FREIE UNIVERSITÄT BERLIN

Robust algorithms for improved reproducible ChIP-seq and ChIP-nexus peak calling

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
Vorgelegt im Jahr 2018, von

Peter Hansen

Fachbereich Mathematik und Informatik

Diese Arbeit wurde betreut von **Prof. Dr. med. Peter N. Robinson, MSc.**

Zweiter Gutachter: Prof. Dr. Miguel Andrade
Datum der Disputation: 11. April 2019

SELBSTÄNDIGKEITSERKLÄRUNG

Ich versichere, dass ich die Dissertation selbständig verfasst, und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit hat keiner anderen Prüfungsbehörde vorgelegen.

Peter Hansen, Berlin, den 28. September 2018

ACKNOWLEDGEMENTS

Ich danke Jochen Hecht, Ilja Demuth, Jonas Ibn-Salem, Daniel Ibrahim, Joachim Klose, Clemens Kühn, Benjamin Mencüç, Nah Zhu, Alexander Krannich, Katrein Sauer, Robin Steinhaus und Gunilla, Matthias Truss, Stefan Mundlos und Peter Robinson.

CONTENTS

1

GENERAL INTRODUCTION

1.1 TRANSCRIPTIONAL GENE REGULATION

Following the sequencing of the human genome and the characterization of its complement of genes [24] and regulatory elements [41], the field of epigenetics has become a central topic in biological research. Cells with identical genomic DNA may have very different characteristics. For instance, all cells of a given human individual originate from a common ancestor, which is the fertilized egg [3]. Because the cells proliferate via cell division, the cells in the fully developed organism contain identical copies of the original genomic DNA sequence. At the same time, different cell types perform very specific tasks in very specific places, e.g. red blood cells or neurons. Epigenetics attempts to explain this phenomenon on the basis of modifications of the DNA and associated proteins that are transmissible to daughter cells and do not involve a change of the nucleotide sequence.

Cell types within a given species differ with regard to gene expression, i.e. which gene products – most commonly proteins – are actively produced. A few thousand genes referred to as housekeeping genes are expressed in all cells of a given organism because they are required to maintain basic cellular function [33]. Beyond that, there are genes that are specifically expressed only within particular cell types as well as genes that are expressed only upon certain environmental stimuli or during specific developmental stages [43]. Expression of genes in the wrong place or at the wrong time may have severe consequences indicating complex and fine-tuned underlying mechanisms of regulation.

Gene expression is a complicated and still only partially understood topic whose details are beyond the scope of this thesis. On a very abstract level, in order to become active first of all genes have to be transcribed. This process is catalyzed by a multiprotein complex referred to as RNA polymerase that first binds to promoter regions upstream of genes and then moves in downstream direction thereby assembling a complementary RNA strand using the genomic DNA as a template. After some post-transcriptional modifications, the resulting transcripts may be functional molecules, such as regulatory small interfering RNAs or nuclear RNAs. In contrast to that, messenger RNAs (mRNAs) serve as templates that are exported to the cytoplasm and further translated into amino acid sequences that subsequently form proteins. Regulation may occur at various stages of expression, for instance, the levels of mRNAs depend on the rates of transcription, post-transcriptional processing, transport from the nucleus to the cytoplasm, translation, decay as well as active degradation [112].

At the transcriptional level, the expression of genes is primarily limited by the accessibility of the individual genes and their associated regulatory DNA elements such as promoters or enhancers, whereby accessibility is linked to the state of packaging of the DNA. The nucleus of each diploid human cell contains about 6 billion base pairs that, if they were arranged in a row, would make up 2 meters of DNA [6]. On the other hand, the diameter of the nucleus is only 10 to 20 μm. In order to fit into this tiny space, the DNA needs to be compacted. At the lowest level, this involves the formation of protein-DNA complexes referred to as nucleosomes each consisting of eight histone proteins around which the DNA is wrapped 1.65 times. Along the DNA, the nucleosomes occur at regular and short distances of about 165 base pairs, which corresponds to the length of the DNA wrapped around one nucleosome plus the average length of intermediate linker DNA. Beyond that, there are further compaction levels, and the highest degree of compaction can be observed within the metaphase of mitosis.

The compacted DNA together with all associated proteins including the histone molecules is referred to as chromatin, which can be further distinguished into the more tightly packed heterochromatin and the more loosely packed euchromatin. Genes and regulatory regions within tightly packed regions are not accessible for RNA polymerases as well as associated regulatory proteins and, therefore, cannot be transcribed. However, the structure of chromatin can be modulated through a number of different modifications of the histone proteins within the nucleosomes involving the recruitment of remodeling enzymes and repositioning of nucleosomes [11].

In recent years, increasingly more importance has been attached to chromatin structure. For instance, regions with increased frequencies of pairwise DNA contacts are regarded as functionally isolated topological associating domains comprising genes and regulatory DNA elements that may share functional roles [117]. Furthermore, multiple distal enhancers looped to the promoter of one gene are considered as active chromatin hubs associated with enhanced transcription [85].

At the sequence level, another important mechanism for the regulation of transcription is methylation of cytosines at CG sites that can be propagated through cell division due to the symmetry of the complementary strands [101]. DNA methylation is mainly associated with silencing of genes, whereby promoter or enhancer sequences are methylated making them inaccessible for regulatory proteins such as transcription factors. This mechanism enables that for the two alleles of a given gene – one from the maternal and one from the paternal chromosome – the one is expressed but the other is not. A popular example is the insulin like growth factor 2, for which only the paternal allele is expressed [27]. Moreover, silencing of genes through methylation occurs in a coordinated fashion and on a genome-wide scale, whereby specific genes required for particular states are jointly deactivated or activated.

Transcription factors are the key players in the regulation of transcription. They are known for the formation of the transcription pre-initiation complex and the recruitment of RNA polymerase [116] as well as for setting critical switches in response to external stimuli especially during development.

They co-operate on chromatin in order to initiate or modulate transcription directly or indirectly by inducing changes in the structure of chromatin or methylation status of DNA. The function of transcription factors can be context specific, for instance, a given transcription factor may regulate the transcription of different genes in different cell types [65].

Typical transcription factors bind via sequence specific domains to short nucleotide patterns of 8 to 12 base pairs (bp), whereby the preference can be as much as 1000 times higher as compared to other sequences. Regulatory DNA regions such as promoters or enhancers contain variations of those nucleotide patterns referred to as motifs. In addition to the DNA binding domain, transcription factors have domains that allow for interaction with ligands or other proteins. Once bound to the DNA, they can display enzymatic activities such as the catalysis of conformation changes in nucleosomes [65].

1.2 NEXT-GENERATION SEQUENCING

Sequencing of DNA is one of the key technologies in the life sciences. Over the last decade, next-generation sequencing (NGS) has been developed, with ongoing improvements regarding speed, accuracy and costs. Compared to the conventional Sanger sequencing, NGS enables sequencing of large amounts of DNA at affordable prices and has been used in many areas of biological research such as genetics, microbiology and oncology [13].

There is a variety of sequencing technologies being used that all have their individual strengths and weaknesses, whereby the most frequently used platform is Illumina producing massive quantities of relatively short *reads* of DNA sequences from 30 up to 300 bp [70]. For Illumina and other short-read platforms, DNA is typically fragmented into small pieces of a few hundred base pairs and only the outermost ends are sequenced – either one (single-end) or both ends (paired-end) of each fragment. In addition, there are platforms such as Pacific Biosciences that produce comparatively small numbers of long reads up to 40 kbp in length. Long reads allow for the investigation of scientific questions that are difficult to address using short reads, for instance, long reads are better suited for the detection of larger structural variants such as inversions. However, these technologies are less established and, besides the lower throughput, they still have the disadvantage of higher error rates [70].

Over the years, numerous different applications of short-read NGS have been developed, whose primary purpose is not always to decode unknown DNA sequences but also to verify or even quantify the presence of specific RNA or DNA molecules in order to draw conclusions about processes that take place within cells. The most prominent example is RNA-seq [78] that can be used to analyze the set of all transcripts present in cells under certain conditions [119]. For this purpose, RNA is extracted, fragmented and converted into cDNA. The ends of the fragments are sequenced and mapped to the corresponding reference transcripts on the basis of sequence identity. The read counts can then be used to quantify the level of transcription. Alter-

natively, the reads can be mapped to the corresponding reference genome allowing for the identification of novel transcripts arising from alternative splicing. In comparison to the conventional hybridization-based microarray technology, RNA-seq is more generic and therefore allows a wider range of scientific questions to be addressed. However, the complexity and the sheer amount of data requires for expertise in computational biology as well as for substantial computing and storage resources.

Besides RNA-seq, there is a number of applications that can be used to investigate the regulation of gene expression with a particular focus on epigenetics. For instance, the sequencing of genomic sodium bisulfite treated DNA (BS-seq) can be used to detect methylated cytosines using the fact that, at the sequencing level, only those are retained whereas unmethylated cytosines are converted to thymines (T) that will form mismatches with their methylated counterparts (C) in the sequences of the untreated DNA [123]. Another example is Hi-C [115] that can be used to measure the genome-wide interactions between distal genomic sequences and thereby to characterize compartments of increased pairwise interaction frequency referred to as topological associating domains (TADs) that often contain functional entities consisting of genes and regulatory sequences such as enhancers [117]. Or, to give a final example, the accessibility of chromatin can be evaluated using ATAC-seq [19] that requires only low amounts of starting material and is relatively simple to do. In order to address meaningful biological questions, often more than one NGS application is applied to the same tissue or population of cells. For example, to investigate the effect of methylation on gene expression RNA-seq and BS-seq can be performed simultaneously [83].

1.3 CHIP-SEQ, CHIP-EXO AND CHIP-NEXUS

This thesis is about the primary analysis of ChIP-seq [60, 92] and ChIP-nexus [47] data. For this reason, these methodologies are described on a more detailed level in the following paragraphs. The classical **Ch**romatin **I**mmuno**p**recipitation (ChIP) procedure (Figure 1) can be used to verify interactions between specific genomic sites and a target protein in vivo. For this purpose, all proteins bound to the DNA of a population of cells are first cross-linked using formaldehyde. Subsequently, chromatin is extracted, sheared into small fragments of 100 to 500 bp and subjected to immunoprecipitation using a specific antibody directed against the target protein, thereby co-enriching DNA fragments that are bound to the protein of interest. Finally, the cross-linking is reversed, and polymerase chain reaction (PCR) with appropriate primers can be used to verify specific protein-DNA interactions. This procedure is referred to as ChIP quantitative PCR. In contrast, for ChIP-seq, the co-precipitated DNA fragments are sequenced, and the reads are mapped to the corresponding reference genome, which allows the genome-wide identification of interacting sites without prior knowledge about specific binding sites.

In the course of the preparation of sequencing libraries, the DNA fragments are subjected to PCR amplification in order to obtain a sufficient

overall amount of DNA required for technical reasons. The common read length used for ChIP-seq is between 20 and 100 bp, which is smaller than the average length of the fragments (100-500 bp). Therefore, the reads represent only the ends of fragments, whereby the 5′ end positions of reads represent breakpoints introduced by the shearing of DNA. Each fragment within the library can be assumed to be sequenced with equal probability. Furthermore, if single-end sequencing is performed, which is usually done for ChIP-seq, both ends of each given fragment can be assumed to be sequenced with equal probability resulting in an approximately even overall number of reads from both strands.

The reads are mapped to the associated reference genome on the basis of sequence identity. At binding sites, the mapped reads form bimodal clusters termed ChIP-seq peaks, whereby half of the reads map to the forward strand directly before the binding site and the other half to the reverse strand directly after the binding site (Figure 2A). The more precise distribution of breakpoints around binding sites depends on the distribution of fragment lengths in the sequencing library. Assuming that shearing is independent of the interactions between target protein and DNA, the breakpoints may occur with equal probability at each position within a region of length two times the average fragment length around given binding sites. However, due to the bimodal strand specific distribution of ChIP-seq reads, binding sites can be predicted with great accuracy, especially for strong peaks.

There are some details of the ChIP-seq protocol that must be taken into account for data analysis. One aspect affects almost all NGS applications and is due to the PCR amplification that is performed prior to sequencing. With increasing sequencing depth there will be more and more reads with identical sequence coming from PCR duplicated fragments. The smaller the number of fragments used as starting material, the lower the complexity of the library and the sooner this will happen.

Another critical point is the evenness of PCR amplification. Ideally, each fragment is amplified with the same efficiency, which is not entirely true, because the efficiency i.a. depends on the base composition of the individual fragments, especially on the GC content [1]. Consequently, this step potentially distorts the composition of fragments. On the other hand, breakpoints may occur by chance in different cells at the same position, which results in reads with identical sequence that cannot be distinguished from reads originating from PCR duplicated fragments. It is common practice to treat both categories in the same way by keeping only one read for each given sequence. This approach is based on the assumption that identical reads are much less likely to arise from identical breakpoints than from PCR-overamplification, i.e. that the loss of informative reads is acceptable.

Aside from the PCR amplification step, the mapping of reads involves some difficulties. Repetitive regions are difficult to evaluate when using ChIP-seq, because ambiguously mapped reads are not suitable for the determination of binding sites. Furthermore, repetitive regions are prone to mapping artifacts [21], whereby the mapped reads form large clusters that are sometimes difficult to distinguish from ChIP-seq peaks (Figure 2B). In general, the longer the reads the larger the proportion of the genome to

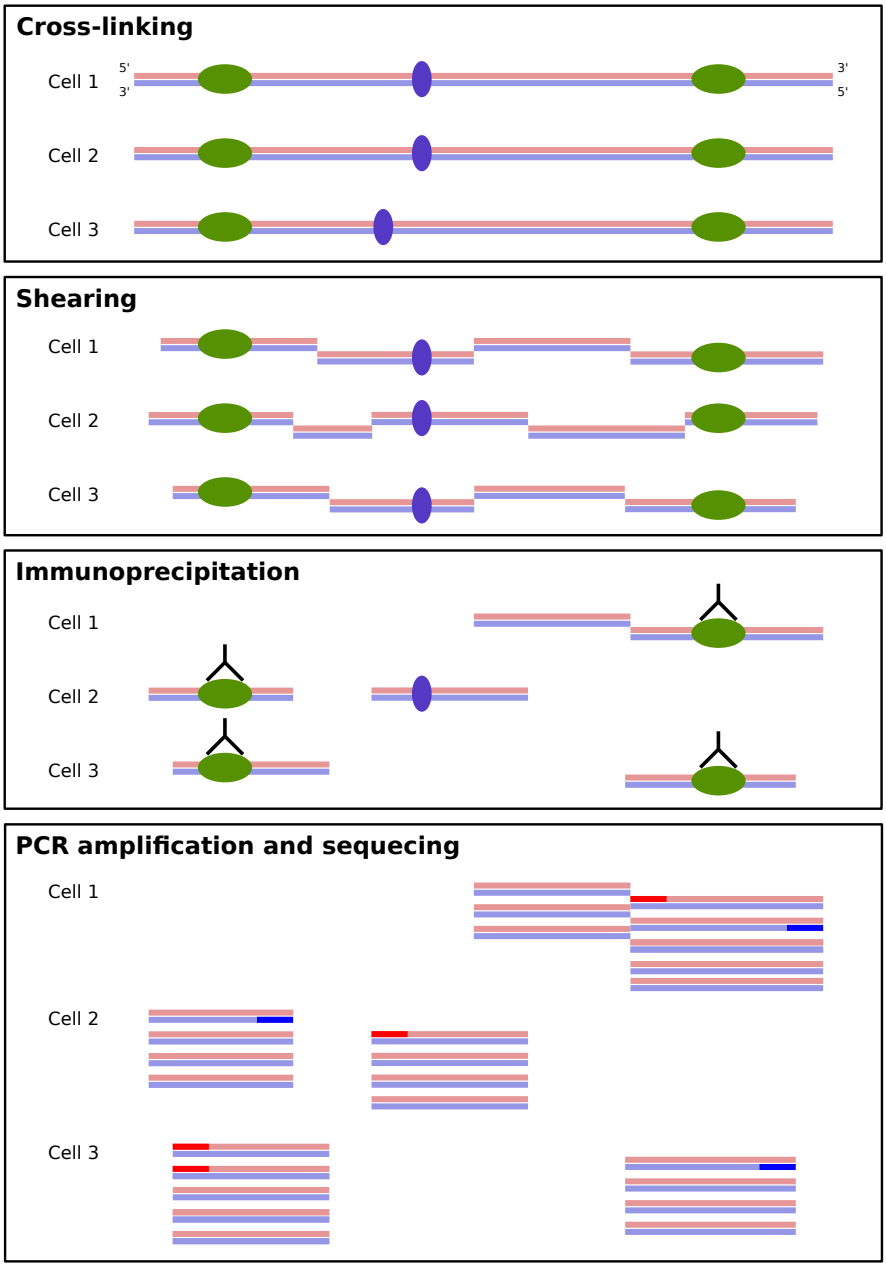


Figure 1: **Schematic representation of the ChIP-seq protocol:** Double stranded DNA is depicted in red and blue. The chromatin of a population of cells is cross-linked in vivo, the DNA is sheared into fragments of 100 to 500 bp, which are subjected to immunoprecipitation using an antibody directed against the protein of interest (green). Depending on the efficiency of the immunoprecipitation, fragments that were bound by any protein or other proteins (purple) may remain and constitute the background. During the preparation of the sequencing library, the co-precipitated DNA fragments are amplified using PCR. Each fragment within the library can be assumed to be sequenced with equal probability (stronger colored). Furthermore, both ends of a given fragment can be assumed to be sequenced with equal probability resulting in an approximately even number of reads from both strands. Overrepresented PCR-duplicated fragments can lead to redundant reads with identical sequence.

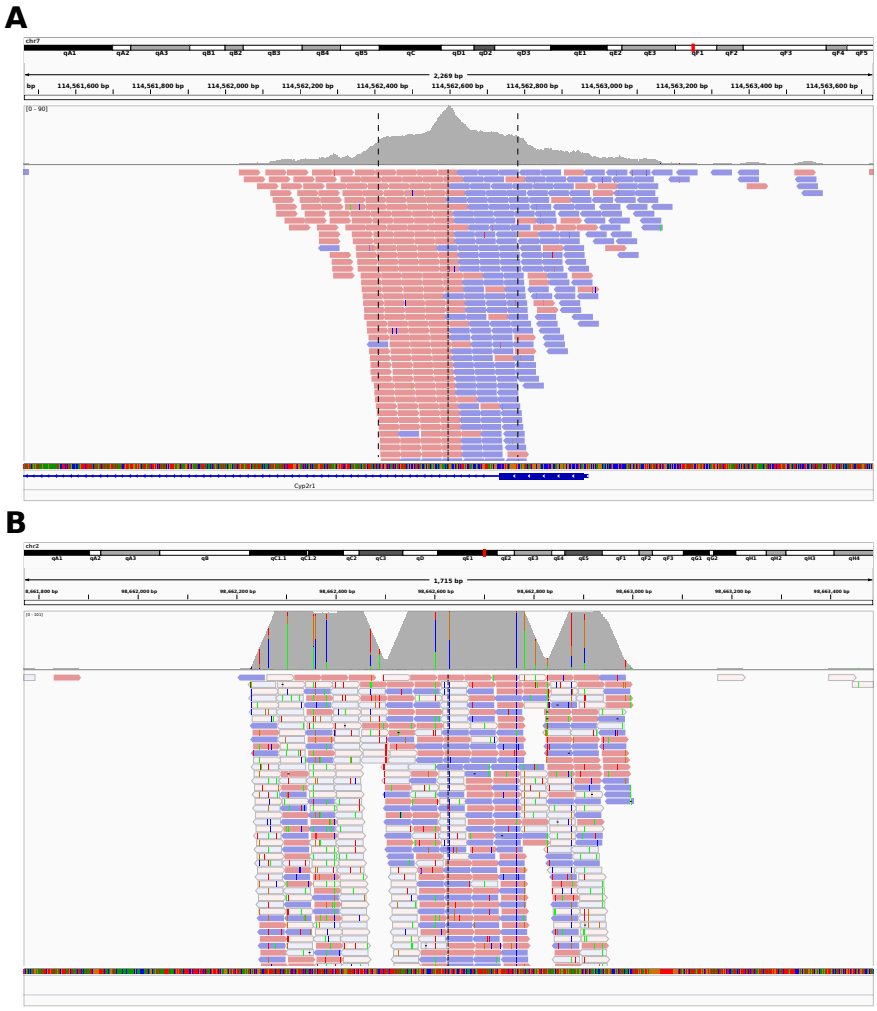


Figure 2: **Mapping of ChIP-seq reads:** ChIP-seq reads are mapped to the associated reference genome on the basis of sequence identity. The mappings can be visualized using genome browsers such as the Integrated Genomics Viewer (IGV) [93]. (**A**) The displayed region shows a typical ChIP-seq peak for E2F6 in mouse embryonic stem cells. At binding sites, the mapped reads form bimodal clusters, whereby half of the reads map to the forward strand directly before the binding site (red) and the other half to the reverse strand directly after the binding site (blue). Due to the strand specific distribution, the binding sites can be predicted with great accuracy (chain line in the center), especially for strong peaks. The precise shape of peaks depends on the overall distribution of fragment lengths within the library. The adjusted fragmentation size for this experiment was 185 bp (flanking dashed lines). The isolated mapped reads in the surrounding region correspond to background fragments. (**B**) Repetitive regions are prone to mapping artifacts, whereby the mapped reads form large clusters that can be difficult to distinguish from ChIP-seq peaks.

which reads can be unambiguously mapped.

Another issue with ChIP-seq is the uneven accessibility of DNA [38] due to varying degrees of compaction that also affects shearing of DNA [109]. As a consequence, the background distribution of mapped reads is not uniform across the genome.

GC and repeat content as well as DNA compaction are local potentially overlapping characteristics of DNA that affect the depth of mapped reads [38]. Therefore, control experiments are usually performed in order to generate data that can be used to correct for these biases. Most commonly, the control data is derived from input chromatin, i.e. chromatin from the same batch that has been cross-linked and sonicated but not immunoprecipitated. Alternatively, an IgG control antibody that binds only to non-nuclear proteins is used place of the specific antibody.

ChIP-seq is a valuable and well-established method for the genome-wide identification of protein-DNA interactions. Since its introduction, thousands of experiments have been performed in individual laboratories as well as in large scale projects such as the ENCODE project [41]. The results contributed significantly to the creation of epigenomic maps for different cell types that now serve as reference for the scientific community.

The major shortcoming of ChIP-seq is the limited accuracy of binding site prediction, which depends on the distribution of the lengths of fragments in the sequencing library and on the strength of the individual peaks. For strong peaks, the binding site can be precisely predicted, but for weak peaks the data simply does not provide enough information to narrow down the exact binding location. Another related question is resolution, i.e. how well two adjacent peaks can be distinguished from one another. If the distance between two peaks becomes smaller than one average fragment length, there will be reads that cannot be unambiguously assigned to one or the other peak.

On the other hand, accuracy is important for many biological questions such as the search for sequence motifs associated with a given transcription factor. The more precisely the peaks are predicted, the more often motif sequences will occur in the sequences beneath the peaks that are subsequently used for de novo motif analysis. Moreover, DNA binding proteins commonly form complexes such as the polycomb repressive complex [75, 22] consisting of multiple proteins that coordinately interact with the DNA. A high resolution is essential for drawing conclusions about the architecture of such protein-DNA complexes. Finally, resolution also matters for proteins that tend to produce broader ChIP-seq peaks. For example, the peaks for some histone modifications can be seen as series of adjacent peaks. In such cases, the picture is inverted, and the task is to identify the gaps between peaks, which can correspond, for instance, to nucleosome depleted regions.

The ChIP-exo methodology is a further development of ChIP-seq that can be used to predict binding sites with much greater accuracy [39, 102, 106] and therefore allows for a characterization of cooperative binding of proteins to DNA to a level of detail that was not feasible before. For instance, the profile-based analysis of ChIP-exo data [104] was used to discriminate between

direct and indirect binding of the glucocorticoid receptor. For ChIP-exo, the increased accuracy is achieved by 5′-3′ (λ) exonuclease digest of the 5′ end segments of fragments up to about 5-6 base pairs to the location at which target protein and DNA were cross-linked [90]. In theory, fragments that were not cross-linked are digested completely, which contributes to a reduction of background noise [90, 91]. The remaining intact 3′ end segments of fragments are sequenced and mapped to the reference genome, which results in very sharp peaks at cross-linked locations.

Besides the advantage of a massively increased accuracy, a number of shortcomings of the ChIP-exo method have also been reported [47] including high amounts of input DNA required to avoid PCR-overamplification artifacts. Just as for ChIP-seq, it cannot be determined whether two identical reads originate from different fragments or not. On the other hand, an increased number of the sequenced fragment ends must end up at the same genomic position, given the extremely narrowed range around the cross-linked interacting proteins (5-6 bp). Assuming the 5′-3′ (λ) exonuclease perfectly extended to the cross-linked location at the exact same position for each protein-DNA binding event, then each signal would consist of duplicated reads only. In this situation, it would be nonsensical to remove all duplicated reads because this would eliminate the actual signal.

ChIP-nexus (ChIP-**N**ucleotide resolution through **EX**onuclease, **U**nique barcode and **S**ingle ligation) is a ChIP-exo protocol with improved efficiency. In addition, random barcodes are introduced that allow to distinguish PCR duplicated reads from identical reads emerging from different fragments. Due to the protocol, PCR duplicated reads originating from the same DNA fragment have the same random barcode. The random barcodes require appropriate preprocessing. At the time ChIP-nexus was introduced [47], this was done using a number of scripts invoking available tools. Furthermore, the peak caller MACS2 was used that was originally developed for ChIP-seq data, which has different characteristics as compared to ChIP-exo and ChIP-nexus data.

For ChIP-seq, the distribution of fragment lengths affects the the shape of peaks, especially the width. This is no longer valid for ChIP-exo and ChIP-nexus because the 5′ end segments of fragments are digested up to the position at which the 5′-3′ (λ) exonuclease encounters an obstacle and falls off. At binding positions, the regions between exonuclease stop positions on the forward and reverse strand are ideally free of signal because they are occupied by the investigated proteins or by co-regulator proteins that can also be part of the same complex. These regions appear to be protected from 5′-3′ (λ) exonuclease digestion and, therefore, are here referred to as protected region.

1.4 OUTLINE

This thesis deals with the primary analysis of ChIP-seq and ChIP-nexus data. The work on these topics lead to three publications that are presented in

chronological order each in a separate chapter.

The first part (Chapter 2) provides an application example of ChIP-seq that was used in this case in order to elucidate the pathomechanism underlying the phenotype of a patient with severe hand and foot malformations carrying a mutation of the gene encoding for the transcription factor HOXD13 [54]. This work with a focus on developmental biology was done in the preliminary phase of this thesis, and the computational contributions were largely limited to the use of existing software in accordance with the standards for ChIP-seq data analysis as defined by the ENCODE project consortium [66]. Nevertheless, it is presented here because it provides insights to the practical benefit of ChIP-seq. Furthermore, the main ideas for this thesis emerged from practical application of the recommended standard software, which is why it is introduced in detail.

In the next part (Chapter 3), the ChIP-seq peak caller Q [46] is introduced that addresses shortcomings of the recommended software identified in the course of practical applications. Improvements regarding efficiency and reproducibility were verified within the framework of the ENCODE standards using 38 publicly available datasets. Furthermore, Q was used to characterize a signature of RNA polymerase II (RNAPII) and histone modification H3K4me3 peaks that is consistent with the concept of paused open promoters.

In the final part (Chapter 4), the first bespoke software package for the analysis of ChIP-nexus data is presented [45]. The ChIP-seq caller Q was extended by additional modules that are required for the analysis of ChIP-nexus data. The software makes use of the random barcodes introduced with ChIP-nexus and was released under the name Q-nexus. Finally, Q-nexus was compared to two other peak callers with respect to reproducibility of peak calling.

2

APPLICATION EXAMPLE OF CHIP-SEQ

2.1 INTRODUCTION

In this chapter an exemplary application of ChIP-seq [54] is presented in order to provide insight into the practical benefit of ChIP-seq and to introduce concepts and software for data analysis relevant for the problems addressed within the scope of this thesis. The analysis of the ChIP-seq experiments were performed in the preliminary phases of this thesis project and contributed great to the development of ideas. Several established methods were used for the analysis that were subsequently also used benchmark the methods developed in this thesis, which is why they are explained in detail.

