JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Dissertation

Identification and modelling of genetic features for genomic position profiling to support future approaches in gene therapy

in fulfilment of the requirements for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

by
Michael Menzel

submitted to the
Faculty of Biology and Chemistry
Justus Liebig University Giessen
Giessen, Germany

prepared at the
Faculty of mathematics, natural sciences and informatics
University of Applied Sciences Giessen
Giessen, Germany

in conjunction with the
Research Campus of Central Hessen
Giessen, Germany

Giessen, Mai 2021

Reviewers

First reviewer:	Prof. Dr. Andreas Gogol-Döring, University of Applied Sciences, Giessen
Second reviewer:	Prof. Dr. Alexander Goesmann, Justus Liebig University, Giessen
Examiner:	Prof. Dr. Stefan Janssen, Justus Liebig University, Giessen
Examiner:	Prof. Dr. Marek Bartkuhn, Justus Liebig University, Giessen

Date of defense:	October 6th, 2021

Declaration of Authorship

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived verbatim from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen "Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis" in carrying out the investigations described in the dissertation.

Gießen, Mai 2021

Michael Menzel

Abstract

Genetic variability and mutations are a fundamental necessity for living organisms. In the context of evolution, both factors facilitate survival and enable the adaption of life to the environment. However, genetic variability also leads to risk factors for various medical conditions, and mutation can lead to divergent behavior, such as unsupervised division and proliferation of cells, known as cancer. With the cause being genetic, the most effective treatment and therapy is to modify the corresponding genetic sequence to repair or augmenting erroneous genes and silence risk factors, known colloquially as gene therapy. Besides applications in therapy, genetic modification is also a breakthrough technology for biotechnological engineering where cell cultures are utilized for drug production.

Modifying genomes is a complex challenge because the correlation within the genome is neither fully understood, nor is the modification itself based on reliable mechanisms. Initial approaches to gene therapy in humans have shown that it is a potential game changer for many targets. However, immune reaction and insertional mutagenesis is a significant concern and universal application in therapy is only possible if major side-effects can be avoided.

The foundation of gene therapy is the understanding of genomic function. Thus, in this dissertation, the aspect of transcript factor binding specificity is closely examined to gain new insights using a novel technology to analyze ChIP-Seq datasets. My approach is based on inferring binding models directly from the distributions of reads in relation to nearby sequences. This novel approach is capable of analyzing data sets that did not yield results using established methods.

Furthermore, gene therapy relies on vectors that deliver genetic elements and insert them based on given targets. Therefore, a platform is presented to review insertional characteristics of genomic positions based on viral integration and transposases. Fundamental for the analysis is a mechanism to create computational background models that can be adapted for technological factors, as well as other known covariates. The applicability of the platform is shown in several publications that review genomic insertion preferences of delivery vectors.

Acknowledgements

With the completion of this thesis, I want to express my sincere thanks to my direct supervisor Prof. Dr. Andreas Gogol-Döring for his continuous efforts in supporting my scientific work, his imaginative research ideas, and assistance in the manifold challenges of my work. I would like to extend my thanks to Prof. Dr. Alexander Goesmann, whoes supervision greatly influenced the success of my work in scientific development, for his guidance through my studies, and help in improving the thesis. My sincere thanks also goes to Prof. Dr. Andreas Dominik for his support in my academic studies from my first semester until now, as well as his aid in my research projects.

I gratefully acknowledge the funding received from the Hessian Ministry of Higher Education, Research and the Arts and the Strategischer Forschungsfond of the University of Applied Sciences. I thank the Research Campus of Central Hessen and the divison of Forschung, Transfer und wissenschaftlicher Nachwuchs for their support. I want to thank Sabine Hurka for her efforts in reviewing my dissertation and her prior research, which greatly advanced my work, as well as Stefan Glasenhardt for his assistance with the development of numerous auxiliary components. Furthermore, I want to thank Eric Hartmann for his help and technical assistance in our shared office.

Finally, I want thank my parents for their love and constant support in my life, which enables me to pursue my goals.

Contents

1 Introduction

The intricate systems that each cell contains are the foundation of living organisms and their proliferation. Cell regulation enables cells to uphold physiological equilibrium, respond to changing external influences, and perform cell division. Each cell and each mechanism within constitute a part of organism survival. Evolved mechanisms ensure the adaptability of life to numerous environments as well as the adaption of individual cells to their requirements. Metabolic pathways and regulatory complexity needed to uphold physiological balance and cell function are the *sine qua non* of biological life. However, failures in cellular processes are inseparable from complex regulatory systems. Single-cause errors that happen by mischance are a constant occurrence. Evolution has led to the prevalence of various correction mechanisms that are able to repair damage or induce apoptosis of cells for irreparable states. Nonetheless, combinations of genetic damage, failure of correction mechanisms, and specific mutations lead to acquired genetic disorders. For example, cancerous cells average four mutations, whereby mutations ranging from one to ten per tumor are commonly found [80], illustrating the impact of minor divergence. Further, inherited genetic diseases appear as systemic divergence, such as chromosome abnormalities, susceptibility to certain ailments and other genetic conditions are commonly present.

While building the substructure of biological life, cellular mechanisms are also a root cause for many diseases, and varying cell responses influence disease susceptibility. Current approaches to curing genetic diseases and reducing risk factors for other diseases rely on the ability to alter genetic elements. Today, gene therapy, including gene alteration, insertion, and silencing is a highly specialized curative therapy. Numerous possible side-effects and the lack of understanding of genetic features impede approval for a multitude of potential targets.

This dissertation is focused on two aspects of gene therapy applications: first, the evaluation of insertional elements that are used as delivery vectors for genetic elements; and second, a re-evaluation of Chromatin immunoprecipitation followed by sequencing (ChIP–Seq) [107] experiments using a novel k-mer based technology to increase understanding of transcription factor mechanisms and therefore cell regulation. Hereafter, an overview of foundations and recent advances in gene therapy as well as cell regulation by transcription factors is provided, followed by a summary of objectives and open questions.

1.1 Curing diseases with gene therapy

Mutation is a naturally occurring process in cell lifetime and reproduction caused by cellular processes and outside influence. The majority of genetic alterations are inconsistent. Different mechanisms exist for detection and corrective repair. The proliferation of excessive damage is prevented by apoptosis. Nonetheless, certain genetic alterations elicit diseases. Cell regulation is a balanced process working in complex coherency and those complex relationships are strongly affected by alterations of genetic function and regulatory mechanisms. Alterations can lead to divergence from expected behavior and subsequently to diseases that affect cells and consecutively organisms. This includes common genetic diseases, such as cystic fibrosis, haemophilia, sickle-cell anaemia and aneuplodies [53]. Furthermore, rare genetic diseases exist that are rare in relative appearance, yet common when considered collectively [70]. To date, over 4,000 genes associated with disease phenotypes have been identified [1]. Additionally, genetic constitution exerts an influence on almost all courses of prevalent diseases with varying intensity [20, 53]. For example, a hypercholesterolemia mutation results in a three-fold increased risk of coronary artery disease (CAD) [2]. Genetic predisposition is known to influence the onset of various diseases, such as Huntington's disease associated with a genetic defect in the HD gene [92], up to 15% of cases of Parkinson [129], Amyotrophic lateral sclerosis (ALS) [83], and increased risk of type 2 diabetes [34, 114], among others.

Commonly-used therapy approaches for such diseases focus on suppressing adverse implications and symptoms. For genetic variation in a contained set of cells, such as tumors, the removal of mutated cells is a possibility. However, in the case of metastasis or genetic diseases that affect the majority of cells, removal is impossible. Actual curative therapy for genetic diseases and acquired disorders is the modification of DNA, known as gene therapy. This is the general term for different procedures that aim to treat a genetic disease by altering the genome. The conception of altering DNA has been around since the structure of information retention on DNA was discovered [33, 8]. Even though prospects are wide-ranging and over 2,500 clinical studies

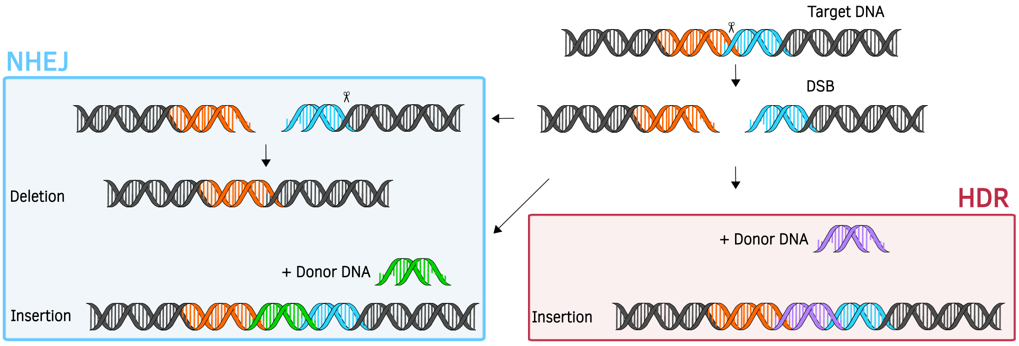


Figure 1.1: Comparison of NHEJ and HDR repair of a DSB. Both mechanisms are able to insert a donor DNA into the DSB. For HDR, a donor DNA is necessary, while with NHEJ both strands can be joined directly (Adapted from [103, 60]).

have been conducted to date, only six gene therapies are approved by the European Medicines Agency (EMA) and US Food and Drug Administration (FDA) for different cancer typess, cystic fibrosis, AIDS and deficiency diseases [8]. The complexity of development, adverse events in clinical trials and a lack of knowledge of cellular interactions impede their development, whereby each of these factors will be reviewed in the following.

Within the broad term of gene therapy, different strategies to alter genomic features exist according to Anguela et al. (2019). Non-functional genes can be replaced by inserting a functional copy of the gene that compensates for a loss-of-function mutation of the original gene. Gene silencing inhibits gene expression and is suitable to correct for gain-of-toxicity mutations. Gene addition inserts genes which modulate diseases and gene editing specifically repairs gene mutations [8]. Each strategy has different challenges for research and clinical application, as well as possible side-effects.

Novel genome editing technology relies on the initiation of a DNA double strand break (DSB) at a specified loci as defined by Cox et al. (2015). The DSB leads to endogenous repair mechanisms that work with either non-homologous end-joining (NHEJ) or homology directed repair (HDR) to join the break (Figure 1.1). DSBs are initiated by nucleases, of which four classes are relevant for editing applications, including zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly-interspaced short palindromic repeats (CRISPR)/Cas9 [21].

Editing can be conducted *ex vivo*, where cells are modified outside of the body and transplanted back. There are several delivery platforms available for *ex vivo* transfer. High editing rates as well as accurate control of dosage make *ex vivo* transfer advantageous. However, not all tissues are capable of surviving extraction and manipulation, and the failed re-introduction leads to a diminished effect. *In vivo* transfer has greater potential for numerous applications, yet, poses new challenges in development and application. Especially delivery mechanisms for *in vivo* transfer are more difficult in fabrication and regulation [21]. A viral vector or non-viral gene carrier is needed to transport new genes or other DNA compounds [90]. Commonly-used vectors are viruses, such as the Adeno-associated virus (AAV) [134]. A closer examination of viral vectors will take place in the following section.

Gene therapy was successfully used in several cases of cellular disorders, first in 2000 when X-linked severe combined immunodeficiency (SCID-X1) was treated by restoring the Il-2 receptor γ (IL2RG) gene using a viral vector [18]. However, among nine successfully treated patients, of ten in total, four developed T cell leukemia after therapy [40]. In a trial conducted in 2009, adenoise deaminase (ADA)-deficient SCID could be corrected by a gene transfer without leukemia side-effect and patients were able to live a normal life [5]. A study from 2007 showed that the motor functions in six patients with Parkinson's disease could be improved using AAV serotype 2 vectors [57]. Further, a degenerative disease of inherited blindness of several children benefited from gene therapy without major side-effects [12, 19, 78].

Although gene therapy has been utilized with success, the main limitation lies with its safety and tolerability. The main aim of future development is to improve integration mechanisms and vectors to mitigate side-effects such as immune system response and improved insertion specificity. Significant effort is undertaken to modify existing vectors, such as AAV, to increase target site specificity and reliability [134]. Additionally, new vectors are developed based on other platforms. With new vectors and modified targeting mechanisms comes the need to evaluate those altered vectors for their integration profile to obtain a risk measurement. As outline before, in many regions integration into DNA induces adverse effects on cell functionality. Especially integration into genes or regulatory regions interferes with cell functionality, so-called insertional mutagenesis.

A recently published review by Bushman (2020) defines four mechanisms responsible for retroviral-caused mutagenesis in humans. First, enhancer insertion boosting cell proliferation, which was observed with early gammaretroviral vectors in the aforementioned SCIDX1 trial. Vectors caused integration of strong enhancers in close vicinity of LMO2, a gene associated with cell proliferation. Second, promotor insertion increasing oncogene activity. Human immunodeficiency virus (HIV) insertions are found enriched near genes associated with leukemia and mRNA sequences of those genes are observed to contain HIV RNA sequences appended at the 5' end. Third, gene inactivation with chimeric antigen receptor (CAR) targeting cancer with T cells. CAR T cell therapy utilizes modified T cells to attack cancer cells. However, loss of function by vector insertion into the intron of TET2 was apparent, a gene involved in myelopoiesis. Finally, mutagenesis by mRNA 3' end substitution reported in a successful correction of gene beta-thalassemia. Vector integrations were observed in the gene HMGA2, associated with various cancers. The genes mRNA contains a 3' untranslated region that is targeted by microRNA, resulting in RNA degradation. With the 3' end mutation, through lentiviral integration the mRNA was stable and resulted in increased HMGA2 protein levels. Additional potential mechanisms are likely to occur with advances in therapy development [17].

To evaluate insertional risks, the annotations for cells such as genetic elements, their expression levels, protein binding sites and histone modifications are commonly intersected with known integration sites, as shown in many publications [37, 22, 109, 26]. Based on statistical evaluation of integration sites in relation to genetic annotations, preferences of a virus or transposable element can be determined. An increased occurrence of sites around highly-expressed genes can often be identified, as shown for HIV [22].

To measure integration enrichment, viral or other genetic sites are compared to a background model. Ideally, background models are an experimentally determined set of genomic sites, subject to equal experimental conditions. However, wet lab costs and time expenditure often prevent their generation and therefore computational generated background models are used. The model can be a random selection of sites throughout the genome. However, the significance of findings is strongly increased by adjusting random sites to fit experimental parameters, as well as known viral preferences. Background models can be adjusted to mimic integration preferences of virus sites over a specific annotation, a so-called covariate. By selecting covariates, this approach enables scientists to formulate specific questions in relation to their dataset. For example, if a virus is known to integrate close to highly expressed genes, the preferences can be used to adjust the background model to mimic the integration near genes and help to identify more faint integration preferences that remain masked by the stronger effect otherwise.

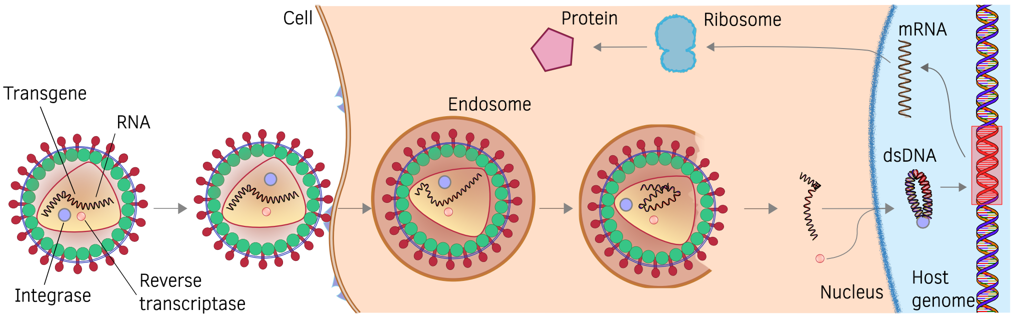


Figure 1.2: Schematic overview of viral integration into a host genome with subsequent translation and transcription of the protein (Adapted from [31, 89, 99, 30]).

1.2 Delivery vectors for in vivo gene therapy

Retroviruses are viruses known to integrate into DNA of host organisms to adapt its functionality for their own replication [16]. This is accomplished by inserting their viral RNA into the host genome using a reverse transcriptase protein, then representing the provirus [16]. After insertion, viral elements are translated alike to other proteins produced by the cell (Figure 1.2) and lead to further spreading of the viral infection.

The sites that viruses prefer for integration are of major interest. They are identified by sequencing cells with next-generation sequencing (NGS) in conjunction with polymerase chain reaction (PCR) amplification [115]. These methods allowed to gather large amounts of viral integration sites in recent years. Those are available online, and organized in different databases [117, 124]. Locations of integration events by retroviruses and transposable elements (TEs) are predominantly purposefully selected, based on genetic features in their vicinity. Between groups of retroviruses the preferences for genomic features can differ [28]. Based on genomic positions obtained from virus integration, it is possible to create a preference profile for each virus that represents features relevant for their integration site selection.

The preferences of several retroviruses are well known. For example, HIV is directed to active transcription units which allows the provirus to be transcribed more frequently [22]. Murine Leukemia Virus (MLV) shows a bimodal integration pattern around transcription start sites (TSS), as well as a preference for chromatin marks near TSS (e.g. H3K4me3) [37]. AAV serotype 2 has a specific integration site, called AAVS1, which is located on chromosome 19 [51, 50]. Although the preferences of many viruses are known, new variants appear and known viruses as well as TEs are modified to abide selected integration preferences. Computational investigations into site selection and integration mechanisms is therefore a continuous effort.

Besides viral vectors, properties of certain TEs form a possible mechanism for genomic alteration [128], such as Sleeping Beauty (SB) [58] and PiggyBac (PB) [88]. TEs are genetic elements that are able to move within the genome from one position to another. It has been shown that TEs form up to 50% of our genes [67]. TEs are naturally involved in regulatory processes of cells, including epigenetic activation and acting as a promotor or enhancer. Further, their ability to mutate genes, as well as chromatin modification and imprinting affect cell regulation [122].

As previously described, the integration of genetic elements at pseudo-random positions in the genome is known to have adverse effects on cells. Instead of using vectors that select their integration positions based on genomic features, a vector targeting a specific sequence can be used. The target is usually selected such that it occurs infrequently and in a safe region for integration. Artificially altered vectors with targeted insertion exist as well as natural viruses that prefer safe integration regions. Nevertheless, those vectors do not show integration solely at intended target locations, but rather off-target sites are also possible. Development and testing of those delivery mechanisms are still subject to recent research [45, 15, 6]. Furthermore, with improved specificity targeted vectors bring the capability to not only add a missing gene in a remote location but specifically edit sequences to restore nominal conditions.

1.3 Cell regulation and transcription factor binding

Each cell in an organism fulfils a specific function, e.g. structural roles like epithelial cells, signal transmission of nerve cells or different blood cells. Their individual function is largely determined by expression based on endogenous factors and their response to environmental influences, such as interaction with other cells and external signals. Expression of genes is controlled by genetic structures on the DNA itself, which provides instructions of regulatory proteins, binding locations and instructions for genes. Other regulating mechanisms are chromatin remodeling complexes [93] and non-DNA-binding co-factors [95]. Regulating transcription of DNA and cell cycle enables cells to adapt to their environment, uphold physiological balance, ensure functionality and enable cellular differentiation. Regulation of gene expression and thereby cell functionality holds significant interest to understand cell behavior in normal and abnormal performance and it is necessary for all therapeutic approaches [131, 120, 3]. Among other influences, the transcription of DNA into mRNA is controlled by proteins that bind to DNA at specific locations called transcription factors (TF). This ephemeral binding induces or inhibits mRNA transcription. Sites where TFs bind are called transcription factor binding sites (TFBS). The relation between TF and gene expression is often not a simple one-to-relation, but rather a complex set of interactions of different TFs and TFBS that increase or inhibit transcription, and affect each other. Even for single genes, a complex network of inhibiting and promoting factors can be relevant [118]. For example, TF binding affinity is not only moderated by the direct DNA binding site, but also the flanking DNA structure, as shown for Gcn4 [98] and Cbf1 in yeast [38]. Likewise, protein binding specificity can be altered by co-factors that bind in an adjacently protein-protein interaction. Cooperative binding does change the affinity to known binding sites and can result in new binding sites. Prominent examples are Matα2, Matα1 and Mcm1 in yeast [139].

The main elements of gene regulation in DNA are promotors and enhancers. Promotor sequences contain TFBS for protein-DNA interaction and are in the close vicinity of TSS. Enhancers are regions found further down- or upstream from TSS containing binding sites for multiple activators [61].

Interaction between enhancers, TFs and promotors is still under investigation. Several mechanisms are proposed that explain the communication structure, including linking of transcription factors between both sites or loops in the DNA that form to bring both genetic elements together (Figure 1.3 right) [35]. It is not fully understood how the mechanism of loop formation works that associates promotors and enhancer functions [35].

Binding sites, to which TFs bind, are determined through a sequence motif, a short sequence whose counterpart is embedded in the TF. Figure 1.3 left shows a simple model of TF and TFBS interaction. Binding proteins recognize a set of similar sequences with varying binding affinity, dependent on the similarity between the binding site sequence and TF sequence. With higher affinity, TFs bind more firmly and their influence is increased [52].

Sequence binding motifs are usually represented by a sequence logo, which is a combined visualisation of bases observed at each binding position. The underlying model of a sequence motif is a position-weight-matrix (PWM), comprising a data matrix

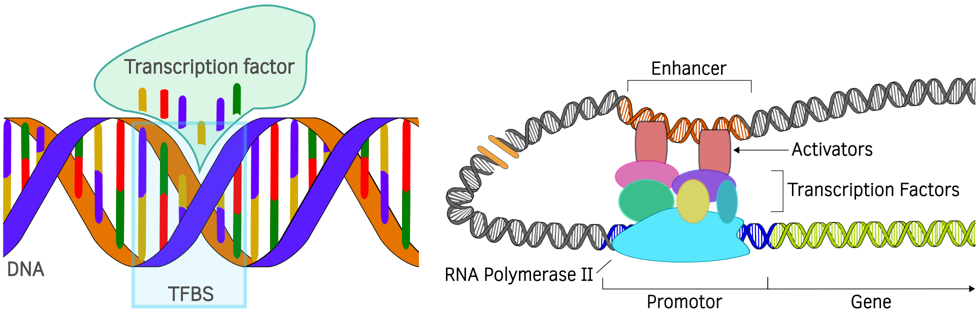


Figure 1.3: Left: Schematic overview of TF binding to DNA. The transcription factor has a complementary sequence to the binding site. Colors represent different nucleobases. Right: TF binding complex with DNA loop (Adapted from [35]).

that represents the observed base composition. The logo visually presents information about the conservation of bases and motif structure. At each position of the sequence logo, letters display the potential binding sequence by letter size, which is calculated with the relative frequency and information content of the bases [113, 123]. Even though sequence logos provide a visually simple and intuitive representation of binding preferences, they inherently show limitations in motif representation as information is highly condensed [39]. Algorithms for motif generation are prone to over-valuing consensus sequences [29, 112]. PWMs are unable to capture variably gapped motifs [9] and interdependent bases cannot be displayed, as PWMs assume an independent contribution of each binding site [39].

Datasets on TF structure and binding site selection are available in different online databases, including Factorbook [137], JASPAR [111], HOCOMOCO [65], CIS-BP [138], and UniPROBE [97]. As of 2018, there are 1,639 known or probable human TFs, of which 93% are expected to bind to DNA features as a monomer or homomultimer [66, 131].

Genetic alteration variants in TFs also account for genetic diseases; for example, associations between TF alterations and cancer are well known [96]. Furthermore, in a recent review, van der Lee et al. [71] compiled a list of 46 regulatory variants with 40 TF genes associated with rare diseases with a focus on deregulated regulatory elements. Identification of variants in regulatory components is becoming common practice in disease identification [82]. However, correlations between changes in regulatory sequences and human disease are rare, and identification of therapy targets relies on comprehensive knowledge of functions and relations [132].

1.4 Peak-based motif discovery

The commonly-used technology to identify protein binding sites and subsequently sequence motifs is ChIP-sequencing [107], with the first publications in 2007 [13, 130, 87, 101]. ChIP-Seq is still used today to identify genetic sections that are bound by a protein, such as TFBS, and histone modifications. Proteins in question are bound to DNA and fixed using formaldehyd, followed by sonication of DNA strands to shear them. Antibodies are used to select DNA fragments with the bound protein, which allows separating unbound fragments. Selected DNA strands can be sequenced, resulting in a set of several million reads where a binding is observed. The raw sequences obtained are then mapped back onto the known sequence of the genome. From the mapping, a collection of sites in the genome is created where each read is found [101]. Commonly-used tools for read mapping are BWA [74] and Bowtie2 [68].

Due to molecular and technical variation, such as sequencing errors and incidental binding of proteins, reads are known to contain a distorted representation of actual binding events [126]. It is therefore not possible to directly identify a region selected by a read as a protein binding site and analysis methods are designed to adapt to the high signal-to-noise ratio. Futher, experiments are replicated and utilize high-throughput technology to increase coverage. A common approach to identify binding regions from reads is peak calling with subsequent motif discovery (Figure 1.4) [127]. Reads from ChIP-Seq experiments, which are known to bind the target protein, are accumulated based on their position on the genome. Using the distribution of those reads, locations on the genome can be identified that have an increased coverage of reads. An enriched number of reads at a position form a peak. Read accumulating reduces outlier signals and increases the probability of detecting TFBS in the near vicinity [127]. To discard read accumulations that took place due to experimental bias, peak calling tools measure significance of peaks to separate signals from noise [55]. Widespread tools for peak calling are MACS [145], SISSRS [94], QuEST [130], and Hpeak [104]. ChIP-Seq experimental data, including raw data reads and their analysis, is available online. The Encode Encyclopedia of DNA elements (ENCODE) [24, 125] currently lists about 130,000 released datasets from various species, cell lines and conditions.

Based on peak locations, binding motifs can be derived that represent binding affinity of the proteins tested. The sequence motifs are short, related patterns that are repeatedly found and express a biological meaning [42]. Different approaches exist

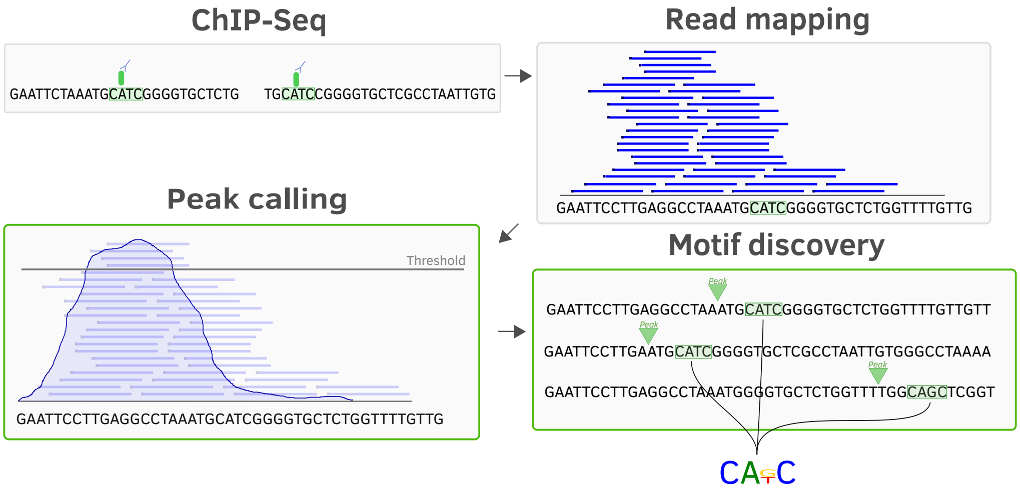

ChIP-Seq
GAATTCTAAATGCATCGGGGTGCTCTG
TGCATCCGGGGTGCTCGCCTAATTGTG
Read mapping
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTG
Peak calling
Threshold
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTG
Motif discovery
Peak
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTGTT
Peak
GAATTCCTTGAATGCATCGGGGTGCTCGCCTAATTGTGGGCCTAAAA
Peak
GAATTCCTTGAGGCCTAAATGGGGTGCTCTGGTTTTGGCAGCTCGGT
CAGC

Figure 1.4: Overview of ChIP-Seq data analysis using peak calling and motif discovery (Adapted from [85]).

to identify overrepresented sequences around peaks. They can be grouped into enumeration approaches that identify consensus sequences and probabilistic approaches that construct probability-based models [23]. Popular motif discovery tools are Homer [43], ChIPMunk [72], and MEME-ChIP [10]. There are numerous challenges in motif discovery based on ChIP-Seq data. As previously described, observations of complex protein-DNA interactions are made with multiple affinities, TF complexes of multiple TF and epigenetic features [52]. Furthermore, understanding of regulatory mechanisms is spare and an absolute standard is missing to thoroughly benchmark existing algorithms [52].

1.5 Objectives and lack of knowledge

Gene therapy will presumably be a key technology in future therapy approaches, significantly augmenting the armamentarium beyond current capabilities. The foundation for genetic alterations are delivery vectors that enable the addition and change of genomic sequences, including retroviral elements and transposases. Both systems are not artificially built, but rather their integration mechanisms evolved through the need for proliferation. They show different preferences for genetic elements, such as highly-expressed genes, histone modifications or sequence motifs. Knowledge on integration preferences is crucial to evaluate vector safety. Nonetheless, not all preferences for different vectors are known. While collections of genetic annotations increase in size, the necessary effort to review each annotation in relation to vectors also increases. Integration sites are therefore often aligned to only a small set of annotations selected as probable preferences using custom scripts or genome browsers. Moreover, integration mechanisms are constantly modified to increase safety and effectiveness. Increasing amounts of annotation and integration data reveal a demand for a software platform to perform recurring analysis of large amounts of annotation sites. Existing applications are not capable of handling the quanity of annotation data, both from a computational perspective and in reporting relevant discoveries from large-scale analysis. The lack of a scalable application leads to an insufficient usage of available data.

Analysis of genetic positions is based on comparing integration characteristics to background models. In an ideal setting, those random sites are generated using the same wet lab protocols to ensure control for experimental factors. However, additional experiments are cost- and work-intensive, and thus capturing more signal sites instead of controls is usually prefered. For this reason, random background models are often created artificially based on known experimental factors. Applications exist that create random genomic sites, although they are not capable of directly creating background models based on multiple annotations and they require the usage of programming to be combined.

