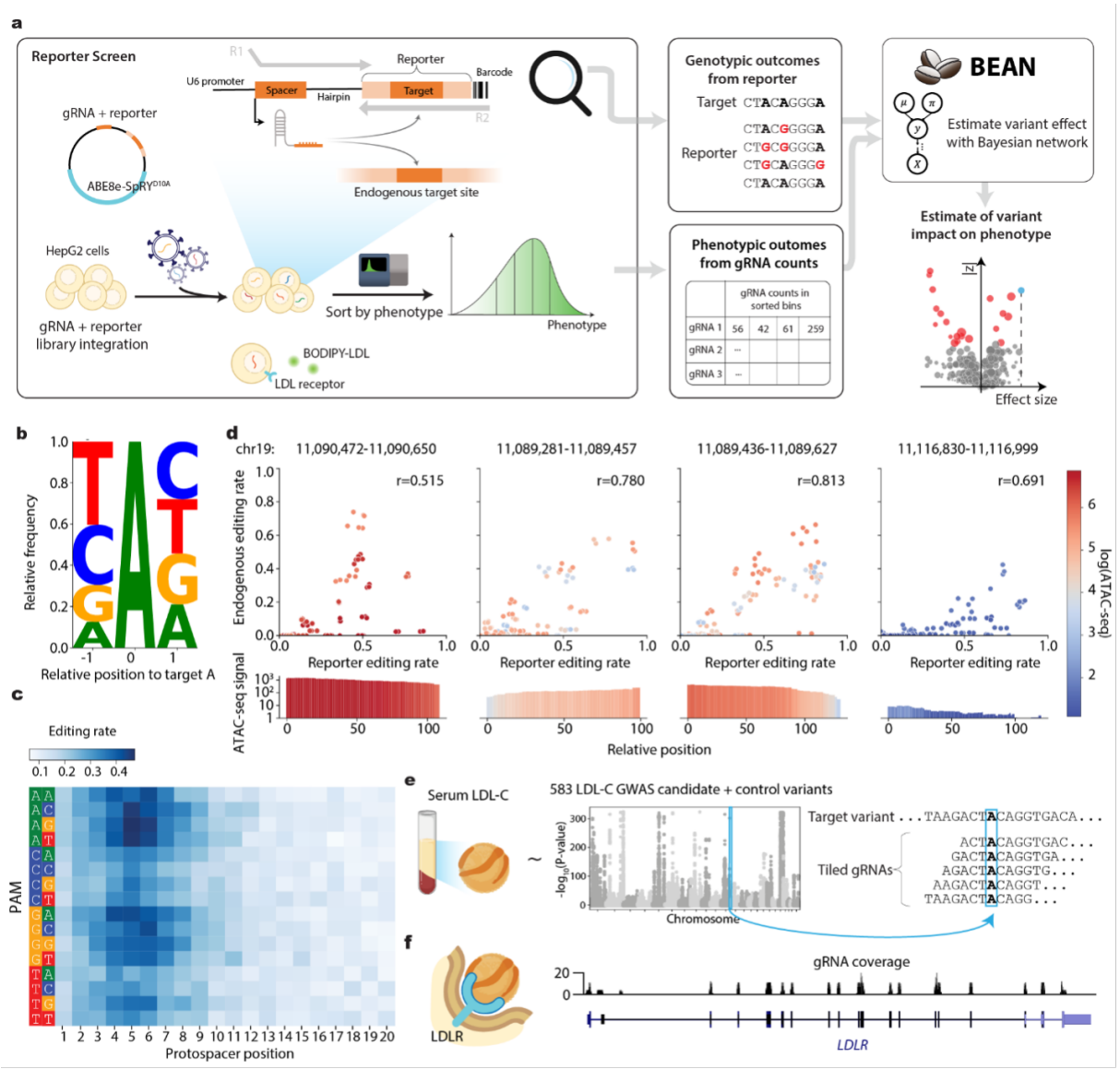


Figure 6.27. Activity-normalized base editing screening pipeline. (A)) Schematic of activity-normalized base editing screening process and analysis by BEAN. A library of gRNAs, each paired with a reporter sequence encompassing its genomic target sequence, is cloned into a lentiviral base editor expression vector. Lentiviral transduction is performed in HepG2, followed by flow cytometric sorting of four populations based on fluorescent LDL-cholesterol (BODIPY-LDL) uptake. The gRNA and reporter sequences are read out by paired-end NGS to obtain gRNA counts and reporter editing outcomes in each flow cytometric bin. BEAN models the reporter editing frequency and allelic outcomes and gRNA enrichments among flow cytometric bins using BEAN to estimate variant phenotypic effect sizes. **(B)** Adjacent nucleotide specificity of ABE8e-SpRY editing represented as a sequence logo from 7,320 gRNAs; the height of each base represents the relative frequency of observing each base given an edit at position 0. **(C)** Average editing efficiency of ABE8e-SpRY by protospacer position and PAM sequence **(D)** Scatterplots comparing nucleotide-level editing efficiency between the reporter and endogenous target sites for a total of 49 gRNAs across four loci across 3 experimental replicates. The accessibility of the four loci as measured by ATAC-seq signal in HepG2 is shown in the top panel, and the scatterplot markers are colored by the accessibility of each nucleotide. Pearson correlation coefficients are shown as r. **(E)** Schematic of the LDL-C variant library gRNA design for selected GWAS candidate variants with a Manhattan plot showing variant P-values from a recent GWAS study (Klimentidis *et al.*, 2020). gRNAs tile the variant at five positions with maximal editing efficiency (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.

knockout screens (Hamilton *et al.*, 2023; Emmer *et al.*, 2021). We designed five tiled gRNAs for each variant allele that place the variant in positions shown to induce most efficient editing with ABE8e (**Fig. 6.27(E)**) (Arbab *et al.*, 2023). Positive control gRNAs which ablate splice donor and acceptor consensus sites in six genes found to have significantly altered LDL-C uptake upon knockout (Hamilton *et al.*, 2023) and 100 non-targeting negative control gRNAs that tile 20 synthetic variants were included, for a total of 3,455 gRNAs.The second library (LDLR tiling library) targeted the LDLR gene. Taking advantage of the flexible PAM recognition of SpRY, every possible gRNA targeting the LDLR coding sequence on both strands was included. Lower density gRNAs, tiled every 2-3-nt, targeted the the 50-nt flanking

each LDLR exon, the LDLR 5' and 3' UTR, promoter, and two intronic enhancers (**Fig. 6.27(F)**). This library also contained 150 non-targeting negative control gRNAs, for a total of 7,500 gRNAs.We first assessed editing outcomes through lentiviral transduction of each library in HepG2 cells followed by NGS of gRNA-reporter pairs 10-14 days afterwards. We developed an end-to-end computational toolkit for base-editing screens, BEAN, which includes the ability to perform quality control and quantify editing outcomes from raw reads among other functionalities. Importantly, the quantification step is designed to account for self-editing of the spacer sequence, which we found to occur at appreciable frequency and with modest correlation with reporter editing frequency (LDL-C GWAS library median 31%, Pearson r=0.36, LDLR tiling library median 18%, r=0.31). We used BEAN to profile the previously uncharacterized PAM-less base editors ABE8e-SpRY and AID-BE5-SpRY on reporter data from the >10,000 gRNAs in both libraries (**Fig.6.27(B)** and **(C)**),. The result clearly recapitulated the hallmark positional preferences of these base editors (Myers *et al.*, 1986; Wang *et al.*, 2014) the NRY PAM preference of the SpRY enzyme (Komor *et al.*, 2016; Gaudelli *et al.*, 2017), and the relative depletion of editing at AA dinucleotides by ABE8e. Notably, the average maximal positional ABE8e-SpRY editing frequency at protospacer positions 3-8 across dinucleotide PAM sequences ranges from 32% to 46%, indicating the ability of this enzyme to install variants efficiently across a wide variety of genomic locations.To validate that editing of the reporter provides an accurate surrogate for endogenous editing, we generated a library where both the reporter and endogenous target site are sequenced following the editing by 49 gRNAs across four loci surrounding LDLR with varying levels of HepG2 chromatin accessibility. We demonstrate that nucleotide-level and allele-level reporter editing fractions correlate well with endogenous target site editing fractions ((**Fig.6.27(D)**, average Pearson correlation across 4 loci is r=0.70 for per-nucleotide editing rate r=0.70, per-allele editing rate r=0.69), and the reporter shows higher correspondence than BE-Hive predictions29 (Nucleotide r=0.44, allele r=0.64) (**Fig.6.28**). Notably, while reporter editing correlates with endogenous editing at all four loci, we found that endogenous editing frequency also depends on the accessibility of the target region, as has been previously reported for Cas9-nuclease (Schep *et al.*, 2021; Ding *et al.*, 2019; Liu *et al.*, 2019) and base editors (Shin *et al.*, 2021; Yang *et al.*, 2023). Yet, current computational analyses do not model these dependencies, motivating the development of a tailored modeling framework. We then performed fluorescent LDL uptake screens with each library in $\geq$ 5 biological replicates, ensuring >500 cells per gRNA at all stages. We used simulation to determine the optimal flow cytometric sorting scheme, accounting for variability in gRNA editing rate, gRNA coverage, gDNA sampling and PCR amplification (`https://github.com/pinellolab/screen-simulation`). Based on our simulation result that finer bin widths improves sensitivity (Supplementary Fig. 6, Supplementary Note 1), we flow cytometrically isolated four populations per replicate with the very low (0-20% percentile), low (20-40%), high (60-80%), and very high (80-100%) LDL uptake (**Fig.6.27(A)**), performing NGS on gRNA and reporter pairs in each sorted population. We observed robust replicability (median Spearman ρ=0.84 for LDL-C GWAS library, 0.88 for LDLR tiling library) in gRNA counts across replicates (Supplementary Fig. 7), indicating technical reproducibility. the variant at five positions with maximal editing efficiency (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.

6.3.2 Activity-normalized base editing screen analysis with BEAN

We postulated that the gRNA editing outcomes provided by the reporter together with the accessibility of the target region could improve the quantification of variant phenotypic effects in our pooled BE screens. To do so, we developed a novel analysis method, BEAN (Base Editor screen analysis with Activity Normalization), to quantify the effect of each variant from gRNA abundance in sorted populations along with genotypic outcome information provided by reporter editing. BEAN assumes that the observed phenotypic distribution in a population of cells for each gRNA derives from a mixture of cells with unedited and edited alleles (**Fig. 6.30**). The proportion of cells carrying a given gRNA that possess a particular genotype is inferred based on the editing outcome observed in reporter as well as chromatin accessibility of the target locus using a Bayesian network. The distribution of cells with each gRNA prior to sorting is modeled as a Gaussian mixture for each underlying genotype produced by that gRNA. Because multiple gRNAs may induce the same genotypic outcome at different frequencies, BEAN uses this redundancy to build confidence in the predicted phenotypic impacts of a given genotype. As the output for each variant, BEAN provides its effect size i.e. the posterior mean phenotypic shift along with the corresponding z-score, and 95% credible interval (CI). We also note that BEAN can be adapted to an arbitrary number and arrangement of sorting bins and other base editing enzymes including those with uncharacterized editing preferences, and can accommodate screens without reporter or accessibility