Ibrahim et al. [54] investigated the pathomechanism underlying a novel heterozygous missense mutation in a gene encoding for the transcription factor HOXD13. Affected individuals show severe hand and foot malformations (Figure 3). On the protein level, this mutation leads to a glutamine to lysine substitution at position 317 within the DNA binding domain (Q317K). For bicoid-type homeodomain proteins including PITX1, a lysine at position 317 is a typical feature, and misexpression of PITX1 in the developing chick wing bud had been shown earlier to lead to a partial fore-to-hindlimb transformation accompanied by the formation of hindlimb characteristics [74, 107] reminiscent of the phenotype that was observed for the patient carrying the Q317K mutation. Based on this, Ibrahim et al. postulated a pathomechanism according to which the mutation of the binding domain alters the binding specificity of HOXD13 resulting in a regulation of an abnormal set of target genes that are naturally regulated by PITX1. This hypothesis was verified using a variety of molecular biological techniques including the NGS applications ChIP-seq and RNA-seq.

In principle, ChIP-seq is applicable to every protein for which a specific antibody is available. However, in order to apply ChIP-seq to the Q317K mutation, mainly two difficulties had to be overcome. First, for heterozygous mutations wild type and mutant proteins both are present in the cell, but the available antibodies do not distinguish between these two forms. Second, it is virtually impossible to obtain sufficient amounts of tissue for the relevant anatomical site and developmental time point. Because of this, Ibrahim et al. [54] developed a cell culture-based retroviral overexpression system that allows for specific targeting of the mutated factor and yields sufficient starting material required for ChIP-seq (Figure 4). Roughly speaking, the coding sequence of the gene of interest is fused with a triple FLAG tag sequence and inserted into a vector, which is then transferred into an RCASBP virus. Upon infection of a cell, the vector construct will be integrated into the genome

Figure 3: **Phenotype of Q317K missense mutation.** Photographs and associated radiographs of hand (upper panels) and foot (lower panels) of a patient carrying a Q317K mutation of the transcription factor HOXD13. The mutation leads to severe malformations with missing and shortened digits. *This figure was originally published in Ibrahim et al., 2013 [54].*

and thus passed on to all daughter cells that will express the triple FLAG tagged protein of interest. After culturing and harvesting of the infected cells, ChIP-seq is performed using a standard anti-FLAG antibody. Due to the triple FLAG tag, the ChIP-enrichment can be performed with increased efficiency. Furthermore, the protein of interest is moderately overexpressed. Therefore, binding sites of the analyzed factor can be detected with high sensitivity. Ibrahim et al. used chicken mesenchymal limb bud cells, but in principle the system is applicable to all culturable cell types.

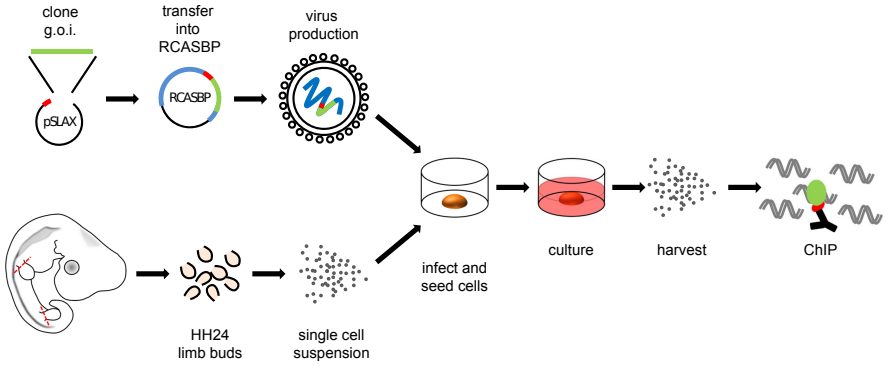


Figure 4: **Retroviral expression system:** The sequence of a gene of interest (green) is inserted into a vector in frame with a triple FLAG tag (red). The vector is then transferred into an RCASBP virus. Embryonic chick mesenchymal cells isolated from limb buds are infected with the virus and subsequently cultured. After harvesting of cells, ChIP-seq with an anti-FLAG antibody is performed. Due to moderate overexpression of triple FLAG tagged proteins, increased efficiency of the ChIP-enrichment is achieved. *This figure was originally published in Ibrahim et al., 2013 [54].*

The retroviral expression system was applied to the wild type proteins Hoxd13^{wt} and PITX1^{wt} as well as to the mutant proteins Hoxd13^{Q317K} and Hoxd13^{Q317R}, and the harvested cells were used to perform ChIP-seq and RNA-seq. The next sections will show how the data was processed. The focus is on the main steps of the primary analysis of ChIP-seq data. Special emphasis is put on applied methods that initiated further developments or were subsequently used for validation purposes. The detailed descriptions are intended to serve as a reference for the interpretation of the results presented in this thesis (Chapters 3 and 4). If appropriate, the individual analysis results of Ibrahim et al. [54] will be presented briefly along with the corresponding methods and jointly discussed at the end of this chapter.

2.2 CHIP-SEQ DATA ANALYSIS

2.2.1 *ENCODE guidelines and best practices*

Large scale projects such as the ENCODE project [41] necessarily use standards for applied methods and experimental setups as well as for the accessibility and evaluation of experimental results [23, 16]. After the raw and processed data is released it serves as a reference for the scientific community.

Compliance with such standards opens up the possibility of comparing the quality of one's own results to those of many others. Furthermore, if experimental data comply with ENCODE standards, this can be taken as a strong argument for the quality of the results.

At the end of the first production phase of the ENCODE project [34, 31], the final results were released accompanied by a series of publications. One of these publications was entirely dedicated to the working standards and guidelines for ChIP experiments used throughout the project [66], whereby particular focus was put on data analysis. These guidelines resulted from experiences gained in the course of the project for which more than thousand ChIP-seq experiments for a variety factors and cell types were performed in different laboratories.

In accordance with these standards, Ibrahim et al. [54] performed each ChIP-seq experiment in two biological replicates. Additionally, associated input control experiments were performed. Furthermore, all sequencing libraries were sequenced to an appropriate depth exceeding the defined specification in all cases. The raw sequencing data was submitted to the Short Read Archive [68] and the primary results of the data analysis were deposited at the Gene Expression Omnibus database [32]. Also in terms of data analysis, Ibrahim et al. adhered to the standards. The efficiency of ChIP-enrichment was assessed using the cross-correlation method of the SPP package (Section 2.2.4), and peak calling was performed with special attention to reproducibility using the peak caller MACS2 (Section 2.2.5) within the framework of the Irreproducibility Discovery Rate (IDR) procedure (Section 2.2.6).

2.2.2 *Preprocessing and mapping of ChIP-seq reads*

Short reads are typically given in FASTQ format. Each record consists of three data fields containing a unique identifier, the nucleotide sequence and the Phred quality scores [35] for the individual read positions. A frequently used software for the first inspection of reads is FastQC [4] which generates a report providing graphical presentations of summary statistics including per base quality statistics (Figure 5). In the first step of the analysis for Ibrahim et al., the average Phred score was calculated for each read and those with an average score below 28 were discarded. For fragments smaller than the uniform read length, the corresponding reads contain adapter sequence at the 3' ends that has to be clipped off. This step was skipped for the data analysis of Ibrahim et al. because the read length was only 36 bp.

In the next step of ChIP-seq data analysis, the reads are mapped to the corresponding reference genome using Burrows-Wheeler aligners such as bowtie [67] or BWA [72] also referred to as short read mappers. Such mappers typically produce output files in SAM/BAM format [71], which consist of a comparatively small header section containing primarily information about chromosome lengths and a large section for the individual alignments of reads. For ChIP-seq and many other NGS applications, ambiguously mapped

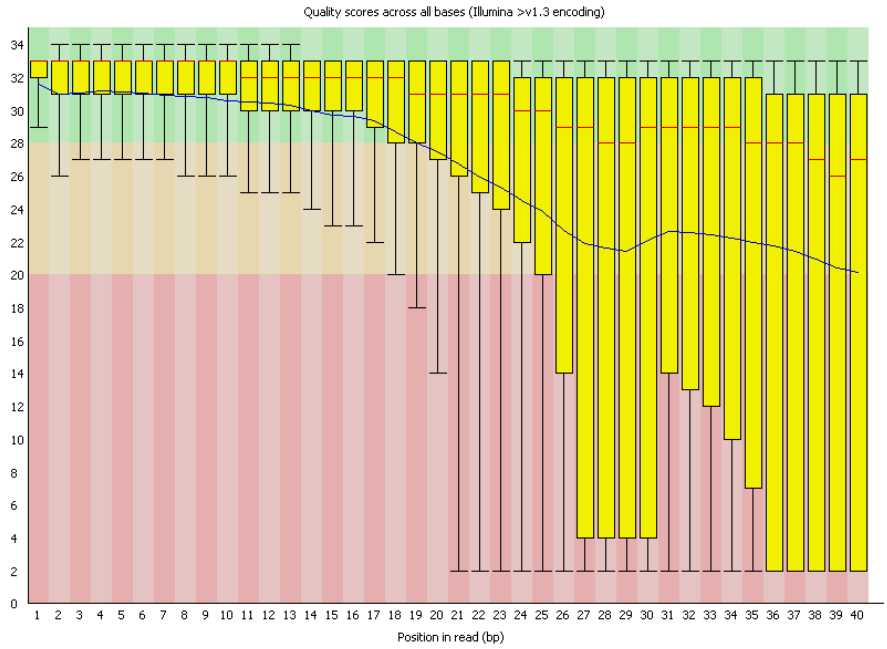


Figure 5: **Per base quality plot of FastQC:** Example for a FASTQ dataset of low overall quality. For each read position summary statistics are calculated including median (red) and mean (blue) Phred quality scores. *This figure was taken from the documentation of the FastQC software [4].*

reads complicate the analysis, because it cannot be determined from which region of the genome they originate from. Therefore, such reads are typically removed. For the analysis presented here, the quality filtered reads were mapped to the chicken reference genome (genome build *galGal3*) using BWA with at most two allowed mismatches per read, and ambiguously mapped reads were discarded using the BWA specific SAM tag XT:AU.

Library complexity is an import issue in ChIP-seq and other NGS applications (Section 1.3), which is why the report of FastQC also includes a plot for sequence duplication levels (Figure 6). PCR duplicated reads are not informative and artificially influence read depth. Therefore, all duplicated sequences are typically removed prior downstream analysis which has the side effect that identical reads that do not originate from PCR overamplification are also removed. For the analysis of Ibrahim et al., the removal of duplicates was performed using samtools [71] with the `rmdup` sub command. The average number of unambiguously mapped deduplicated reads is $27,193,329$ which is in compliance with the ENCODE standards (Section 2.2.1).

2.2.3 *Fragment length estimation*

The predominant length of fragments in given ChIP-seq libraries is a crucial parameter for data analysis and visualization. At the same time, it may vary from one experiment to another because it depends on the settings used for sonication. Therefore, the predominant fragment length has to be determined for each experiment individually, and numerous algorithms have

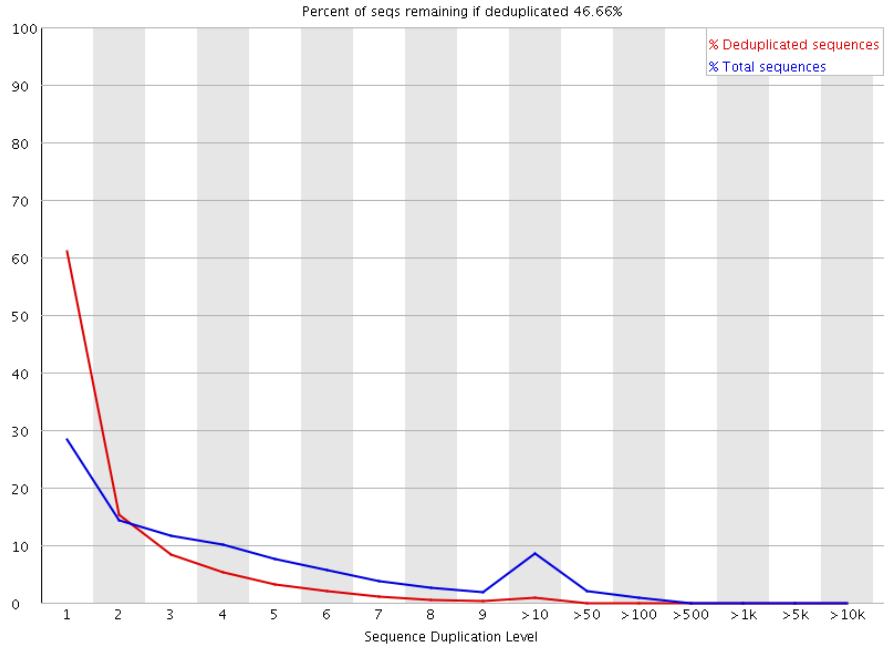


Figure 6: **Duplication level plot of FastQC:** The blue line shows the proportion of reads within a given duplication level amongst all reads (before deduplication), and the red line shows the proportion of reads with distinct sequences within a given duplication level amongst all reads with distinct sequences (after deduplication). The percentage of remaining reads after deduplication is calculated as the ratio of the total numbers reads before and after deduplication. *This figure was taken from the documentation of the FastQC software [4].*

been developed that can be used to estimate this parameter from mapped ChIP-seq reads [114, 125, 61, 18, 69, 48, 49, 88]. An established procedure that was also used throughout the ENCODE project [41, 66] is the cross-correlation method [63]. A special feature of this method is that it is applied genome-wide and has no assumptions about *true* binding sites of the target protein which are subject of investigation.

For the cross-correlation method, each chromosome of the genome is divided up into equally sized bins. Subsequently, the 5′ end positions of mapped reads within each bin are counted. This is done separately for the forward and the reverse strand, which results in two count vectors $n_c^+(x)$ and $n_c^-(x)$, where x denotes the position, s the strand and c the chromosome. Finally, the two count vectors are shifted against each other, and for each shift δ the strand cross-correlation $X(\delta)$ is calculated as follows[1]:

$$X(\delta) = \sum_{c \in C} \frac{N_c}{N} \cdot P[n_c^+(x + \frac{\delta}{2}), n_c^-(x - \frac{\delta}{2})] \tag{1}$$

whereby $P[a, b]$ is the Pearson linear correlation coefficient for the vectors a and b, C is the set of all chromosomes, N_c the total number of 5′ end positions for a given chromosome and N the total number of 5′ end positions for the entire genome.

For a typical ChIP-seq experiment, the cross-correlation curve has a global maximum usually between 90 and 250 bp (Figure 7A), and the corresponding shift size is interpreted as the predominant length of fragments [66] here

[1] This formula was transferred one-to-one from the original publication [63].

denoted by ℓ. For greater strand shifts, the curve gradually decreases and finally converges to a certain correlation coefficient that depends on the noise level of a given ChIP-seq experiment. To gain intuition, due to the strand specific distribution of the detected breakpoints at binding sites (Figure 2) clusters of bins with increased counts for $n_c^+(x)$ and $n_c^-(x)$ are opposite to each other at a shift of 0, and mainly the bins between binding sites contribute to the correlation. With increasing strand shifts, the clusters on the forward and reverse strand start to overlap and increasingly contribute to the correlation until they start to pass which is when the correlation starts to decrease.

If the cross-correlation method is applied to data of input control experiments, i.e. no ChIP-enrichment was performed, there is no distinct peak at a shift size of one predominant fragment length (Figure 7B). Instead, there is a sharp peak at a shift of one read length rl. This peak, referred to as *phantom peak* [66], is also identifiable in curves derived from experiments with ChIP-enrichment although less pronounced (Figure 7A). The phantom peak has been associated with mapping artifacts arising from genomic regions that are difficult to map [88]. Blacklisting of such regions as well as the removal of duplicated reads has an effect on the height of the phantom peak [110, 21]. Visual inspection of the read alignments in regions that most likely contribute to the phantom peak revealed pile-ups of mapped reads arranged in a way that the 5' end positions of reads mapped to the forward and reverse strands occur at a distance of one read length (Figure A3).

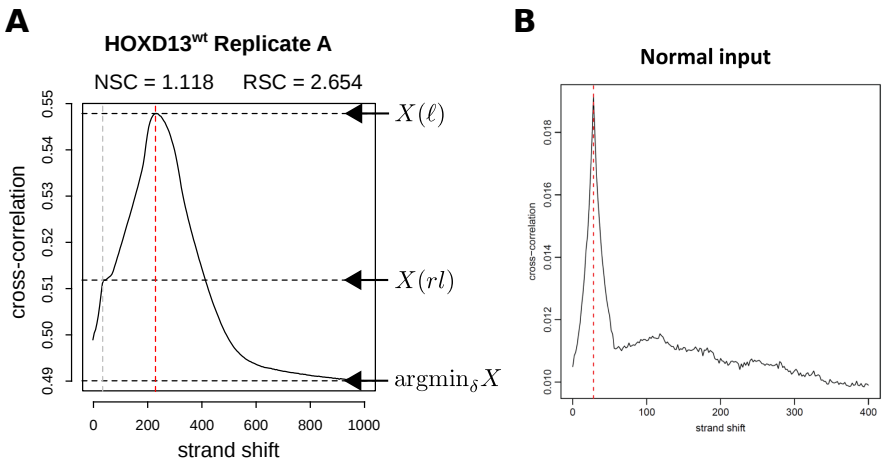


Figure 7: **Cross-correlation curves for a HoxD13 replicate and input control:** The cross-correlation method can be used to estimate the fragment length and to assess the noise level of given ChIP-seq experiments. (**A**) With increasing strand shifts, the correlation rises until a maximum is reached (red dashed line). The corresponding shift size is interpreted as the predominant fragment length (ℓ). For shift sizes greater than ℓ, the correlation falls away and finally converges to a certain value that can be interpreted as the noise level ($\text{argmin}_\delta X$). At a shift size of one read length (rl, gray dashed line), there is a unevenness that corresponds to the phantom peak (see text). The poorer the enrichment the more pronounced is the phantom peak. $X(\ell)$ is related to $X(rl)$ to calculate the RSC and to $\text{argmin}_\delta X$ to calculate the NSC (see text). (**B**) For input controls (no enrichment) the phantom peak constitutes the maximum. *The plot in panel* **A** *was taken from the doctoral thesis of Daniel Ibrahim [53] and the plot in panel* **B** *from Landt et al, 2012 [66].*

2.2.4 *Assessment of ChIP-enrichment*

Two quality metrics, referred to as the normalized and relative strand cross-correlation coefficient (NSC and RSC), can be derived from the cross-correlation curve (Figure 7). Both of them reflect the efficiency of ChIP-enrichment, which essentially depends on the quality of the used antibody but also on other experimental details such as washing steps. Therefore, these metrics may provide valuable feedback for optimization.

The NSC puts the height $X(\ell)$ at the maximum of the cross-correlation curve into relation to the correlation at the background level.

$$\text{NSC} = \frac{X(\ell)}{\text{argmin}_\delta X} \tag{2}$$

By contrast, for the RSC the background level is eliminated by subtraction from the numerator and denominator, and the height $X(\ell)$ is put into relation to the height of the phantom peak $X(rl)$ at one read length.

$$\text{RSC} = \frac{X(\ell) - \text{argmin}_\delta X}{X(rl) - \text{argmin}_\delta X} \tag{3}$$

Since the height of the phantom peak depends on mapping artifacts arising from individual characteristics of the given reference sequence, RSC values should be compared only within the same genome build.

Based on practical experience, the ENCODE project consortium set the following rules. Experiments have to be repeated for which the NSC is smaller than 1.05 and the RSC smaller than 0.8. If the NSC and RSC cannot be improved in additional experiments, the corresponding datasets are flagged as marginal in quality. Exceptions from these rules are allowed in particular cases. For instance, transcription factors with few genuine binding sites may show low NSC values even for high-quality datasets.

2.2.5 *Peak calling using MACS2*

For ChIP-seq, the genome-wide prediction of DNA sites that interact with a given target protein is referred to as peak calling. Roughly speaking, regions with increased coverage of mapped reads (Figure 2A) are summarized in genomic coordinates with associated significance scores. Looking more carefully, it becomes apparent that ChIP-seq peak calling involves a number of known challenges that require for sophisticated solution approaches. A variety of factors influence local read depth such as accessibility of chromatin or GC content (Section 1.3), which is why the background distribution of mapped reads is not uniform but varies throughout the genome. There is also systematic variation. For instance, the fragment length may vary from one experiment to another. Furthermore, there are mapping artifacts that can be easily mistaken for genuine ChIP-seq peaks (Figure 2B). With that in mind, the task is to mark out individual peaks with greatest possible accuracy and to evaluate binding frequencies within their respective context and with regard to the biological question as well as statistical significance.

Due to the influence of the fragment length on the peak width (Section1.3), virtually all peak callers incorporate an associated parameter or even an extra routine for estimation (Section 2.2.3). The determined fragment length can be used to enhance the coverage at binding positions with respect to background. This is typically accomplished either by shifting all mapped reads by the half of the fragment length towards 3' direction [114, 125, 59, 82] or by extending all mapped reads to an entire fragment length also towards 3' direction [36, 18, 95, 113, 86, 89, 121, 26] (Figure 8A). This effectively means that reads mapped to different strands are shifted or extended in opposite directions. At binding positions, 5' ends of reads mapped to different strands are more likely to occur at a distance of a fragment length as compared to other regions (Figure 2). Therefore, the coverage profiles for shifted or extended reads have depth distributions in which ChIP-seq peaks are more pronounced as compared to the read coverage profiles.

The prepared coverage profiles are subsequently searched for local maxima referred to as *summit* positions. For each summit position, a raw signal score is reported such as the read coverage or the fold change, if data for a second condition is available. In addition, an associated P-Value is reported which reflects the probability to observe a given or even greater signal score just by chance, whereby the meaning of *by chance* depends on the setup of the analysis. For instance, if the data is evaluated with respect to input control data (Section 1.3), it means: *What is the probability to observe this score given that no ChIP has been performed.* By contrast, if the data was derived from two cell populations that were treated in different ways but ChIP has been performed in either case, by chance means: *The treatment has no impact on binding of the target protein.*

P-values are often calculated using discrete probability distributions providing the probabilities for all possible outcomes of a random experiment. For example, the probability that k reads map to a given genomic interval is often modeled using a Poisson distribution, which involves the parameter λ that defines mean and variance of the distribution at the same time and is typically estimated from the entire data assuming a uniform distribution of mapped reads across the genome. Once a λ is determined, P-values can be calculated by summing up the probabilities for observing k or more reads mapping to a given interval. Alternatively, the probabilities for less than k reads can be summed up and subtracted from 1 using the fact that the sum of the probabilities for all possible outcomes must be 1. Another probability distribution that is used in this context is the negative binomial distribution [84]. A fundamental property of the Poisson distribution is that its variance is equal to its mean. For many NGS-applications including ChIP-seq, the variance of the read counts is often much larger than the mean. This phenomenon is called *overdispersion*. The negative binomial distribution is similar to the Poisson distribution but has an extra parameter for dispersion to model the variance, which provides a better model of the background [59].

In order to decide if an observed signal score is significant or not a threshold has to be defined. Within the classical framework of statistical hypothesis testing, the P-value threshold is set to 0.05, whereas for ChIP-seq peak calling much smaller values such as 10^{-6} are used as a threshold. Peaks

with P-values greater than the specified threshold are considered to be false discoveries. Depending on the experimental conditions and the binding properties of the target protein, there may be up to tens of thousands of ChIP-seq peaks, and for each peak an individual test is performed that may erroneously discard the null hypothesis. Therefore, the P-values are typically corrected for multiple testing. The Bejamini-Hochberg procedure [14] provides a simple solution for this. First, the P-values are sorted in ascending order, and then each P-value is multiplied by the ratio of the total number of tests and the corresponding rank in the sorted list.

One of the first algorithms introduced for ChIP-seq peak calling was MACS (Model-based analysis of ChIP-Seq) [125]. Due to the early introduction and good usability it became the standard application, and the further developed version MACS2 [38] was also used in the course of the ENCODE project [66]. MACS2 includes a routine for the estimation of the average fragment length ℓ. The estimated length was formerly used to define an optimal shift size $\ell/2$. For MACS2 the shifting of reads was replaced by extension to a full fragment length (Figure 8A). The most special feature of MACS2 is the use of a dynamic local lambdas that are intended to account for the uneven background distribution of reads (Figure 8B). Instead of estimating only one global λ_{BG} from the entire genome, MACS2 estimates lambdas for windows of different sizes (λ_{1k}, λ_{5k}, and λ_{10k} kbp) centered at peaks. The local lambdas are subsequently used to approximate a binomial distribution for the statistical evaluation of peaks [38]. This has the effect that peaks within regions of low mapped read density become statistically significant even if they are below the global background level (λ_{BG}). Finally, if a control sample is available an empirical false discovery rates (FDR) is reported, which is defined as the expected proportion of peaks that were erroneously evaluated as significant amongst all peaks evaluated as significant.

2.2.6 *Irreproducible Discovery Rate (IDR)*

In accordance with the standards set by the ENCODE project consortium, Ibrahim et al. used MACS2 in conjunction with the irreproducible discovery rate (IDR) procedure [73] in order to evaluate the reproducibility of their ChIP-seq experiments and to define reproducible peaks sets. The IDR procedure quantifies the consistency between replicates and assigns each overlapping peak a posterior probability of being irreproducible. Technically speaking, the scores of the replicates are modeled as a mixture of two groups – a *reproducible* and an *irreproducible group*.

From a practical point of view, the use of the IDR procedure has the advantage that the distributions and scales of the significance scores may be different between replicates. Only the ranks in the peak lists sorted by score are taken into account, which is useful in the context of collaborative projects because it provides a flexible criterion for the reproducibility of ranked peak lists contributed by different participants. Selecting overlapping peaks at a threshold IDR ≤ 0.01 ensures that the proportion of peaks from

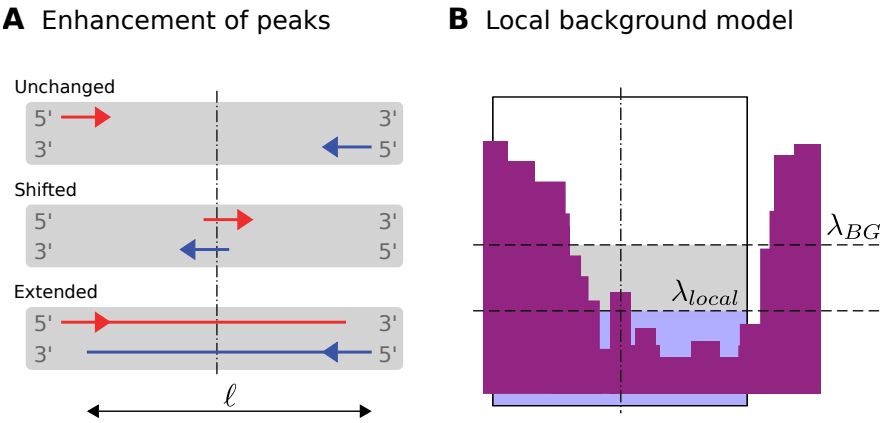


Figure 8: **Enhancement of coverage at binding sites and local background model of MACS:** (**A**) ChIP-seq peaks can be enhanced with respect to background exploiting of the characteristic strand specific distribution of mapped reads at binding sites. This is typically accomplished either by shifting all reads by $\ell/2$ towards the 3′ direction, or by extending all reads by an entire fragment length ℓ towards the 3′ direction. (**B**) Schematic representation of the local background model implemented in the peak caller MACS. The coverage profile is depicted in purple. In order to evaluate peaks, MACS dynamically derives local background parameters λ_{local} from the surrounding coverage (box). This has the advantage that peaks below the global background level (λ_{BG}) can also be detected and false-positive peak calls may be prevented in areas of high background.

the irreproducible group amongst all selected peaks is less than 1% (Figure 9A), which is similar to the FDR.

The R package for the IDR procedure includes a second component that implements the change of correspondence method. As well as the IDR procedure, this method can be used to evaluate the reproducibility of replicates but it takes a completely different approach. In principle, the change of correspondence method is based on the following assumptions. Given two replicates that measure the same underlying stochastic processes and an appropriate scoring system, the significance scores of true signals are expected to be higher and more consistent between replicates as compared to the scores of spuriously detected peaks [73]. This implies that there is a larger degree of consistency for the higher ranked peaks as compared to the lower ranked peaks, and that if we move down the ranks, the consistency will drop at the transition from signal to noise. This conception is captured by the correspondence curve Ψ and its first derivative Ψ' (Figure 9B).

The correspondence curve is constructed by successively determining the proportion of overlapping peaks for increasing fractions of peaks in the upper ranks, i.e. the proportion of overlapping peaks in the top 1% ranks, in the top 2%, and so on, up to 100%. To gain intuition, take an example with two identical replicates. In this case, the proportion of overlapping peaks would be 100% in each step, i.e. the correspondence profile Ψ would form a diagonal with a slope of one 1, and consequently Ψ' would be always 0. However, for real use cases the proportion of overlapping peaks rapidly decreases at some point referred to as *breakdown of consistency*. At this point, the correspondence curve Ψ moves away from the diagonal accompanied by a decrease in slope, i.e. the slope becomes smaller than 1. Since the correspondence analysis is

performed on overlapping peaks only, the proportion of overlapping peaks in the top 100% ranks must be 100%, i.e. the curve must return to the diagonal at some point, and the slope becomes greater than 1. The later this happens, the more peaks were consistently identified for the two replicates. A large fraction of consistently identified peaks indicates high reproducibility.

