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 (**?**). 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 (**?**). 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 (**?**). 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 (**?**) and MotifRaptor (**?**). GRAFIMO searches for occurrences of known TFBSs genome graphs (**?**), 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 (**?**), 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 (**???**). In human, approximately 1600 proteins are recognized as TFs (**?**). 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 (**?**). TFs exhibit a modular structure, that is divided into three distinct domains (**?**). (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 (**?**). (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 (**?**), 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 (**?**)), known as transcription factor binding sites (TFBSs), they finely regulate gene expression in a cell-specific manner. TFBSs are located within gene promoters (**?**) or in more distant regulatory elements such as enhancers, silencers, or insulators (**???**). 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 (**??**). 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 (**???**). 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 on the complex mechanisms governing gene expression and the cell environment.

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

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*) (**?**). Early methods, like electrophoretic mobility shift assay (EMSA) (**?**) or footprinting (**?**), 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 (**???**), 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 (**??????**). 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) (**?**) 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 the binding site positions. Therefore, several alternative motif models have been proposed (**???**) , 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.

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. *In vivo* and *in vitro* experimental assays to identify and validate transcription factor binding sites

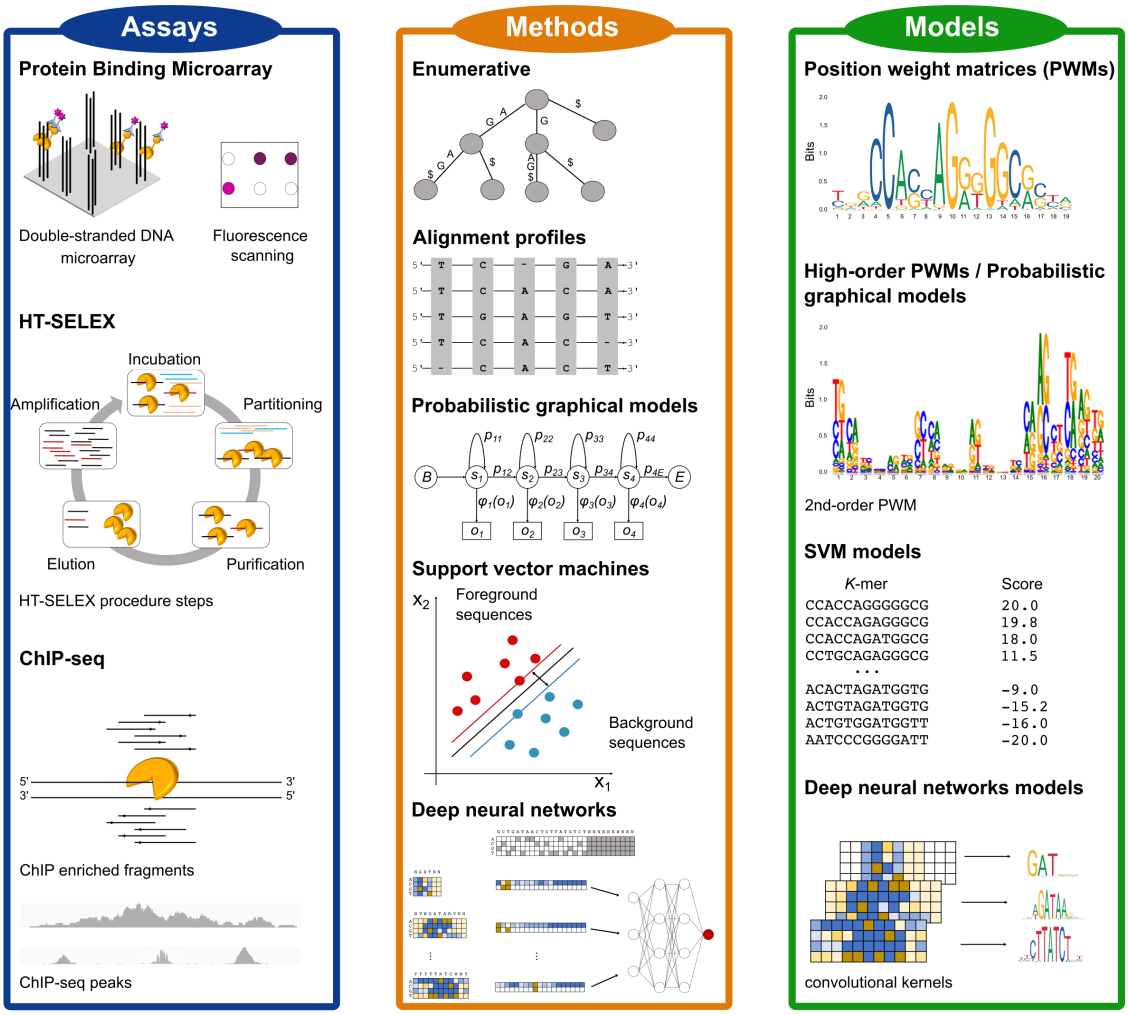


Figure 2.2. Experimental and computational methods to discover TFBS and popular models to represent binding site motifs. Protein binding microarray (PBM), HT-SELEX and ChIP-seq have become the most popular assays to determine TF binding preferences and identify their target sites (TFBS) in recent years.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

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 (**?**) (**Fig.2.2** and **Table 2.1**). Early studies on TF binding focused their analysis on gene promoters (**?**) and employed in vitro methods, such as Electro-Mobility Shift Assay (EMSA) (**?**) or DNase footprinting (**?**). 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 (**?**). 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 advantages: (i) they do not require any prior knowledge on the binding site sequence (**??**) and (ii) produce datasets of thousands of bound sequences allowing a better characterization of TF binding preferences (**?**). Protein binding microarrays (PBMs) (**??**) 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 (**??**). PBM analysis is usually constrained to binding sites ∼10–12 bp long. HT-SELEX (**??**) 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 (**??**). SELEX does not require any prior knowledge on the target sites of the investigated factor (**?**). 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 (**?**). 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 (**??**)) or sequencing (ChIP-seq (**??**)). To locate the binding regions, the recovered DNA fragments are mapped onto the genome. After ChIP-seq reads mapping, peak calling algorithms (**???**) 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 (**?**). 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 (**?**). (ii) ChIP-seq peaks may be false positives, recovered because of poor antibody quality (**?**). (iii) ChIP-seq returns low-resolution datasets, whose sequences include several nucleotides flanking the target TFBS. ChIP-exo (**?**) 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 (**??**) 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.

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 (**?**). 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 (**??**). 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 (**???**). In a recent study (**?**), 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 (**?**), PWMs (**??**), high-order PWMs (**??**),, SVM-based (**?**) and deep neural network-based (**?**) 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 (**?**), 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) (**?**). RE quantifies the difference between computed nucleotide frequencies and those obtained from aligning random sequences. PWMs are visualized as logos (**?**), 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) (**????**). 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 (**?**). 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 (**?**) 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 (**?**). Many frameworks visualize the discovered motifs as PWMs, computed aligning the sequences activating the convolutional kernels of the DNN (**?**). 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 (**?**) 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 (**?**) 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 $s \in S$ (**?**) or restricting mismatching locations to specific motif positions (**?**). However, mismatches can occur at any motif position. Weeder (**??**) and SMILE (**?**) proposed using suffix trees (STs) (**?**) 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 (**?**). Therefore, more efficient approaches specifically tailored to work on large datasets were proposed. MDscan (**?**) and Amadeus (**?**) 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 (**?**) 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 (**???**) 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.

DREME (**?**) searches motif candidate sequences using regular expressions. To statistically evaluate the discovered motifs, DREME employs the Fisher's exact test, comparing the number of sequences in S and B, in which the motifs occur. Trawler, HOMER and STREME (**???**) use STs similarly SMILE or Weeder. Trawler enumerates all sequence patterns using STs indexing S and B, and measures patterns frequencies to find overrepresented sequences in the input dataset. To match motif sequences, Trawler employs degenerate consensuses, allowing mismatching positions. To statistically evaluate the prioritized patterns, Trawlers uses the z-scores, derived from the normal-approximation to the binomial distribution. Since the function does not correct for the effect of overlapping motif instances, the computations are significantly faster. The similar and significant patterns are clustered to form motifs. HOMER algorithm was specifically designed to discover TFBS motifs in ChIP-seq data. HOMER indexes the foreground and background datasets using STs and searches for k-mers overrepresented in S, through approximate pattern matching on the ST. HOMER employs the hypergeometric or binomial test to statistically evaluate each motif candidate. Using the hypergeometric or the binomial tests reduces the running time, by avoiding the explicit count of k-mers frequencies. STREME builds a ST from S and B, and counts the number of exact matches of seed words of different lengths in both datasets, evaluating the statistical significance of their enrichment using Fisher's exact test or the binomial distribution. Then, STREME counts the number of the approximate matches of the most significant words on the ST. The prioritized words are then grouped to derive the corresponding motifs. Importantly, STREME requires one single tree visit to discover motifs of different lengths, significantly speeding-up the discovery of different TFBS motifs.

Alignment-based motif discovery algorithms compute alignment profiles to describe motif binding preferences (**Fig.2.2 (B)**), avoiding the exhaustive k-mers enumeration. The rationale is to build an alignment from k-mers selected from the input dataset S, and score the resulting profile through appro-

priate measures, like nucleotide conservation, information content, or profile statistical significance. Motif statistical significance is based on the probability of obtaining the same alignment from random sequences or from a background dataset B. Alignment-based motif discovery algorithms usually assume that the motif length $|M|$ is known *a priori*. For alignment-based algorithms motif discovery can be formalized as a combinatorial problem. Let us assume $|M| = k$, the goal is to find the best alignment profile built combining k-mers from S, according to a scoring criterion. The best alignments are then used to derive the corresponding PWMs. Most alignment-based algorithms assume that each sequence in S contains one or zero binding sites. 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 dataset, alignment-based algorithms explore the solution space using heuristics, such as greedy (**?**), expectation-maximization (EM) (**?**), stochastic (e.g. Gibbs sampling) (**?**), or genetic algorithms (**?**). CONSENSUS (**?**) proposed a greedy approach to incrementally build the motif alignment profiles. The problem is initially solved on two sequences, then is progressively solved by adding the remaining sequences $s \in S$ one by one. The MEME algorithm (**???**) proposed a different strategy to explore the solution space by iteratively refining an initial profile, substituting some k-mers in the profile with others more likely to produce better solutions. By employing an EM strategy, MEME evaluates how well each k-mer in $s \in S$ fits the current alignment profile, rather than a background model. The algorithm has two main steps: the E-step and the M-step. During the E-step, MEME computes a likelihood score for each k-mer in S, using the current alignment profile. In the M-step, MEME assigns a weight to each k-mers in the current profile proportional to the scores computed during the E-step and updates the alignment by substituting low scoring k-mers with others better fitting the current profile. However, the EM algorithm can prematurely converge to local maxima and convergence depends on the algorithm starting conditions. Stochastic optimization strategies like Gibbs sampling (**?**) address these limitations. Given an initial profile built with k-mers randomly selected from each $s \in S$, at each iteration the algorithm removes from the profile the k-mer coming from a certain s. Then, the algorithm assigns a likelihood score to each k-mer in s using the modified profile. Similarly to MEME, the score describes how well each k-mer fits the profile rather than a background model. Then, a new k-mer from s is chosen to replace the removed subsequence in the profile, with probability proportional to its likelihood score. The procedure is repeated until a fixed number of iterations or no further updates applied to the profile. Therefore, while EM optimization strategies like MEME update the profile selecting the k-mers deterministically according to how well they fit the alignment, Gibbs sampling algorithms optimize the profile selecting the subsequences to remove and insert using a stochastic approach. However, the basic Gibbs sampling algorithm assumes that each sequence contains exactly one binding site. In (**?**) the authors extended the algorithm to consider sequences containing one, multiple, or no binding sites. AlignACE (**?**) and ANN-spec (**?**) furtherly modified the Gibbs sampling algorithm, enabling the simultaneous search for TFBS on both strands. Moreover, ANN-spec coupled the stochastic motif search with a perceptron artificial neural network learning the best alignment from the binding specificities observed in S. BioProspector (**?**) and MotifSampler (**?**) used third order Markov models as background, improving the predictive performance of Gibbs sampling. GLAM (**?**) modified the Gibbs sampling strategy to estimate the optimal alignment length, and employed simulated annealing for profile optimization. GLAM algorithm was then furtherly extended to also consider gapped motifs (**?**). Although stochastic and EM are the most popular optimization strategies for alignment-based algorithms, researchers also focused on adopting other optimization methods. GADEM (**?**) paired EM local search with genetic algorithms for profile refinement. However, using alignment profiles the solution search space grows exponentially with the size of S. Even employing heuristics, the analysis of the datasets produced by NGS assays rapidly becomes impractical and requires large amounts of time and computational resources (**?**). Therefore, researchers focused on developing new alignment-based motif discovery algorithms tailored to analyze the datasets produced by high-throughput protocols. MEME-ChIP (**?**) and STEME (**?**) improved the original MEME algorithm for the analysis of ChIP datasets. MEME-ChIP focuses the analysis on a random subset of sequences in S, reducing the theoretical size of the solution space. STEME indexes the sequences in S using suffix trees and avoids to evaluate the likelihood of all possible k-mers during MEME's E-step, by employing a branch and bound strategy on the ST. ChIPMunk (**?**) proposed a greedy profile optimization similar to EM, to discover motifs in large ChIP datasets, while accounting for ChIP peaks shape. ChIPMunk explores the solution space to find the alignment maximizing the profile discrete information content (**?**), and uses peaks shape to weight the contribution of each sequence $s \in S$ to the motif. XXmotif (**?**) combined enumerative motif discovery with profile refinement, by iteratively selecting sets of k-mers from S maximizing their fitness to the profile, and increasing the motif fitness to S. Similarly, ProSampler (**?**) proposed a highly optimized hybrid method to discover motifs in large ChIP-seq datasets, combining

motif enumeration with Gibbs sampling to refine preliminary motif profiles.