To simplify the generation of background models, a platform is needed that is able to build adapted background models based on selected annotations. Only a few studies make use of adjusted background models for their experiments, which is potentially partly due to the complexity of creating them. This prompts the question of whether more reliable information can be retrieved from integration sites if adapted background models are utilized. Furthermore, it holds interest whether other insights can be gained from positional analysis by using adapted background models not only for experimental conditions, but also for known preferences. Integration preferences are not fully understood, partly due to confounding factors that interfuse effects of different genetic elements. For example, it is challenging to separatly investigate integration effects into TSS and certain histone modifications that occur near TSS. Combining different covariates and comparing results between the models can possibly reveal subtle binding preferences and help to further identify confounding factors. A software capable of instantly yielding results and revealing differences between results from varying background models can clearly improve the analysis of genomic positions and therefore integration vectors for gene therapy.

The application of gene therapy certainly relies on a deeper understanding of regulatory structures and relationships. Therefore, the extended focus of this dissertation lies on TF binding characteristics, their binding motifs and ChIP-Seq data analysis. As previously described, knowledge on TF binding is incomplete. Wet lab experiments are known to be expensive and time-consuming and have an error rate that is reflected in the lacking availability of high-quality datasets. Even with increasing data availability, the task of unbiased motif discovery is a challenging task. *De novo* motif discovery currently relies on peak calling together with motif discovery, which has been the standard protocol for ChIP-Seq data analysis. As stated before, the accumulation of peaks reduces noise in the data. Nonetheless, it also removes biologically relevant signals, which reduces the distinctivenes of the results and with low-quality datasets it can lead to a complete removal of biological signals. In addition, motif discovery uses synthetic background models that tend to overvalue the significance of motifs as they are too random [119].

Based on known limitations of the peak-based process, a novel approach to improve the detection of TF binding motifs in ChIP-Seq data is build. It implements an alternative to peak calling that derives k-mers and subsequently binding motifs directly from read distributions without the need to accumulate peaks. This prompts the question of how well a k-mer based approach on ChIP-Seq data can identify binding characteristics of TF to TFBS interaction. The usage of all available reads for TF modelling is supposedly advantageous for low-quality datasets to discovery faint binding effects, in contrast to read accumulation into peaks, where many reads are removed as outliers and low-signal peaks are omitted. The prevalence of low-quality ChIP-Seq data is well known [79] and presumably even more common when unpublished data is also considered. Therefore, a novel algorithm with new capabilities to handle low-quality data would be beneficial for future research and the re-evaluation of existing data. Furthermore, compared to common motif discovery, a direct derivation of binding sequences allows for the usage of actual background models, which would improve the statistical significance of identified motifs.

I aim to systematically investigate these assumptions in comparison to established methods and demonstrate the benefits and limitations of the k-mer-based approach. Previous analyses have shown that protein binding microarray (PBM) k-mer based models outperform PWMs in their ability to capture TF preference [29], whereby we know that k-mer based models of TF specificity are beneficial. As previously described, prior knowledge on TFs is adjuvant to perform motif discovery. An alternative to existing tools could provide support for established motif discovery pipelines and peak assessment. Moreover, direct derivation of k-mer to read distances could reveal binding site characteristics that are hidden in the read distributions, as well as faint co-factor bindings besides the primary motif.

Both publications are included in the following and supplements of the first publication can be found in the Appendix. After the original publications, the main results are summarized in Chapter 3, followed by a discussion of the results, as well as an outlook on future challenges in Chapter 4.

2 Publications

2.1 NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling

Michael Menzel[1], Sabine Hurka[1,2], Stefan Glasenhardt[1], Andreas Gogol-Döring[1]

Affiliations	[1] MNI, Technische Hochschule Mittelhessen - University of Applied Sciences, Giessen, Germany. [2] Justus Liebig University, Giessen, Germany.
Journal	Bioinformatics, *Oxford Journals*
Date, Issue	Sept 2020, 37(5)
Pages	596–602
DOI	10.1093/bioinformatics/btaa845
Supplementary data	Partialy attached. Full document available online

Contributions

The following contributions are attributed to the thesis author:

Conceptualization	Definition of project goals
Methodology	Defined metrics and evaluation
Investigation	Performed all analysis
Software	Implemented the NoPeak software package
Visualization	Created all Figures and Supplementary files
Writing	Wrote the manuscript

Bioinformatics, 37(5), 2021, 596–602
doi: 10.1093/bioinformatics/btaa845
Advance Access Publication Date: 29 September 2020
Original Paper

Genome analysis

NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling

Michael Menzel[1], Sabine Hurka [2], Stefan Glasenhardt[1] and Andreas Gogol-Döring[1,*]

[1]MNI, Technische Hochschule Mittelhessen, University of Applied Sciences, Giessen 35390, Germany and [2]Institute for Insect Biotechnology, Justus Liebig University, Giessen 35392, Germany

*To whom correspondence should be addressed.

Associate Editor: Berger Bonnie

Received on February 5, 2020; editorial decision on August 31, 2020; accepted on September 14, 2020

Abstract

Motivation: The discovery of sequence motifs mediating DNA-protein binding usually implies the determination of binding sites using high-throughput sequencing and peak calling. The determination of peaks, however, depends strongly on data quality and is susceptible to noise.

Results: Here, we present a novel approach to reliably identify transcription factor-binding motifs from ChIP-Seq data without peak detection. By evaluating the distributions of sequencing reads around the different k-mers in the genome, we are able to identify binding motifs in ChIP-Seq data that yield no results in traditional pipelines.

Availability and implementation: NoPeak is published under the GNU General Public License and available as a standalone console-based Java application at https://github.com/menzel/nopeak.

Contact: andreas.gogol-doering@mni.thm.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Regulation of gene expression is an important factor in the control of cellular processes. Thus, the elucidation of mechanisms behind gene regulation is a central research topic in the study of the cell as a complex system. Transcription factors (TFs) that interact with DNA at specific genomic regions—the TF-binding sites (TFBS)—play an essential role in this process. The location of TFBS is mostly determined by the respective DNA sequence to which the factor binds. The identification of TF-binding motifs, that is, short sequence patterns that are presumably causative for a close interaction between TF and TFBS, has long been a subject of research (Sandve and Drabløs, 2006; Stormo, 2000).

The general approach to identify sequence motifs utilizes chromatin immunoprecipitation sequencing (ChIP-Seq) (Nakato and Shirahige, 2016). This method uses antibodies that bind specifically to target proteins for extracting protein–DNA complexes. Extracted DNA fragments are sequenced and those sequencing reads are mapped to a reference genome. Due to the antibody-mediated enrichment process reads map more frequently to genomic regions with strong protein binding. TFBS can therefore be identified by detecting areas with significantly increased numbers of reads. Several tools for finding peaks in ChIP-Seq reads have been developed in recent years including MACS2 (Zhang *et al.*, 2008), SISSRs (Narlikar and Jothi, 2012), QuEST (Valouev *et al.*, 2008) and Hpeak (Qin *et al.*, 2010). Detected TFBS can then be searched for sequence motifs using motif discovery tools like HOMER (Heinz *et al.*, 2010), MEME (Bailey *et al.*, 2009), MotifSuite (Claeys *et al.*, 2012), Trawler (Dang *et al.*, 2018). These tools are compared on performance in different publications (Keilwagen *et al.*, 2011s Wilbanks and Facciotti, 2010).

Although peak calling has been established in virtually all standard methods for the analysis of ChIP-Seq data, the procedure still presents with difficulties and reproducibility of the identified peaks is not always guaranteed (Li *et al.*, 2011). The data quality supplied by ChIP-Seq often poses a problem (Marinov *et al.*, 2014), and it is particularly crucial for accurate peak calling, which may fail if, for example, the degree of accumulation by chromatin immunoprecipitation is too low or noise in the data is too high (Nakato and Shirahige, 2016). There is a strong publication bias against datasets whose analysis encounters any problems. If usual data processing methods yield no results, the data will likely be discarded from the study. There might be a considerable number of ChIP-Seq experiments that failed to be analyzed and remained unpublished. Those data would get a second chance for a reasonable evaluation if an alternative method of analysis was available.

In this publication, we present a novel approach to identify TF-binding motifs also from datasets in which peak calling only yields very few or even no peaks. ChIP-Seq ensures that the DNA is increasingly likely to be sequenced near TFBS, and thus also near TF-binding motifs. Binding motifs in the genome are characterized by an increased read frequency. It is therefore not necessary to search for peaks in the reads before searching for motifs. Instead, we may skip the intermediate step of peak calling and try to deduce TF-binding motifs directly from the distribution of reads within the genome.

Our motif discovery workflow operates on the analysis of genome-wide read distribution profiles. To generate these profiles, it is not necessary to determine TFBS, and therefore our method does not require peak calling. For each k-mer, that is, each possible DNA sequence of length *k*, we calculate one profile that represents a measure of the read density in the vicinity of the respective k-mer. The profiles have a characteristic shape for k-mers that resemble a TF-binding motif (Fig. 1). Unlike other approaches, which consider the shape of individual peaks in order to enhance peak calling (Nakato and Shirahige, 2018; Strino and Lappe, 2016; Wu and Ji, 2014), our method is based on the shape of read profiles that reflect all reads around all occurrences of the respective k-mer in the genome.

To evaluate the performance, we tested our method against common motif discovery tools in regards to peak prediction, synthetically degraded datasets and protein-binding microarray (PBM) experiments. The algorithm and other tools are combined as an application called NoPeak published under GNU General Public License version 3. Software and source code is available at https://github.com/menzel/nopeak.

2 Material and methods

Our method is separated in three phases. At first, we build read distribution profiles for each k-mer. Secondly, the profiles are filtered for quality to remove k-mers with low information content or coming from repetitive regions. The remaining profiles are scored based on their height. Finally, the k-mers of high-scoring profiles are combined into position weight matrixes (PWMs). Each phase is explained in detail in the following passage. Figure 2 gives an overview of our method in comparison to common motif discovery. We observed an average runtime <1 h on our hardware to build profiles using NoPeak, which is on a similar scale to peak calling combined with motif discovery.

Building profiles. Reads are mapped to the genome using a common mapping tool (e.g. Bowtie 2; Langmead and Salzberg, 2012) for both signal and control reads, if control reads are available. Reads that are not uniquely mapped are discarded. The mean fragment length (FL) for each dataset is estimated using the method from Gogol-Döring and Chen (2010).

Next, using mapped reads, read distribution profiles for both control and signal data are built using k-mers and their distance to the read start. To build profiles, the distribution of mapped reads around each k-mer is examined inside a fixed window (default 500 bp). The distance between each read to the k-mer is one data point for the profile that represents relative occurrence of reads around the k-mer. High profiles are observed for k-mers with higher frequencies of reads nearby, while k-mers without increased occurrence of nearby reads produce profiles that are flat and irregular. Profiles from reverse complement k-mers are combined by adding the profiles.

Reads generated by ChIP-Seq analysis are over-represented around TFBS. This fact is used in common motif discovery to identify peaks, which are small regions with an enriched read count. As shown in Figure 2, instead of identifying read-enriched regions, we use each read to analyze observable k-mers. Reads from ChIP-Seq are more abundant upstream of the TFBS. This is due to the technical process of ChIP-Seq where only DNA-fragments that are bound to the protein are kept. Therefore, the TF-specific k-mer to which the TF is attracted is assumed to be downstream of the reads.

The profile shape is determined by the distribution of reads relative to k-mers. For example, a sharp profile peak with little variance shows that all mapped reads are located in a similar distance to the observed k-mer, which could represent a biological meaning. In Section 4.2, we will discuss the possible relationship between profile shapes and biological interpretation.

In our profile model (Fig. 1), we assume two underlying technical induced factors that influence the shape and make the profile round and broader. First, all methods for DNA fragmentation produce pieces of DNA not only of a fixed length but from a spectrum of different lengths. Second, the cross-linking between proteins and DNA with formaldehyde typically used in ChIP protocols causes the TF to bind to DNA over a wider range than just at the short-binding position. In this way, also DNA fragments downstream from the actual binding site are captured by ChIP and appear in profiles as enrichment on the right of position 0. The effect is modeled by a statistical deviation of the read localization.

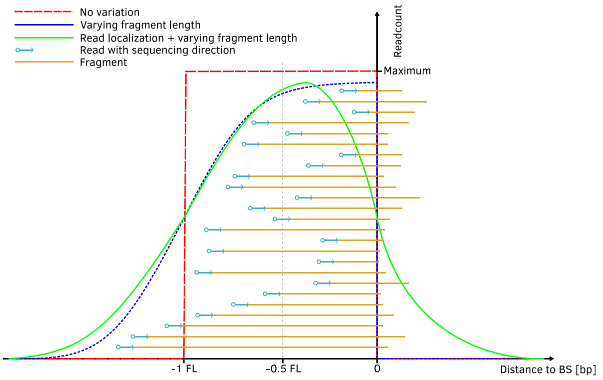


Fig. 1. Model for k-mer profiles. A profile represents the read frequency at each distance from the k-mer. Reads are cumulated over all occurrences of a k-mer in the genome. Reads from the negative strand were mirrored at the origin such that all reads point to the right. The position of a read is defined as the position of its 5′ end (blue circles). If the k-mer acts as a TF-binding motif, then the frequency of DNA fragments crossing the k-mer at position 0 is increased. The ideal profile (red line) corresponds to a rectangle the width of the fragment length (FL). However, we assume that the profile shape is influenced by two sources of variation: Varying fragment lengths and read localization deviation. The blue shape shows the impact of the varying fragment length and green the expected shape with both factors

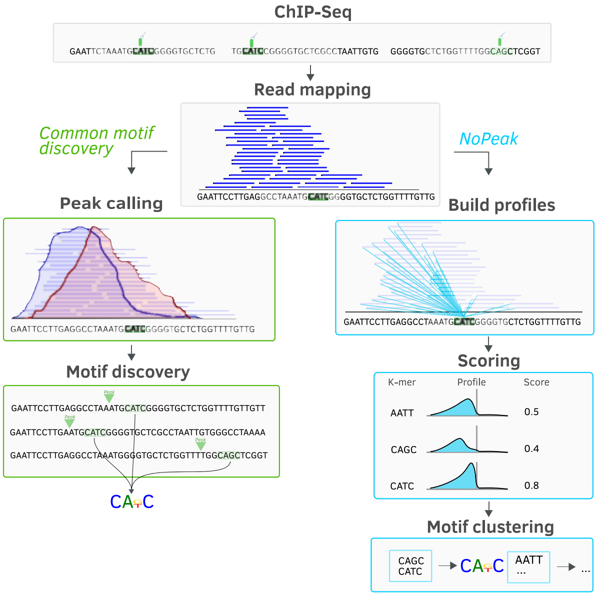

ChIP-Seq
GAATTCTAAATGCATCGGGGTGCTCTG
TGCATCCGGGGTGCTCGCCTAATTGTG
GGGGTGCTCTGGTTTTGGCAGCTCGGT
Read mapping
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTG
Common motif discovery
NoPeak
Peak calling
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTG
Build profiles
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTG
Motif discovery
Peak
GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTGTT
Peak
GAATTCCTTGAATGCATCGGGGTGCTCGCCTAATTGTGGGCCTAAAA
Peak
GAATTCCTTGAGGCCTAAATGGGGTGCTCTGGTTTTGGCAGCTCGGT
CA C
Scoring
K-mer
Profile
Score
AATT
0.5
CAGC
0.4
CATC
0.8
Motif clustering
CAGC
CATC
CA C
AATT
...
...

Fig. 2. Comparison between peak calling with motif discovery and NoPeak. For both analyses, reads extracted from ChIP-Seq experiments are mapped to the genome. Following this step, common motif discovery and NoPeak use different approaches to identify the motif. $CA(T \mid G)\ C$ is highlighted as an example sequence motif throughout the steps

We used the software ChIPulate (Datta *et al.*, 2019) to simulate reads with different fragment length and fragment jitter to observe the influence on profile shape (see Supplement Section S1.4). The simulation shows that two distributions influence variance and mean distance between reads and k-mers. Examples of actual profiles from different k-mers are shown in Figure 3.

Using control data. For a better evaluation of ChIP-Seq studies, the experimental design usually includes additional control datasets that are generated without the selection of reads by specific binding antibodies. Control data can be used to remove background noise from the actual signals data as follows. Both control and signal datasets are equally processed independently by mapping and building profiles for each k-mer. For normalizing control profiles with respect to size of the two datasets, profile heights are divided by the ratio of total read counts in control versus signal data. Then, for each k-mer, the profile from the signal dataset is divided at each position by the corresponding control profile, resulting in a profile of fold changes.

Occasionally, control and signal data generate almost identical profiles. This would cause the signal to be removed if the control file is applied. To improve automatic logo detection, we therefore implemented a filter that disables the use of control reads if high-ranking k-mers are below a fixed threshold.

Scoring profiles. So far, NoPeak has built and normalized read distribution profiles for each k-mer. These profiles are now used to score and filter the k-mers. Profiles are filtered by several criteria to remove those based on repetitive regions, technical artifacts or noise. According to our profile model (Fig. 1), a profile that corresponds to a k-mer with significant binding is assumed to have a single peak upstream and close to the BS with no drop below the surrounding level. If the shape of a profile clearly deviates from this, the corresponding k-mer is discarded.

The filters are implemented using a fixed set of rules based on our expectations toward the shape as described before. First, the mode of the curve is checked to be upstream and within the estimated fragment length of the binding site. Secondly, no sharp drop up- or downstream of the mode is allowed. The profile smoothness over a frequency of 50 bp is measured and profiles are filtered out if above a relative threshold. Finally, the profiles are filtered according to a shape score which is a measure of the correspondence between the width of the profile and FL. According to our profile model (Fig. 1), we expect the read frequency in FL distance upstream from the putative binding site to be approximately half of the maximum read frequency. The shape score is therefore defined as the difference between the profile value at position -FL and one half of the profile maximum. The profile is filtered if this score is above a certain threshold. Examples for profiles that were filtered because of their shape are added in Supplements Section S1.3.

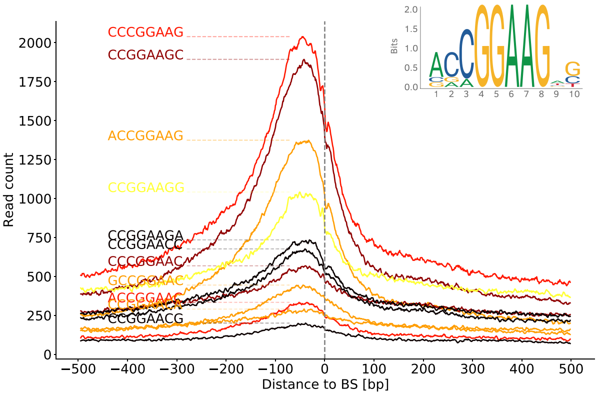


Fig. 3. Read distribution profiles for the transcription factor GABPA. A sequence logo in the upper right visualize the binding motif of GABPA as it appears in the database JASPAR (Khan *et al.*, 2018). The diagram shows the profiles of eleven 8mers that correspond most closely to the GAPBA-binding motif. Each profile shows the read start count for each position relative to the corresponding 8mer (BS) within a 1000 bp window. All reads are oriented in positive *x*-direction. Profiles are colored for better differentiation and smoothed with a moving average of window size 20 bp

From the remaining profiles, we define the height to be the difference between the maximum value and the mean of the two borders of the profile. The height of a profile is assumed to correlate with binding affinity of the corresponding k-mer and therefore serves as a score for further analysis.

At this step of the analysis, we have produced a list of scored k-mers that resembles the output produced by PBM experiments. Comparisons between the results from NoPeak and PBM are shown in Section 3.2.

Creation of sequence logos. Lists of scored k-mers contain all relevant information that is observed by NoPeak. However, they are lacking a graphic representation often needed to visually interpret TF-binding patterns. We therefore added an additional step to build sequence logos from k-mer score lists.

TFs can express multiple targeting sites through co-factors and chromatin context (Arvey *et al.*, 2012; Badis *et al.*, 2009). We observe this in our results with a set of different sequence clusters. It is therefore necessary to group observed k-mers to create sequence logos for different sequence motifs. The k-mers are combined into groups based on sequence similarity (default up to *k*/2 mismatches). k-mers from each group are aligned to the highest scoring k-mer. Using the results of the alignment, a position count matrix is generated that can easily be converted to a sequence logo for visual representation, for example, by using WebLogo (Crooks *et al.*, 2004).

3 Results

3.1 Reproducibility of results

To verify that results from NoPeak are reproducible, we used NoPeak to generate 8mer score lists for different ChIP-Seq experiments of the same TF. Figure 4 shows scores from different datasets as color intensities for top-scoring k-mers from the TF MAX. The vertical alignment of scores shows that scores are reproducible between the experiments.

While most TFs were found to be consistent across experiments and cell lines, we also identified some TFs such as GATA3 or JUN with considerably lower score similarity between replicates. When we analyzed these experiments separately, we found that the differing datasets feature varying sequence motifs as it is also present in the motifs enrichment in Factorbook (Wang *et al.*, 2012) for those TFs. This is illustrated in Supplement Section S1.2 that shows the intra-TF k-mer correlation for several relevant TFs.

3.2 Comparison to other technologies

We used NoPeak to analyze ChIP-Seq data for a set of 13 well-researched TFs to show that NoPeak is capable to capture known sequence motifs. In total, NoPeak processed 90 datasets downloaded from ENCODE (Davis *et al.*, 2018), encompassing several replicates

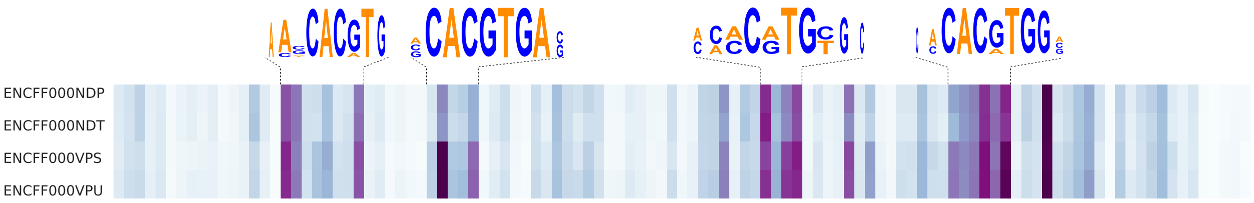


Fig. 4. This heatmap shows 109 k-mer scores of four normalized ChIP-Seq datasets of the TF MAX that are present in each experiment. The color intensity represents the score. k-mer columns are ordered by sequence similarity using CLUSTAL W (Thompson *et al.*, 1994). High-ranking k-mer groups are combined to a sequence logo by WebLogo (Crooks *et al.*, 2004). The expected motif for TF MAX is shown in Figure 5. Vertical alignment of k-mer scores illustrates that k-mer scores are similar across experiments

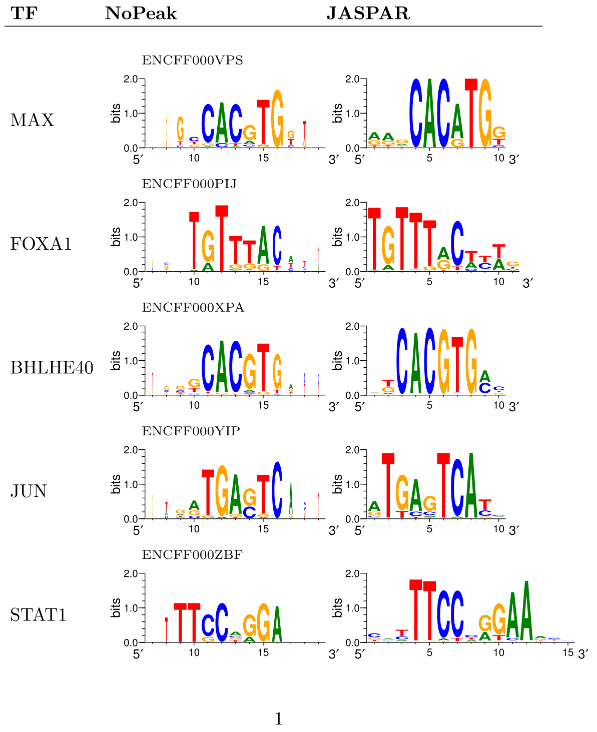


Fig. 5. Comparison of sequence logos build using NoPeak and from JASPAR for five well-analyzed TFs. ENCODE experiment IDs are noted above each logo

and experiments from different cell lines per TF. Reads were mapped using Bowtie 2 and sequence motifs were created using NoPeak. Figure 5 shows that NoPeak is able to reproduce the expected sequence logos. Additional sequence logos for more TFs and datasets for each TF are available in Supplementary Material (Supplementary Section S1.6).

To evaluate our results further, we compared NoPeak to results from PBM. It is a well-established procedure for measuring binding affinities of a given TF (Berger and Bulyk, 2006) to different DNA sequences. From PBM experiments, it is possible to derive scores for binding affinity of TFs to each k-mer (e.g. for $k = 8$). In order to compare the results from our method to PBM, we searched the Cis-BP database (Weirauch *et al.*, 2014) for TFs for which both ChIP-Seq and PBM data are available.

ChIP-Seq reads were mapped to the hg38 reference genome (Schneider *et al.*, 2017) and k-mer scores generated using NoPeak. NoPeak scores were then correlated to the corresponding *Z*-scores from PBM experiments for each TF and sequence logos generated from both techniques. We found a correlation of k-mer scores and visual similarity between the sequence logos (Supplement Section S1.1).

3.3 Discriminating peaks from controls

In order to find out how accurately NoPeak describes binding preferences of TFs compared to a conventional peak-based motif discovery, we have examined the capability of different tools to distinguish between DNA sequences at ChIP-Seq peaks and random control sequences. For this, we analyzed all SYDH datasets from ENCODE (Davis *et al.*, 2018; ENCODE Project Consortium *et al.*, 2012) comprising 806 experiments from 91 TFs. A full list of experiments is available in Supplement Section S1.10. As positive sets, we selected 100 bp sequences around the top 500 peaks identified by MACS2 (Zhang *et al.*, 2008). The control sequences were generated as described before (DREAM5 Consortium *et al.*, 2013) by using three methods: (1) randomly sampling from the genome, (2) randomly sampling from promotor regions, and (3) shuffling the sequences from the positive set in a way that preserves dinucleotide frequencies. Each method was used to generate 500 control sequences, so in total, we generated 1500 control sequences per dataset, each of length 100 bp.

A prediction model was trained on each experiment with specific control experiments. Subsequently, based on leave-one-out cross validation, we used each model to predict each experiment within the TFs and calculated the area under curve for the receiver-operating characteristics curve (AUC-ROC). The mean of the AUC-ROCs for each experiment was noted as prediction score.

For the comparison, we used three well-established motif discovery tools (HOMER, Heinz *et al.*, 2010; ChIPMunk, Levitsky *et al.*, 2014; MEME, Bailey *et al.*, 2009). Each tool was used to generate PWMs based on the discovered peaks. For MEME and ChIPMunk, we limited the training peaks to 1000 top-scoring peaks to reduce runtime. HOMER was configured to build a model with three motifs. MAST from the MEME Suite (Bailey *et al.*, 2009) was used to score the sequences from both positive and controls sets against the PWMs.

Using NoPeak, we trained a k-mer-based model on each experiment. The model consisted of a list of k-mers and their associated scores. To evaluate a potential peak sequence, we identified all 8mers inside the sequence with a sliding window and summed the k-mer scores learned by NoPeak for this TF. We allowed up to one mismatch and reverse-complements and used only the 100-top k-mers to account for the selection of top peaks from the experiments.

A flowchart of the analysis is added to Supplements Section S1.8. A heatmap showing the mean AUC-ROC scores for each TF is given in Figure 6.

3.4 Analysis of degraded datasets

Because our method does not rely on peak calling, we hypothesized that NoPeak may be able to successfully analyze ChIP-Seq datasets where common peak calling pipelines struggle. To evaluate this hypothesis, we iteratively removed reads from ChIP-Seq experiments and tested NoPeak and HOMER on motif quality. In each iteration, we used MACS2 (Zhang *et al.*, 2008) for peak calling and removed all reads near discovered peaks. The process was repeated until no detectable peaks were left. In each round, we used NoPeak to predict motifs from the remaining reads as well as HOMER to discover motifs within the remaining peaks. Motifs from NoPeak and HOMER were scored against a reference motif from Factorbook (Wang *et al.*, 2012) using the Jaccard distance calculated with MACRO-APE (Vorontsov *et al.*, 2013).

The scores for three TFs are shown in Figure 7. With each iteration, the score of HOMER decreased until MACS2 could not identify any peaks and consequently, HOMER could not discover any sequence motifs. NoPeak could keep a high score throughout the peak removal test for all three TFs.

We analyzed the read distribution of the highest profiles generated by NoPeak in each round and observed that profile heights were only decreased in the first round of read removal when the largest number of reads was removed. In the following rounds, only few reads were removed for eliminating remaining peaks, which had only minor impact on read distribution profiles. Additional information and sequence logos for each reduction step from both NoPeak and HOMER are added in Supplement Section S1.5.

Further, we measured the correlation of k-mer scores between the first and last experiment which resulted in a Pearson Correlation Coefficient (PCC) of 0.98 for FOXA1, 0.91 for MAX and 0.97 for JUN. This indicates that read removal from peaks does not affect k-mer scores generated using NoPeak.