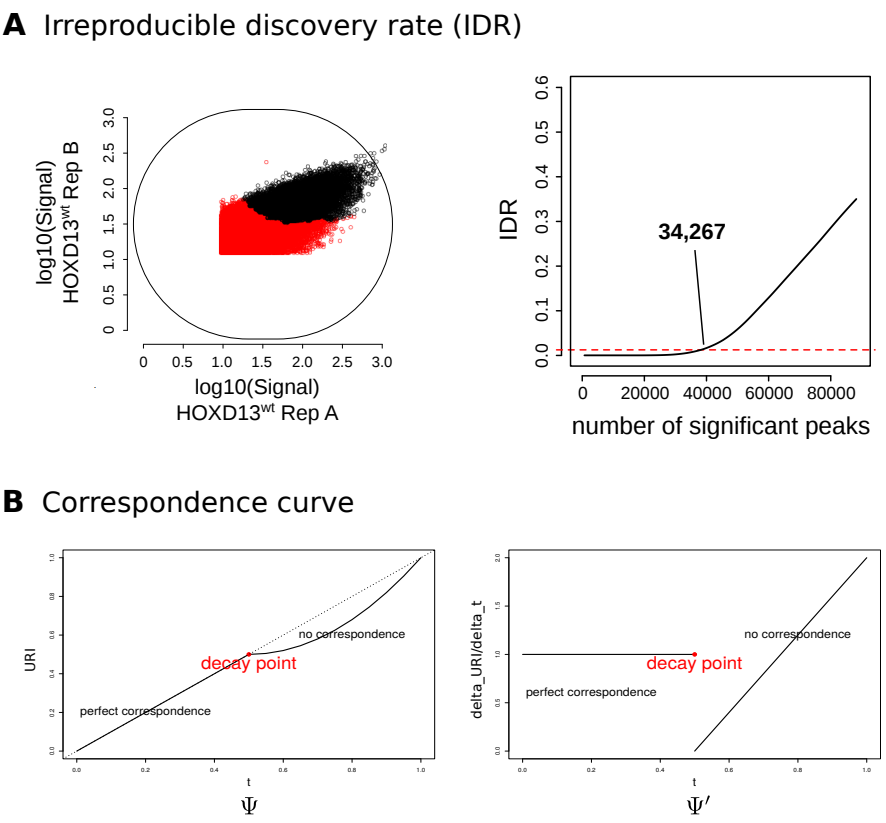


Figure 9: **Explanation of IDR and correspondence curve:** The IDR procedure includes two independent components for the evaluation of the reproducibility of ranked peak lists. (**A**) Scatterplot for logarithmized significance scores of two biological Hoxd13^{wt} replicates (left panel). Dots corresponding to overlapping peaks with a posterior probability for being from the irreproducible group greater than 0.01 are colored in red. The panel on the right-hand side shows the numbers of selected overlapping peaks at various IDR thresholds. 34,267 overlapping peaks can be selected at a threshold of IDR $\leq$ 0.01 (black dots in the scatterplot). The proportion of overlapping peaks erroneously classified as reproducible amongst all selected peaks can be assumed to be below 0.01. (**B**) Idealized example of the correspondence curve. The 50% top ranked overlapping peaks each have the same rank, whereas the ranks of the bottom 50% are shuffled. At the decay point the slope drops below 1 and from then on increases. At some point the slope must become greater than 1, because the analysis is performed on overlapping peaks only. The later this happens the more peaks are considered to be consistently identified. *The figures in the upper two panels were taken from the doctoral thesis of Daniel Ibrahim 2015 [53] and the figures in lower two panels from Li et al., 2011 [73].*

2.2.7 *Motif analysis of ChIP-seq peaks*

Many proteins bind sequence specific to DNA often collaboratively with other proteins or as part of a complex. Therefore, peak regions are typi-

cally enriched for recurring variations of short nucleotide patterns (6-8 bp) that can be combined into motifs representing sequence specific preferences of DNA binding proteins. If ChIP-seq is applied to a transcription factor, the motifs are generally associated with the analyzed factor and co-factors thereof. DNA-binding specificity is an intrinsic property of transcription factors, and the determination of the binding preferences of a given factor yields valuable input for further analyses [65]. Therefore, motif analysis is one of the key objectives in ChIP-seq data analysis [81, 12].

In order to identify motifs associated with a given factor, the nucleotide sequences beneath peaks are extracted and searched for recurring nucleotide patterns that are evaluated with respect to control data. In this context, a popular software package is the MEME Suite which provides a wide range of tools for de novo motif discovery [9, 8] and downstream analyses such as scanning extracted sequences for given motifs [42] or checking identified motifs against motif databases [97, 10].

Motifs can be represented in different ways. A simple representation method makes use of the IUPAC nucleotide code [56] which includes degenerate base symbols for all possible combinations of the four bases. For instance, TGC**K**AT stands for TGC**G**AT or TGC**T**AT (Figure 10A). Other representations such as positions weight matrices [105] or sequence logos [98] additionally take into account base frequencies at individual positions (Figure 10B).

The motif finder DREME, available as part of the MEME Suite, was used in order to derive motifs from the peaks of the ChIP-seq experiments performed by Ibrahim et al. This turned out to be a suitable choice because DREME is optimized for finding short core motifs. For wild type Hoxd13, a motif with a length of 8 bp was derived that is almost perfectly in line with a previously published Hoxd13 motif [15] indicating a valid analysis. The peaks for the mutants HoxD13^{Q317K} and HoxD13^{Q317R} as well as for PITX1 were analyzed in the same way, which resulted in motifs incompatible with the motif derived for wild type Hoxd13. However, the motif derived for the mutant HoxD13^{Q317K} is partially compatible with the motif derived for PITX1, which is not the case for the motif derived for the mutant HoxD13^{Q317R} (Figure 10B).

2.2.8 *Higher level analyses of ChIP-seq and RNA-seq data*

The lists of reproducible peaks obtained through the IDR procedure were further evaluated with respect to their distribution across the genome (Figure 11A). For this purpose, the chicken genome was subdivided into promoter, exon, intron, gene flanking and intergenic regions using annotation data from Ensembl BioMart [2]. Subsequently, the peaks were counted for each category using a custom-made PERL script that builds on BEDTtools [87] a multipurpose program for processing files in BED and related formats.

It turned out that all four analyzed proteins distribute in similar fashion over the individual regions, but peaks for wild type Hoxd13 and the mutants HoxD13^{Q317K} and HoxD13^{Q317R} occur more often in conserved regions as compared to PITX1 (Figure 11A).

For a principle component analysis of the read coverage profiles (Figure

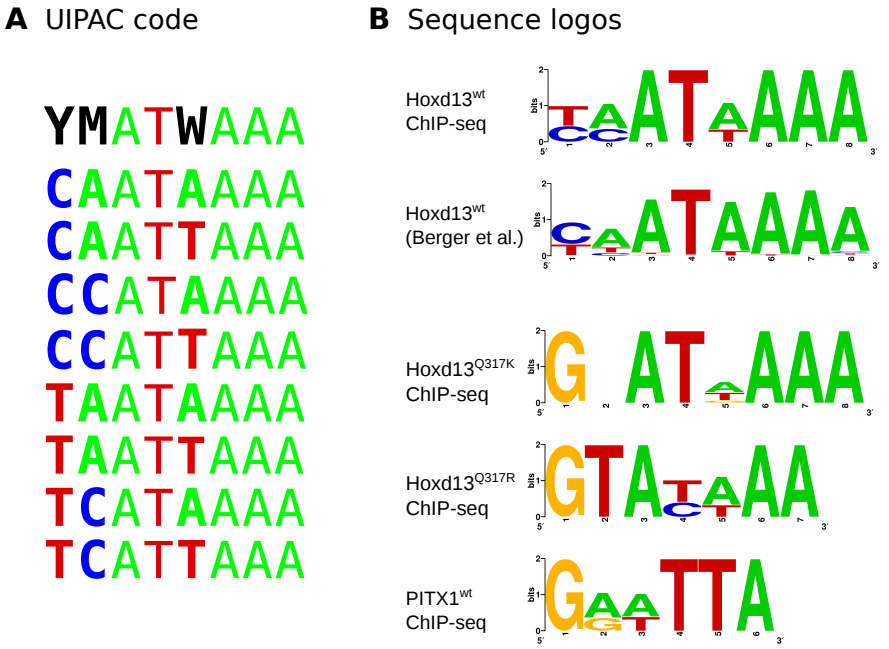


Figure 10: **Representations of transcription factor binding motifs:** ChIP-seq peak regions are enriched for variations of short nucleotide patterns which can be combined into motifs using different representation techniques. (**A**) Degenerate base symbols of the IUPAC code can be used represent ambiguous positions. For instance, `Y` stands for `C` or `T`. (**B**) Sequence logos additionally take base frequencies into account. The sequence logo for wild type HoxD13 (HoxD13^{wt}) at the top corresponds to the sequences shown in panel A and is almost identical with a previously published HoxD13 motif (Berger et al.) [15]. The motif derived for the Q317K mutant is partially compatible with the motif derived for wild type PITX1 (bottom). By contrast, the motif of the additionally analyzed mutant Q317R requires an `T` at the second position that is incompatible with the PITX1 motif. *The sequence logos in the right panel were contributed by Daniel M. Ibrahim.*

11B), the genome was windowed, and for each 500 bp window the number of mapped reads was determined for each analyzed factor. The obtained count vectors were used as input for a principle component analysis that was carried out within the open source statistical environment R [111] using the function `prcomp`. The coverage profiles for biological replicates of HoxD13^{wt} and the two mutants form separate cluster, which is in line with the postulated pathomechanism according to which the mutation within the binding domain alters the binding specificity of the mutant HoxD13^{Q317K} proteins. Furthermore, biological replicates group together indicating good reproducibility.

A similar analysis was performed for RNA-seq data but this time vectors of the base-2 logarithms of fold-changes of 3118 up- or down-regulated transcripts were used as input for the PCA (Figure 11C). The fold-change vector for HoxD13^{wt} forms a cluster with that of the mutant HoxD13^{Q317R}, and the vectors for PITX1 and HoxD13^{Q317K} form a separate cluster. This can be interpreted as a further evidence for the postulated pathomechanism according to which the HoxD13^{Q317K} mutant regulates a subset of genes that are normally regulated by PITX1.

Finally, also a combined analysis of ChIP-seq and RNA-seq data was carried out. For this purpose, HoxD13^{Q317K} /PITX1 co-bound genes were defined as genes that share a peak for the mutant HoxD13^{Q317K} and wild type PITX1, and co-regulated genes were defined as genes that are either twofold up- or -down-regulated for HoxD13^{Q317K} and PITX1. These definitions were used to break down the set of all genes into four subsets: co-bound and co-regulated, not co-bound but co-regulated, co-bound but not co-regulated, and neither co-bound nor co-regulated (contingency table). Out of all 436 co-regulated genes 57 were also co-bound (Fisher's exact P-value $= 1.295 \cdot 10^{-3}$) showing that co-regulation and co-binding do not occur independently. The same analysis was performed for HoxD13^{Q317R} and PITX1. Out of all 237 co-regulated genes 12 were also co-bound which corresponds to a number that one would expect only by chance (P-value $= 0.1335$). This shows that co-binding and co-regulation occurs independently for the HoxD13^{Q317R} and PITX1 but not for HoxD13^{Q317K} and PITX1.

For illustration purposes, a simulation study with $100,000$ iterations was performed selecting randomly from 436 co-regulated genes. For each iteration the number of selected genes that were also co-bound was determined, and the combined counts for all iterations were presented in a histogram (Figure 11D). The determined empirical P-values correspond to those obtained using Fisher's exact test.

2.3 DISCUSSION

The quality of the data and the standardized analyses (Section 2.2.1-2.2.6) contributed to the credibility of the work of Ibrahim et al. [54, 53]. Beyond that, motif analyses supported the hypothesis with the altered binding preferences of mutant HOXD13 transcription factors (Section 2.2.7). Furthermore, results of the higher-level analyses of ChIP-seq and RNA-seq data (Section

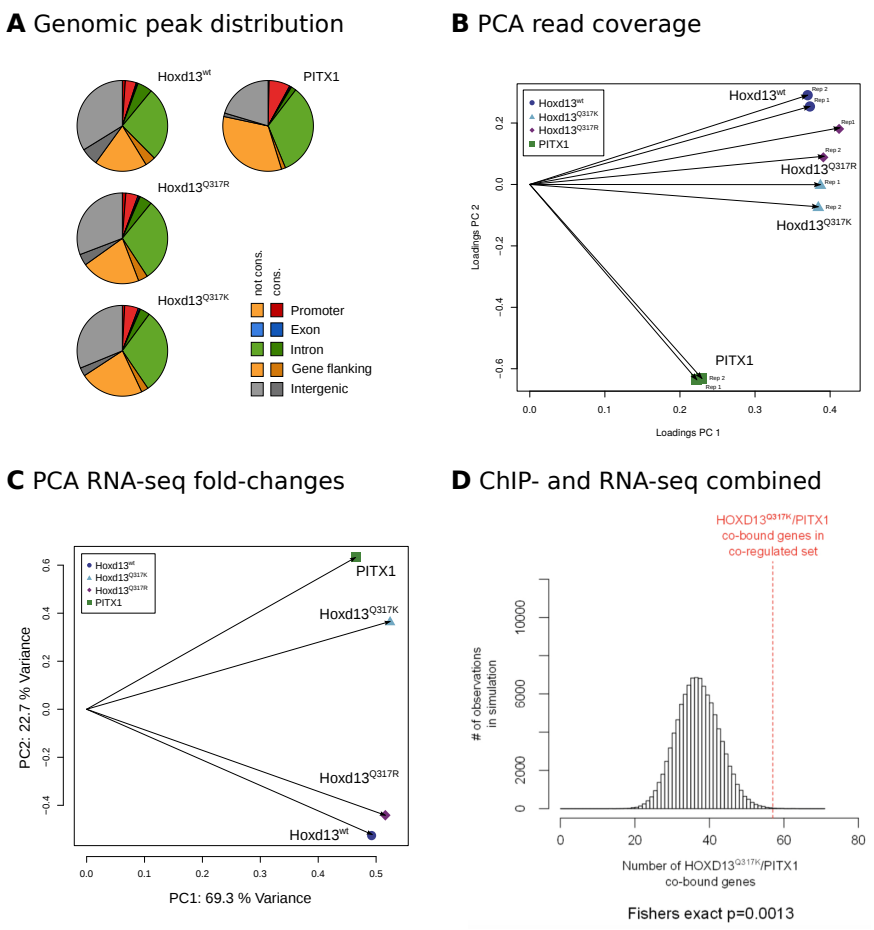


Figure 11: **Higher-level analyses of ChIP-seq and RNA-seq data:** (**A**) The pie charts show the genomic distribution of peaks. Based on annotation data, the chicken genome was subdivided into promoter, exon, intron, gene flanking and intergenic regions. Subsequently, the peaks were counted within each region. The peaks of the four proteins distribute in a similar fashion over the individual regions. (**B**) For a principle component analysis of the read coverage profiles for windows of 500 bp, the Hoxd13 proteins (wildtype and mutant proteins) and PITX1 form two distinct clusters. Furthermore, biological replicates (same symbol and color) group together. (**C**) A principle component analysis was also performed for the RNA-seq data, but this time for the base-2 logarithms of fold-changes of 3118 up- or down-regulated transcripts. The Q317R mutant forms a cluster with the HoxD13 wild type protein, whereas the Q317K mutant forms a cluster with the PITX1 wild type protein. (**D**) Illustrative presentation of the Fisher's exact test of co-bound and co-regulated genes that were defined using ChIP-seq and RNA-seq data (see text). *These figures were originally published in Ibrahim et al., 2013 [54].*

2.2.8) are in line with the hypothesis with a global shift of the regulatory role of the mutant HoxD13^{Q317K} towards that of PITX1.

In the course of the analyses that were carried out in collaboration with Daniel M. Ibrahim and Jochen Hecht, there were regular meetings with wet-lab biologists performing ChIP-seq and active exchange about the ENCODE guidelines. All tools required for data analyses in compliance with the standards were installed on a locally installed GALAXY instance [17, 40] at the Charité. This provided scientists from various institutes with the possibility to use this resource for data analysis [28, 53, 58]. One the other hand, it contributed to deeper understanding of the data.

In the course of practical applications, it became evident that the usage of the peak caller MACS2 (Section 2.2.5) in combination with the IDR procedure (Section 2.2.6) does not yield optimal results in terms of resolution and reproducibility. Manual inspection of the coverage at peak regions revealed that MACS2 occasionally tends to combine adjacent peaks in close proximity into one peak. This unstable characteristic affects the ranking of peaks and thus potentially also affects the results of the IDR procedure. Another drawback of the recommended pipeline was that the estimation of the predominant fragment length using the cross-correlation method (Section 2.2.4) is very time-consuming. This can be overcome by using larger step sizes, e.g. 10 bp instead of 1 bp, but this will result in coarser estimates which is not acceptable, because the fragment length is an important parameter for downstream analyses, especially peak calling. To overcome these drawbacks, an improved peak calling algorithm was developed that is presented in the next chapter.

3

REPRODUCIBLE CHIP-SEQ PEAK CALLING

3.1 INTRODUCTION

Here, a peak caller named Q [46] was developed to address all the shortcomings that became apparent in the course of the practical application of available software (Chapter 2). Q was implemented in C++ using the SeqAn library [30] that enables efficient analysis of next-generation sequencing data. The result is a fast and user-friendly software package that allows for accurate and reproducible identification of ChIP-seq peaks.

The implementation of Q includes a module for the estimation of the predominant fragment length (Section 3.2.1). To a great extent, the methodology was inspired by the cross-correlation method and yields almost equivalent results but three times faster. For ChIP-seq, enhancement of read coverage at peak positions is typically achieved by either shifting or extending reads by the half or the entire fragment length towards 3' direction. In Q, an alternative approach was implemented that makes use of *qfrags* that are similar to extended reads but the coverage profiles for qfrags are quadratically amplified in peak regions (Section 3.2.2). The improved coverage profiles allow more accurate prediction of binding sites. Another innovation of Q is the statistical evaluation of peaks that shifts the focus from peak height towards saturation of genomic positions around binding positions (Section 3.2.3).

Q was compared to three other peak callers recommended by the ENCODE project consortium with regard to runtime (Section 3.2.4), reproducibility (Section 3.2.5) as well as motif content of peaks (Section 3.2.6). Furthermore, Q was used to characterize the architecture of paused open promoters using data for RNA polymerase II (RNAPII) and the histone modification H3K4me3 (Section 3.2.7).

3.2 METHODS AND RESULTS

3.2.1 *Fragment length estimation*

The predominant fragment length in ChIP-seq sequencing libraries is a crucial parameter for peak calling and downstream analyses. The well accepted cross-correlation method (Section 2.2.3) can be used to estimate this parameter from mapped read data. In addition, the cross-correlation curve allows the assessment of ChIP-enrichment (Section 2.2.4).

For the cross-correlation method, the Pearson correlation coefficient between count vectors for the forward and reverse strand is calculated for

increasing strand shifts, and the shift size with the highest correlation is interpreted as the fragment length. The count vectors contain the numbers of 5′ end positions of mapped reads at any given genomic position. If duplicated reads are removed beforehand, the counts are either 0 or 1. Based on this fact and in order to improve efficiency, an alternative metric for similarity between the two strands was implemented in Q. Instead of the Pearson correlation the Hamming distance is calculated for each strand shift, and the shift size with the minimal distance is interpreted as the fragment length.

Analogously to the cross-correlation method, initially two bit vectors $n_c^f(x)$ and $n_c^r(x)$ for the two strands are created for each chromosome c. The bits of these vectors are set, if the corresponding genomic position is covered by at least one 5′ end position of a mapped read. Subsequently, the two vectors are shifted against each other, and for each shift size σ the Hamming distance $H(\delta)$ is calculated as follows:

$$H(\delta) = \sum_{c \in C} d_H[n_c^+(x+\delta), n_c^-(x)], \tag{1}$$

where $d_H[X, Y]$ is the Hamming distance between the vectors $X = x_1, ..., x_n$ and $Y = y_1, ..., y_n$ and C is the set of all chromosomes. The shift size δ that corresponds to minimum Hamming distance is taken as the estimated fragment length

$$\ell = \text{argmin}_\delta H(\delta). \tag{2}$$

The Hamming distance corresponds to the number of positions by which the vectors differ and can be calculated by applying the logical operator XOR and summing up the number of set bits in the resulting vector. Efficient implementations of bit vectors and operations of the Boost C++ library were used for the implementation in Q.

For verification purposes, the two methods were applied to 38 ENCODE datasets *(Supplemental Table S1* of the original publication [46]). The used datasets were generated in the course of the ENCODE project [41, 66] and already mapped reads in BAM format were downloaded from UCSC [94]. In order to ensure comparable results, duplicated reads were removed, and a bin size of 1 bp was used.

In most cases, the two methods produce curves of almost equivalent shape, and the estimate for the fragment length differs by exactly 1 bp, which can be explained by an index shift (Figure 12).

For datasets with good ChIP-enrichment, also the values for the quality metric RSC (Section 2.2.4) are comparable. However, for datasets with poor ChIP-enrichment the values may differ more significantly. To compensate for this, Q was applied to 392 available ENCODE datasets so as the results can be used as a reference for the RSC as well as for other quality metrics

derived by Q[1].

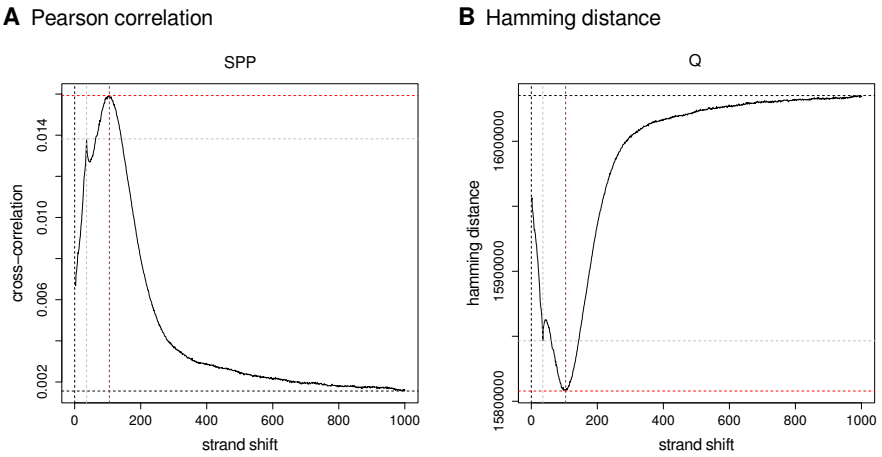


Figure 12: **Fragment length estimation:** (**A**) Cross-correlation and (**B**) Hamming distance curve for the dataset GM12878-BATF-REP1. In most of the cases, the two methods produce equivalent curves that are horizontally flipped. Additional plots can be found in *Supplemental Fig. S1* of the original publication. *This figure was originally published in Hansen et al., 2015 [46].*

3.2.2 *Concept of qfrags*

In this section, an innovative measure of coverage is introduced. In order to selectively increase coverage within peak regions, reads are typically shifted by half a fragment length or extended by an entire estimated fragment length towards 3′ direction (Section 2.2.5). Those methods use the fact that, within peak regions, two reads mapped to the forward and reverse strand are more likely to occur at a distance of about one fragment length to one another (Figure 2). The same fact is used in Q, but instead of shifting or extending reads, the 5′ ends of any pair of reads are connected to a *qfrags*, if they are at a distance of about one fragment length, and the first read maps to the forward and the second to the reverse strand (Figure 13). Intuitively, this method should yield a quadratic increase in coverage at binding sites in contrast to the merely linear increase that can be achieved through shifting or extension of reads.

More formally, a *qfrag* is defined as the genomic interval between any pair of 5′ end positions of reads mapped to the forward and reverse strand at a distance of at least $q_{min} = \ell - x$ and at most $q_{max} = \ell + x$, where ℓ is the estimated fragment length (Section 3.2.1) and x is intended to reflect deviations from ℓ (Figure 13). The qfrag coverage at any given position in a genome equals the number of qfrags that cover the position, and the consecutive qfrag coverage along all positions is referred to as qfrag coverage profile (Figure 13B). The centers of local maxima in the qfrags coverage profile are defined as predicted binding positions that will be statistically tested in the subsequent step of the algorithm (Figure 13C).

[1] http://charite.github.io/Q/tutorial.html#statistics_and_quality_metrics

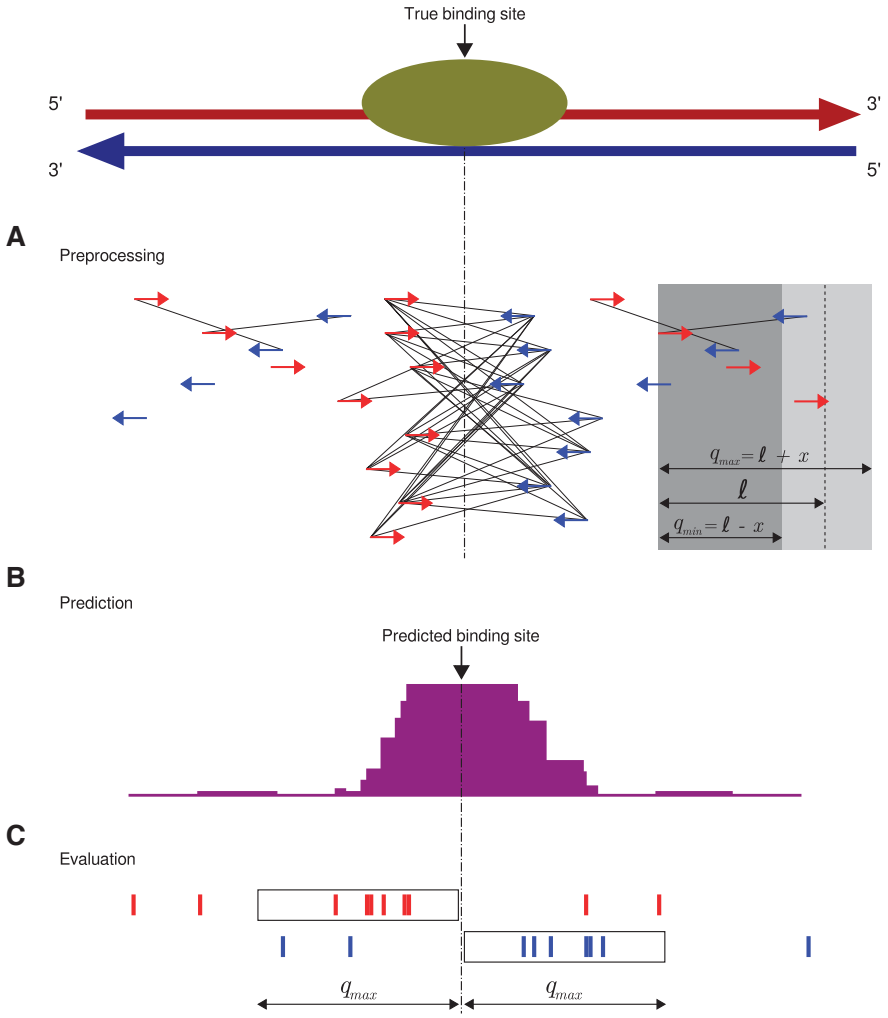


Figure 13: **Concept of qfrags and process description for Q:** (**A**) qfrags (black lines) are formed between 5′ end positions of read pairs that map to the forward (red arrows) and reverse strand (blue arrows) at a distance of at least $q_{min} = \ell - x$ and at most $q_{max} = \ell + x$ (gray box), e.g. the read at the left edge of the gray box (red) forms qfrags with all compatible reads (blue) within light gray area of the box, which results in only one qfrag in this case. (**B**) The qfrag coverage (purple) is calculated for each genomic position using a sliding window approach (gray box). Summit positions are defined to be the center of local maxima in the in the qfrag coverage profile. (**C**) Candidate regions for statistical evaluation are defined as the summit position $\pm q_{max}$. *This figure was originally published in Hansen et al., 2015 [46].*

In order to explore the qfrag coverage profile and also to compare it with the conventional coverage profiles, all profiles were visualized using IGV [55]. For this purpose, appropriate files were prepared using an extra subroutine of Q that estimates the fragment length ℓ from the mapped reads (Section 3.2.1), and subsequently writes out four BED files for reads, shifted reads ($\ell/2$), extended reads (ℓ) and qfrags (ℓ and $x = 50$). The BED files were converted into IGV's binary tiled data format using the sub command `count` of `igvtools` [55] with a bin size of 1. The height of each bin corresponds to the number of features that cover the bin.

