White Paper

White paper on the Transcription Factor ChIP-Seq

September 11, 2015

Sample to Insight

CLC bio, a QIAGEN Company
Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · Fax: +45 86 20 12 22
www.clcbio.com · info@clcbio.com

Contents

1 Abstract 3

2 Introduction 3

2.1 Background . . . 3

2.2 State of the art . . . 4

2.3 Aims for a next-generation peak caller . . . 5

2.4 New approach . . . 6

3 Signal detection using a Hotelling-filter 7

3.1 Quality control of ChIP-seq data . . . 8

3.2 Normalization . . . 9

3.3 Discovering obvious peaks . . . 10

3.4 Learning the peak shape . . . 11

3.5 Peak Shape Score . . . 13

3.6 Peak-detection . . . 13

3.7 The Transcription Factor ChIP-Seq . . . 13

4 Performance evaluation 14

4.1 Gold-standard dataset . . . 15

4.2 Calculating performance metrics . . . 16

4.3 Running Transcription Factor ChIP-Seq with different input datasets . . . 16

4.4 Results . . . 16

5 Conclusions 20

5.1 Discussion . . . 21

5.2 Future Developments . . . 21

6 Materials and Methods 23

References 24

1 Abstract

In this White Paper we present a performance analysis on the new Transcription Factor ChIP-seq tool available in CLC Genomics Workbench and CLC Genomics Server (version 7.5 and up), and in Biomedical Genomics Worbench and Server. Using an independent manually curated benchmark dataset [Rye et al., 2011] we show that the CLC implementation of the shape-based peak caller ranks well among popular state-of-the-art peak callers while requiring a minimum of intervention and parameterization from the user. We describe the algorithmic principles underlying the implementation and provide an evaluation in terms of receiver-operator characteristic (ROC) plots. Finally, we identify the most promising avenues for further developments and future application areas.

2 Introduction

2.1 Background

In order to identify functional elements in a genome, a number of experimental high-throughput techniques have been developed for investigating specific interactions between proteins and DNA. These protocols provide us with a deeper understanding of gene-regulatory and epigenetic mechanisms by identifying, for example, Transcription-Factor Binding Sites (TFBS) or the location of epigenetic marks. In this paper, we focus on transcription factors. An example is the Octamer Transcription Factor 1 (OCT1), which recognizes an 8 bp long, conserved DNA motif in promoter regions and activates the associated genes. The three dimensional structure of the transcription factor OCT1 bound to a DNA fragment is shown in figure 1.

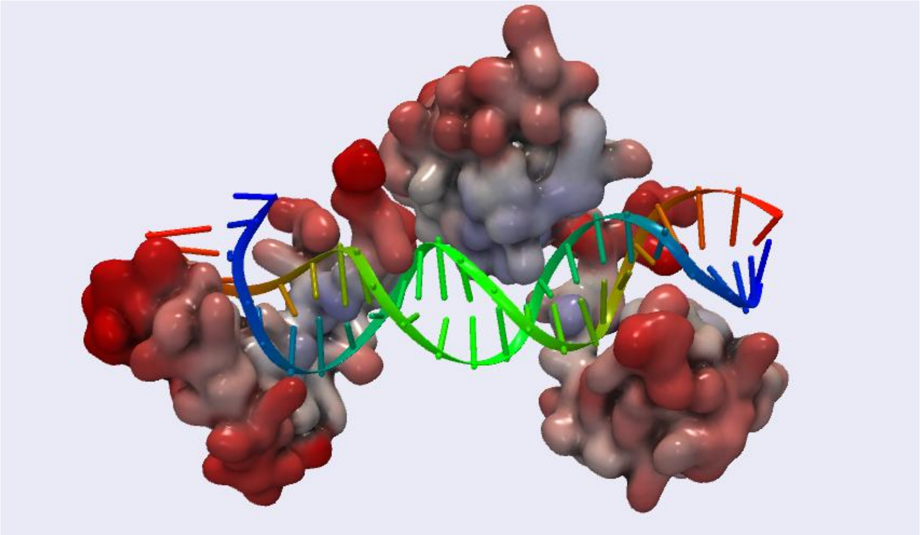

Figure 1: *A transcription factor binding to a fragment of DNA. This illustration is based on PDB entry 1gt0 of the Octamer-binding transcription factor OCT1, rendered in space-filling surface representation with blue/gray/red coloring indicating lower/medium/higher structural flexibility. The DNA fragment is shown in stick/ribbon representation. The figure was created using CLC Drug Discovery Workbench 1.5.*

In broad terms, these techniques chemically cross-link proteins to those stretches of DNA they are bound to *in vivo*. After shearing the DNA, a protein of interest is extracted along with the cross-linked DNA fragments from the cell-lysate using specific antibodies. Following this Chromatin Immuno-Precipitation (ChIP) step, the short stretches of DNA attached to the protein of interest are identified by high-throughput sequencing.

For any targeted protein and a given cell-line or condition, this results in several million reads of raw sequencing data. Usually a control experiment is performed where the immuno-precipitation step is left out. For example, the ENCODE Project (ENCyclopedia Of DNA Elements) has produced data on hundreds of regulatory factors (see http://encodeproject.org/) in mouse and human. For more in-depth information we recommend the "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia" [Landt et al., 2012, Marinov et al., 2014].

The initial step for all downstream computational analyses of the sequencing data starts by mapping the resulting set of raw reads to a reference genome. Obviously, the quality of the read mapping has an impact on the downstream analysis results. However, details of the mapping process are beyond the scope of this white paper. In this context it suffices to highlight that the CLC read mapper is one of the most accurate and best performing tools available for this task (see http://www.clcbio.com/files/whitepapers/whitepaper-on-CLC-read-mapper.pdf). In this paper we will assume that an accurate read mapping is provided. For the performance comparison, all peak callers used the same read mappings as input.

2.2 State of the art

By plotting the number of reads mapped to genomic coordinates as a so-called coverage graph, consistent and specific binding sites of the protein of interest become visually apparent as peaks (see figure 2). Rather than identifying such regions by eye, subsequent bioinformatics analysis aims at the reliable automated identification of protein-DNA binding events from the read mapping - a process referred to as peak calling.

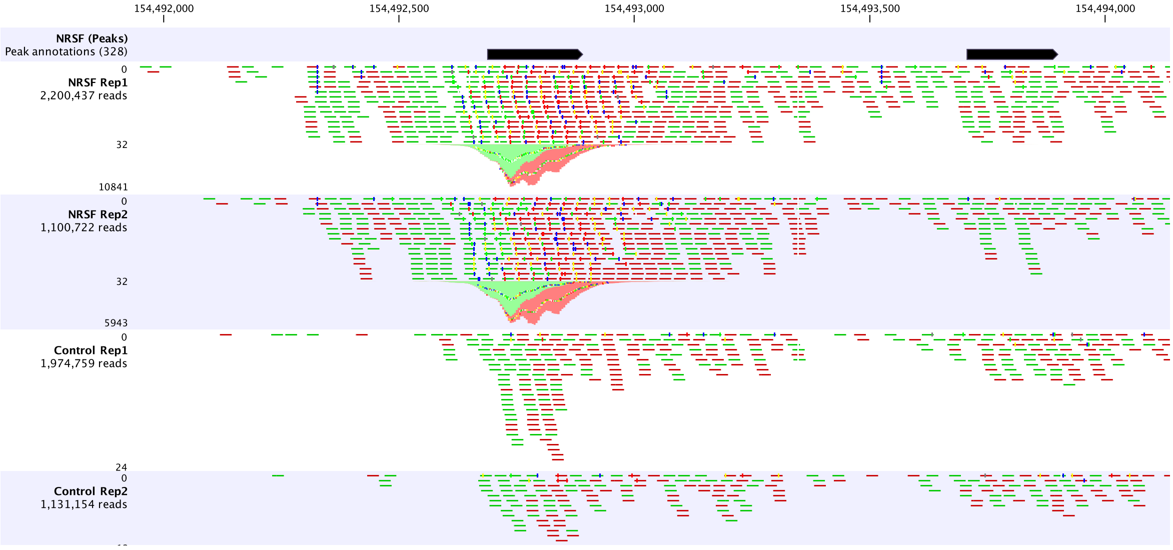

Figure 2: *Coverage graphs and mapped reads from the NRSF dataset, showing a region of human Chromosome 1. The uppermost track marks two regions classified as positive peaks in the reference dataset with blue arrows. The two tracks below visualize the read mapping of the two replicate ChIP-samples. The two lowermost tracks display the reads from the two control samples. In all tracks, forward reads are shown in green and reverse reads are shown in red. Below the reads of the ChIP-seq sample data at the peak regions, the coverage graph of forward and reverse reads is shown.*

From a computational viewpoint, the main lessons learned from the current generation of peak calling algorithms can be summarized as follows: The success of peak calling depends on how well the statistical model of the ChIP-seq signal can be fitted to the data under consideration. In this context, parameterizing a peak caller can be seen as tweaking its intrinsic model to improve the fit to the data. However, this requires in-depth knowledge of the underlying algorithm and statistical model and a good grasp of how the behavior is affected by the parameters. Therefore the fine-tuning of parameters remains a "black art" to most biologists who want to analyze the results of their ChIP-seq experiments. Since parameter optimization is a hard and time-consuming task, it is recommended [Landt et al., 2012] to focus on improvements in experimental design in order to obtain better input data rather than attempting to optimize the downstream bioinformatics pipeline.