4 Discussion

4.1 Performance of NoPeak

We have shown that NoPeak is able to capture TFBS information similarly well as conventional tools that are based on peak calling and subsequent motif discovery. The results calculated by NoPeak are reproducible and correspond to PWMs available in databases (Khan *et al.*, 2018; Wang *et al.*, 2012). One advantage of our

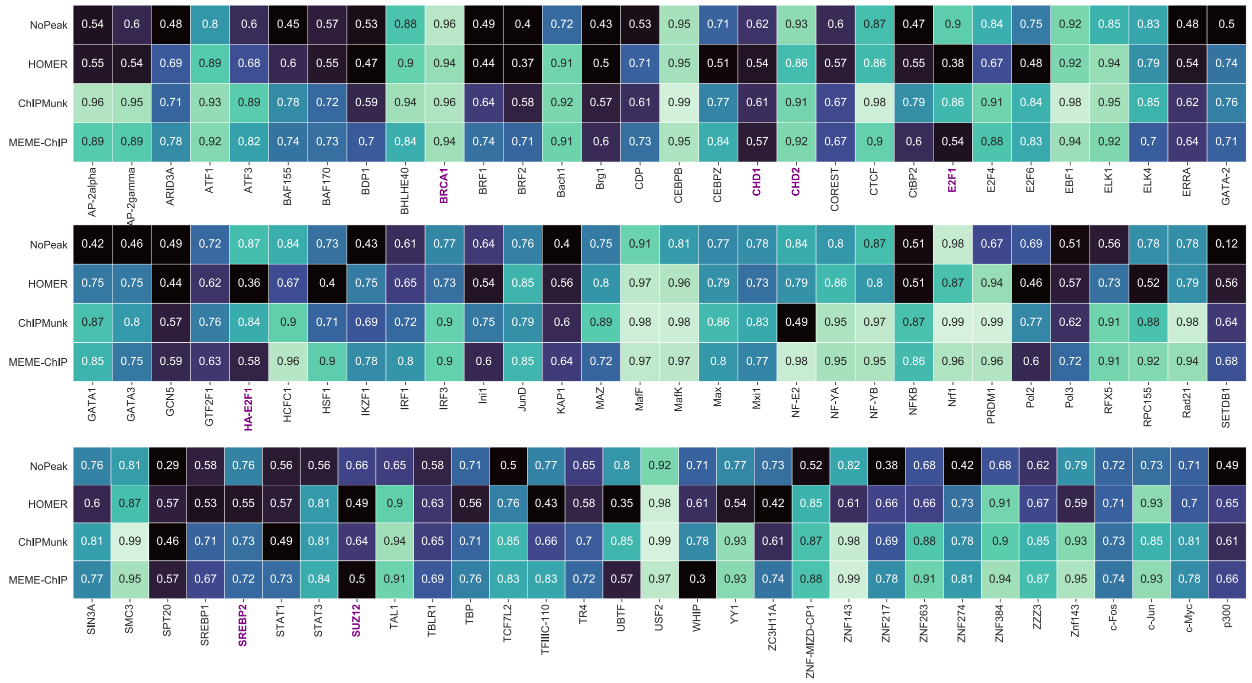


Fig. 6. Mean AUC-ROC scores for prediction of peaks from ChIP-Seq experiments. NoPeak and common motif discovery tools are compared on peak prediction capabilities of 100 bp sequences around top peaks from each experiment and random sequences in a leave-one-out cross validation setting. TFs where NoPeak outperformed all other tools are highlighted with bold font

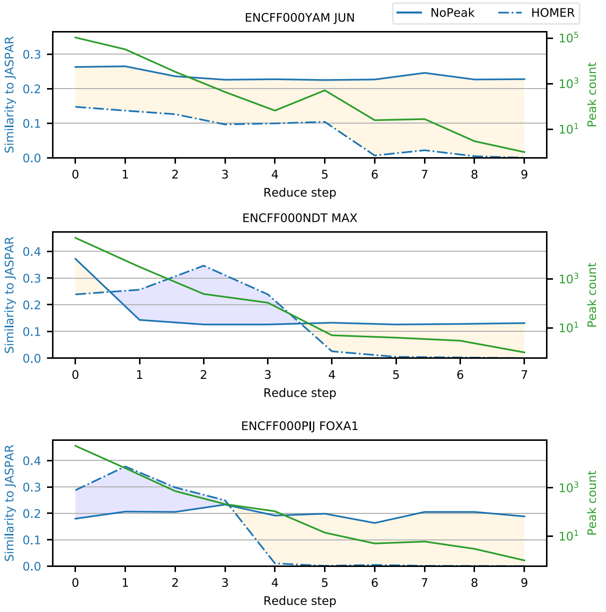


Fig. 7. Motif score of NoPeak and HOMER on degraded datasets. Three series of different TFs (FOXA1, JUN, MAX) were degraded step-wise by removing reads from identified peaks until no peaks were identified. In each step, the remaining reads were used to identify sequence motifs with NoPeak and HOMER and scored against the expected motif for the corresponding TF

approach is that it is independent of peak calling. Even after the removal of all peaks from datasets, the information about TF-binding preferences still remains subtly hidden in the remaining reads and can be revealed using by NoPeak. This implies that our approach is suitable for experiments where peak calling finds no or too few reliable TFBS.

The consistency of NoPeak results and PBM experiments and between different ChIP-Seq replicates shows that our method is robust for high-scoring k-mers. The scores of low-scoring k-mers on the other hand are less reproducible and may vary between different ChIP-Seq experiments as well as in comparison to PBM. For continuative analyses of k-mer-to-score results from NoPeak, only high-scoring k-mers should be used. Low-scoring k-mers are subject to higher noise and are more specific to the dataset than the biological entity.

Although NoPeak was not designed for the identification of individual peaks and it does not consider information about peak positions in its analysis, the performance of our tool in predicting peak regions is comparable to peak-based methods and it even outperformed the motif discovery tools in 7% of the experiments. Averaged over all examined TFs, NoPeak achieved a mean AUC-ROC of 0.68, which is similar for HOMER (0.67) but below ChIPMunk (0.8) and MEME-ChIP (0.79). However, it should be noted that TFBS detected in ChIP-Seq replicates usually overlap, hence the motif discovery tools had the advantage of partially predicting peaks that they already used to build the model, while NoPeak could not benefit from this inevitable mixing of training and test data.

PWMs are a widely used motif model (Stormo, 2000), yet are known to not capture all TFBS information, as they lack the ability to represent variable gaps, position dependencies and multiple targets. Comparative studies have shown that k-mer-based methods can perform better in capturing TFBS information content (DREAM5 Consortium *et al.*, 2013), and that scored k-mer lists are more robust against single-base differences (Guo *et al.*, 2018). NoPeak is a k-mer-based approach, and we encourage to directly use the k-mer-based model for interpreting TFBS.

A possible disadvantage of NoPeak could be that our approach is less able to handle very wide sequence motifs, as the number of k-mers and thus of profiles grows exponentially with the length *k*. At the same time, the average number of occurrences of each k-mer in the genome decreases, so that on average fewer reads are counted in each profile and the profiles become more noisy. We tested k-mers

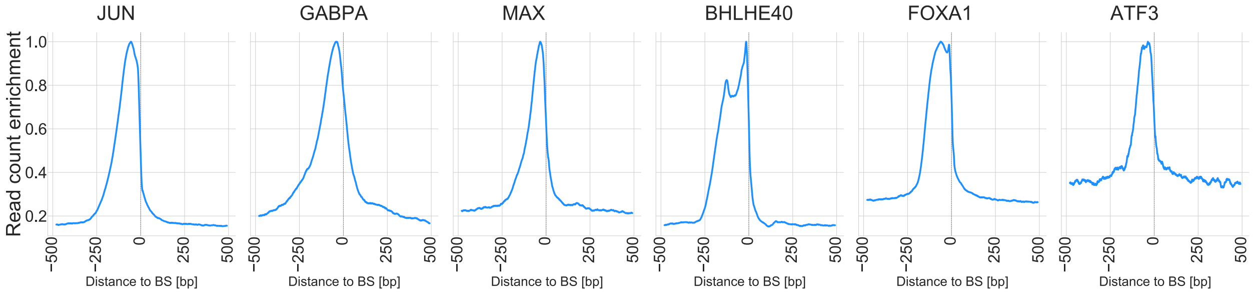


Fig. 8. k-mer profiles for six different TFs. Different TFs show distinguishable shapes. We assume that profile shape corresponds with binding characteristics of the TF

up to 12 bp, yet found that lower lengths lead to more distinctive profiles. Longer motifs could be discovered through the combination of shorter overlapping k-mers.

4.2 Further evaluation of profile shapes

Read distribution profiles show distinct shapes that could possibly be used to reveal further binding site characteristics of TF binding. The profile shape is a direct product of the distribution of reads around binding sites and corresponding k-mers. Peak shape properties that might be useful for more detailed characterization of TF binding include multiple local peaks, change in profile width, plateaus and their width, downstream decline and drop-off ratio. These profile shape characteristics are applicable as an indication on binding site structure.

Figure 8 shows top profiles for different TFs. It is apparent that profile shape is diverse between TF: The shape for GABPA for example is a uniform profile for up- and downstream enrichment, while MAX presents as skewed in up-stream direction with more reads up-stream of the peak. Furthermore, FOXA1 shows a two-peak profile with peaks at a similar height. This could be related to the shadow peaks that are induced by estrogen-mediated chromatin loops (Glont *et al.*, 2019). BHLHE40 also has two peaks, but the first of the two peaks is significantly below the peak closer to the binding site. ATF3 is known to bind as a dimer (Jadhav and Zhang, 2017) and has a double peak for the top motif. Additional profiles for all experiments are available in Supplements Section S1.7. A deeper analysis and interpretation of different profile shapes will be the subject of future work. We also analyzed data from CUT&RUN and ChIP-exo experiments and found significant read profiles for the expected k-mers (Supplement Sections S1.13 and S1.14).

5 Conclusion

With NoPeak, we present a novel method to analyze data from ChIP-Seq experiments that do not require determination of peaks and instead rely on the analysis of genome-wide profiles. The shape of these profiles depends on the TF and contains information about binding characteristics. NoPeak is capable of finding valid TF-binding motifs in experiments that are not utilizable with traditional motif discovery methods. The correspondence between PBM and NoPeak k-mer scores for high-ranking k-mers, as well as the ability of NoPeak to distinguish real from random ChIP-Seq peak sequences supports that the results delivered by NoPeak are biologically relevant. We therefore consider NoPeak to be usable as a second-stage tool for motif discovery if common pipelines fail due to low read count or peak quality.

By involving ChIP-Seq control datasets, NoPeak provides the unique feature of using an empirical background model in motif discovery, while conventional tools have to rely on potentially questionable artificially constructed control sequences to identify overrepresented motifs (Wilbanks and Facciotti, 2010).

NoPeak generates scored k-mer lists, which opens the possibility to utilize methods from PBM analysis such as Seed-and-Wooble (Berger and Bulyk, 2006) or BEEML-PBM (Zhao and Stormo, 2011).

Since NoPeak's analysis is not limited to peak regions but extracts information from complete ChIP-Seq datasets, we believe that our approach may be more robust than conventional methods, both with respect to noise in the data and being less susceptible to possible collocations of binding motifs with sequence patterns that are irrelevant to the actual binding. With NoPeak, it might even be possible to detect subtle binding patterns that are too weak to cause distinct peaks.

The capabilities of NoPeak could be further enhanced by building a more intrinsic model for profile shape evaluation that is able to allow TF-specific variation, yet filters artifacts. Currently, NoPeak is set up to work with a fixed k-mer length in each run, an additional approach to combine different lengths k-mers could easily be evaluated. This is also associated with the capability to identify long motifs, as shorter k-mers can be combined to identify larger motifs. To predict binding sites, we scored sequences by the sum of k-mers found in the sequence in question. This approach could be enhanced by using a framework that uses energy-scoring like BEEML-PBM to better capture binding preferences. As described, sequence logos present a visual way to store binding site information, yet it cannot capture all information NoPeak is able to identify. Instead of using PWMs as a result, we recommend to use scored k-mer lists. However, a pipeline is needed to work with those k-mer lists as well as a new visual representation. An application to analyze k-mer lists together with an advanced sequence scoring framework could form a new application to improve motif discovery with NoPeak and subsequently be used to identify peaks using k-mer lists.

Acknowledgements

This publication represents a component of my doctoral thesis in the Faculty of Biology and Chemistry at the Justus-Liebig-University Giessen, Germany and the Research Campus of Central Hessen (RCCH).
Financial Support: none declared.

Conflict of Interest: none declared.

References

Arvey,A. *et al.* (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.

Badis,G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Berger,M.F. and Bulyk,M.L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. In: *Gene Mapping, Discovery, and Expression*. Vol. **338**. Humana Press, New Jersey. pp. 245.

Claeys,M. *et al.* (2012) MotifSuite: workflow for probabilistic motif detection and assessment. *Bioinformatics*, **28**, 1931–1932.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Dang,L.T. *et al.* (2018) TrawlerWeb: an online de novo motif discovery tool for next-generation sequencing datasets. *BMC Genomics*, **19**.

Datta,V. *et al.* (2019) ChIPulate: a comprehensive ChIP-seq simulation pipeline. *PLoS Comput. Biol.*, **15**, e1006921.

Davis,C.A. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

DREAM5 Consortium *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.

ENCODE Project Consortium *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Glont,S.-E. *et al.* (2019) Comprehensive genomic analysis reveals that the pioneering function of FOXA1 is independent of hormonal signaling. *Cell Rep.*, **26**, 2558–2565.e3.

Gogol-Döring,A. and Chen,W. (2010). Finding optimal sets of enriched regions in ChIP-Seq data. In: *German Conference on Bioinformatics 2010*. Gesellschaft für Informatik eV.

Guo,Y. *et al.* (2018) A novel *k*-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Res.*, **28**, 891–900.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Jadhav,K. and Zhang,Y. (2017) Activating transcription factor 3 in immune response and metabolic regulation. *Liver Res.*, **1**, 96–102.

Keilwagen,J. *et al.* (2011) De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput. Biol.*, **7**, e1001070.

Khan,A. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Levitsky,V.G. *et al.* (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics*, **15**, 80.

Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Marinov,G.K. *et al.* (2014) Large-scale quality analysis of published ChIP-Seq data. *G3 (Bethesda)*, **4**, 209–223.

Nakato,R. and Shirahige,K. (2016) Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief. Bioinformatics, **18**(2), 279–290, bbw023.

Nakato,R. and Shirahige,K. (2018) Sensitive and robust assessment of ChIP-Seq read distribution using a strand-shift profile. *Bioinformatics*, **34**, 2356–2363.

Narlikar,L. and Jothi,R. (2012). ChIP-Seq data analysis: identification of protein–DNA binding sites with SISSRs peak-finder. In *Next Generation Microarray Bioinformatics*. Humana Press, 2012.. 305–322.

Qin,Z.S. *et al.* (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

Sandve,G.K. and Drabløs,F. (2006) A survey of motif discovery methods in an integrated framework. *Biol. Direct*, **1**, 11.

Schneider,V.A. *et al.* (2017) Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.

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

Strino,F., and Lappe,M. (2016) Identifying peaks in *-Seq data using shape information. BMC Bioinformatics,**17**, 343–361.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

Vorontsov,I.E. *et al.* (2013) Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.*, **8**, 23.

Wang,J. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

Weirauch,M. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS One*, **5**, e11471.

Wu,H. and Ji,H. (2014) PolyaPeak: detecting transcription factor binding sites from ChIP-Seq using peak shape information. *PLoS One*, **9**, e89694.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS.). *Genome Biol.*, **9**, R137.

Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.

2.2 Enhort: a platform for deep analysis of genomic positions

Michael Menzel[1], Peter Koch[1], Stefan Glasenhardt[1], Andreas Gogol-Döring[1]

Affiliations	[1] MNI, Technische Hochschule Mittelhessen - University of Applied Sciences, Giessen, Germany.
Journal	PeerJ Computer Science
Date, Issue	2019, 5
Pages	e198
DOI	https://doi.org/10.7717/peerj-cs.198
Supplementary data	Available online

Contributions

The following contributions are attributed to the thesis author:

Conceptualization	Definition of project
Methodology	Definied software and capabilities
Investigation	Performed all analysis
Software	Implemented the Enhort platform
Visualization	Created Figures 2,3,4,5 and Supplementary files
Writing	Wrote the manuscript

Enhort: a platform for deep analysis of genomic positions

Michael Menzel, Peter Koch, Stefan Glasenhardt and Andreas Gogol-Döring

MNI, Technische Hochschule Mittelhessen—University of Applied Sciences, Giessen, Hessen, Germany

ABSTRACT

The rise of high-throughput methods in genomic research greatly expanded our knowledge about the functionality of the genome. At the same time, the amount of available genomic position data increased massively, e.g., through genome-wide profiling of protein binding, virus integration or DNA methylation. However, there is no specialized software to investigate integration site profiles of virus integration or transcription factor binding sites by correlating the sites with the diversity of available genomic annotations. Here we present Enhort, a user-friendly software tool for relating large sets of genomic positions to a variety of annotations. It functions as a statistics based genome browser, not focused on a single locus but analyzing many genomic positions simultaneously. Enhort provides comprehensive yet easy-to-use methods for statistical analysis, visualization, and the adjustment of background models according to experimental conditions and scientific questions. Enhort is publicly available online at enhort.mni.thm.de and published under GNU General Public License.

Subjects Bioinformatics, Computational Biology

Keywords Virology, Data analysis, Genome annotation, Next-generation sequencing, Integration profiling

Submitted 13 November 2018
Accepted 12 May 2019
Published 10 June 2019

Corresponding author
Andreas Gogol-Döring,
andreas.gogol-doering@mni.thm.de

Academic editor
James Procter

Additional Information and Declarations can be found on page 10

DOI 10.7717/peerj-cs.198

INTRODUCTION

Some viruses like HIV (*Craigie & Bushman, 2012*) and AAV (*Deyle & Russell, 2009*) are able to copy their genomic sequence into the genome of an infected cell. This can have severe impact on host cell stability as the integration may hit and disable a gene or a regulatory region. The investigation of characteristics and underlying driving factors for virus integration is not only relevant for virology and infectious diseases research but also for approaches in gene therapy that apply virus-derived vectors and transposons to deliver functional DNA fragments into host cells (*Riviere, Dunbar & Sadelain, 2012*; *Li et al., 2015*). Each gene delivery system has its own mechanisms for genomic integration and preferences for choosing integration sites, hence different systems may have different risks for causing undesired side effects.

Next Generation Sequencing (NGS) facilitates the genome-wide profiling of integration sites, as they are collected e.g., in investigations of protein binding, virus/transposon integration or DNA methylation. Integration sites are available from databases like the Retrovirus Integration Database (*Shao et al., 2016*) and are regularly created for novel targeted vectors. Typically, the identified sites are related to a variety of genomic features and any integration preferences are determined by a comparison of actual integration sites to a set of random control sites (*Gogol-Döring et al., 2016*). A proper background

How to cite this article Menzel M, Koch P, Glasenhardt S, Gogol-Döring A. 2019. Enhort: a platform for deep analysis of genomic positions. *PeerJ Comput. Sci.* **5**:e198 http://doi.org/10.7717/peerj-cs.198

model should mimic all known biases of the signal data originating from experimental or laboratory conditions. If, for example, a profiling method is only capable of detecting integration events that are close to certain enzyme restriction sites then the control sites should also be selected accordingly.

Several tools have been published that are capable of processing genomic positions and annotations, like the Genomic HyperBrowser (*Sandve et al., 2013*). Genome browsers like the UCSC Genome Browser (*Kent et al., 2002*), IGV (*Robinson et al., 2011*) or Artemis (*Carver et al., 2011*) are designed for inspecting single genomic locations. Also custom written scripts are commonly used for the analysis of genomic positions (*Cook et al., 2014*) or libraries like PyBedTools (*Janovitz et al., 2014*; *Dale, Pedersen & Quinlan, 2011*). Once written these scripts have the benefit of being a reusable option to conduct a specific set of analysis on recurring data. However, they are limited by the available functionality because each function has be newly developed. Additionally, comparability across laboratories is afflicted by varying functionality and different implementations of background models. There is yet no specialized tool for genomic positions analysis that combines the features of instant analysis and user defined adaptable background models that mimic known biases.

In this paper we present Enhort, a user-friendly web-platform for deep analysis of large sets of genomic positions. Our aim is to accelerate and simplify the data analysis process as well as to standardize it in order to increase reproducibility. Enhort is capable of adjusting background sites used for comparison by user selected covariates. This includes annotation tracks like restriction sites or chromatin accessibility, gene expression tracks and sequence motifs. With covariates it is possible to adjust the background sites selection in a way that they match the investigated sites for a specific track. The adaptation rules out the effects of this annotation for the background. This feature can be used to adjust for experimental bias as well as specific questions. Figure 1 shows the schematic process of data gathering and the usage of Enhort in the workflow of analyzing genomic positions.

METHODS

Integration sites of viruses are gathered by sequencing infected cells and preprocessing as shown in Fig. 1. These sites are uploaded to Enhort and are intersected with each annotation file to compute fold-change enrichment and χ^2 test in comparison to control sites, yielding a measure for effect strength and significance of each annotation respectively. Figure 2 shows the schematic analysis pathway for sites uploaded by a user. Statistical analysis depends on the Apache Commons Math library (https://commons.apache.org/proper/commons-math/) and uses Bonferroni correction for multiple hypothesis testing. The libraries plotly.js (https://plot.ly/javascript/) and Circos (http://circos.ca/) are used for visualization. The results are sorted according to their relevance and presented in conjunction with appropriate figures. Example results for a virus can be seen in Fig. 3A. The software has been designed in a way that analysis results are almost immediately available after upload.

In many cases a background model consisting of random sites is not sufficient for an adequate analysis. Some protocols, for example, can only detect integration events that

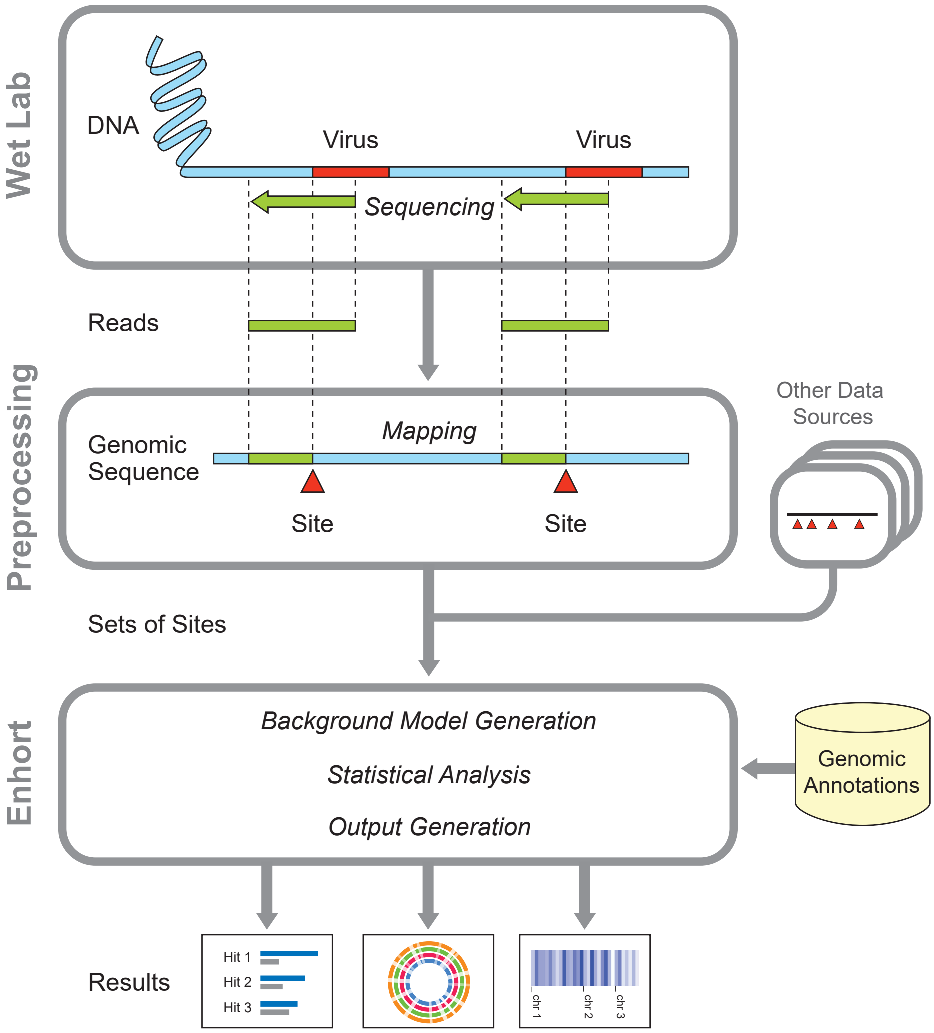


Figure 1 **Overview of preparatory work and data gathering for analysis in Enhort.** Reads containing viral integration sites are identified and sequenced in the WebLab and mapped to a reference genome. Identified insertion sites are converted to a BED file for the usage in Enhort. Together with genomic annotations from public database the analysis in Enhort is conducted to generated analysis of the given integration sites.

Full-size DOI: 10.7717/peerjcs.198/fig-1

occurred in close proximity to a restriction site of a specific enzyme, like EcoRI, which cuts inside of GAATTC hexamers (*Pingoud & Jeltsch, 2001*). Background models should be adapted to mimic the actual integration pattern with regard to any known technical bias. In this case, the control sites should also be selected to be near restriction sites. This can be achieved in Enhort by setting the appropriate genome annotation as a covariate. When selecting the track that contains all possible genomic positions of GAATTC hexamers as covariate, Enhort will generate a set of control sites having exactly the same distribution of distances to the enzyme restriction sites as the actual virus integration sites.

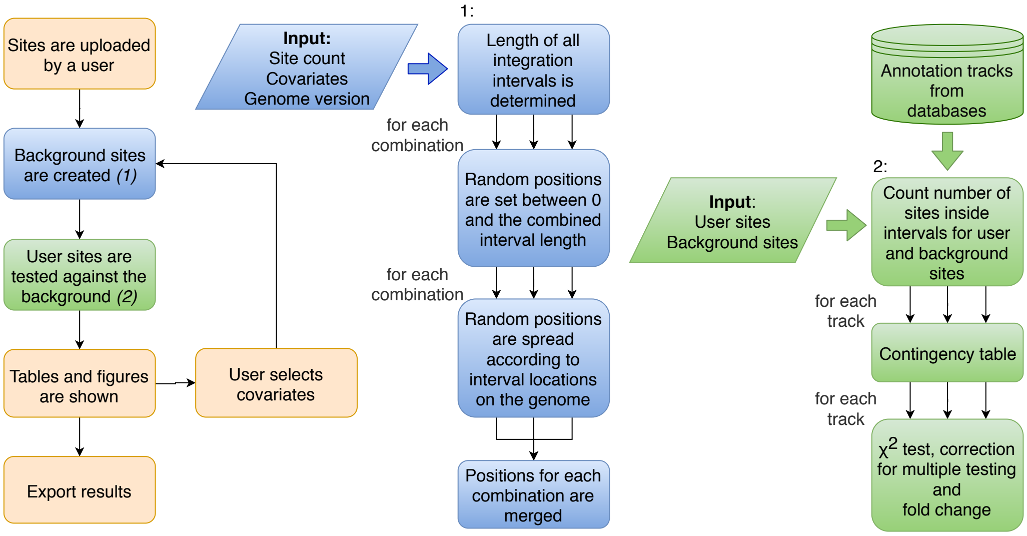

Sites are uploaded by a user
Background sites are created *(1)*
User sites are tested against the background *(2)*
Tables and figures are shown
User selects covariates
Export results
Input: Site count Covariates Genome version
1:
Length of all integration intervals is determined
for each combination
Random positions are set between 0 and the combined interval length
for each combination
Random positions are spread according to interval locations on the genome
Positions for each combination are merged
Input: User sites Background sites
Annotation tracks from databases
2:
Count number of sites inside intervals for user and background sites
for each track
Contingency table
for each track
χ^2 test, correction for multiple testing and fold change

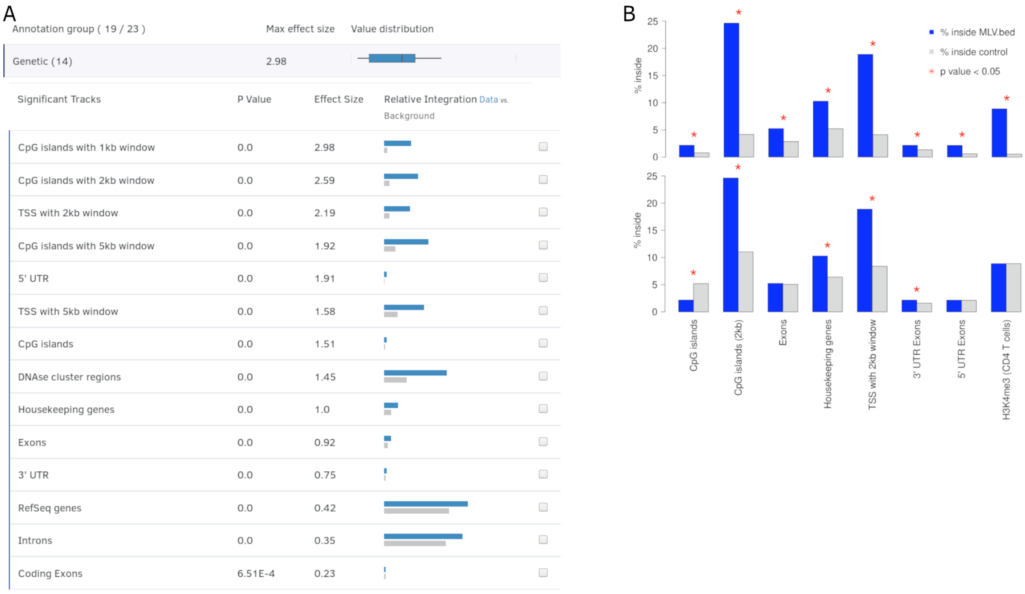
Figure 2 **Flowchart of the procedure of analysis performed by Enhort.** Blue boxes show the steps to create a background model based on multiple covariates. Random positions have to be set for each combination of covariates. Green boxes show the steps to test the user sites against the background sites. The results are returned as a table and converted into figures for the user.

Full-size DOI: 10.7717/peerjcs.198/fig-2

Covariates help to adapt the background model both for technical circumstances, for example, restriction sites and for eliminating a bias or biological preferences such as motifs or genetic features. Covariates can also be used to identify dependent or separate weak integration preferences that are covered by stronger effects, as shown in Fig. 3B. MLV integration sites are compared to two different control sets: A random and an altered background, to identify the actual integration preferences; e.g., for histone mark H3K4me3, which is a known preference of MLV (*Gogol-Döring et al., 2016*).

For the validity of statistical testing it is usually indispensable to normalize the background model relative to multiple covariates. For that purpose, Enhort supports the selection of multiple covariates simultaneously in order to further investigate the integration site characteristics. For example, Enhort may create a control set that considers chromatin accessibility, restriction site distance as well as several histone modifications simultaneously. This functionality is needed to build background models for sites that are influenced by multiple factors, e.g., biological and technical biases. A set of additional features listed in the following table:

1. Statistical analysis for annotation tracks:
   (a) Fold change
   (b) χ^2 test
   (c) Kolmogorov–Smirnov test
2. Hotspot analysis (Fig. 4C)
3. Position depended enrichment (Fig. 4A)
4. Background models based on:



Annotation group (19 / 23)	Max effect size	Value distribution
Genetic (14)	2.98	

Significant Tracks	P Value	Effect Size	Relative Integration Data vs. Background
CpG islands with 1kb window	0.0	2.98	□
CpG islands with 2kb window	0.0	2.59	□
TSS with 2kb window	0.0	2.19	□
CpG islands with 5kb window	0.0	1.92	□
5' UTR	0.0	1.91	□
TSS with 5kb window	0.0	1.58	□
CpG islands	0.0	1.51	□
DNAse cluster regions	0.0	1.45	□
Housekeeping genes	0.0	1.0	□
Exons	0.0	0.92	□
3' UTR	0.0	0.75	□
RefSeq genes	0.0	0.42	□
Introns	0.0	0.35	□
Coding Exons	6.51E-4	0.23	□

Figure 3 Output view example, generated by Enhort when analyzing Murine Leukemia Virus (MLV) integration sites in CD4 $^+$ T cells (*Roth, Malani & Bushman, 2011*). (A) The results are presented in a table containing for each annotation the *p* value, effect size and a visual representation of the integration. The annotations are ranked by effect strength. (B) Effect of covariate selection. The upper diagram contains integration frequencies of MLV compared to random sites for a selection of annotations. This virus is known for preferentially integrating near transcription start sites (TSS) and H3K4me3 histone marks (*LaFave et al., 2014*). The lower diagram shows the same data after selecting H3K4me3 as covariate. The adapted background model is generated in a way that control sites and MLV integration sites have the same frequency relative to H3K4me3. This also changed the control site frequencies for other annotations: MLV integration is no longer enriched but depleted in CpG islands when compared to the adapted background model.