The four coverage profiles were constructed for a dataset derived from a ChIP-seq experiment with RNAPII in HeLa-S3 cells *(Supplemental Table S1* in [46]). All four profiles show peaks at the transcription start sites (Figure 14), which is a typical feature of RNAPII ChIP-seq data. Apart from that, the different read transformations lead to coverage profiles with different depth distributions. The profiles for extended reads and qfrags look smoother compared to those for reads and shifted reads. For extended reads, there is two-fold increase in the maximum coverage as compared to the maximum coverage for reads or shifted reads, whereas for qfrags the increase is more than quadratic. In proportion to background regions, the peaks in the qfrags coverage profile are most pronounced and more distinct.

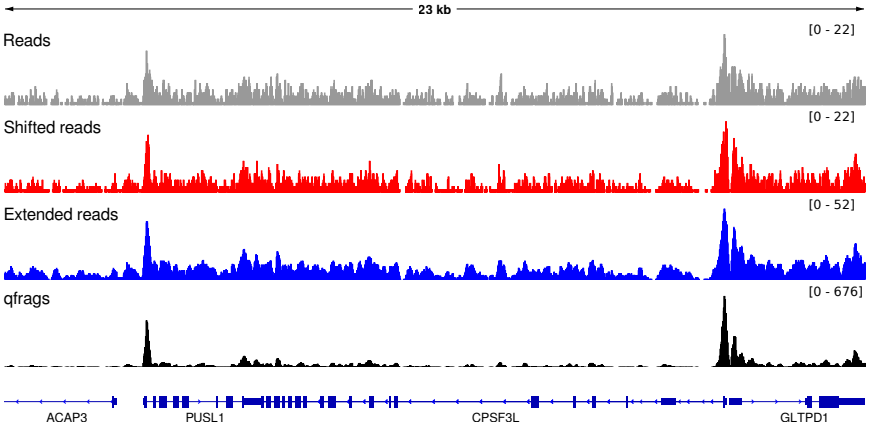


Figure 14: **Coverage profiles for a ChIP-seq experiment with RNA polymerase II in HeLa-S3 cells:** The coverage profiles for reads (gray), shifted reads (red), reads extended to fragments (blue) and qfrags (black) were derived using Q [46] and visualized using the IGV [55]. The same estimated fragment length of $\ell = 120$ bp was used for shifting ($\ell/2$), extension (ℓ) and formation of qfrags (ℓ and $x = 50$). Each track is scaled to its maximum value within the displayed region (chr1:1,240,182-1,263,869) given in square brackets. *This figure was originally published in Hansen et al., 2015 [46].*

Due to the quadratic nature of qfrags, naive sequential approaches cannot be used for efficient construction of the qfrags coverage profile and prediction of peaks therein. Writing to and reading from disk is too time consuming, and the memory of a standard computer is typically too small[2]. Therefore, the identification of local maxima in the qfrags coverage profile was implemented using a sliding window approach.

Within the sliding window three tasks are performed: formation of qfrags, construction of the coverage profile and detection of local maxima therein.

[2] 2 or 4 GB RAM were standard at the time of implementation.

The chromosomes are processed separately, and the values for the local qfrag coverage are stored in arrays of much smaller size (1000 bp) as compared to those of typical chromosomes. Previous values no longer required are repeatedly overwritten by applying the modulo operator with the array size to the current chromosomal position. Due to this implementation detail, chromosomes can even be processed in parallel on standard computers without running into memory issues.

In what follows, a more formal description of the algorithm is provided. The 5′ end positions of reads mapped to the forward (f) or reverse (r) strand of a target sequence of length l are here referred to as *hits* and defined as:

$$h = (\text{pos} \in \{1, ..., l\}, \text{strand} \in \{f, r\}). \tag{3}$$

The algorithm iterates over the sorted hits on the forward strand, instead of iterating over each chromosomal position. For a current hit (g_c, f) on the forward strand, and for each subsequent hit (g_n, r) on the reverse strand within the region $(g_c + q_{min}, ..., g_c + q_{max})$, the current qfrag coverage qc_0 at the genomic position g_c is incremented by 1, and the qfrag coverage at position $g_n + 1$ required for future calculations is decremented by 1. The previous two window positions are kept track of together with the corresponding qfrag coverages $qc_{(-2)}$ and $qc_{(-1)}$. If $qc_{(-1)}$ is greater than $qc_{(-2)}$ and $qc_{(0)}$, the center position of the local maximum with a coverage of $q_{(-1)}$ is defined and reported as a *raw* summit.

The number of raw summits potentially can become large, because every unevenness in the profile constitutes a summit (Figure 15). Therefore, the set of raw summits is further refined by discarding summits that are not freestanding, defined as follows. A summit at position s_i is not freestanding, if there is another summit with a greater qfrag coverage in one of the adjacent regions $(s_i - q_{min}, ..., s_i - 1)$ or $(s_i + 1, ..., s_i + q_{min})$. In a second refinement, step all adjacent summit positions s_i and s_{i+1} with the same qfrag coverage are combined into a single new summit position $s_j = s_i + \lceil (s_{i+1} - s_i)/2 \rceil$ that is located in center between the two original summits. Finally, the center positions of all remaining freestanding summits and combined summits are defined to be the predicted binding positions in the following also referred to as summit positions (Figure 13B).

3.2.3 *Saturation-based evaluation of ChIP-seq peaks*

3.2.3.1 *Introduction of saturation*

Traditionally, peak calling algorithms are geared towards peak height, i.e. candidate regions are tested for significantly increased read or extended read depth using mostly a Poisson or negative binomial distribution in order to model the background distribution of reads [84]. In this section, an alternative concept is introduced that is different with respect to several aspects.

Hits most naturally represent breakpoints in genomic DNA (Section 1.3), whereas the read or fragment length are experimental parameters that may vary from one experiment to another. For Q, only hits that belong to qfrags

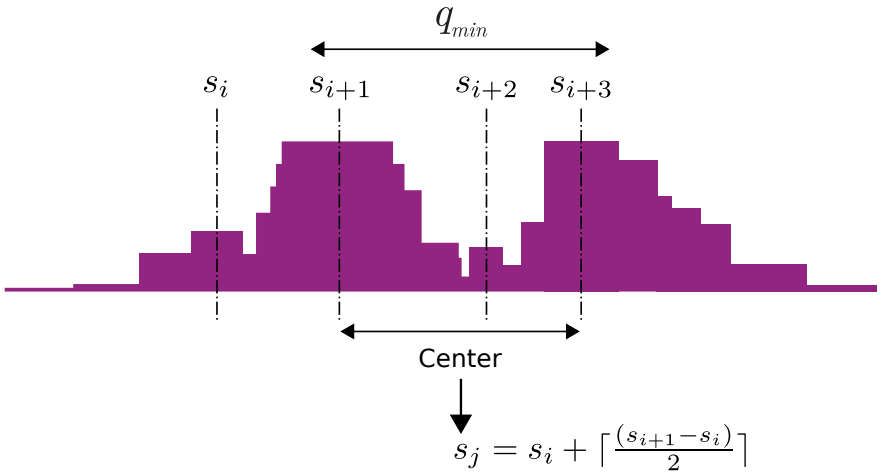


Figure 15: **Schematic illustration of summit refinement:** Q identifies all local maxima in the qfrag coverage profile. The centers of the maxima are defined as raw summits, in this example s_i, s_{i+1}, s_{i+2} and s_{i+3}. In a first refinement step, all raw summits that have adjacent summits with a greater qfrag coverage within a distance of at most q_{min} bp are discarded, in this case s_i and s_{i+2}. In a second step, adjacent summits with equal qfrag coverage within a distance of at most q_{min} bp are combined into a single new summit that is located in the center between the original two summits. In this example, s_{i+1} and s_{i+3} are combined into s_j.

are taken into account. Given the strand specific distribution at peaks, these hits mutually confirm each other inasmuch as that they more likely to originate from peak regions.

Another innovation of Q is that the peak height measure is replaced by saturation. A genomic position is here referred to as saturated, if it is occupied by at least one hit that belongs to a qfrag, and the strength of any given peak is measured as the proportion of saturated positions surrounding the summit position (Figure 13C). Due to the characteristic strand specific bimodal distribution of forward and reverse strand hits at peaks (Figure 2A), these positions tend to be well saturated.

Identified candidate regions are tested statistically with regard to the number of saturated positions using a binomial test with the probability p that is modeled within the framework of the classical occupancy problem [37]. Furthermore, an extension of the binomial test allows for taking into account data derived from a control experiment. In this case, it is tested for the difference of saturated positions between ChIP and control data within a given candidate region.

3.2.3.2 *Statistical model without control data*

Given a refined summit s_i (Section 3.2.2), the genomic candidate region that will be statistically tested for saturation is defined as $s_i - q_{max}, ..., s_i + q_{max}$ (Figure 13C). The output of a ChIP-seq experiment can be seen as a set of hits:

$$T = \{h = (\text{pos}, \text{strand}) | \text{pos} \in \{1, ..., l\} \wedge \text{strand} \in \{f, r\}\}, \qquad (4)$$

where T stands for treatment (i.e., use of a specific antibody for enrichment of a ChIP-seq target protein). This set can be divided into hits on the

forward and on the reverse strand here referred to as T_f and T_r. Since both ends of each fragment in the sequencing library are sequenced with equal probabilities, the number of hits for two strands can be assumed to be approximately equal, i.e. $|T_f| \approx |T_r|$.

For the null model it is assumed that hits are evenly distributed across the genome and independently of the two strands. A position i is saturated, if it is covered by a hit on the forward strand that is part of a qfrag. This can only be the case, if there is at least one hit on the reverse strand within the range $i + q_{min}, ..., i + q_{max}$. Given the null model, the expected number of reverse strand hits at each position of a given reference sequence of length l is $|T_r|/l$, and, by linearity of expectation, the expected number of such hits at positions $i + q_{min}, ..., i + q_{max}$ is

$$\lambda_t = (q_{max} - q_{min}) \cdot \frac{|T_r|}{l}, \tag{5}$$

where the subscript t stands for treatment. The probability that there is at least one hit on the reverse strand within the range $i + q_{min}, ..., i + q_{max}$ is the same as the probability for a hit on the forward strand at position i to be part of a qfrag and is calculated using a Poisson distribution.

$$P(h_i \text{ is part of a qfrag} \mid h_i.\text{strand} = f) = 1 - \text{Pois}(0, \lambda_t). \tag{6}$$

At any given position, the expected number of hits on the forward strand is $|T_f|/l$. Therefore, the rate of qfrag start positions, is

$$\frac{|T_f|}{l} \cdot (1 - \text{Pois}(0, \lambda_t)). \tag{7}$$

Since the total number of hits are assumed to be approximately equal for the forward and reverse strand, also the rates for the two strands can be assumed to be approximately equal. Therefore, if it is not distinguished between strands, the expected joint rate of qfrag start and end positions is calculated as

$$r_t = 2 \cdot \frac{|T_f|}{l} \cdot (1 - \text{Pois}(0, \lambda_t)). \tag{8}$$

Using this rate r_t, the saturation of genomic positions with qfrag start or end positions is then modeled within the framework of the occupancy problem [37], which can be expressed as: *Placing m balls randomly into n bins, what is the probability that exactly $n - k$ bins remain empty?* The probability that a given ball is placed in one particular bin out of n bins is $1/n$. Vice versa, the probability that the ball is not placed in this bin is $(1 - 1/n)$, and the probability that any of the m balls is placed into this bin is

$$P(\text{bin remains empty}) = (1 - \frac{1}{n})^m \cong e^{-m/n}. \tag{9}$$

See Appendix B for a proof of the approximation above. This link can be used to estimate the probability that a given position is saturated as

$$p_t = 1 - e^{r_t}. \tag{10}$$

A random variable Q_t is defined on the sample space $\Omega = \{0, ..., 2 \cdot q_{max}\}$ that represents all possible numbers of saturated positions within a window

of length $2 \cdot q_{max}$. The saturation of each individual position is then modeled as independent and identically distributed Bernoulli trial. Given this, the random variable Q_t follows the binomial distribution

$$Q_t \sim \text{Bin}(n = 2 \cdot q_{max}, p = p_t). \tag{11}$$

Given the null model, the probability that exactly k positions within a window of length $2 \cdot q_{max}$ are saturated by chance is then calculated as:

$$P(Q_t = k) = \binom{2 \cdot q_{max}}{k} \cdot p_t^k \cdot (1 - p_t)^{2 \cdot q_{max} - k}, \tag{12}$$

and the probability that at least k window positions are saturated is calculated as:

$$P(k \leq Q_t \leq q_{max}) = \sum_{i=k}^{2 \cdot q_{max}} \binom{2 \cdot q_{max}}{i} \cdot p_t^i \cdot (1 - p_t)^{2 \cdot q_{max} - i}. \tag{13}$$

This probability is also referred to as P-vaule which is corrected for multiple testing using the Benjamini-Hochberg procedure.

3.2.3.3 *Statistical model with control data*

In order to extend the model to the case with control data, another set C for control hits is defined analogously to T, and, as for T, it can be assumed that $|C_f| \approx |C_r|$. Furthermore, a second random variable Q_c on the sample space $\Omega = \{0, ..., 2 \cdot q_{max}\}$ is defined analogously to Q_t:

$$\begin{aligned} \lambda_c &= (q_{max} - q_{min}) \cdot \frac{|C_r|}{l}, \\ r_c &= 2 \cdot \frac{|C_r|}{l} \cdot (1 - \text{Pois}(0, \lambda_c), \\ p_c &= 1 - e^{-r_c}. \end{aligned} \tag{14}$$

For the null model, Q_t and Q_c are assumed to be independent. Based on this, a third random variable $Q_d = Q_t - Q_c$ is defined on the sample space $\Omega = -2 \cdot q_{max}, ..., 0, ..., 2 \cdot q_{max}$. Q_d represents all possible differences between the numbers of saturated positions within a window of length $2 \cdot q_{max}$. That is, if there are more saturated window positions for the treatment data, then the difference d will be greater than 0. Conversely, if there are more saturated window positions for the control data, d will be less than 0.

The derivation of the probability for observing a difference of $Q_d = d$ requires the analysis of the convolution of Q_t and Q_c. To gain intuition, imagine a random experiment in which one random number is drawn from Q_t and another one from Q_c. For didactic purposes, assume $d \geq 0$ for a start. There are $2 \cdot q_{max} - d + 1$ possible outcomes. There can be d saturated positions for the treatment dataset, i.e. $Q_t = d$, and zero saturated positions for the control dataset, i.e. $Q_c = 0$. Or it can be $Q_t = d + 1$ and $Q_c = 1$, and so up, until the window is completely saturated for the treatment dataset, i.e. $Q_t = 2 \cdot q_{max}$ and $Q_c = 2 \cdot q_{max} - d$. Therefore, the products for all possible

outcomes need to be summed up in order to calculate the probability for observing a difference of $Q_d = d$:

$$P(Q_d = d) = \sum_{i=0}^{2 \cdot q_{max} - d} P(Q_t = i + d) \cdot P(Q_c = i). \tag{15}$$

However, if $d < 0$, the index i can become larger than $2 \cdot q_{max}$. Furthermore, it is possible that $i + d < 0$ or $i + d > 2 \cdot q_{max}$. Therefore, the following equation is used for $d < 0$:

$$P(Q_d = d) = \sum_{i=0}^{2 \cdot q_{max} - |d|} P(Q_t = i) \cdot P(Q_c = i + |d|). \tag{16}$$

The probability to observe a difference in saturated positions of at least d is calculated as follows:

$$P(d \leq i \leq 2 \cdot q_{max}) = \sum_{i=d}^{2 \cdot q_{max}} P(Q_d = i). \tag{17}$$

And, as for the case without control, this P-value is corrected for multiple testing using the Benjamini-Hochberg procedure.

3.2.3.4 *Influence of the control dataset size on P-values*

Treatment and control samples are typically not sequenced to the same depth. The saturation model with control data takes this into account, because the rate parameters r_t and r_c are estimated from the treatment and control data separately (Formulas 8 and 14). Therefore, no downsampling or scaling is necessary, if the number of hits differs for the treatment and control dataset.

To analyze the effect of different control dataset sizes on P-values, a ChIP-seq dataset for RNAPII in HeLa S3 cells was used *(Supplemental Table S1* in [46]). If duplicated reads are removed, there are 23,494,468 hits for the treatment and 29,454,439 hits for the associated control dataset. From these two datasets, four datasets were derived for testing. First, the treatment dataset was downsampled to $n = 11,747,234$ hits, which is exactly half of the original number. The resulting dataset is here referred to as $\mathcal{T}$. Second, the control dataset was downsampled to $2 \cdot n$, n and $n/2$ hits. The resulting datasets are here referred to as $\mathcal{C}_2$, $\mathcal{C}_1$ and $\mathcal{C}_{1/2}$. Next, the Q software was applied to the three treatment-control dataset pairs $(\mathcal{T}, \mathcal{C}_{1/2})$, $(\mathcal{T}, \mathcal{C}_1)$ and $(\mathcal{T}, \mathcal{C}_2)$. For Q, the identification of candidate regions is performed on the treatment dataset only, and for this analysis, no P-value cutoff was used. Therefore, the identical peak set comprising 746,755 peaks was derived in all three cases.

For visual inspection, the P-values were plotted against each other for different ratios of treatment and control hits (Figure 16). For the predominant portion of peaks, approximately the same P-value is determined for different control dataset sizes, indicating that the evaluation of peaks is largely independent of the control dataset size. For smaller control dataset sizes, some peaks are assigned P-values much smaller (more significant) than those obtained for larger control dataset sizes. Manual inspection revealed that

such peaks almost exclusively correspond to mapping artifacts due to repeats mainly in centromeric regions. All things considered, the saturation-based evaluation of peaks seems to be quite robust against varying sizes of control datasets, and the more control data there is, the fewer artifacts will be detected as significant.

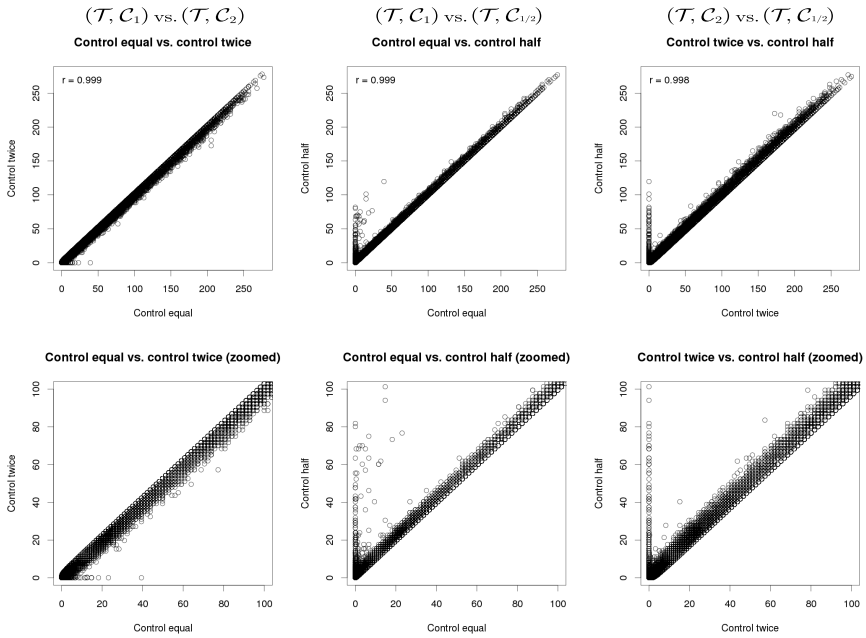


Figure 16: **Effect of different control dataset sizes on P-values:** The Q software was applied to the same treatment dataset $\mathcal{T}$ using control datasets of different sizes (see text). $\mathcal{C}_2$ contains twice as much, $\mathcal{C}_1$ the same number and $\mathcal{C}_{1/2}$ half as much hits as $\mathcal{T}$. In each case, the same set of 746,755 peak regions is identified, because the identification step is independent of the control dataset, and no P-value cutoff was used in this case. The P-values ($-\log_{10}$) were plotted against each other for different ratios of treatment and control hits. The plots in the upper row show the full range of determined P-values for $(\mathcal{T}, \mathcal{C}_1)$ vs. $(\mathcal{T}, \mathcal{C}_2)$ (left), $(\mathcal{T}, \mathcal{C}_1)$ vs. $(\mathcal{T}, \mathcal{C}_{1/2})$ (middle), and $(\mathcal{T}, \mathcal{C}_2)$ vs. $(\mathcal{T}, \mathcal{C}_{1/2})$ (right). The plots in the lower row show the same data but only for the range 0 to 100. The predominant portion of peaks are assigned approximately the same P-values (dots on the diagonal). Only a small number of peaks are assigned much more significant P-values, if smaller control datasets are used (dots along the y-axes) *This figure was originally published in Hansen et al., 2015 [46].*.

3.2.4 *Implementation and runtime analysis*

The estimation of the average fragment length via Hamming distance (Section 3.2.1), the identification of peak regions using qfrags (Section 3.2.2) and the subsequent saturation-based evaluation of peaks (Section 3.2.3) is implemented in a C++ command line program named Q. The software makes use of the SeqAn library [30] which provides routines for command line parsing, reading and writing of standard formats as well as efficient data structures and algorithms for sequence analysis. The source code, executable binaries, a detailed documentation and a tutorial is available on GitHub[3].

[3] http://charite.github.io/Q/

Q can be applied with or without control data. The hits can be read from SAM or BAM formatted files, whereby prior sorting or removal of duplicates is not necessary. If no value for the fragment length ℓ is specified, this parameter will be estimated from the input data and used for the identification and evaluation of peaks. The identified peaks are written to ENCODE narrowPeak formatted files [94] containing the coordinates, the raw values for saturation, associated P-values and Q-values corrected for multiple testing. Beyond that, there is a number of features and additional output files that are useful for quality assessment and visualization of ChIP-seq data.

Q has a small memory footprint and can be used on standard desktop computers. In addition, Q supports parallel computing and can utilize multiple processors on larger computers. However, the runtime of Q [46] was compared to that of MACS2 [125, 38], SPP [63] and PeakSeq [96] using only a single thread omitting differences regarding parallelization. Individual runtimes were determined for 38 datasets *(Supplemental Table S1* in [46]) and average runtimes were used for comparison.

First, Q's runtime for the estimation of the fragment length was compared to that of SPP (Section 3.2.1). This was done on a desktop computer and on a rack server before and after removal of duplicates. For all test setups, the fragment length can be estimated at least three times faster using Q instead of SPP (Table 1).

Next, the runtime for peak calling was compared to that of MACS2, SPP and PeakSeq. This comparison was performed after removal of duplicates and on a rack server only. For each given dataset, the same fragment length ℓ, previously estimated with Q, was used for all peak callers. The internal estimation routines of SPP and MACS2 were omitted. In this way, only the runtime for peak calling is taken into account. On average, peak calling with Q can be performed in only two minutes. Compared to the other three peak callers, Q shows a three to 19-fold improvement in runtime (Table 2).

CPU	Remove dup.	N-Fold improvement	SPP (m)	Q (m)
Intel	no	3.12	45.56	15.28
Intel	yes	3.09	45.12	15.11
AMD	no	3.61	138.76	44.25
AMD	yes	4.74	144.69	31.33

Table 1: **Runtime comparison - Fragment length estimation with SPP and Q:** The comparisons were performed on a desktop computer with Intel® Core™ i7-3770 (3.4 GHz) processors and on a rack server with AMD Opteron™ 6172 processors (2.1 GHz) before and after removal of duplicates. The average runtimes (user time plus system time) for 38 datasets are indicated in minutes.

Q (m)	MACS2 (m)	SPP (m)	PeakSeq (m)
2.06	10.92	38.11	7.27

Table 2: **Runtime comparison - Peak calling with Q, MACS2, SPP and PeakSeq:** The same 38 datasets as for the previous runtime analysis were used (Table 1). The comparison was performed after duplicate removal on a rack server only. Internal routines for fragment length estimation were omitted.

3.2.5 *Reproducibility of ChIP-seq peak calling*

3.2.5.1 *Comparison framework*

The reproducibility of peak evaluation (Section 3.2.3) was compared for the peak callers Q, MACS2, SPP and PeakSeq. This was done within the framework of the IDR procedure (Section 2.2.6). In practice, the IDR procedure is used for the assessment of the biological reproducibility of ChIP-seq experiments [66]. The proposed workflow includes multiple branches in which additional datasets are derived from the mapped reads of two biological replicates by pooling and random sampling. Peak calling is then performed on each prepared dataset, and pairs of ranked peak lists are used as input for the IDR procedure. For the analysis presented here, only *pseudo-replicates* were used, which were derived by splitting the set of mapped reads for given biological replicates randomly into two halfs. The same pair of pseudo-replicates was then used as input for each of the four peak callers in conjunction with the IDR procedure. In this way, the reproducibility of peak calling can be compared between peak callers.

To be more precise, given the mapped reads for a ChIP-seq and associated control experiment, four pseudo-replicates were derived (Figure 17A). In order to obtain the same number of reads for the ChIP-seq and the control dataset, the larger dataset was downsampled before splitting. This was done to eliminate effects arising from up- or down-scaling. For instance, MACS2 performs linear upscaling of the smaller dataset by default. Furthermore, the cross-correlation method of the SPP package was used to estimate the fragment length ℓ and the window half size (*whs*), whereby the *whs* corresponds to the width of the cross-correlation peak at $1/3$ of the peak height and is used by SPP as a parameter analogous to ℓ.

In a next step, peak calling was performed using the prepared pseudo-replicates for treatment and control as input (Figure 17B). If appropriate for this particular application, the parameter settings for the individual peak callers were chosen according to the recommendations of Anshul Kundaje [64][4]. For better comparability, the same previously estimated parameter ℓ (or *whs* only for SPP) was used for all peak callers in appropriate ways. The derived peak lists were sorted by significance and only the top $100,000$ peaks were used as input for the IDR procedure.

3.2.5.2 *Application of the comparison framework to 38 datasets*

The framework for comparison of reproducibility of peak calling was applied to the 38 datasets that were also used for the runtime analyses (Section 3.2.4), and the results of the IDR analyses were compared for the four peak callers.

Figure 18 shows detailed results for a dataset derived for a ChIP-seq experiment with RNAPII in HeLa S3 cells. The first step of the IDR procedure consists in the determination of overlapping peaks. A large proportion of overlapping peaks indicates good reproducibility. For the RNAPII dataset, Q

[4] Find a detailed listing of the applied software and all chosen parameters in the supplementary methods section of the original publication [46].

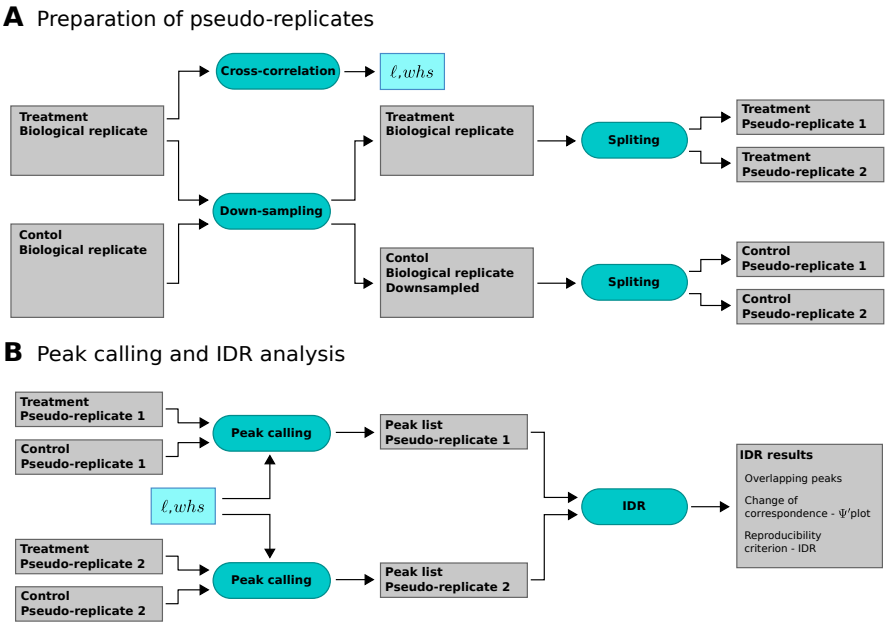


Figure 17: **Reproducibility of peak calling - Comparison framework:** (**A**) The workflow begins with the mapped reads for a ChIP-seq and an associated control experiment given as BAM formatted files. The larger dataset is downsampled in order to obtain the same number of reads for the ChIP-seq and control dataset. Pseudo-replicates are derived by splitting the sets of mapped reads randomly into two halves of equal size. The parameters ℓ and whs are estimated using the cross-correlation method of the SPP package. (**B**) The pseudo-replicates are grouped into two pairs of ChIP-seq and control datasets and used as input for peak calling. The previously estimated parameters ℓ and whs are used in appropriate ways as input for the individual peak callers. The derived peak lists are sorted by significance and only the top $100,000$ peaks are used as input for the IDR procedure.

shows the largest number of overlapping peaks (60,450) as compared to the peak callers MACS2 (46,976), SPP (45,759) and PeakSeq (45,022). Moreover, similar significance scores of overlapping peaks indicate good reproducibility. The Pearson correlation coefficient for the scores of overlapping peaks is 0.97 for Q, whereas for the other three peak callers it is only between 0.88 and 0.90 (Figures 18A-D). Also with respect to the correspondence method, Q outperforms the other three peak callers. The transition from signal to noise occurs at around 35,000 top ranked peaks for Q, whereas for the other three peak callers the transition occurs between 15,000 and 20,000 (Figure 18E). Furthermore, at arbitrary IDR thresholds, the largest number of peaks is selected for Q (Figure 18F). For the recommended default threshold IDR $\leq$ 0.01, 27,284 overlapping peaks are selected for Q, 17,015 for MACS2, 11,618 for SPP and 16,318 for PeakSeq.