The importance of dependencies between TFBS nucleotides and how including them in motif models would improve the models' performance has been controversially debated (**???**). However, some studies demonstrated the existence of such dependencies in TFBS, between neighboring and non-neighboring nucleotides (**??**). Enumerative and alignment-based algorithms represent motifs without accounting for potential dependencies between the binding site positions, modeling TFBS as PWMs. PWMs can be extended to count the frequency di- or tri-nucleotides, producing high-order PWMs, like dinucleotide weight matrices (DWMs) (**?**). Dimont (**?**) and diChIPMunk (**?**) employed DWMs to discover and represent motifs. However, dependencies may exist also between non-neighboring positions and not between neighboring nucleotides. Therefore, the models could learn dependencies not in the data, often overfitting the training set S. Probabilistic graphical models (**Fig.2.2 (B)**) like Bayesian networks (BNs) or Markov models (MMs) provide flexible and efficient frameworks to capture and encode dependencies between TFBS nucleotides. Barash and coworkers (**?**) proposed to model TFBS motifs as BNs trained through an EM strategy. Importantly, BNs can capture dependencies between neighboring and non-neighboring motif positions. However, the model assumes the same order of dependence throughout the entire binding site. Similarly, in (**?**) the authors introduced VOBN models representing TFBS as BNs, but accounting for variable orders of dependencies between neighboring and non-neighboring TFBS positions. However, BNs training is often computationally demanding and requires large amounts of data to avoid the model to overfit S. MMs and Hidden Markov models (HMMs) provide efficient frameworks to include dependencies between motif positions, and more scalable and efficient training procedures compared to BNs. TFFMs (**?**) proposed a HMM-based model capturing dinucleotide dependencies between neighboring motif positions. TFFM models also capture the properties of the sequences flanking the binding site and accommodate changes in the motif length, by employing the background and insertion/deletion states, respectively. TFFMs are trained using a set of previously prioritized motif candidate sequences and employing the Baum-Welch algorithm. TFFM models assume the same order of dependency between neighboring positions across the whole TFBS. Slim models (**?**) introduced a framework to learn neighboring and non-neighboring dependencies between TFBS nucleotides, using different orders of dependency. Dependency orders are pruned in a data-driven fashion, establishing the TFBS positions on which the motif nucleotides depend on. The models are trained using the motif candidates prioritized using Dimont. Discrover (**?**) proposed a discriminative motif discovery method designed to analyze ChIP-seq datasets and modeling TFBS as HMMs. Given S and B, Discrover searches for overrepresented motifs in S using regular expressions. Then, an initial HMM is trained using the prioritized motif candidates as seeds. To identify good seeds, Discrover uses different objective functions such as likelihood, or mutual information. The initial HMM is then iteratively refined through a gradient optimization of likelihood. In (**?**) the authors proposed a method to discover CTCF (**?**) motif using variable-order MMs to capture different orders of dependency between neighboring nucleotides. The model is trained using an EM strategy iteratively refining the MM parameters, to optimize its fitness to the input dataset S. BaMMotif (**?**) introduced an efficient method to discover TFBS motifs in large datasets, employing a Bayesian approach to train Markov models. To avoid model overfitting, the motif model is trained using the conditional probabilities of $(q-1)$th order as priors for the qth order conditional probabilities. The algorithm iteratively optimizes the model parameters until convergence through an EM strategy. In the E-step, the algorithm estimates the probability that each position of each $s \in S$ is a motif starting position, using the current model. During the M-step, BaMMotif uses the motif candidates identified in the previous step to refine the model. Moreover, the algorithms dynamically adapt the model parameters during training to capture variable orders of dependencies between motif positions. Recently, the authors developed a faster and improved of the algorithm (**?**).

Support Vector Machines (SVMs) (**?**) have been successfully applied to different problems in computational biology (**?**), including TFBS motif discovery (**Fig.2.2 (B)**). The general idea is to decompose the bound sequences (foreground dataset S) and the unbound sequences (background dataset B) in sequence elements of length k (k-mers). The k-mers frequencies are then used as features to train a sequence similarity kernel (**?**), to discriminate between bound and unbound sequences. 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. Although early k-mer based kernels (**???**) were originally developed to tackle protein sequence homology problems, recently several SVM-based algorithms extending the original kernels were proposed to discover TFBS motifs. Kmer-SVM (**??**) proposed a SVM-based method to discover TFBS motifs in ChIP-seq datasets. Kmer-SVM employs the spectrum kernel (**?**), which counts the exact matches for all contiguous k-mers in S and B, building the k-mers feature space. By using a trie (**?**) indexing S and B, the algorithm runs in linear time. Although TFBS

motifs contain highly degenerate non-informative positions, kmer-SVM does not explicitly accommodate such sequence variability by counting the frequency of exact k-mers matches. The mismatch and wildcard kernels (**??**) introduced k-mers frequencies counting allowing a fixed number of mismatching positions for each k-mer. Agius and coworkers (**?**) proposed the di-mismatch kernel to discover TFBS on PBM and ChIP-seq datasets. The di-mismatch kernel is a first order Markov mismatch kernel based on dinucleotides alphabet, handling sequence variability and accounting for dependencies between neighboring nucleotides. However, these methods generally employ small k ($\sim 6-8$), discovering short motifs. Since TFBS lengths range between $6-20$ bp, longer TFBS like CTCF (19bp) cannot be exhaustively characterized by short k-mers, without tiling overlapping k-mers across the motif. Moreover, simply increasing the k often results in sparse feature vectors, overfitting the training dataset. Gapped k-mers (**?**) represent longer motifs as k-mers with gaps in the non informative or degenerate TFBS positions. Gapped k-mers accounts for both variability in the motif sequence and length. Gkm-SVM (**??**) extended kmer-SVM to train SVM kernels employing gapped k-mers as feature. The algorithm considers larger k avoiding overfitting and reducing the dependency of method's performance from the parameters choice. Recently, the algorithm was optimized in LS-GKM (**?**), enabling SVM kernel training on large scale ChIP-seq datasets.

In recent years, Deep Neural Networks (DNNs) have received increasing attention from computational biology community (**??????????**). DNNs provide a powerful framework to learn complex patterns at multiple layers (**?**), exploiting the ever growing omics data (**?**). Recently, convolutional Neural Networks (CNNs) (**?**) have been successfully employed to characterize the complex patterns underlying in vivo TF-DNA interactions (**????**) (**Fig.2.2 (B)**). CNNs were originally introduced in computer vision to solve the two-class image classification task (**???**). CNNs adaptively learn the important features from the data while training the model. CNNs map input data to high-dimensional representations by applying nonlinear transformations and simplifying classification tasks (**?**). By representing sequences as matrices through one hot encoding transformation, CNNs can analyze genomic data. Therefore, sequences are represented as 1-D images with four associated channels (`A`, `C`, `G`, `T`), instead of 2-D images with the canonical color channels (**?**). Discovering and TFBS in genomic sequences becomes a problem like two-class image classification. Generally, CNNs architecture in motif discovery context has one or more sets of four layers: the convolutional layer, the max-pooling layer, the fully connected NN layer, and the output layer (**?**). The convolutional stage scans windows of the input sequences $s \in S$ with a set of convolutional filters (motif kernels). The response values are transferred to the max-pooling layer, which keeps only the maximum kernel responses for each convolutional layer. The maximal responses are then embedded in feature vectors. The NN stage transforms the feature vectors into scalar scores, proportional to the contribution of each site to the TFBS. Often this stage is followed by a backdrop layer, which randomly mask portions of the NN output to avoid model overfitting. The output layer generally consists of two neurons fully connected to the previous layer. The output layer returns the model predictions, identifying potential motif occurrences. DNNs can present multiple sets of each layer. Moreover, the convolutional, pooling, and NN layers have tunable parameters, dynamically learned from the input data. Often CNNs employ backpropagation and parameters update stages after the output layer to iteratively improve the model and optimize its parameters. In motif discovery context, the backpropagation and parameters tuning stages update the discovered motifs, response values, and NN weights to improve the model accuracy. The operation is performed until convergence or after a determined number of iterations. DeepBind (**?**) employed CNNs to discover TFBS motifs on ChIP-seq, HT-SELEX, and PBM datasets. Deepbind introduced two variants to basic CNN architecture: a rectification layer, between the convolutional and pooling stages, and the averaging and maximization steps in the pooling layer. The rectification stage isolates those sequences sites showing good matches to the convolutional kernel, shifting the response by some nucleotides and masking bad matching positions. In the pooling stage, DeepBind computes the maximum and average kernel responses, discovering cumulative effects of short motifs and the locations of longer TFBS, respectively. To visualize the discovered motifs, DeepBind computes weighted ensembles of PWMs, obtained aligning the subsequences in S activating the kernels response. The alignment is centered around the position with the maximum response. Basset (**?**) added three convolutional layers after the pooling stage, followed by two fully connected NNs. Basset randomly initializes the starting parameters. Basset adaptively updates the parameters using backpropagation. Basset represents the discovered motifs as PWMs, computed grouping and aligning the sequences activating the kernels response in the first convolutional layer. However, TF-DNA interactions are also governed by multiple binding subregions (long-term interactions), and by interactions between the nucleotides and high-order structures of TFs (short-term interactions). CNNs cannot capture such long-term and short-term dependencies. Recurrent Neural Networks (RNNs) were introduced to model sequential signals exhibiting stationary features over time. Long Short Term Memory Networks (LSTMs) (**?**), a variant of RNNs, efficiently capture long-term

and short-term dependencies, by learning the positional dynamics of sequential signals. Bi-directional long/short-term memory (BLSTMs) networks are variants of standard LSTMs, which combine the outputs of two RNNs: one analyzing sequential data from left to right and the other from right to left (in genomic context the forward and reverse strands, respectively). DanQ (**?**) proposed a hybrid architecture, employing CNNs and BLSTMs. The BLSTM layer replaced the fully connected NN, followed by a dense layer of rectified linear unit and a multi-task sigmoid output stage (**?**). Placing the BLSTM layer between the pooling and the dense rectified stages, DanQ captures the positional dynamics of input sequences. Moreover, DanQ captures complex motif grammars determining the spatial arrangements and frequencies of combinations of TFBS in vivo, which are fundamental characteristics of genomic regulatory elements like enhancers. DanQ parameters can be initialized with random values or known motifs. The discovered motifs are represented as PWMs, computed aligning the sequences activating the kernels. FactorNet (**?**) extended DanQ framework by incorporating additional features during model training like DNase-seq signals. To avoid overfitting and reduce training complexity, FactorNet employs a Siamese architecture accounting for the reverse complement. The siamese architecture applies identical networks sharing the weights, to the forward and reverse strands. This ensures that both networks return the same output and reduces the amount of the required training data. DeeperBind (**?**) proposed a hybrid CNN-LSTMs architecture, capturing long-term and short-term positional dependencies. Importantly, DeeperBind omits the pooling layer, to avoid losing positional information of subregions activating the kernels in the convolutional stage.