In recent literature, it has been observed that current peak callers may miss peaks that are clearly visible to the human eye. More precisely the peaks exhibit distinct shapes which act as visual cues. However, as stated by Rye et al., 2011, "peak shape information is not fully exploited in the evaluated programs." There is still a gap between the peaks apparent to expert users and what is recognized algorithmically. In a recent paper, Heydarian et al., 2014 report: "This discordance led us to revisit our analyses of the ChIP-seq data, but we were unable to find parameters for the MACS algorithm that successfully discriminated many promoters as active, even though visual inspection clearly showed they had peaks in both chromatin modifications."

One approach to improve peak calling is to take more of the observed specific characteristics of existing ChIP-seq datasets into account and encode them directly into the algorithm (see e.g. Kornacker et al., 2012). However, hard-coding more specific models into the heuristic is not a sustainable path for future developments in the long run, as this procedure may be overfitting certain kinds of datasets. Specially given the increasing variety of ChIP-seq related experimental protocols, this approach would ultimately lead to a similar variety of specialized heuristics. This variety and specialization will make it increasingly difficult to update, test, and release algorithms while leaving users confused as to which tool and version is optimally suited to their data at hand.

Nevertheless, novel peak-shape recognition algorithms have been shown to outperform existing heuristics, identifying peaks that were missed previously [Mendoza-Parra et al., 2013, Kumar et al., 2013]. Generally speaking, in contrast to a hard-coded internal model these approaches "learn" the peak-shape from the underlying data. In the "learning-phase" an initial set of positive examples is identified, i.e. regions that unambiguously contain peaks. From this initial set, a computational representation of the specific peak-shape is constructed. This representation is then applied to the entire dataset to perform the automated peak calling, based on a suitable statistical framework.

2.3 Aims for a next-generation peak caller

Reflecting on the recent developments in the field combined with the practical lessons learned from a number of different current peak calling algorithms and the potential advances offered by the aforementioned shape-based approaches, we formulated the main requirements for our new peak calling toolset as follows:

Generality In order to keep up with the steadily rising number of experimental protocols that require peak calling for data-analysis, the algorithmic engine has to be general enough to be applicable across datasets from many different *-seq technologies (i.e. ChIP-seq, FAIRE-seq, DNase-seq etc.). The toolset should be swiftly adaptable to new datasets as they become available without expensive recoding efforts.

Specificity For achieving optimal results, the specific characteristics of the peaks need to be recognized by the algorithm. The optimization and parameterization for the task at hand should not sacrifice the generality of the underlying algorithmic implementation.

Robustness Rather than inventing ad-hoc scoring schemes, the algorithms need to be built on a mathematically and statistically well-founded framework. In particular, we employ methodologies from digital signal processing and machine learning, which have been extensively studied and are deeply understood.

Simplicity Despite the algorithmic and statistical complexity of the data-analysis task, the implementation needs to be suitable for a general audience. This translates into minimizing or even eliminating the need for parameterization and automating the most common tasks. At the same time, the algorithm needs to be transparent about its results and intuitive to use, such that advanced users can adopt the tools easily to their needs.

2.4 New approach

Some of the aims formulated above may seem to be contradictory or even mutually exclusive to each other at first sight. However, recently a general approach has been described, based on the observation that the peak calling constitutes a special case of signal detection algorithms and that it "can be solved by adapting ’uniform’ and ’formally optimal’ techniques from the signal processing literature" [Kumar et al., 2013]. In addition to being general such that signals of arbitrary shape can be processed, it is based upon a well studied framework from signal processing theory. In broad terms, the specific shape of the signal is learned from a number of "positive" regions and "negative" regions (or noise) where the signal is absent or is the result of a sequencing artifact. The resulting shape of the signal minus the noise is encoded in a vector (the so-called Hotelling-observer, named after the mathematical statistician Harold Hotelling [Hotelling, 1931]), which is then evaluated against the data-stream. This resulting filter contains the information needed for peak detection and can easily be transferred and applied to other datasets as well. This approach lends itself to visual parameterization by example such that advanced users can define the set of regions from which the filter is constructed, making it transparent as to what pattern the algorithm is detecting. Examples of approaches along this rationale are described in Kornacker et al., 2012, Kumar et al., 2013, and Stanton et al., 2013. In contrast to the "black box approach" of current peak callers, these methods lend themselves to visual and interactive parameterization by displaying positive and negative examples in a user-friendly way. At the same time, the underlying implementation remains independent of the peak-shape it is detecting - analogous to text-search algorithms being independent of the text pattern or regular expressions.

In order to build specific peak-shape filters without extensive manual annotation of positive and negative regions, we take into account that the vast majority of peak callers hardly disagree about top-scoring peaks [Wilbanks and Facciotti, 2010]. The differences in performance become apparent only for less obvious peaks; it is in this "gray zone" where the fit between the data and the intrinsic statistical model of the algorithms decides about their relative performance. This observation suggests a strategy for boot-strapping the shape-based approach: There are sufficient positive regions that can be safely and unambiguously identified by any of the currently available methods in a first pass. A more specific model of the peak-shape is then inferred from these clear-cut examples, allowing the algorithm to tune itself automatically to the data at hand. A second peak-detection step is performed, resulting in a much more sensitive peak-detection overall.

The components of the underlying "learn and apply"-cycle for peak-shapes are available as advanced tools, allowing for more manual intervention. These modules can be combined to produce more complex workflows or for iterative refinement of algorithmic performance. We defer further details to an advanced tutorial, here we stick to the automated ChIP-seq functionality, akin to running existing algorithms with standard parameters. A more detailed description of the algorithmic basis and our implementation will be given in the next section, followed by a section on the performance evaluation, benchmarking the Transcription Factor ChIP-Seq against CisGenome [Ji et al., 2008], HOMER [Heinz et al., 2010], and MACS [Zhang et al., 2008].

3 Signal detection using a Hotelling-filter

Since the shape of the signal from ChIP-seq data depends on which protein was targeted in the immuno-preciptation reaction [Stanton et al., 2013, Kumar et al., 2013], the Transcription Factor ChIP-Seq is designed to take the characteristic peak-shape into account. For example, the typical signal shape of a transcription factor binding site like NRSF shows a high concentration of forward reads followed by a high concentration of reverse reads (figure 3).

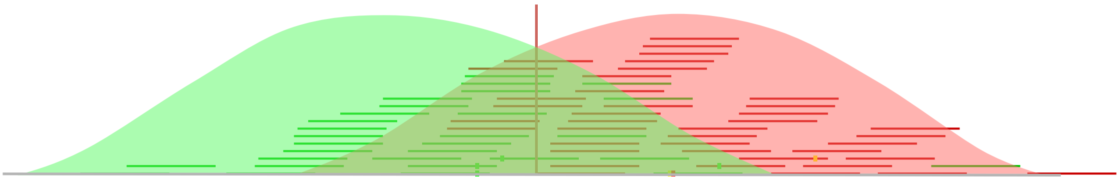

Figure 3: *Distribution of forward (green) and reverse (red) reads around a binding site of the transcription factor NRSF. The centre of the putative binding site is indicated by a red vertical line.*

The average shape of the positive regions of the NRSF transcription factor for the forward and reverse strands is shown in figure 4.

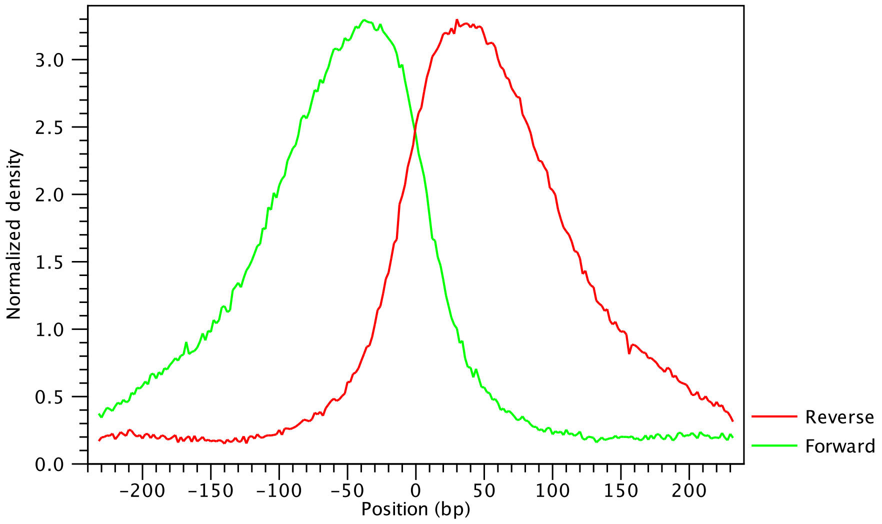

Figure 4: *Average peak shape of the transcription factor NRSF.*

The Transcription Factor ChIP-Seq makes use of both the characteristic peak shape and enriched read coverage to identify peaks in ChIP-seq data. Next, we will outline the individual steps of the entire ChIP-seq peak calling pipeline, starting from quality control and normalization of the data, describing the fundamentals of using a filter, learning a characteristic peak shape from highly enriched regions and calling peak regions including boundary refinement. Finally, we describe how these steps are working together to result in a highly automated peak detection pipeline that provides near optimal results without the need for extensive parameterization by the user.