Full-size DOI: 10.7717/peerjcs.198/fig-3

(a) Inside and outside of annotations
(b) Distance to annotations
(c) Scored annotations
(d) Sequence logo

5. Upload background sites
6. Comparing effects of different background models
7. Batch analysis of multiple integration sets
8. Heatmaps to compare integration sets (Fig. 4B)
9. Custom annotation tracks
10. Blend annotation tracks
11. Export results as R code and CSV files

Enhort is separated into a lightweight, web-based user interface and a high performance back-end server attached to a SQLite database storing meta-information about the annotations fetched from DeepBlue (*Albrecht et al., 2016*). Results from Enhort are instantaneously available as seen in Table 1 where the run times for different input sizes are shown. Our application currently offers 1402 annotation tracks from 97 cell

Table 1 **Analysis execution times for different usual site counts, annotation tracks from hg19 and covariate counts.** (Back-end server: SuperMicro SuperServer 4048B-TRFT 4x Intel Xeon E7-8867v3 with 2048GB DDR3 ECC LR).

Track count	**23**			**1,127**		
Covariate count	**0**	**2**	**5**	**0**	**2**	**5**
Site count	**Execution time (ms)**					
150k	877	1,188	4,668	8,538	10,436	12,540
125k	717	1,103	5,628	5,509	7,552	7,975
100k	749	817	5,085	3,724	4,672	8,673
75k	624	571	4,019	4,905	4,397	9,633
50k	470	555	5,455	4,736	5,844	10,451
25k	308	351	4,628	3,246	3,091	8,111

lines and tissues for human genome assemblies hg19 and hg38, downloaded from UCSC Genome Browser (*Fujita et al., 2011*), Encode (*ENCODE Project Consortium, 2004*), ChIP-Atlas (http://chip-atlas.org), BLUEPRINT Epigenome (*Adams et al., 2012*) and Roadmap Epigenomics (*Roadmap Epigenomics Consortium et al., 2015*) using the DeepBlue Epigenomic Data Server (*Albrecht et al., 2016*).

RESULTS AND DISCUSSION

Literature review

We reviewed the relevance of Enhort for contemporary research by systematically searching PubMed, Google Scholar, and several review articles for publications concerning the analysis of genomic integration sites. The publications include virus integration site analysis for HIV, MLV, HRP-2, SIV, foamy virus, HPV, AAV and transposons such as piggyBac, LINE-1, Alu and sleeping beauty. In total we identified 59 relevant publications. Details on the reviewed publications and methodological analysis are available in the Table S1. Of these publications 19 used completely random control sites, only six used adapted control sites. The data analyses presented in 37 (63%) publications could have been entirely performed with our tool. Six further publications use at least some methods provided by Enhort. We assume that if they had the opportunity to use Enhort the authors would have saved a lot of effort writing custom analysis scripts.

Data re-analysis

To further present the capabilities of Enhort we re-analyzed integration sites of the PiggyBac transposon (PB) published by *Gogol-Döring et al. (2016)* using Enhort. Results from *Wilson, Coates & George (2007)* are used for comparison. PB integration characteristics show a preference for genes, exons, introns, highly expressed genes, DNase I hypersensitive sites, H3K4me3 and open chromatin structures (*Wilson, Coates & George, 2007*; *Li et al., 2013*). We uploaded the PB integration sites to Enhort, selected all relevant tracks and finally exported the results. Figure 5A shows the log fold changes for a selection of annotations for PB against a random background in grey. Figure 5B shows the sequence logos for the

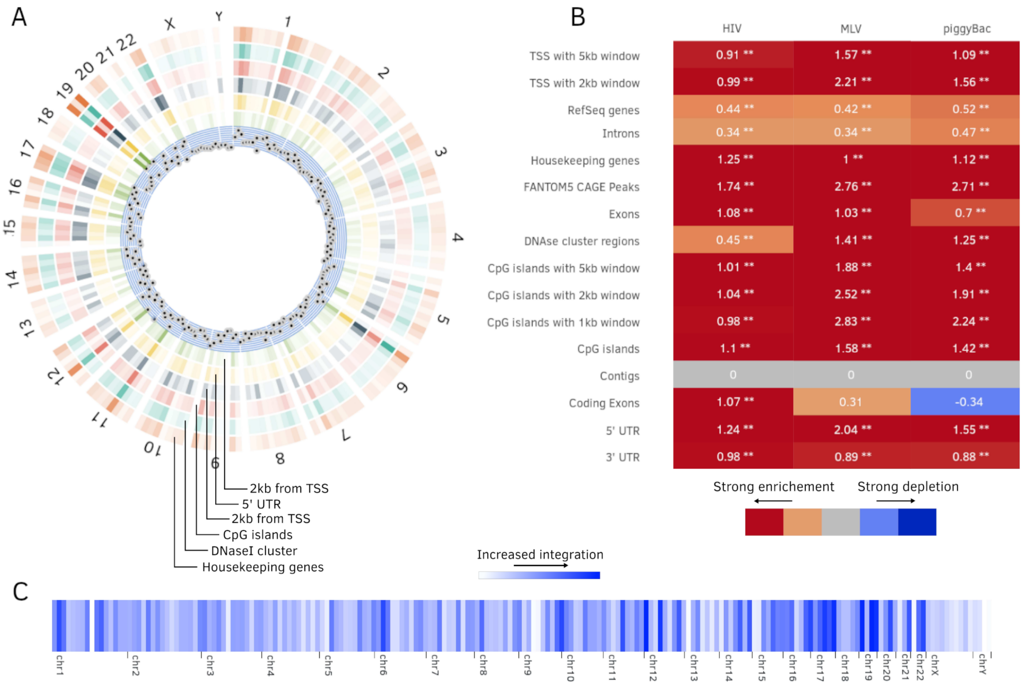


Figure 4 Additional plots generated by Enhort. (A) Circos plot (*Krzywinski et al., 2009*) of position dependent enrichment over all chromosomes for MLV for the most significant tracks. (B) Heatmap for a set of three integration data sets against various annotations. The values are $\log_2$-fold changes of the numbers of integration vs control sites falling into a given annotation. Star symbols mark statistically significant changes. The same background sites are used for the comparisons. The background sites are adapted to integrate only inside the sequence contigs. (C) Integration hotspots across the genome for MLV. The color intensity of the thin bars show the integration ratio inside of the respective genomic region.

Full-size DOI: 10.7717/peerjcs.198/fig-4

PB integration sites and the random background. The barplots were created using the R-export feature of Enhort.

The key feature of the PB integration preference is the TTAA motif in which all integrations occur. To precisely analyze the preferences of PB integration the background model has to be adapted to replicate the TTAA motif preference. This can be achieved using Enhort by creating a set of pseudo-random control sites that are located only inside a TTAA sequence. To achieve this, we simply selected the sequence logo as a covariates. Enhort takes genomic positions from a pre-sampled set of positions where each position has a probability based on the similarity between the surrounding sequence and the TTAA sequence. The results are shown in Fig. 5C where the background sites and PB show a similar motif after the motif is added as a covariate using Enhort. The motif adaption also changes the observed integration characteristics seen in Fig. 5A. The relative decreased integration of PB into coding exons is changed to a significant preference, because CpG islands are less likely to be hit by a site from the adapted background model, as TTAA occurs relatively less frequent in CpG islands. The same applies to DNAse cluster regions, TSS and exons, where the significance of integration is enriched in comparison to a random background. Only a small change for the enrichment in introns and genes is visible. Overall

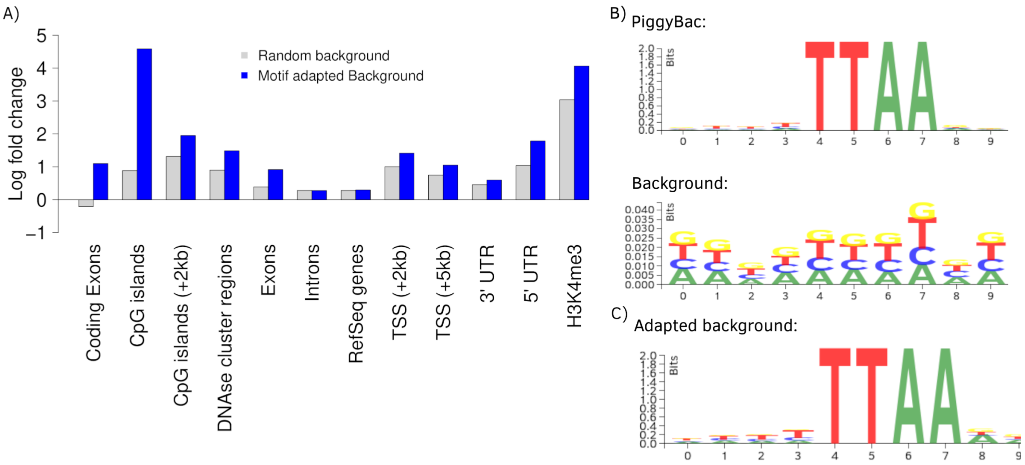


Figure 5 **Analysis of PB integration sites.** (A) Log fold changes of PB integration sites in relation to several annotations against a random and an adapted background model. Changing the background model to adapt the TTAA motif changes the observation of several integration preferences. (B) The PB motif and random sites motif, corresponding with the random background bars in (A). (C) Motif of the random sites after adaption to the PB motif using Enhort.

Full-size DOI: 10.7717/peerjcs.198/fig-5

Table 2 **Log fold changes and integration ratios of *Wilson, Coates & George (2007)* in comparison to Enhort for two PB integration site sets.**

Annotation track	Enhort Fold change	Wilson et al. Fold change	Enhort PB (%)	Wilson et al. PB (%)	Enhort Random (%)	Wilson et al. Random (%)
RefSeq genes	1.32	1.46	63.08	48.8	47.93	33.2
TSS (±5 kb)	2.14	3.00	20.8	16.2	9.7	5.4
CpG islands (±1 kb)	5.52	2.00	12.99	3.8	2.35	1.9
CpG islands (±5 kb)	2.82	0.96	22.85	7.7	8.09	8.3
Repeats:						
LINE	0.71	0.76	7.72	12.7	10.90	16.7
SINE	0.50	0.54	3.8	6.0	7.64	11.1
LTR	0.56	1.84	2.79	6.8	5.0	3.7
DNA	1.61	1.18	1.87	4.0	1.61	3.4

this indicates that beside the TTAA preference of PB there are additional mechanisms that alter the integration preferences. Using the background adaption feature of Enhort it would be possible to test different hypothesis against the data and build a model that explains the integration preferences.

To further review the analytic capabilities of our software, the integration counts of PB sites are compared to published results from *Wilson, Coates & George (2007)*. The comparison can be seen in Table 2. An increased integration of PB into RefSeq genes, inside the 5kb-TSS window, as well as a preference for CpG islands is observable for both analyses.

Wu et al. (2003) published a study on MLV and HIV stating that MLV favors TSS regions, whereas HIV does not display a strong preference towards TSS regions. The

Table 3 **Comparison between fold changes of *Wu et al. (2003)* and Enhort over different annotations on the same integration sites.**

	Wu et al.		Enhort					
	HIV	**MLV**	**HIV**	**MLV**	**HIV[a]**	**MLV[a]**	**HIV[b]**	**MLV[b]**
RefSeq genes	2.58*	1.5*	1.7*	1.4*	1	1	1	1
Housekeeping genes	–	–	3.7*	1.36	2.22*	1.12	2.05*	1.04
CpG islands (±1 kb)	1	8*	0.41	6.24*	0.35	6.17*	0.31	4.09*
TSS (±5 kb)	2.5*	4.7*	1.34	2.3*	1.14	2.02*	1	1
H4K20me1	–	–	1.71*	1.56*	1.34*	1.52*	1.36*	1.42*
H3K4me2	–	–	1.23	21.7*	1.48	21.29*	1.09	15.2*
H3K27ac	–	–	0.9	24.52*	1.01	22.79*	0.83	20.12*

Notes.

$^*P < 0.002$.

[a]with RefSeq genes as covariate.

[b]with RefSeq genes and TSS (± 5 kb) as covariates.

available integration sites were uploaded to Enhort and analyzed using the batch tool with a random 10,000 site background model. The results from Enhort show a similar integration pattern as stated in *Wu et al. (2003)* (Table 3). Except for CpG islands for HIV where Wu et al. found a near random integration and we found a decreased integration.

For further review, HIV and MLV integration sites were uploaded independently to Enhort, and RefSeq genes added as covariate. This background model had only a little effect on MLV as the preference for TSS and CpG islands only changed slightly, indicating that the preference for TSS is not due to a preference for RefSeq genes. For the HIV integration sites the housekeeping genes, which are a known preference of HIV (*Craigie & Bushman, 2012*), are still statistically significant against this background model.

Finally, RefSeq genes and TSS (±5 kb) were both used as covariates together, showing that the integration ratio of MLV into CpG islands with a (±1 kb) window decreases slightly. This shows that the integration into the CpG islands is probably not a side effect of the preference for TSS or genes. The combined background model with RefSeq genes and TSS does not have any influence on the HIV fold changes compared to the previous background model.

The creation of each background model and comparing the results was possible using built-in features of Enhort. We further added histone modifications to the analysis showing that H4K20me1 is significantly enriched for both integration sets and does not change significantly for the different background models. This indicates that the histone modification preferences is an additional effect, only slightly influenced by the preference for genes and TSS. H3K4me2 and H3K27ac are known preferences of MLV (*De Ravin et al., 2014*) and show a high fold change for all background models. With the available database it would be easy to add numerous additional annotations for comparison.

We have shown that Enhort is capable of reproducing integration site analysis with less effort and additionally offers easy-to-use mechanisms to create more sophisticated analysis using adaptable background models. The exact annotation files were not available for comparison, so it was not possible to produce the exact numbers. However, Enhort uses the same calculation principle. With the same annotations and sites the results by Enhort would be the same as in the referenced publications.

CONCLUSION

In this publication we present Enhort, a fast and easy-to-use analyzing platform for genomic positions. Based on a comprehensive library of genomic annotations, Enhort provides a wide range of methods to analyze large sets of sites. In contrast to multi-purpose software such as bioconductor, Enhort enables scientists to analyze data without programming effort or extensive manual work.

Our literature review shows that Enhort is able to perform most of the analyses commonly used in the investigation of integration sites. The re-analysis of *Wilson, Coates & George (2007)* and *Wu et al. (2003)* demonstrates that Enhort is able to reproduce analyses from literature with little effort. It was not possible to reproduce the exact values, because the version of the annotation was not recorded in the publications. However, more detailed insights can be made using adaptable background models. This was shown in the comparison of HIV and MLV from Wu et al. against different control sites.

Most publications use very simple background models for statistical analysis of integration data and could potentially be improved using better background models. Enhort provides methods to easily create more sophisticated background models for improving both the accuracy and the range of possible analyses. Complex background models can be used to identify weak effects and segregate driving factors for integration, find a minimal set of annotations to mimic integration characteristics, as well as to eliminate technical biases. In conclusion, this shows that Enhort will be a valuable tool for further analyses of genomic positions, no matter if these positions are derived from virus integration, sequence motifs, enzyme restrictions, histone modifications, or protein binding.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Hessen State Ministry for Higher Education, Research and the Arts. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Hessen State Ministry for Higher Education, Research and the Arts.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Michael Menzel conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, performed the computation work, authored or reviewed drafts of the paper, approved the final draft.

- Peter Koch contributed reagents/materials/analysis tools, performed the computation work.
- Stefan Glasenhardt contributed reagents/materials/analysis tools, performed the computation work.
- Andreas Gogol-Döring prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The source code and build instructions are available at https://git.thm.de/mmnz21/Enhort.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.198#supplemental-information.

REFERENCES

Adams D, Altucci L, Antonarakis S, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, Dahl F, Dermitzakis E, Enver T, Esteller M, Estivill X, Ferguson-Smith A, Fitzgibbon J, Flicek P, Schacht C, Willcocks S. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology* **30(3)**:224–226 DOI 10.1038/nbt.2153.

Albrecht F, List M, Bock C, Lengauer T. 2016. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research* **44(W1)**:W581–W586 DOI 10.1093/nar/gkw211.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2011. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28(4)**:464–469 DOI 10.1093/bioinformatics/btr703.

Cook Lucy B, Melamed A, Niederer H, Valganon M, Laydon D, Foroni L, Taylor GP, Matsuoka M, Bangham CRM. 2014. The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. *Blood* **123(25)**:3925–3931 DOI 10.1182/blood-2014-02-553602.

Craigie R, Bushman FD. 2012. Hiv dna integration. *Cold Spring Harbor Perspectives in Medicine* **2(7)**:Article 006890 DOI 10.1101/cshperspect.a006890.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27(24)**:3423–3424 DOI 10.1093/bioinformatics/btr539.

De Ravin SS, Su L, Theobald N, Choi U, Macpherson JL, Poidinger M, Symonds G, Pond SM, Ferris AL, Hughes SH, HL M, X W. 2014. Enhancers are major targets for murine leukemia virus vector integration. *Journal of Virology* **88(8)**:4504–4513 DOI 10.1128/JVI.00011-14.

Deyle DR, Russell DW. 2009. Adeno-associated virus vector integration. *Current Opinion in Molecular Therapeutics* **11**(**4**):442–447.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**(**5696**):636–640 DOI 10.1126/science.1105136.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research* **39**(**suppl 1**)D876–D882 DOI 10.1093/nar/gkq963.

Gogol-Döring A, Ammar I, Gupta S, Bunse M, Miskey C, Chen Wei, Uckert W, Schulz TF, Izsvák Z, Ivics Z. 2016. Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the MLV retrovirus and the piggyBac transposon in primary human CD4+ T cells. *Molecular Therapy* **24**(**3**):592–606 DOI 10.1038/mt.2016.11.

Janovitz T, Oliveira T, Sadelain M, Falck-Pedersen E. 2014. Highly divergent integration profile of adeno-associated virus serotype 5 revealed by high-throughput sequencing. *Journal of virology* **88**(**5**):2481–2488 DOI 10.1128/JVI.03419-13.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle Tom H, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**(**6**):996–1006 DOI 10.1101/gr.229102.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**(**9**):1639–1645.

LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. 2014. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Research* **42**(**7**):4257–4269 DOI 10.1093/nar/gkt1399.

Li MA, Pettitt SJ, Eckert S, Ning Z, Rice S, Cadianos J, Yusa K, Conte N, Bradley A. 2013. The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Molecular and Cellular Biology* **33**(**7**):1317–1330 DOI 10.1128/MCB.00670-12.

Li L, Zhang D, Li P, Damaser M, Zhang Y. 2015. Virus integration and genome influence in approaches to stem cell based therapy for andro-urology. *Advanced Drug Delivery Reviews* **82–83**:12–21 DOI 10.1016/j.addr.2014.10.012.

Pingoud A, Jeltsch A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research* **29**(**18**):3705–3727 DOI 10.1093/nar/29.18.3705.

Riviere I, Dunbar CE, Sadelain M. 2012. Hematopoietic stem cell engineering at a crossroads. *Blood* **119**(**5**):1107–1116 DOI 10.1182/blood-2011-09-349993.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R,

Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang Wei, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518(7539)**:317–330 DOI 10.1038/nature14248.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29(1)**:24–26 DOI 10.1038/nbt.1754.

Roth SL, Malani N, Bushman FD. 2011. Gammaretroviral Integration into Nucleosomal Target DNA In Vivo. *Journal of Virology* **85(14)**:7393–7401 DOI 10.1128/JVI.00635-11.

Sandve GK, Gundersen S, Johansen M, Glad I, Gunathasan K, Holden L, Holden M, Liestl K, Nygrd S, Nygaard V, Paulsen J, Rydbeck H, Trengereid K, Clancy T, Drabls F, Ferkingstad E, Kala M, Lien T, Rye MB, Frigessi A, Hovig E. 2013. The Genomic HyperBrowser: an analysis web server for genome-scale data. *Nucleic Acids Research* **41(W1)**:W133–W141 DOI 10.1093/nar/gkt342.

Shao W, Shan J, Kearney MF, Wu X, Maldarelli F, Mellors JW, Luke B, Coffin JM, Hughes SH. 2016. Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology* **13(1)**:Article 47 DOI 10.1186/s12977-016-0277-6.

Wilson MH, Coates CJ, George AL. 2007. PiggyBac transposon-mediated gene transfer in human cells. *Molecular Therapy* **15(1)**:139–145 DOI 10.1038/sj.mt.6300028.

Wu X, Li Y, Crise B, Burgess SM. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300(5626)**:1749–1751 DOI 10.1126/science.1083413.

3 Summary

3.1 k-mer-based motif discovery

Transcription factors are a vital component of cellular regulation and even though considerable effort has been devoted to their discovery, knowledge remains far from complete. As previously described, peak calling together with motif discovery are the common analysis methods for ChIP-Seq datasets. Although widely used, the typical two-step process presents limitations as many signals are removed during peak generation by discarding reads below certain thresholds. The accumulation of reads to peaks reduces noise, but also removes subtle binding information. To provide an alternative method for the discovery of TF target sequences, NoPeak was developed, a software tool that identifies motifs based on the enrichment of k-mers near each read. A k-mer is a short genetic sequence of *k* bases. The usage of k-mer-based approaches in motif discovery has been successfully established by different applications [44, 144, 32, 39, 116].

The key distinction between our approach and past publications is the absence of peak calling. We offer an approach that is independent from peaks and thus is able to capture more information by not filtering reads through consolidating them into peaks. Certainly, besides the signal amounts of noise also increase and poses new challenges for evaluation techniques. By using all reads available, our method is able to find motifs in datasets with low coverage where peak calling is unable to identify peaks. The algorithm and software NoPeak itself was published in 2020 [85]. My work encompassed the implementation of a new software package to generate k-mer profiles, new algorithms to generate sequence logos from k-mer profiles and k-mer lists, and the analysis of profile shapes. Moreover, the design of two challenges where NoPeak was tested on depleted datasets and comparison of NoPeak against common peak calling tools on peak identification, as well as the implementation of additional tools to work with intermediate outputs. In the following the method of NoPeak will be summarized, as well as findings and implications that can be drawn from the novel approach.

3.1.1 Transcription factor models based on k-mers

Different established models to identify motifs from ChIP-Seq data exist. They apply read mapping and accumulate the mapped reads into peaks. After peak calling, the identified peak regions are scanned for enriched motifs, often compared to a null model based on random generated or random selected sequences from the genome not involved in regulation. Thus, motifs are searched position-dependent and combined into PWMs at a later stage.

Our novel method directly combines observations for k-mers by searching the vicinity of each read for observed sequences of a specified length (NoPeak Bioinformatics Fig. 2). The distance from each observed sequence to the read position is registered for that sequence. Aggregated distances build a profile for each sequence and show their distribution in relation to reads. Subsequently, those profiles are evaluated to identify the relevance of each sequence and discard noise. Examples for discarded profiles are shown in the Appendix in NoPeak Supplement Section 1.3. Based on control reads a background model is generated with the same procedure. Control profiles are substituted at each profile position from the signal profiles to remove background effects from each k-mer, respectively. In contrast to PWMs, where interdependencies are not resolved and variable gapped motifs are difficult to model, our approach directly identifies sequences from the genome and only actually present k-mers are recorded.

The method is able to assign scores to each k-mer that represents the observed frequency of occurrence corresponding with binding activity. k-mer scores are based on relative profile height and multiple filters assure that k-mers based on repetitive regions and other low-mappability regions are omitted. Profiles are also capturing expedient information on binding site structure, which will be reviewed hereafter.

Within the publication, it could be shown that the algorithm is capable of producing meaningful results for TFBS motifs based on k-mer profiles with position-independent analysis (NoPeak Bioinformatics Section 3.1, Supplements Section 1.6). To make comparisons to traditional analysis methods and provide visual results, NoPeak is able to generate PWMs and sequence logos. Binding motif from NoPeak were compared to motifs from databases based on well-established methods and identified a close resemblance between high-ranking motifs (NoPeak Bioinformatics Section 3.2).

The comparison shows that results from NoPeak closely resemble the main motif by significant bases that are visible at the expected locations. NoPeak Bioinformatics

Fig. 5 contrasts the results for several TFs with clearly visible similarities between the main motifs. Besides the main motif, less significant bases closer to the borders and inside the main motif can differ between the methods, which is expected, given that less signal is available to support base occurrence. Further comparisons between NoPeak and common databases are available in Supplements Section 1.6, demonstrating the similarity of a larger set of TFs. Dataset-dependent divergence between results is observable for different TFs, for example visible for several GATA3 and BRCA1 data sets, from which several are marked as low or insufficient read length in Encode. Additionally, NoPeak is able to reveal possible secondary motifs based on k-mer clustering. However, it remains unclear which of those weak motifs are actual weak binding sites or whether their appearance is coincidental.

Results from PBM [91, 14] are also *k-mer-to-score* lists, indicating possible synergies in software usage. Results from NoPeak and PBM for different TFs found in UniPROBE [97] were compared exemplary to highlight similarities (Supplements Section 1.1). Nonetheless, further research is needed to evaluate the possibilities and limitations of their compatibility.

New technologies to gather information on regulation structures have been developed in recent years, including ChIP-Seq variants such as ChIP-exo [108] and CUT&RUN [121]. Fundamental evaluations using NoPeak showed that it is also possible to evaluate raw data from both technologies using a genome-wide k-mer profiles approach (Supplements Section 1.13 and 1.14).

Overall, the findings demonstrate that the direct inference of binding motifs from read distributions yields good results for TF binding models. The results demonstrate that our implementation NoPeak is appropriate to be used as *de novo* motif discovery tool. Further applications are described in the following sections.

3.1.2 k-mer profile shapes reveal transcription factor characteristics

As previously described, k-mer profiles represent the distribution of k-mers in relation to reads. The distribution of reads is influenced by technical variation, such as fragment length and scatter. Additionally, the influence of biological factors including binding site characteristics such as chromatin structure or dimer binding can be assumed.

Within the publication, it could be demonstrated that k-mer profiles provide a novel insight into TFBS characteristics and showed that profile shapes relate to known binding site structures. Characteristic profile shapes are observable within each TF and remain visible with the decreasing strength of k-mers. While comparison between TFs shows divergent profile shapes. For example, the dimer binding of ATF3 shows a double peak in the profile, while FOXA1 shows a secondary peak, possibly related to a known shadow peak induced by estrogen-mediated chromatin loops (NoPeak Bioinformatics Fig. 8) [54, 36]. This indicates that k-mer profiles can be used to predict biological structures as k-mer profile shapes indicate analog binding modalities. In a future evaluation, the systematic search for profile structures in ChIP-Seq experiments could reveal unknown binding characteristics.

Figure 3.1 shows the top scoring k-mer profiles for two further evaluated TFs. BHLHE40 shows a twin-peak shape, most prominent in the top profile of the right-most data set (ENCFF216ZWY), but also visible in less prominent profiles.

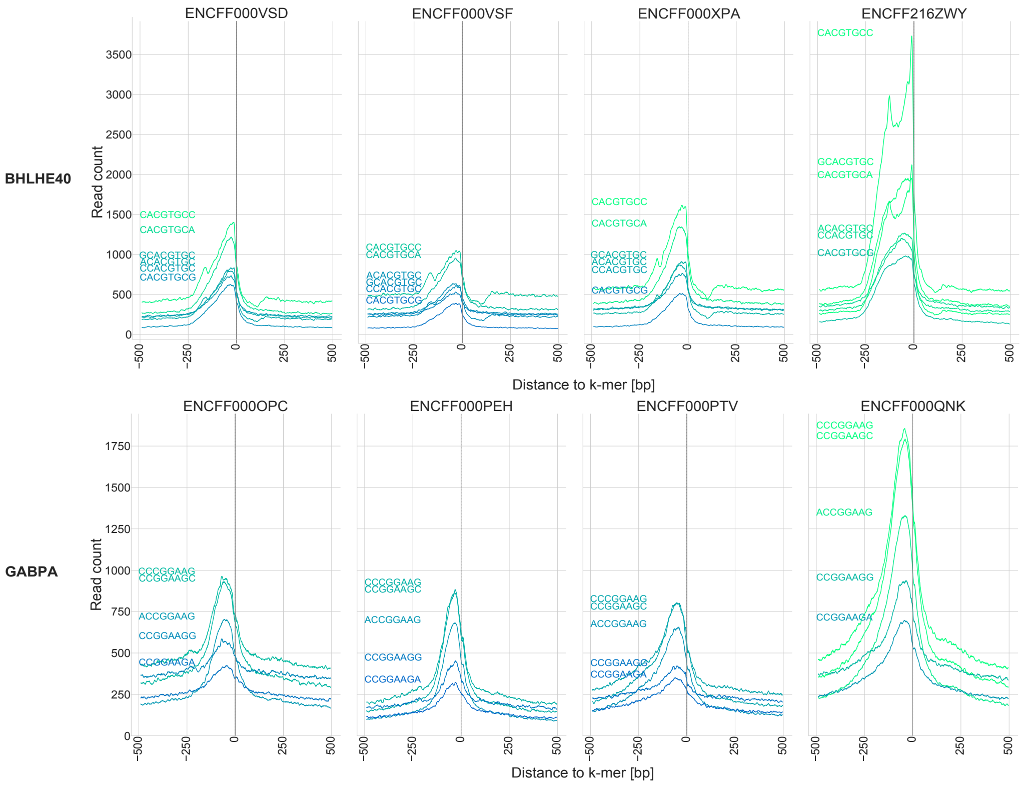


Figure 3.1: Different profile shapes for the top scoring k-mers of two TFs (BHLHE40 and GABPA) in different datasets. Colors correspond to the read count. Encode IDs of each experiment are noted above. Adapted from NoPeak Supplements Section 1.7.