2.1.3 Computational Transcription Factor binding sites models

Different computational models have been proposed to represent TFBS structure and binding preferences (**Fig.2.2 (C)**). Currently, the most popular models are *consensus sequences*, *PWMs*, *probabilistic graphical models* (e.g. MMs and HMMs), *k-mer-based* and *DNN-based models*. Generally, enumerative and alignment-based algorithms employ consensus sequences (**?**) or PWMs (**??**). Consensus sequences denote at each motif position the most frequent nucleotide observed in the set of prioritized sequences, providing a sequence summarizing the discovered TFBS. Although TFBS have conserved positions not tolerant to mutations (**?**), other binding site locations admit alternative nucleotides. Degenerate consensus accommodate ambiguous motif positions using the 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 the contribution of each motif position to the binding events. 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 assessed through nucleotide conservation (similarity between the aligned k-mers), and difference from PWMs computed aligning randomly selected k-mers. Relative Entropy (RE) (**?**) is often employed to assess PWMs statistical significance. RE measures how much the computed nucleotide frequencies would differ from values obtained aligning random sequences. PWMs are visualized as logos (**?**), where the height of each motif nucleotide is proportional to its RE. PWMs models for many TFBS are available in several motif databases, like JASPAR, TRANSFAC, or HOCOMOCO (**??????**). However, PWMs assume independence between motif positions. Probabilistic graphical models, like Dinucleotide weight matrices (DWMs), high-order PWMs, BNs, MMs, or HMMs, represent TFBS integrating dependency between the motif nucleotides (**????**). Interestingly, JASPAR and HOCOMOCO provide also DWMs for many TFBS motifs. DWMs and high-order PWMs are visualized as logos with q-mers replacing the single nucleotides, where q is the dependency order between neighboring nucleotides. However, the model complexity and number of parameters exponentially grow with q, often resulting in the model overfitting the input dataset. K-mer-based models represent TFBS listing all k-mers with an associated weight, assigned by the SVM kernel trained on the S and B datasets. The weight of each k-mer reflects its contribution to the motif definition. Importantly, k-mers indirectly capture k-th order dependencies between neighboring nucleotides. Breaking motifs in k-mers prioritize the informative subsequences governing TF-DNA binding events, without assuming fixed spacing between informative sequence features. However, simple k-mer-based models are limited to consider short k ($\sim$ 10bp) and cannot explicitely represent longer motifs. Gapped k-mers addressed this limitation (see **Section 2.1.2**), handling longer TFBS and sequence degeneration in noninformative motif positions. Since k-mers are features that are either present or absent, k-mer-based models can be trained on small datasets (**?**), unlike PWMs which require large datasets to optimize their parameters (**?**). To visualize the discovered motifs, k-mer-based models are reduced to PWMs computed aligning the informative k-mers. Many studies showed that in vivo TF-DNA binding is influenced by diverse facets of genomic sequences, such as chromatin accessibility (**?**) and local GC content (**?**). DNN-based models integrate the diverse,

complex and hierarchical patterns governing TF-DNA binding events from the input nucleotide sequences. Although DNN-based are accurate and powerful models, their "black box" nature is a major limitation, and the discovered motifs cannot be easily visualized and interpreted (**?**). Many frameworks address the issue visualizing the discovered motifs as PWMs, computed aligning the sequences activating the convolutional kernels (**?**). 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, reducing the model interpretability. DeepLIFT (**?**) addressed the issue proposing a method assigning 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 (**?**) extended the idea by clustering and aggregating the discovered motifs, using the importance scores assigned to the kernels. However, computing interpretable models without losing some of the information learned in the DNN is still an open challenge.

Motif Graphs

Several studies showed that TFs present population-specific (**?**), cell-type-specific (**??**), and even individual-specific (**?**) binding sites. Algorithms analyzing only the reference genome would provide general models, which could return wrong TFBS predictions when analyzing personal genomic sequences. Motif models integrating data from different populations, cell-types, or individuals would recover more reliable TFBS occurrence predictions (**Fig.?? (A)**). Moreover, such models would better predict the consequences of noncoding variants on TF-DNA binding events, and they could encode variable orders of nucleotide dependencies. Importantly, graph data structures are often interpretable and intuitive. With this aims in mind we developed the Motif Graph, a framework integrating the advantages of *probabilistic graphical models-based* and *SVM-based* motif discovery algorithms, without sacrificing model interpretability. A Motif Graph model G is defined by a set of vertices V, a set of edges E, and a set of paths P. Each vertex $v \in V$ on the graph is labeled with a nucleotide, or mathematically $label(v) \in \{A, C, G, T\}$. Each edge $e \in E$ represents the allowed links between consecutive nucleotides in the TFBS motif. Each path $p \in P$ represents a *haplotype* embedded in G, which correspond to one sequence used to train the Motif Graph model. Currently, the model is limited to encode 1st order dependencies between consecutive

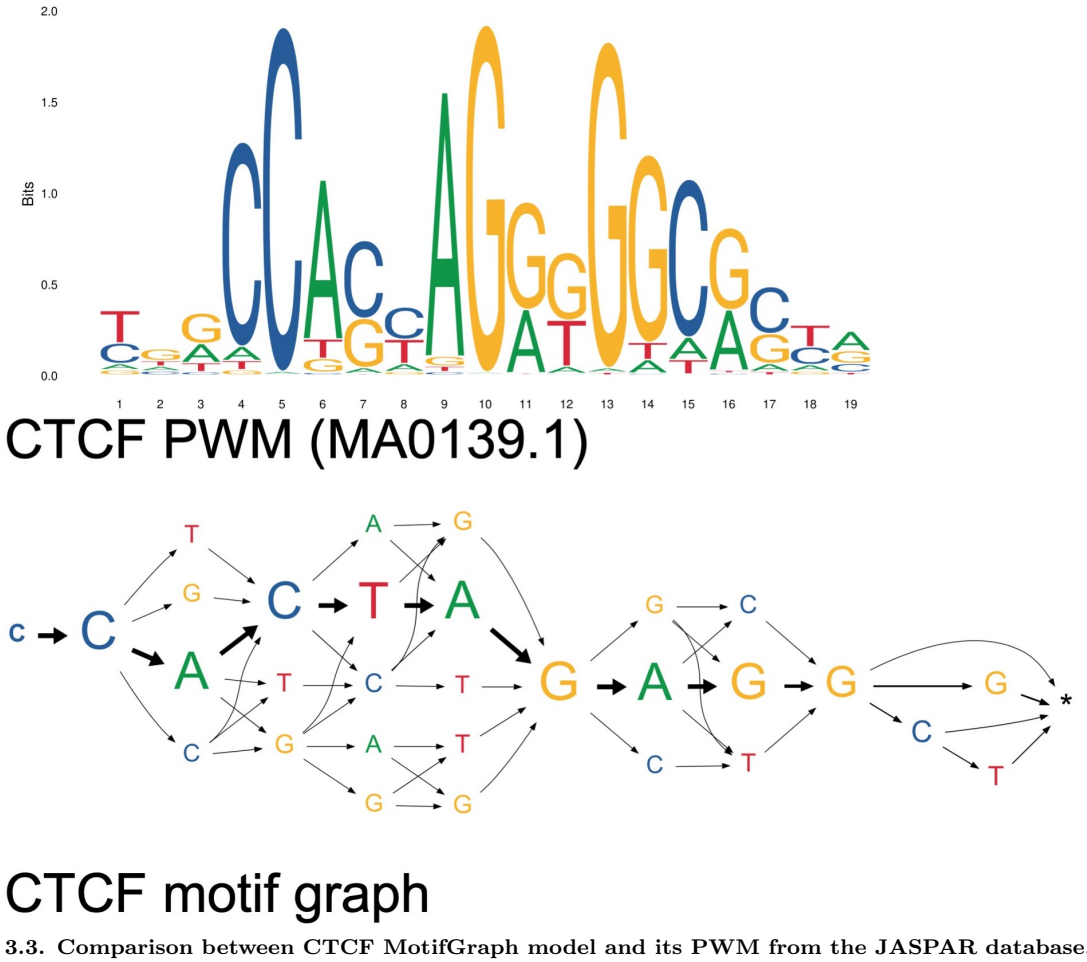


Figure 3.3. Comparison between CTCF MotifGraph model and its PWM from the JASPAR database. On top the CTCF motif available on JASPAR database (MA0139.1). On the bottom the Motif Graph model trained using 100 k-mers obtained from a ChIP-seq experiment trageting CTCF binding site on HepG2 cell line.

nucleotides. However, the model is flexible and accomodate variable length motifs, beside recording the training sequences.

3.1 Motif Graph model construction

Motif Graph motif discovery procedure (**Algorithm ??**) is constitued of two main steps: k-mers prioritization, and graph model training and construction. To prioritize the k-mers Motif Graph currently employs the k-mer based motif discovery procedure implemented in Gkm-SVM (**?**). Therefore, given a positive sequence dataset S and a background dataset B, for each k-mer of length k in S, the algorithm counts the number of matches in S and in B, allowing mismatching positions (see **Section 2.1.2** for further details). Then, a similarity kernel is trained using the recovered k-mers frequencies. The trained kernel assigns to each k-mer in S and in B a weight proportional to its contribution in defining the foreground or the background dataset. The algorithm then ranks the k-mers according to their weight scores. The Motif Graph model is iteratively trained employing a greedy approach, which adds the top ranked k-mers to G incrementally, one by one (see **Algorithm ??** for details). Each k-mer is aligned to the current G to maximize the number of nucleotides matching the current model. While aligning the k-mers to the current Motif Graph model, the algorithm shifts the input k-mer on the right and on the left up to a defined offset number of nucleotides. In our experiments with set the offset to 3. Once built the model, we construct a scoring matrix similar to the widely used $PSSM$. However, our scoring matrix account for 1st order dependencies between nucleotides, recalling the well-known DWMs (see **Section 2.1.3** for details). The scoring matrix is used to assign a likelihood score and classify new sequences as potentially bound or not bound by the investigated factor. In other words, the score describes how likely is the scanned sequence to contain a potential binding site. To score a sequence we slide the scoring matrix along the string employing a procedure similar to classical PWM scanning tools like FIMO (**?**).