3.1 Quality control of ChIP-seq data

During the first step of the analysis, the Transcription Factor ChIP-Seq computes several quality measures to check whether the input data satisfy the assumptions made by the algorithm. These measures can be derived from the cross-correlation profile between reads mapping to the forward and to the reverse strand. This plot is often used to investigate the quality of ChIP-seq experiments [Landt et al., 2012, Marinov et al., 2014]. A correlation profile is shown in figure 5.

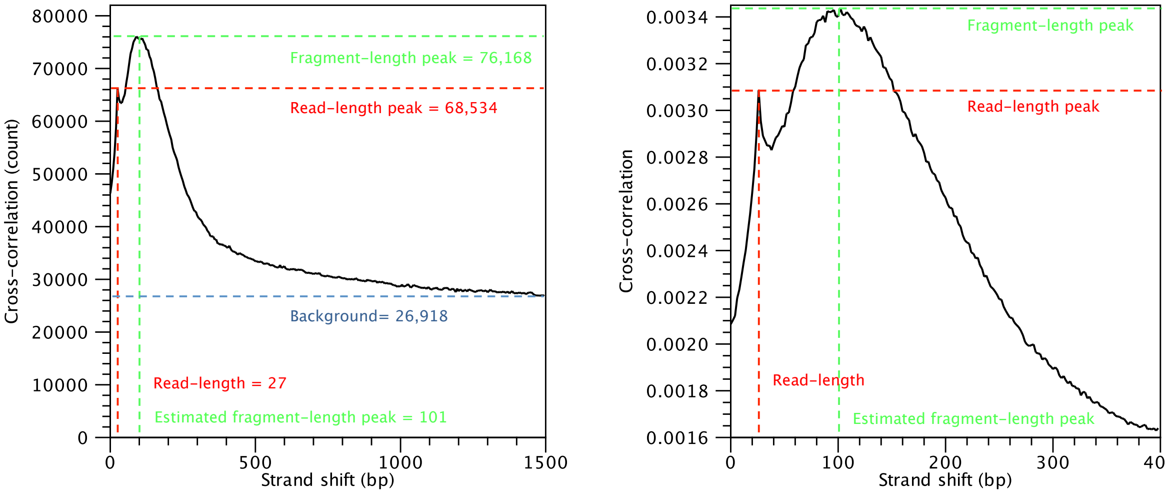

Figure 5: *Inspection of the cross-correlation plot of a ChIP-seq experiment. On the left, full correlation profile. On the right, blow-up of the interesting region between 0 and 400 bp. Note that in the right plot, the cross-correlation values have been normalized so that the area under the curve is 1.*

Several features of the cross-correlation plot are connected to the quality of the ChIP-seq experiment:

Read-length peak The cross-correlation function typically has a maximum (often called a phantom peak [Landt et al., 2012]) when the strand shift value is equal to the length of the sequenced reads. In our example (figure 5, red lines), the read length is known to be 27 and the height of the read-length peak is 68,534.

Fragment-length peak The cross-correlation function typically has a maximum when the value of the strand-shift is close to the length of the DNA fragment being sequenced. This is indicated in green in figure 5, where the height of the peak is 76,168. This peak is a characteristic feature of a ChIP-seq experiment. One can expect a pronounced peak around the fragment length because the frame shift between reads mapping to the forward and to the reverse strand near a typical transcription factor binding site (figure 3) is on average equal to the fragment length [Stanton et al., 2013, Landt et al., 2012, Kharchenko et al., 2008]. Therefore, the (relative) height of this peak can be considered a proxy for the quality of the ChIP-seq experiment. Finally, the location of this peak can be used to estimate the average length of the DNA fragments after the fragmentation step (e.g. sonication or MNAse digestion). The average fragment length in our example is estimated to be 101 bp.

Background The value of the cross-correlation at a large distance. The Transcription Factor ChIP-Seq uses the value at a strand shift of 1,500, which can be considered to be far enough from the fragment length distribution [Kharchenko et al., 2008]. This level is shown in blue in figure 5, where it is equal to 26,918. This value can be considered a proxy for the background noise of the experiment.

Normalized strand coefficient The normalized strand coefficient describes the ratio between the fragment-length peak and the background cross-correlation. In our example (figure 5), $NSC = \frac{76{,}168}{26{,}918} \approx 2.8$. This value should be greater than 1.05 for ChIP-seq experiments [Landt et al., 2012].

Relative strand correlation The relative strand correlation describes the ratio between the fragment-length peak and the read-length peak in the cross-correlation plot obtained after correcting for the background. In our example (figure 5), $RSC = \frac{76{,}168-26{,}918}{68{,}534-26{,}918} \approx 1.2$. This value should be high (at least 0.8) for transcription factor binding sites, which have a concentrated signal. However, it should be noted that this value can be low even for successful ChIP-seq experiments on histone modifications, whose signal can span large genomic regions [Landt et al., 2012].

Those quality measures have been investigated by the modENCODE consortium and are described in more detail in Landt et al., 2012. The cross-correlation profile shown in figure 5 is typical of a successful ChIP-seq experiment. On the other hand, cross-correlation plots without a pronounced fragment-length peak are typically reflective of poor quality and ragged cross-correlation profiles are typically caused by low yield (figure 6).

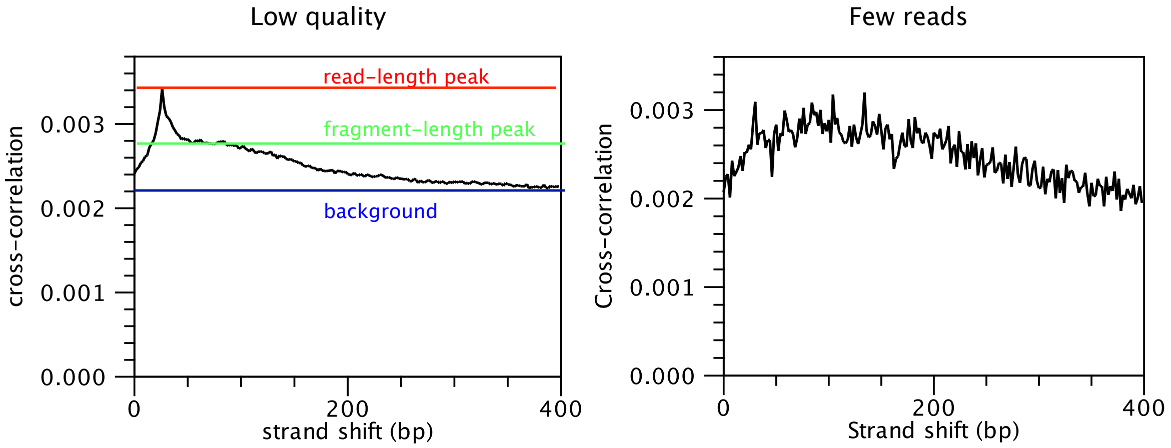

Figure 6: Cross-correlation plots of two low-quality ChIP-seq datasets. On the left, the read-length peak is significantly higher than the fragment-length peak (relative strand correlation of around 0.5), indicating potential problems in the immuno-precipitation step. On the right, a very noisy cross-correlation profile indicates a ChIP-seq experiment where a very small number of reads was sequenced.

3.2 Normalization

The Transcription Factor ChIP-Seq analyzes the genomic coverage of the reads. For each read mapping, the 5' position of the reads mapping to the forward strand and the 3' position of the reads mapping to the reverse strand are extracted. For each genomic position x, we define $f(x)$ as the number of reads mapping to the forward strand where x is the 5' position and $r(x)$ as the number of reads mapping to the reverse strand where x is the 3' position. A quantile standardization is then applied to $f(x)$ and $r(x)$ such that the normalized coverage functions f' and r' follow a standard normal distribution, i.e. $f'(x), r'(x) \sim \mathcal{N}(0,1)$.

3.3 Discovering obvious peaks

The next step of the Transcription Factor ChIP-Seq is to build a filter, which can be used to identify genomic regions whose read coverage profile matches the characteristic peak shape and to determine the statistical significance of this match. In order to build such a filter, examples of positive (e.g. ChIP-seq peaks) and negative (e.g. background noise, PCR artifacts) profiles are needed as input. The rationale is that regions with very high coverage in the ChIP-seq experiment are positive examples and regions with high coverage in the control and low in the experimental ChIP-seq data are negative examples, as they most likely originate from regions with strong sequencing biases or from PCR artifacts. Positive regions are generally easy to find and are typically found by every peak caller [Wilbanks and Facciotti, 2010]. The Transcription Factor ChIP-Seq finds these peaks by building a Gaussian filter based on the mean and variance of the fragment length distribution, which are inferred from the cross-correlation profile (figure 6). An example of a filter is shown in figure 7.