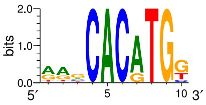
Moreover, a drop below the background after the peak is observable for most profiles. By contrast, the shape of the GABPA profiles shows a smooth single peak for each k-mer and no drop below the background (additional profiles from more TFs are in Supplements Section 1.7).

3.1.3 Motif discovery in low-quality datasets

Low-quality datasets from ChIP-Seq experiments are presumably more common than apparent when reviewing databases, as they are commonly discarded during analysis and not published. Besides single datasets, complete experiments often remain unpublished if no significant results are found. Within the publication, it is shown that the NoPeak approach is able to capture binding motifs in low-quality ChIP-seq data even if other methods are unable to resolve peaks. This improvement was expected because NoPeak uses more signal from the given data by relating each read to surrounding k-mers, in contrast to peak calling, where only peak positions are used to identify the binding motif.

To prove our claims, two challenges were designed in which reads were iteratively removed from ChIP-Seq datasets. The first challenge removed reads from peak regions while the second removed reads at random. Each degraded dataset was evaluated using peak calling with *de novo* motif discovery and our approach. The results from each round for both technologies were then compared to the database motif. Evaluations show that NoPeak is able to better perform identifications of relevant k-mers for low-quality datasets. While the motif produced by NoPeak remains almost constant, the motif produced by Homer further deteriorates with each step. For each experiment, a reduction step after which peak calling does not yield any more peaks is reached and no further motif can be generated by common motif discovery, while NoPeak is still able to return relevant k-mers and the expected motif.

Figure 3.2 exemplary shows the effect of read reduction from peak regions and resulting sequence logos from both tools. Within the fourth reduction step the sequence logo deteriorates from the inital result for Homer, while the motif produced by NoPeak is constant. The peak count significantly decreases, showing the dependence between peak calling and subsequent motif discovery. By contrast, the read count is only slightly reduced after the first round. Results from more experiments corresponding to the NoPeak Bioinformatics Fig. 7 are in NoPeak Supplement Section 1.5 in the Appendix.

Database motif: 

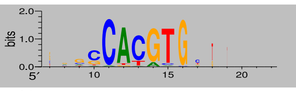
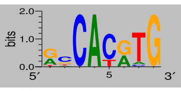
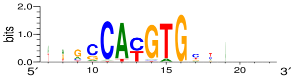
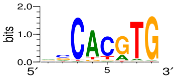
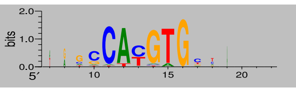
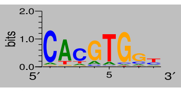
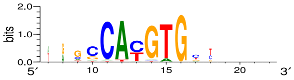
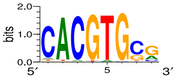
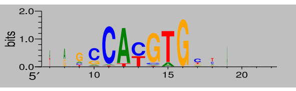
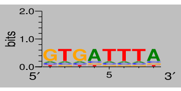
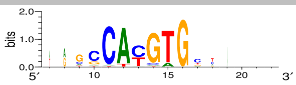
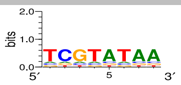
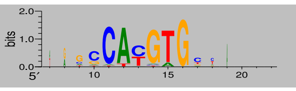
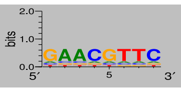
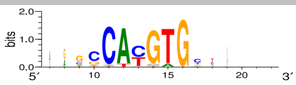
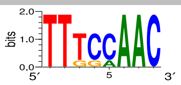
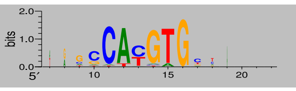
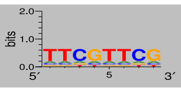
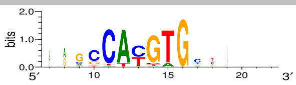
Round	Peak count	Read count	Logo NoPeak	Logo HOMER
0	54250	29729565		
1	6467	24909911		
2	728	24696286		
3	210	24673770		
4	106	24667343		
5	13	24664359		
6	4	24663836		
7	5	24663678		
8	2	24663499		
9	0	24663459		

Figure 3.2: Comparison between sequence logos generated for each round of read reduction of the TF MAX. The expected motif from JASPAR is shown above. Redrawn after NoPeak Supplement Section 1.5.

3.1.4 Peak identification using k-mers

Our approach operates without the use of peak calling. However, beside TF binding motifs the peak positions and therefore binding positions are crucial in understanding regulatory structures. Therefore, we evaluated our approach based on the capabilities of identifying peak regions based on k-mer results. We replicated the DREAM5 Motif Recognition Challenge [29] to evaluate whether our k-mer based approach is able to differentiate peak from random regions (NoPeak Bioinformatics Section 3.3, Supplement Section 1.11).

The challenge is based on 2,000 regions of 100 bp sequences, with 500 of them being sampled from peak regions, while the others are non-peak regions built by different methods. Each tool evaluates given regions based on a model that was trained a priori and assigns a confidence value that states how likely the tool estimated the region to be in the vicinity of an actual peak.

To differentiate peak and random regions, NoPeak used resulting k-mers together with their associated score as a model. Each sequence was evaluated based on known k-mers that were found, together with the certainty that this k-mer is relevant for binding affinity. The relevance of each k-mer was expressed through its score. Statistical evaluation of the prediction results reveals that our approach is able to produce equal results to Homer, although it falls behind ChIPMunk and MEME-ChIP. We suspect that the unavoidable mix of test and training data for the peak-calling-based approaches improves peak prediction for the other tools, while NoPeak does not have this benefit. Further, the results show that datasets tend to yield similar results for different software tools, as is visible in the vertical similarity for many TFs. Besides those similar results, each software tool shows outliers with low predictability for several TFs.

3.2 Genomic position profiling

The safety of gene therapy applications relies on the capabilities of delivery vectors. In early gene therapy research, viral vectors were commonly used. Today, they are complemented by transposable elements and technologies like CRISPR/Cas9. Irrespective of the vector origin, insertion characteristics including genetic feature preferences, targeted sequences or integration loci need to be determined. To build a platform for the analysis of genomic positions in relation to annotations and genome sequence, we developed Enhort. The platform was published in 2019 [86] based on previous work on background models in my master's thesis [84]. The platform combines easy-to-use methods for statistical analysis of genomic sites by providing a comprehensive database of annotations. At present, 1,402 annotation tracks from 97 cell lines are available.

Within the scope of this dissertation background model creation for multiple tracks, as well as sequence motifs was developed. The database of annotation tracks that contain publically-available annotations for commonly-used cell-lines was built. The hotspot analysis was implemented, as well as batch analysis of multiple genomic site tracks, additional tools for detailed graphic representation, export of results, sequence logo analysis, sequence logo-based background models and a guided analysis tool to simplify user access. Hereafter, the main aspects of Enhort are recapitulated followed by summaries of publications where the platform was used and the contributions made using Enhort.

3.2.1 Easy-to-use genomic position analysis

The platform is available online as a website (https://enhort.mni.thm.de). Users can upload genomic positions, such as viral integration sites, TEs sites or TFBS in single or multiple files. Enhort determines the genome version and intersects given sites with available annotations. To determine enrichment and statistical significance, a random background model is created based on positions inside the *contig* regions. Counts from given sites and background model are evaluated by statistical significance and fold change in relation to annotations. The results are presented to the user in various diagrams. Through custom development and optimization of algorithms, the platform is able to return results in real time, which we were able to show for various sizes of input data (Enhort PeerJ Table 1).

After reviewing the results, the user is able to select annotations, such as cell-line specific features, expression data or custom annotation tracks as covariates to create a background model based on scientific questions or experimental limitations. Custom generated background sites are expedient to thoroughly dissect genomic site selection, and they offer the possibility to study effects of faint integration effects, which are otherwise covered by large-scale influence. Enhort is able to instantly create background models based on user selected covariates and thus it provides an easy access to advanced analysis that previously needed complicated scripting and understanding of the underlying models.

We were able to reproduce results from well-known viral integration sites such as AAV, HIV and MLV that are backed up by various publications [142, 22, 26]. We systematically reviewed literature for publications using genomic position analysis to support our claims of application possibilities. Out of 59 identified publications, 37 could be performed using methods from Enhort and six further publications apply methods that are partly reproducable utilizing Enhort. As previously stated, only a minority of six publications make use of adapted control sites based on experimental parameters or known preferences to identify integration preferences. The analysis steps described in most publications could have been performed with our platform without any additional programming, including the analysis of integration near genetic features, histone modifications and sequence motifs.

3.2.2 Multi-factor background models

Enhort likewise allows setting multiple covariates to control background sites for various annotations in each experiment. For example, a virus is known to integrate near active TSS and specific histone modifications and sites have to be near restriction enyzme cutting sites. Background sites mimic integration of given virus sites for each annotation, as well as each combination of annotations. For the mentioned example, that would include TSS without histone modification, histone modification without TSS, both modifications, and corrected for intervals near restriction enzyme cuts. Computational complexity rises exponentially with an increasing number of selected covariates, and thus suitable algorithms are implemented to uphold a fast response time.

Although, background models based on multiple covariates are achievable using custom scripts, such effort is rather uncommon for positional analysis, as seen in various publications that we examined during the literature review. Numerous significant integration effects are observable for viruses and transposable elements besides those mentioned in the literature. Faint integration preferences are visible and can be found by explorative analysis of annotation data using Enhort and the built-in mechanism of adaptable background models.

3.2.3 RNA-guided retargeting of Sleeping Beauty transposition in human cells

The subsequent section is based on the publication: RNA-guided retargeting of Sleeping Beauty transposition in human cells, published in eLife by Adrian Kovač, Csaba Miskey, Michael Menzel, Esther Grueso, Andreas Gogol-Döring and Zoltán Ivics [63].

As previously described, gene therapy and genetic engineering holds strong potential for developing new treatments and applications in biotechnology. Many challenges are present that currently prevent universal application, including expression levels control to prevent overexpression, as well as low expression levels and insertional mutagenesis. Besides retroviruses as delivery vectors, the usage of transposons shows promising results as immune reactions are lower and they perform better in handling [48]. One of the transposons used is Sleeping Beauty (SB), a Class II DNA transposon. SB has a random genome-wide integration at the specific target sequence *TA*. This pseudo-random integration is also the main issue present, as random integration can disrupt genetic functionality by integrating inside a gene or promotor region and although this is rare for SB integration it nonetheless occurs [41]. An analysis of 59,169 SB sites revealed that 2.43 % of SB integration sites are located in 5kbp TSS regions, which is near random (2.64 %) [7].

Apart from integration vectors, the mechanisms of targeted nucleases like ZFNs and TALENs and the CRISPR/Cas system can be utilized for genetic modification, as shown by Aird et al. (2018). Each enzyme functions by a DNA-binding domain (DBD) with a target sequence and a nuclease domain for DNA cleavage. The Cas nuclease stands out with its need for a single guide RNA (sgRNA) that determines the target sequence. This mechanism allows for specific targeting inside the genome. However, the insertional mechanism of nucleases is of poor efficiency because the more frequent non-homologous end joining (NHEJ) repair-mechanism leads to insertions or deletions [4].

Each technology shows a drawback: SB integrates randomly at a common sequence motif and the integration mechanism of the Cas system has low integration efficiency. Nevertheless, these issues are complementary, each showing a benefit for the other's weakness regarding insertional features. To develop an integration system that is able to be targeted like the Cas system yet showing the delivering capabilities of SB, both were successfully combined and the findings were published [63].

Three targeting factors were built to investigate the feasibility of this approach. SB100X, a hyperactive SB transposase was fused with dCas9 at the C-terminus and N57, an N-terminal fragment of SB was fused to dCas9 at both the C- and N-terminus, each with a flexible linker. Two integration targets were selected, the HPRT gene on the X chromosome and AluY, a highly conserved ALU retrotransposon for which sgRNA was provided. The fusions were cloned into all-in-one expression plasmids allowing expression of the targeting factors and sgRNAs and subsequently transposition was tested in human HeLa cells. Integration datasets were then generated for several fusions and targets.

The identification of target preferences was conducted by using Enhort, as positional enrichment is the core functionality of our tool. A custom database was built using Bowtie1 [69] based on the target sequences of AluY and HPRT. The resulting annotation tracks were added as custom tracks to the Enhort database. Both sets of targeted positions (dCas9-SB100X and dCas9-N57 + SB100X) were uploaded using the batch feature and untargeted background sites set as a reference model. With Enhort, we were able to detect the weak preference for the targeted region as described in the publication and reveal that SB fused with dCas9 is a potential RNA-guided vector for gene therapy.

3.2.4 Reprogramming triggers mobilisation of endogenous retrotransposons in human-induced pluripotent stem cells with genotoxic effects on host gene expression

The following segment is based on a contribution at the European Society of Gene and Cell Therapy (ESGCT) Congress by Sabine Klawitter et al. [56], supported by my computational analysis.

Modification of human-induced pluripotent stem cells (hiPSCs) holds interest as they are capable of unlimited proliferation and generation of *in vivo* derivatives [62]. However, reprogramming-induced activation of endogenous mobile retrotransposons LINE-1, Alu, and SINE-VNTR-Alu (SVA) is observable [62]. Therefore, a study was designed to identify possible activated transposons and other interferences of L1 insertions with gene expression. Here, Enhort was used to determine the integration preferences of L1 for different genomic features, selected genes and sequences.

3.2.5 Engineered transposases and transposons enforce integration into highly active genomic loci and facilitate optimal transgene expression

The following section is based on the manuscript in preperation entitled Engineered transposases and transposons enforce integration into highly active genomic loci and facilitate optimal transgene expression, written by Sven Krügener, Thomas Rose, Michael Menzel, Anneliese Krüger, Fränzi Creutzburg, Annette Knabe, Andreas Gogol-Döring and Volker Sandig [64].

Besides the application of *in vivo* gene therapy to cure illnesses, genetic modification of pharmaceutical cell lines holds strong interest for industrial purposes. Usually, those cell lines are grown in suspension cultures and utilized for industrial production of biopharmaceuticals such as monoclonal antibodies [73]. A suitable vector for genetic engineering of such cell lines is the piggyBac (PB) transposase, a naturally active transposase targeting *TTAA* sequences. PB has proven useful in different biotechnological applications with a capacity for larger insertions compared to other vectors, while insertions show a preference for gene regions, resulting in improved transgene expression [140]. Transposase-based delivery systems for Chinese hamster ovary (CHO) cell lines have recently been developed [81, 106], increasing the expression compared to conventional transfection, although integration is still random.

Similar to SB, insertions are randomly distributed over the genome at *TTAA* sequences [37]. As active transcription regions are rare and integrations occur nearby, efficient cell cultures are infrequent and need to be selected by a screening process. To increase favorable integration, hyperactive variants of PB have recently been developed [143, 75], showing increased insertional activity. An overall increase in insertion raises the chance of insertions into active transcription regions. Selecting highly transcriptional regions for insertion is an important challenge to increase effectiveness in genetic engineering.

To develop an insertion system capable of inserting a given transgene at transcriptional highly active regions, a novel delivery system based on PB was designed and described in the manuscript [64]. For targeting the characteristics of histone modifications are utilized. Histone modifications play a vital role in cell regulation, including the triple methylation of lysine 4 of histone 3 (H3K4me3), which is frequent in highly active promotor regions. H3K4me3 is recognized by the PHD domain of the TF TAF3. The TAF3 PHD domain could successfully be fused to a PB wildtype and an hyperpactive PB variant (haPB), both showing significant enrichment in H3K4me3 peak regions. To verify transcriptional activity, the monoclonal IgG4 antibody Nivolumab was used. Experimental results show a higher viability with increased IgG titers for both technologies compared to wildtype and non-fused haPB. haPB fused to TAF3 exhibits especially high titer concentrations.

Enrichment analysis was performed using our position analysis platform Enhort. Custom annotations were generated using H3K4me3 peaks from the CHO-K1 cell line for different window sizes around peaks. Additionally, annotation tracks for the CHO cell line were created based on genes, exons, CpG and transcriptional regions. Annotations were loaded as custom tracks into Enhort and their insertion evaluation as well as statistical evaluation was performed [64]. The results reveal the suitability of epigenetic targeted PB for genetic engineering by improved transgene expression and the applicability of Enhort to custom data analysis.

4 Discussion

While being a key-technology for future medicine, the complex topics of gene therapy and genetic engineering present a multitude of challenges. In this dissertation two aspects were examined to improve the understanding of delivery vectors in relation to genetic features and TF binding models. Hereafter, an integration of findings to current research, limitations and outlook on future challenges are provided.

4.1 k-mer-based motif discovery

Based on known limitations of common ChIP-Seq analysis tools, NoPeak was developed as an alternative to peak calling approaches. With the evaluation of numerous datasets and by comparison with known motifs, it could be shown that NoPeak is able to reproduce expected sequence motifs. Partial differences between results and databases are observable that can be accounted to the usage of single datasets for the comparison with JASPAR [111], which uses manual curation. Datasets that show deviant sequence motifs are often marked as having insufficient coverage and read depth. Further, differences in low-certainty bases are expectable. As described in the literature, sequence motifs overvalue the significance of consensus sequences [29] and manual curation poses a non-negligible effect on the results. Bias in sequence representation due to selected prior knowledge and data handling is inevitably built-in in all methods. We speculate that NoPeak is less likely to discard contrary findings, as the results are based on a table of k-mers associated with their individual score, while other methods usually combine observations directly into single sequence motifs. k-mer lists are also inherently able to represent secondary motifs, variable gapped motifs and positional dependencies.

The main difference between NoPeak and peak calling is the extent to which reads are used. Through peak calling, large amounts of the signals are reduced. Our approach utilizes each mapped read, which enables us to better detect binding motifs in data with low signal content. However, it also introduces the need for enhanced noise management as noise peak accumulation reduces experimental and biological noise. To differentiate between signal and noise, the shape of k-mer profiles is used, introducing new parameters for noise removal. Each parameter leads to a more complex usage and increases the potential for inaccurate settings or an unintended introduction of bias. Experimental processes are known to be influenced by observer biases, most prominently when certain results are expected [46]. Each software used is dependent on numerous parameters, either alterable by users or hidden. For computational discoveries and especially explorative data analysis, the impact of parameter selection has to be closely examined to prevent the introduction of incentives. To reduce the influence of fixed parameters, an interactive version of the motif building step was implemented, allowing researchers to test parameter combinations and review different outcomes.

Our approach shows restrain for long motifs that are difficult to identify due to computational limitations for large values of *k*. By combining short overlapping k-mers to longer motifs by methods like *seed-and-wobble* [14], size constraints can be shifted. With increasing motif overhang, uncertainty and instability increases, although large motifs are also problematic in standard motif discovery tools [47].

The results confirm that NoPeak is able to reproduce known TF preferences and that k-mer representations serve as an beneficial model of TF affinity, while NoPeak is able to perform *de novo* motif discovery in ChIP-Seq and related data. NoPeak is position-independent, whereby direct identification of binding locations is not possible. Therefore, NoPeak is not able to replace peak calling as a default method of ChIP-Seq analysis, as binding positions are also an important measurement. Nevertheless, publications have shown improvements in peak detection using prior knowledge on TFs [119]. This suggests that using prior knowledge obtained through NoPeak can support peak analysis. A further benefit of using a different method for the provision of a priori knowledge is the prevention of circular reasoning, which could occour if the same peak evaluation method is used twice. Further information on improving peak discovery with NoPeak is presented in the following Section 4.1.4.

Based on the algorithmic complexity, the groundwork of k-mer to read relation is presumably more robust than peak-to-motif relations. NoPeak explores the direct relation of read signals to surrounding sequences in contrast to the two-step process of peak calling and separate motif discovery. In future development, representation of k-mers as a sequence logo could be improved by including auxiliary information from base dependencies and frequencies with presentations similar to dependency logos by Keilwagen and Grau [59] or Guo et al. [39]. Similar to our observations, several ChIP-Seq control datasets show extensive read clustering [79] similar to signal values. A filter was implemented to notify the user if control data too closely resembles the signal. The relation between control and signal data should be further evaluated for possible errors in experimental design.

4.1.1 Better background control for motif discovery

De novo motif discovery tools utilize random genetic sequences without TFBS as a background model for motif identification. Motifs identified inside the peak region that do not occur in the background are labeled as enriched. It is known that these background models are often insufficient for motif discovery as their sequences are too random compared to peak regions, which leads to overvaluing of motif sequences [119].

By using k-mer profiles from actual ChIP-Seq controls, NoPeak is able to apply real background data for each k-mer in motif discovery. Background profiles are applied to each k-mer profile using all genomics regions, in contrast to motif discovery, where genomic regions are artificially selected to represent the background signal. Background k-mer profiles are generated with the same procedure as signal values and directly substracted from the signal profiles. Additionally, background profiles are viewable and can help to identify control datasets that show unexpected read clustering, as previously described. Depending on analysis needs, the influence of background substraction can easily be adjusted.

4.1.2 Profile shape evaluation

Intermediate data from NoPeak revealed that binding characteristics were embedded in the distribution of reads, and these characteristics are visible in the k-mer profile shape. By simulating ChIP-Seq reads, it could be shown how motif shape is influenced by fragment length and variation (Supplement Figure 1.4). Besides technical influences, a strong biological influences on profile shape is probable. By comparing profiles from different TFs and experiments for each TF, characteristic profile shapes could be identified (NoPeak Bioinformatics Figure 8). Based on known binding characteristics of the TFs, the observed profile shapes could be correlated with biological features. For instance, the dimer-binding of ATF3 or the chromatin *shadow loop* by FOXA1 which are presumably related to the profile displayed by the significant k-mers. Closely related profile shapes appear for all top k-mers found in different datasets for the same TF (Supplement Section 1.7). Further analysis of profile shapes could yield insights into binding characteristics such as chromatin structure, sequence features and binding properties, as those are known to be influental on peak and therefore also read distribution [136]. Observations on systematic profile shape variation tie well with previous studies wherein ChIP-Seq data was evaluated using *TFBS-Landscapes* [141]. Discovered sequence motifs were used to create a comparable representation of motif distances to the peak position to derive quality and enrichment without selecting a threshold. Similar properties such as shape and density differences between TFs were discovered by this approach. Further, analyses performed by Bailey et al. [11] showed that binding properties as well as quality could be derived from distances between motif and reads.

The analysis of k-mer profile shapes is still an initial phase as technical and biological influences can only be partly dissected. Future work should focus on understanding influential factors on peak shape to improve noise removal and enable the derivation of biological properties.

4.1.3 Improved motif discovery in low-quality datasets

Previous studies have emphasized that ChIP-Seq data quality in available databases is not continuously sound. Marinov et al. found 20% of evaluated datasets to be of poor quality, while about 25% were of intermediate quality [79]. Furthermore, these findings only apply to published datasets, it can be expected that large quantities of data are not published when certain quality criteria are not met. Therefore, a method capable of identifying motifs in low-quality data could be an asset to build motif models of unknown TF or improve existing models. NoPeak combines the observations of each k-mer across the whole genome for each read. The number of data points collected for each observation is improved at two stages. First, all reads are used that would otherwise be discarded by the accumulating peaks. Second, all k-mer-to-read relationships across the genome are combined for each k-mer and reverse complement. For this reason, NoPeak is able to identify more faint effects than common motif discovery, as shown in different evaluations. Performance was evaluated for different types of quality-modified datasets and could support the claim. Moreover, the models based on k-mer profiles can easily be combined for replicates, enhancing the extent of the data. Software needed for this task was published alongside the main application.

Based on achieved expectations, it can be assumed that the re-evaluation of large quanitities of unpublished datasets and poor-quality datasets could be performed using NoPeak. Especially lesser covered TFs could benefit from results of additional datasets and analysis with a second method that supports uncertain results. A more thorough dissection of existing datasets is a fast and profitable usage of resources as no further wet lab experiments need to be conducted.

4.1.4 Assisted peak selection

As previously described, *de novo* motif discovery without prior knowledge on affinities can be improbable due to several factors [119]. Using k-mer models by NoPeak, it could be shown that differentiation between peak and random regions in ChIP-Seq data is possbile. For several cases, NoPeak could outperform common motif discovery tools. These results ties well with previous studies on evaluation of TF models where certain algorithms TF-specific outperformed other approchaes [29]. Incorporating advantages for different TF from multiple algorithms will likely outperform single approaches This suggest that models based on k-mers can be used as an asset in motif discovery pipelines. For the utilization, parameters for method combination need to be defined that combine the advantages of k-mer based results with peak calling. Further, utilizing known motifs for motif identification is already common practice. By using a k-mer based approach, circular reasoning using the same discovery algorithm in both steps can be prevented.

4.2 Genomic position analysis

Continuous efforts in the development of safe insertional vectors prompt the need to review new vectors and their integrational preferences. The development of targeted vectors, for genetic features [77], specified sequences [15] or safe harbors [100, 110, 102] is a foundation of gene therapy approaches and close inspection of their actual insertional behavior is required to prevent insertional mutagenesis [17, 15]. Based on the requirements, an online platform called Enhort was implemented. Our literature review and recent publications [63, 56, 64] that use the platform show its practicability and capabilities. The extensive number of annotations simplify analysis for users by providing data that otherwise needs to be gathered manually. Furthermore, adaptable background models inside the real-time platform are a novel innovation and extend the analytical capabilities for integration site analysis.

For a static analysis of few annotations or different species, the usage of BEDTools [105] is certainly more convenient. By allowing the download of background sites and upload of custom annotations to Enhort, both tools can utilize their advantages and be used in conjunction. However, for explorative data analysis with unknown integrational properties, various cell lines and annotation categories, the usage of a platform combining these with statistical analysis is clearly superior to manual processing and BEDTools.

With progress in development of sequence targeted vectors, as seen with CRISPR/Cas9, the focus on insertion site sequence analysis becomes more important. Therefore, Enhort is able to identify integration motifs based on given sites and display sequence motifs. Additionally, background models based on sequence motifs can be generated, alike to using annotations as covariates. Further development could include the detection of preferences of mismatch sites and estimation of the impact of off-target integrations. To enable the analysis of complex integration motifs custom-build tracks for target sequences can be uploaded individually for analysis. Viral vectors are also used as delivery vehicles in novel gene therapy mechanisms like CRISPR/Cas9 as non-viral delivery proves to be difficult to utilize[76], which shows the importance of viral mechanisms in future applications and therefore the relevance of the Enhort analysis platform.

Genomic position analysis using Enhort is limited by the availability and quality of annotations. Future work to improve Enhort encompasses the addition of annotations for various species and maintenance of the current annotation database.

4.2.1 Adaptive background models to improve site analysis

The benefits of background correction with covariates has already been shown in different publications. De Jong et al. 2014 [25] analyzed SB and PB with a background model controlled for their respective integration motif combined with the distance to the nearest restriction site and unmappable regions. This is similar to Li et al. [75], where background sites were selected to adhere to the motif that PB integration sites express. Hüser et al. [49] used adapted control sites for restriction enzymes and mappability. Wang et al. [135] and Roth et al. [109] used background sites that were controlled for the distance from restriction enzyme sites. Adapted background models to TSS regions were used to detect insertion sites by de Ridder et al. [27]. Nonetheless, the literature review showed that adapted background models were only used to improve the analysis in a minority of publications. Even though in many cases specificially adapted background models are not strictly necessary, it can be assumed that they would increase the reliability of findings. Their spare usage is potentially due to the effort needed to generate them and the complexity in calculation.

With the Enhort analysis platform, the effort and complexity is substantially reduced and researchers are enabled to easily utilize advanced background models. Complex models with different incorporated factors can be built with little effort. In addition to the usage of known influences, such as restriction sites, covariates can be selected iteratively and thus remove the most prominent effects on site distribution. This method can be used to build a background model that closely resembles the integration sites. At this step, the neccessary influence factors that explain observed integrations can be taken from the set of selected covariates. Adaptable background models can help to differentiate integration effects. However, genomic structures are known to be highly dependent and entwined. It is therefore not possible to resolve all confounding factors based only on computational analysis. With an increasing number of covariates, the possible regions for background integration are reduced. By selecting covariates too strictly, sites are forced into the same regions as the given sites and thereby removing the observational significance for any annotation. Nonetheless, the usage of specifically adapted background models enhances the quality of gathered results and the extent of results extracted from the given data. Further, the usage of the platform helps to standardize genomic site analysis and enhances the comparability of results.

4.2.2 Explorative positional data analysis

Large amounts of annotation data on genetic structure have been collected in recent years. Traditional scientific reasoning is based on a hypothesis-driven method. It starts by an unexplainable observation, for which the most-likely unproven and testable explanation is evaluated in experiments. If experimental results are supportive of the hypothesis, new insights are gained, otherwise experiments or hypotheses are altered. With the usage of high-throughput technologies, however, the scientific method shifted to data-driven explorative analysis where numerous measurements are evaluated without prior hypothesis [133]. The applicability of this method relies on computational platforms that are able to handle the large amounts of data and can judge relevance based on statistical models.

Enhort supports this new method of explorative analysis by allowing repeatedly altering parameters and instantaneously displaying results. As shown, Enhort is capable of handling large amounts of integration sites and annotation data. The filtering of most significant findings to present the user with the relevant results could also be shown in numerous analysis. Furthermore, this approach can be used to generate novel hypothesis, like observed preferences for certain genetic features of a retrovirus. Based on hypotheses found in explorative analysis, experiments for validation can be designed that verify the observation, conjoining traditional scientific reasoning with high-throughput explorative analysis.