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 ← ∅
for kmer in rankedKmers do
    G ← 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 ← insertKmer(G, alignment)
return G
```

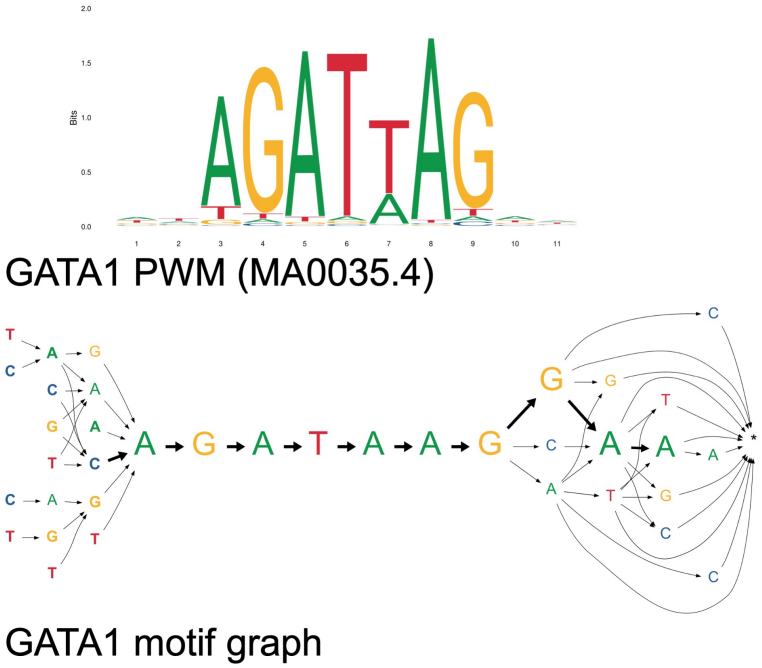


Figure 3.4. Comparison between GATA1 MotifGraph model and its PWM from the JASPAR database. On top the GATA1 motif available on JASPAR database (MA0035.4). On the bottom the Motif Graph model trained using 100 k-mers obtained from a ChIP-seq experiment trageting GATA1 binding site on K562 cell line.

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

3.2 Results

To test our motif discovery algorithm we obtained 10,000 ChIP-seq peak sequences from the ENCODE Project database (**?**) for CTCF and GATA1 Transcription Factors, obtained on the HepG2 and K562 cell lines, respectively. The original ChIP-seq datasets were sorted according to the peaks enrichment score in decreasing order, in order to test our algorithm on reliable peaks. Interestingly, the trained Motif Graph models were closed to the motifs PWM available on the JASPAR database (**?**) for both CTCF and GATA1 (**Fig.??** and **Fig.??**). For both TFs, the Motif Graph models captured the main motif sequence. The main motifs are also enforced by the edge thickness which is proportional to the number of training k-mers supporting each $p \in P$. Then, we tested the discriminative performance of both Motif Graph model, with different number of training k-mers (**Fig.??**), to establish the optimal number of training sequences for the CTCF and GATA1 Motif Graph models. To compare the models discriminative performance, we performed a cross-validation experiment using splitting the S and B dataset in training and testing set (75% and 25%, respectively). We trained the models with 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 training the Motif Graph on 100 k-mers (**Table ??**). 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 ??**). Then, we compared the Motif Graph models discriminative performance against the corresponding PWM and DWM models, recovered from JASPAR (**?**) and HOCOMOCO (**?**) databases, respectively. For CTCF, both the PWM and the DWM models returned better predictive performance than our model (**Fig.?? (A)**). On the other hand, on GATA1 data our model showed better performance than PWMs, but still performed worse than DWMs (**Fig.?? (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 3.3. GATA1 Motif Graph model AUC and F1-scores values with different number of training k-mers.

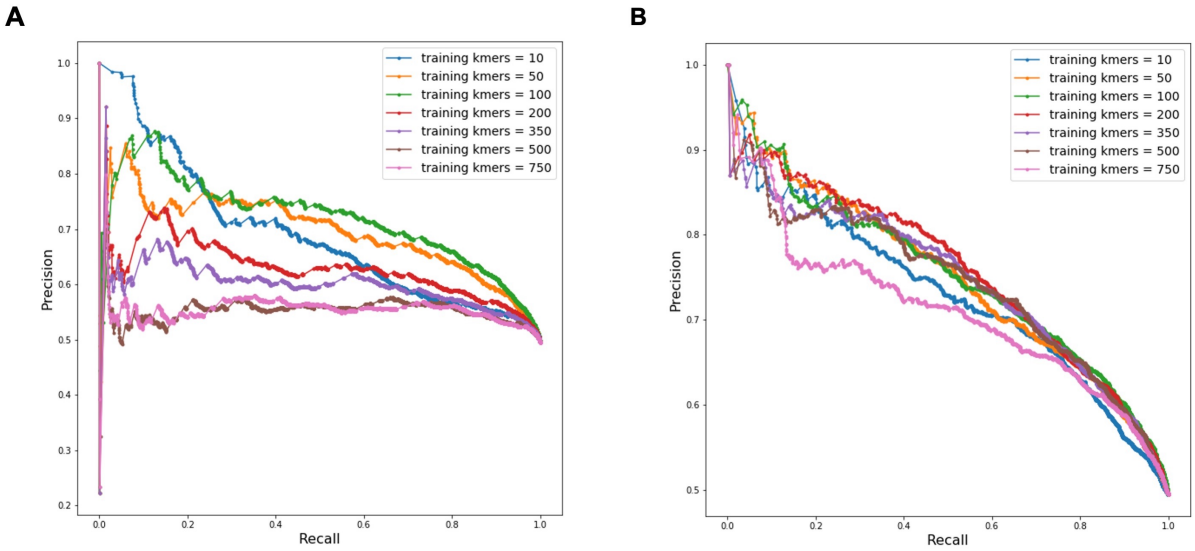


Figure 3.5. Precision-Recall curves obtained varying the number of k-mers used to train the Motif Graph models. To establish the number of training k-mers returning the best discriminative performance we computed the Precision-Recall curves of different Motif Graph models trained using 10, 50, 100, 200,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

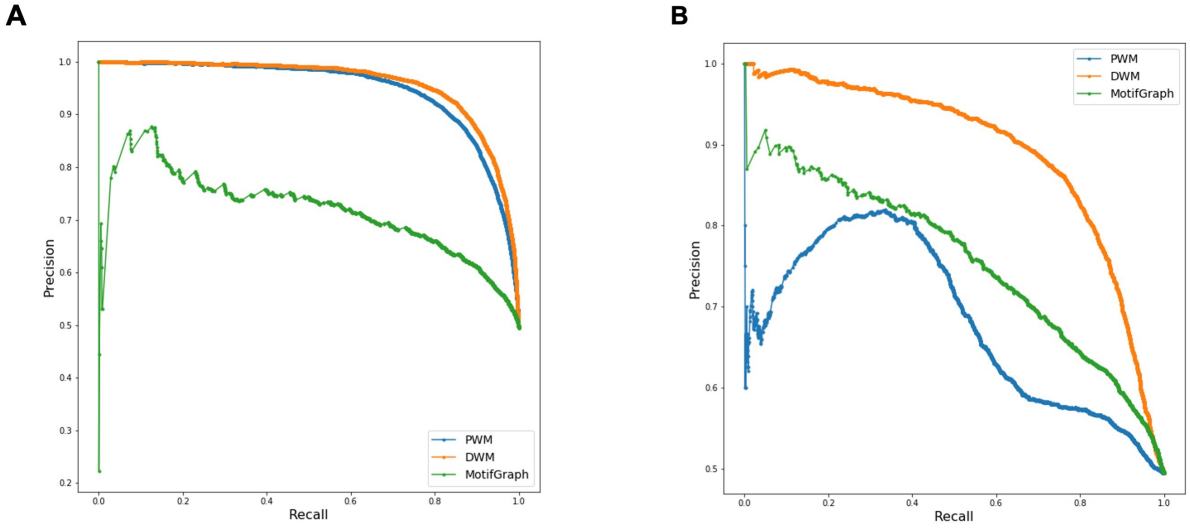


Figure 3.6. Comparing Motif Graph, PWM, and DWM Precision-Recall curves. We compared the discriminative power of the Motif Graph models against that of PWMs and DWMs for both **(A)** CTCF and **(B)** GATA1.

Variant effects prediction on Transcription Factor Binding Sites

Several studies showed that genetic variants can significantly impact TF-DNA binding events (**???**). Genome-wide association studies (GWASs) uncovered thousands of genetic variants (SNPs), associated with complex human traits, located in noncoding regions of the genome. Most of these variants are located within functional regulatory elements, like enhancers (**?**). Therefore, misregulations in gene expression may be mediated by SNPs modulating TF-DNA interactions. Such variants may perturb TF-DNA binding sequences, potentially changing the downstream gene expression (**?**). Importantly, mutations altering TFBS can occur in haplotypes conserved within a population of individuals (**?**), resulting in population specific TFBS. Similarly, cell-type genetic variability can create cell-type specific TF target sequences. Therefore, the development of software designed to predict the potential effects of genetic variation on TFBS specificity accounting for haplotype- and cell-type specific mutations becomes fundamental to increase our knowledge about the cell mechanisms governing gene expression. To this aim, we developed two tools predicting mutations effects on TFBS in haplotypes and different cell-types. GRAFIMO (**?**) is a variant- and haplotype-aware motif scanning tool searching potential occurrences of known TF motifs on genome graphs (**?**). 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 (**?**) (**Fig.??**). Variants like SNPs and indels form bubbles in the graph, where diverging paths through the graph are anchored by a common start and end sequence on the reference [138]. MotifRaptor (**?**) investigates the impact of genetic variants on TF, exploiting cell-type specific information regarding chromatin accessibility and gene expression. While the original MotifRaptor used PWM models to assess genetic variants effects, we extended the framework to integrate k-mer-based motif models (see **Section 2.1.3** for details), which provide more reliable predictions.

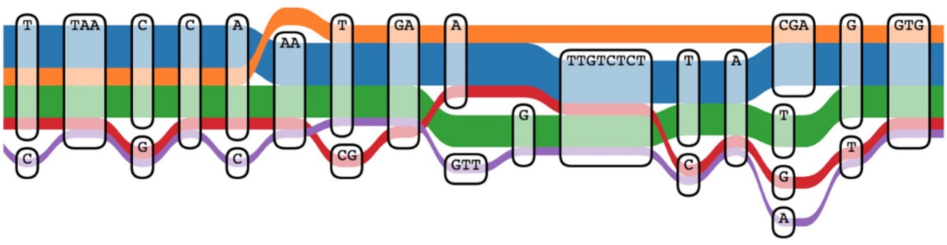


Figure 5.7. Genome Graphs data structure visualization. Each color corresponds to a path in the graph. Each path represents the genomic sequences of one of the individual genomes encoded in the Genome Graph structure.

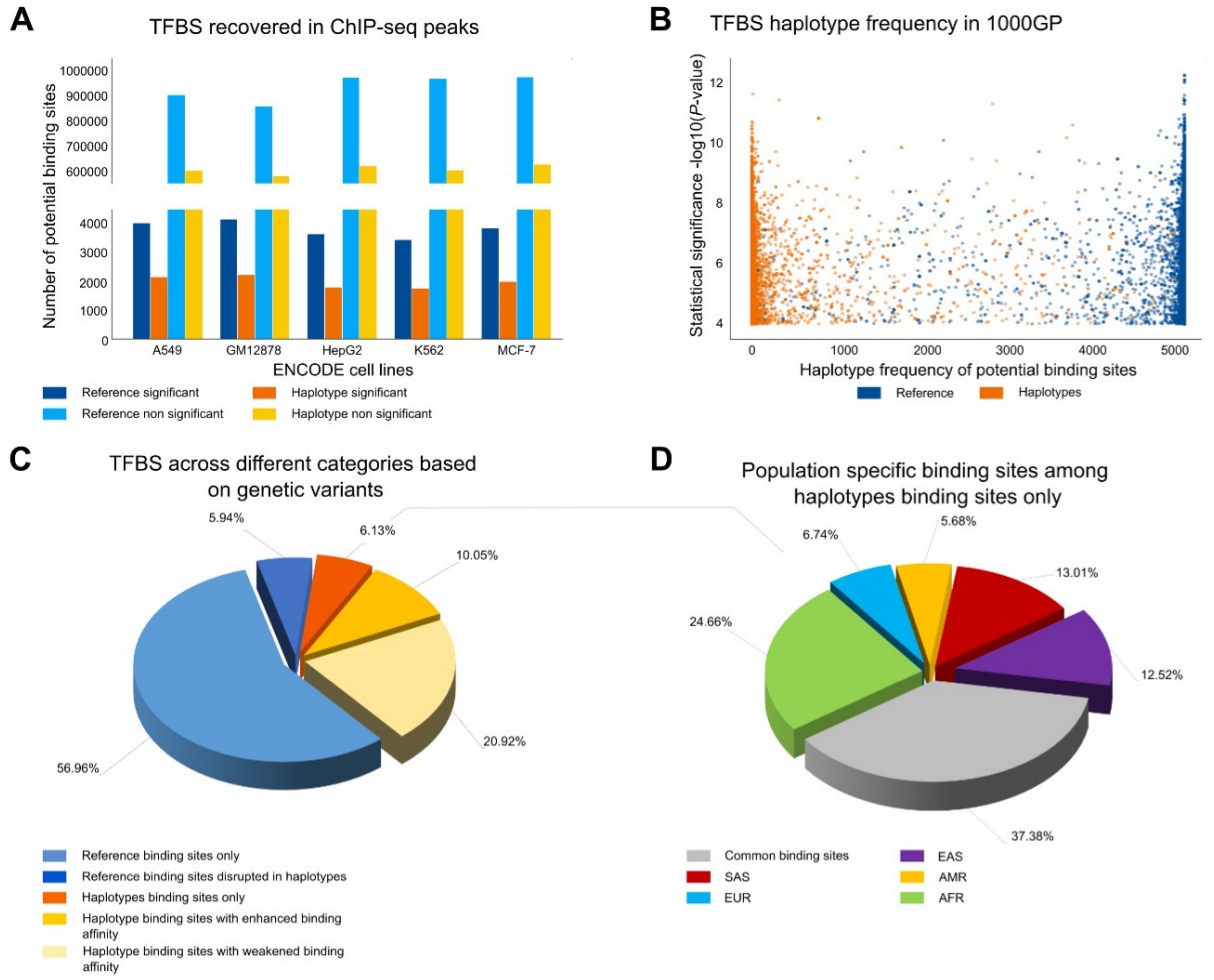
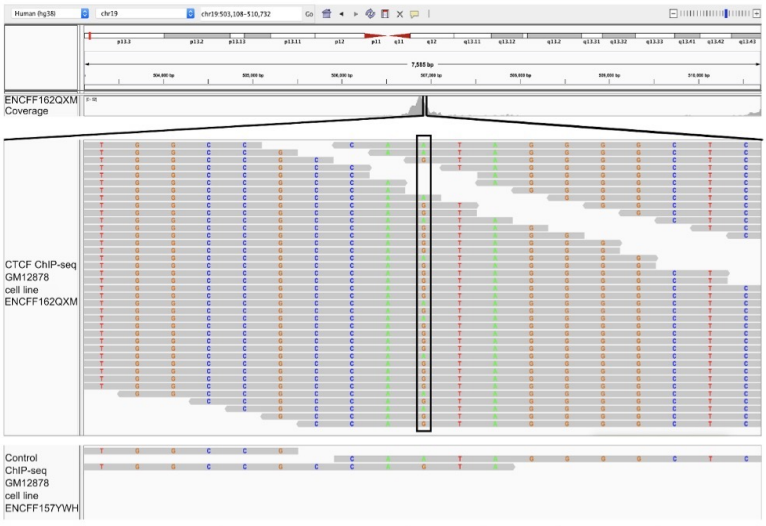


Figure 5.8. Searching CTCF motif on VG with GRAFIMO provides an insight on how genetic variation affects putative binding sites. **(A)** Potential CTCF occurrences statistically significant (P-value ¡ 1e-4) and non-significant found in the reference and in the haplotype sequences found with GRAFIMO oh 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 GRAFIMO