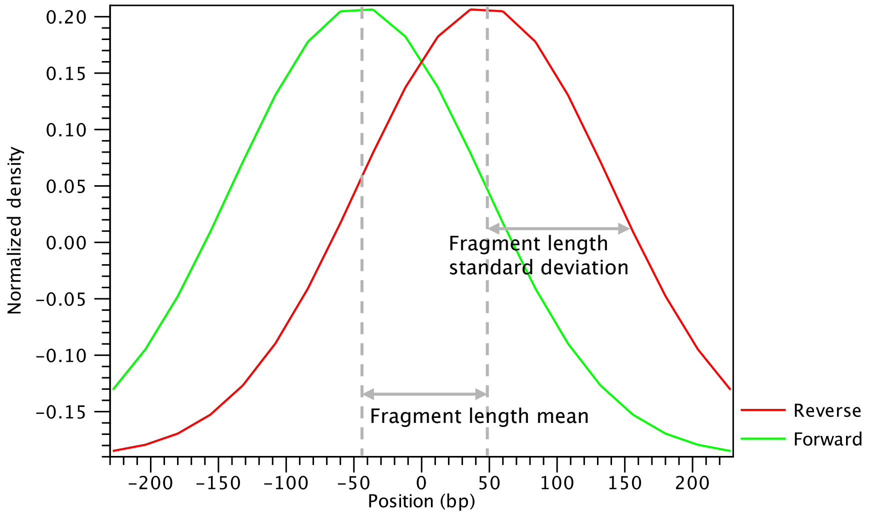

Figure 7: *A Gaussian filter for the transcription factor NRSF.*

The filter is then applied to the input data as shown in figure 8 and the result is a score that indicates how likely a genomic position is to be a center of a peak. In detail, the score is calculated as

$$\text{score} = \text{genomic coverage} \star \text{filter}, \tag{1}$$

where $\star$ denotes the cross-correlation operator. The cross-correlation between a function and a filter can be described as follows: For each genomic position x, we extract the genomic coverage profile of a window centered at x. We multiply this profile by the peak shape filter and we sum the result. The resulting number indicates how well the shape of the filter is matched. The score will reach a maximum at the center of a peak. Peaks are then identified as the regions whose centers are the genomic positions with highest score and whose size is the size of the filter.

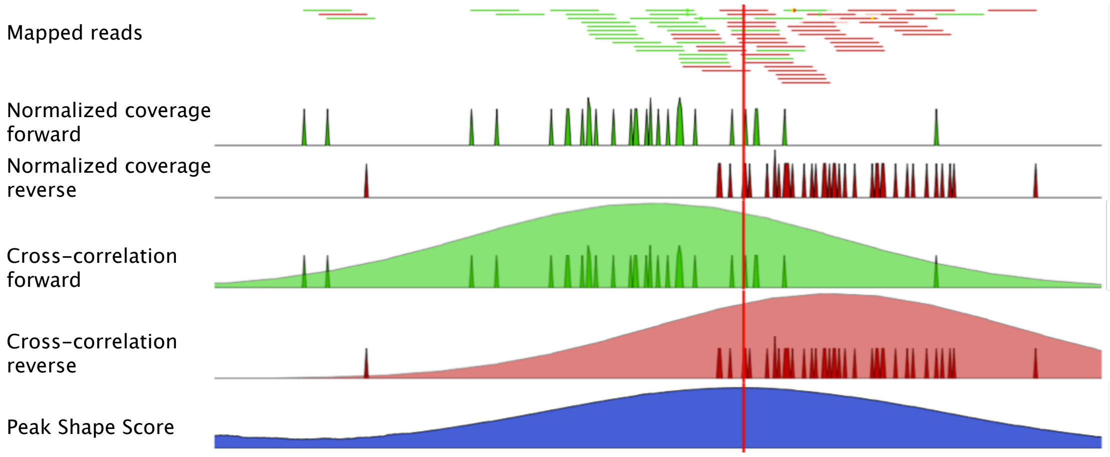

Figure 8: *Application of a Gaussian filter (figure 7) to ChIP-seq data. In the first step, the normalized coverage for reads mapping to the forward (green) and reverse (red) strand are computed. Later, for a genomic position, the cross-correlation between the forward coverage and the filter shape centered at that position is computed. An analogous procedure is performed for reads mapping to the reverse strand. Their sum constitutes the Peak Shape Score (blue).*

Similarly, the set of negative examples is identified by running a Gaussian filter. If control ChIP-seq data is available, the negative examples are identified as regions where the genomic coverage in the control dataset is higher than the one in the ChIP dataset. However, if there is no information to build a negative profile from, the negative profile is estimated from the sequencing noise.

3.4 Learning the peak shape

After identification of positives and negatives, outliers are removed. The Mahalanobis distance [Mahalanobis, 1936] between each example and the group of positives and negatives is computed and candidate regions with highest Mahalanobis distance are removed (see figure 9). The Mahalanobis distance was chosen as metric because it gives more importance to the most conserved part of the filter (typically, the maxima) than to the less conserved (typically, the noisy edges) and corrects for correlation between genomic positions.

The threshold is chosen using a robust estimator and a confidence level of $\alpha = 0.95$ as

$$\text{threshold} = \text{median}(d_i) + \Phi^{-1}(\alpha)\frac{\text{m.a.d.}(d_i)}{\Phi^{-1}(0.75)}, \tag{2}$$

where d_i is the Mahalanobis distance between the region i and its reference group, Φ^{-1} indicates the quantile of the standard normal distribution, m.a.d. indicates the median absolute deviation, and the term $\frac{\text{m.a.d.}(d_i)}{\Phi^{-1}(0.75)}$ is a robust estimate of the standard deviation under the assumption of normal distribution [Huber, 1981].

Once the positive and negative regions have been identified, the Transcription Factor ChIP-Seq learns a filter that matches the average peak shape, which we term Peak Shape Filter. The filter implemented is called Hotelling Observer [Hotelling, 1931] and was chosen because it is the matched filter that maximizes the Area Under the Curve of the Receiver Operator Characteristic (AUC-ROC), one of the most widely used measures for algorithmic performance.

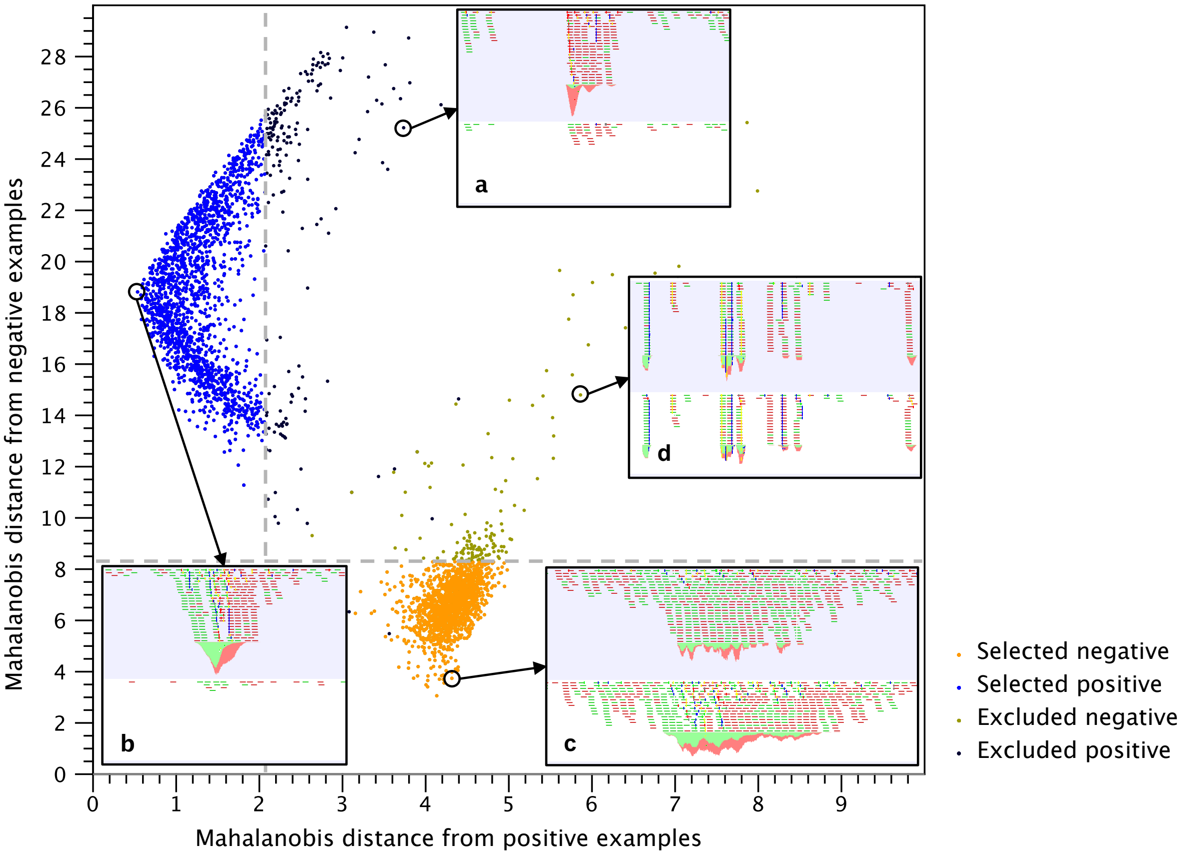

Figure 9: *Outlier detection. The plot shows the Mahalanobis distance of each candidate region to the set of positive and negative profiles. The threshold values for positive and negative samples are indicated by a vertical and a horizontal line, respectively. Inner figures a, b, c, and d show examples of profiles, displaying reads matching to ChIP-seq data (top part, light blue background) and reads matching to the genomic control (bottom part, white background):* ***a)*** *A noisy outlier is excluded from the set of positive profiles;* ***b)*** *A very clear peak is classified as positive profile;* ***c)*** *A peak with a high coverage in the control is classified as negative profile;* ***d)*** *A region with abnormally high coverage is excluded from the set of negative profiles.*