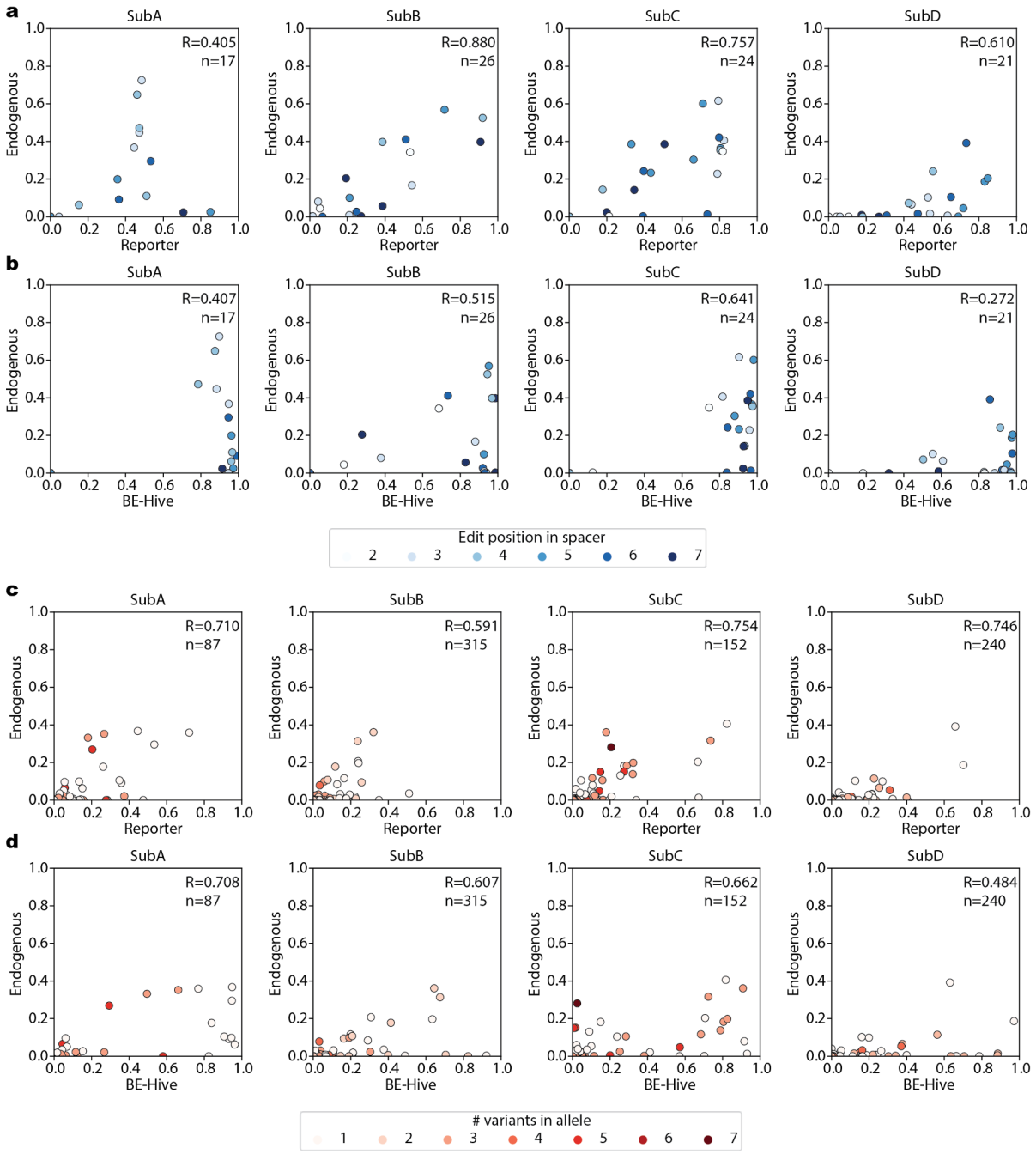


Figure 6.28. Endogenous target site editing rate comparison with reporter and BE-Hive predicted editing outcomes. Mean endogenous and reporter edit rates across three replicates are plotted. **(A)-(B)** Nucleotide-level editing rates in **(A)** reporter and **(B)** BE-Hive plotted against those in endogenous loci within protospacer position 2 to 7 (1-based, inclusive). **(B)** Allele level editing rates in **(C)** reporter and **(D)** BE-Hive plotted against those in endogenous loci. Alleles are defined within editing window within protospacer 1 to 19 (1-based, inclusive). SubA-D denotes sublibraries A-D. Pearson correlation coefficients are denoted as R and the number of points is denoted as n.

information.

BEAN only assumes population-level consistency between editing of the reporter and endogenous target site. We hypothesized that variation in editor expression or cellular state may lead certain cells to be more amenable to editing than others. In this scenario, "jackpot" cells would be more likely to have editing at both endogenous and reporter loci. To assess this possibility, we compared the enrichment of a gRNA in the highest vs. lowest sorted LDL uptake quantile bin with the difference in reporter editing observed in cells sorted into these bins, reasoning that endogenous editing should be highest in the cells sorted into the enriched bin. We indeed observed such correlation for LDLR and MYLIP splice-ablating gRNAs (Spearman ρ=0.32, **Fig. 6.31**), suggesting the existence of cell-level factors leading to "jackpot" cells with higher editing at both endogenous and reporter loci. However, the correlation

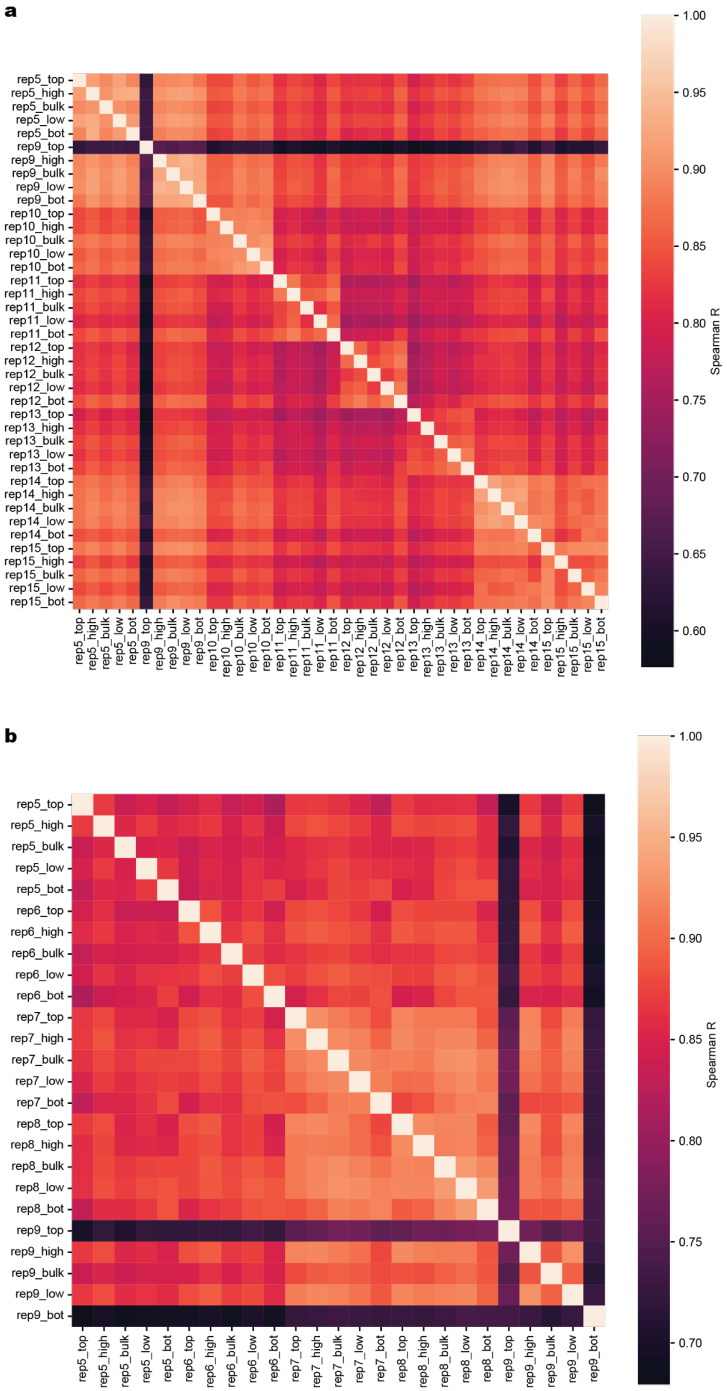


Figure 6.29. Base editing screen read correlations. Spearman correlation coefficient of **(A)** gRNA counts of LDL-C GWAS library and **(B)** LDLR tiling library samples after filtering out for gRNAs with less than 10 read counts. Top; 80-100%, high; 60-80%, low; 20-40%, bot: 0-20% percentile LDL-C uptake. Rep; experimental replicate.

between phenotypic and reporter editing enrichment was weaker when considering all positive control gRNAs (Spearman ρ=0.13). We thus concluded that incorporating the jackpot effect into BEAN would be unlikely to improve model performance.

6.3.3 BEAN identifies LDL uptake altering GWAS variants

We applied BEAN to the LDL-C GWAS library screen. From the reporter data, variant editing efficiency per gRNA is highly variable with average edit fraction of 34.0%. Encouragingly, most target variants are edited at high efficiency by at least one of the five targeting gRNAs (median maximal editing of 60.4%, **Fig. 6.32**). First, we compared the performance of BEAN and five published CRISPR screen analysis methods at distinguishing the effects of positive control splice-altering variants versus nega-