GRAFIMO (GRAph-based Finding of Individual Motif Occurrences) (**?**) is a command-line tool to scan known TF DNA motifs represented as PWMs in genome variation graphs (VGs) (**?**), a scalable and efficient implement of genome graph framework. GRAFIMO extends the standard PWM scanning procedure by considering variations and alternative haplotypes encoded in a VG. Using GRAFIMO on a VG based on individuals from the 1000 Genomes project (**?**) we recover several potential binding sites that are enhanced, weakened or missed when scanning only the reference genome, and which could constitute individual-specific binding events. During the last decade, several methods have been developed to search TFBS on linear reference genomes, such as FIMO (**?**) and MOODS (**?**) or to account for SNPs and short indels such as is-rSNP, TRAP and atSNP (**???**), however these tools do not account for individual haplotypes nor provide summary on the frequency of these events in a population. We observed that several CTCF motif occurrences are found only in non-reference haplotypes, therefore a consistent number of potential motif occurrences are potentially lost when scanning only the reference sequence (**Fig.?? (A)**). Moreover, we found several potential CTCF occurrences with high statistical significance occurring in rare haplotypes, which could modulate gene expression in those subjects (**Fig.?? (B)**). We also investigated how often regulatory mutations disrupt, create or modulate TFBS binding affinity. Interestingly, we observed that many TFBS can be found only on individual haplotypes. Moreover, 6% of TFBS are disrupted (do not survive the ss threshold) by the presence of genetic variants, while 30% of potential CTCF TFBS are still significant in non-reference haplotypes, but with enhanced or weakened ba score (**Fig.?? (C)**). We also observed that several putative non-reference TFBS are population-specific



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 5.9. 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.

(**Fig.?? (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.??**). Taken together these results highlight the importance of considering non-reference genome data when searching for potential TFBS or to characterize their potential activity in a population of individuals.

5.2 MotifRaptor2

Several studies showed that genetic variants can enhance or disrupt TF-DNA binding affinities. Moreover, most of the SNPs ($\sim$ 90%) associated to common diseases and complex phenotypical traits thorugh GWAS analyses have been located within REs, such as enhancers or promoters. MotifRaptor exploits cell-type specific data, like chromatin accessibility and gene expression, to predict the impact of SNPs prioritized via GWAS on TFBS sequence and the corresponding TF-DNA interactions. MotifRaptor workflow is currently constitued of three steps (**Fig.??**). (i) TCharacterize important cell-types through the enrichment of the phenotype associated SNPs in open chromatin regions. (ii) Search TFBS 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. The original MotifRaptor employed TFBS PWM models to predict SNP impact on the binding sites of the investigated factor. However, PWMs have several limitations, like the assumption of independence between neighboring and non-neighboring nucleotides within TF target sites, and often they require large amount of training data to return reliable models (**?**) (see **Section 2.1.3** for details). K-mer-based motif models address some of PWMs' limitations. (i) K-mers are features that are either present or not in the training dataset, therefore, the resulting models do not require large amount of training data. (ii) K-mers capture kth order dependencies between neighboring nucleotides. Therefore, we extended the MotifRaptor analysis

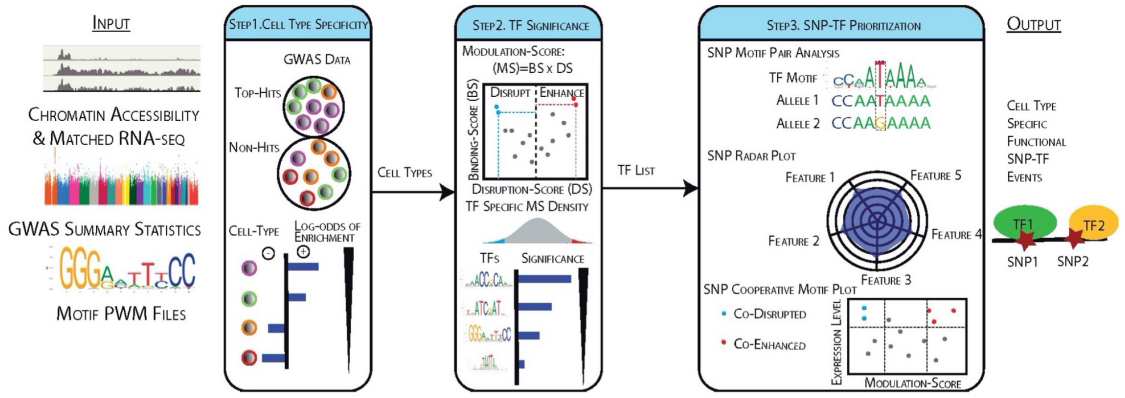


Figure 5.10. MotifRaptor analysis workflow. Three steps are performed: (1) characterize relevant cell types based on the enrichment of phenotype associated SNPs in chromatin accessible sites, (2) find TFs with binding sites that are significantly modulated by genetic variants in these cell types and (3) identify and visualize individual TF-SNP regulation events

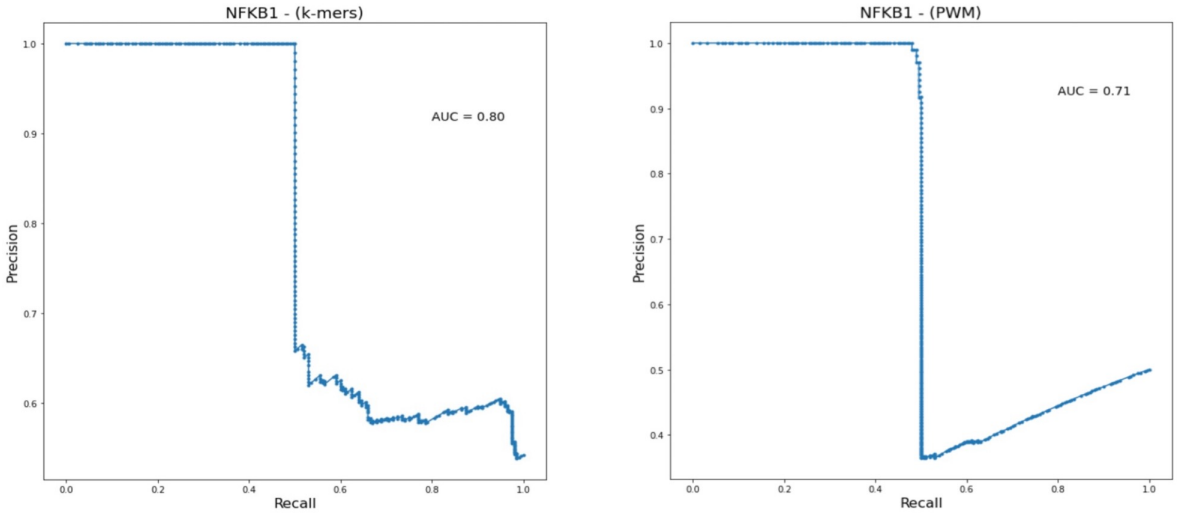


Figure 5.11. Precision-Recall curves for NFKB1 models (k-mers and PWM) and AUC.

to consider k-mer-based motif models to assess the impact of genetic variants on TFBS. With this aim we integrated Gkm-SVM analysis pipeline in MotifRaptor framework, to compute sequence scores reflecting SNPs impact on binding sites. We tested MotifRaptor2 by predicting how NFKB1, RUNX3 and STAT1 binding landscapes are affected by genetic variation, considering SNPs linked to rheumatoid arthritis trait. The k-mer based models for the investigated TFs were obtained using Gkm-SVM on ChIP-seq peakssequences from the ENCODE Project database. Interestingly, we observed that k-mer-based motif models show a higher discriminative power than PWM when identifying bound and unbound sequences (ChIP-seq peak and random sequences, respectively) (**Fig.??**). Therefore, MotifRaptor2 is a promising tool to produce reliable predictions regarding genetic variants impact on transcription factors binding landscapes.

CRISPR off-targets

CRISPR gene editing (**?**) 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 (**?**). 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 (**?**). To exploit the exonucleasic function Cas9 recognize its PAM sequences. PAMs are very short nonspecific sequences, occurring in several locations along the genome (**?**). 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.??**). 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.

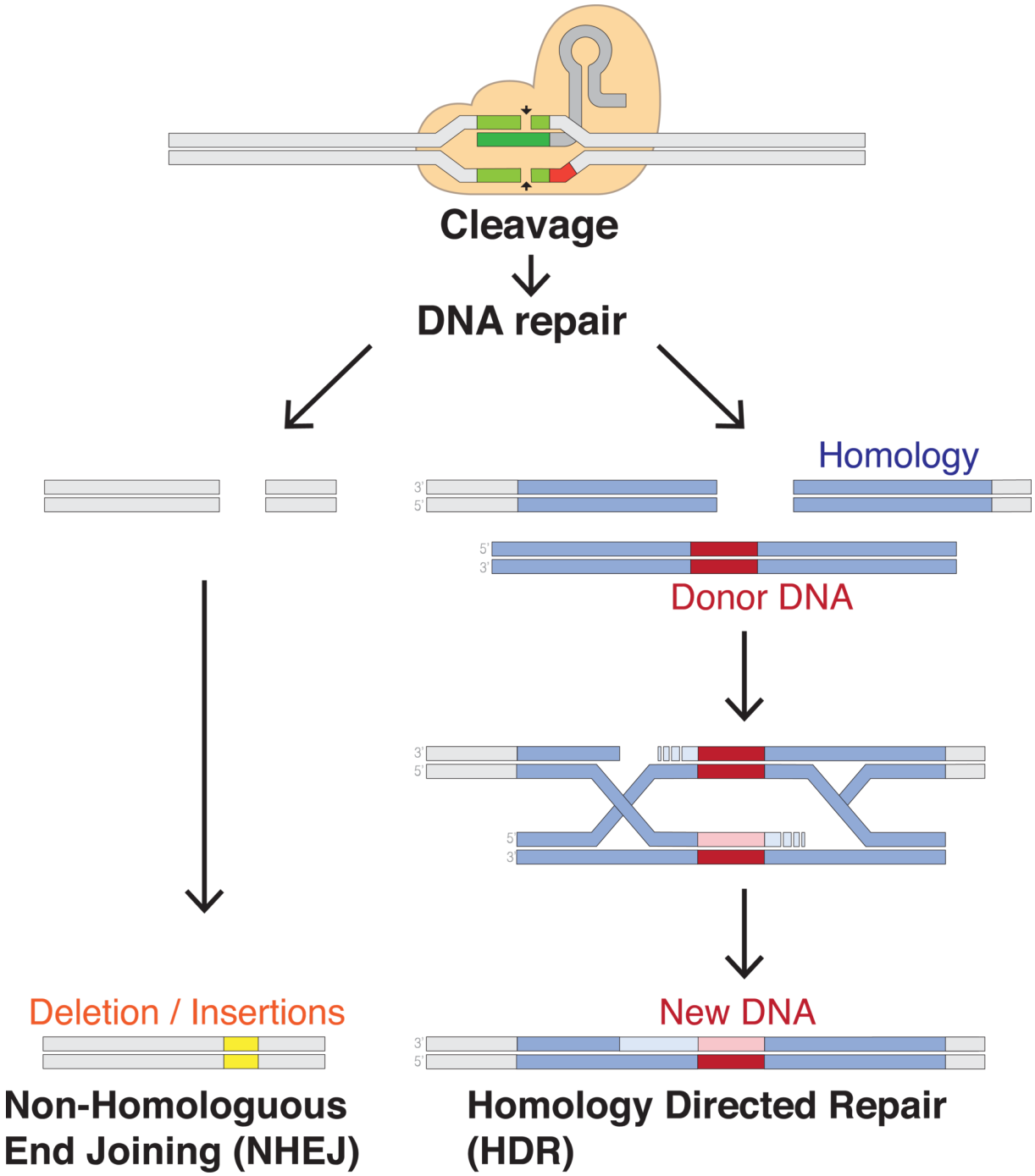


Figure 6.12. CRISPR gene editing via HDR and NHEJ.

CRISPRme

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 (**?**). 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 (**????**). Computational models can complement experimental approaches to off-target nomination in several respects: to triage gRNAs prior to experiments by predicting the number and cleavage potential of off-target sites and to prioritize target sites for experimental scrutiny. 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 (**??**). 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 (1000G) (**?**) and the Exome Aggregation Consortium (**?**) have highlighted how genetic variants can significantly alter the off-target landscape by creating novel and personal off-target sites not present in a single reference genome (**??**). 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 (**????**). 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 (**?**). This tool uses optimized data structures to efficiently account for single variants, mismatches and bulges but with significant limitations (**?**). We substantially extended CRISPRitz 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. 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. It can integrate population genetic variants from sets of phased individual variants (like those from 1000G), unphased individual variants (like those from the Human Genome Diversity Project, HGDP (**?**)) and population-level variants (like those from the Genome Aggregation Database, gnomAD (**?**)). 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 SCD and β-thalassemia (**???**). We identify possible off-targets introduced by genetic variants included within and extending beyond 1000G. 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+ hematopoietic stem and progenitor cells. Furthermore, we demonstrate that alternative allele-specific off-target potential is widespread across various gRNAs and editors.