The Hotelling observer h is defined as:

$$h = \left(\frac{R_p + R_n}{2}\right)^{-1} (\mathbb{E}[X_p] - \mathbb{E}[X_n]) \,, \qquad (3)$$

where $\mathbb{E}[X_p]$ is the average profile of the *positive* regions, $\mathbb{E}[X_n]$ is the average profile of the *negative* regions, while R_p and R_n denote the covariance matrices between the *positive* and *negative* profiles, respectively. The Hotelling Observer has already previously been successfully used for calling ChIP-seq peaks [Kumar et al., 2013]. An example of the Hotelling observer is shown in figure 10.

Even though the shape of the Hotelling Observer is typically similar to the average profiles (figure 4), it is in fact modeling the shape that is maximally discriminative between positive and negative example and is therefore more similar to the difference between positive and negative examples.

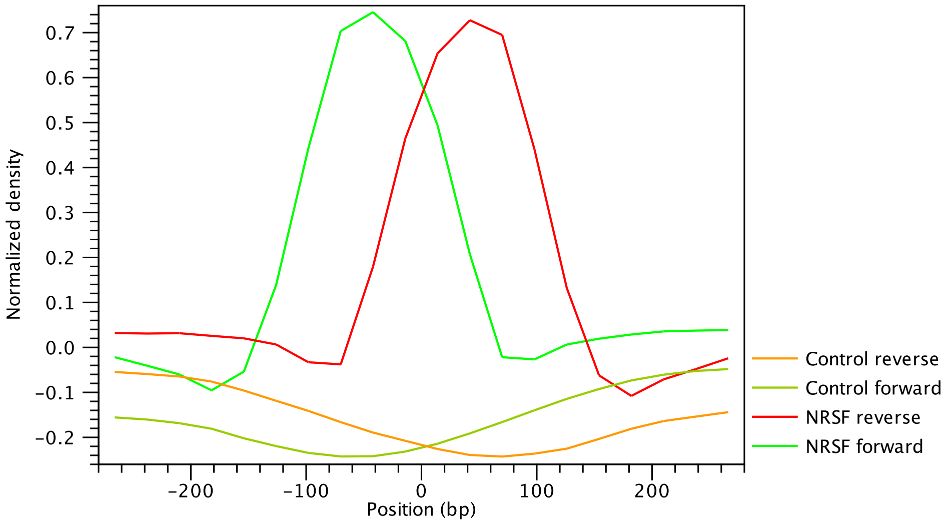

Figure 10: *Peak Shape Filter for the transcription factor NRSF.*

3.5 Peak Shape Score

The Peak Shape Filter is applied to the experimental data and a score is calculated at each genomic position (figure 8). The score is obtained by extracting the genomic coverage profile of a window centered at the genomic position and then comparing this profile to the Peak Shape Filter. The result of this comparison defines the Peak Shape Score. The Peak Shape Score is standardized and follows a standard normal distribution, so a p-value for each genomic position can be calculated as p-value $= \Phi(-$Peak Shape Score of the peak center$)$, where Φ is the standard normal cumulative distribution function.

3.6 Peak-detection

Finally, peaks are called by first identifying the genomic positions whose p-value is higher than the specified p-value threshold and which do not have any higher value in a window around them. The size of this window is determined by the filter as the longest distance between two positive values in the filter. These maxima define the center of the peak, while the peak boundaries are identified by expanding from the center both left and right until either the score becomes 0 or the peak touches a window boundary (figure 11).

3.7 The Transcription Factor ChIP-Seq

The Transcription Factor ChIP-Seq implements all the steps previously described in a single and easy-to-use algorithm. The input to the algorithm is mapped reads for ChIP-seq and genomic controls. The only parameter required is a p-value threshold. The results of the algorithm are:

QC Report The QC report contains metrics about the quality of the ChIP-seq experiment. It lists the number of mapped reads, the normalized strand coefficient, and the relative strand correlation for each mapping. Furthermore, the QC report shows the mean read length, the inferred fragment length, and the window size used to model the signal shape. In case the input contains paired-end reads, the report will also contain the empirical fragment length distribution. For details, see section 3.1.

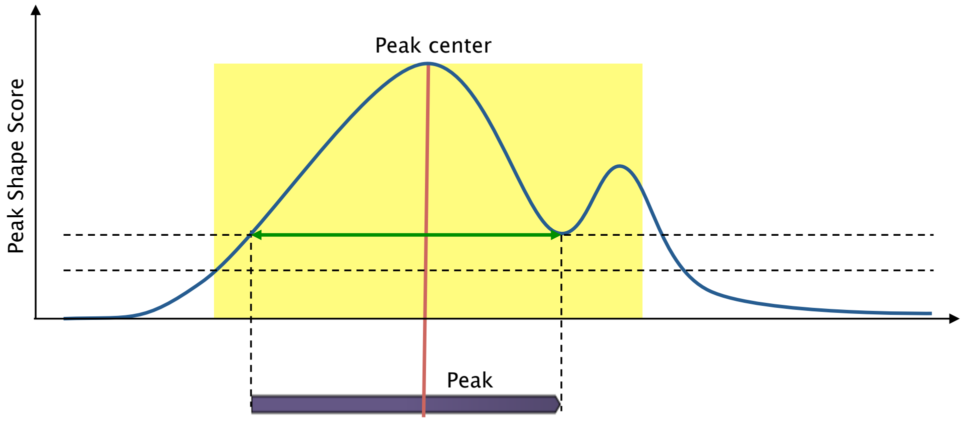

Figure 11: *Peak calling. After the center of the peak is identified (red line), the values in a window around the center are analyzed (yellow). The minima on the left and right side are identified (dashed horizontal lines) and the higher value is used as threshold. Next, the peak is expanded (green arrows) by adding all genomic position whose value is higher than the threshold to the left and to the right side of the center. The final peak is shown in purple.*

Peak Shape Filter the Hotelling Observer filter that was learned by the Transcription Factor ChIP-Seq (see section 3.4).

Peak Shape Score The peak shape score value for every genomic position (see section 3.5).

Peaks the list of all called peaks (see section 3.6).

While the Transcription Factor ChIP-Seq is the preferred way to analyze ChIP-seq data in CLC Genomics Workbench, we remark that advanced users interested in performing more complex analyses may also access the main components of the algorithm through the Advanced Peak Shape Tools Plugin. The following tools are provided:

The Learn Peak Shape Filter tool creates a new Peak Shape Filter from sequencing data and a set of positive and negative regions.

The Apply Peak Shape Filter tool applies a Peak Shape Filter to sequencing data to discover regions (peaks), which match a given peak shape.

The Score Regions tool scores genomic regions according to how well they match a given peak shape.

These tools are designed to be modular and can be easily combined to perform ad-hoc analyses. For example, the same Peak Shape Filters can be used for several experiments following similar experimental protocols, which allows the analysis of datasets with a small number of mapped reads or to compare datasets obtained in different biological conditions.

4 Performance evaluation

There is a number of ChIP-seq peak calling packages available, each with slightly different implementation-dependent strengths and weaknesses. See for example the extensive comparison conducted by Wilbanks and Facciotti [Wilbanks and Facciotti, 2010]. For clarity, in this paper we limit the comparative analysis of the Transcription Factor ChIP-Seq to three of the most popular state-of-the-art peak callers: the seqpeak tool included in the CisGenome software collection [Ji et al., 2008], the findPeaks tool included in the HOMER software collection [Heinz et al., 2010] (`http://biowhat.ucsd.edu/homer/chipseq/`), and the MACS software [Zhang et al., 2008, Feng et al., 2011, Feng et al., 2012]. These algorithms consistently rank among the top performing implementations [Rye et al., 2011] and will serve as reference points.

4.1 Gold-standard dataset

In this white paper we present benchmark results from calling peaks in experiments targeted at identifying transcription factor binding sites (TFBS). Benchmark datasets often provide control data together with data from specific experiments (see Landt et al., 2012 for guidelines on how to construct control samples). We will refer to the experiment samples as ChIP sample and to the control as control sample.