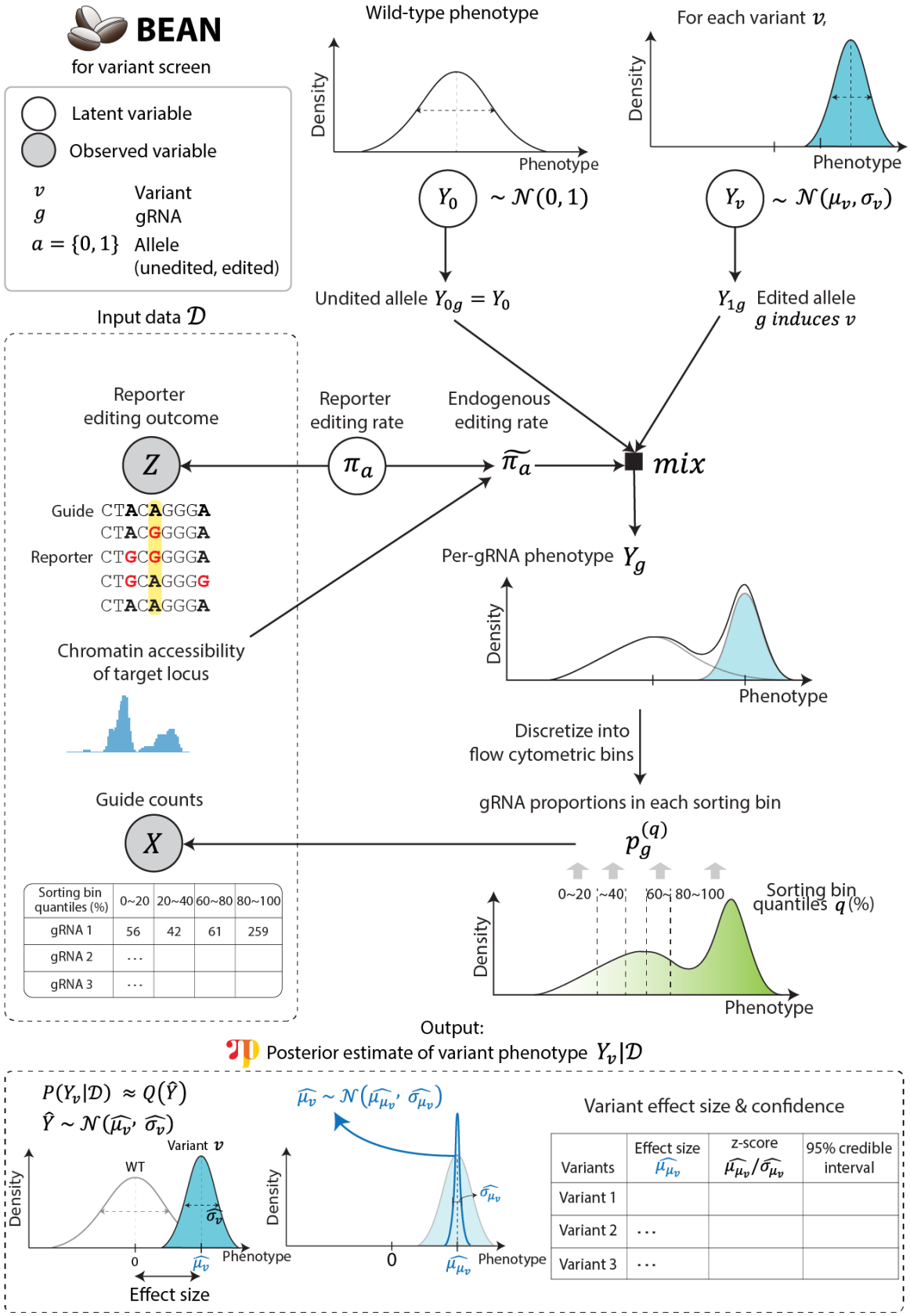


Figure 6.30. BEAN models variant effects from activity-normalized base editing screens. Simplified schematic of BEAN Bayesian network that models input reporter editing outcomes and gRNA counts. The Bayesian network model recapitulates the data generation process starting from a variant-level phenotype Y_v and models per-gRNA phenotypes as a Gaussian mixture distribution of edited and unedited (wild-type) allele phenotypes. The weights of the mixture components are modeled to generate reporter editing outcomes. gRNA abundance in each sorting bin is then calculated by discretizing the gRNA phenotype based on the experimental design into the phenotypic quantiles, and is modeled to generate the observed gRNA counts using an overdispersed multivariate count distribution (see Methods). BEAN outputs the parameters of the posterior distribution of mean phenotypic shift as Gaussian distribution with mean $\hat{\mu_\mu}$ (effect size), along with negative-control adjusted z-score and credible interval (CI), where $\mathcal{D}$ is the input data.

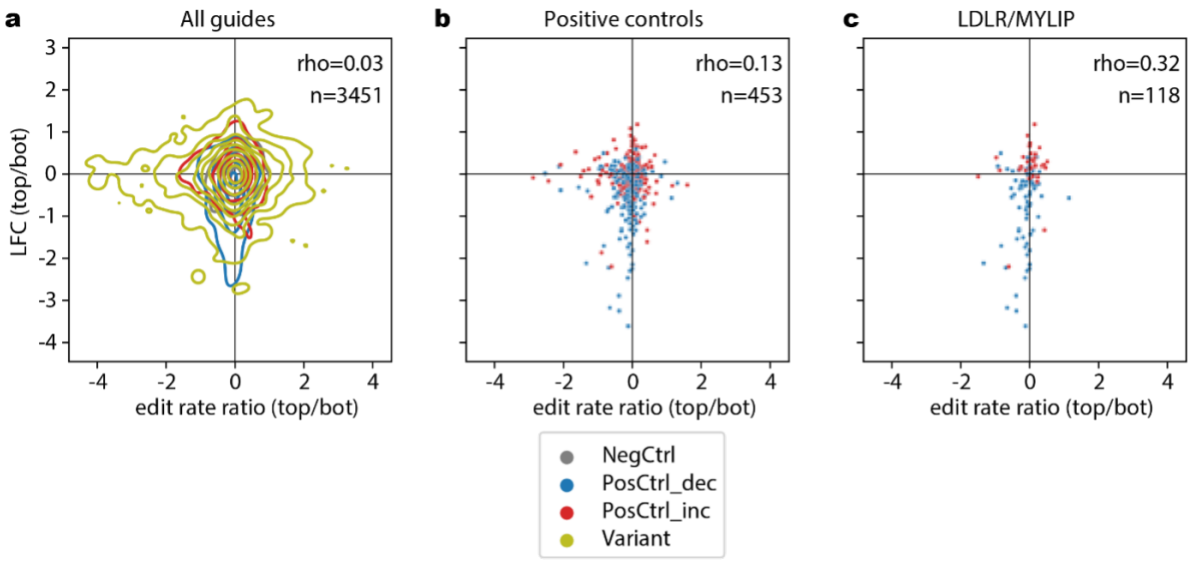


Figure 6.31. Jackpot analysis. Scatterplots of editing rate enrichment plotted against gRNA abundance enrichment of **(A)** all gRNAs, **(B)** positive control gRNAs, **(C)** strongest positive control gRNAs targeting LDLR or MYLIP splicing sites. Enrichment of both editing efficiencies and gRNA abundance is calculated as the log fold chance of the measures in the highest and lowest 20% quantile bin. Spearman correlation coefficients are shown as rho. NegCtrl; negative control gRNAs, PosCtrl_dec; positive control gRNAs that is expected to decrease LDL-C uptake, PosCtrl_inc; positive control gRNAs that is expected to increase LDL-C uptake.

tive control non-targeting gRNAs (Huang *et al.*, 2021a; Li *et al.*, 2014, 2015; Jeong *et al.*, 2019; Daley *et al.*, 2018) **Fig. 6.33(A)**). To dissect the contributions of individual features to BEAN performance, we included two reduced versions of BEAN: one that considers reporter editing but not chromatin accessibility (BEAN-Reporter), and another that ignores the reporter, assuming uniform gRNA editing efficiency (BEAN-Uniform). BEAN outperforms other evaluated methods at this classification task (**Fig. 6.33(B)**)), and this improved performance is accentuated when the data is subsampled for fewer replicates, demonstrating its ability to maintain robustness even with less data. Importantly, BEAN shows improved performance (mean AUPRC=0.90 across 15 2-replicate subsamples) over BEAN-Reporter (mean AUPRC=0.87), which in turn outperforms BEAN-Uniform (mean AUPRC=0.85), supporting the value of accurately modeling target site editing. Intriguingly, even BEAN-Uniform outperforms alternative approaches, likely due to more accurate modeling of sorting bins, suggesting the utility of BEAN in sorting screens without reporter. Having demonstrated robust performance of BEAN, we evaluated our ability to characterize common variants that alter LDL-C uptake. We identified 54 variants that significantly alter LDL-C uptake (95% CI does not contain 0). These variants include intronic variants in well-known LDL-C uptake mediators whose knockout altered LDL-C uptake in a recent genome-scale CRISPR screen37 such as ABCA1, LDLR, and SCARB1 (**Fig. 6.33(E)**). We additionally identified coding/intronic variants in APOE, CCND2, GAS6, and FBLN1 with strong genetic likelihood of causality (UKBB SUSIE fine-mapping PIP ¿ 0.99 and/or the only variant in a fine-mapped credible set(Graham *et al.*, 2021)), indicating that the effect of these variants on serum LDL-C is at least partially mediated by LDL-C uptake. To validate the inferred effect sizes, we performed individual lentiviral ABE8e-SpRY transduction of HepG2 cells with gRNAs targeting 22 variants and 4 positive controls. We performed fluorescent LDL-C uptake profiling of each edited cell line mixed with an in-well control cell line in 6 biological replicates, allowing us to compare changes in LDL-C uptake with matched data from the screen. The individual LDL-C uptake log-fold-change (LFC) values showed strong correlation to the BEAN effect sizes (μ, Spearman R=0.69, **Fig. 6.33(C)** and **D**), showing more robust correlation than BEAN-Uniform (R=0.68), log fold change based on MAGeCK-RRA (R=0.51), and regression coefficients β of MAGeCK-MLE (R=0.61). These data demonstrate that BEAN enables accurate inference of variant effects on LDL-C uptake over a wide dynamic range.

To gain insight into a set of 20 variants for which the mechanism of LDL-C uptake alteration is less clear, we developed a pipeline to assess the cellular effects of variant installation (**Fig. 6.34(A)**). First, we asked which of these variants impact chromatin accessibility. We established an approach to perform pooled variant editing followed by ATAC-seq. High multiplicity of infection (MOI) lentiviral delivery of a pool of 20 ABE8e-SpRY gRNAs to HepG2 cells was followed by ATAC-seq and paired genomic DNA collection in three biological replicates in standard and serum-starved conditions. We performed multiplexed

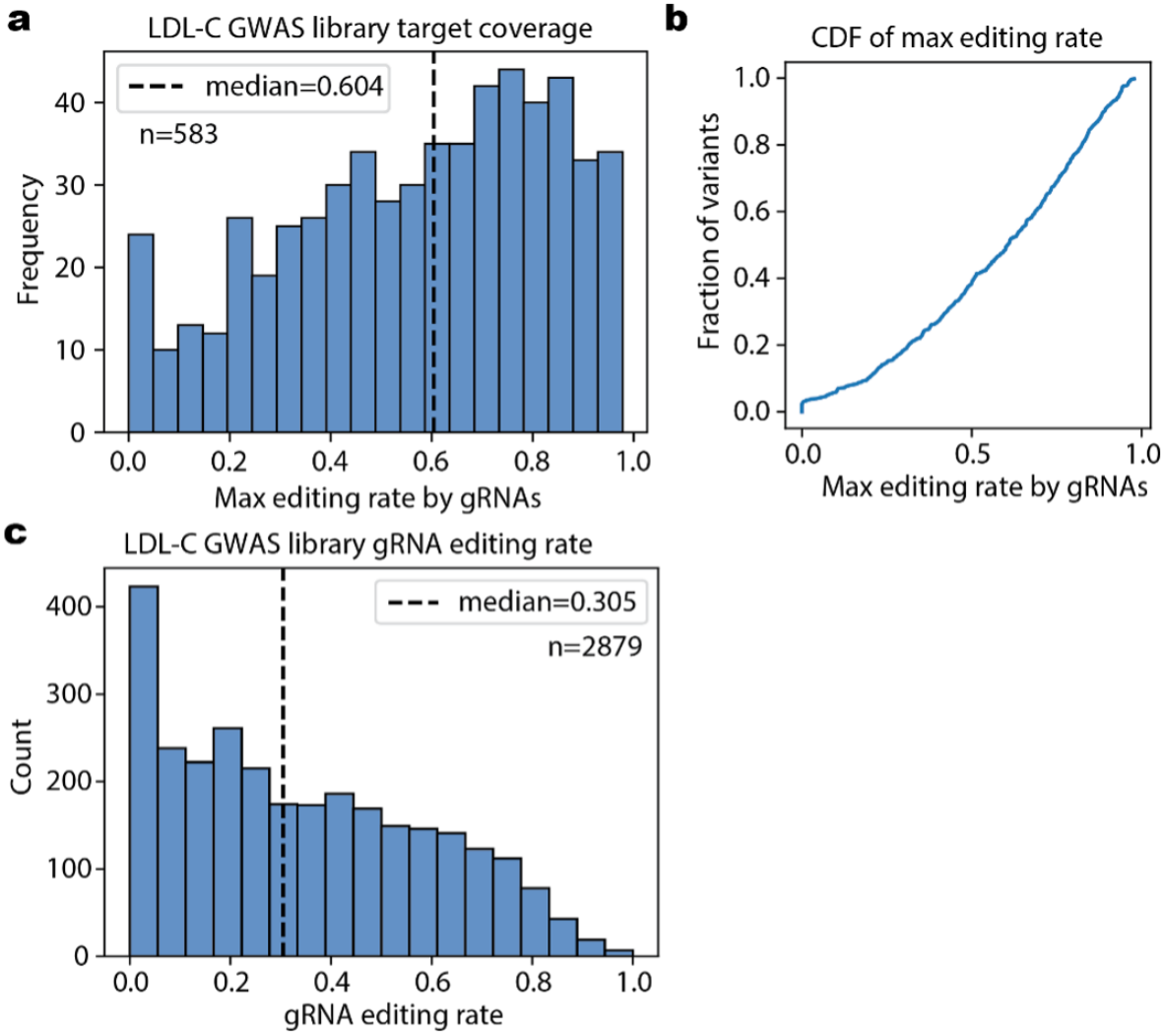


Figure 6.32. LDL-C GWAS variant editing coverage. (A) Histogram of per-target variant coverage calculated by the maximal editing rate of the variant by any gRNAs targeting the variant, where gRNA editing rates are calculated as the mean editing rate in bulk samples across replicates where more than 10 reads are observed. **(B)** Cumulative distribution function of the same distribution plotted in **(A)**. **(C)** Per-gRNA editing rate of the target variant in bulk samples across replicates where more than 10 reads are observed.