Figure 19 shows a summary of the numbers of overlapping peaks for all 38 datasets. Between 5991 and 70,663 overlapping peaks among the top 100,000 peaks derived from the two pseudo-replicates are identified by all four peak callers (Fig. 19A). For 31 datasets, Q shows the largest number of overlapping peaks and, for 33 datasets, the Pearson correlation coefficient for the significance scores is the highest for Q *(Supplemental Table S3 and S4 in [46])*. In order to emphasize the differences between peak callers for given datasets, the mean number of overlapping peaks was subtracted from total numbers of peaks for the individual peak callers. For Q, the deviation from the mean is above zero in all cases (Figure 19B). Altogether, the mean normalized peak numbers are significantly larger for Q as compared to MACS2, SPP and PeakSeq (Figure 19C).

3.2.5.3 *Compatibility with the IDR procedure*

A closer inspection of the results of the IDR procedure revealed that for some datasets overlapping peaks with exceptional small significance scores for both replicates were reported as reproducible regarding a cutoff of IDR $\leq$ 0.01. This observation contradicts one of the basic assumptions of the IDR procedure according to which higher ranked signals are more reproducible than lower ranked signals [73]. Therefore, this phenomenon was further investigated.

For each dataset i, the mean value μ_i of the significance scores of all overlapping peaks that were rejected as irreproducible at a threshold of IDR $\leq$ 0.01 was calculated and used as a reference point (Figure 20A). Furthermore, the significance scores of overlapping peaks classified as irreproducible (IDR $>$ 0.01) and reproducible (IDR $\leq$ 0.01) were plotted separately. For the RNAPII dataset, the vast majority of peaks that were classified as reproducible has significance scores greater than μ_i for both replicates, and only a small dissociated fraction has smaller scores (Figure 20B). Those peaks were defined as consistently weak overlapping peaks (CWOP), and the proportion of CWOP amongst all peaks selected at a threshold of IDR $\leq$ 0.01 was determined for each of the 38 datasets (Figure 20C). For 31 datasets, no CWOP were identified for Q and only for one dataset more than 10%. In comparison to this, there are 3 datasets for SPP with more than 10% CWOP 14 for MACS2 and 13 for PeakSeq.

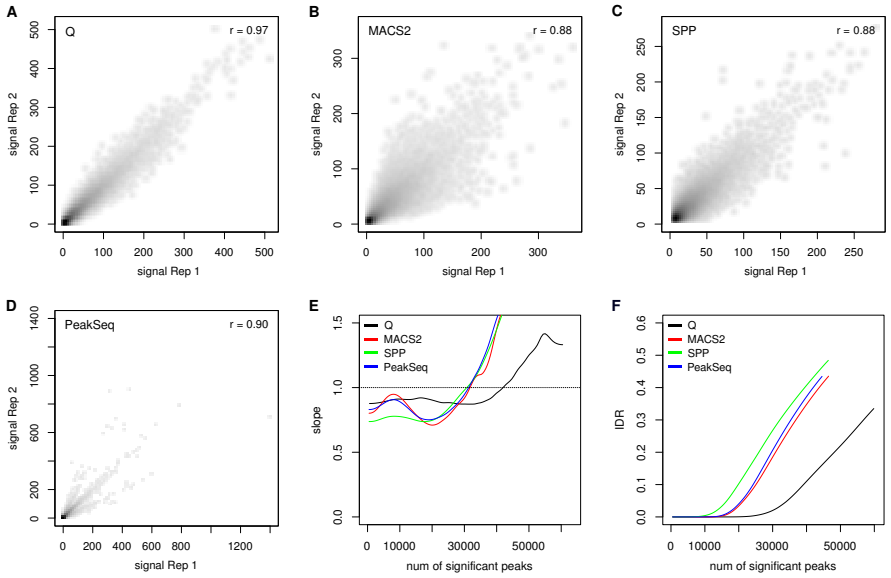


Figure 18: **Reproducibility of peak calling - Detailed results for a RNAPII dataset:** The comparison framework was applied to the peak callers Q, MACS2, SPP and PeakSeq. (**A-D**) Scatterplots of significance scores of overlapping peaks derived from the two pseudo-replicates. For Q (A), MACS2 (B) and PeakSeq (D) the negative decadic logarithm of P-values were used as scores, whereas for SPP (C) the signal values were used. Q shows the largest number of overlapping peaks (60,450) as compared to the three other peak callers (between 45,022 and 46,976). In addition, Q shows the highest Pearson correlation coefficient (0.97). (**E**) Change of correspondence curve (Ψ' plot). The latest transition from signal to noise is observed for Q at around 35,000 peaks. (**F**) For Q, the largest number of peaks is selected at arbitrary IDR thresholds. For the recommended default threshold IDR $\leq$ 0.01, the largest number of peaks is selected for Q (27,284) as compared to the three other peak callers (between 11,618 and 17,015). Find a detailed explanation of the IDR plots in panel E and F in Section 2.2.6. *This figure was originally published in Hansen et al., 2015 [46].*

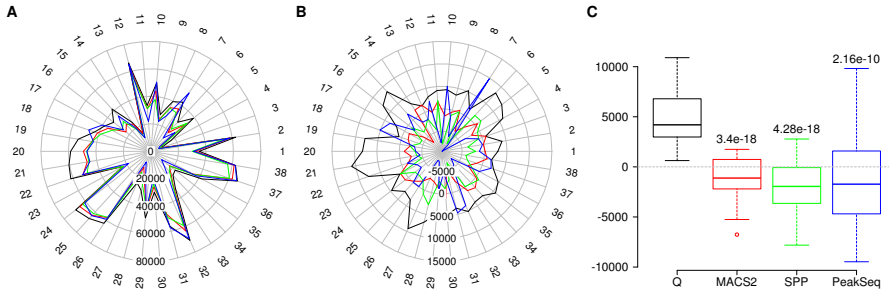


Figure 19: **Reproducibility of peak calling - Summary of overlapping peaks for 38 datasets** (**A**) Numbers of overlapping peaks for Q (black), MACS2 (red), SPP (green) and PeakSeq (blue). The individual datasets are indicated by the numbers around the radar plot *(Supplemental Table S1* in [46]). (**B**) Mean normalized numbers of overlapping peaks (shown in A). For each dataset, the mean of the numbers from all four peak callers was subtracted from the individual total numbers. (**C**) Boxplots of mean normalized peak numbers (shown in B). The P-values were determined with respect to Q using a two-sample, two-sided Wilcoxon tests. *This figure was originally published in Hansen et al., 2015 [46].*

For some datasets, the IDR procedure fails completely when applied in conjunction with PeakSeq, i.e. most of the peaks are classified as irreproducible, and only CWOP are classified as reproducible *(Supplemental Table S5 in [46])*. A detailed examination of these extreme cases revealed that there are many ties, especially at the lower ranks, i.e. overlapping peaks with the same significance scores for both replicates. This problem is due to a too small range of significance scores for weak peaks. The developers of the IDR procedure recommend only ranking systems that produce scores without ties [73]. For this reason, some significance measures are considered to be more compatible with the IDR procedure than others [122]. For instance, the enrichment value of SPP is considered to be well compatible with the IDR, which could be confirmed by the analysis of CWOP presented here.

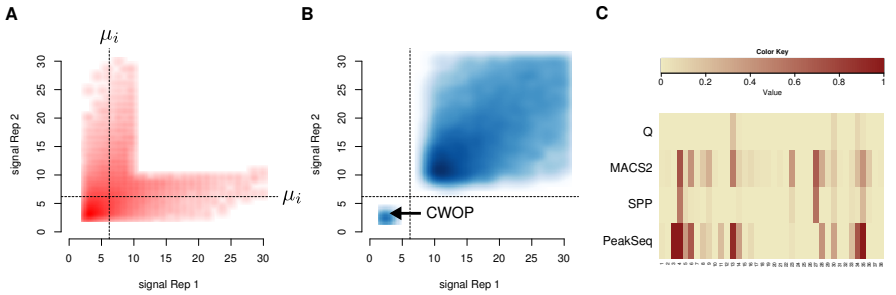


Figure 20: **Reproducibility of peak calling - IDR compatibility:** Scatter plots show the negative decadic logarithm of P-values derived by Q for the RNAPII dataset. (**A**) Overlapping peaks classified as irreproducible (IDR $>$ 0.01; red). The dashed lines represent the mean value μ_i of the significance scores of all overlapping peaks with IDR $>$ 0.01. (**B**) Overlapping peaks classified as reproducible (IDR $\leq$ 0.01; blue). Overlapping peaks with IDR $\leq$ 0.01 and significance scores less than μ_i are defined as consistently weak overlapping peaks (CWOP). Such peaks contradict one of the fundamental assumptions of the IDR procedure, whereby reproducible signals are higher ranked as compared to irreproducible signals. (**C**) The proportion of CWOP was determined for each dataset. Q shows the best compatibility with the IDR procedure with only small proportions of CWOP for a few datasets. Closer inspection of the significance scores of CWOP revealed that the incompatibility is due to ties, especially at the lower ranks. *This figure was originally published in Hansen et al., 2015 [46].*

3.2.5.4 *Results of the IDR procedure for unproblematic datasets*

For 21 datasets, all four peak callers have less than 10% CWOP. In order to eliminate the effect of IDR compatibility, only these datasets were used for downstream analyses. Figure 19 shows the numbers of overlapping peaks selected at a threshold of IDR $\leq$ 0.01. In the majority of cases, Q identifies the largest number of reproducible peaks (Figure 21A). Furthermore, the mean normalized peak numbers are significantly larger for Q as compared to the other three peak callers (Figure 21B,C).

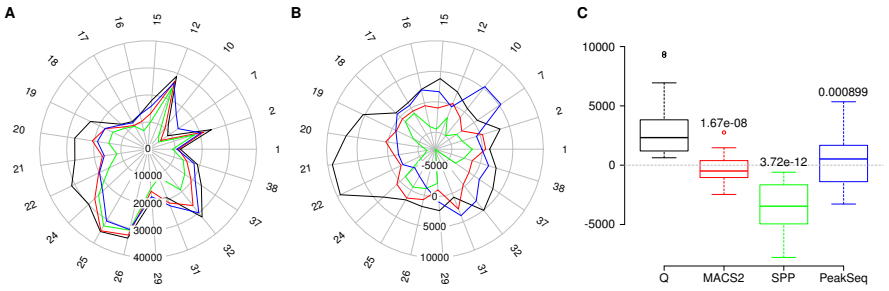


Figure 21: **Reproducibility of peak calling - Summary of overlapping peaks selected at a threshold of IDR $\leq$ 0.01 for 21 datasets:** This figure is analogous to 19A-C except that overlapping peaks were selected according to the reproducibility criterion IDR $\leq$ 0.01. 17 datasets were excluded from the analysis in order to eliminate effects of IDR compatibility. *This figure was originally published in Hansen et al., 2015 [46].*

3.2.6 *Motif content of peaks*

One of the main applications of ChIP-seq is the analysis of transcription factor binding sites (Section 2.2.7). In this section, a comparison regarding motif content of peaks derived by the four peak callers is presented. For this purpose, a framework was developed that aims to highlight only the differences between the peak callers by excluding as many other influencing factors as possible.

In a first step, a set of reference motifs is defined by performing a de novo analysis on the 4-way intersection of the top 50,000 peak sequences derived by the individual peak callers (Figure 22A). The initial peak lists are prepared as for the IDR comparison (Section 3.2.5) but all mapped reads are used for peak calling. The summit positions of all peaks are extended in upstream and downstream direction by a previously estimated fragment length ℓ, and the extended peaks are used to derive the 4-way intersection with bedtools [87]. The genomic sequences between the most upstream and downstream positions of overlapping peaks are extracted and used as input for a de novo motif analysis with DREME [8]. The 10 most significant motifs are defined to be the reference motifs.

In a second step, the number of peaks that contain at least one occurrence of a reference motif is determined for the individual peak lists (Figure 22B). The same peak lists that are used for the determination of the reference motifs are used for this step of the analysis. But this time the summit positions are extended only by $\ell/2$ in upstream and downstream direction. Using the FIMO software [42], the corresponding genomic sequences are extracted and searched for occurrences of the reference motifs in order to identify peaks that contain at least one motif occurrence. Since all lists contain the same number of peaks, the absolute numbers are directly comparable between the individual peak callers. Find a listing of the applied software and chosen parameters in the methods section of the original publication [46].

The comparison framework was applied to the same 38 datasets that were also used for the reproducibility analysis (Section 3.2.5). For 33 datasets, the top 50,000 peaks obtained from Q contain the largest proportion of peaks with at least one occurrence of a reference motif (Figure 23A, *Supplemental*

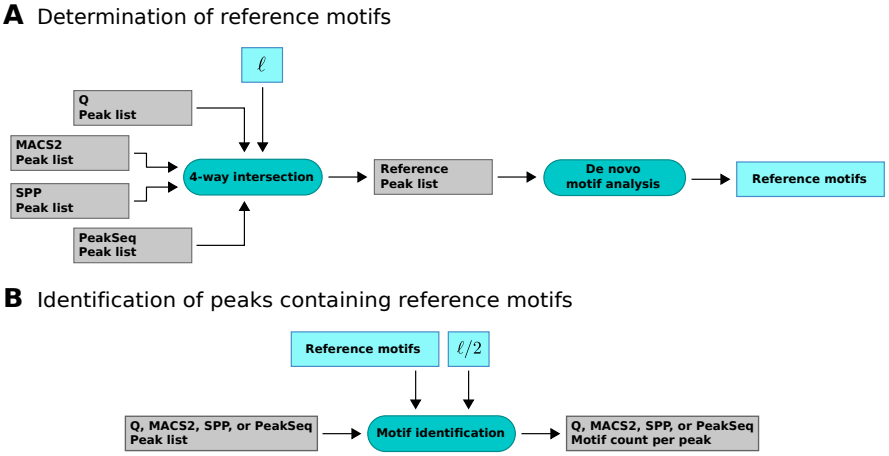


Figure 22: **Motif content of peaks - Comparison framework:** (**A**) Determination of reference motifs. The summit positions of the top 50,000 peaks derived by Q, MACS2, SPP and PeakSeq are extended in upstream and downstream direction by an estimated fragment length ℓ. Peaks consistently identified by all four peak callers are defined as reference peaks. The genomic sequences of the reference peaks are extracted and used as input for a de novo motif analysis. The ten most significant motifs are defined to be the reference motifs. (**B**) Identification of reference motifs. The summit positions of the same peak lists that are used for the determination of the reference motifs are extended by $\ell/2$ in both directions. The corresponding sequences are extracted, and the number of peaks containing at least one occurrence of a reference motif is determined for each peak caller.

Table S5 in [46]). In order to highlight the differences between peak callers, for each given dataset, the mean number of peaks containing at least one reference motif was subtracted from corresponding peak numbers of individual peak callers (Figure 23B). For Q, the mean normalized peak numbers are significantly larger as compared to MACS2, SPP, and PeakSeq (Figure 23C).

3.2.7 *Signature of paused open promoters*

The analyses regarding reproducibility of peak calling (Section 3.2.5) and motif content of peaks (Section 3.2.6) are rather technical and aim at performance comparison between peak callers. In this section, biologically relevant findings that were obtained using Q are presented. In the course of the reproducibility analysis, it became apparent that Q performs particularly well on RNAPII datasets (Figure 21 datasets 19 to 22). Visual inspection of RNAPII peaks at TSS revealed many cases in which Q identifies two peaks directly in upstream and downstream direction of the TSS, whereas the other peak callers only identify a single peak.

For a systematic investigation, TSS flanking double summits (TFDS) were defined as adjacent pairs of peak summit positions that flank TSS and occur within the range of TSS $\pm$1500. Note that these ranges are henceforth are referred to as promoters. Furthermore, the frequencies of TFDS at each position within this range were determined and plotted as histograms. Note that one TFDS consists of two individual summits in upstream and downstream direction of the TSS. Therefore, two bins of the histogram are incremented for each TFDS. The determination and visualization of summit frequencies

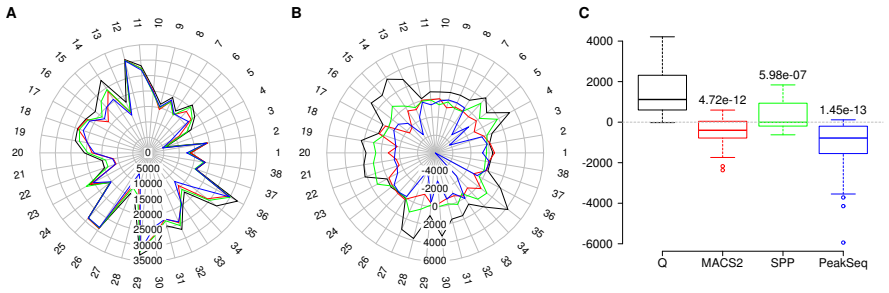


Figure 23: **Motif content of peaks – Summary of peaks with at least one occurrence of a reference motif for 38 datasets:** (**A**) The radar plot shows the absolute numbers of peaks among the top 50,000 that contain at least one occurrence of a reference motif for Q (black), MACS2 (red), SPP (green) and PeakSeq (blue). (**B**) Mean normalized peak numbers. For each dataset, the mean number of peaks with at least one reference motif occurrence was subtracted from the corresponding numbers of the individual peak callers. (**C**) Boxplots for mean normalized peak numbers. The P-values were calculated relative to Q using a two-sample, two-sided, Wilcoxon test. *This figure was originally published in Hansen et al., 2015 [46].*

was done using self-developed PERL and R scripts as well as TSS annotation data from NCBI (build 37.2; NCBI Homo sapiens annotation release 104). To avoid double counting, TSS for which the two promoters (TSS $\pm$1500) overlap were excluded for this analysis which results in 19,722 TSS.

For the analysis of the RNAPII data, the overlapping peaks of the top 100,000 peaks derived from the pseudo-replicates of the reproducibility analysis (Section 3.2.5) were used. The summit positions of overlapping peaks between pseudo-replicates were replaced the by the position in the center between the original summit positions. For the RNAPII datasets, Q identifies at least one summit in 37.4%-42% of all 19,722 non-overlapping promoters and in 39.5%-48.4% of these a TFDS *(Supplemental Table S8 in [46])*. The distribution of TFDS around TSS shows two clearly separated peaks (Figure 24A), one sharp peak 50-100 nucleotides (nt; used synonymously with bp) in downstream direction and a second peak 150-250 nt in upstream direction of the TSS. The median distances between the two peaks in the distribution of TSDS range from 375 to 426 nt *(Supplemental Table S9 in [46])*. Next, the same analysis was performed for datasets derived from ChIP-seq experiments with the histone modification H3K4me3 in HCT-116 and HeLa-S3 cells *(Supplemental Table S1 in [46])* which were processed in the same way as for RNAPII. Q identifies at least one summit in 39.4%-43.2% of all non-overlapping promoters and in 59.2%-70.6% of these a TFDS *(Supplemental Table S9 in [46])*. As for RNAPII, the distribution of TFDS around TSS shows two clearly defined peaks that are separated by a larger distance as compared to those obtained for RNAPII (Figure 24B). The peak in upstream direction of the TSS is slightly sharper and located at a distance of 250-300 nt, and the peak in downstream direction is located at a distance of 300-400 nt. The two peaks in the TFDS distribution are separated by at least 400 nt, and the median distances range from 710 to 778 nt ((*Supplemental Figure S11 in [46])*.

The same TFDS analyses were also performed for the other peak callers *(Supplemental Figures S8-S11 and Tables S8-S9 in [46])*. Even though MACS2, SPP and PeakSeq identify a similar amount of RNAPII and H3K4me3 bound promoters, they fail to produce comparable results due to poor detection

of TFDS. In contrast to that, Q reproducibly identifies TFDS for biological replicates. The overlap of promoters with TFDS between replicates amounts to 78.1% and 82.9% for RNAPII and to 85.3% and 90.6% for H3K4me3 *(Supplemental Table S10 in [46])*.

A combined analysis of TFDS for RNAPII and H3K4me3 revealed a signature that is in line with the notion of paused open promoters with large nucleosome depleted regions (NDR) interspersed with RNAPII and flanked by H3K4me3 modified histones [20, 7, 103]. For the four analyzed datasets, 78.1%-90.6% of the promoters that have a TFDS for RNAPII also have a TFDS for H3K4me3, and for 63.5%-68.7% of these the summits of the TFDS for RNAPII are located within the range between the summits of the TFDS for H3K4me3. In order to decide whether the proportion of promoters showing the signature among those sharing TFDS for RNAPII and H3K4me3 is greater than expected by chance, a simulation study with $10,000$ iterations was performed *(Supplemental Table S11 in [46])*. For the promoters that contain TFDS for RNAPII and H3K4me3 the intervals of TFDS for H3K4me3 were randomly shuffled among the promoters, and the number of occurrences of the paused open promoter signature was determined for each iteration. For no iteration larger overlaps as for the original data were observed (empirical P-value of at most 10^{-4}).

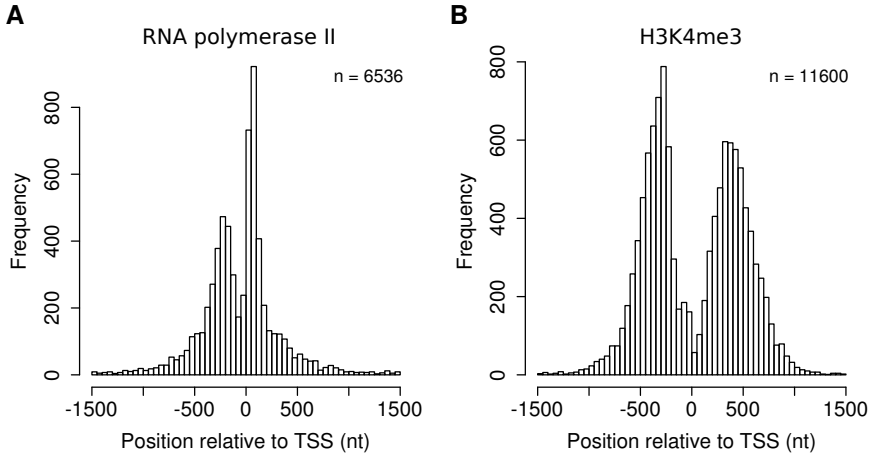


Figure 24: **Signature of paused open promoters - TSS flanking double summits for RNAPII and H3K4me3:** The summit positions of TFDS were integrated over all promoters centered at the TSS. *n* equals the total number of integrated summits. The distributions of TFDS are shown for (**A**) RNAPII and (**B**) H3K4me3. *This figure was originally published in Hansen et al., 2015 [46].*

3.3 DISCUSSION

The methods and software introduced in this chapter provide innovations and improvements at all levels of ChIP-seq peak calling, from the estimation of the fragment length (Section 3.2.1) to the detection (Section 3.2.2) through to the statistical evaluation (Section 3.2.3) of peaks. Q can be executed on desktop computers but also on larger machines and outperformed MACS2, SPP and PeakSeq with respect to runtime (Section 3.2.4). Q was tested and

verified by comparison with the other peak callers regarding reproducibility of peak calling (Section 3.2.5) and motif content of peaks (Section 3.2.6). Finally, Q was used to detect a signature of biological relevance, whereby the other peak caller failed to produce similar results (Section 3.2.7).

The predominant fragment length of the sequencing library may vary from one experiment to another and, at the same time, is an essential parameter for ChIP-seq peak calling and downstream analyses (Section 2.2.3). Furthermore, the quantification of ChIP-enrichment is important for the evaluation of the experiment and experimental trouble shooting (Section 2.2.4). Virtually every peak caller either includes a routine for the estimation of the fragment length or requires this parameter to be specified (Section 2.2.5). The cross-correlation method is a valid method for the estimation of the fragment length and additionally provides quality metrics that reflect the quality of enrichment. Q can be used to produce equivalent estimates of the fragment length and similar results for the evaluation of ChIP-enrichment but three times faster (Section 3.2.4).

The estimated fragment length can be used to enhance the coverage at peaks exploiting the characteristic strand specific distribution of ChIP-seq reads at binding sites (Section 2.2.5), which is often accomplished by shifting or extension of reads (Figure 8). The concept of qfrags (Section 3.2.2) is an innovative further development that involves qfrags coverage profiles which can be searched for peaks as the profiles for shifted or extended reads but have a different depth distribution (Figure 14). At peaks, the qfrag method yields approximately quadratic increase in coverage with respect to the read coverage, whereas the increase is merely linear for the conventional methods.

5′ end positions of mapped reads, here referred to as *hits*, represent the breakpoints in the DNA introduced by shearing of chromatin and thus the outcome of a ChIP-seq experiment, whereas the read or fragment length are experimental parameters that may vary from one experiment to another (Section 1.3). Furthermore, duplicated reads are typically removed in order to avoid PCR-overamplification artifacts and, after this, each position of the genome can be covered by at most one hit (Section 2.2.2). The saturation-based evaluation of peaks (Section 3.2.3) takes these facts into consideration and provides an alternative to the conventional approaches that test peak height for statistical significance using a Poisson or negative binomial distribution [84]. In contrast to height measures, the saturation score is limited to values between 0 and the number of positions within the window that is evaluated, which might mitigate overdispersion effects typically observed for count data derived from NGS applications [126]. The concept of saturation was supplemented by a statistical test modeled within the framework of the classical occupancy problem [37]. Two models for the cases without and with control data were developed. The model for the case with control data also takes into account different sequencing depths for treatment and control data, whereby different amounts of control data have only a marginal effect on P-values (Figure 16).

Q was implemented in a memory efficient fashion and outperformed the peak callers MACS2, SPP and PeakSeq with respect to runtime. The C++ source code was deposited at GitHub along with a tutorial about ChIP-seq data analysis[5]. For a typical ChIP-seq experiment, no parameters need to

[5] http://charite.github.io/Q/

be specified. The fragment length ℓ is estimated from the input data, and an R script is generated that can be executed in order to create the Hamming distance plot (Figure 12) in pdf format. The peak lists are written to standard formatted files so as they can be used for downstream analyses. In addition, Q provides additional output that is useful for documentation, quality assessment and visualization. For instance, Q can be instructed to generate a bedGraph file for the fragment coverage that can be uploaded to UCSC's genome browser [62], which is a popular way of presenting one's own ChIP-seq data in the context of a variety of genomic data.

If high-throughput techniques such as ChIP-seq are applied, reproducibility is an important subject, and the ENCODE project consortium has defined standards concerning this matter (Section 2.2.6). According to these standards, the reproducibility of biological replicates is evaluated using the IDR procedure that is applied multiple times to the original data as well as to differently prepared datasets that are derived from two biological replicates by pooling and random sampling. Since given pairs of pseudo-replicates are derived from the same biological replicate at the level of the mapped reads, the reproducibility analysis cannot reflect experimental differences that occurred before peak calling. Therefore, the results should only reflect differences with respect to peak calling, assuming sufficient sample sizes and validity of the IDR procedure. The reproducibility of peak calling using Q was compared to that of the recommended peak callers MACS2, SPP and PeakSeq using pseudo-replicates derived from 38 individual biological replicates (Section 3.2.5). In most cases, Q identifies the largest number of overlapping peaks and reproducible overlapping peaks selected at a threshold of IDR ≤ 0.01. Furthermore, the correspondence analyses indicate better reproducibility for Q as compared to the three other peak callers. Finally, Q showed the best overall compatibility with the IDR procedure.