As a benchmark dataset we use the data published by Rye et al., 2011 as our gold-standard of truth. This dataset is based on expert curation of several hundred regions, each manually classified as either positive, negative, or ambiguous. The classification is done for three different ChIP-seq experiments, namely for the Transcription factors MAX, NRSF, and SRF. In contrast to synthetic "spike-in" data generated by some authors, the use of manually annotated real-world data has the advantage of a blind experiment in the sense that the gold-standard classification is not produced by the very same persons developing the peak calling algorithm.

The ChIP-seq datasets are:

MAX Published by Michael Snyder's lab at Yale University and generated from cell-line K562 targeting Myc-Associated factor X.

NRSF Published by the Myers Lab at the HudsonAlpha Institute for Biotechnology and generated from cell-line Gm12878 targeting the Neural Restrictive Silencer Factor (NRSF or REST) transcription factor.

SRF Published by the Myers Lab at the HudsonAlpha Institute for Biotechnology and generated from cell-line Gm12878 targeting the Serum Response Factor (SRF) transcription factor.

In table 1 we list the amounts of reads in each ChIP and control dataset (for MAX only one control replicate is available).

	MAX	NRSF	SRF
Experiment Rep. 1	7,916,698	16,145,592	12,750,756
Experiment Rep. 2	5,947,320	26,619,271	12,291,355
Control Rep. 1	13,510,465	16,377,339	14,164,649
Control Rep. 2	N/A	14,363,052	16,222,442

Table 1: *Reads in benchmark datasets used in Rye et al., 2011, available from ENCODE.*

The published manually classified peaks for the above three publicly available ChIP-seq datasets are summarized in table 2.

	MAX	NRSF	SRF
#Positive peaks	225	138	134
#Negative peaks	62	66	46

Table 2: *Peaks for benchmark datasets manually classified and used in Rye et al., 2011.*

4.2 Calculating performance metrics

The output from most peak callers is typically represented in the form of a ranked list of genomic sites, which are considered as peaks by the algorithm. This list of candidate peaks are ranked according to a statistical measure or score calculated by the algorithm, i.e. p-value, maximum coverage, fold enrichment, or a (log-transformed) combination thereof. Based on such ranked lists, the general framework for comparing prediction results of different predictors - with respect to a given gold-standard - is using Receiver-Operator Characteristic (ROC) curves, which plot the true positive over the false positive rate. Such plots are produced by going down the ranked list and evaluating each entry against the gold-standard. Every correctly called peak increases the number of true positives and the curve moves further upward. For every false prediction, the false positive rate increases and the curve moves to the right. Peaks without a reference classification are ignored in generating the ROC plots. Hence a perfect prediction method (which assigns a higher score to all positive peaks compared to any negative peak) would reach a 100 percent true positive rate before encountering the first false negative. A perfect classifier would directly jump to the upper left corner of the graph, while a random classifier would produce a curve close to the diagonal.

For the Transcription Factor ChIP-Seq, we investigated both the performance of the peak caller and the performance of the Peak Shape Score. The main advantage of using the Peak Shape Score is that the result provides single nucleotide precision. Regions from the gold standard are scored as depicted in figure 12.

4.3 Running Transcription Factor ChIP-Seq with different input datasets

First, we investigated how the presence of a control experiment and the treatment of replicate experiments would affect the performance of the Transcription Factor ChIP-Seq. The analysis was performed using single replicates, using both replicates, and using a single file containing the reads of both replicates. The algorithm was then run without using control reads. In each analysis, the gold-standard regions were scored using the maximum of the Peak Shape Score in the region (see figure 12) and AUC-ROC (Area Under the Curve of the Receiver Operator Characteristic) values were estimated (table 3).

The results in table 3 show small differences between the performances obtained by using different choices of input, suggesting that the performances of the Transcription Factor ChIP-Seq do not degrade significantly when fewer data is available and even when no control data is available. As expected, running the algorithm using all available data consistently gave top performances, indicating that there is no need to perform pre- or post-processing steps when analyzing datasets where replicates are available.

4.4 Results

We ran all peak callers (CLC bio, CisGenome, HOMER, and MACS) on the three datasets (table 1). All tools were run initially with their respective default parameters, as this is how the software

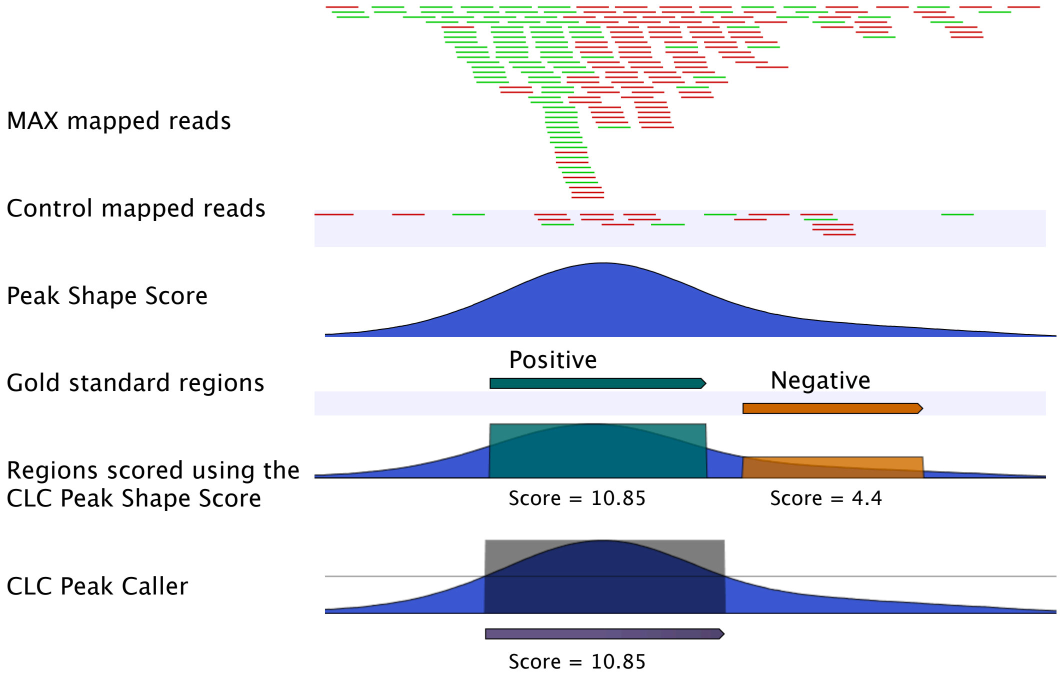

***Figure 12:** Assigning scores to validated regions. Two nearby regions, which are part of the gold standard, were annotated as positive (cyan) and negative (orange), respectively. Two methods for assigning scores are shown. **1) Transcription Factor ChIP-Seq:** The result of the peak calling is a single peak (purple bar) overlapping the region validated positive and the score associated with it is the maximum score within the peak, i.e. 10.85. No peaks overlapping with the negative were found so no score will be assigned to the negative region. **2) CLC Peak Shape Score:** The Peak Shape Score values within the positive and negative regions are extracted and the maximum values are used as score for the region. Note that this procedure always assigns a score to every region in the gold standard.*

Experiment ChIP	Control ChIP	MAX	NRSF	SRF
Rep. 1 and Rep. 2	Rep. 1 and Rep. 2*	0.94	0.98	0.97
Merged	Merged*	0.93	0.98	0.97
Rep. 1	Rep. 1	0.93	0.99	0.96
Rep. 2	Rep. 2*	0.95	0.97	0.97
Rep. 1 and Rep. 2	none	0.92	0.95	0.95
Merged	none	0.93	0.94	0.94
Rep. 1	none	0.90	0.95	0.94
Rep. 2	none	0.94	0.94	0.95

***Table 3:** Area Under the ROC Curve values for different input settings.*
** For MAX the only available control replicate Rep. 1 was used as input (see table 1).*

is used mostly. However, since all the peak callers by default do not output ambiguous peaks, many ROC curves would plateau and result in an unfairly low AUC-ROC value. Therefore, we relaxed parameters related to filtering of non-significant peaks, while leaving the other parameters unchanged. For CisGenome, we relaxed the cutoff for defining peak boundaries from 3 to 2.5; for HOMER, we relaxed the fold enrichment over input tag threshold from 4.0 to 2.0, the fold enrichment over local tag count threshold from 4.0 to 2.0, and removed the filtering step based on expected unique tag positions; for MACS we relaxed the q-value threshold from the default value of 0.05 to 0.25; and for the Transcription Factor ChIP-Seq, we relaxed the p-value threshold from 0.05 to 0.25 (see section 6). In all cases, the resulting AUC-ROC values obtained with these relaxed parameters were greater or equal to the AUC-ROC obtained using default parameters. We note that it was not feasible to remove the filtering steps altogether or to further relax the thresholds, because the resulting AUC-ROC values for CisGenome, MACS, and HOMER degraded in at least one dataset.

For CisGenome, we considered both the peaks (which we refer to as CisGenome) and the refined peak regions (which we refer to as CisGenome Refined), since the peak regions are typically very large. For MACS and HOMER, peaks were called after merging the reads from both replicates. For the Transcription Factor ChIP-Seq, both the Peak Shape Score and the results of the peak calling were collected as shown in figure 12. Then, for each dataset, an ordered list of peaks of decreasing confidence was obtained using the reported p-values of the peak regions. In this way, for each algorithm, the classified peaks are ordered according to the best scoring (lowest p-value) intersecting called peak or the worst p-value (1.0) if no called peak intersects. If ties exist in this ordering, peaks classified as negative are considered before peaks classified as positive. From this ordered list of positive and negative classified peaks, ROC curves and AUC-ROC values are produced. We discuss the results in the following section.