A Appendix

A.1 NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling – Supplements

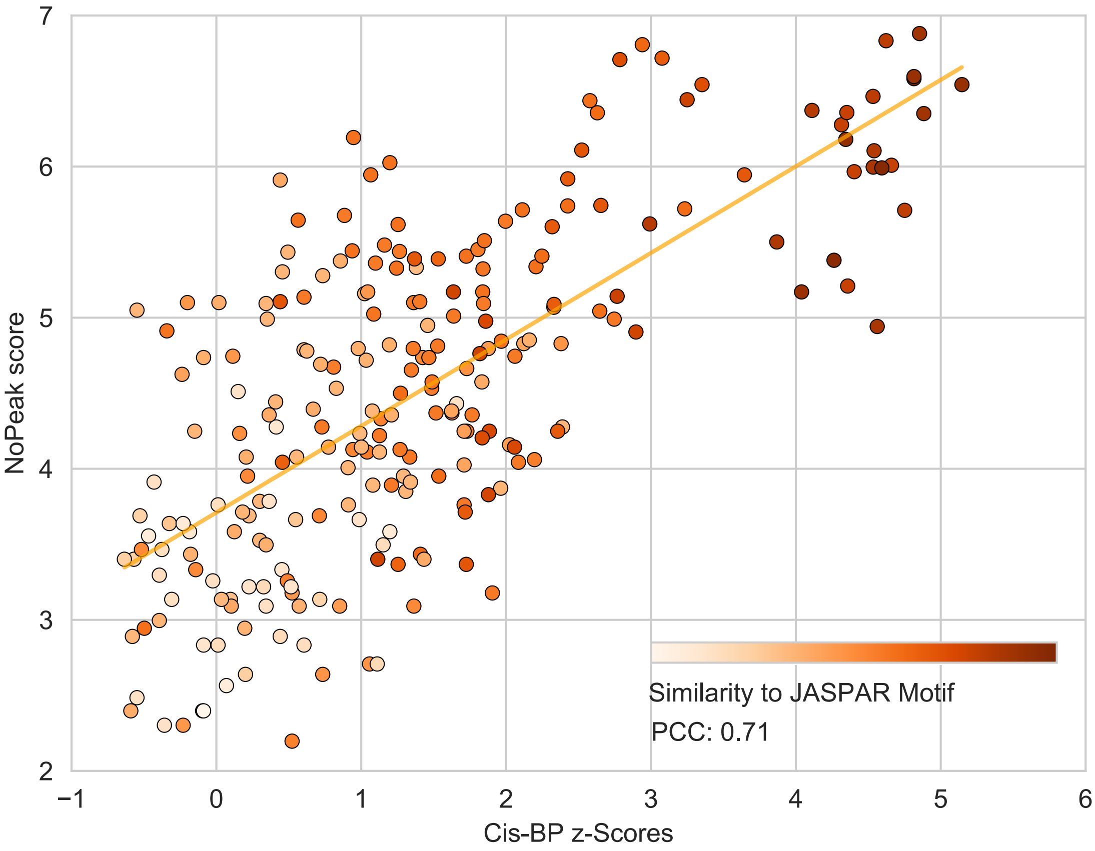
The supplements shown here are an abridged version of the full document. Tables with Encode experiment IDs (Section 1.10, 1.12), correlation plots of ChIP-Seq experiments (Section 1.2) and a flow chart for the peak identification challenge (Figure 1.8) were removed. The complete Supplements are published online with the original paper[1].

[1] Available here: https://doi.org/10.1093/bioinformatics/btaa845

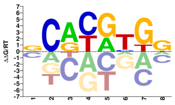
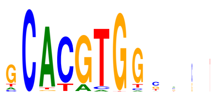
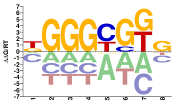
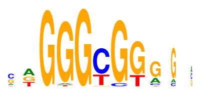
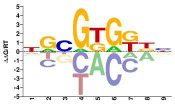
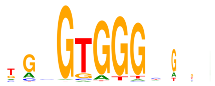
1 Supplements NoPeak

1.1 Similarity between PBM and NoPeak Scores

The scatterplot shows the correlation between 8-mer PBM scores and 8-mer NoPeak scores for the transcription factor (TF) MAX. Additionally, 8-mers are colored based on the Jaccard distance between the 8-mer and the PWM from JASPAR calculated by MACRO-APE.

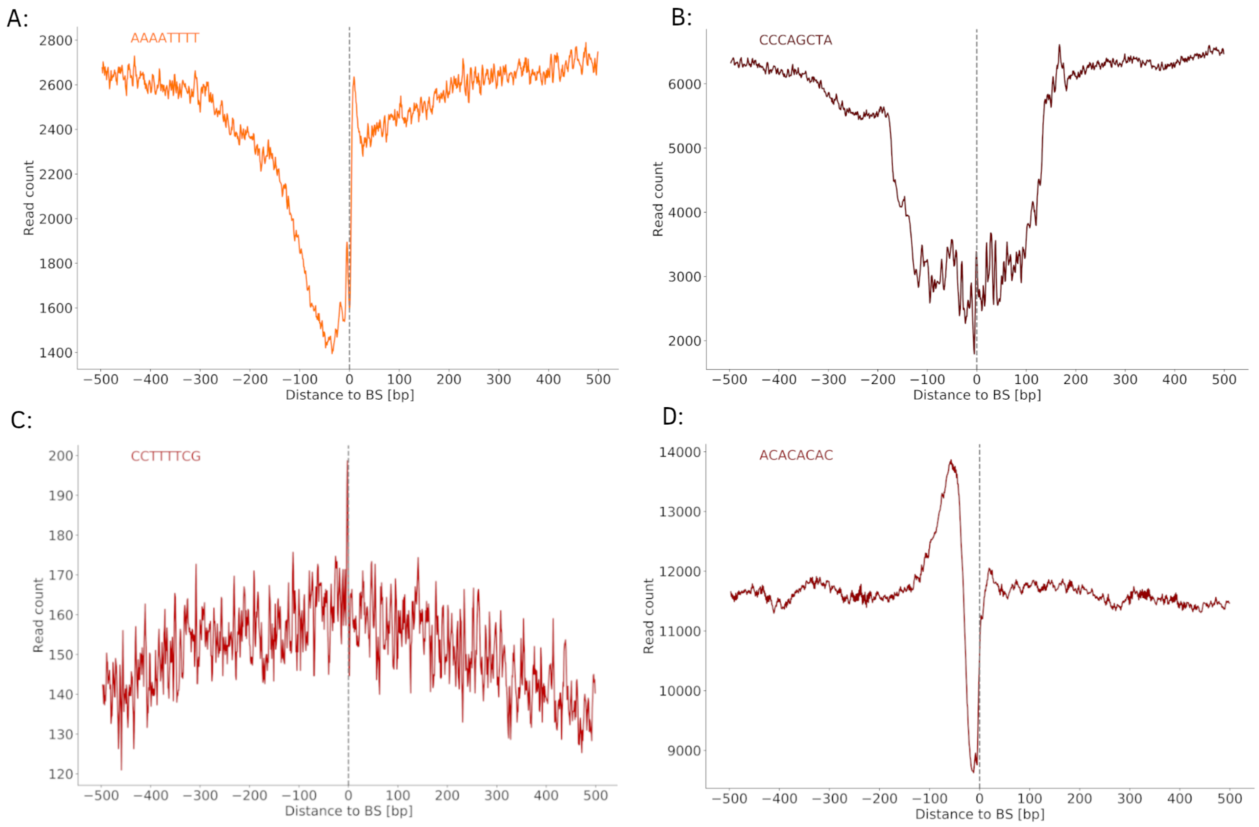


The table shows the correlation between k-mer scores and logos from PBM experiments against NoPeak with ChIP-Seq data. Three PBM 8-mer score data sets from Cis-BP data base are correlated using the Pearson correlation coefficient against scores from NoPeak based on ChIP-Seq data from ENCODE and Short Read Archive (SRA). Additionally, sequence logos generated by both are shown.

Name	PBM dataset	NoPeak dataset	PCC	Logo PBM	Logo Nopeak
MAX	MAX_L (Guo et al. 2014)	ENCFFF000NDT	0.71		
KLF1	KLF1 REF (Barreara et al. 2016)	SRR3083226	0.38		
EGR2	EGR2_D383Y (Barreara et al. 2016)	SRR3083136	0.30		

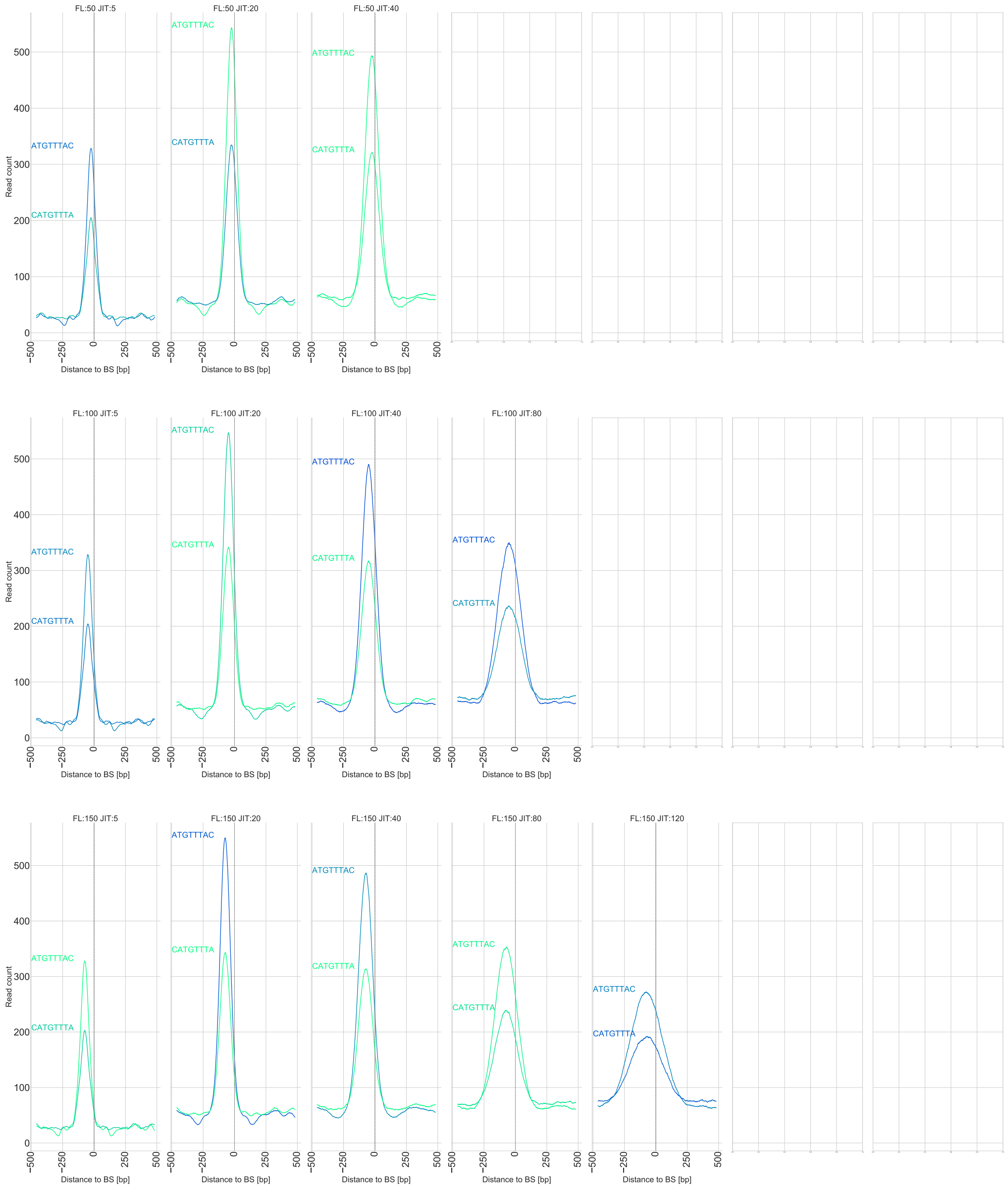
1.3 Filtered profiles

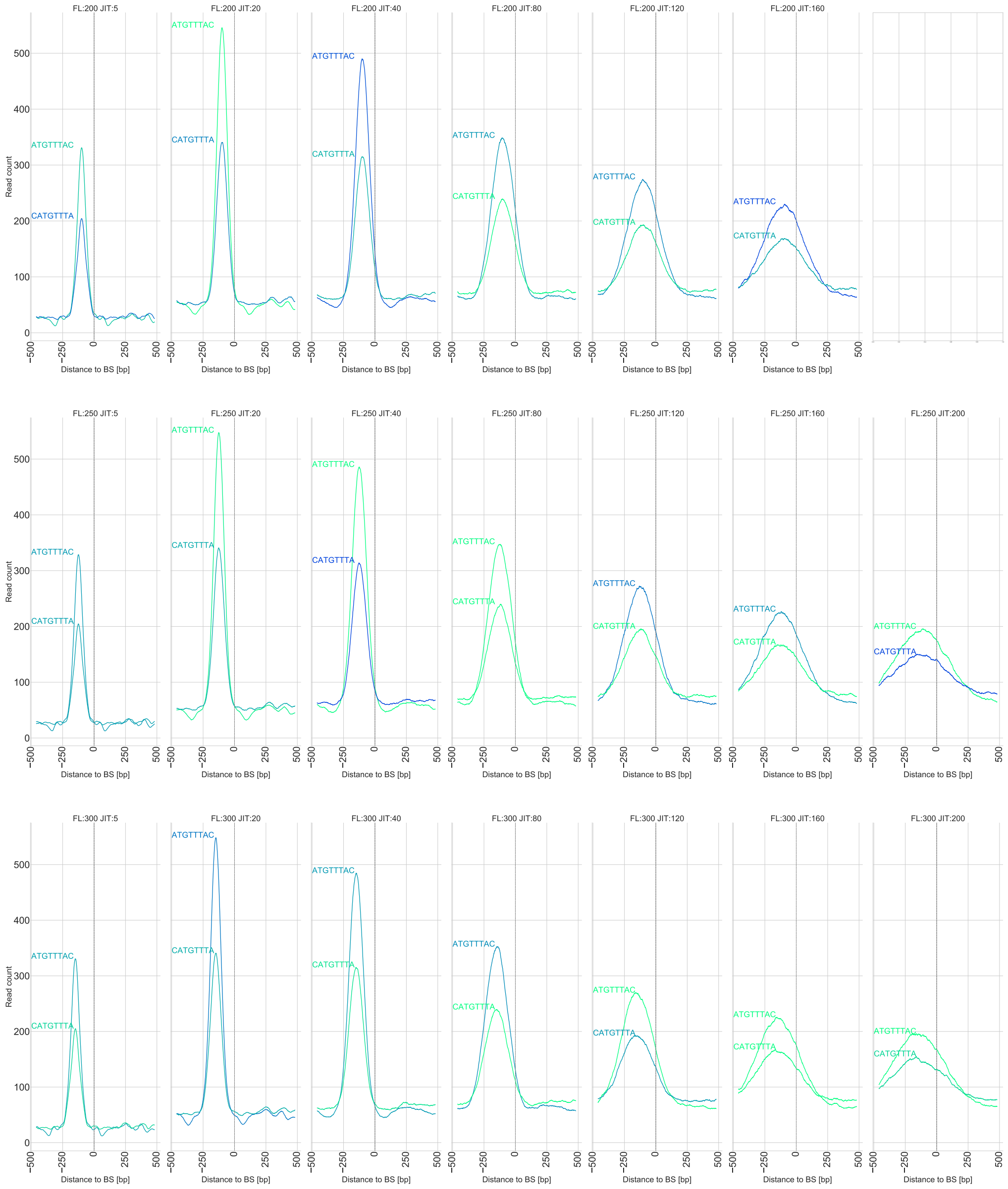
We assume that profile shapes give insights on binding characteristics. The shape can also be used to filter k-mers that are located in repetitive regions, such as $AAAATTTT$. Several shapes for the TF GABPA are shown here that are deviant from our expectation and are filtered by NoPeak as explained in the main manuscript.

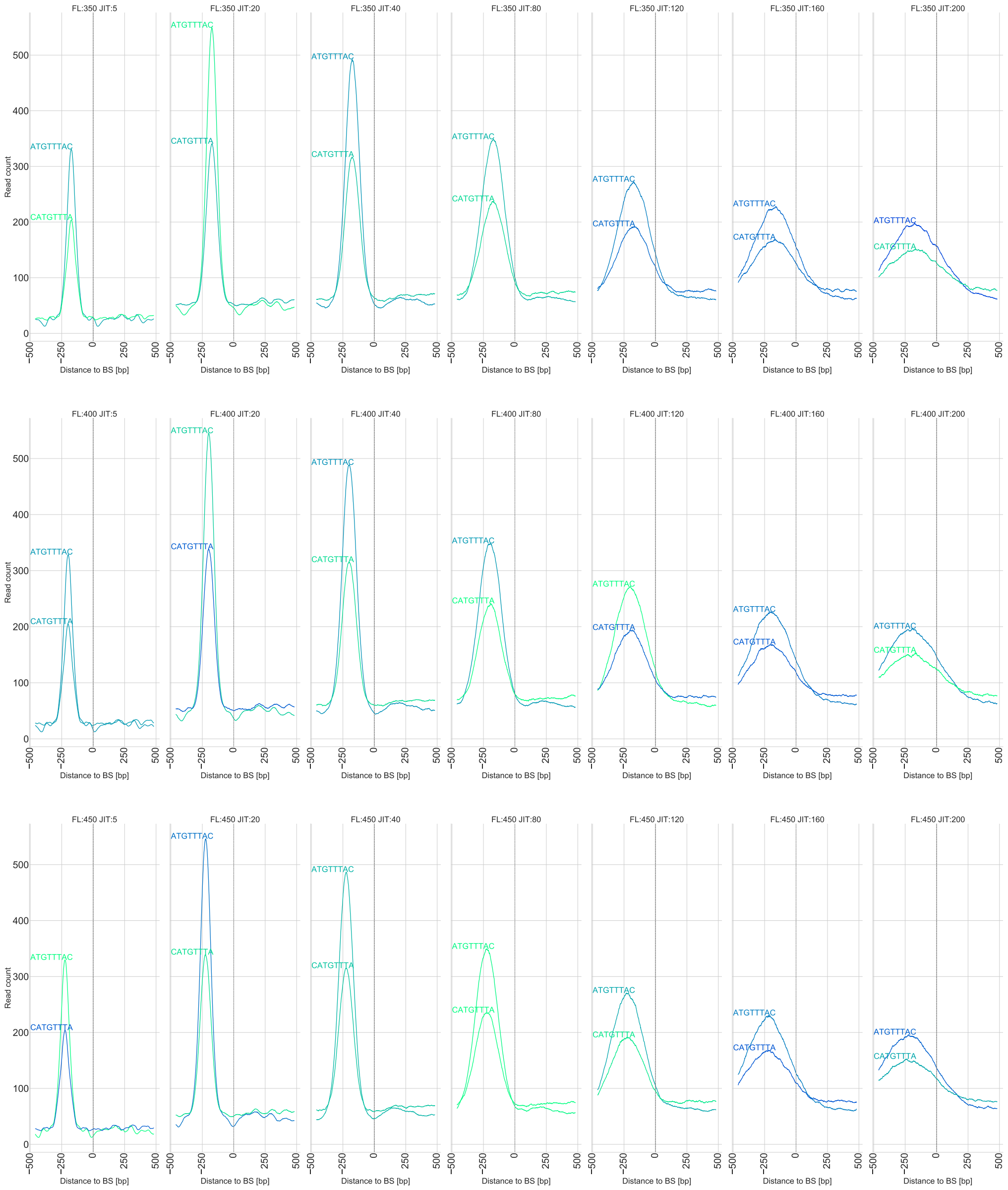


1.4 ChIPulate simulated read profiles

Using ChIPulate (Datta et al. 2019) synthetic reads were created with changing fragment length (FL) and jitter (JIT). Subsequently, the reads were mapped and k-mer profiles were build using NoPeak. For each combination of FL and jitter the k-mer profiles for two selected k-mers are plotted to visualize the effect of both parameters on profile shape and location.

FL:50 JIT:5
FL:50 JIT:20
FL:50 JIT:40
ATGTTTAC
CATGTTTA
Read count
Distance to BS [bp]
FL:100 JIT:5
FL:100 JIT:20
FL:100 JIT:40
FL:100 JIT:80
FL:150 JIT:5
FL:150 JIT:20
FL:150 JIT:40
FL:150 JIT:80
FL:150 JIT:120
0
100
200
300
400
500
−500
−250
0
250
500

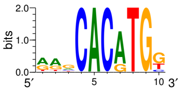
FL:200 JIT:5
FL:200 JIT:20
FL:200 JIT:40
FL:200 JIT:80
FL:200 JIT:120
FL:200 JIT:160
FL:250 JIT:5
FL:250 JIT:20
FL:250 JIT:40
FL:250 JIT:80
FL:250 JIT:120
FL:250 JIT:160
FL:250 JIT:200
FL:300 JIT:5
FL:300 JIT:20
FL:300 JIT:40
FL:300 JIT:80
FL:300 JIT:120
FL:300 JIT:160
FL:300 JIT:200
ATGTTTAC
CATGTTTA
Read count
0
100
200
300
400
500
−500
−250
0
250
500
Distance to BS [bp]

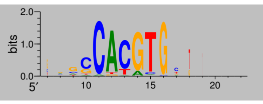
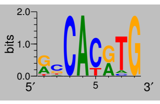
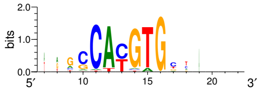
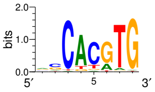
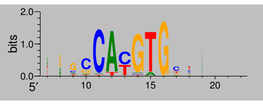
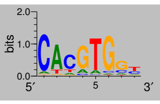
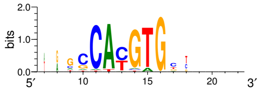
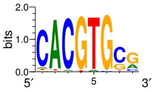
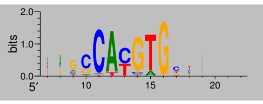
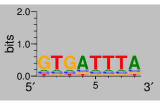
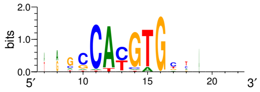
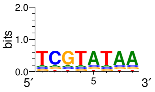
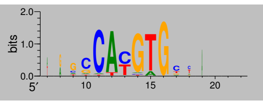
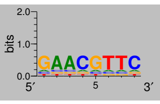
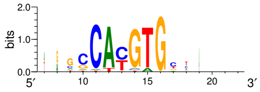
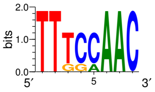
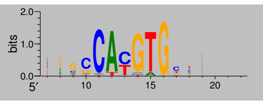
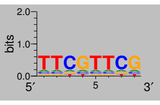
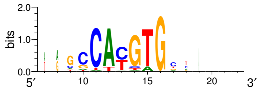

FL:350 JIT:5
FL:350 JIT:20
FL:350 JIT:40
FL:350 JIT:80
FL:350 JIT:120
FL:350 JIT:160
FL:350 JIT:200
FL:400 JIT:5
FL:400 JIT:20
FL:400 JIT:40
FL:400 JIT:80
FL:400 JIT:120
FL:400 JIT:160
FL:400 JIT:200
FL:450 JIT:5
FL:450 JIT:20
FL:450 JIT:40
FL:450 JIT:80
FL:450 JIT:120
FL:450 JIT:160
FL:450 JIT:200
ATGTTTAC
CATGTTTA
Read count
0
100
200
300
400
500
Distance to BS [bp]
−500
−250
0
250
500

1.5 Read reduce from peaks

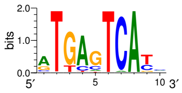
The table shows the performance of NoPeak and Homer in regard to ChIP-Seq analysis where the reads were removed iteratively. Scores and peak count are the same as shown in Fig. 7 in the main manuscript. Sequence logos were generated using WebLogo with the best fitting PWM from Homer and the top-scoring from NoPeak in each round.

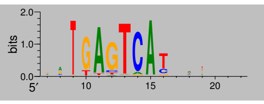
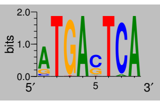
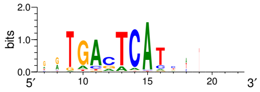
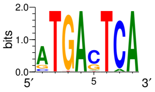
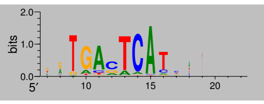
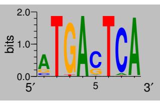
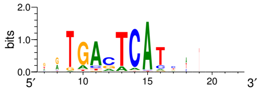
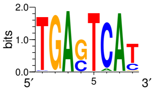
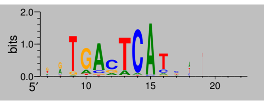
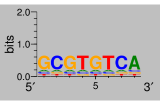
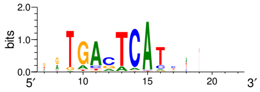
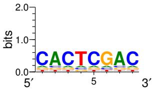
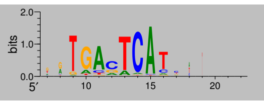
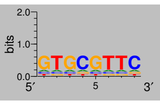
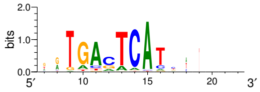
ENCODE ID: ENCFF000NDT
TF: MAX
JASPAR sequence logo:



Round	Score NoPeak	Score HOMER	Peak count	Read count	Logo NoPeak	Logo HOMER
0	0.18	0.29	54250	29729565		
1	0.21	0.38	6467	24909911		
2	0.21	0.3	728	24696286		
3	0.23	0.25	210	24673770		
4	0.19	0.01	106	24667343		
5	0.2	0.0	13	24664359		
6	0.16	0.0	4	24663836		
7	0.21	0.0	5	24663678		
8	0.21	0.0	2	24663499		
9	0.19	0.0	0	24663459		

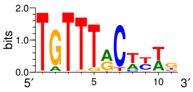
ENCODE ID: ENCFF000YAM
TF: Jun

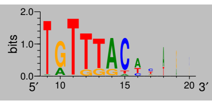
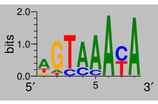
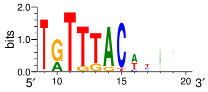
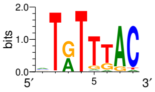
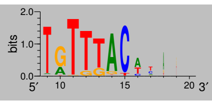
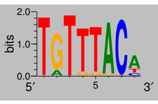
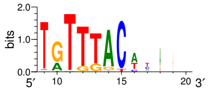
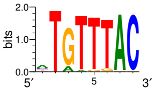
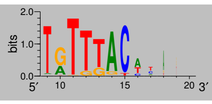
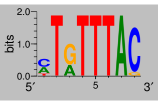
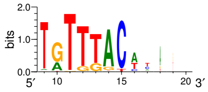
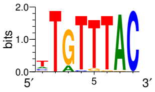
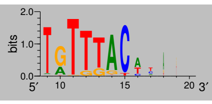
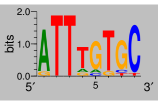
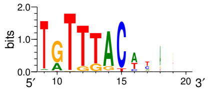
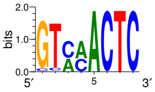
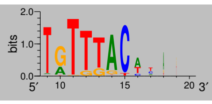
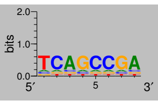
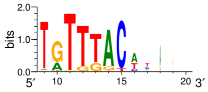


Round	Score NoPeak	Score HOMER	Peak count	Read count	Logo NoPeak	Logo HOMER
0	0.37	0.24	47665	17233661		
1	0.14	0.26	3102	14877486		
2	0.13	0.35	242	14818454		
3	0.13	0.24	106	14813203		
4	0.13	0.03	4	14811614		
5	0.13	0.0	3	14811526		
6	0.13	0.0	2	14811434		
7	0.13	0.0	0	14811392		

ENCODE ID: ENCFF000PIJ
TF: FOXA1

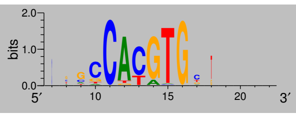
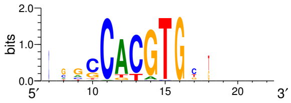
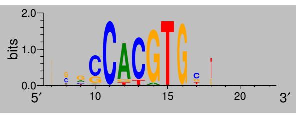
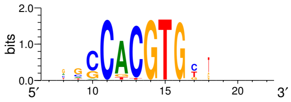
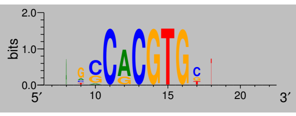
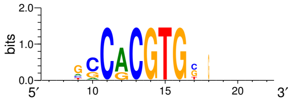
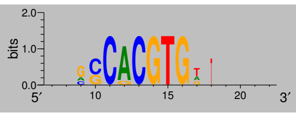
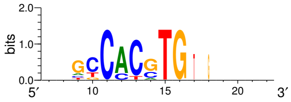
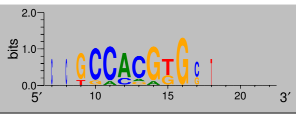
JASPAR sequence logo:



Round	Score NoPeak	Score HOMER	Peak count	Read count	Logo NoPeak	Logo HOMER
0	0.26	0.15	105973	24886701		
1	0.26	0.14	31937	16402986		
2	0.24	0.13	3320	15829289		
3	0.23	0.1	430	15775679		
4	0.23	0.1	65	15768441		
5	0.23	0.1	510	15767268		
6	0.23	0.01	24	15761070		
7	0.25	0.02	27	15760719		
8	0.23	0.0	2	15760358		
9	0.23	0.0	0	15760344		

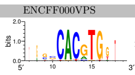
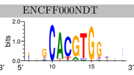
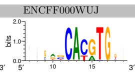
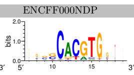
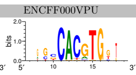
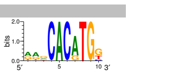
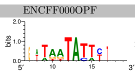
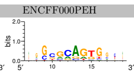
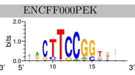
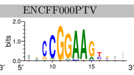
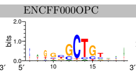
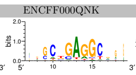
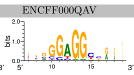
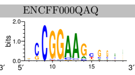
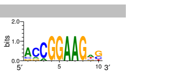
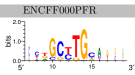
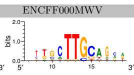
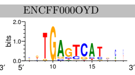
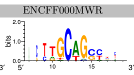
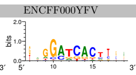
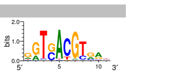
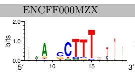
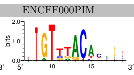
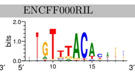
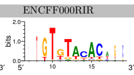
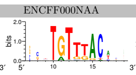
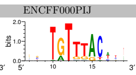
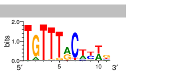
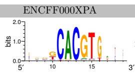
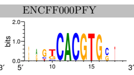
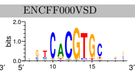
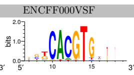
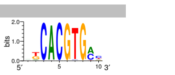
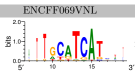
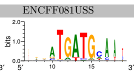
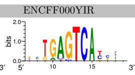
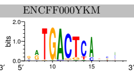
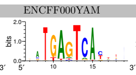
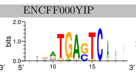
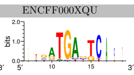
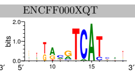
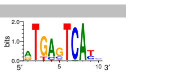
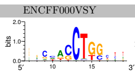
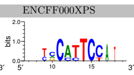
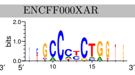
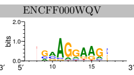
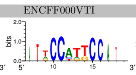
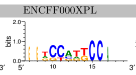
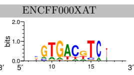
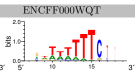
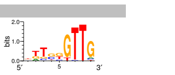
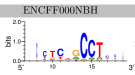
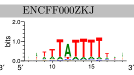
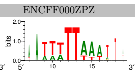
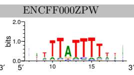
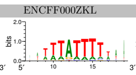
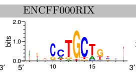
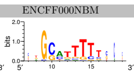
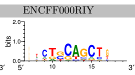
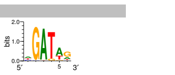
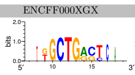
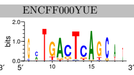
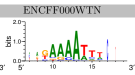
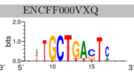
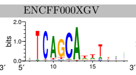
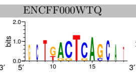
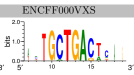
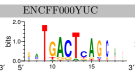
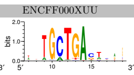
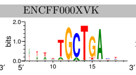
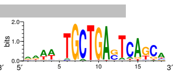
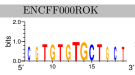
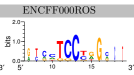
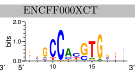
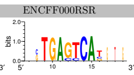
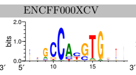
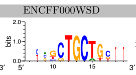
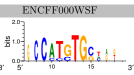
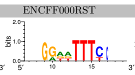
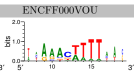
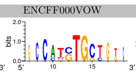
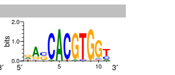
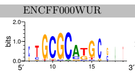
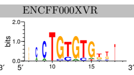
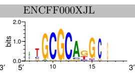
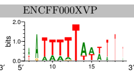
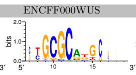
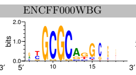
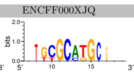
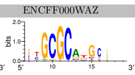
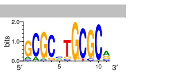
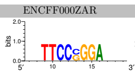
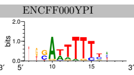
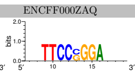
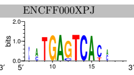
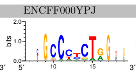
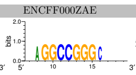
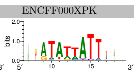
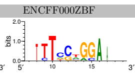
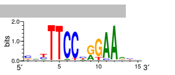
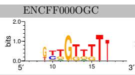
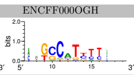
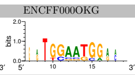
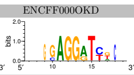
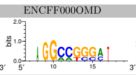
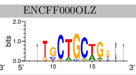
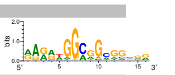
1.5.1 Read reduce TF MAX

Logos generated using NoPeak when reducing not only reads from peaks but all reads for the TF MAX. The table shows the sequence logo genereated by NoPeak if reads are removed.