PCR enrichment of the regions surrounding each of the 20 edited variants followed by targeted amplicon sequencing by NGS. Differential representation of an alternate allele in ATAC-seq relative to gDNA sequencing implies differential accessibility of the alternate allele than the reference (**Fig. 6.34(B)**). Eight of the 20 variants are heterozygous in HepG2, and thus we could assess whether these variants reside in chromatin accessibility quantitative trait loci (caQTL) (Degner *et al.*, 2012), showing differential relative accessibility of the two haplotypes irrespective of base editing. We found five of these eight variants to be caQTLs (**Fig. 6.34(C)**). Two of these loci (rs35081008 and rs2618566) were also identified as caQTLs in a recent analysis of 20 human liver tissue samples (Currin *et al.*, 2021). Importantly, caQTL analysis cannot address the causality of the evaluated variant due to the presence of linked variants which could contribute to the differential ATAC-seq signal. To assess whether individual variants alter chromatin accessibility, we evaluated whether base editing induces differential accessibility for any of the 20 tested variants. Technical issues including insufficient representation of the region, insufficient editing, and inability to phase heterozygous loci prevented assessment of five of the variants. Of the 15 remaining variants, eight significantly altered chromatin accessibility when edited (family-wise error rate 0.1, **Fig. 6.34(C)**). Four such variants (rs11149612, rs35081008, rs8126001, and rs2618566) were in loci identified as liver tissue caQTLs. Because base editing only alters a single variant in a locus, this analysis establishes at least partial causality to the tested variant. We performed deeper characterization of three variants whose editing alters both LDL-C uptake and chromatin accessibility. Rs704 is a missense coding variant in VTN and is the only variant in a fine-mapped credible set from LDL-C GWAS30, suggesting high likelihood of causality. The other two variants are in gene promoters—rs35081008 is in

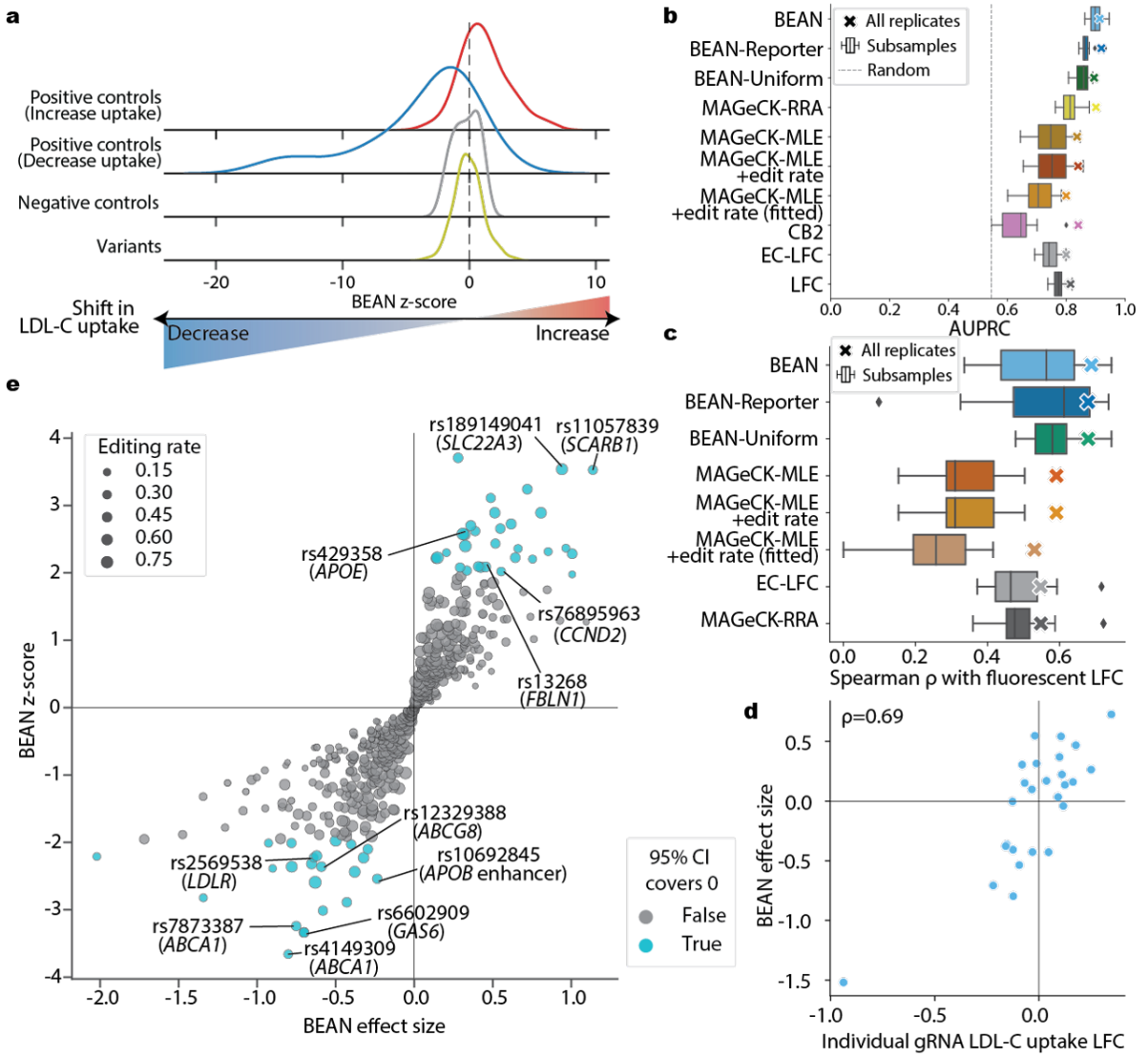


Figure 6.33. BEAN improves variant impact estimation from the LDL-C GWAS library screen. (A) Ridge plot of BEAN z-score distributions of positive controls, negative controls, and test variants. **(B)** AUPRC of classifying LDLR and MYLIP splicing variants vs. negative controls. Metrics for all 5 replicates are shown as markers and metrics of 15 two-replicate subsamples among the 5 replicates are shown as box plots. **(C)** Spearman correlation coefficient between BEAN effect size and log fold change of LDL-C uptake following individual testing of 26 gRNAs. Metrics for all 5 replicates are shown as marker and metrics of 15 two-replicate subsamples among the 5 replicates are shown as box plots. **(D)** Scatterplot of BEAN effect size and log fold change of LDL-C uptake following individual testing of 26 gRNAs. Spearman correlation coefficient is denoted as ρ. **(E)** Scatterplot of variant effect size and significance estimated by BEAN. Labels show rsIDs of selected variants and dbSNP gene annotations and a manual annotation for APOB enhancer in the parenthesis.

the ZNF329 promoter, and rs8126001 is in the shared OPRL1/RGS19 promoter **Fig. 6.34(D)**. Both variants have moderate probability of causality from GWAS evidence (SUSIE PIP=0.49 for rs35081008, PIP=0.25 for rs8126001), with the remaining probability in the rs35081008 locus deriving from a linked variant (rs34003091) 19-nt upstream in the ZNF329 promoter. None of the putative target genes has been previously found to alter LDL-C uptake, nor do they show significance in LDL-C burden analyses. To investigate how the prioritized variants might affect transcription factors (TF) binding sites and thereby regulate proximal genes involved in LDL-C uptake, we adapted the MotifRaptor pipeline (Yao *et al.*, 2021). For each variant, we retrieved genomic sequences spanning 61 bp centered around the SNP location, using the hg38 genome assembly as a reference. Each sequence was mutated by substituting the major allele with the minor allele at the SNP position, yielding both a reference and an alternative sequence for each variant. Subsequently, to evaluate the potential for transcription factor (TF) binding, we employed all the human TF position weight matrices (PWMs) from the CIS-BP database (Weirauch *et al.*, 2014) to scan each pair of reference and alternative sequences. This motif scanning generated binding scores at each sequence position, serving as predictive indicators of TF binding potential. We then compared these scores for each TF across the reference and alternative alleles within every sequence