Transcription factors bind sequence specific to DNA. Therefore, ChIP-seq peaks are typically enriched for binding motifs (Section 2.2.7). The peak callers Q, MACS2, SPP and PeakSeq were compared with respect to motif content of peaks (Section 3.2.6). To ensure a fair comparison, reference motifs were derived only from peaks that are identified by all four peak callers. Subsequently, the reference motif content was determined for each of the top $50,000$ peaks derived by the individual peak callers. This procedure was applied to the same 38 peak lists that were also used for the reproducibility analysis. In most cases, Q shows the largest number of peaks containing at least one reference motif. Taken together, the number of such peaks is significantly larger for Q as compared to the three other peak callers.

In the course of the reproducibility analysis, it became apparent that Q performs especially well on RNAPII datasets (Section 3.2.5). Visual inspection revealed TSS flanking double summits (TFDS) that were often exclusively identified by Q, whereby the other three peak callers reported only single summits instead of TFDS. Beyond that, a systematic analysis incorporating additional ChIP-seq datasets for the histone modification H3K4me3 revealed a TFDS signature that is consistent with the conception of paused open promoters (Section 3.2.7).

4

SOFTWARE FOR CHIP-NEXUS DATA ANALYSIS

4.1 INTRODUCTION

Detailed investigations of protein-DNA binding architectures require binding site predictions at high resolution (Section 1.3). ChIP-nexus is a further development of the ChIP-seq and ChIP-exo protocol that provides improved resolution as compared to ChIP-seq and addresses shortcomings of ChIP-exo regarding efficiency and monitoring of PCR-overamplification.

In this chapter, the innovations that allow Q to be applied to ChIP-nexus data are described. As a groundwork, a preprocessing module was implemented that takes into account the specific structure of ChIP-nexus reads and makes use of the random barcodes (Section 4.2.1). Based on this, a plot was developed that enables unbiased monitoring of PCR duplication rates (Section 4.2.2). Furthermore, a novel method was implemented that can be used to estimate the width of the protected region (Section 4.2.3), which is a parameter similar to the fragment length in ChIP-seq. The estimated parameter is used as an argument for Q to perform peak calling (Section 4.2.4). The software named Q-nexus [45] along with a documentation and a tutorial is available on GitHub[1]. Finally, Q-nexus was compared to MACS2 and MACE with respect to reproducibility of ChIP-nexus peak calling.

4.2 METHODS AND RESULTS

4.2.1 *Preprocessing and mapping of ChIP-nexus reads*

ChIP-nexus reads have a specific and consistent structure consisting of a fixed and a random barcode, an adapter as well as the fragment sequence originating from the actual biological sample. Just as for many other NGS applications, all sequences unlike the sequences of the fragments have to be clipped off before mapping. The removal of adapters is especially important for ChIP-exo reads because due to the 5′-3′ (λ) exonuclease digest the fragments often become shorter than reads. For ChIP-nexus, the random barcodes require additional efforts because they serve for the identification of PCR duplicated reads and therefore have to be preserved beyond mapping. Alternatively, it is possible to perform the random barcode processing before mapping using an index structure comprising all reads, but this approach was not implemented in Q-nexus.

All tasks arising from the specific structure of ChIP-nexus reads up to the

[1] http://charite.github.io/Q/tutorial_chip_nexus.html

BAM files and read counts that can be used for further processing including monitoring of PCR-overamplification (Section 4.2.2) were accomplished by Benjamin Mencüç under the supervision of Peter Robinson and myself in the scope of a Dr. med. thesis [76].

The workflow for ChIP-nexus preprocessing consists of three steps: removal of barcodes and adapters, mapping and selective removal of PCR duplicated reads (Figure 25). In the first step, the five-nucleotide random barcodes and the adjacent four-nucleotide fixed barcodes at the 5′ ends of reads are clipped off, and the random barcodes are inserted into the ID field of the FASTQ records. Furthermore, the adapters at the 3′ ends of reads are clipped off, and reads without barcodes or with more than one mismatch within the fixed barcode sequence are discarded. The clipping was carried out using FLEXBAR [29] for which some adjustments had to be made in order to remove multiple adapter sequences on either side of the reads and to conserve the information of the random barcodes.

In the next step, the trimmed reads are mapped to the reference genome using bowtie [67] with settings appropriate to ensure that for each read only the best alignment will be reported, whereby the best alignment may also include reads mapping to multiple position. This is an adapted form of a more stringent approach, whereby only uniquely mapped reads are used for downstream analysis because ambiguously mapped reads can lead to false positive predicted binding sites.

In the final step, selective removal of PCR duplicated reads is performed using the information of the random barcodes that was passed on with the read IDs to the BAM files containing the alignments. Multiple reads that map to the same genomic position and have an identical barcode are discarded except for one read. This was implemented in a SeqAn application called NEXCAT that additionally reports duplication levels for different categories of duplicated reads that were subsequently used for further analysis (Section 4.2.2).

Using 10 datasets of the original ChIP-nexus publication (GEO accession code: GSE55306), an average runtime of 17 minutes was determined using four threads on a desktop computer with an Intel® Core™ i7-3770 (3.4 GHz) processor.

4.2.2 *Unbiased monitoring of PCR-overamplification*

The plot for sequence duplication levels of the FastQC package [4] is often used to assess PCR-overamplification. This plot shows the distribution of percentages of reads with a given degree of sequence duplication. For instance, a value of 10% at the duplication level 2 means that 10% of all reads exist in exact two copies.

With the ChIP-nexus protocol random barcodes were introduced that allow for selective removal of PCR-duplicated reads (Section 4.2.1) and, additionally, provide the opportunity to monitor PCR-overamplification in an unbiased fashion. In this section, the reconstruction of FastQC's plot for

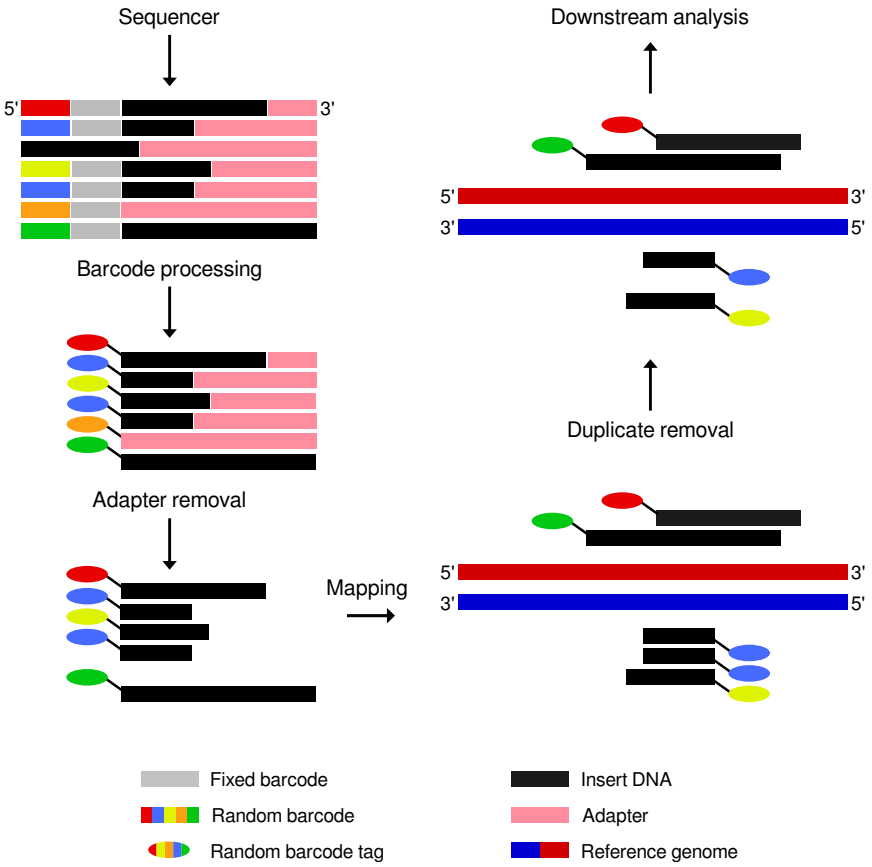


Figure 25: **Preprocessing of ChIP-nexus reads:** The workflow consists of three steps: At first, the barcode (varicolored, gray) and adapter (pink) sequences are removed while random barcodes are preserved. Faulty reads such as those consisting of adapter sequence only are filtered out (left). Subsequently, the trimmed reads (black) are mapped to the reference genome using a read mapper such as bowtie with appropriate parameter settings (middle). Finally, PCR-duplicated reads are selectively removed in consideration of random barcodes (right). *This figure was originally published in Hansen et al., 2016 [45].*

sequence duplication levels is presented. The percentages of the various levels are calculated with and without consideration of random barcodes. Duplicated reads can be identified before mapping on the basis of sequence identity only but this requires an index structure on all reads, which comes at the cost of high memory requirement. Alternatively, duplicates can be identified after mapping on the basis of identical mapping positions. Since the reads have to mapped either way, the latter approach was implemented in Q-nexus.

Without additional information the set of all mapped reads can be divided into two categories: unique reads that are mapped to genomic positions to which no other read can be mapped, and other duplicated reads, which are mapped to the same position as at least one other read. The latter are here referred to as identically mapped reads (IM), which can be further subdivided into levels, e.g. all reads that are mapped to positions to which exactly one other read is mapped belong to level 2 (Figure 26A). Unique reads belong to level 1. IM reads correspond to the conventional duplicated reads, and the percentages obtained for the various levels conform to those in the original plot of FastQC.

Using the information of the random barcodes, a subcategory of IM reads can be identified consisting of reads that are mapped to the same genomic position as at least one other read that additionally has an identical barcode. Those reads are here referred to as identically mapped reads with identical barcode (IMIB). For IMIB reads it can be assumed that they originate from from PCR-overamplification, and the subdivision into levels yields an unbiased equivalent of the conventional plot for sequence duplication levels. The remaining IM reads are here referred to as identically mapped reads with unique barcode (IMUB). In contrast to IMIB reads, IMUB reads can be assumed to originate from different fragments. Therefore, only those are used together with unique reads for downstream analysis.

A quality metric that is often used in order to assess complexity of sequencing libraries is the overall duplication level calculated as the proportion of duplicated reads amongst all reads [4]. The method introduced here provides separate overall duplication levels for IM, IMIB and IMUB reads defined as proportion reads of the respective category within the levels greater or equal 2 amongst all mapped reads.

The sequence duplication level plots were derived [45] for eight datasets of the original ChIP-nexus publication [47]. The conventional overall duplication levels are between 54% and approximately 100% , which is exceptionally high. For instance, at a threshold of 20% the module of FastQC issues a warning, and at a threshold of 50% datasets completely fail this quality check. Furthermore, also the unbiased overall duplication levels between 42% and 99% are very high, and only one dataset would pass the quality check. The progression of the duplication level curves is as expected (Figure 26B-C). For high overall duplication levels, there is a shift for IMIB reads towards the lower levels with respect to IM reads.

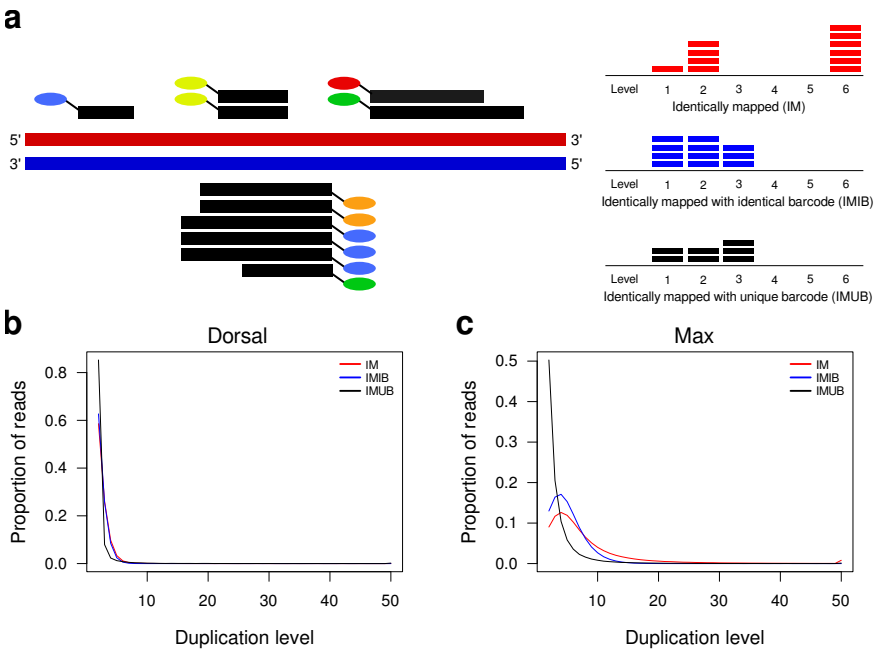


Figure 26: **Unbiased monitoring of PCR-overamplification:** (**A**) Demonstration example of mapped reads tagged with random barcodes and the related level counts for IM (red), IMIB (blue) and IMUB reads (black). The number of horizontal bars for a given level correspond to the number of reads with this level of duplication. A more detailed example and an extensive explanation can be found in *Additional file 1: Figure S1* of the original publication. (**B,C**) Duplication level plots for Dosal and Max. For IM reads, i.e. without using random barcode information, the overall duplication level is 54% for Dorsal and 95% for Max. For IMIB reads, overall duplication levels of 50% and 95% were determined. Additional plots for the other datasets can be found in *Additional file 1: Figure S2* of the original publication. *This figure was originally published in Hansen et al., 2016 [45].*

4.2.3 *Protected region width estimation*

Another challenge in the analysis of ChIP-exo and ChIP-nexus data is that the average fragment length, which is a crucial parameter for ChIP-seq, is no longer a relevant factor. For ChIP-exo and ChIP-nexus, the equivalent of the average fragment length is the width of the region occupied by the target protein or an associated complex, which is here referred to as "protected-region width". Such regions typically have a width of 6-20 bp, which is much shorter than the fragment lengths observed for ChIP-seq (90-230 bp).

The curves produced by the cross-correlation method (Section 2.2.3) as well as by its variation using Hamming distance (Section 3.2.1) are compromised by an artefactual peak at one read length termed "phantom peak" [66]. This peak has already been associated with repetitive regions, and the influence of the exclusion of known artifact regions using so called blacklists was investigated [21]. Such regions often occur on non-canonical chromosomes, which serve as reservoir for scaffolds partly of low quality that could not unambiguously assigned to a chromosome arm. For instance, for Drosophila melanogaster (dm3) these chromosomes are referred to as U and chrUextra [50]. To prevent complications, reads that can be mapped to non-canonical chromosomes are often excluded from the analysis. In general, the most generic approach to correct for such artifacts is the inclusion of appropriate control data. However, no control data is available for ChIP-exo and ChIP-nexus [108].

For the estimation of the fragment length from ChIP-seq data, the phantom peak does not pose a problem because the second maximum that corresponds to the fragment length occurs in a different range. However, for ChIP-exo and ChIP-nexus the phantom peak masks the range of interest. In this section, a method is presented that can be used to estimate the protected region width denoted as ℓ''', which is subsequently used for the construction of the qfrag coverage profile. This method is similar to the cross-correlation method inasmuch that it operates genome-wide and has no assumptions about binding sites of the target protein, but it circumvents the problem with the phantom peak.

For a first investigation, the cross-correlation method was applied to all datasets. Without exception, the curves are strongly dominated by the phantom peak, and the maximum is at one read length. The removal of reads mapping to the non-canonical chromosomes U and Uextra has only a minor effect (*Additional file 1: Figure S5* in [45]).

Furthermore, the distributions of 5′ end positions of mapped reads around preselected sites were determined for the ChIP-nexus datasets, which is a popular way of presenting ChIP-exo data [104] that was also used for the original ChIP-nexus publication. The approach turned out to be unstable when using standardized parameter settings for all samples. Only for six out of ten datasets useful results are obtained (*Additional file 1: Figure S4* in [45]). Furthermore, the derived distributions slightly differ from those shown in the original ChIP-nexus publication [47]. The results heavily depend on the motif used for selection and centering of peaks and could possibly be improved using prior knowledge about the preferred motifs of the analyzed

factors. However, this was not done because the focus of this investigation was on the unbiased and automated estimation of the protected region width.

The method introduced here is based on the concept of qfrags (Section 3.2.2). The empirical distribution of qfrag-lengths is derived from the data by counting the number of qfrags for given lengths, and the qfrag-length with the highest count is determined. The basic idea is similar as for ChIP-seq. At at binding sites, 5′ end positions of reads mapped to the forward and reverse strand are likely to occur at a distance that is related to the width of the protected region (Figure 27A).

The qfrag-length distribution method was applied to the ten ChIP-nexus datasets, and it turned out that it is also affected by the phantom peak although to a lesser extent (*Additional file 1: Figure S6* in [45]). In six cases, the qfrag-length with the highest count is smaller than one read length, and after removal of reads mapping to non-canonical chromosomes there are three more such cases. However, given the strong bias introduced by the phantom peak, the method needed to be further improved.

Given the strand specific distribution of 5′ end positions at binding sites and the operating principle of the cross-correlation method (Section 2.2.3), it appeared plausible that the phantom peak corresponds to clusters of reads that map to the same region with a length of about two times the read length but to the forward and reverse strands so as most of the 5′ positions on different strands occur at a distance of about one read length. For a closer examination, a standard Q peak calling was performed with the parameters ℓ set to one read length and $x = 5$, which is appropriate for the identification of such regions. Visual inspection of the peaks indeed revealed a number of such clusters (Appendix A3).

With this in mind, a pseudo-control was developed that can be used to subtract the proportion of signal caused by the artifacts within the range of interest. The pseudo-control is derived from the original data by switching the strands of all reads and, subsequently, shifting the 5′ end positions by one read length towards 3′ direction (Figure27B). The qfrag-length distribution is then derived from the original data as well as from the pseudo-control (Figure 27B), and the counts for the pseudo-control are subtracted for each qfrag-length (Figure 27C). The qfrag-length ℓ''' with the largest difference in counts between original data and pseudo-control is taken as the estimated width of the protected region.

The qfrag-length distribution method presented here was implemented as an additional module of the Q software (Section 3.2.4), and the differences between qfrag-length counts for original data and pseudo-control were derived for all datasets (*Additional file 1: Figure S6* in [45]). For the cases in which a usable footprints of 5′ end positions around predefined binding sites is available, the qfrag-length method reproducibly produces consistent estimates for ℓ''' (Figures 28A-D). For the other cases, the estimates are also reproducible and within the expected range (6-20 bp) (Figures 28E-F). Finally, the curves show characteristic progressions for the individual factors (Figures C,D and G,H).

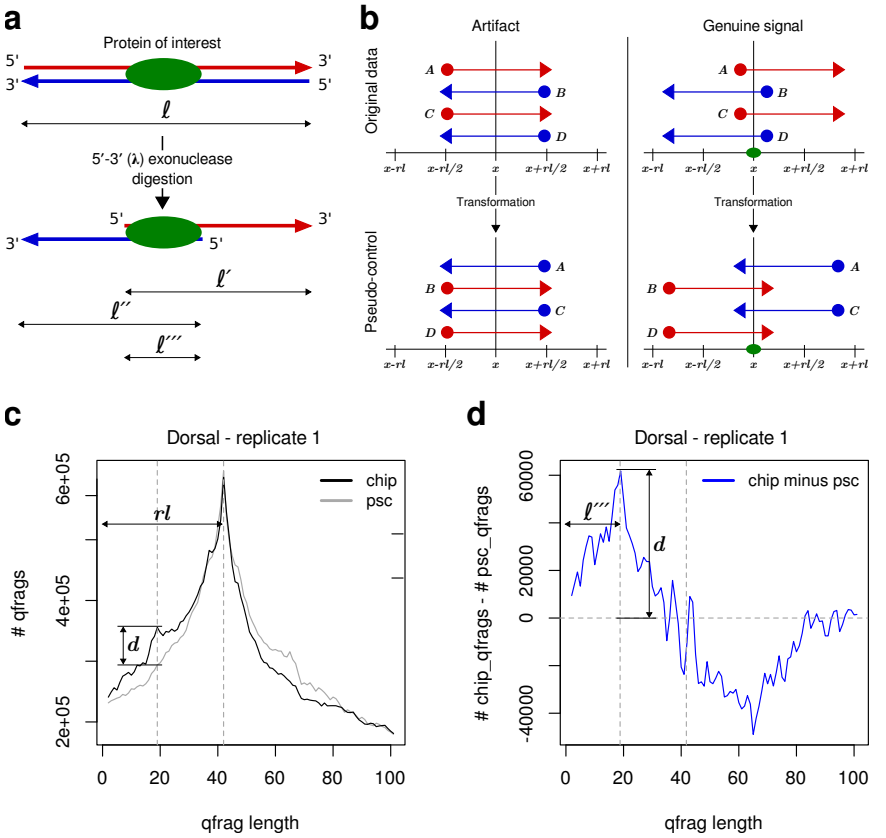


Figure 27: **Protected region width estimation - Methods:** (**A**) Schematic representation of a fragment of length ℓ that is bound to a protein of interest. 5′-3′ (λ) exonuclease is employed to digest the 5′ ends to within a few bp at which the protein is bound, which can result in a fragment of length ℓ' or ℓ''. The relevant length for ChIP-exo and ChIP-nexus is ℓ''', which corresponds to the width of the region that is protected from digestion. (**B**) Schematic representation of a mapping artifact most likely causing the phantom peak (top left) and a genuine ChIP-nexus peak (top right). The original data is transformed into a pseudo-control by switching the strands of all reads and shifting the 5′ end positions by one read length towards 3′ direction. The mapping artifact remains in the pseudo-control (bottom left), whereas the 5′ end positions of genuine peaks are shifted outwards so as the relevant range (6-20 bp) becomes accessible (bottom right). (**C**) qfrag-length distribution derived from a Dorsal dataset with a pronounced phantom peak at one read length (black) and from the corresponding pseudo-control (gray). The second local maximum at a qfrag-length of 19 bp corresponds to the actual signal. The difference between the qfrag-length counts for the original data and for corresponding pseudo-control data is depicted as d, which will be reused in (D) for didactic purposes. (**D**) Difference between qfrag-length counts for original and pseudo-control data. The qfrag-length with the largest difference is defined to be the estimated width of the protected region ℓ''' in this case 19. *This figure was originally published in Hansen et al., 2016 [45].*

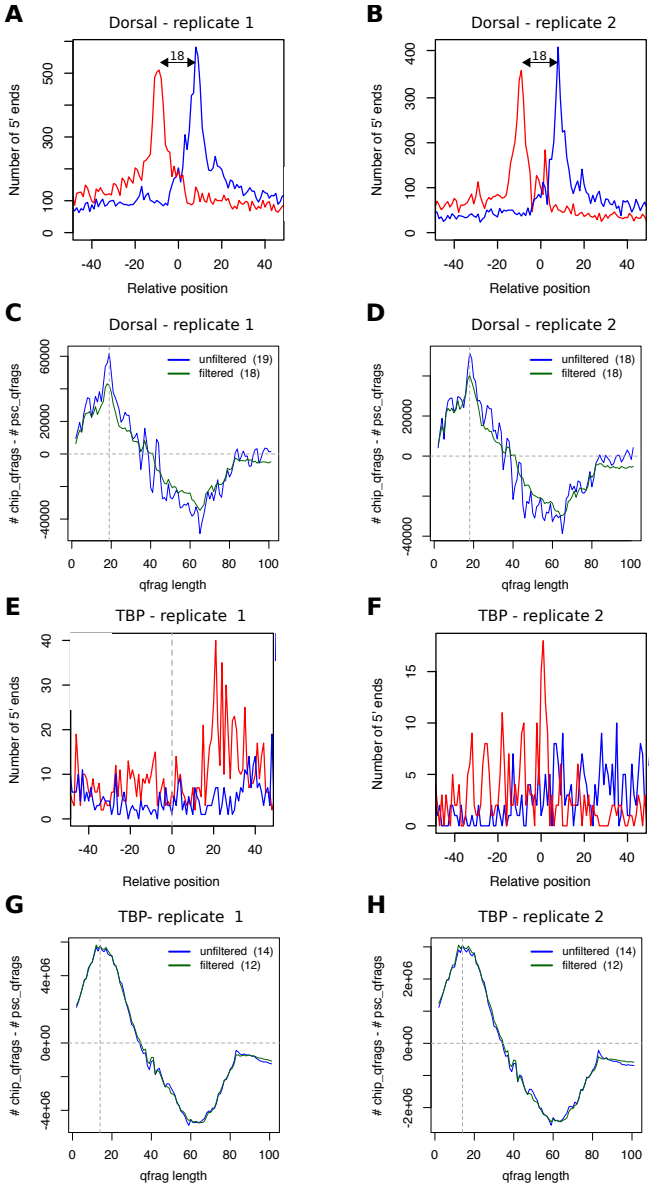


Figure 28: **Protected region width estimation - Results:** (**A-B**) Footprints of 5′ end positions of mapped reads around preselected sets of binding sites for two biological replicates of ChIP-nexus experiments with the Dorsal factor. The distance between the two peaks on the forward and reverse strand is line with a previous analysis of the same data [47]. (**C,D**) Qfrag-length counts after subtraction of the pseudo-control for the same data that was used for (A,B). The analysis was performed before (blue) and after removal of reads mapping to non-canonical chromosomes (green). The estimated parameter $\ell''' = 18$ is consistent with the footprint of 5′ positions. (**E-H**) The same analyses were performed as for A-D but for the TATA-binding protein (TBP). Using standardized parameter settings, no useful footprints of 5′ end positions can be obtained in this case. The qfrag-length distribution method reproducibly yields estimates that are plausible from a biological point of view, and the curve progressions are clearly different from those obtained for Dorsal (C,D). *This figure was originally published in Hansen et al., 2016 [45].*

4.2.4 *ChIP-nexus peak calling*

The Q software was extended by a nexus mode (Figure 29) that is optimized for the analysis of ChIP-exo and ChIP-nexus data. If Q is executed in this mode, and except the data no further parameters are specified, duplicates are not removed, and the width of the protected region ℓ''' is estimated using the qfrag-length distribution method (Section 4.2.3) instead of the Hamming distance method (Section 3.2.2). Furthermore, a simpler model is used for the evaluation of peaks in place of the saturation-based approach. The software was released under the name Q-nexus.

The qfrag coverage profile is constructed and searched for peaks as for Q (Section 3.2.2). However, due to the increased resolution of ChIP-nexus, the previously estimated ℓ''' instead of ℓ and a much smaller argument of $x = 5$ for the allowed deviation from ℓ''' are used. Positions covered by at least one qfrag and with no higher qfrag depth at a distance of q_{min} are predicted as binding positions and the surrounding regions $\pm q_{max}$ are subjected to statistical evaluation.

For preprocessed ChIP-nexus reads (Section 4.2.1) it can be assumed that they are free of PCR duplicates, and reads that map to the same position constitute a large portion of the signal. For the saturation approach (Section 3.2.3) it makes no difference if one or more reads map to the same position. Therefore, it is not suitable for the evaluation of ChIP-nexus peaks. Instead, for each summit s_i, the number of 5′ end positions of mapped reads within the range $s_i - q_{max}, ..., s_i + q_{max}$, denoted as k, is determined. The P-values are calculated using a Poisson distribution $\text{Pois}(k, \lambda)$ assuming that 5′ end positions are evenly distributed across the genome, i.e. $\lambda = 2 \cdot q_{max} \cdot (|T_f| + |T_r|)/l$, where $|T_f| + |T_r|$ is the total number of 5′ end positions and l the length of the genome. Finally, the Benjamini-Hochberg procedure is used to correct for multiple testing (Section 2.2.5).

4.2.5 *Reproducibility of ChIP-nexus peak calling*

Q-nexus was compared to MACS2 [38] and MACE [118] with regard to reproducibility, because MACS2 was also used in the original ChIP-nexus publication [47] and MACE [118] was the only publication mentioning the analysis of ChIP-nexus data at that time.

The comparison framework using the IDR procedure (Figure 17) was reused but with biological replicates in place of technical replicates. Furthermore, instead no uniform argument for the protected region width and similar parameters for the other two peak callers were used. This was done because the estimation of the width of the protected region (Section 4.2.3) is one of the main innovations in Q-nexus.