Round	Peak count	Read count	Logo NoPeak
0	54250	29729565	
1	41226	14663914	
2	27956	7260641	
3	15801	3594423	
4	7151	1780026	
5	2175	881200	
6	320	436357	
7	25	216222	
8	3	107298	

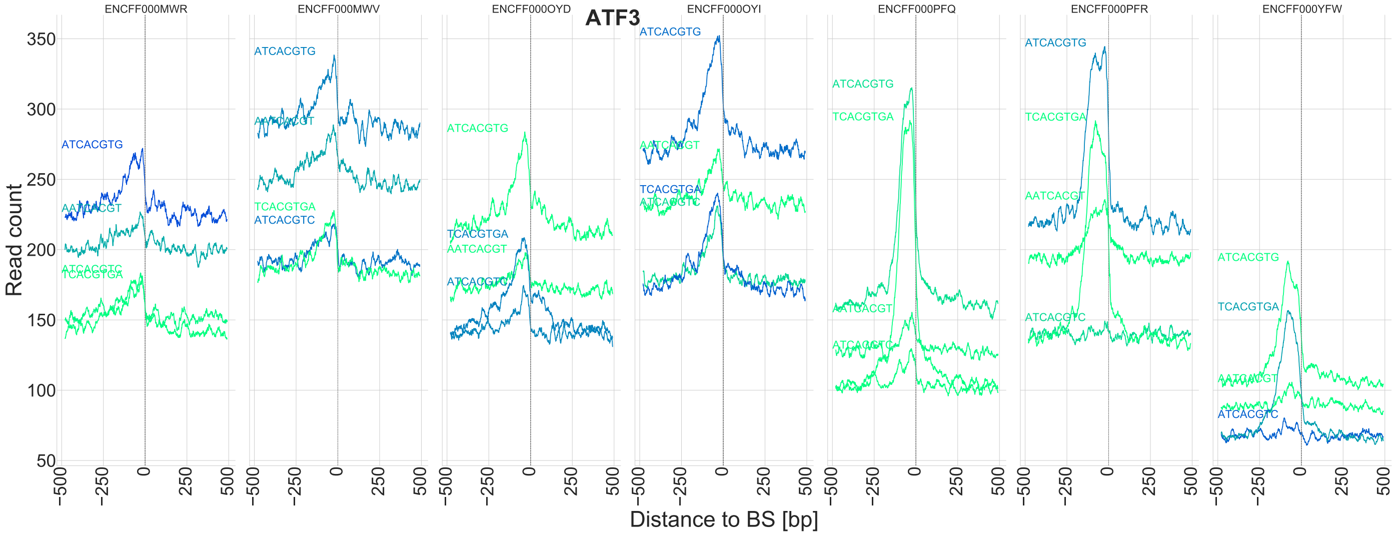
1.6 Sequence logos build by NoPeak

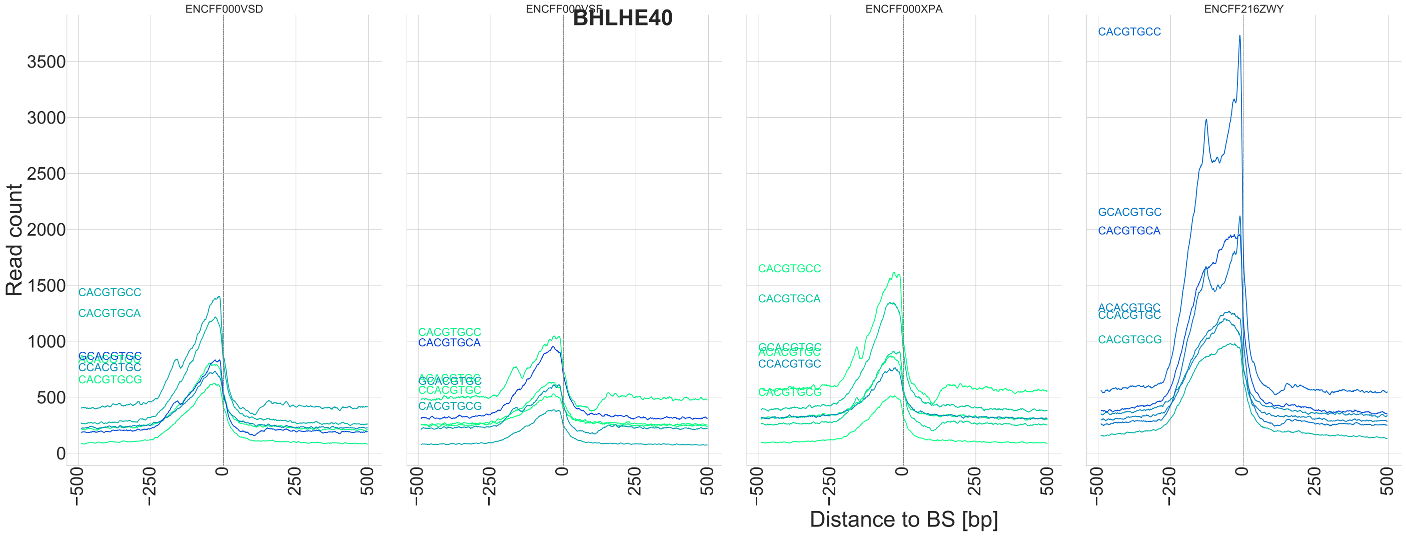
The following table contains sequence logos build using NoPeak based on single ChIP-Seq experiments. The ENCODE ID of each experiment is noted above the sequence logo. The rightmost column contains the sequence logo from the JASPAR database or from HOCOMOCO (Kulakovskiy et. al 2018) if the motif was not available in JASPAR (MAFK and TAF1). All logos were build using WebLogo.

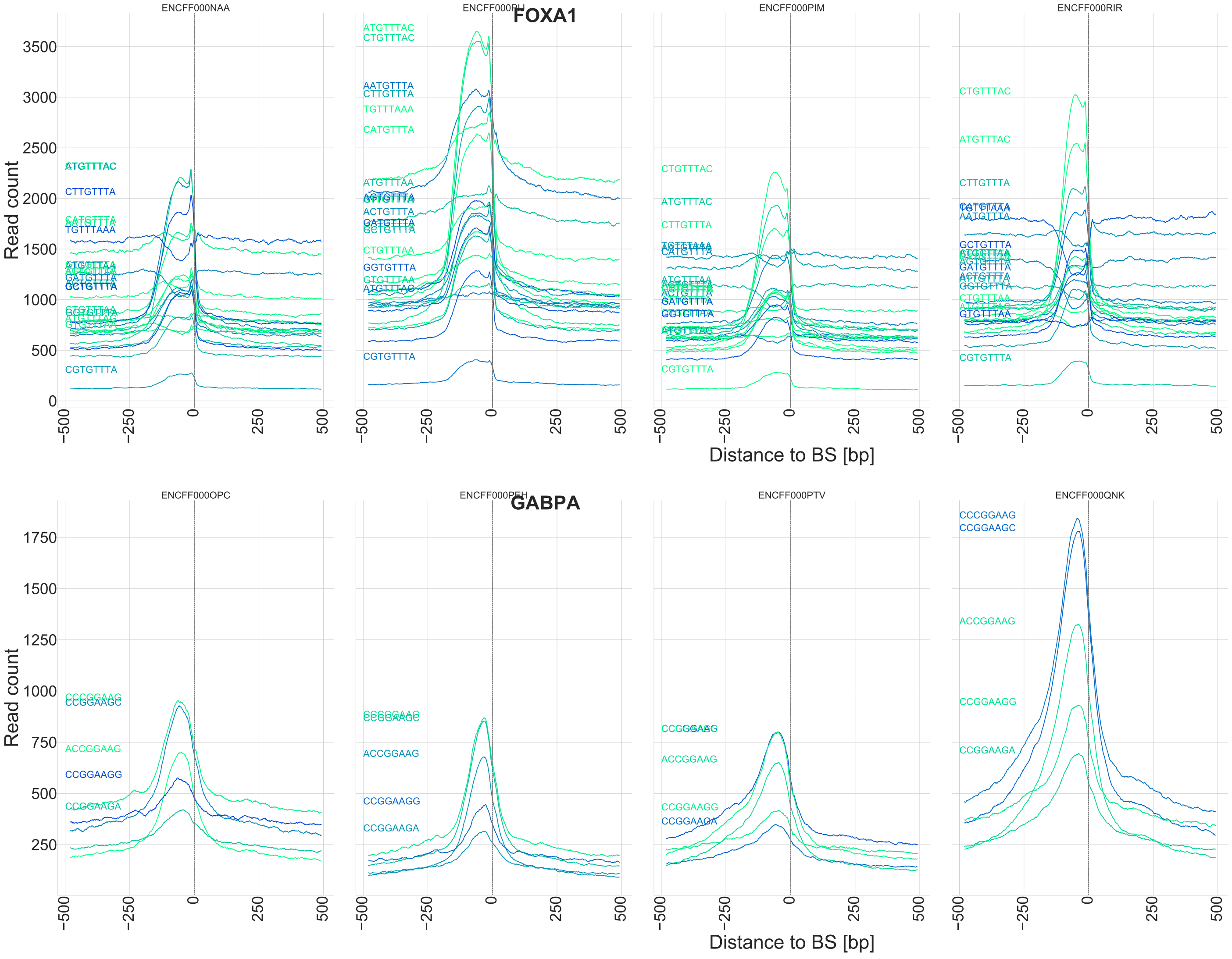
TF											JASPAR / HOCOMOCO
MAX	ENCFF000VPS 	ENCFF000NDT 	ENCFF000WUJ 	ENCFF000NDP 	ENCFF000VPU 						
GABPA	ENCFF000OPF 	ENCFF000PEH 	ENCFF000PEK 	ENCFF000PTV 	ENCFF000OPC 	ENCFF000QNK 	ENCFF000QAV 	ENCFF000QAQ 			
ATF3	ENCFF000PFR 	ENCFF000MWV 	ENCFF000OYD 	ENCFF000MWR 	ENCFF000YFV 						
FOXA1	ENCFF000MZX 	ENCFF000PIM 	ENCFF000RIL 	ENCFF000RIR 	ENCFF000NAA 	ENCFF000PIJ 					
BHLHE40	ENCFF000XPA 	ENCFF000PFY 	ENCFF000VSD 	ENCFF000VSF 							
JUN	ENCFF069VNL 	ENCFF081USS 	ENCFF000YIR 	ENCFF000YKM 	ENCFF000YAM 	ENCFF000YIP 	ENCFF000XQU 	ENCFF000XQT 			
BRCA1	ENCFF000VSY 	ENCFF000XPS 	ENCFF000XAR 	ENCFF000WQV 	ENCFF000VTI 	ENCFF000XPL 	ENCFF000XAT 	ENCFF000WQT 			
GATA3	ENCFF000NBH 	ENCFF000ZKJ 	ENCFF000ZPZ 	ENCFF000ZPW 	ENCFF000ZKL 	ENCFF000RIX 	ENCFF000NBM 	ENCFF000RIY 			
MAFK	ENCFF000XGX 	ENCFF000YUE 	ENCFF000WTN 	ENCFF000VXQ 	ENCFF000XGV 	ENCFF000WTQ 	ENCFF000VXS 	ENCFF000YUC 	ENCFF000XUU 	ENCFF000XVK 	
MYC	ENCFF000ROK 	ENCFF000ROS 	ENCFF000XCT 	ENCFF000RSR 	ENCFF000XCV 	ENCFF000WSD 	ENCFF000WSF 	ENCFF000RST 	ENCFF000VOU 	ENCFF000VOW 	
NRF1	ENCFF000WUR 	ENCFF000XVR 	ENCFF000XJL 	ENCFF000XVP 	ENCFF000WUS 	ENCFF000WBG 	ENCFF000XJQ 	ENCFF000WAZ 			
STAT1	ENCFF000ZAR 	ENCFF000YPI 	ENCFF000ZAQ 	ENCFF000XPJ 	ENCFF000YPJ 	ENCFF000ZAE 	ENCFF000XPK 	ENCFF000ZBF 			
TAF1	ENCFF000OGC 	ENCFF000OGH 	ENCFF000OKG 	ENCFF000OKD 	ENCFF000OMD 	ENCFF000OLZ 					

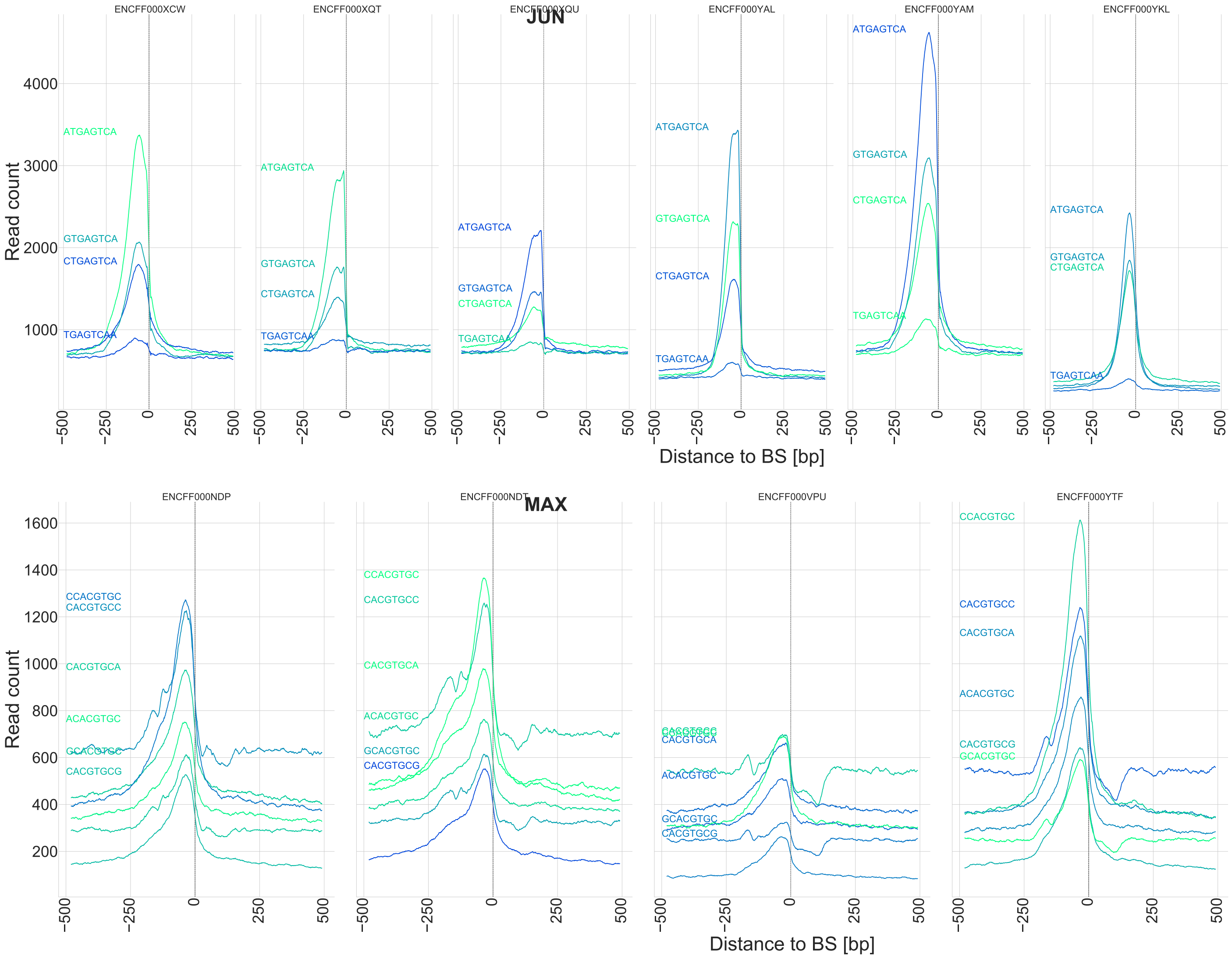
1.7 Read distribution profiles

Read distribution profiles for several TFs from different experiments. For each TF profiles from k-mers close to the expected motif sequence are shown.


ATF3
ENCFF000MWR
ENCFF000MWV
ENCFF000OYD
ENCFF000OYI
ENCFF000PFQ
ENCFF000PFR
ENCFF000YFW
Read count
350
300
250
200
150
100
50
ATCACGTG
AATCACGT
ATCACGTC
TCACGTGA
ATCACGTG
TCACGTGA
ATCACGTC
ATCACGTG
TCACGTGA
AATCACGT
ATCACGTC
ATCACGTG
TCACGTGA
ATCACGTC
ATCACGTG
AATCACGT
ATCACGTC
ATCACGTG
TCACGTGA
AATCACGT
ATCACGTC
ATCACGTG
TCACGTGA
AATCACGT
ATCACGTC
−500
−250
0
250
500
Distance to BS [bp]

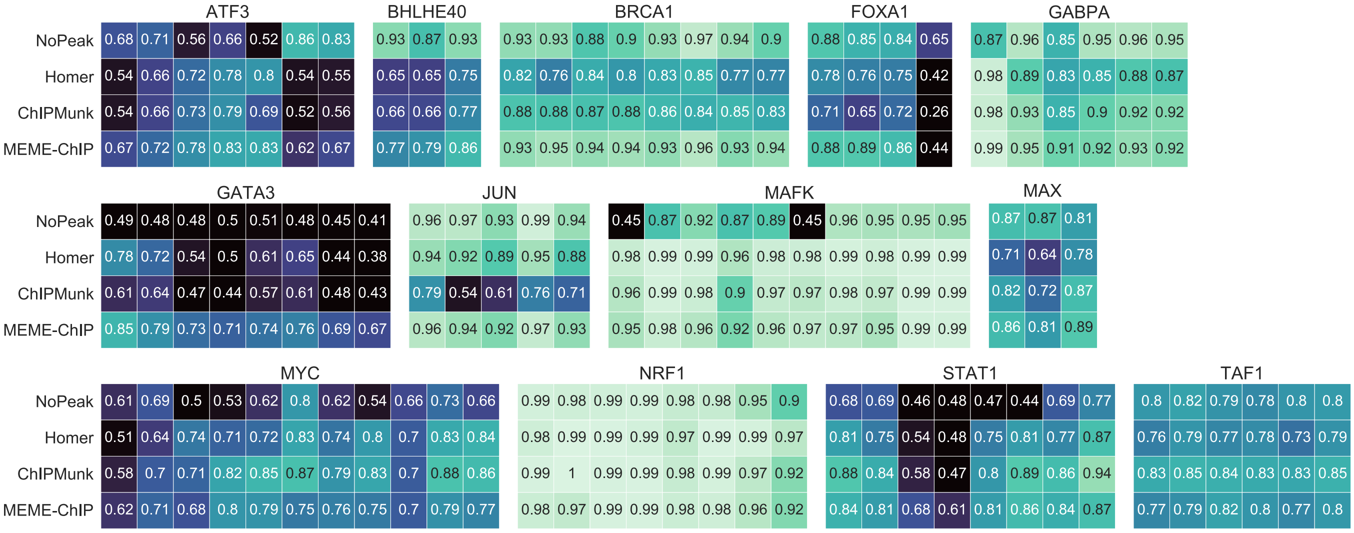

BHLHE40
ENCFF000VSD
ENCFF000VSE
ENCFF000XPA
ENCFF216ZWY
Read count
3500
3000
2500
2000
1500
1000
500
0
CACGTGCC
CACGTGCA
GCACGTGC
CCACGTGC
CACGTGCG
CACGTGCC
CACGTGCA
ACACGTGC
GCACGTGC
CCACGTGC
CACGTGCG
CACGTGCC
CACGTGCA
GCACGTGC
ACACGTGC
CCACGTGC
CACGTGCG
CACGTGCC
GCACGTGC
CACGTGCA
ACACGTGC
CCACGTGC
CACGTGCG
−500
−250
0
250
500
Distance to BS [bp]


FOXA1
ENCFF000NAA
ENCFF000PIM
ENCFF000RIR
Read count
3500
3000
2500
2000
1500
1000
500
0
−500
−250
0
250
500
Distance to BS [bp]
ATGTTTAC
CTTGTTTA
CATGTTTA
TGTTTAAA
GATGTTTA
GCTGTTTA
CGTGTTTA
CTGTTTAC
AATGTTTA
TGTTTAAA
CATGTTTA
ATGTTTAA
ACTGTTTA
GATGTTTA
GCTGTTTA
CTGTTTAA
GGTGTTTA
GTGTTTAA
ATGTTTAG
ATGTTTAA
GATGTTTA
GGTGTTTA
ATGTTTAC
CTTGTTTA
GCTGTTTA
ACTGTTTA
GTGTTTAA
GABPA
ENCFF000OPC
ENCFF000PTV
ENCFF000QNK
1750
1500
1250
1000
750
500
250
CCCGGAAG
CCGGAAGC
ACCGGAAG
CCGGAAGG
CCGGAAGA


JUN
ENCFF000XCW
ENCFF000XQT
ENCFF000XQU
ENCFF000YAL
ENCFF000YAM
ENCFF000YKL
Read count
4000
3000
2000
1000
ATGAGTCA
GTGAGTCA
CTGAGTCA
TGAGTCAA
-500
-250
0
250
500
Distance to BS [bp]
MAX
ENCFF000NDP
ENCFF000NDT
ENCFF000VPU
ENCFF000YTF
1600
1400
1200
1000
800
600
400
200
CCACGTGC
CACGTGCC
CACGTGCA
ACACGTGC
GCACGTGC
CACGTGCG
GCACGTGC
CACGTGCG
Distance to BS [bp]

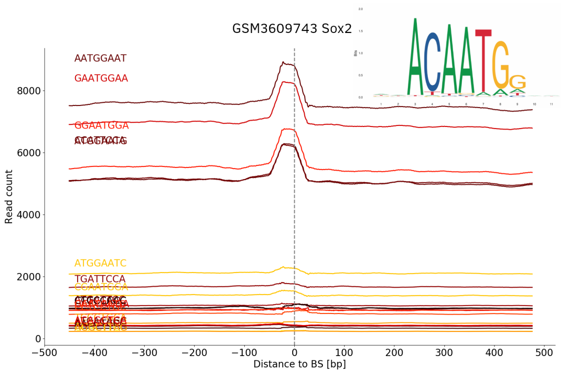
1.11 Peak discrimination of additional experiments

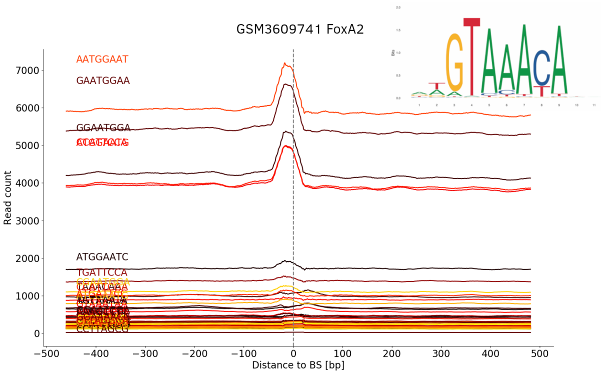
Peaks from 13 commonly used TFs that were also used throughout the manuscript were tested for peak discrimination equivalent to Section 3.4 in the main manuscript.



1.13 CUT&RUN profile shapes

In addition to ChIP-Seq data we used TFBS CUT&RUN[1] data with NoPeak from Meers et al.[2] to build profiles (k: 8bp, radius: 500bp). The 10 most significant k-mers are plotted with their k-mer as well as the expected sequence logo from JASPAR:



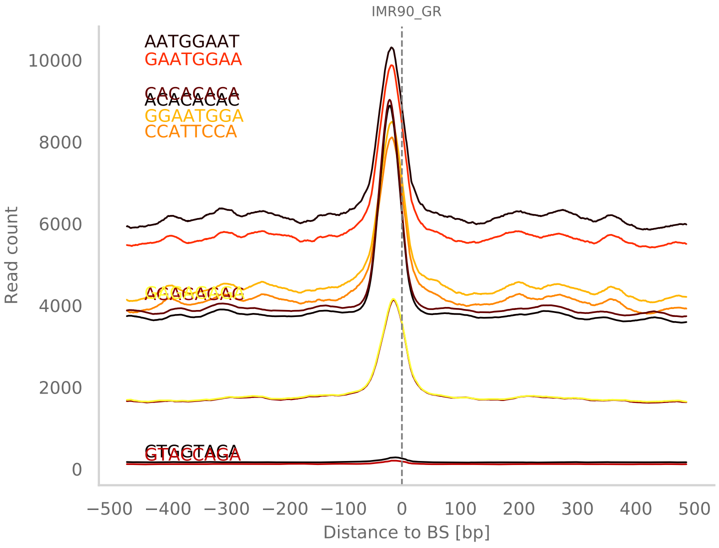


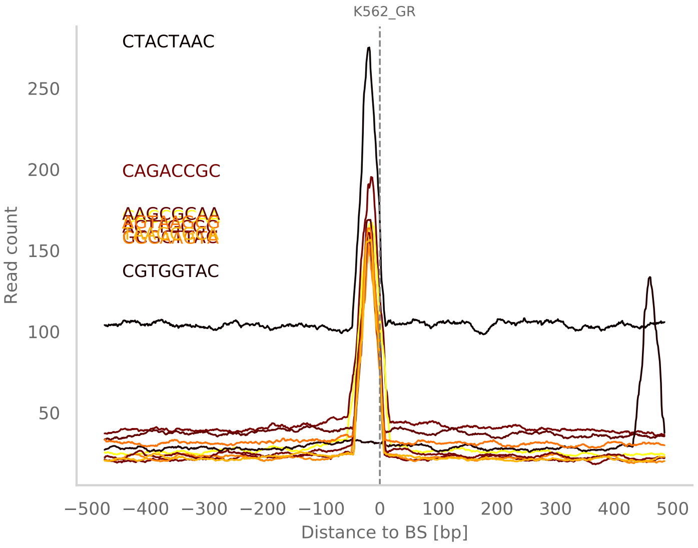
[1] An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites

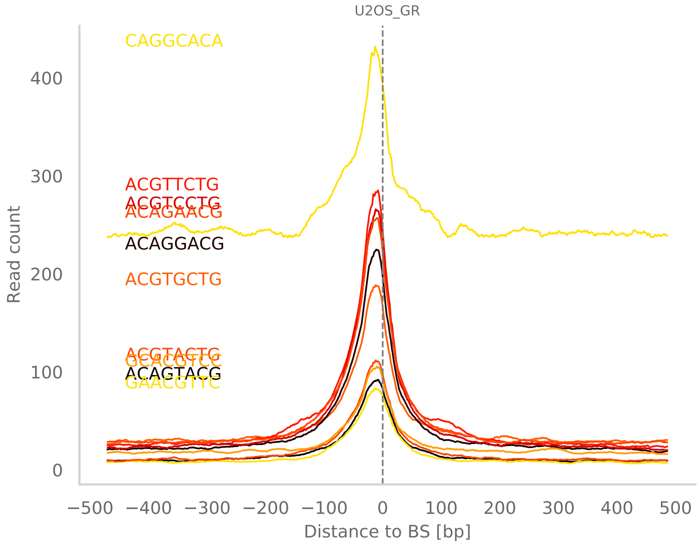
[2] Meers, Michael P et al. "Improved CUT&RUN chromatin profiling tools." eLife vol. 8 e46314. 24 Jun. 2019, doi:10.7554/eLife.46314

1.14 ChIP-Exo Data

Three ChIP-Exo raw datasets were downloaded from Arrayexpress[3], profiles were build using NoPeak (k: 8bp, radius: 500bp) and the 10 most significant profiles are plotted here with their respective k-mer:







[3] https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2956/

Bibliography

[1] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), February 2021. World Wide Web URL: https://omim.org/statistics/geneMap, 2020. 1.1

[2] N. S. Abul-Husn, K. Manickam, L. K. Jones, E. A. Wright, D. N. Hartzel, C. Gonzaga-Jauregui, C. O'Dushlaine, J. B. Leader, H. Lester Kirchner, D. M. Lindbuchler, M. L. Barr, M. A. Giovanni, M. D. Ritchie, J. D. Overton, J. G. Reid, R. P. R. Metpally, A. H. Wardeh, I. B. Borecki, G. D. Yancopoulos, A. Baras, A. R. Shuldiner, O. Gottesman, D. H. Ledbetter, D. J. Carey, F. E. Dewey, and M. F. Murray. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science*, 354(6319), 2016. 1.1