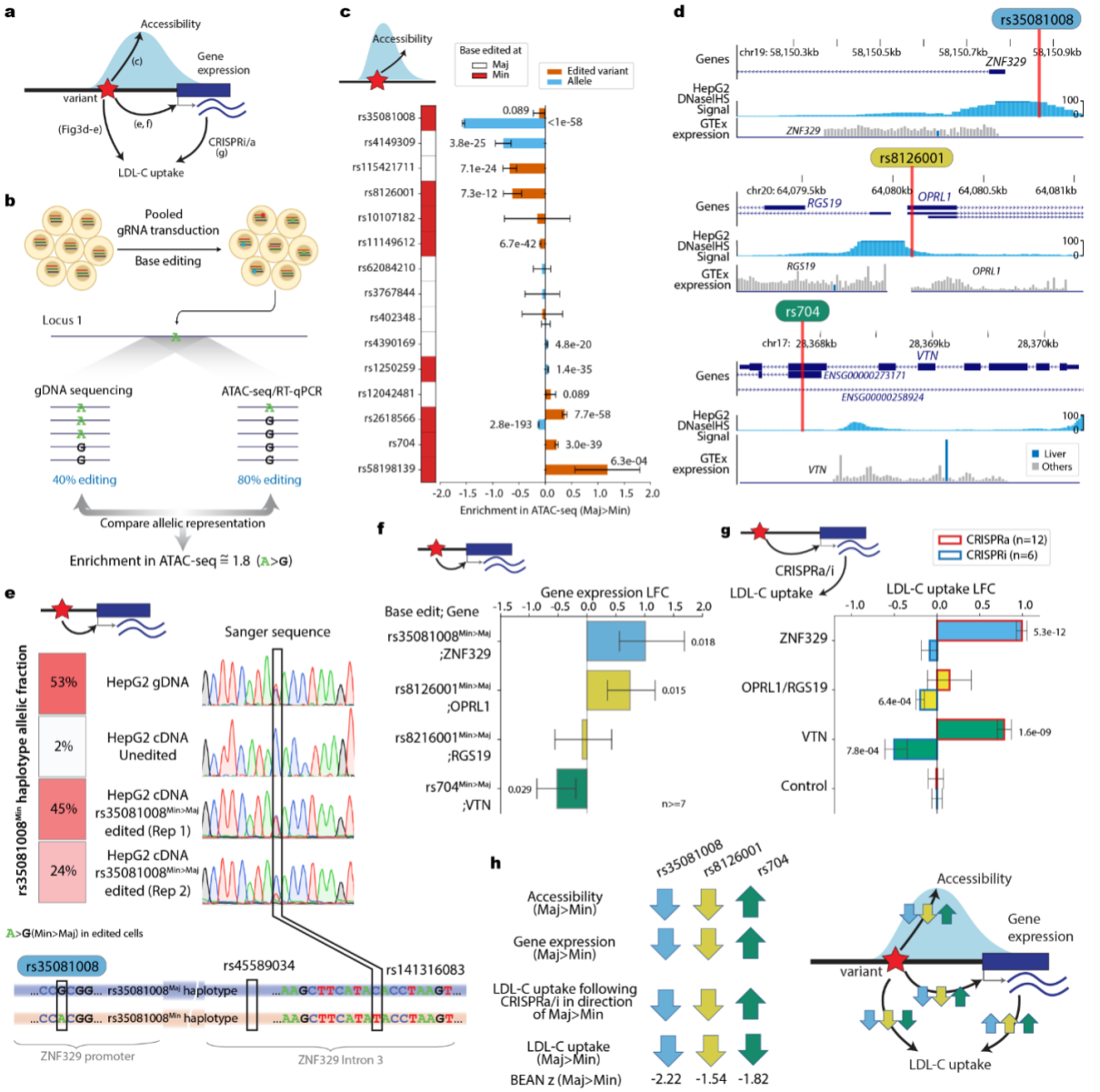


Figure 6.34. Functional characterization of LDL-C GWAS variants. (A) Schematic of potential variant mechanisms and the figure panels showing data from each mechanistic experiment. **(B)** Schematic of pooled ATAC-seq analysis to identify variants impacting accessibility. Differential representation of allele in gDNA and ATAC-seq reflects differential accessibility induced by the base edit or heterozygous reference allele. **(C)** Change in ATAC-seq enrichment from the pooled ATAC-seq experiment. 95% confidence intervals are shown as the error bars. "Edited variant" denotes the enrichment by base edit and "Allele" denotes the enrichment by either of the heterozygous alleles, when available, translated uniformly to major (Maj) to minor (Min) allele direction. Variants where the base edit is conducted from minor allele are denoted as red in the color bar. Family-wise error rate (FWER) with Benjamini-Hochberg multiple correction is shown for each enrichment value where FWER ¡ 0.1. **(D)** Genomic tracks for three selected variants. DNaseIHS; DNase 1 Hypersensitivity. Multiple transcript variants of RGS19 and OPRL1 are shown in the middle panel. **(E)** Fraction of ZNF329 minor (Min) allele haplotype reads in gDNA and cDNA from untreated HepG2 and HepG2 with rs35081008Min¿Maj base editing. **(F)** Change in gene expression following base editing of three selected variants from minor (Min) to major (Maj) allele. P-values of the one sample Student's t-test of LFC vs. mean of 0 that are smaller than 0.05 are shown above each bar. **(G)** Change in cellular LDL-C uptake following CRISPRa/i of proximal genes for three selected variants. P-values of the one sample Student's t-test of LFC vs. mean of 0 that are smaller than 0.05 are shown above each bar. **(H)** Summary schematic of characterization results.

pair. This comparative step is crucial for determining a variant's impact on TF binding. Specifically, higher binding scores for the alternative sequence indicate an increase in TF binding potential, while lower scores suggest a decrease. To quantify these changes, we calculated a 'disruption score' as follows:

$$\text{disruption} = \text{score}(s_{alt}) - \text{score}(s_{ref})$$

This score helps capture the directional change each variant induces, where a negative value signifies reduced TF binding potential and a positive value indicates an increase. For rs8126001, our approach

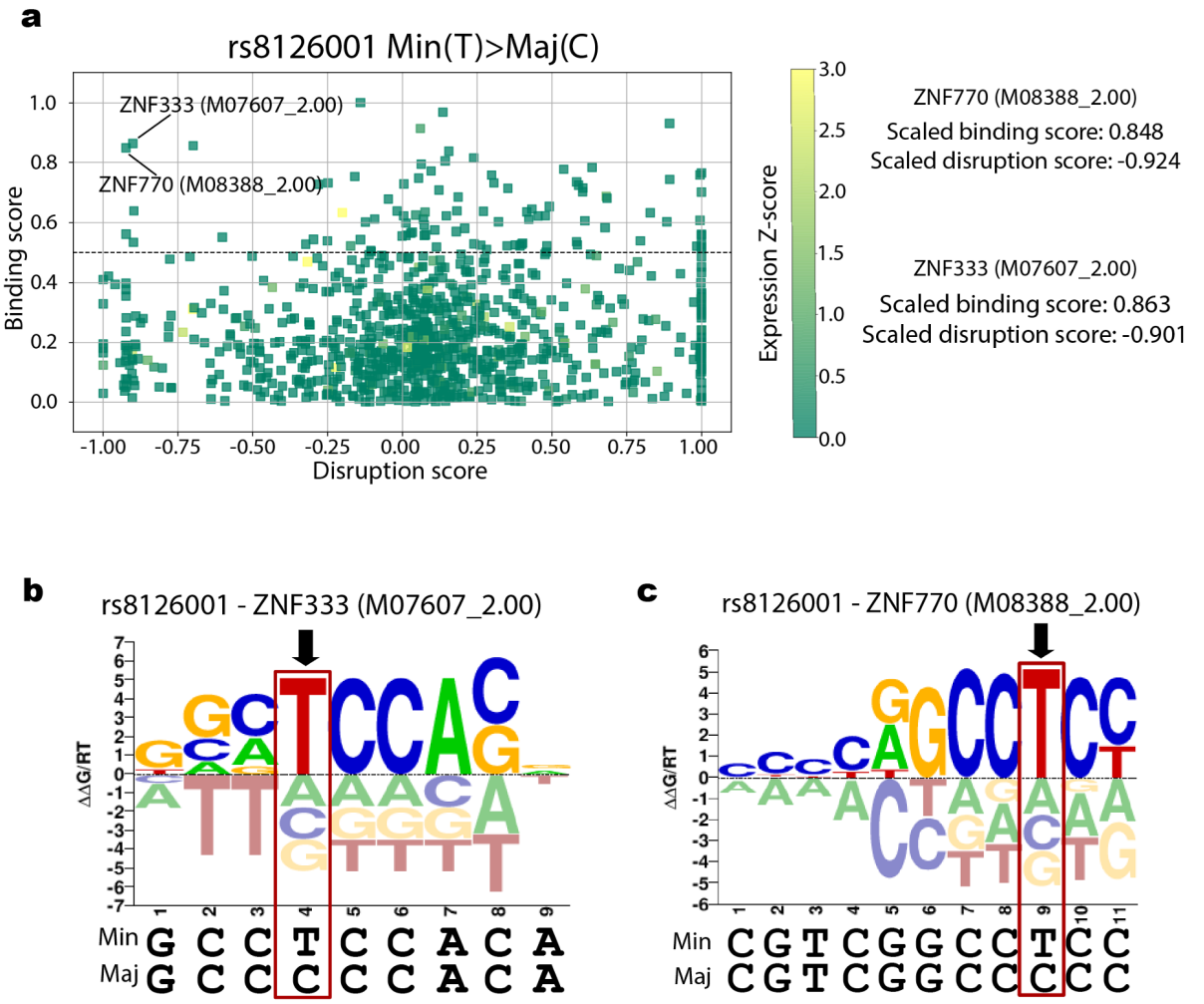


Figure 6.35. MotifRaptor analysis of candidate variant transcription factor binding disruption. (A) Disruption plotted against binding score of motifs of rs8126001 minor allele transition to major. **(B-C)** Identified ZNF333 and ZNF770 motif aligned with rs8126001 loci with minor and major alleles.

prioritized two zinc finger TFs, ZNF333 and ZNF770 with enhanced binding site sequences due to the heterozygous minor allele in HepG2 cells (**Fig. 6.35**). HepG2 ChIP-seq data59 further support the binding of these TFs at this locus, although the variant lies at the edge of the peaks (**Fig. 6.36**). While definitive conclusions about these factors will require further experimental validation, our observations align with previous research (Farh *et al.*, 2015) suggesting that only a minority of causal variants directly alter canonical TF binding sequences and instead affect non-canonical sequences. We confirmed through RT-qPCR analysis that editing the minor to major alleles of rs35081008 and rs8126001 leads to increased expression of ZNF329 and OPRL1 respectively (**Fig. 6.34(F)**), which is consistent with the increased chromatin accessibility induced by these edits. rs35081008 is heterozygous in HepG2, and we used two linked ZNF329 intronic variants to assess allele-specific expression. In wild-type HepG2, only 2% of ZNF329 transcripts derive from the minor allele haplotype (**Fig. 6.34(E)**), consistent with the diminished chromatin accessibility of this allele (**Fig. 6.34(C)**) and the status of rs35081008 as a liver eQTL. Editing rs35081008 from minor to major allele restores expression of this haplotype to 35% of total transcripts (**Fig. 6.34(E)**), providing further evidence that rs35081008Maj results in increased expression of ZNF329.We then performed CRISPRa and CRISPRi targeting to assess whether altered expression of the four candidate target genes alters LDL-C uptake. CRISPRa induction of VTN and ZNF329 significantly increased LDL-C uptake, and CRISPRi repression of VTN and OPRL1/RGS19 reduced LDL-C uptake (**Fig. 6.34(G)**). In our base editing experiments, rs704Min shows decreased LDL-C uptake, so we surmise that this allele must have decreased expression or function, given that decreased VTN expression decreases LDL-C uptake (**Fig. 6.34(H)**). Prior biochemical characterization has shown decreased cellular binding capacity of rs704Min , suggesting a possible mechanistic explanation. Our data are consistent with rs35081008Min decreasing ZNF329 expression, which in turn decreases LDL-

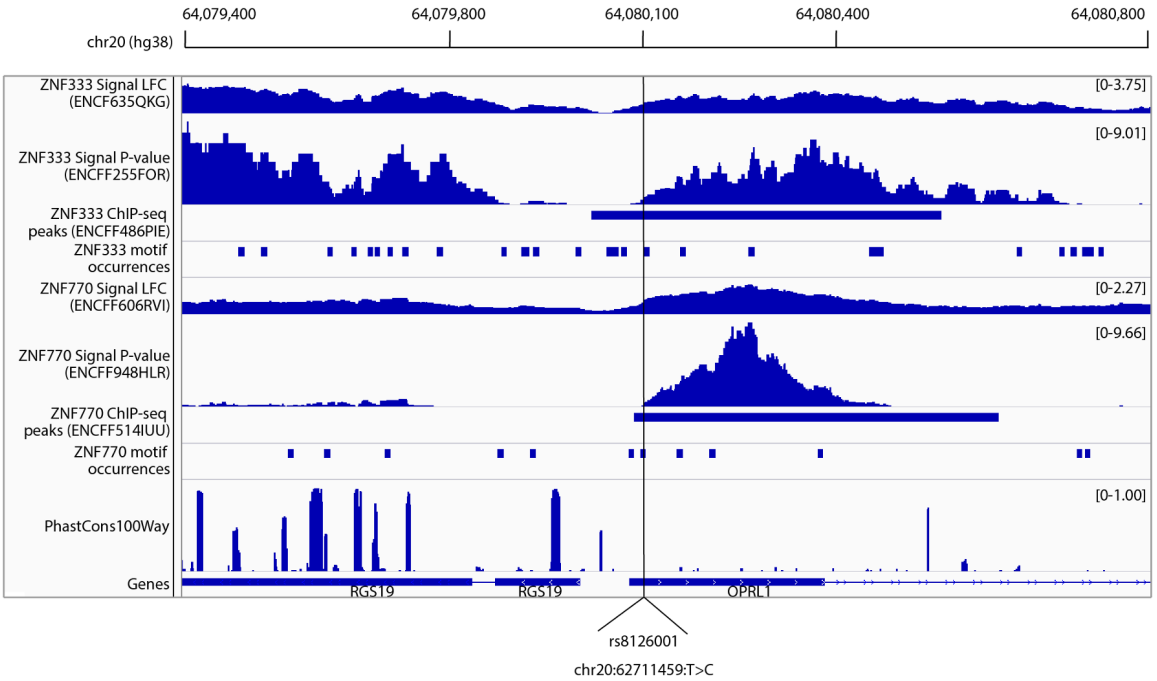


Figure 6.36. ChIP-seq signal LFC, signal P-values, peaks and motif occurrences of ZNF333 and ZNF770 around rs8126001. PhastCons100way conservation scores and gene annotations are displayed together. ENCODE accessions are shown in the parenthesis.

C uptake. Finally, our data are most consistent with rs8126001Min decreasing OPRL1 expression, which leads to decreased LDL-C uptake. This observation aligns with the higher predictive binding affinity of ZNF333, a transcriptional repressor (Jing *et al.*, 2004; Witzgall *et al.*, 1994), with rs8126001Min and the potential disruption of its binding with rs8126001Maj. In summary, through accurately quantifying impacts of disease-associated variants on LDL-C uptake, BEAN reveals genetic mechanisms underlying control of LDL-C levels.