The ROC curves for the experiments using the MAX dataset are shown in figure 13.

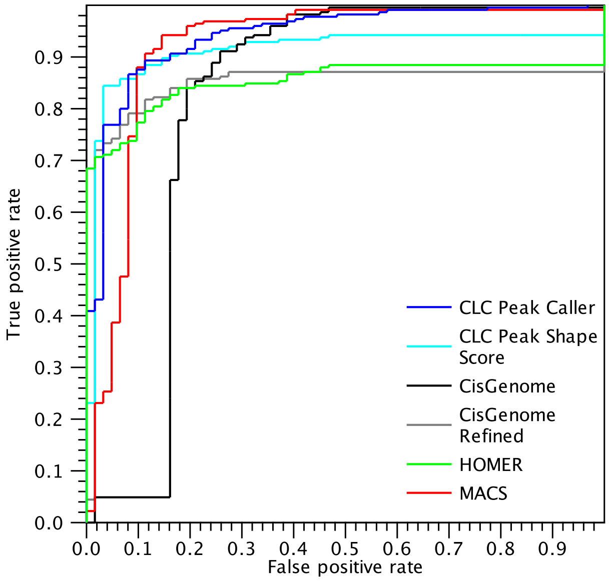

Figure 13: *ROC curves for peaks found in the MAX dataset.*

The first consideration is that HOMER makes its first mistake after reaching a true positive rate of 0.67, significantly later than the other algorithms. However, the performances of HOMER decline for more ambiguous peaks, resulting in a smaller AUC-ROC than MACS and the Transcription Factor ChIP-Seq. On the other hand, MACS makes more mistakes in the beginning, but it is able to identify nearly all peaks, resulting in a higher AUC-ROC value. The Transcription Factor ChIP-Seq behaves similarly to HOMER at low false positive rate but is able to call more peaks. We note that in this dataset there is a difference in the ROC curves of the CLC Peak Shape Score and the Transcription Factor ChIP-Seq. This is due to the fact that many regions of the gold standard are situated near the center of a peak. Therefore, in this context, the performance value depends significantly on the ability to identify the correct peak boundaries. For example, a negative region close to a peak region is shown in figure 12. In this example, the peak caller correctly identifies that there is no peak in the negative region, while the CLC Peak Shape Score approach assigns a moderate score to the negative region because it is in proximity of a peak. Conversely, a situation where a positive region is situated near a strong peak may be missed by the peak caller, which calls only the main peak and may not extend the boundaries enough. This makes the two ROC curves different, making the peak caller more appropriate for small false positive rates, whereas the Peak Shape Score approach assigns a score to every genomic position. The difference between CisGenome and CisGenome Refined is even more pronounced, as the peaks called by the default CisGenome are too large and often include nearby regions annotated as negatives sites. On the other hand, using refined peak boundaries drastically improves the performances of the algorithm for this dataset, although it misses some peaks of lower quality.

Figure 14 shows the ROC curves for called peaks in the NRSF dataset.

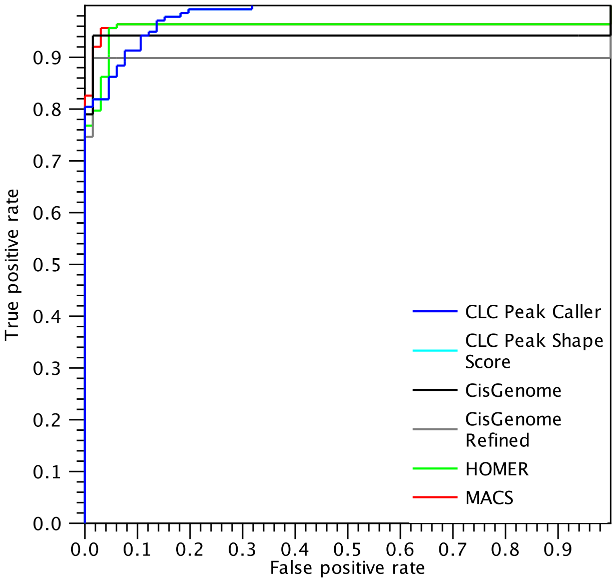

Figure 14: *ROC curves for peaks found in the NRSF dataset.*

The performances of all the algorithms are very good in this dataset and all algorithms make their first mistake after a true positive rate of 0.7. The main difference between the Transcription Factor ChIP-Seq and the other two algorithms is that, although it makes a few mistakes earlier, the peak caller is able to call nearly all peaks, resulting in a higher overall AUC-ROC value. In this dataset, the performances of the Transcription Factor ChIP-Seq and the Peak Shape Score are equal, because there is no ambiguity regarding peak boundaries. Therefore, the two ROC curves completely coincide and only the curve from CLC Peak Caller (blue) is visible. The performances of the two CisGenome variants are quite similar, but CisGenome Refined misses some peaks that the default CisGenome correctly identifies.

Figure 15 shows the ROC curves for called peaks in the SRF dataset.

Similarly to the NRSF dataset, most positive peaks are called by all algorithms and only a few mistakes are made, resulting in good performances for all the algorithms. MACS performs better than the other algorithms for small false positive rates, resulting in the highest AUC-ROC value. In this dataset, the difference between the peak caller and the Peak Shape Score is very small and the curves are very close to each other. The main difference is that the peak caller does not identify all peaks and plateaus slightly before reaching the top of the plot. On the other hand,

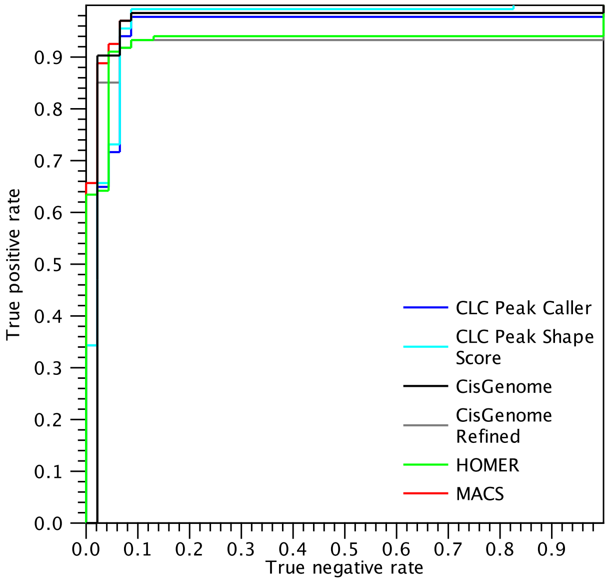

Figure 15: *ROC curves for peaks found in the SRF dataset.*

the Peak Shape Score gives a score to every region in the gold standard, so it is always able to reach a true positive rate of 1.

	MAX	NRSF	SRF
CLC Peak Shape Score	0.94	0.98	0.97
CLC Peak Caller	0.91	0.98	0.95
CisGenome	0.82	0.94	0.96
CisGenome Refined	0.84	0.90	0.91
HOMER	0.85	0.96	0.93
MACS	0.92	0.96	0.97

Table 4: *Area Under the ROC Curve values.*

Table 4 summarizes the results by the area under the ROC-curve values. MACS and the Transcription Factor ChIP-Seq have the highest values, while HOMER is often penalized because it misses several peaks and CisGenome does not often identify the correct peak boundaries. Even though it is hard to directly compare the results of the Peak Shape Score with more traditional peak calling algorithms, we observe that there are clear advantages in having a score with single nucleotide resolution, especially for calling low-quality peaks.

5 Conclusions

In this white paper, we discussed the state-of-the-art and current trends in the field of peak-detection algorithms that lead to the development of the new Transcription Factor ChIP-Seq. We described our implementation of a general and statistically well founded algorithmic engine, applicable to a wide range of different datasets. This flexibility is combined with specificity by automatically learning the characteristics of the signal present in the data.

5.1 Discussion

The systematic performance evaluation of the Transcription Factor ChIP-Seq on a published manually curated reference dataset shows that it compares favourably to popular current algorithms. It is important to note that the existing peak callers are readily tuned to detect narrow peaks at transcription factor binding-sites and hence are very well suited to the benchmark data. Hence, it is reassuring to see the new Transcription Factor ChIP-Seq performs on the same level and sometimes even better than CisGenome, HOMER, and MACS.

Besides the theoretical advantages of the peak-shape approach and the positive results obtained on the benchmark data, in practice it is equally important that the Transcription Factor ChIP-Seq wraps the underlying algorithm and statistics into an automated pipeline, which is simple to use for non-specialists. At the same time, we provide the option of further optimization by manually delineating positive and negative examples - a process that is much easier understood and visualized than abstract parameters. Moreover, it is of great importance to make sure that the input data has sufficient quality and consequently that the results can be trusted. Therefore, potential problems are highlighted by detailed quality reports (see section 3.1).