7.1 CRISPRme: a variant-aware computational tool to nominate candidate off-target sites

CRISPRme is a web-based tool to predict off-target potential of CRISPR gene editing that accounts for genetic variation. CRISPRme can also be deployed locally as a web app or used as a command line utility, both of which respect genomic privacy offline. CRISPRme takes as input a Cas protein, gRNA spacer sequence(s) 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.?? (A)**). We designed CRISPRme to be flexible with support for new gene editors with variable and extremely relaxed PAM requirements (**?**). 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 7) and RNA:DNA bulges (tested with up to 2). Notably, a comprehensive search performed with up to 6 mismatches, 2 DNA/RNA bulges and a fully non-restrictive PAM (`NNN`) takes only 24 hours on a small computational cluster node (Intel Xeon CPU E5-2609 v4 clocked at 2.2 GHz and 128 GB RAM). All the 1000G variants, including both SNVs and indels, can be included in the search together with all the available metadata for each individual (sex, super-population 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 (**??**) 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 graphical user interface. 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.?? (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 GENCODE22 (genomic features) and ENCODE23 (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.??**). (iii) Using 1000G and/or HGDP16 variants, CRISPRme reports the cumulative distribution of homologous sites based on the reference genome or super-population. 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 (**Fig.??**). 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.

7.2 Results

We tested CRISPRme with a gRNA (#1617) targeting a GATA1 binding motif at the +58 erythroid enhancer of BCL11A18,19. A recent clinical report described two patients, one with SCD and one with β-thalassemia, each treated with autologous gene modified hematopoietic stem and progenitor cells (HSPCs) edited with Cas9 and this gRNA, who showed sustained increases in fetal hemoglobin, transfusion-independence and absence of vaso-occlusive episodes (in the SCD patient) following therapy. This study as well as prior pre-clinical 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 (**???**). 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.?? (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 3 mismatches and a CFD score (CFDalt 0.95) approaching that of the on-target site

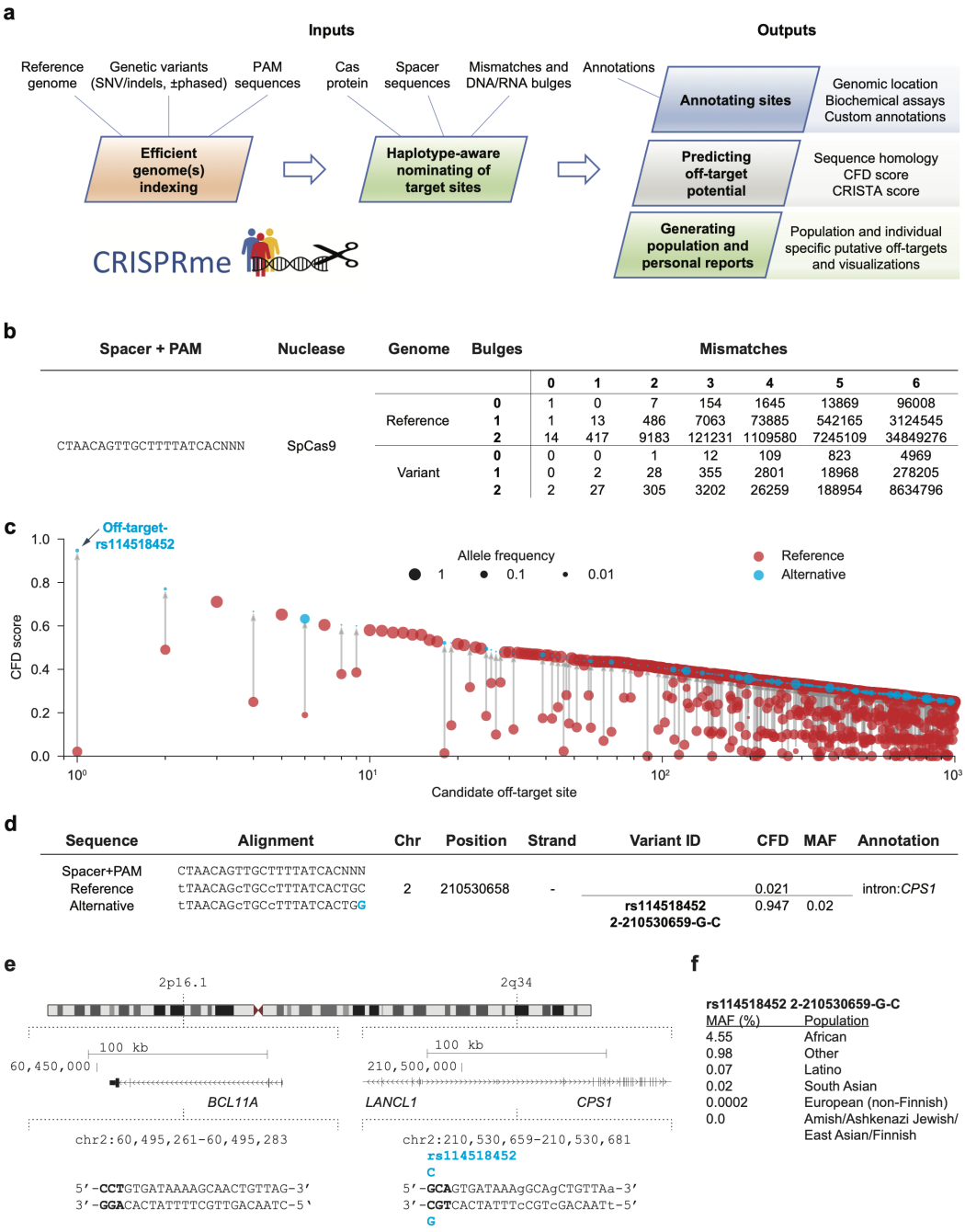


Figure 7.13. CRISPRme provides web-based analysis of CRISPR-Cas gene editing off-target potential reflecting population genetic diversity. **(A)** CRISPRme software takes as input a reference genome, genetic variants, PAM sequence, Cas protein type, spacer sequence, homology threshold and genomic annotations 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.

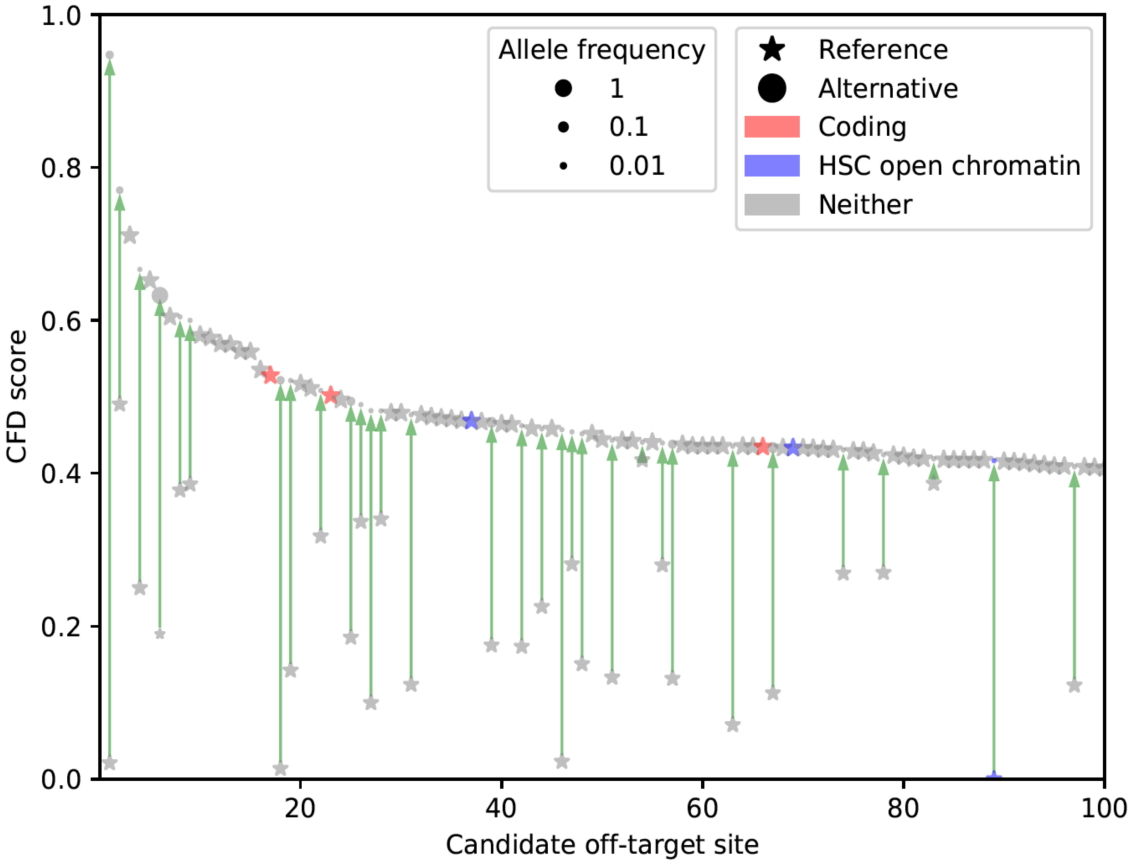


Figure 7.14. Top 100 predicted off-target sites for BCL11A-1617 spacer by CFD score. CRISPRme search as in **Fig.??**. Candidate off-target sites within coding regions based on GENCODE annotations and ATAC-seq peaks in HSCs based on user-provided annotations are highlighted.

(**Fig.?? (E)**). In contrast, the reference allele rs114518452-`G` disrupts the PAM to `TGC`, which markedly reduces predicted cleavage potential (CFDref 0.02). rs114518452-`C` has an overall MAF of 1.33% in gnomAD v3.1, with MAF of 4.55% in African/African-American, 0.9% in Other, 0.07% in Latino/Admixed American, 0.02% in European (non-Finnish) and 0.00% in East Asian populations (**Fig.?? (F)**). To consider the off-target potential that could be introduced by personal genetic variation that would not be predicted by 1000G variants, we analyzed HGDP variants identified from whole genome sequences of 929 individuals from 54 diverse human populations. We observed 249 candidate off-targets for gRNA #1617 with CFD ≥ 0.2 for which the CFD score in HGDP exceeded that found for either the reference genome or 1000G variants by at least 0.1 (**Fig.?? (A)** and **Fig.??**). These additional variant off-targets not found from 1000G were observed in each super-population, with the greatest frequency in the African super-population (**Fig.?? (B)**). 229 (92.0%) of these variant off-targets were unique to a super-population and 172 (69.1%) of these were private to just one individual (**Fig.?? (C)**). Furthermore, single individual focused searches, for example an analysis of HGDP01211, an individual of the Oroqen population within the East Asian super-population, showed that most variant off-targets (with higher CFD score than reference) were due to variants also found in 1000G (n=32369, 90.4%), a subset were due to variants shared with other individuals from HGDP but absent from 1000G (n=3177, 8.9%), and a small fraction were private to the individual (n=234, 0.7%) (**Fig.?? (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.?? (D)** and **(E)**). To 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.?? (C-F)**). We performed 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.?? (A-C)**), suggesting $9.6 \pm 1.0\%$ off-target editing of the alternative allele. In an additional 6 HSPC donors homozygous for the reference allele rs114518452-`G/G`, $0.00 \pm 0.00\%$ indels were

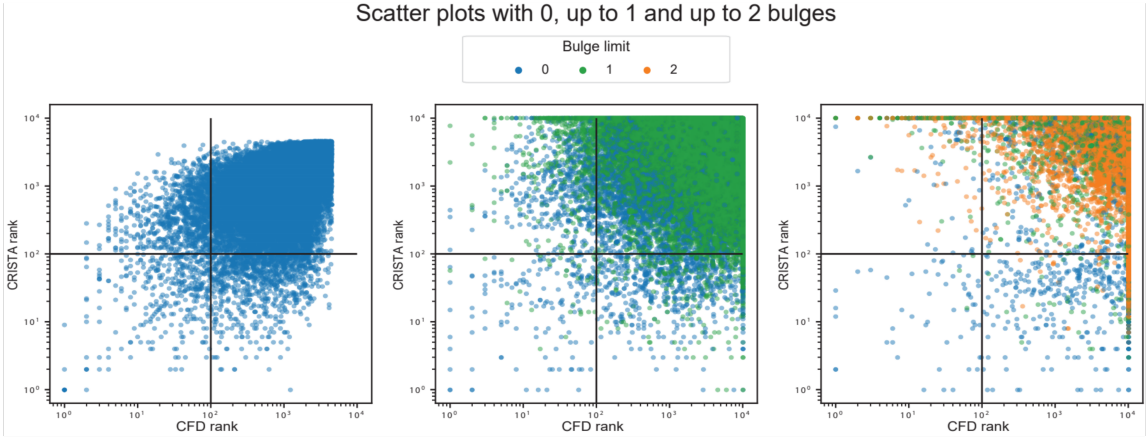
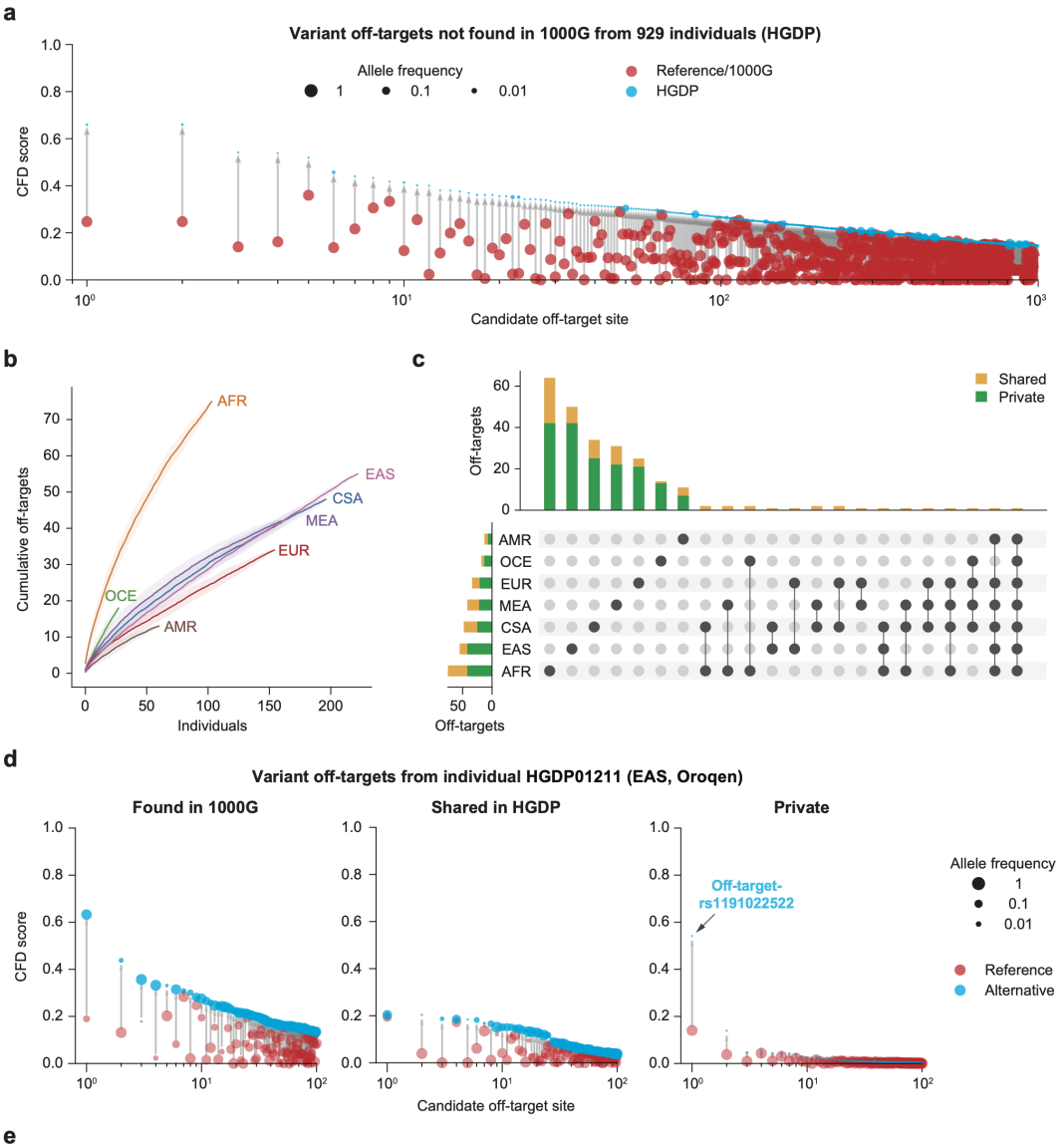


Figure 7.15. Plots with rank ordered correlation between CFD and CRISTA reported targets. Scatter plots show from left to right, the correlation of ranked targets, extracted by selecting top 10000 targets ordered by CFD and CRISTA score, respectively. The left plot shows the rank correlation of targets with 0 bulges (Pearson's correlation: 0.57, $p < 1e^{-10}$, Spearman's correlation: $0.55, p < 1e^{-10}$), the center plot shows 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, since the two scoring methods may prioritize different alignments and thus different number of mismatches and bulges of the same genomic target.