6.3.4 Saturation LDLR coding sequence tiling screening enables quantitative assessment of rare variant deleteriousness

We next adapted BEAN to the LDLR tiling library, enhancing the model to specifically assess the contributions of individual amino acid mutations rather than SNVs, by enabling a more comprehensive understanding of coding region alterations. Previous coding sequence base editing analyses have assumed that all editable bases within a window are edited, which leads to erroneous amino acid mutation assignments, or have analyzed gRNA-level signal only (Martin-Rufino *et al.*, 2023). We aimed to exploit the combination of dense tiling afforded by ABE8e-SpRY and reporter editing outcomes to model the effects of coding variants more accurately. The LDLR tiling screen showed high coverage of edited nucleotides and amino acids (92% of targetable nucleotides and 74% of the 860 LDLR amino acids in the LDLR coding sequence were edited at >10% frequency by at least one gRNA in the reporter,. A total of 2,182 distinct variants were assessed, of which 874 are missense coding variants. Each gRNA produced an average of 2.6 distinct alleles, and each variant was covered by 5.8 gRNAs on average. Thus, ABE8e-SpRY tiling of LDLR resulted in a rich dataset of coding variants for the evaluation of their phenotypic impacts. As opposed to the LDL-C GWAS analysis in which each gRNA was evaluated based on its editing frequency at a single target position, we adapted BEAN to account for multi-allelic outcomes. First, BEAN translates the edited alleles, i.e., aggregates nucleotide-level allele counts that leads to the identical amino acid transition into a single amino-acid level allele counts, while preserving nucleotide transition in non-coding regions. BEAN then filters for the translated alleles that are robustly observed for each gRNA (**Fig. 6.37(A)**). BEAN uses a Bayesian network to combine phenotypic information from all the gRNAs that produce a given allele. Importantly, the phenotype attributed to each gRNA is modeled as a mixture distribution of the alleles it generates, with the contribution of each allele weighted by its corresponding editing frequency.BEAN assigned significant z-scores (<-1.96, equivalent to 95% credible interval not covering 0) to 145 among 2,182 variants assessed from the LDLR tiling library, 131 of which decrease LDL-C uptake. 47 variants that significantly decrease LDL-C uptake are

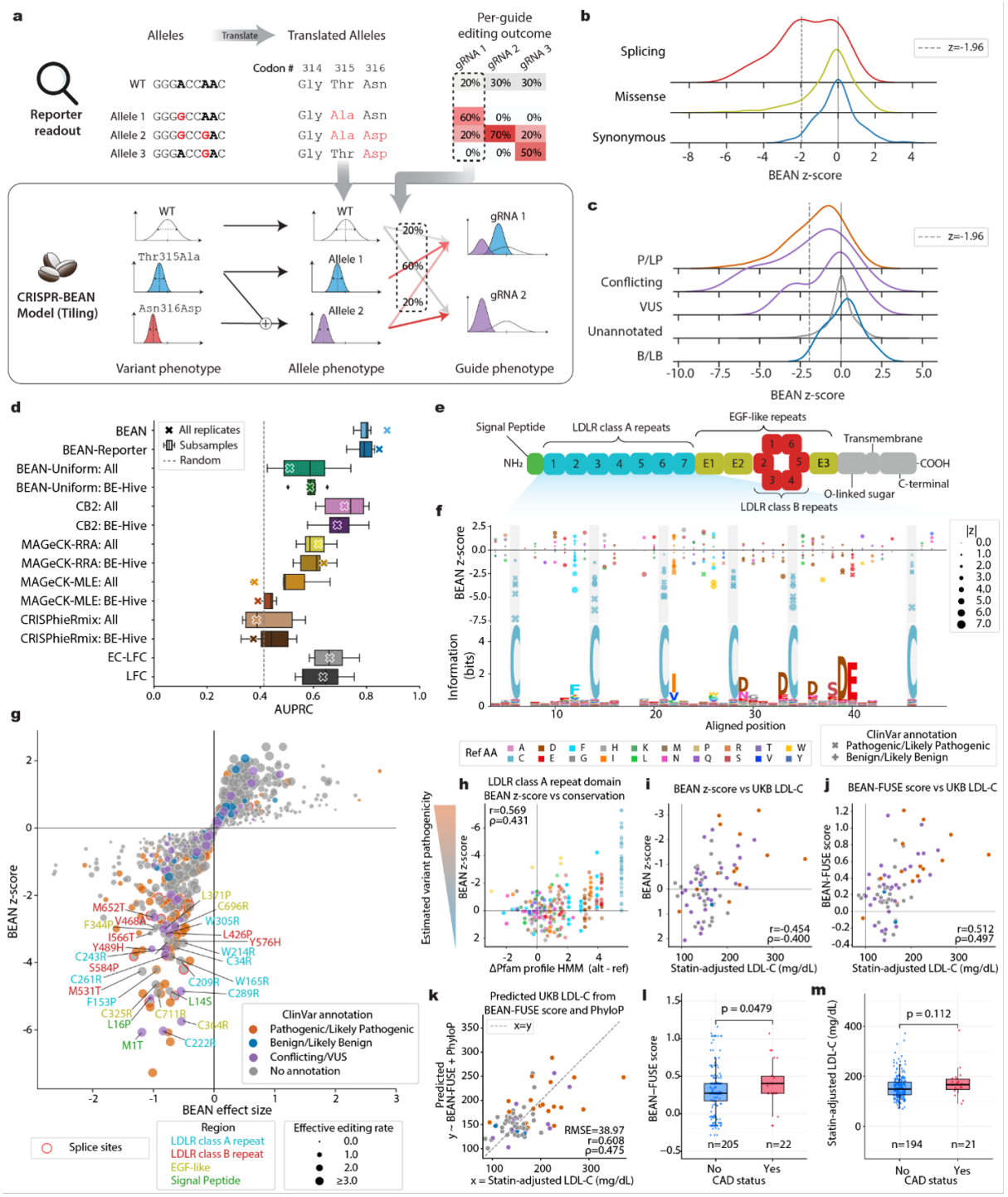


Figure 6.37. Dissection of LDLR variant effects through BEAN modeling of a saturation tiled base editing screen. (A) BEAN model for coding sequence tiling screens. Reporter editing efficiencies are calculated at the amino acid-level when the edited nucleotides are in coding region. Phenotypes of gRNAs with multi-allelic outcomes are modeled as the Gaussian mixture of allelic phenotypes. If an allele consists of more than one variant, the phenotype of the allele is modeled as the sum of the component variants. A Bayesian network is used to model variant-level phenotypes, sharing phenotypic information from all available gRNAs. **(B)** Ridge plot of BEAN z-score distributions of positive controls, negative controls, and variants. **(C)** Ridge plot of BEAN z-score distributions of Clinvar variants annotated as pathogenic/likely pathogenic (P/LP), benign/likely benign (B/LB), conflicting interpretation of pathogenicity (conflicting), and Uncertain significance (VUS), and unannotated variants. **(D)** AUPRC of classifying ClinVar pathogenic/likely pathogenic vs. benign/likely benign variants. The marker shows the metrics of each method run on 4 replicates with no failing samples. Boxplot shows the metrics of 6 2-replicate combinations of the 4 replicates. **(E)** LDLR domain structure adopted from Oommen et al71. **(F)**BEAN *z*-scores for variants in the 7 LDLR class A repeat domains aligned with the Pfam profile HMM logo. Highly conserved cysteines are highlighted in grey. **(G)** Scatterplot of estimated variant effect sizes and z-scores. Labels of selected deleterious variants without ClinVar pathogenic/likely pathogenic annotations are shown. **(H)** Scatterplot of LDLR class A repeat missense variant BEAN *z*-scores and ΔPfam profile HMM scores. Higher ΔPfam scores correspond to substitution from highly conserved to rarely observed amino acids. **(I)** Comparison of mean statin-adjusted LDL-C level and BEAN z-score for variants observed in UKB and base editing. **(J)** Comparison of mean statin-adjusted LDL-C level and BEAN-FUSE scores for variants observed in UKB and base editing. **(K)** LDL-C levels of observed missense variants predicted by a regression model using BEAN-FUSE and PhyloP scores with 10-fold cross validation, compared with mean statin-adjusted LDL-C level in UKB. **(L)** Boxplots of BEAN-FUSE functional scores for UKB individuals with variants observed in our base editing screen with or without CAD. **(M)** Boxplots of statin-adjusted LDL-C levels of UKB invididuals with variants observed in our base editing screen with or without CAD. *P*-value of two-sided Wilcoxon rank-sum test is denoted. *r*; Pearson correlation coefficient, *ρ*; Spearman correlation coefficient

annotated in ClinVar as pathogenic/likely pathogenic, while 17 are ClinVar VUS/conflicting variants and none are ClinVar benign/likely benign **Fig. 6.37(G)**), indicating that BEAN can reliably predict the pathogenicity for variants without a pathogenic or benign classification **Fig. 6.37(B)** and **(C)**). We compared the performance of BEAN at distinguishing ClinVar-annotated pathogenic from benign/likely benign LDLR variants to other available screen analysis methods (Li *et al.*, 2014, 2015; Jeong *et al.*, 2019; Daley *et al.*, 2018). To allow comparison of methods that do not account for editing outcomes, we assigned outcomes to each gRNA either by assuming all editable bases within the maximal editing window are perfectly edited (Hanna *et al.*, 2021) ("All") or by using the most frequent predicted outcome from BE-Hive (Arbab *et al.*, 2020)("BE-Hive"). As in the LDL-C GWAS screen, BEAN showed better performance than any other method **Fig. 6.37(D)**), and BEAN also outperformed the model variants that do not account for accessibility (BEAN-Reporter) or reporter editing outcomes (BEAN-Uniform), further justifying the modeling decision to explicitly leverage editing outcomes and accessibility. BEAN achieves an AUPRC of 0.88 at this task, indicating highly effective distinction of pathogenic and benign LDLR variants through scoring HepG2 LDL-C uptake proficiency. To gain insight into mechanisms by which these variants disrupt LDL-C uptake, we examined BEAN z-scores for variants that reside within conserved functional domains **Fig. 6.37(E)** and **(F)**). LDLR contains seven highly conserved LDLR class A repeats that bind to LDL. The LDLR class A repeat is structurally anchored by six highly conserved cysteines that form three disulfide bonds (Fass *et al.*, 1997). As expected, many of the missense variants with the strongest effects on LDLR function disrupt these cysteines **Fig. 6.37(F)**). We find that cysteine mutating edits in each of the seven LDLR class A repeats disrupt LDLR activity, suggesting that structural integrity of all repeats is required for efficient LDL binding, although disruption is most impactful in repeats 3-7. Truncation experiments have reported that repeats 1 and 2 are dispensable for LDL binding (Russell *et al.*, 1989) in partial accord with our results (Jeon and Blacklow, 2005). To examine the relationship between conservation and function more comprehensively in these repeats, we compared the BEAN z-score of every installed variant with its change in amino acid conservation score from the Pfam profile HMM66. We observed strong concordance (Pearson r = 0.57 **Fig. 6.37(H)**), with N-terminal hydrophobic residues and C-terminal calcium-coordinating acidic residues within the repeats also showing particular functional importance, as expected from the known function of these domains.Encouraged by the concordance between our screen and conservation scores within the LDLR class A repeats, we asked whether BEAN scores could predict functional impairment across the entire LDLR gene. We examined statin-adjusted LDL-C levels (glo, 2013) in the UK Biobank (UKB) for individuals with paired exome sequencing and lipid level data. To control for the contribution of other variants in genes known to impact serum LDL-C level, we filtered out individuals who harbor nonsynonymous APOB or PCSK9 variants or multiple LDLR missense variants, leading to 9,819 individuals harboring 358 distinct LDLR missense variants. There are 76 distinct LDLR missense variants observed in our base editing data with UKB carriers. We observe robust concordance between the average carrier LDL-C and BEAN scores for these variants (Spearman $\rho = 0.40$, Pearson r = 0.45, **Fig. 6.37(I)**, suggesting that BEAN provides accurate quantitative prediction of the impact of LDLR missense variants on control of serum LDL-C levels in the human population. As our base editing screen does not exhaust possible mutation types per position, we used the FUSE (Yu *et al.*, 2023) pipeline to impute the impact of unobserved variants at positions at which a different missense variant is scored. FUSE uses an amino acid substitution matrix derived from 24 deep mutational scanning datasets to impute functional scores for all possible missense variants at positions observed in base editing data (BEAN+FUSE score, see Methods). Applying FUSE to the 76 UKB variants with observed base editing data, BEAN-FUSE shows improved correlation with UKB carrier LDL-C (Spearman $\rho = 0.50$, Pearson r = 0.51, **Fig. 6.37(J)**). BEAN-FUSE correlation with UKB carrier LDL-C was robust but lower at all 358 missense LDLR variants with lipid measurements (Spearman $\rho = 0.37$, Pearson r = 0.35). Altogether, BEAN-FUSE provides a pipeline to extend base editing screening to predict functional impairment for unobserved missense variants, although our data suggest that accuracy does decrease for unobserved variants. As base editing provides orthogonal functional assessment to conservation, we asked whether the LDL-C levels of UKB variant carriers could be predicted with BEAN-FUSE scores and PhyloP 100way vertebrate conservation scores. Using XGBoost regression (Chen and Guestrin, 2016), we achieved more robust correlation with UKB carrier LDL-C than either BEAN-FUSE or PhyloP alone at the 76 variants observed in the base editing screen (Spearman $\rho = 0.48$, Pearson r = 0.61, RMSE=39.0, **Fig. 6.37(K)**) and at 358 variants with BEAN-FUSE score (Spearman $\rho = 0.37$, Pearson r = 0.31, RMSE=51.1). This result demonstrates the potential utility of base editing data to improve quantitative phenotype prediction combined with computational prediction methods. Individuals with pathogenic FH variants are at higher risk of coronary artery disease (CAD), even after controlling for LDL-C levels (Clarke *et al.*, 2022). However, the vast majority of rare LDLR