Apart from that, MACS2 and MACE were run with appropriate parameter settings. For instance, the option `--call-summits` was set to instruct MACS2 to search primary peak regions post hoc for *sub peaks*, which is recommended for the detection of adjacent binding sites [38], or the feature for nucleotide composition bias correction of MACE was switched off because it

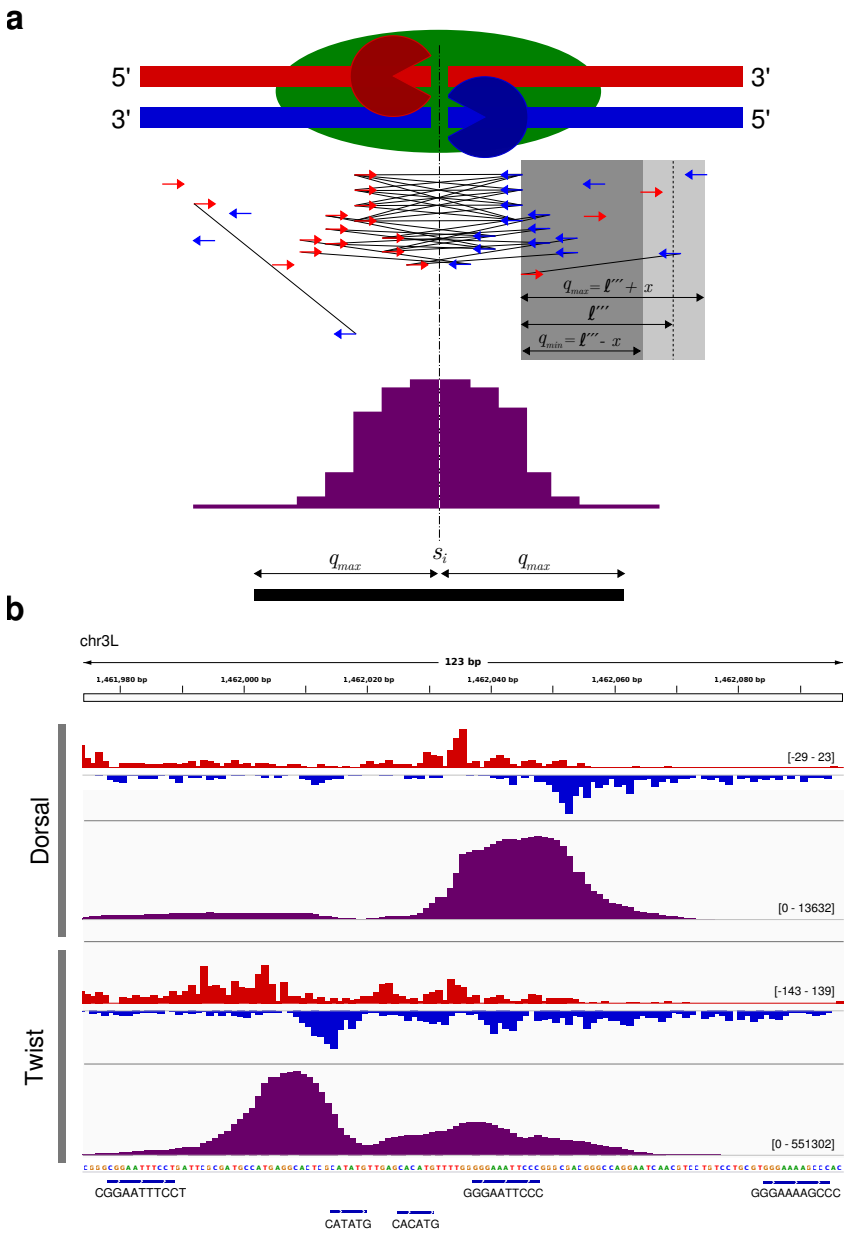


Figure 29: **ChIP-nexus peak calling:** (**A**) Description of the Q-nexus workflow. For ChIP-nexus, 5′-3′ (λ) exonuclease ('Pac-Man' icons) is employed to digest the 5′ ends of fragments up to about 5-6 base pairs to the location at which target protein (green) and DNA were cross-linked. The mapped reads (red and blue arrows) correspond to 5′ ends of fragments. If Q is executed in nexus mode, duplicates are not removed. The qfrag-length distribution method (Figure 27) is used to estimate the length of the protected region (ℓ'''), which is used along with a default argument of $x = 5$ to construct qfrags (gray box and black lines). Summit positions in the qfrag coverage profile (purple) are defined as for Q (Section 3.2.2). For each summit position s_i, the number of 5′ end positions within the region $\pm q_{max}$ is tested for statistical significance using a Poisson distribution. (**B**) Coverage profiles for 5′ end positions (red and blue) and qfrags (purple) at the rho NEE enhancer for Dorsal and Twist. The same region is shown in the original ChIP-nexus publication [47]. The qfrag coverage profile shows two distinct peaks for Dorsal and Twist. *This figure was originally published in Hansen et al., 2016 [45].*

is not appropriate for ChIP-exo and ChIP-nexus data[2] and, additionally, lead to a significant loss of reads utilized for the analysis.

The peak callers were applied to the BAM files (Section 4.2.1) for the ChIP-nexus replicates for Dorsal, Twist, Max, Myc and TBP. Without exception, Q-nexus identifies significantly more overlapping peaks than MACS2 or MACE (Figures 30A-C,E). This also applies to numbers of selected peaks at a threshold of IDR ≤ 0.01 (Figure 30D,E,G). MACS2 reproducibly identifies peaks for all analyzed factors and also performs well in conjunction with the IDR procedure. However, to a much lower extend as compared to Q-nexus. For MACE, there are many overlapping peaks with identical significance scores at the lower ranks (Figure 30C) that must lead to ties known to be incompatible with the IDR procedure (Section 2.2.6), which could be an explanation for the poor performance in conjunction with IDR procedure.

4.3 DISCUSSION

The Q-nexus software introduced in this chapter was the first comprehensive software package for the analysis of ChIP-nexus data. Furthermore, a detailed duplication analysis was performed making use of the random barcodes. Beyond that, Q-nexus implements a novel method for the estimation of the protected region width that yields estimates which are plausible from a biological point of view and consistent previously published results. Using the estimated parameter along with parameter settings optimized for ChIP-nexus, Q-nexus outperformed MACS2 and MACE with respect to reproducible peak calling.

The preprocessing of ChIP-nexus reads (Section 4.2.1) was implemented by Benjamin Mencüç as part of a Dr. med. thesis [76]. The focus of this work was on efficiency which is why there was also active communication with the developers of the adapter clipping software FLEXBAR [29] and the C++ library SeqAn [30]. The adjustments that were made to FLEXBAR initiated further developments whose details[3] are beyond the scope of this thesis. The software permits fast processing of ChIP-nexus reads, which includes clipping of barcodes and adapters, mapping to a reference genome, selective removal of PCR duplicated reads and calculation of sequence duplication levels. In contrast to the preprocessing performed for the original ChIP-nexus publication using a number of scripts and available tools [47], Q-nexus comes with a clearly defined user-interface and a detailed documentation including a tutorial[4].

Besides the fact that random barcodes allow selective removal of PCR-duplicated reads, they enabled unbiased monitoring of PCR-overamplification of ChIP-exo reads for the first time (Section 4.2.2). A sequence duplication

[2] The nucleotide composition bias correction is intended to correct for bias emerging from random hexamer priming [118] which is performed for RNA-seq [44] but neither for ChIP-exo nor for ChIP-nexus.

[3] https://github.com/seqan/flexcat

[4] http://charite.github.io/Q/tutorial_chip_nexus.html#preprocessing

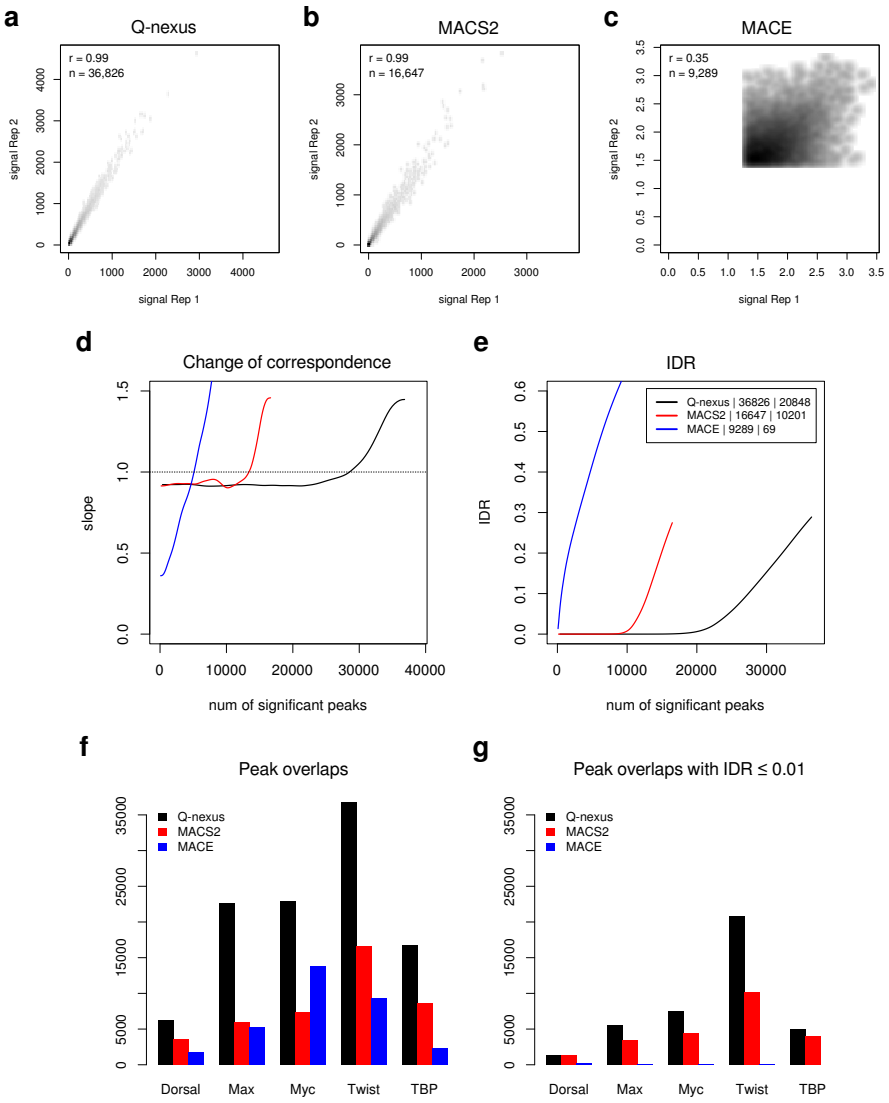


Figure 30: **Reproducibility of ChIP-nexus peak calling:** The peak callers Q-nexus (black), MACS2 (red) and MACE (blue) were compared with regard to reproducibility of peak calling using the same comparison frame work as for ChIP-seq (Section 3.2.5) but biological instead of technical replicates were used. Furthermore, no uniform argument for the protected region width and similar parameters for MACS2 and MACE was used, but these parameters were estimated by the corresponding internal routines of the individual peak callers. (**A-C**) Scatterplots of the significance scores of overlapping peaks amongst the top 100,000 peaks derived for the two replicates of the transcription factor Twist. The Pearson correlation coefficients and the numbers of overlapping peaks are indicated in the upper-left corners of the plots. Q-nexus identifies more than twice as much overlapping peaks as compared to MACS2. (**D**) Change of correspondence curve (Ψ' plot). The latest transition to a segment with positive slope is observed for Q-nexus indicating higher reproducibility as compared to the two other peak callers. Find a detailed description of this and the next plot in (Section 2.2.6). (**E**) Numbers of selected peaks at various IDR thresholds. For Q-nexus, the largest number of reproducible peaks is selected at arbitrary cutoffs. For the default threshold IDR $\leq$ 0.01, Q-nexus identifies about twice as much peaks as compared to MACS2. (**F, G**) Similar results as shown in A-E were obtained for the other factors Dorsal, Max, Myc and TBP. The two bar charts show the numbers of overlapping peaks (F) and numbers of overlapping peaks selected at IDR $\leq$ 0.01 (G). *This figure was originally published in Hansen et al., 2016 [45].*

plot adapted to ChIP-nexus was developed. The duplication levels are determined for three categories of reads: PCR duplicated reads, identical reads originating from different fragments and the union of both (Section 4.2.2). Furthermore, the overall duplication levels for the three categories are calculated. The provided framework for duplication analysis in consideration of random barcodes may be useful for the evaluation of experiments and troubleshooting. For all analyzed ChIP-nexus datasets [47], high overall duplication levels were observed, whereby the differences between the conventional and unbiased overall duplication levels were only marginal. In general, high duplication levels indicate libraries of low complexity due low amounts of input DNA.

For ChIP-exo and ChIP-nexus the equivalent of the average fragment length (ℓ) is the width of the occupied region (ℓ''') which is protected from 5'-3' (λ) exonuclease digest. The estimated parameter ℓ''' is intended to maximize the number qfrags that are used to construct the qfrag coverage profile which is subsequently searched for peaks. Regions with clusters of reads on the forward and reverse strand whose 5' end positions occur at a distance of about ℓ''' will be selectively emphasized (Figure 29). The default argument for the allowed deviation from ℓ''' is $x = 5$, and smaller arguments can be used in order to increase specificity. Peaks in the qfrag coverage profile are identified as for Q, but for Q-nexus a different measure of enrichment is used. The saturation approach is not suitable for ChIP-nexus data, because multiple 5' end positions at individual genomic positions form a part of the signal. Therefore, the number of hits that map to given peak regions is determined and tested for statistical significance using a Poisson distribution.

One of the key innovations of Q-nexus is the estimation of the parameter ℓ'''. As the cross-correlation method, the qfrag length-distribution accumulates strand specific signal characteristics from all over the genome including regions prone to mapping artifacts. As a result, the relevant range (6-20 bp) in the qfrag-length distribution is masked by the phantom peak. The pseudo-control makes this range accessible. However, the drawback of this approach is that the clusters of reads at genuine binding sites introduce new bias to a range upwards of one read length that is difficult to delimit (*Additional file 1: Figure S4* in [45]). One advantage of this approach is that it does not require the exclusion of genomic regions prone to mapping artifacts using blacklists that are not available for all species and genome builds. Furthermore, no prior knowledge about the binding sites of the target protein is required as it is the case for the conventional footprints of 5' end positions at preselected sites. Interestingly, the curve progressions for the differences between qfrag-length counts of the original data and pseudo-control reproducibly show characteristic signatures for the individual factors. However, these plots have to be carefully interpreted for the reasons discussed above. The aim was to estimate a parameter ℓ''' that can be used for the construction of the qfrag coverage profile, and the length that maximizes the number of qfrags seemed to be a good choice. For the analyzed ChIP-nexus datasets, the estimated widths are between 9 and 18 bp, which is within the expected range from a biological point of view. Furthermore, the estimates are by and large consistent with the footprints of 5' end positions centered at predefined

sets of binding positions (See *Table 3* in [45]).

Another question is how well the qfrag coverage profile for a given ℓ reflects the binding architecture of a given target protein. The notion that binding positions are flanked by two pile-ups of reads on the forward and reverse strand is a simplification that is not applicable to all cases. At binding positions, the succession of exonuclease stop positions on the forward and reverse strand depends on the preferred cross-linking configurations of the target protein or complex as well as on proteins that co-bind to DNA nearby. Furthermore, there may be subpopulations of binding sites that show different architectures [104]. In this context, the qfrag-length distribution may represent a mixture of architectures, and a single qfrag coverage profile might not be sufficient in order to draw meaningful conclusions. In such cases, it might be useful run Q-nexus peak calling with differently specified ℓ''' and smaller x.

The reproducibility of Q-nexus peak calling was compared to that of MACS2 and MACE using the biological replicates of the original ChIP-nexus publication, whereby the estimation of the protected region width or comparable parameters for MACS2 and MACE was left to the build-in modules individual peak callers. Q-nexus outperforms the MACS2 with respect to the number of reproducible peaks selected at a threshold of IDR ≤ 0.01, and MACE completely fails to identify reproducible peaks. The poor performance of MACS2 and MACE can be explained by the estimation of inappropriate arguments for the extension size and border pair size, which are the equivalents of the protected region width. For MACS2 as well as MACE, the estimates are shifted towards one read length that is 42 bp in this case (See *Table 3* in [45]).

For MACS2, the smoothing effect on the coverage profile increases with the extension size. If the extension size exceeds the width of the protected region, the prediction of binding sites will start become imprecise. Moreover, adjacent biding sites will start to get merged, if the extension exceeds the distance between them. Finally, depending on the overall noise level, at some point flanking noise reads will start to overlap peak regions. However, it has to be said that the estimates of MACS2 could be possibly improved by using a more suitable bandwidth parameter which is used for the preselection of regions used for the estimation [38], but this would require prior knowledge.

For MACE, the small number of overlapping peaks can be explained similarly. MACE estimates optimal border pair size that is conceptually similar to the width of the protected region [118]. Furthermore, a border detection (peak calling) is performed for both strands separately, and forward and reverse strand peaks at a distance of about the border pair size are matched. Using the inappropriate border pair size, MACE fails to detect reproducible peaks. Apart from that, MACE seems to use an internal lower significance threshold that cannot be overridden by users[5]. This may lead to ties at the lower ranks that are known to be incompatibility with the IDR procedure.

[5] For the comparisons the user defined threshold was released by setting a P-value cutoff of 0.99.

5

FINAL DISCUSSION

Genetics is a fundamental part of biological research, and epigenetics shifts the focus beyond the sequence level to the configuration of DNA that determines the transcription levels of genes within given cells. The accessibility and three-dimensional layout of genomic DNA are important modulators of gene expression that involve modifications of DNA and proteins that interact with DNA. The technology of DNA sequencing has become a popular and valuable instrument that enables researches to answer a variety of related questions. The next-generation sequencing application ChIP-seq can be used for the genome-wide identification of sites that interact with target proteins including key players such as transcription factors and histone proteins.

This bioinformatic thesis is about the primary analysis of ChIP-seq and ChIP-nexus data. The main ideas and concepts emerged from the practical application of bioinformatic software and guidelines for the assessment of data quality and reproducibility (Chapter 2). A thorough study of the standards for ChIP-seq data analysis as defined by the ENCODE project consortium was done in collaboration with scientists performing ChIP-seq and DNA sequencing, and a local instance of the GALAXY platform was setup that allowed for primary data analyses compliant with the standards. This infrastructure was of mutual benefit. On the one hand, the users were enabled to perform this part of their investigation autonomously, which resulted in greater transparency and a better ability to interpret the results [53]. On the other hand, the steady and active communication with the users contributed to a deeper understanding of the data and the additional resources could be used to focus on bioinformatic questions [46, 45]. The analysis of the HoxD13^{Q317K} mutant represents an application example of ChIP-seq that could be used in this case to characterize and compare global binding properties of wildtype and mutant proteins, which contributed to the elucidation of the pathomechanism underlying a phenotype with severe hand and foot malformations.

For this thesis, a ChIP-seq peak caller named Q was developed (Chapter 3) addressing shortcomings identified during practical applications of existing software. Three methodological innovations were implemented in Q: the estimation of the fragment length using Hamming distance (Section 3.2.1), the concept of qfrags (Section 3.2.2) and the saturation-based evaluation of peaks (Section 3.2.3).

The method for the estimation of the fragment length largely recapitulates the cross-correlation plot, but Hamming distances instead of Pearson correlation coefficients are used in order to measure similarity between shifted strands. The implementation in Q yields equivalent estimates of the fragment length but three times faster as compared to the conventional method

implemented in the SPP package [63].

The degree of ChIP-enrichment is an important quality feature of ChIP-seq experiments. The RSC that can be derived from the cross-correlation curve is a well accepted metric. The RSC derived from the Hamming distance curves are comparable for high to moderate degrees of enrichment. In addition, Q reports own enrichment metrics. However, RSC values and Q's enrichment scores are influenced by a number of factors including artifact regions in which the mapped reads form large clusters (Figure 2) as well as details of data processing such as removal of duplicated reads [21]. Furthermore, there is also a dependency on the number of peaks. Factors that tend to bind at many sites in the genome will yield better enrichment scores as compared to others. Therefore, these metrics are best suited for comparisons of replicates performed for the same factor within the same experimental setup, and any deviations from this should be taken into account for interpretation.

Within Q, the estimated fragment length is used to create qfrag coverage profiles that have different depth distributions as compared to the conventional profiles for shifted and extended reads, whereby qfrags at peaks show a quadratic instead of only a linear increase in coverage. Intuitively, this should also contribute to a better centering of peaks at binding positions, which was the decisive argument for the development of Q-nexus.

For NGS applications, DNA is typically randomly sheared into fragments, and the 5' end positions of reads represent breakpoints in the DNA of given cells. Furthermore, removal of duplicated reads is a common procedure applied in order to eliminate PCR duplicated reads. The saturation-based evaluation of peaks takes both aspects into account and thus provides a more robust alternative to conventional approaches that focus on peak height, which depends on experimental parameters such as the read or fragment length as well as on whether duplicated reads were removed or not. Finally, the saturation-based evaluation was supplemented with statistical tests for the cases with and without control.

Q was implemented in C++, and the software along with a tutorial is available on GitHub (Section 3.2.4). A memory efficient implementation allows Q to be run on desktop computers. Beyond that, Q is also fast. For typical datasets, peak calling can be performed within two minutes on average, which is advantageous for large scale projects with large numbers of ChIP-seq datasets. Given the well-defined command line interface and the ability to read and write standard formats, Q blends in well with the landscape of bioinformatic tools that are typically involved in ChIP-seq data analysis.

Reproducibility is an important issue in ChIP-seq. The peak callers Q, MACS2, SPP and PeakSeq were compared with respect to reproducibility of peak identification and ranking (Section 3.2.5). Q showed the best compatibility with the IDR procedure, and as measured by the ENCODE standards, identified significantly more reproducible peaks, whereby datasets for which at least one of the four peak callers showed problems with IDR compatibility were excluded from the analysis. SPP is considered as well compatible with the IDR procedure and this could be confirmed by the IDR compatibility analysis. PeakSeq did not perform well with regard to IDR compatibil-

ity. However, after exclusion of problematic datasets, PeakSeq yielded the second-best results with respect to the identification of reproducible peaks. Finally, the most widely used peak caller MACS2 showed a moderate overall performance.

The elucidation of transcription factor binding preferences is one of the main applications of ChIP-seq. Transcription factors and other proteins bind sequence specific to DNA which is why genuine ChIP-seq peaks are typically enriched for binding motifs. Therefore, the peak callers were also compared with respect to motif content of peaks. For this purpose, reference motifs were derived from peak regions identified by all four peak callers, and the occurrences of reference motifs were determined for the top 50,000 peaks of the individual peak callers. As compared to the other peak callers, Q showed significantly more peaks with at least one reference motif occurrence (Section 3.2.6), whereby SPP also showed above average performance.

The TFDS signature of paused open promoters derived from RNAPII and H3K4me3 ChIP-seq data (Section 3.2.7) opens up new perspectives for future investigations. Regulatory DNA elements such as promoters and enhancers are often brought into contact with one another by the formation of chromatin loops. The NGS application Hi-C can be used be used to identify all pairwise DNA interactions in a genome-wide fashion [115]. Capture Hi-C is an extension of Hi-C that allows to focus on preselected target regions referred to as viewpoints. Similar to exome sequencing, this reduces the sequencing depth which is required to achieve a resolution that is sufficient to assign gene promoters their regulatory elements [51, 77, 99, 100, 57]. Such analyses are often complemented with other NGS applications. For instance, additional RNA-seq and histone ChIP-seq experiments were performed in order to characterize observed contacts more thoroughly [5]. For future investigations, ChIP-seq experiments with RNAPII and H3K4me3 could be performed in addition to Hi-C, and Q's TFDS signature could be used to identify interacting paused open promoters and associated regulatory elements.

Altogether, it could be shown that Q is suitable for large scale analyses such as the ENCODE project and additionally has advantages with respect to runtime and reproducibility. At the same time, Q can also be used for the detailed analyses of combined binding patters of different target proteins.

At the outset of the development of Q-nexus [45], there was no comprehensive software package for the analysis for ChIP-nexus data. The supplied software provides innovative solutions for the most important subproblems, from the preprocessing of ChIP-nexus reads (Section 4.2.1) and unbiased monitoring of PCR-overamplification making use of the random barcodes (Section 4.2.2), over the estimation of the protected region width from genome-wide data avoiding the problem with the phantom peak (Section 4.2.3), up to the reproducible identification of transcription factor binding sites reusing the concept of qfrags (Section 4.2.4).

FastQC's familiar plot for duplication levels was translated into its unbiased counterpart using the information of the random barcodes (Section 4.2.2). The concept of this plot is not restricted to ChIP-nexus but also applicable to other NGS applications that make use random barcodes in order to

identify PCR duplicated reads. Besides the PCR duplication levels, this plot also includes the duplication levels of ChIP-nexus reads after removal of PCR duplicated reads, which could be useful for the evaluation of experiments and troubleshooting whenever pile-ups of reads form a part of the signal, for instance, for the sequencing of small RNAs.

The qfrag-length distribution takes up the idea of the cross-correlation method inasmuch as the binding characteristics are derived in a genome-wide fashion. The problem with the phantom peak is circumvented using pseudo-controls that are derived from the original data by applying simple *strand switch* and *shift* operations to all mapped reads. The estimates of the protected region width ℓ''' range from 9 to 18 bp, which is plausible from a biological point of view. Furthermore, the estimated widths are largely in line with previously published footprints of 5′ end positions at preselected binding sites [47]. Interestingly, the qfrag-length distribution reproducibly shows very characteristic signatures for individual transcription factors. However, these plots have to be interpreted carefully, because the method with the pseudo-control introduces new bias to the range upwards of one read length.

The estimated protected region width is used to construct the qfrag coverage profile, and regions in which pile-ups of reads on the forward and reverse strand occur at a distance of ℓ''' will be selectively emphasized. However, how well the qfrag coverage profile reflects the precise binding architecture of given target proteins or complexes remains an open question. The estimation procedure integrates qfrags from the entire genome, but there may be subpopulations of binding sites with different architectures [104], which is a general problem with the interpretation of ChIP-exo and ChIP-nexus data. For simpler architectures such as seen for Dorsal, it seems to be sufficient to use only a single ℓ''', but for more complex architectures it might be useful to apply Q-nexus repeatedly with various protected region widths, whereby appropriate settings of ℓ''' could be read off from the qfrag-length distribution or conventional footprints of 5′ end positions.

Using the datasets of the original ChIP-nexus publication, Q-nexus was compared to MACS2 and MACE with respect to reproducibility of peak calling (Section 4.2.5). This comparison was performed within almost the same framework that had been used before for the analysis of reproducible ChIP-seq peak calling (Section 3.2.5) with the difference that no uniform argument was used for the protected region width and the analogous parameters of the other two peak callers. Within this setting, peaks can be identified most reproducibly using Q-nexus. However, it turned out that the build-in estimation routines of MACS2 and MACE did not perform well for the chosen parameter settings and on the analyzed datasets. Without exception, the estimates of MACS2 and MACE are larger than Q-nexus' protected region width and seem to be shifted towards one read length, similar to those derived from the cross-correlation curve. Furthermore, there is no obvious way to bring these estimates in line with underlying molecular mechanisms. Therefore, no valid conclusions can be drawn with regard to the reproducibility of peak calling using MACS2 or MACE. In contrast to that, Q-nexus showed good IDR compatibility and identified large numbers of reproducible peaks using the protected region width derived from the qfrag-length distribution.

Taken everything together, Q-nexus is good for a first exploratory data

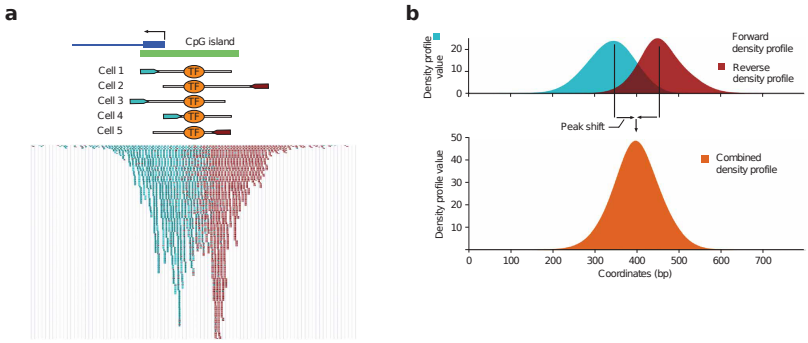
analysis without prior knowledge about the analyzed factor. The BAM file free of PCR duplicates, the unbiased plot for duplication levels, the estimation of the width of the protected region and the lists of reproducible peaks form a good basis for further investigations. In that respect, Q-nexus can be used to speed up and complement the conventional analysis pipelines for ChIP-exo and ChIP-nexus data analysis.