More advanced users can enter into an optimization cycle by iterating through the individual steps available in the advanced toolset as described in section 3.7. It is easy to see that the set of regions that was used to build the filter in the first place can be re-evaluated with the resulting filter, producing a ranked list of scores for each region. The user can then quickly inspect these regions visually and decide to exclude regions that do not clearly exhibit the peak-shape characteristics on a case-by-case basis. Similarly, other peaks detected by applying the filter to the entire dataset can be included into the training set if they display the characteristic shape. This way an updated set of regions is assembled from which an updated filter is generated, which can be applied and re-scored yet again. This learn-apply-score cycle converges very quickly and helps users to visually fine-tune the peak detection to find the exact signal they are looking for. It is important to note that our current implementation of the ChIP-seq peak-detection automatically performs a single round of the steps outlined above with very satisfactory results. To outperform the Transcription Factor ChIP-Seq in its automated mode users will need a significant amount of training and experience. A more detailed description of the advanced toolset will be available in a forthcoming advanced tutorial.

To our knowledge, the Transcription Factor ChIP-Seq is unique in its combination of flexibility and accuracy through the ability of building optimized filters for different analysis tasks and datasets. This enables sustained development and maintenance of the tool set in exploratory research, since the adaptation to different filters does not require changes to the underlying code base. At the same time, existing filters are transferable and can be applied to new datasets as they become available in production settings, so there is no need to re-learn and optimize the filters from scratch if the same peak-shape is to be detected.

5.2 Future Developments

The Transcription Factor ChIP-Seq is not a stand-alone tool, but represents an important building block within a larger ecosystem of NGS data analysis software. Hence considerations for further developments have to take into account the interplay with other related components. From a user centric view, major improvements will be gained by seamless interoperability with both upstream and downstream analysis steps. Since peak detection is now handled by a flexible and generally applicable algorithmic engine, the focus shifts from algorithmic issues to data integration, downstream analyses, and visualization capabilities.

A great deal of interoperability is already taken care of by the integration of the Transcription Factor ChIP-Seq into the track-based framework of CLC Genomics Workbench. By taking its inputs from tracks and producing outputs in the same format, it can be freely combined with other track-based tools. Therefore it is easily integrated into larger analysis pipelines and workflows while benefiting from the ongoing improvements to the visualizations and operations available for track-based data. In this context, the visual potential of the peak shape could be exploited even further (for example by displaying the shapes directly on the called peak regions) to arrive at an even better and more intuitive grasp of how well the applied filter fits the signal in the data.

A major advantage of the Transcription Factor ChIP-Seq is that it is not exclusively designed to detect signals in transcription factor ChIP-seq datasets. In principle, the methodology is applicable to detect signals in data from a wide range of different sequencing protocols, including FAIRE-seq, broad-peaks from ChIP-seq of histone-modifications, other epigenetic marks such as DNA methylation, and even DNaseI hypersensitivity footprints.

Building on the peak shape detection algorithm described here, we have recently developed a Histone ChIP-Seq plugin functionality for both Genomics Workbench (versions 8.5 and above) and Biomedical Research Workbench (versions 2.5 and above). It is well-sutied for detection of broad peaks characteristic of these other types of ChIP-seq datasets, striking a good compromise between sensitivity of detection, and computational feasibility of a genome-wide scale task of detecting peaks of undertermined and variable width. While a detailed discussion and quantitative benchmarking of this new tool is beyond the scope of this white paper, we find that its performance on the H3K36 trimethylation data is very satisfactory.

In addition, the methodology is capable of integrating multidimensional signals. Already the algorithm uses multidimensional inputs, for example by combining ChIP and control data or by using replicates. In fact, the filters could use any combination of read coverages or any other numeric value. Hence further developments could enable users to integrate the signal present in several datasets simultaneously, which will increase confidence and statistical power of the results.

For transcription factor ChIP-seq data, investigating the relationship between the detected peaks and TFBS motifs is one of the obvious next steps in the analysis. This means that the found peaks need to be easily searched and annotated with known motifs. In the absence of known motifs, the de-novo finding of over-represented motifs is of primary interest. While it is already possible to extract the peak-regions and subject them to motif finding with external tools (such as MEME [Bailey et al., 2006]), deeper integration with TRANSFAC® [Matys et al., 2006] will make both tasks much simpler in future.

Looking further ahead with respect to downstream integration, systems biology approaches such as the integration with metabolic or signaling pathways will become relevant. The details of mapping information from genomic coordinates onto biological networks is non-trivial, but can be broadly described as follows: First the summary statistics derived from the called peaks need to be mapped to the associated (closest) gene. Then the genes in turn can be used to link the information to nodes in the biological network (e.g. the associated proteins and their interactions). In the space of biological networks, an entirely new set of analytical options become available to researchers, such as disease associations, links to relevant scientific literature, and causal inference tools.

To put things in a broader perspective, we expect that the performance improvements to the

Transcription Factor ChIP-Seq described in this paper will in turn lead to improved performances of entire bioinformatics pipelines ranging from processing raw sequencing data via read-mapping, peak finding, and heterogeneous data integration all the way to biological network information. Providing this functionality in a robust and easily accessible way is a prerequisite to help researchers generate novel biological insights and testable hypotheses from their data.

6 Materials and Methods

The manually curated peak annotation by Rye et al., 2011 are obtained via `http://tare.medisin.ntnu.no/chipseqbenchmark/downloads/ChIPSeq_files_in_bed_format/`

The corresponding original data from ENCODE (release 1) are available via the following links: `ftp://encodeftp.cse.ucsc.edu/pipeline/hg18/wgEncodeHudsonalphaChipSeq/release1/` and `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeHudsonalphaChipSeq/release1/`

CisGenome version 2.0 was downloaded from `http://www.biostat.jhsph.edu/~hji/cisgenome/index_files/download.htm`. The seqpeak tool was run with parameters "`-c 2.5`". The columns "`start`" and "`end`" were chosen to define the peak regions and the columns "`left_peakboundary`" and "`right_peakboundary`" were chosen to define the refined boundaries. In both cases, the "`rank`" column was used to build the ROC curve.

HOMER version 4.7 was downloaded from `http://homer.salk.edu/homer/download.html`. The findPeaks tool was run with parameters "`-style factor -F 2 -L 2 -C 0`" the column "`findPeaks Score`" was used to build the ROC curve. Note: removing the filtering options -F -L -C resulted in dramatically reduced performances, so the threshold for -F and -L was relaxed.

MACS version 2.1.0 was obtained from `https://github.com/taoliu/MACS/`. The software was run with the parameter "`-qvalue 0.25`" and the column "`q-value`" was used to build the ROC curve.

The Transcription Factor ChIP-Seq version 1.0 was run with the parameter "`-p-value 0.25`". The Score Regions tools from Advanced Peak Shape Tools Plugin version 1.0 beta 3 was used to assign Peak Shape Scores to gold standard regions.

References

[Bailey et al., 2006] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(suppl 2):W369–W373.

[Feng et al., 2012] Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat Protoc*, 7(9):1728–1740.

[Feng et al., 2011] Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*, 34(2):2–14.

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

[Heydarian et al., 2014] Heydarian, M., Romeo Luperchio, T., Cutler, J., Mitchell, C., Kim, M.-S., Pandey, A., Soliner-Webb, B., and Reddy, K. (2014). Prediction of gene activity in early B cell development based on an integrative multi-omics analysis. *J Proteomics Bioinform*, 7(2):050–063.

[Hotelling, 1931] Hotelling, H. (1931). The generalization of Student's ratio. *Ann Math Statist*, 2(3):360–378.

[Huber, 1981] Huber, P. (1981). *Robust Statistics*, page 181. Springer.

[Ji et al., 2008] Ji, H., Jiang, H., Ma, W., Johnson, D., Myers, R., and Wong, W. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology*, 26(11):1293–1300.

[Kharchenko et al., 2008] Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26(12):1351–9.

[Kornacker et al., 2012] Kornacker, K., Rye, M. B., Håndstad, T., and Drabløs, F. (2012). The Triform algorithm: improved sensitivity and specificity in ChIP-Seq peak finding. *BMC Bioinformatics*, 13:176.

[Kumar et al., 2013] Kumar, V., Muratani, M., Rayan, N. A., Kraus, P., Lufkin, T., Ng, H. H., and Prabhakar, S. (2013). Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol*, 31(7):615–22.

[Landt et al., 2012] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, 22(9):1813–31.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

[Marinov et al., 2014] Marinov, G. K., Kundaje, A., Park, P. J., and Wold, B. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, 4(2):209–23.

[Matys et al., 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–10.

[Mendoza-Parra et al., 2013] Mendoza-Parra, M.-A., Nowicka, M., Van Gool, W., and Gronemeyer, H. (2013). Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics*, 14:834.

[Rye et al., 2011] Rye, M. B., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res*, 39(4):e25.

[Stanton et al., 2013] Stanton, K. P., Parisi, F., Strino, F., Rabin, N., Asp, P., and Kluger, Y. (2013). Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures. *Nucleic Acids Res*, 41(16):e161.

[Wilbanks and Facciotti, 2010] Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137.