missense variants lack ClinVar pathogenic/likely pathogenic designations, preventing information about these potentially disease-causing variants from being shared with patients. Therefore, we asked whether CAD incidence within LDLR variant carriers could be stratified by functional scores. We found that for individuals with rare LDLR variants, functional scores processed by BEAN-FUSE were significantly higher for patients with prevalent or incident CAD (Wilcoxon rank-sum test, p = 0.0479, **Fig. 6.37(L)**). BEAN-FUSE scores provided more robust stratification of individuals with CAD than statin-adjusted LDL-C values for individuals with variants covered in the screen (Wilcoxon rank-sum test, p = 0.112, **Fig. 6.37(M)**). This demonstrates the advantage of quantifying genetic risk, which has a lifelong impact on LDL-C levels, over the snapshot provided by a single LDL-C measurement. Overall, we show that activity-normalized base editing screening can yield accurate quantitative estimation of LDLR variant pathogenicity in a large human cohort.

6.3.5 Structural basis of LDLR missense variants

We further analyzed LDLR missense variants identified to significantly impair LDL-C uptake by BEAN to gain insight into mechanisms of their pathogenicity. We first examined variants with top z-scores that are unannotated or annotated as conflicting, or VUS in ClinVar. The top ranked variant, which shows even more significant loss-of-function than splice-ablating variants, alters the start codon, preventing full-length LDLR translation. Other top variants such as C222R, C261R, C289R, and C364R disrupt conserved disulfide bond-forming cysteines in LDLR class A repeats and EGF-like domains. Top-ranked variant in the of the signal peptide L16P substitute hydrophobic leucines with prolines in the transmembrane alpha helix, which is likely to distort the alpha helix (Kim and Kang, 1999) and the neighboring L15P has been shown to reduce LDLR transport to the plasma membrane (Pavloušková *et al.*, 2016) Neighboring L14S that also ranks high substitutes hydrophobic leucine with serine in the hydrophobic h-region central to the signal peptide (Von Heijne, 1985). Additionally, multiple variants disrupt calcium ion binding, which is key to LDLR class A repeat folding (Pena *et al.*, 2010) through the conversion of negatively charged amino acids (D/E) to glycine (G), thereby disrupting ionic interactions with side-chain carboxylates and calcium ions (D94G, E101G, E179G, D307G) in LDLR class A repeats. We also found that L371P, a VUS, disrupts a calcium ion interaction in the EGF-like domain by breaking the coordinate covalent bond between the calcium ion and the carbonyl group within the L371 main chain due to backbone distortion. Finally, we found that F153P significantly interfered with hydrophobic interactions between the aromatic ring and the attached saccharide on Q182. We noticed that an appreciable number of deleterious variants that lack ClinVar pathogenic designation reside in the six LDLR class B repeats. The LDLR class B repeats, also known as YWTD repeats, form a propeller-like structure involved in the release of LDL following its endocytosis. To gain insights into unannotated variant impact, focusing on the LDLR class B repeats, we used the full wild-type LDLR structure from the AlphaFold Protein Structure Database (Jumper *et al.*, 2021; Varadi *et al.*, 2022) and the MODELLER (Webb and Sali, 2016) -generated mutant structures to calculate changes in interatomic interactions using Arpeggio (Jubb *et al.*, 2017). Additionally, we predicted the effects of variants on protein stability ($\Delta\Delta$G, negative value indicates destabilization) with DDMut (Zhou *et al.*, 2023). We found that the 26 significant LDLR class B variants induce more destabilizing effects, disrupt more hydrophobic interactions, have lower relative solvent accessibility (Rose *et al.*, 1985) (0.041 of maximum residue solvent accessibility), and have higher wild-type residue depth as compared to the other observed variants in this region (**Fig. 6.38(A)-(D)**). Collectively, these observations strongly indicate that these significant LDLR class B repeat variants are predominantly buried within the protein core where they engage in extensive hydrophobic interactions essential for protein folding. Moreover, we found a conserved interaction across repeat domains in which a tyrosine holds neighboring propeller blades together through interactions with a hydrophobic residue of the neighboring repeat (**Fig. 6.38(E)-(F)**.

We identified five of these variant pairs (Y442C with V481A, Y442C with V468A, Y489H with M531T, Y532C with L568P, and Y576H with V618A), where all nine positions have at least one variant that weakens their hydrophobic interaction and has a significant BEAN z-score. Among the top-ranked unannotated or ClinVar VUS and conflicting variants within LDLR class B repeats, the six most significant variants (L426P, Y489H, M531T, I566T, S584P, and Y576H) all disrupt residues that hold the propeller blades together through hydrophobic interactions (**Fig. 6.38(G)-(I)**). Further supporting the importance of hydrophobic interactions, the base editing screen installed additional missense variants at positions 531, 566, and 652 that conserve hydrophobicity. In all cases, mutation into hydrophobic residues has less severe impact from the base editing screen and DDMut-predicted destabilization than mutation into non-hydrophobic residues (**Fig. 6.38(J)**). For example, while we find M652T to be highly

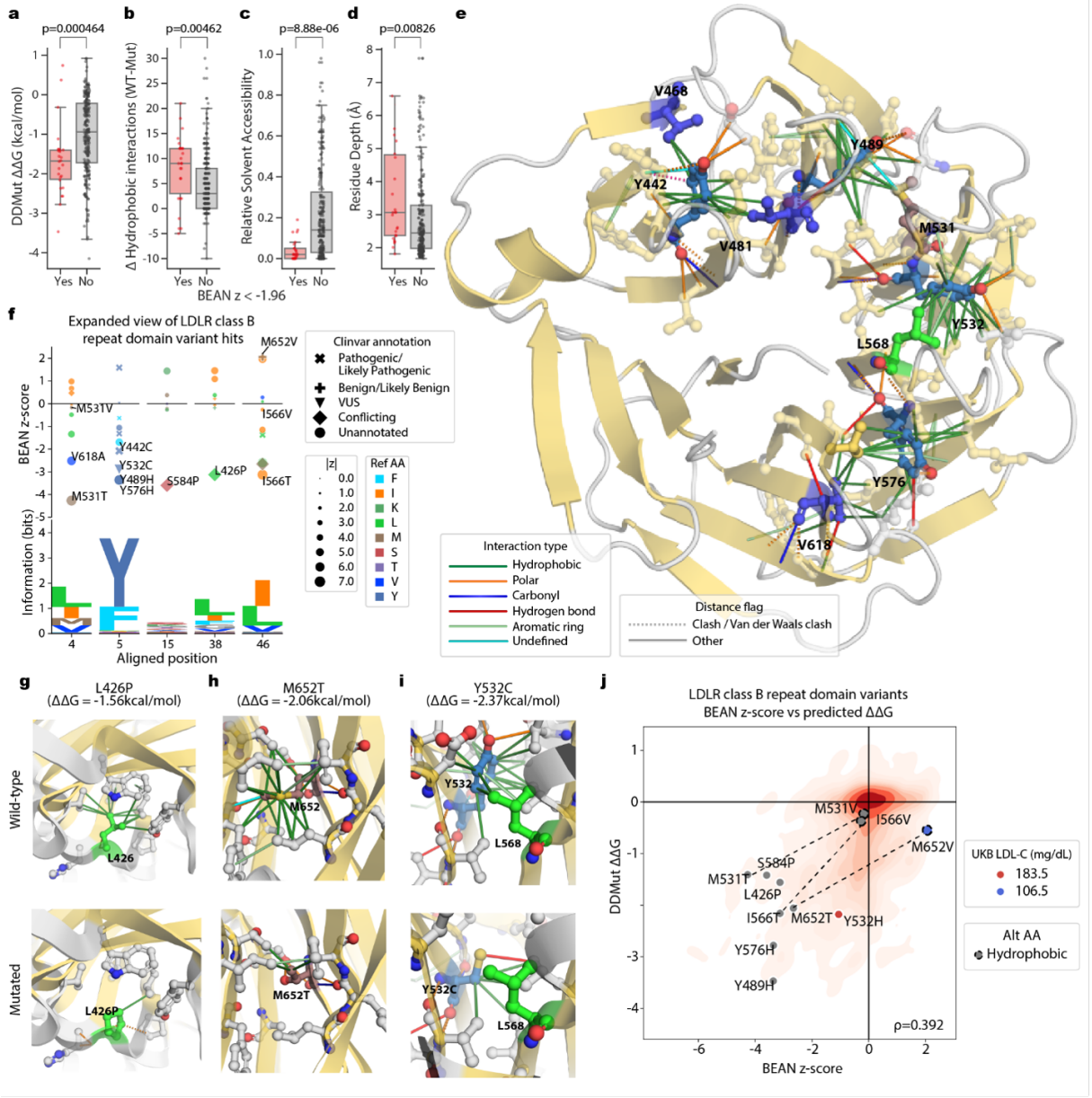


Figure 6.38. Deleterious variants in LDLR class B repeats weaken hydrophobic interactions (A)-(D) Boxplots of 26 significant (z ¡ -1.96) and the rest of 259 variants observed in LDLR class B repeats. P-values of two-sided Wilcoxon rank-sum test are denoted. WT; wild-type, Mut; mutated. **(E)** Conserved interactions involving tyrosine of which mutation showed significant BEAN scores. Simplified interaction types and distance flags as annotated by Arpeggio are shown in the legend. **(E)** BEAN z-scores of positions with conserved hydrophobic residues are shown along with the LDLR class B repeat PFAM HMM logo. **(G)-(I)** Local atomic interaction in wild-type and mutated structure for ClinVar conflicting variants or VUS L426P, M652T, and Y532C. Residues in the variant positions are colored by the reference amino acids. Residues that interact with the variant position are shown. Variant position and interacting residues are colored by elements (O: red, N: blue, S: yellow). **(J)** Contour plot of BEAN z-score against ΔΔG predicted by DDMut for 872 missense variants. Positions with distinct observed missense variants that disrupt and conserve hydrophobic sidechains are connected by dashed line.

deleterious (BEAN z=-2.65), we find no functional disruption from the hydrophobicity-conserving M652V variant (BEAN z=+2.06). This analysis is supported clinically, as M652V is designated in ClinVar as "Likely Benign," and the average UKB carrier LDL-C is below average (106mg/dL). In summary, structural analysis of rare LDLR variants identified by BEAN provides a basis for the missense variant impact through affecting structural integrity of LDLR, highlighting a central role for hydrophobic interactions that hold together adjacent beta blades of the LDLR class B repeat domain

References

(2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, **45**(11), 1274–1283.