[3] D. Accili and K. C. Arden. FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell*, 117(4):421–426, 2004. 1.3

[4] E. J. Aird, K. N. Lovendahl, A. St. Martin, R. S. Harris, and W. R. Gordon. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Communications Biology*, 1(1), Dec. 2018. 3.2.3

[5] A. Aiuti, F. Cattaneo, S. Galimberti, U. Benninghoff, B. Cassani, L. Callegaro, S. Scaramuzza, G. Andolfi, M. Mirolo, and I. Brigida. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *New England Journal of Medicine*, 360(5):447–458, 2009. 1.1

[6] M. Amberger and Z. Ivics. Latest Advances for the Sleeping Beauty Transposon System: 23 Years of Insomnia but Prettier than Ever. *BioEssays*, 42(11):2000136, 2020. 1.2

[7] I. Ammar, A. Gogol-Döring, C. Miskey, W. Chen, T. Cathomen, Z. Izsvák, and Z. Ivics. Retargeting transposon insertions by the adeno-associated virus Rep protein. *Nucleic Acids Research*, 40(14):6693–6712, Aug. 2012. 3.2.3

[8] X. M. Anguela and K. A. High. Entering the Modern Era of Gene Therapy. *Annual Review of Medicine*, 70(1):273–288, Jan. 2019. 1.1

[9] M. Annala, K. Laurila, H. Lähdesmäki, and M. Nykter. A Linear Model for Transcription Factor Binding Affinity Prediction in Protein Binding Microarrays. *PLoS ONE*, 6(5):e20059, May 2011. 1.3

[10] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server):W202–W208, July 2009. 1.4

[11] T. L. Bailey and P. Machanick. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):e128–e128, Sept. 2012. 4.1.2

[12] J. W. Bainbridge, A. J. Smith, S. S. Barker, S. Robbie, R. Henderson, K. Balaggan, A. Viswanathan, G. E. Holder, A. Stockman, and N. Tyler. Effect of gene therapy on visual function in Leber's congenital amaurosis. *New England Journal of Medicine*, 358(21):2231–2239, 2008. 1.1

[13] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007. 1.4

[14] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, Nov. 2006. 3.1.1, 4.1

[15] G. Blattner, A. Cavazza, A. J. Thrasher, and G. Turchiano. Gene Editing and Genotoxicity: Targeting the Off-Targets. *Frontiers in Genome Editing*, 2, Dec. 2020. 1.2, 4.2

[16] P. O. Brown, B. Bowerman, H. E. Varmus, and J. M. Bishop. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proceedings of the National Academy of Sciences*, 86(8):2525–2529, 1989. 1.2

[17] F. D. Bushman. Retroviral Insertional Mutagenesis in Humans: Evidence for Four Genetic Mechanisms Promoting Expansion of Cell Clones. *Molecular Therapy*, 28(2):352–356, Feb. 2020. 1.1, 4.2

[18] M. Cavazzana-Calvo, S. Hacein-Bey, G. d. S. Basile, F. Gross, E. Yvon, P. Nusbaum, F. Selz, C. Hue, S. Certain, J.-L. Casanova, P. Bousso, F. L. Deist, and A. Fischer. Gene Therapy of Human Severe Combined Immunodeficiency (SCID)-X1 Disease. *Science*, 288(5466):669–672, Apr. 2000. 1.1

[19] A. V. Cideciyan, T. S. Aleman, S. L. Boye, S. B. Schwartz, S. Kaushal, A. J. Roman, J.-j. Pang, A. Sumaroka, E. A. Windsor, and J. M. Wilson. Human gene therapy for RPE65 isomerase deficiency activates the retinoid cycle of vision but with slow rod kinetics. *Proceedings of the National Academy of Sciences*, 105(39):15112–15117, 2008. 1.1

[20] M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. MacArthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, and M. I. McCarthy. A brief history of human disease genetics. *Nature*, 577(7789):179–189, Jan. 2020. 1.1

[21] D. B. T. Cox, R. J. Platt, and F. Zhang. Therapeutic genome editing: prospects and challenges. *Nature Medicine*, 21(2):121–131, Feb. 2015. 1.1, 1.1

[22] R. Craigie and F. D. Bushman. HIV DNA Integration. *Cold Spring Harbor Perspectives in Medicine*, 2(7), July 2012. 1.1, 1.2, 3.2.1

[23] M. K. Das and H.-K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(S7), Dec. 2007. 1.4

[24] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. Strattan, O. Jolanki, F. Y. Tanaka, and J. Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, Jan. 2018. 1.4

[25] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. A. Wessels, and J. de Ridder. Chromatin Landscapes of Retroviral and Transposon Integration Profiles. *PLOS Genetics*, 10(4):e1004250, Apr. 2014. 4.2.1

[26] S. S. De Ravin, L. Su, N. Theobald, U. Choi, J. L. Macpherson, M. Poidinger, G. Symonds, S. M. Pond, A. L. Ferris, S. H. Hughes, H. L. Malech, and X. Wu.

Enhancers Are Major Targets for Murine Leukemia Virus Vector Integration. *Journal of Virology*, 88(8):4504–4513, Apr. 2014. 1.1, 3.2.1

[27] J. de Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels. Detecting Statistically Significant Common Insertion Sites in Retroviral Insertional Mutagenesis Screens. *PLOS Computational Biology*, 2(12):1–13, 2006. 4.2.1

[28] S. Desfarges and A. Ciuffi. Retroviral Integration Site Selection. *Viruses*, 2(1):111–130, Jan. 2010. 1.2

[29] DREAM5 Consortium, M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, Feb. 2013. 1.3, 1.5, 3.1.4, 4.1, 4.1.4

[30] I. Dufait, T. Liechtenstein, A. Lanna, C. Bricogne, R. Laranga, A. Padella, K. Breckpot, and D. Escors. Retroviral and Lentiviral Vectors for the Induction of Immunological Tolerance. *Scientifica*, 2012:1–14, 2012. 1.2

[31] E. Fanales-Belasio, M. Raimondo, B. Suligoi, and S. Buttò. HIV virology and pathogenetic mechanisms of infection: a brief overview. *Annali dell'Istituto superiore di sanita*, 46:5–14, 2010. 1.2

[32] E. Fratkin, B. T. Naughton, D. L. Brutlag, and S. Batzoglou. MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, 22(14):e150–e157, July 2006. 3.1

[33] T. Friedmann and R. Roblin. Gene Therapy for Human Genetic Disease? *Science*, 175(4025):949–955, 1972. 1.1

[34] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Müller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram,

B.-J. Kim, Y. Kim, Y. J. Kim, M.-S. Kwon, J. Lee, S. Lee, K.-H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B.-G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stančáková, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S.-H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C.-Y. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farook, S. P. Fowler, B. I. Freedman, M. Griswold, D. E. Hale, P. J. Hicks, C.-C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S.-T. Tan, H. A. Taylor, F. Thameem, G. Wilson, T. Y. Wong, P. R. Njølstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jørgensen, T. Jørgensen, C. Ladenvall, J. M. Justesen, A. Käräjämäki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. Hrabé de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O'Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A.-C. Syvänen, R. N. Bergman, D. Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Mägi, L. Lind, Y. Farjoun, K. R. Owen, A. L. Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J.-Y. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M.

Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burtt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler, and M. I. McCarthy. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, Aug. 2016. 1.1

[35] E. E. M. Furlong and M. Levine. Developmental enhancers and chromosome topology. *Science*, 361(6409):1341, Sept. 2018. 1.3, 1.3

[36] S.-E. Glont, I. Chernukhin, and J. S. Carroll. Comprehensive Genomic Analysis Reveals that the Pioneering Function of FOXA1 Is Independent of Hormonal Signaling. *Cell Reports*, 26(10):2558–2565.e3, Mar. 2019. 3.1.2

[37] A. Gogol-Döring, I. Ammar, S. Gupta, M. Bunse, C. Miskey, W. Chen, W. Uckert, T. F. Schulz, Z. Izsvák, and Z. Ivics. Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4+ T Cells. *Molecular Therapy*, 24(3):592–606, Mar. 2016. 1.1, 1.2, 3.2.5

[38] R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. Bulyk. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports*, 3(4):1093–1104, Apr. 2013. 1.3

[39] Y. Guo, K. Tian, H. Zeng, X. Guo, and D. K. Gifford. A novel k -mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Research*, 28(6):891–900, June 2018. 1.3, 3.1, 4.1

[40] S. Hacein-Bey-Abina, A. Garrigue, G. P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B. H. Belohradsky, U. Wintergerst, M. C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F. D. Bushman, A. Fischer, and M. Cavazzana-Calvo. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *Journal of Clinical Investigation*, 118(9):3132–3142, Sept. 2008. 1.1

[41] P. B. Hackett, S. C. Ekker, D. A. Largaespada, and R. S. McIvor. Sleeping Beauty Transposon-Mediated Gene Therapy for Prolonged Expression. In *Advances in Genetics*, volume 54, pages 189–232. Elsevier, 2005. 3.2.3

[42] F. A. Hashim, M. S. Mabrouk, and W. Al-Atabany. Review of different sequence motif finding algorithms. *Avicenna journal of medical biotechnology*, 11(2):130, 2019. 1.4

[43] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, May 2010. 1.4

[44] J. v. Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies11Edited by G. von Heijne. *Journal of Molecular Biology*, 281(5):827 – 842, 1998. 3.1

[45] P. C. Hendrie and D. W. Russell. Gene Targeting with Viral Vectors. *Molecular Therapy*, 12(1):9–17, July 2005. 1.2

[46] L. Holman, M. L. Head, R. Lanfear, and M. D. Jennions. Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol*, 13(7):e1002190, 2015. 4.1

[47] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899–4913, Sept. 2005. 4.1

[48] M. Hudecek and Z. Ivics. Non-viral therapeutic cell engineering with the Sleeping Beauty transposon system. *Current Opinion in Genetics & Development*, 52:100–108, 2018. 3.2.3

[49] D. Hüser, A. Gogol-Döring, W. Chen, and R. Heilbronn. Adeno-associated virus type 2 wild-type and vector-mediated genomic integration profiles of human diploid fibroblasts analyzed by third-generation PacBio DNA sequencing. *Journal of virology*, 88(19):11253–11263, Oct. 2014. 4.2.1

[50] D. Hüser, A. Gogol-Döring, T. Lutter, S. Weger, K. Winter, E.-M. Hammer, T. Cathomen, K. Reinert, and R. Heilbronn. Integration Preferences of Wild-type AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome. *PLoS Pathogens*, 6(7):e1000985, July 2010. 1.2

[51] D. Hüser and R. Heilbronn. Adeno-associated virus integrates site-specifically into human chromosome 19 in either orientation and with equal kinetics and frequency. *Journal of General Virology*, 84(1):133–137, Jan. 2003. 1.2

[52] S. Inukai, K. H. Kock, and M. L. Bulyk. Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, 43:110–119, Apr. 2017. 1.3, 1.4

[53] M. Jackson, L. Marks, G. H. May, and J. Wilson. The genetic basis of disease. *Essays in Biochemistry*, 62(5):643–723, Dec. 2018. 1.1

[54] K. Jadhav and Y. Zhang. Activating transcription factor 3 in immune response and metabolic regulation. *Liver Research*, 1(2):96–102, Sept. 2017. 3.1.2

[55] H. Jeon, H. Lee, B. Kang, I. Jang, and T.-Y. Roh. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. *Genomics & Informatics*, 18(4):e42, Dec. 2020. 1.4

[56] S. Jung-Klawitter, N. Fuchs, K. Upton, A. Froemmrich, C. Miskey, M. Munoz-Lopez, R. Shukla, J. Wang, A. Sebe, S. Merkert, et al. Reprogramming triggers mobilisation of endogenous retrotransposons in human induced pluripotent stem cells with genotoxic effects on host gene expression. *Human Gene Therapy*, 28(12):A2–A2, 2017. 3.2.4, 4.2

[57] M. G. Kaplitt, A. Feigin, C. Tang, H. L. Fitzsimons, P. Mattis, P. A. Lawlor, R. J. Bland, D. Young, K. Strybing, D. Eidelberg, et al. Safety and tolerability of gene therapy with an adeno-associated virus (aav) borne gad gene for parkinson's disease: an open label, phase i trial. *The Lancet*, 369(9579):2097–2105, 2007. 1.1

[58] P. Kebriaei, Z. Izsvák, S. A. Narayanavari, H. Singh, and Z. Ivics. Gene Therapy with the Sleeping Beauty Transposon System. *Trends in Genetics*, 33(11):852–870, Nov. 2017. 1.2

[59] J. Keilwagen and J. Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, 43(18):e119–e119, Oct. 2015. 4.1

[60] S. Khan, M. S. Mahmood, S. u. Rahman, H. Zafar, S. Habibullah, Z. khan, and A. Ahmad. CRISPR/Cas9: the Jedi against the dark empire of diseases. *Journal of Biomedical Science*, 25(1), Dec. 2018. 1.1

[61] G. Khoury and P. Gruss. Enhancer elements. *Cell*, 33(2):313–314, 1983. 1.3

[62] S. Klawitter, N. V. Fuchs, K. R. Upton, M. Muñoz-Lopez, R. Shukla, J. Wang, M. Garcia-Cañadas, C. Lopez-Ruiz, D. J. Gerhardt, A. Sebe, I. Grabundzija, S. Merkert, P. Gerdes, J. A. Pulgarin, A. Bock, U. Held, A. Witthuhn, A. Haase, B. Sarkadi, J. Löwer, E. J. Wolvetang, U. Martin, Z. Ivics, Z. Izsvák, J. L. Garcia-Perez, G. J. Faulkner, and G. G. Schumann. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nature Communications*, 7(1), Apr. 2016. 3.2.4

[63] A. Kovač, C. Miskey, M. Menzel, E. Grueso, A. Gogol-Döring, and Z. Ivics. RNA-guided retargeting of S*leeping Beauty* transposition in human cells. *eLife*, 9:e53868, Mar. 2020. 3.2.3, 4.2

[64] S. Krügener, T. Rose, M. Menzel, A. Krüger, F. Creutzburg, A. Knabe, A. Gogol-Döring, and V. Sandig. Engineered transposases and transposons enforce integration into highly active genomic loci and facilitate optimal transgene expression. *(In preperation)*. 3.2.5, 4.2

[65] I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, F. A. Kolpakov, and V. J. Makeev. 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, Jan. 2018. 1.3

[66] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018. 1.3

[67] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001. 1.2

[68] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357, Mar. 2012. 1.4

[69] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):1–10, 2009. 3.2.3

[70] C. E. Lee, K. S. Singleton, M. Wallin, and V. Faundez. Rare Genetic Diseases: Nature's Experiments on Human Development. *Iscience*, page 101123, 2020. 1.1

[71] R. v. d. Lee, S. Correard, and W. W. Wasserman. Deregulated Regulators: Disease-Causing cis Variants in Transcription Factor Genes. *Trends in Genetics*, 36(7):523–539, 2020. 1.3

[72] V. G. Levitsky, I. V. Kulakovskiy, N. I. Ershov, D. Y. Oshchepkov, V. J. Makeev, T. C. Hodgman, and T. I. Merkulova. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC genomics*, 15(1):80, 2014. 1.4

[73] F. Li, N. Vijayasankaran, A. Y. Shen, R. Kiss, and A. Amanullah. Cell culture processes for monoclonal antibody production. *mAbs*, 2(5):466–479, Sept. 2010. 3.2.5

[74] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. 1.4

[75] X. Li, E. R. Burnight, A. L. Cooney, N. Malani, T. Brady, J. D. Sander, J. Staber, S. J. Wheelan, J. K. Joung, and P. B. McCray. piggyBac transposase tools for genome engineering. *Proceedings of the National Academy of Sciences*, 110(25):E2279–E2287, 2013. 3.2.5, 4.2.1

[76] C. A. Lino, J. C. Harper, J. P. Carney, and J. A. Timlin. Delivering CRISPR: a review of the challenges and approaches. *Drug Delivery*, 25(1):1234–1257, Jan. 2018. 4.2

[77] K. Lundstrom. Viral Vectors in Gene Therapy. *Diseases*, 6(2):42, May 2018. 4.2

[78] A. M. Maguire, F. Simonelli, E. A. Pierce, E. N. Pugh, F. Mingozzi, J. Bennicelli, S. Banfi, K. A. Marshall, F. Testa, E. M. Surace, S. Rossi, A. Lyubarsky, V. R. Arruda, B. Konkle, E. Stone, J. Sun, J. Jacobs, L. Dell'Osso, R. Hertle, J.-x. Ma, T. M. Redmond, X. Zhu, B. Hauck, O. Zelenaia, K. S. Shindler, M. G. Maguire, J. F. Wright, N. J. Volpe, J. W. McDonnell, A. Auricchio, K. A. High, and J. Bennett. Safety and Efficacy of Gene Transfer for Leber's Congenital Amaurosis. *New England Journal of Medicine*, 358(21):2240–2248, May 2008. 1.1

[79] G. K. Marinov, A. Kundaje, P. J. Park, and B. J. Wold. Large-Scale Quality Analysis of Published ChIP-seq Data. *G3&amp;#58; Genes|Genomes|Genetics*, 4(2):209–223, Feb. 2014. 1.5, 4.1, 4.1.3

[80] I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21, Nov. 2017. 1

[81] M. Matasci, L. Baldi, D. L. Hacker, and F. M. Wurm. The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability. *Biotechnology and Bioengineering*, 108(9):2141–2150, Sept. 2011. 3.2.5

[82] A. Mathelier, W. Shi, and W. W. Wasserman. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76, Feb. 2015. 1.3

[83] R. Mejzini, L. L. Flynn, I. L. Pitout, S. Fletcher, S. D. Wilton, and P. A. Akkari. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Frontiers in Neuroscience*, 13, Dec. 2019. 1.1

[84] M. Menzel. *Background Models for Genomic Position Analysis*. Master's thesis, Justus Liebig University Giessen, 2016. 3.2

[85] M. Menzel, S. Hurka, S. Glasenhardt, and A. Gogol-Döring. NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling. *Bioinformatics*, 37(5):596–602, 09 2020. 1.4, 3.1

[86] M. Menzel, P. Koch, S. Glasenhardt, and A. Gogol-Döring. Enhort: a platform for deep analysis of genomic positions. *PeerJ Computer Science*, 5:e198, June 2019. 3.2

[87] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, Aug. 2007. 1.4

[88] H. Mirzaei, A. Sahebkar, M. R. Jaafari, J. Hadjati, S. H. Javanmard, H. R. Mirzaei, and R. Salehi. PiggyBac as a novel vector in cancer gene therapy: current perspective. *Cancer Gene Therapy*, 23(2):45–47, 2016. 1.2

[89] K. Miyauchi, Y. Kim, O. Latinovic, V. Morozov, and G. B. Melikyan. HIV Enters Cells via Endocytosis and Dynamin-Dependent Fusion with Endosomes. *Cell*, 137(3):433–444, May 2009. 1.2

[90] R. Mohammadinejad, A. Dehshahri, V. Sagar Madamsetty, M. Zahmatkeshan, S. Tavakol, P. Makvandi, D. Khorsandi, A. Pardakhty, M. Ashrafizadeh, E. Ghasemipour Afshar, and A. Zarrabi. In vivo gene delivery mediated by non-viral vectors for cancer therapy. *Journal of Controlled Release*, 325:249–275, Sept. 2020. 1.1

[91] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36(12):1331–1339, Dec. 2004. 3.1.1

[92] R. H. Myers. Huntington's disease genetics. *NeuroRx*, 1(2):255–262, 2004. 1.1

[93] G. J. Narlikar, H.-Y. Fan, and R. E. Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, 2002. 1.3

[94] L. Narlikar and R. Jothi. ChIP-Seq Data Analysis: Identification of Protein–DNA Binding Sites with SISSRs Peak-Finder. In J. Wang, A. C. Tan, and T. Tian, editors, *Next Generation Microarray Bioinformatics*, volume 802, pages 305–322. Humana Press, Totowa, NJ, 2012. 1.4

[95] L. Narlikar and I. Ovcharenko. Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics and Proteomics*, 8(4):215–230, July 2009. 1.3

[96] D. W. Nebert. Transcription factors and cancer: an overview. *Toxicology*, 181-182:131–141, 2002. 1.3

[97] D. E. Newburger and M. L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(Database):D77–D82, Jan. 2009. 1.3, 3.1.1

[98] R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge. Direct measurement of DNA affinity landscapes on a

high-throughput sequencing instrument. *Nature Biotechnology*, 29(7):659–664, July 2011. 1.3

[99] E. P. O'Keefe. Nucleic acid delivery: Lentiviral and retroviral vectors. *Materials and Methods*, 3:174, 2013. 1.2

[100] E. P. Papapetrou and A. Schambach. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. *Molecular Therapy*, 24(4):678–684, Apr. 2016. 4.2

[101] P. J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, Oct. 2009. 1.4

[102] S. Pellenz, M. Phelps, W. Tang, B. T. Hovde, R. B. Sinit, W. Fu, H. Li, E. Chen, and R. J. Monnat. New Human Chromosomal Sites with "Safe Harbor" Potential for Targeted Transgene Insertion. *Human Gene Therapy*, 30(7):814–828, July 2019. 4.2

[103] K. R. Pritchett-Corning and C. P. Landel. Chapter 32 - Genetically Modified Animals. In J. G. Fox, L. C. Anderson, G. M. Otto, K. R. Pritchett-Corning, and M. T. Whary, editors, *Laboratory Animal Medicine (Third Edition)*, American College of Laboratory Animal Medicine, pages 1417–1440. Academic Press, Boston, third edition edition, 2015. 1.1

[104] Z. S. Qin, J. Yu, J. Shen, C. A. Maher, M. Hu, S. Kalyana-Sundaram, J. Yu, and A. M. Chinnaiyan. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC bioinformatics*, 11(1):369, 2010. 1.4

[105] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar. 2010. 4.2

[106] Y. Rajendra, R. B. Peery, and G. C. Barnard. Generation of stable Chinese hamster ovary pools yielding antibody titers of up to 7.6 g/L using the piggyBac transposon system: Biotechnol. Prog. *Biotechnology Progress*, 32(5):1301–1307, Sept. 2016. 3.2.5

[107] B. Ren. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309, Dec. 2000. 1, 1.4

[108] H. S. Rhee and B. F. Pugh. ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy. *Current Protocols in Molecular Biology*, 100(1), Oct. 2012. 3.1.1

[109] S. L. Roth, N. Malani, and F. D. Bushman. Gammaretroviral Integration into Nucleosomal Target DNA In Vivo. *Journal of Virology*, 85(14):7393–7401, July 2011. 1.1, 4.2.1

[110] M. Sadelain, E. P. Papapetrou, and F. D. Bushman. Safe harbours for the integration of new DNA in the human genome. *Nature reviews Cancer*, 12(1):51–58, 2012. 4.2

[111] A. Sandelin. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D–94, Jan. 2004. 1.3, 4.1

[112] T. D. Schneider. Consensus sequence zen. *Applied bioinformatics*, 1(3):111, 2002. 1.3

[113] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990. 1.3

[114] R. A. Scott, L. J. Scott, R. Mägi, L. Marullo, K. J. Gaulton, M. Kaakinen, N. Pervjakova, T. H. Pers, A. D. Johnson, J. D. Eicher, A. U. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N. R. Van Zuydam, A. Mahajan, H. Chen, P. Almgren, B. F. Voight, H. Grallert, M. Müller-Nurasyid, J. S. Ried, N. W. Rayner, N. Robertson, L. C. Karssen, E. M. van Leeuwen, S. M. Willems, C. Fuchsberger, P. Kwan, T. M. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H. A. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Frånberg, R. J. Strawbridge, R. Benediktsson, A. B. Hreidarsson, A. Kong, G. Sigurðsson, N. D. Kerrison, J. Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov, C. Fox, C.-T. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D. J. Couper, J. S. Pankow, N. Grarup, C. T. Have, M. E. Jørgensen, T. Jørgensen, A. Linneberg, M. C. Cornelis, R. M. van Dam, D. J. Hunter, P. Kraft, Q. Sun, S. Edkins, K. R. Owen, J. R. Perry, A. R. Wood, E. Zeggini, J. Tajes-Fernandes, G. R. Abecasis, L. L. Bonnycastle, P. S. Chines, H. M. Stringham, H. A. Koistinen, L. Kinnunen, B. Sennblad, T. W. Mühleisen, M. M. Nöthen, S. Pechlivanis, D. Baldassarre, K. Gertow, S. E. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, R. Wennauer, J. G. Eriksson, S. Männistö, L. Peltonen, E. Tikkanen, G. Charpentier, E. Eury, S. Lobbens, B. Gigante, K. Leander, O. McLeod, E. P. Bottinger, O. Gottesman, D. Ruderfer, M. Blüher, P. Kovacs, A. Tonjes, N. M. Maruthur, C. Scapoli, R. Erbel, K.-H. Jöckel, S. Moe-

bus, U. de Faire, A. Hamsten, M. Stumvoll, P. Deloukas, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, S. Ripatti, V. Salomaa, N. L. Pedersen, B. O. Boehm, R. N. Bergman, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. Hansen, O. Pedersen, I. Barroso, L. Lannfelt, E. Ingelsson, L. Lind, C. M. Lindgren, S. Cauchi, P. Froguel, R. J. Loos, B. Balkau, H. Boeing, P. W. Franks, A. Barricarte Gurrea, D. Palli, Y. T. van der Schouw, D. Altshuler, L. C. Groop, C. Langenberg, N. J. Wareham, E. Sijbrands, C. M. van Duijn, J. C. Florez, J. B. Meigs, E. Boerwinkle, C. Gieger, K. Strauch, A. Metspalu, A. D. Morris, C. N. Palmer, F. B. Hu, U. Thorsteinsdottir, K. Stefansson, J. Dupuis, A. P. Morris, M. Boehnke, M. I. McCarthy, and I. Prokopenko. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*, 66(11):2888–2902, Nov. 2017. 1.1

[115] E. Serrao and A. N. Engelman. Sites of retroviral dna integration: From basic research to clinical applications. *Critical reviews in biochemistry and molecular biology*, 51(1):26–42, 2016. 1.2

[116] M. Setty and C. S. Leslie. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology*, 11(5):e1004271, May 2015. 3.1

[117] W. Shao, J. Shan, M. F. Kearney, X. Wu, F. Maldarelli, J. W. Mellors, B. Luke, J. M. Coffin, and S. H. Hughes. Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology*, 13, July 2016. 1.2

[118] T. Siggers and R. Gordân. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*, 42(4):2099–2111, Feb. 2014. 1.3

[119] D. Simcha, N. D. Price, and D. Geman. The Limits of De Novo DNA Motif Discovery. *PLoS ONE*, 7(11):e47836, Nov. 2012. 1.5, 4.1, 4.1.1, 4.1.4

[120] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, and T. S. Jaakkola. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708, 2001. 1.3

[121] P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, 6:e21856, 2017. 3.1.1

[122] R. K. Slotkin and R. Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285, Apr. 2007. 1.2

[123] G. D. Stormo and Y. Zhao. Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics*, 11(11):751–760, Nov. 2010. 1.3

[124] D. Tang, B. Li, T. Xu, R. Hu, D. Tan, X. Song, P. Jia, and Z. Zhao. VISDB: a manually curated database of viral integration sites in the human genome. *Nucleic Acids Research*, 48(D1):D633–D641, Jan. 2020. 1.2

[125] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012. 1.4

[126] R. Thomas, S. Thomas, A. K. Holloway, and K. S. Pollard. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, page bbw035, May 2016. 1.4

[127] M. Thomas-Chollier, E. Darbo, C. Herrmann, M. Defrance, D. Thieffry, and J. van Helden. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, 7(8):1551–1568, Aug. 2012. 1.4

[128] J. Tipanee, T. VandenDriessche, and M. K. Chuah. Transposons: Moving Forward from Preclinical Studies to Clinical Trials. *Human Gene Therapy*, 28(11):1087–1104, Nov. 2017. 1.2

[129] J. Tran, H. Anastacio, and C. Bardy. Genetic predispositions of Parkinson's disease revealed in patient-derived brain cells. *npj Parkinson's Disease*, 6(1), Dec. 2020. 1.1

[130] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, Sept. 2008. 1.4

[131] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, Apr. 2009. 1.3, 1.3

[132] A. Visel, E. M. Rubin, and L. A. Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, Sept. 2009. 1.3

[133] E. O. Voit. Perspective: Dimensions of the scientific method. *PLoS Computational Biology*, 15(9):e1007279, 2019. 4.2.2

[134] D. Wang, P. W. L. Tai, and G. Gao. Adeno-associated virus vector as a platform for gene therapy delivery. *Nature Reviews Drug Discovery*, 18(5):358–378, May 2019. 1.1

[135] G. P. Wang, A. Ciuffi, J. Leipzig, C. C. Berry, and F. D. Bushman. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research*, 17(8):1186–1194, Aug. 2007. 4.2.1

[136] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, Sept. 2012. 4.1.2

[137] J. Wang, J. Zhuang, S. Iyer, X.-Y. Lin, M. C. Greven, B.-H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil, E. Birney, J.-H. Hung, and Z. Weng. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*, 41(D1):D171–D176, Jan. 2013. 1.3

[138] M. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. Najafabadi, S. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. Walhout, F.-Y. Bouget, G. Ratsch, L. Larrondo, J. Ecker, and T. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, Sept. 2014. 1.3

[139] C. Wolberger. MULTIPROTEIN-DNA COMPLEXES IN TRANSCRIPTIONAL REGULATION. *Annual Review of Biophysics and Biomolecular Structure*, 28(1):29–56, June 1999. 1.3

[140] L. E. Woodard and M. H. Wilson. piggyBac-ing models and new therapeutic strategies. *Trends in Biotechnology*, 33(9):525–533, Sept. 2015. 3.2.5

[141] R. Worsley Hunt, A. Mathelier, L. Del Peso, and W. W. Wasserman. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC genomics*, 15(1):472–472, June 2014. 4.1.2

[142] X. Wu, Y. Li, B. Crise, and S. M. Burgess. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 300(5626):1749–1751, 2003. 3.2.1

[143] K. Yusa, L. Zhou, M. A. Li, A. Bradley, and N. L. Craig. A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences*, 108(4):1531–1536, Jan. 2011. 3.2.5

[144] F. Zambelli, G. Pesole, and G. Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237, Mar. 2013. 3.1

[145] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, and W. Li. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137, 2008. 1.4