observed at the off-target site, suggesting strict restriction of off-target editing to the alternative allele (**Fig.?? (D)**). The on-target BCL11A intronic enhancer site is on chr2p while the off-target-rs114518452 site is on chr2q within an intron of a non-canonical 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. 4a,b). Deep sequencing of the inversion junction showed that inversions were restricted to the alternative allele in the heterozygous cells (**Fig.?? (C)** and **(D)**). Droplet digital PCR revealed these inversions to be present at $0.16 \pm 0.04\%$ allele frequency (**Fig.?? (E)**). Various high-fidelity Cas9 variants may improve the specificity of gene editing, although at the possible cost of reduced efficiency (**?**). Gene editing following the same electroporation protocol using a HiFi variant 3xNLS-SpCas9 (R691A) (**?**) 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, i.e. a $\sim$ 48-fold reduction compared to SpCas9 (**Fig.?? (C)**). Inversions were not detected following HiFi-3xNLS-SpCas9 editing (**Fig.?? (B)** and **(E)**).

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 (**??????????**) and 6 gRNAs for non-SpCas9-based editing such as for SaCas9 and Cas12a (**?????**). CRISPRme analysis including the 1000G 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.?? (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 candidate cis-regulatory elements and protein coding sequences (including putative tumor suppressor genes (**?**)) and/or which involve PAM creation events (**Fig.?? (B-C)**). For example, within the top 20 candidate off-targets nominated by CRISPRme for a SpCas9 gRNA targeting EMX135, 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, while the second introduces two PAM-proximal matches to the gRNA (**Fig.?? (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 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 AAV. For a SaCas9-associated gRNA targeting CEP29040 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 5 (REF) to 4 (ALT) which are predicted to produce cleavages within coding sequences (**Fig.?? (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 PCSK937 that has been used with SpCas9-nickase adenine base editor in vivo in preclinical studies to reduce LDL cholesterol levels, 4 of the top 5 candidate off-target sites involve alternative alleles, including one with CFDref 0.2 and CFDalt 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) which would be predicted to produce missense mutations in EPHB3, a putative tumor suppressor gene (**Fig.?? (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 (RNAi) or antisense oligo (ASO) therapies (**Fig.??**). We performed a variant-aware search (without PAM restriction) for the FDA-approved antisense oligonucleotide Nusinersen (**??**), 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 3 (REF) to 2 (ALT). Similarly, analysis of the FDA-approved RNAi therapy Inclisiran (**?**), which targets PCSK9 mRNA to treat hypercholesterolemia, revealed that its antisense strand has a candidate off-target in the 3' UTR of the ribosomal gene RPP14 for which a common insertion variant (MAF 36%) reduces the number of mismatches and bulges from 7 (REF) to 4 (ALT).



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 7.16. 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 1000G were plotted and sorted by CFD score, with HGDP variant off-targets shown in blue and reference or 1000G variant off-targets shown in red. **(B)** HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD of ≥ 0.1. Individual samples from each of the seven super-populations were shuffled 100 times to calculate the mean and 95% confidence interval. **(C)** Intersection analysis of HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD of ≥ 0.1. Shared variants were found in 2 or more HGDP samples while private variants 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 1000G but shared with other HGDP individuals (center panel), and higher CFD score compared to reference genome and 1000G 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 1000G 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.

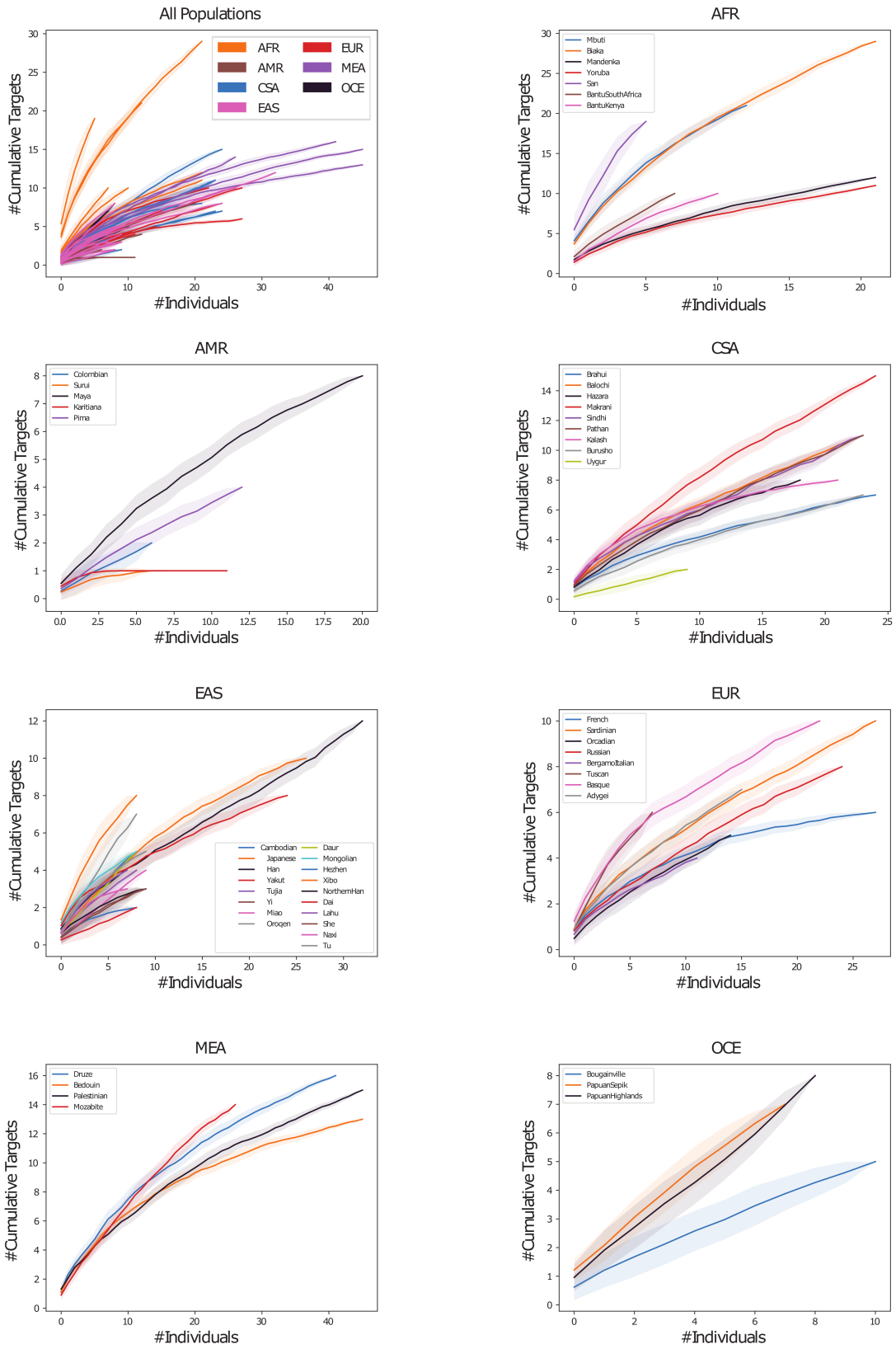


Figure 7.17. HGDP super-population distribution plots. HGDP variant off-targets with CFD ≥ 0.2 and increase in CFD of ≥ 0.1. Individual samples from each of the seven super-populations were shuffled 100 times to calculate the mean and 95% confidence interval. First panel shows distribution within all 54 discrete populations, colored by super-population. Additional seven panels show distribution of discrete populations within each listed super-population.

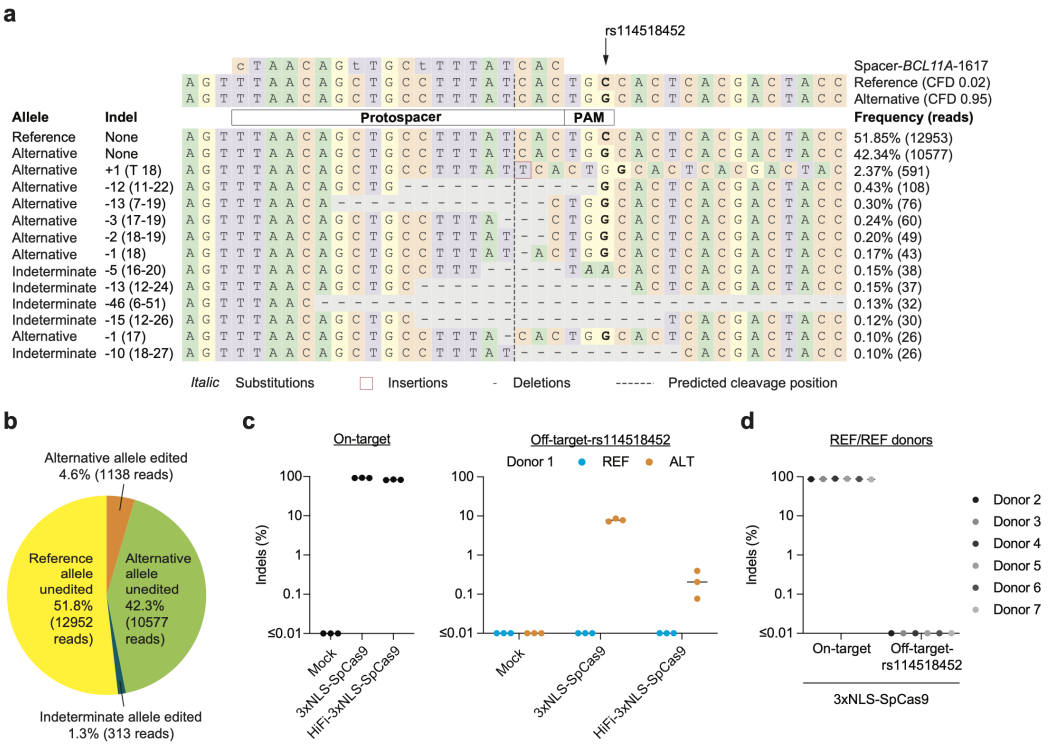


Figure 7.18. Allele-specific off-target editing by a BCL11A enhancer targeting gRNA in clinical trials associated with a common variant in African-ancestry populations. (A) Human CD34+ HSPCs from a donor heterozygous for rs114518452-G/C (Donor 1, REF/ALT) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation 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)** Human CD34+ 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 (mock) followed by amplicon sequencing of the on-target and off-target-rs114518452 sites. Each dot represents an independent biological replicate ($n = 3$). Indel frequency was quantified for reads aligning to either the reference or alternative allele. **(D)** Human CD34+ HSPCs from 6 donors homozygous for rs114518452-G/G (Donors 2-7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation with 1 biological replicate per donor followed by amplicon sequencing of the on-target and off-target-rs114518452 sites.

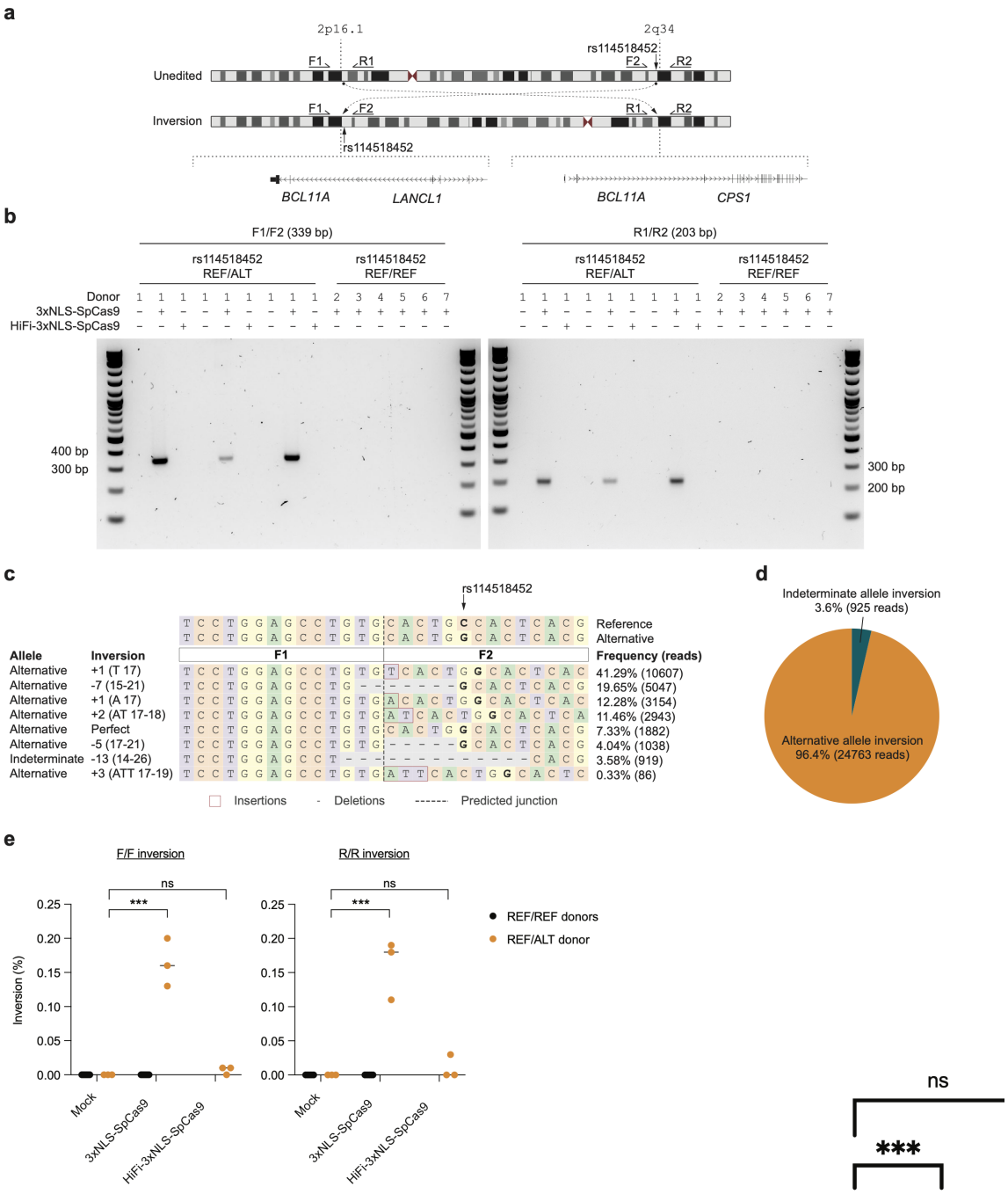


Figure 7.19. Allele-specific pericentric inversion following BCL11A enhancer editing due to off-target cleavage. **(A)** Concurrent cleavage of the on-target and off-target-rs114518452 sites could lead to pericentric inversion of chr2 as depicted. PCR primers F1, R1, F2, and R2 were designed to detect potential inversions. **(B)** Human CD34+ 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 3 biological replicates. Human CD34+ HSPCs from 6 donors homozygous for rs114518452-G/G (Donors 2-7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation with 1 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 ddPCR from same samples as in **(B)**. F/F indicates forward and R/R reverse inversion junctions as depicted in **(A)**.