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.

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.

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.

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.

Anzalone, A. V., Koblan, L. W., and Liu, D. R. (2020). Genome editing with crispr–cas nucleases, base editors, transposases and prime editors. *Nature biotechnology*, **38**(7), 824–844.

Araya, C. L. and Fowler, D. M. (2011). Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology*, **29**(9), 435–442.

Arbab, M., Shen, M. W., Mok, B., Wilson, C., Matuszek, Ż., Cassa, C. A., and Liu, D. R. (2020). Determinants of base editing outcomes from target library analysis and machine learning. *Cell*, **182**(2), 463–480.

Arbab, M., Matuszek, Z., Kray, K. M., Du, A., Newby, G. A., Blatnik, A. J., Raguram, A., Richter, M. F., Zhao, K. T., Levy, J. M., *et al.* (2023). Base editing rescue of spinal muscular atrophy in cells and in mice. *Science*, **380**(6642), eadg6518.

Arvey, A., Agius, P., Noble, W. S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research*, **22**(9), 1723–1734.

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.

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.

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.

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.

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.

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.

Bialk, P., Rivera-Torres, N., Strouse, B., and Kmiec, E. B. (2015). Regulation of gene editing activity directed by single-stranded oligonucleotides and crispr/cas9 systems. *PloS one*, **10**(6), e0129308.

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

Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., *et al.* (2022). High-content crispr screening. *Nature Reviews Methods Primers*, **2**(1), 8.

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

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.

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.

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.

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.

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., 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.

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, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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.

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.

Clarke, S. L., Tcheandjieu, C., Hilliard, A. T., Lee, K. M., Lynch, J., Chang, K.-M., Miller, D., Knowles, J. W., O'Donnell, C., Tsao, P. S., *et al.* (2022). Coronary artery disease risk of familial hypercholesterolemia genetic variants independent of clinically observed longitudinal cholesterol exposure. *Circulation: Genomic and Precision Medicine*, **15**(2), e003501.

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.

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

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.

Currin, K. W., Erdos, M. R., Narisu, N., Rai, V., Vadlamudi, S., Perrin, H. J., Idol, J. R., Yan, T., Albanus, R. D., Broadaway, K. A., *et al.* (2021). Genetic effects on liver chromatin accessibility identify disease regulatory variants. *The American Journal of Human Genetics*, **108**(7), 1169–1189.

Daley, T. P., Lin, Z., Lin, X., Liu, Y., Wong, W. H., and Qi, L. S. (2018). Crisphiermix: a hierarchical mixture model for crispr pooled screens. *Genome Biology*, **19**, 1–13.

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

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., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., *et al.* (2012). Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, **482**(7385), 390–394.

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.

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.

Ding, X., Seebeck, T., Feng, Y., Jiang, Y., Davis, G. D., and Chen, F. (2019). Improving crispr-cas9 genome editing efficiency by fusion with chromatin-modulating peptides. *The CRISPR journal*, **2**(1), 51–63.

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., 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.

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.

Emmer, B. T., Sherman, E. J., Lascuna, P. J., Graham, S. E., Willer, C. J., and Ginsburg, D. (2021). Genome-scale crispr screening for modifiers of cellular ldl uptake. *PLoS Genetics*, **17**(1), e1009285.

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.

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.

Fass, D., Blacklow, S., Kim, P. S., and Berger, J. M. (1997). Molecular basis of familial hypercholesterolaemia from structure of ldl receptor module. *Nature*, **388**(6643), 691–693.

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.

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.

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.

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.

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M., and Joung, J. K. (2014). Improving crispr-cas nuclease specificity using truncated guide rnas. *Nature biotechnology*, **32**(3), 279–284.

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.

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*.

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., 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.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. *Nature protocols*, **11**(10), 1782–1787.

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.

Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., Cooper, G. M., Reddy, T. E., Crawford, G. E., and Myers, R. M. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular cell*, **52**(1), 25–36.

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.

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.

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.

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.

Guo, Y. A., Chang, M. M., Huang, W., Ooi, W. F., Xing, M., Tan, P., and Skanderup, A. J. (2018). Mutation hotspots at ctcf binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature communications*, **9**(1), 1–14.

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

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., 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.

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.

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.

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

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.

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.

Inoue, F. and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, **106**(3), 159–164.

Jeon, H. and Blacklow, S. C. (2005). Structure and physiologic function of the low-density lipoprotein receptor. *Annu. Rev. Biochem.*, **74**, 535–562.

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.

Jing, Z., Liu, Y., Dong, M., Hu, S., and Huang, S. (2004). Identification of the dna binding element of the human znf333 protein. *BMB Reports*, **37**(6), 663–670.

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.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *Journal of molecular biology*, **429**(3), 365–371.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.* (2021). Highly accurate protein structure prediction with alphafold. *Nature*, **596**(7873), 583–589.

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.

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.

Kim, M. K. and Kang, Y. K. (1999). Positional preference of proline in α-helices. *Protein Science*, **8**(7), 1492–1499.

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.

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.

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.

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.

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.

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., 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., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-Alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A., *et al.* (2016). Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research*, **44**(D1), D116–D125.

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.

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.

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.

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.

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. 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.

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.

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., 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.

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, G., Yin, K., Zhang, Q., Gao, C., and Qiu, J.-L. (2019). Modulating chromatin accessibility by transactivation and targeting proximal dsgrnas enhances cas9 editing efficiency in vivo. *Genome biology*, **20**, 1–11.

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.

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.

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.

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.

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.

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.

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.

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.

Myers, R. M., Tilly, K., and Maniatis, T. (1986). Fine structure genetic analysis of a β-globin promoter. *Science*, **232**(4750), 613–618.

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.

Nobles, C. L., Reddy, S., Salas-McKee, J., Liu, X., June, C. H., Melenhorst, J. J., Davis, M. M., Zhao, Y., and Bushman, F. D. (2019). iguide: an improved pipeline for analyzing crispr cleavage specificity. *Genome biology*, **20**, 1–6.

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.

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).

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

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.

Pavloušková, J., Réblová, K., Tichỳ, L., Freiberger, T., and Fajkusová, L. (2016). Functional analysis of the p.(leu15pro) and p.(gly20arg) sequence changes in the signal sequence of ldl receptor. *Atherosclerosis*, **250**, 9–14.

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

Pena, F., Jansens, A., van Zadelhoff, G., and Braakman, I. (2010). Calcium as a crucial cofactor for low density lipoprotein receptor folding in the endoplasmic reticulum. *Journal of Biological Chemistry*, **285**(12), 8656–8664.

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., 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.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, **21**(3), 447–455.

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.

Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A., and Mathelier, A. (2021). Unibind: maps of high-confidence direct tf-dna interactions across nine species. *BMC genomics*, **22**(1), 1–17.

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.

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.

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.

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.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**(4716), 834–838.

Russell, D. W., Brown, M. S., and Goldstein, J. L. (1989). Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. *Journal of Biological Chemistry*, **264**(36), 21682–21688.

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.

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.

Schep, R., Brinkman, E. K., Leemans, C., Vergara, X., van der Weide, R. H., Morris, B., van Schaik, T., Manzo, S. G., Peric-Hupkes, D., van den Berg, J., *et al.* (2021). Impact of chromatin context on cas9-induced dna double-strand break repair pathway balance. *Molecular cell*, **81**(10), 2216–2230.

Schmid-Burgk, J. L., Gao, L., Li, D., Gardner, Z., Strecker, J., Lash, B., and Zhang, F. (2020). Highly parallel profiling of cas9 variant specificity. *Molecular cell*, **78**(4), 794–800.

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.

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

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.

Shin, H. R., See, J.-E., Kweon, J., Kim, H. S., Sung, G.-J., Park, S., Jang, A.-H., Jang, G., Choi, K.-C., Kim, I., *et al.* (2021). Small-molecule inhibitors of histone deacetylase improve crispr-based adenine base editing. *Nucleic Acids Research*, **49**(4), 2390–2399.

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.

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.

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.

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

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.

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

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.

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.

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

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**(8), 467–484.

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.

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.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, **9**(1), 5233.

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.

Vakulskas, C. A., Dever, D. P., Rettig, G. R., Turk, R., Jacobi, A. M., Collingwood, M. A., Bode, N. M., McNeill, M. S., Yan, S., Camarena, J., *et al.* (2018). A high-fidelity cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature medicine*, **24**(8), 1216–1224.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., *et al.* (2022). Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, **50**(D1), D439–D444.

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.

Von Heijne, G. (1985). Signal sequences: the limits of variation. *Journal of molecular biology*, **184**(1), 99–105.

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.

Vorontsov, I. E., Khimulya, G., Lukianova, E. N., Nikolaeva, D. D., Eliseeva, I. A., Kulakovskiy, I. V., and Makeev, V. J. (2016). Negative selection maintains transcription factor binding motifs in human cancer. *BMC genomics*, **17**, 263–276.

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, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **82**(5), 1273–1300.

Wang, R., Lin, D.-Y., and Jiang, Y. (2022). 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, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the crispr-cas9 system. *Science*, **343**(6166), 80–84.

Ward, L. D. and Kellis, M. (2012). Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, **40**(D1), D930–D934.

Webb, B. and Sali, A. (2016). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, **54**(1), 5–6.

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., 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.

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., Van De Geijn, B., Reshef, Y., Márquez-Luna, C., *et al.* (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature genetics*, **52**(12), 1355–1363.

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.

Witzgall, R., O'Leary, E., Leaf, A., Onaldi, D., and Bonventre, J. V. (1994). The krüppel-associated box-a (krab-a) domain of zinc finger proteins mediates transcriptional repression. *Proceedings of the National Academy of Sciences*, **91**(10), 4514–4518.

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.

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.

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, C., Ma, Z., Wang, K., Dong, X., Huang, M., Li, Y., Zhu, X., Li, J., Cheng, Z., Bi, C., *et al.* (2023). Hmgn1 enhances crispr-directed dual-function a-to-g and c-to-g base editing. *Nature Communications*, **14**(1), 2430.

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, 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., 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, T., Fife, J., Adzhubey, I., Sherwood, R., and Cassa, C. (2023). Joint estimation and imputation of variant functional effects using high throughput assay data. *medRxiv*, pages 2023–01.

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.

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.

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.

Zhou, Y., Pan, Q., Pires, D. E., Rodrigues, C. H., and Ascher, D. B. (2023). Ddmut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Research*, page gkad472.

Zia, A. and Moses, A. M. (2012). Towards a theoretical understanding of false positives in dna motif finding. *BMC bioinformatics*, **13**(1), 1–9.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., *et al.* (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, **3**(1), 1–26.

Zuo, C., Shin, S., and Keleş, S. (2015). atsnp: transcription factor binding affinity testing for regulatory snp detection. *Bioinformatics*, **31**(20), 3353–3355.