This thesis has grown out of an exiting environment involving skilled personnel and scientists specialized in medical genetics, developmental biology, information technology, biotechnology, statistics and bioinformatics. Two tools were developed: the ChIP-seq peak caller Q that improves reproducible peak calling and provides opportunities for future investigations as well as the ChIP-nexus analysis pipeline Q-nexus. Both tools implement a number of innovations that possibly could be also applicable to other NGS-applications. For instance, it might be useful to apply the saturation framework of Q to the outermost ends of Hi-C restriction fragments that should be saturated in a similar fashion to ChIP-seq peaks. The saturation scores could then be used to normalize contact frequencies between interacting genomic loci. Or the duplication level plots of Q-nexus may be used whenever pile-ups of reads form a part of the signal and molecular random barcodes are available. Finally, both tools are actively being discussed [120, 80, 79] and used [5, 124, 52] by the scientific community.

Appendices

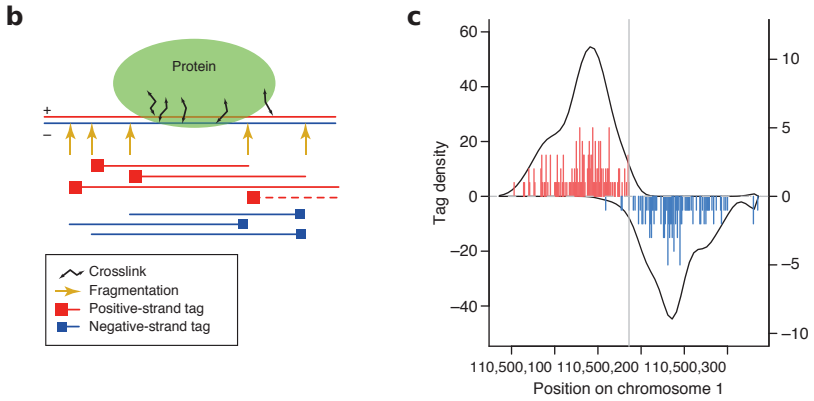
A

MODELS OF CHIP-SEQ READ DISTRIBUTION AT PROTEIN BINDING SITES



Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al. Nat. Methods Sept 2008

Figure A1: **Valouev et al., 2008 [114]:** (**a**) Schematic representation of fragmentation at binding sites together with empirical data is shown. (**b**) Coverage profiles for reads mapped to the forward (blue) and reverse strand (red) are shown (top). These two profiles are combined into one profile by shifting all reads towards the 3′ direction according to the strand to which each individual read is mapped and, from then on, treating them as if they were mapped to the same strand (bottom). The read profiles form two clearly separated bell-shaped curves. This conception involves assumptions regarding the underrepresentation of reads between the two curves which is explained by the fact that only the outer 5′ ends of fragments are sequenced. We observed different read profiles for Hoxd13 in chicken (See original publication of Q in [46] *Supplemental Material: Figure S3*) and E2F6 in mouse cells (Figure 2).



Design and analysis of ChIP-seq experiments for DNA-binding proteins
Kharchenko et al. Nat. Biotech. Nov 2008

Figure A2: **Kharchenko et al., 2008 [63]:** (**b**) Schematic representation of breakpoints at binding positions is shown. In this conception, the distance between the forward and reverse strand peaks is assumed to reflect the width of the region that is occupied by the target protein, whereby the size distribution is taken into consideration as an additional influencing factor. (**c**) Profiles for reads mapped to the forward (red) and reverse strand (blue) are shown separately. The solid curves show the smoothed profiles.

B

APPROXIMATION OF THE PROBABILITY THAT A BIN REMAINS EMPTY

After a conversation about this topic, the following proof was provided by Na Zhu, a former member of the institute for medical genetics and human genetics at the Charité in Berlin. The proof was double checked by myself and Alexander Krannich, a statistician who coordinated clinical studies at the Charité.

Proof. According to the mean value theorem the following holds:

$$\int_{n-1}^{n} \frac{1}{x}\,\mathrm{d}x = \frac{1}{c} \tag{1}$$

$$\frac{1}{n} \leq \int_{n-1}^{n} \frac{1}{x}\,\mathrm{d}x \leq \frac{1}{n-1} \tag{2}$$

$$\begin{aligned} n-1 \leq c \leq \quad & n \\ \tfrac{1}{n} \quad \leq \tfrac{1}{c} \leq \tfrac{1}{n-1} & \end{aligned} \tag{3}$$

The integral of $\frac{1}{x}$ is $\ln(x)$.

$$\begin{aligned} \int_{n-1}^{n} \frac{1}{x}\,\mathrm{d}x &= -\ln(n) - \ln(n-1) \\ &= -\ln(\frac{n-1}{n}) \\ &= -\ln(1-\frac{1}{n}) \end{aligned} \tag{4}$$

Using formula 4 in 3 yields:

$$\begin{aligned} \tfrac{1}{n} \quad &\leq -\ln(1-\tfrac{1}{n}) \leq \tfrac{1}{n-1} \\ -\tfrac{m}{n-1} &\leq m \cdot \ln(1-\tfrac{1}{n}) \leq -\tfrac{m}{n} \\ e^{-\frac{m}{n-1}} &\leq \quad (1-\tfrac{1}{n})^m \quad \leq e^{-\frac{m}{n}} \end{aligned} \tag{5}$$

□

C

CHIP-NEXUS MAPPING ARTIFACTS

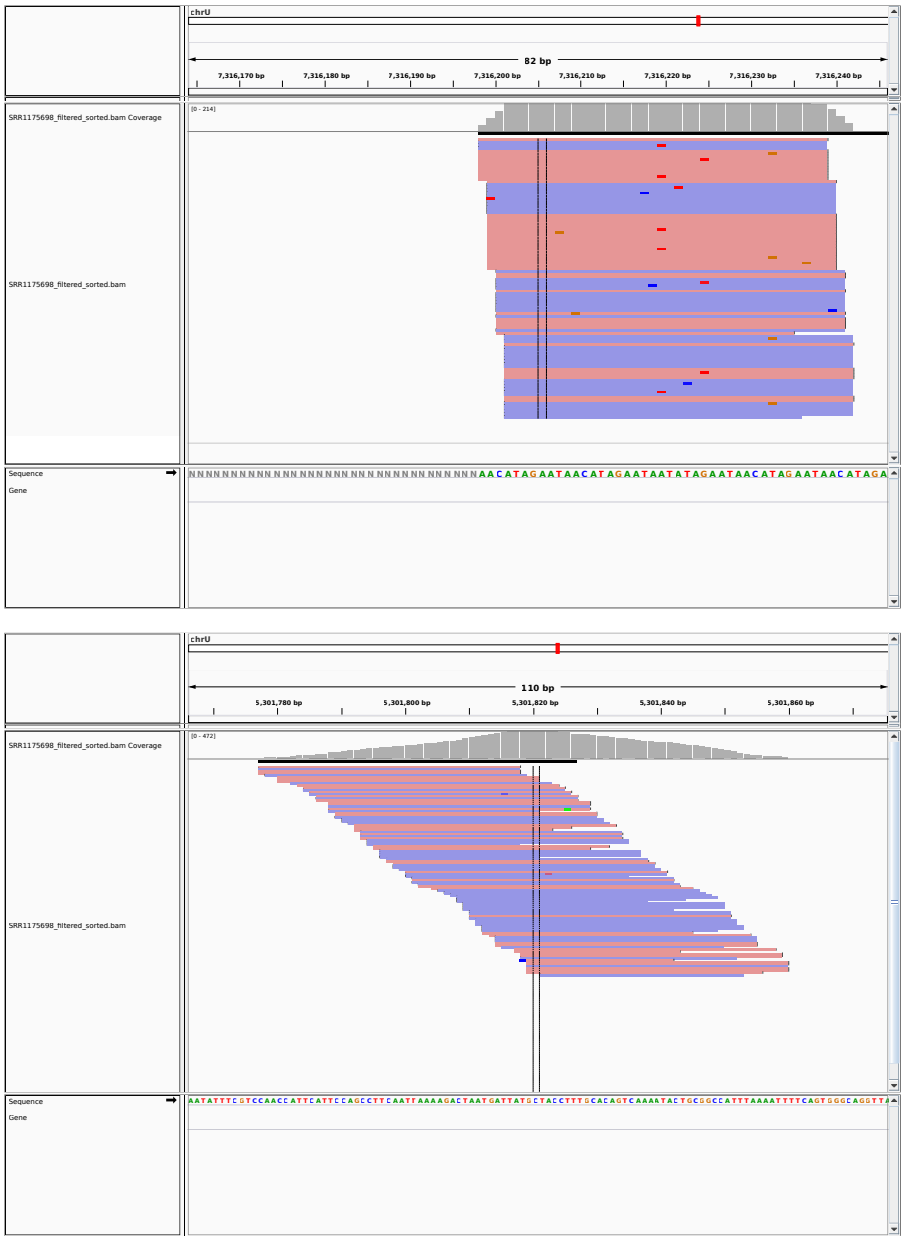


Figure A3: This screenshot was taken from the Integrated Genome Viewer. Two typical mapping artifacts for ChIP-nexus are shown. The data was derived from an experiment with the Dorsal transcription factor in Drosophila Melongaster (GEO: GSE55306). The artifacts are located on chromosome U. Such artifacts most likely contribute to the phantom peak in the cross-correlation plot.

D

ZUSAMMENFASSUNG DER ERGEBNISSE IN DEUTSCHER SPRACHE

Diese Arbeit handelt von der Entwicklung bioinformatischer Methoden und Software zur Vorhersage von DNA-Protein Interaktionen aus ChIP-seq- und ChIP-nexus-Daten.

Die Regulation der Genexpression ist ein zentrales Thema in den Lebenswissenschaften. Die Zellen eines menschlichen Organismus enthalten dieselbe Erbinformation in Form von DNA. Dabei haben verschiedene Zelltypen unterschiedliche Gestalt und Funktion. Auf molekularer Ebene unterscheiden sich Zelltypen vor allem darin, welche der rund 30000 Gene aktiv sind. Damit ein Gen aktiv wird, muss seine genetische Information in funktionelle Moleküle (vorwiegend Proteine) übersetzt werden. Der erste Schritt dieses Vorgangs wird als Transkription bezeichnet und findet direkt an der DNA im Zellkern statt. DNA-bindende Proteine, wie Transkriptionsfaktoren oder Histonproteine, spielen daher eine wichtige Rolle bei der Regulation der Transkription.

Inzwischen werden kostengünstige Hochdurchsatzmethoden zur Sequenzierung von DNA, die üblicherweise als Next-Generation-Sequencing (NGS) bezeichnet werden, auch auf Fragestellungen angewendet, die über das reine Erfassen von Basenabfolgen hinaus gehen. Ein Beispiel einer NGS-Anwendung ist ChIP-seq, welche dazu verwendet werden kann, genomweit Protein-DNA Interaktionen für ein gegebenes Zielprotein zu bestimmen. ChIP-nexus ist eine Weiterentwicklung von ChIP-seq mit deutlich erhöhter Auflösung.

Im Allgemeinen sind NGS-Daten sehr umfangreich und es hängt vom zugrunde liegenden experimentellen Protokoll ab, wie die Daten auszuwerten sind. Dies erfordert effiziente Algorithmen, die individuelle Lösungen umsetzen und typischerweise auch statistische Modelle beinhalten. Für die vorliegende Arbeit wurden eine Reihe von innovativen Algorithmen entwickelt, die verschiedene Teilprobleme bei der Vorhersage von Protein-DNA Interaktionen aus ChIP-seq- und ChIP-nexus-Daten adressieren. Beispielsweise wurde für die Sättigung genomischer Regionen mit mappierten NGS-Reads, die anhand von Sequenzidentität Positionen im Genom eindeutig zugeordnet werden können, im Rahmen des klassischen Occupancy-Problems statistisch modelliert um ChIP-seq peaks zu bewerten. Dabei stellt das Maß der Sättigung eine Alternative zur konventionellen Read-Tiefe dar und ist über ChIP-seq hinaus auch auf andere NGS-Anwendungen anwendbar. Darüber hinaus wurde für diese Arbeit umfangreiche Software entwickelt, die begleitet von zwei von Publikationen in den Fachzeitschriften Genome Research und BMC Genomics auf der Entwickler-Plattform GitHub bereitgestellt wurde: http://charite.github.io/Q/. Diese Software wurde von der wissenschaftlichen Gemeinschaft bereits diskutiert und angewendet.

BIBLIOGRAPHY

[1] Daniel Aird, Michael G. Ross, Wei Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2), 2011.

[2] Bronwen L. Aken, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, et al. Ensembl 2017. *Nucleic Acids Research*, 45(D1):D635–D642, 2017.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, 4th edition*. 2002.

[4] Simon R Andrews. FastQC: a quality control tool for high throughput sequence data., 2010.

[5] Guillaume Andrey, Robert Schöpflin, Ivana Jerković, Verena Heinrich, Daniel Murad Ibrahim, Christina Paliou, Myriam Hochradel, Bernd Timmermann, Stefan Haas, Martin Vingron, and Stefan Mundlos. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome research*, page gr.213066.116, 2016.

[6] Anthony T Annuziato and A Annunziato. DNA Packaging: Nucleosomes and Chromatin. *Nature Education*, 1(1):1, 2008.

[7] G Arya, A Maitra, and S A Grigoryev. A structural perspective on the where, how, why, and what of nucleosome positioning. *J Biomol Struct Dyn*, 27(6):803–820, 2010.

[8] Timothy L Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England)*, 27(12):1653–1659, jun 2011.

[9] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE : tools for motif discovery and searching. *Nucleic acids research*, 37(May):202–208, 2009.

[10] Timothy L. Bailey and Philip MacHanick. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17), 2012.

[11] Andrew J. Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications, 2011.

[12] Anaïs F Bardet, Jonas Steinmann, Sangeeta Bafna, Juergen a Knoblich, Julia Zeitlinger, and Alexander Stark. Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics (Oxford, England)*, 29(21):2705–2713, sep 2013.

[13] Sam Behjati and Patrick S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, 98(6):236–238, 2013.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[15] Michael F. Berger, Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Peña-Castillo, Trevis M. Alleyne, Sanie Mnaimneh, Olga B. Botvinnik, Esther T. Chan, et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell*, 133(7):1266–1276, 2008.

[16] Ewan Birney. The making of ENCODE: Lessons for big-data projects. *Nature*, 489(7414):49–51, 2012.

[17] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: A web-based genome analysis tool for experimentalists, 2010.

[18] Alan Boyle, Justin Guinney, Gregory Crawford, and Terrence Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, nov 2008.

[19] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[20] Bradley R. Cairns. The logic of chromatin architecture and remodelling at promoters, 2009.

[21] Thomas S Carroll, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in genetics*, 5(April):75, jan 2014.

[22] Emily C. Chittock, Sebastian Latwiel, Thomas C.R. Miller, and Christoph W. Müller. Molecular architecture of polycomb repressive complexes. *Biochemical Society Transactions*, 45(1):193–205, 2017.

[23] F S Collins. The heritage of humanity. *Nature*, S1(May 2002):9–12, 2006.

[24] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[25] Timothy Daley, Andrew D Smith, Los Angeles, and Computational Biology. Predicting the molecular complexity of sequencing libraries. 10(4):325–327, 2013.

[26] Bouke a de Boer, Karel van Duijvenboden, Malou van den Boogaard, Vincent M Christoffels, Phil Barnett, and Jan M Ruijter. OccuPeak: ChIP-Seq peak calling based on internal background modelling. *PloS one*, 9(6):e99844, jan 2014.

[27] Thomas M. DeChiara, Elizabeth J. Robertson, and Argiris Efstratiadis. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell*, 64(4):849–859, 1991.

[28] L Dimitrova, V Seitz, J Hecht, D Lenze, P Hansen, M Szczepanowski, L Ma, E Oker, A Sommerfeld, F Jundt, W Klapper, and M Hummel. PAX5 overexpression is not enough to reestablish the mature B-cell phenotype in classical Hodgkin lymphoma. *Leukemia*, jul 2013.

[29] M Dodt, J T Roehr, R Ahmed, and C Dieterich. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel)*, 1(3):895–905, 2012.

[30] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC bioinformatics*, 9:11, jan 2008.

[31] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[32] R. Edgar. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[33] Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited, 2013.

[34] The Encode and Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046, 2011.

[35] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194, 1998.

[36] Anthony Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, aug 2008.

[37] W Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, Inc., New York, 3 edition, 1968.

[38] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirleysvn add Liu. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*, 7(9):1728–1740, sep 2012.

[39] Terrence S Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, (12):840–852, dec.

[40] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[41] P J Good, M S Guyer, S Kamholz, L Liefer, K Wetterstrand, D Kampa, E A Sekinger, J Cheng, H Hirsch, S Ghosh, Z Zhu, et al. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306:636–639, 2004.

[42] Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO : Scanning for occurrences of a given motif. pages 2–3, 2011.

[43] Ava Handley, Tamás Schauer, Andreas G. Ladurner, and Carla E. Margulies. Designing Cell-Type-Specific Genome-wide Experiments, 2015.

[44] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12), 2010.

[45] Peter Hansen, Jochen Hecht, Jonas Ibn-Salem, Benjamin S. Menkuec, Sebastian Roskosch, Matthias Truss, and Peter N. Robinson. Q-nexus: a comprehensive and efficient analysis pipeline designed for ChIP-nexus. *BMC Genomics*, 17(1):873, 2016.

[46] Peter Hansen, Jochen Hecht, Daniel M. Ibrahim, Alexander Krannich, Matthias Truss, and Peter N. Robinson. Saturation analysis of ChIP-seq data for reproducible identification of binding peaks. *Genome Research*, 2015.

[47] Qiye He, Jeff Johnston, and Julia Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol*, 33(4):395–401, apr 2015.

[48] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin Lin, Peter Laslo, Jason Cheng, Cornelis Murre, Harinder Singh, and Christopher 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.

[49] Joshua Ho, Eric Bishop, Peter Karchenko, Nicolas Negre, Kevin White, and Peter Park. ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*, 12(1):134, feb 2011.

[50] Roger A. Hoskins, Joseph W. Carlson, Cameron Kennedy, David Acevedo, Martha Evans-Holm, Erwin Frise, Kenneth H. Wan, Soo Park, Maria Mendez-Lago, Fabrizio Rossi, et al. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science*, 316(5831):1625–1628, 2007.

[51] Jim R Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, Eleni Giannoulatou, Magnus Lynch, Marco De Gobbi, Stephen Taylor, Richard Gibbons, and Douglas R Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, 46(2):205–12, 2014.

[52] Jonas Ibn-Salem and Miguel A. Andrade-Navarro. Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs. *bioRxiv*, page 257584, 2018.

[53] Daniel M. Ibrahim. *ChIP-seq Reveals Mutation-Specific Pathomechanisms of HOXD13 Missense Mutations*. PhD thesis, Humboldt-Universität zu Berlin, 2015.

[54] Daniel M Ibrahim, Peter Hansen, Christian Rödelsperger, Asita C Stiege, Sandra Dölken, Denise Horn, Marten Jäger, Catrin Janetzki, Peter Krawitz, Gundula Leschik, et al. Distinct global shifts in genomic binding profiles of limb malformation associated HOXD13 mutations. *Genome research*, aug 2013.

[55] Competing Financial Interests. Integrative genomics viewer. 29(1):24–26, 2011.

[56] IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and Symbols for Nucleic Acids , Polynucleotides and their Constituents. *European Journal of Biochemisty*, 15:203–208, 1970.

[57] Biola M. Javierre, Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, Mikhail Spivakov, Peter Fraser, Oliver S. Burren, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, 2016.

[58] Ivana Jerković, Daniel M. Ibrahim, Guillaume Andrey, Stefan Haas, Peter Hansen, Catrin Janetzki, Irene González Navarrete, Peter N. Robinson, Jochen Hecht, and Stefan Mundlos. Genome-Wide Binding of Posterior HOXA/D Transcription Factors Reveals Subgrouping and Association with CTCF. *PLoS Genetics*, 13(1), 2017.

[59] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26(11):1293–1300, nov 2008.

[60] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–502, jun 2007.

[61] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo proteinDNA binding sites from ChIP-Seq data. *Nucleic Acids Research*, 36(16):5221–5231, sep 2008.

[62] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, may 2002.

[63] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26(12):1351–1359, dec 2008.

[64] Anshul Kundaje. ENCODE: TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework. *https://sites.google.com/site/anshulkundaje/projects/idr*, 2012.

[65] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R.

Hughes, and Matthew T. Weirauch. The Human Transcription Factors, 2018.

[66] Stephen Landt, Georgi Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley Bernstein, Peter Bickel, James Brown, Philip Cayting, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831, 2012.

[67] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[68] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1), 2011.

[69] Marion Leleu, Grégory Lefebvre, and Jacques Rougemont. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in Functional Genomics*, 9(5-6):466–476, dec 2010.

[70] Shawn E. Levy and Richard M. Myers. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 17(1):95–115, 2016.

[71] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, aug 2009.

[72] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.

[73] Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[74] M. Logan. Role of Pitx1 Upstream of Tbx4 in Specification of Hindlimb Identity. *Science*, 283(5408):1736–1739, 1999.

[75] Raphaël Margueron and Danny Reinberg. The Polycomb complex PRC2 and its mark in life, 2011.

[76] Benjamin Sefa Menküc. *Development of a bespoke algorithm to analyze ChIP-nexus genome-wide protein-DNA binding profiles*. PhD thesis, Charité - Universitätsmedizin Berlin, 2017.

[77] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, 2015.

[78] Ali Mortazavi, Brian a Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, jul 2008.

[79] Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Briefings in Bioinformatics*, 18(2):279–290, 2017.

[80] Ryuichiro Nakato and Katsuhiko Shirahige. Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile. *Bioinformatics*, 2018.

[81] Daniel Newkirk, Jacob Biesinger, Alvin Chon, Kyoko Yokomori, and Xiaohui Xie. AREM: aligning short reads from ChIP-sequencing by expectation maximization. *Journal of computational biology : a journal of computational molecular cell biology*, 18(11):1495–1505, nov 2011.

[82] David Nix, Samir Courdy, and Kenneth Boucher. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, 9(1):523, 2008.

[83] Claire E. Olson and Steven B. Roberts. Genome-wide profiling of DNA methylation and gene expression in Crassostrea gigas male gametes. *Frontiers in Physiology*, 5 JUN, 2014.

[84] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat Meth*, 6(11s):S22—-S32, nov 2009.

[85] Ana Pombo and Niall Dillon. Three-dimensional genome architecture: Players and mechanisms, 2015.

[86] Zhaohui Qin, Jianjun Yu, Jincheng Shen, Christopher Maher, Ming Hu, Shanker Sundaram, Jindan Yu, and Arul Chinnaiyan. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, 11(1):369, jul 2010.

[87] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, mar 2010.

[88] Parameswaran Ramachandran, Gareth A Palidwor, Christopher J Porter, and Theodore J Perkins. Sequence analysis MaSC : mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. 29(4):444–450, 2013.

[89] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67, jan 2011.

[90] Ho Rhee and Franklin Pugh. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6):1408–1419, dec 2011.

[91] Ho Sung Rhee and B Franklin Pugh. ChIP-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*, Chapter 21:Unit 21.24, oct 2012.

[92] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, et al. Genome-wide pro-

files of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657, 2007.

[93] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[94] Kate R Rosenbloom, Cricket a Sloan, Venkat S Malladi, Timothy R Dreszer, Katrina Learned, Vanessa M Kirkup, Matthew C Wong, Morgan Maddren, Ruihua Fang, Steven G Heitner, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research*, 41(Database issue):D56–63, jan 2013.

[95] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, jan 2009.

[96] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, jan 2009.

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

[98] Thomas D Schneider and R Michael Stephens. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[99] Stefan Schoenfelder, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola Maria Javierre, Takashi Nagano, Yulia Katsman, Moorthy Sakthidevi, Steven W. Wingett, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4):582–597, 2015.

[100] Stefan Schoenfelder, Biola-Maria Javierre, Mayra Furlan-Magaril, Steven W Wingett, and Peter Fraser. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *Journal of visualized experiments : JoVE*, (136), jun 2018.

[101] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, jan 2015.

[102] Aurelien A Serandour, Gordon D Brown, Joshua D Cohen, and Jason S Carroll. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol*, 14(12):R147, 2013.

[103] Bernd Stadelmayer, Gaël Micas, Adrien Gamot, Pascal Martin, Nathalie Malirat, Slavik Koval, Raoul Raffel, Bijan Sobhian, Dany Severac, Stéphanie Rialle, Hugues Parrinello, Olivier Cuvier, and Monsef Benkirane. Integrator complex regulates NELF-mediated RNA polymerase II

pause/release and processivity at coding genes. *Nature Communications*, 5, 2014.

[104] Stephan R Starick, Jonas Ibn-Salem, Marcel Jurk, Céline Hernandez, Michael I Love, Ho-Ryun Chung, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H Meijsing. ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Res*, 25(6):825–835, jun 2015.

[105] D Stormo and Thomas D Schneider. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, 10(9):2997–3011, 1982.

[106] J Peter Svensson, Manu Shukla, Victoria Menendez-Benito, Ulrika Norman-Axelsson, Pauline Audergon, Indranil Sinha, Jason C Tanny, Robin C Allshire, and Karl Ekwall. A nucleosome turnover map reveals that the stability of histone H4 Lys20 methylation depends on histone recycling in transcribed chromatin. *Genome Res*, 25(6):872–883, jun 2015.

[107] Daniel P. Szeto, Concepción Rodriguez-Esteban, Aimee K. Ryan, Shawn M. O'Connell, Forrest Liu, Chrissa Kioussi, Anatoli S. Gleiberman, Juan Carlos Izpisúa-Belmonte, and Michael G. Rosenfeld. Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. *Genes and Development*, 13(4):484–494, 1999.

[108] Tommy W. Terooatea, Amir Pozner, and Bethany A. Buck-Koehntop. PAtCh-Cap: Input strategy for improving analysis of ChIP-exo data sets and beyond. *Nucleic Acids Research*, 44(21), 2016.

[109] Leonid Teytelman, Bilge Özaydin, Oliver Zill, Philippe Lefrançois, Michael Snyder, Jasper Rine, and Michael B. Eisen. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE*, 4(8), 2009.

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

[111] The R Core Team. R : A Language and Environment for Statistical Computing, 2015.

[112] Hélène Tourrière, Karim Chebli, and Jamal Tazi. mRNA degradation machines in eukaryotic cells, 2002.

[113] Geetu Tuteja, Peter White, Jonathan Schug, and Klaus H Kaestner. Extracting transcription factor targets from ChIP-Seq data. *Nucleic acids research*, 37(17):e113, sep 2009.

[114] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*, 5(9):829–834, sep 2008.

[115] Nynke L van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S

Lander. Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of visualized experiments : JoVE*, 6(39):1869, 2010.

[116] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: Function, expression and evolution, 2009.

[117] Axel Visel and Stefan Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Article Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. pages 1012–1025, 2015.

[118] Liguo Wang, Junsheng Chen, Chen Wang, Liis Uusküla-Reimand, Kaifu Chen, Alejandra Medina-Rivera, Edwin J Young, Michael T Zimmermann, Huihuang Yan, Zhifu Sun, Yuji Zhang, Stephen T Wu, Haojie Huang, Michael D Wilson, Jean-Pierre A Kocher, and Wei Li. MACE: model based analysis of ChIP-exo. *Nucleic Acids Res*, 42(20):e156, nov 2014.

[119] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics, 2009.

[120] Rene Welch, Dongjun Chung, Jeffrey Grass, Robert Landick, and Sündüz Kele. Data exploration, quality control and statistical analysis of ChIP-exo/nexus experiments. *Nucleic Acids Research*, 45(15), 2017.

[121] Haipeng Xing, Yifan Mo, Will Liao, and Michael Q Zhang. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS computational biology*, 8(7):e1002613, jan 2012.

[122] Yajie Yang, Justin Fear, Jianhong Hu, Irina Haecker, Lei Zhou, Rolf Renne, David Bloom, and Lauren M McIntyre. Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Computational and structural biotechnology journal*, 9(13):e201401002, jan 2014.

[123] Wai-Shin Yong, Fei-Man Hsu, and Pao-Yang Chen. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin*, 9(1):26, 2016.

[124] Yajia Zhang, Sethuramasundaram Pitchiaya, Marcin Cieślik, Yashar S. Niknafs, Jean C.Y. Tien, Yasuyuki Hosono, Matthew K. Iyer, Sahr Yazdani, Shruthi Subramaniam, Sudhanshu K. Shukla, et al. Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. *Nature Genetics*, 50(6):814–824, 2018.

[125] Yong Zhang, Tao Liu, Clifford Meyer, Jerome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nusbaum, Richard Myers, Myles Brown, Wei Li, and Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.

[126] Y H Zhou, K Xia, and F A Wright. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678, 2011.