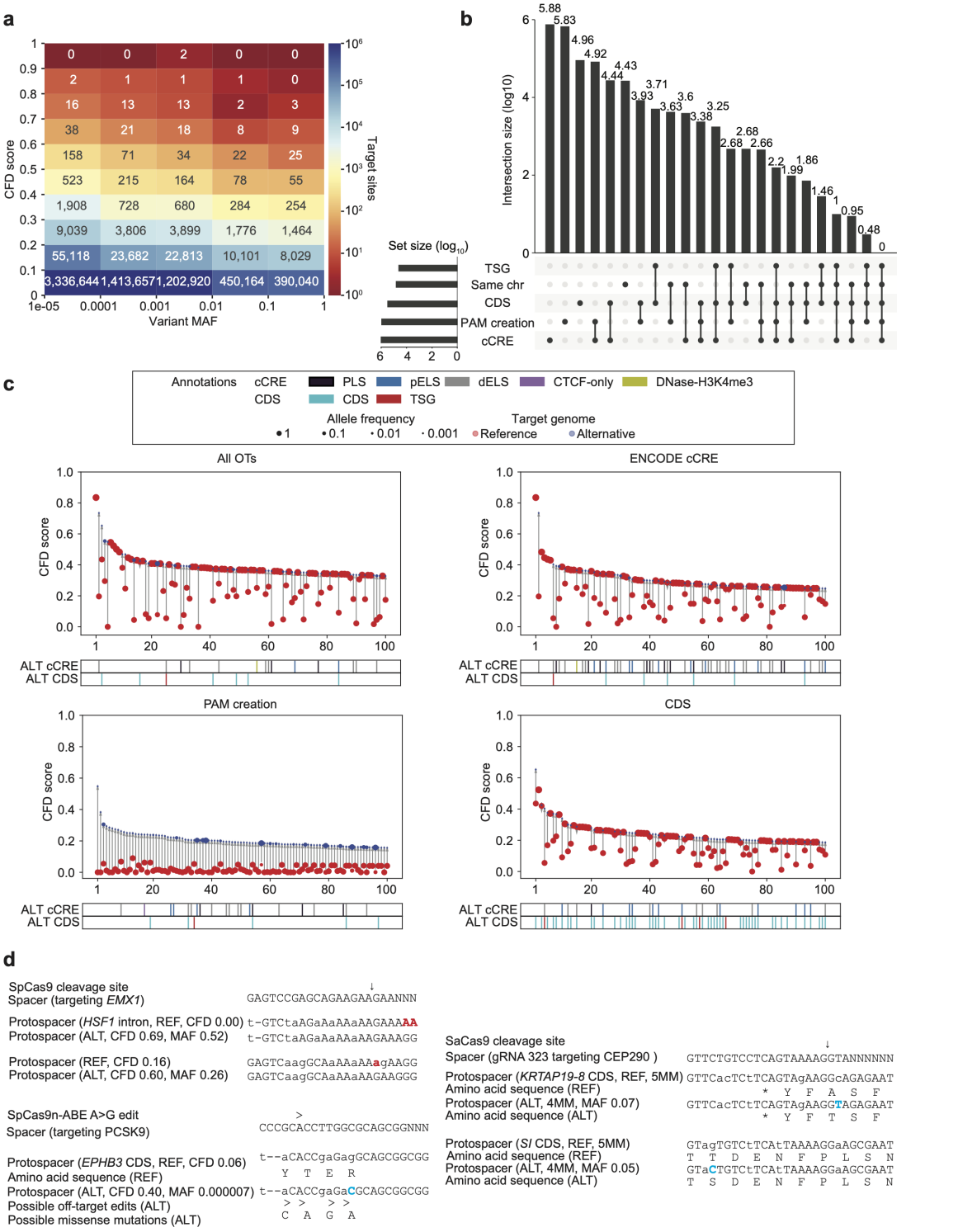


Figure 7.20. CRISPRme illustrates prevalent off-target potential due to genetic variation. (A) Heatmap showing the distribution of alternative allele nominated off-targets for SpCas9 guides by CFD score and MAF. **(B)** UpSet plot showing overlapping annotation categories for candidate off-targets (TSG, tumor suppressor gene; candidate off-targets on the same chromosome 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) Top left: Candidate off-target sites with increased predicted cleavage potential introduced by common (MAF 52% and 26%) indel variants for a SpCas9 gRNA targeting EMX1. Right: 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. Bottom: 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 LDL cholesterol levels. Deletions shown in red, SNPs shown in blue.

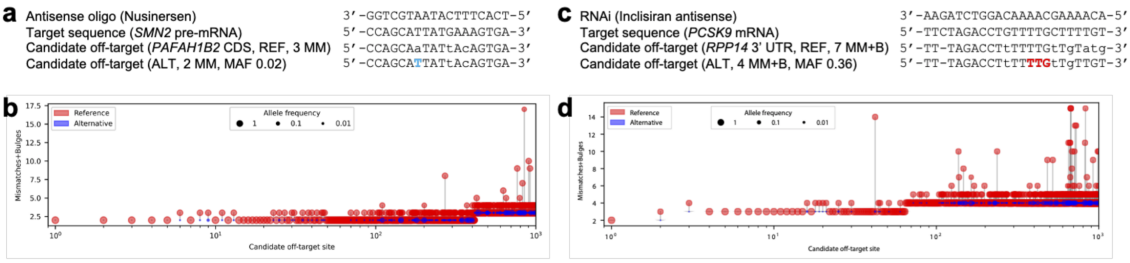


Figure 7.21. Candidate transcript off-targets introduced by common genetic variants for non-CRISPR sequence-based RNA-targeting therapeutic strategies. (A) A common SNP (in blue) introduces a candidate CDS off-target site with 2 mismatches for the FDA-approved antisense oligo Nusinersen. **(B)** Top 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.

Future directions

In the next future, we plan to improve the Motif Graph framework by developing a more efficient and powerful model training procedure, which will accounts for k-mers weights when assigning the scores on the G edges. Moreover, we plan to speed-up the Motif Graph training procedure potenbtially emplyoing string indexing algorithms, like Suffix Arrays or BWT. We plan to integrate the kernel and model training procedures in a single software implementation and extensively test the newly obtained framework on different TFBS data sources , such as ChIP-seq, HT-SELEX, PBM, etc.

In the next months we plan to implement MotifRaptor2 in a comprehensive software suite, and to extensively test the framework on different GWAS datasets, cell-types, and TFs.

We plan to extend CRISPRme framework to account for complex genetic mutation events, such as indels and structural variations. To accomplish this task, we plan to extend CRISPRme search to search off-targets on genome graph data structures, which handle indels and structural variations by definition. Moreover, we also plan to extend CRISPRme to analyze the transcriptome beside the genome. Analyzing the transcriptome would enable to directly estimate potential off-targets on the genome coding regions.

References

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

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

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.

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.* (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, **53**(3), 354–366.

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.

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.

Beer, M. A. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, **117**(2), 185–198.

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

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

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

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

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.

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

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.

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.

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.* (2022). Human genetic diversity alters therapeutic gene editing off-target outcomes (*in press*). *Nature genetics*.

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.

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

Clement, K., 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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., *et al.* (2017). The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, **45**(D1), D840–D845.

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.

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

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

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

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

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., 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.

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., 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.

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, 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, 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., 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.

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

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

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

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

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

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.

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.

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.

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, 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.

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

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

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.

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.

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.

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., *et al.* (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, **22**(9), 1798–1812.

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.

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.

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.

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.

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.

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.

Zaret, K. S. and Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current opinion in genetics & development*, **37**, 76–81.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, **32**(12), i121–i127.

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.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, **12**(10), 931–934.

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

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