Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts

Adam M. Szalkowski and Christoph D. Schmid

Abstract
The current understanding of the regulation of transcription does not keep the pace with the spectacular advances in the determination of genomic sequences. Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) promises to give better insight into transcription regulation by locating sites of protein–DNA interactions. Such loci of putative interactions can be inferred from the genome-wide distributions of ChIP-seq data by peak-calling software. The analysis of ChIP-seq data critically depends on this step and a multitude of these peak-callers have been deployed in the recent years. A recent study reported severe variation among peak-calling results. Yet, peak-calling still lacks systematic quantitative benchmarking. Here, we summarize benchmarking efforts and explain potential drawbacks of each benchmarking method.

Keywords: ChIP-sequencing; peak-calling; benchmarking

INTRODUCTION
Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) can be applied to the analysis of chromatin structure [1] and to the identification of protein binding sites [2]. Thereby, it promotes the understanding of transcriptional regulation, which is fundamental to biotechnology and medical research. Compared to ChIP-chip, the micro-array based predecessor technology, ChIP-seq features a better coverage, enhanced resolution and an increased signal-to-noise ratio [3]. Peak-calling is a critical layer in the processing of ChIP-seq data. Its role is to infer the actual binding loci from the positional distribution of tags, i.e. sequenced DNA fragments mapped onto a reference genome sequence. Development in this domain is advancing at a staggering pace; however, it is unfortunate that it is not often accompanied by adequate benchmarking.

Here, we briefly summarize ChIP-sequencing and survey recent developments in peak-calling. Furthermore, we will review common methods for verification of peak-calling results and point out strengths and limitations.

PEAK-CALLING FOR CHIP-SEQ
In ChIP-seq, chromatin from living cells is cross-linked with interacting proteins and sheared (Figure 1). Specific antibodies are used to bind the target proteins in the chromatin to enrich for DNA segments attached to these complexes. After the reversal of the cross-linking, DNA fragments are purified, for instance, by gel electrophoresis. The majority of the current massively parallel sequencing technologies include PCR-like amplification steps prior to sequencing. The resulting pool of DNA fragments contains specifically immunoprecipitated DNA at an only minor fraction (reportedly <1% [4]) within a large proportion of background DNA.

After sequencing and base-calling, the resulting short reads are mapped onto a reference genome assembly. This is usually the computationally most
expensive part, particularly because the mapping has to account for a few mismatches which are due to either sequencing errors or to single nucleotide variations (SNVs). Typically, only tags with a single matching locus on the genome are considered. The positions of mapped tags are then the input data for peak-calling algorithms, which predict protein binding sites from the distribution of tags on the genome.

The major difficulty in peak-calling is to determine the significance of peaks, i.e. local clusters of tags, over a background distribution of tags. Due to a high variability in the background distribution and the relatively small amount of enriched DNA, it appears advantageous to model the background distribution statistically and to search the data set for local deviations from this distribution. Several models explaining tag densities in the background have been proposed including the Poisson distribution [5], the Negative Binomial distribution [6] and Hidden Markov Models [7]. Genomic DNA not enriched for binding fragments, i.e. input DNA or control data, should conceivably contain only

**Figure 1:** Workflow of a ChIP-seq analysis: Chromatin in the nucleus (1) is cross-linked and sheared (2), followed by enrichment of complexes containing the target protein using immunoprecipitation (3). Short reads obtained from massively parallel sequencing (4) are mapped to a reference genome (5) yielding in a distribution of tags on the genome (adapted from http://en.wikipedia.org/wiki/File:ChIP_sequencing2.png 2008-02-29).
background distributions, leading to approaches fitting the parameters for the background model based on control samples and empirically estimating false discovery rates (FDRs) at the corresponding genomic loci [8]. Accordingly, additional types of control samples are commonly used including ChIP without antibodies (mock IP), or with antibodies lacking specificity against DNA binding proteins. Put in a simplified way, if clusters of reads are also observed in background distributions, a co-localization of two such peaks in both experiment and control would lower the likelihood for the detection of a binding site.

Signal density is further influenced by the presence of short subsequences occurring multiple times in the genome. Tags mapping to such non-unique sequences are typically excluded and therefore some authors [9] propose a correction by computing an index called ‘mappability’ along the reference genome. Copy number variations, i.e. duplications or deletions in the sample compared to the reference genome, may be additional sources of artificial read densities. To consider several different biases, a number of recent tools [8, 10] combine several model parameters derived from control data in the peak-calling.

Peak-calling in a strict sense requires precisely localized binding sites. Such defined and consistently occupied sites may not exist for certain transcription factors with diffuse binding footprints (e.g. E2f1 and RNA Pol II) or for nucleosomes carrying modified histones. Another difficulty is the recognition of multiple partially overlapping peaks. The ChIP-seq procedure offers the possibility to recognize and separately process sequence reads from each end of the pulled down fragments, exploiting differentiated mapping to the + and the − strand. The exact mapping of the ChIP fragment ends allows consequently for more accurate delineation of DNA regions that interact with the targeted protein [11]. A mirrored distribution resulting from both ends of immunoprecipitated fragments is commonly exploited as criterion for true signal in peak-calling. Recent approaches involve an iterative blind deconvolution of overlapping peaks [12] or a modeling of peak shape and estimation of peak parameters using an Expectation–Maximization algorithm [13]. A number of recent reviews focus on ChIP-seq technology, applications and software [3, 14–16].

A very important, yet largely overseen challenge in peak-calling is benchmarking. Due to the lack of adequate and comprehensive performance evaluations of current peak-calling software, it is difficult for users to select appropriate tools for their ChIP-seq data processing.

**BENCHMARKING OF PEAK-CALLING**

Several methods for assessing the performance of peak-calling algorithms have been proposed. As a common scheme, the presentation of novel tools usually displays a selection of performance measures, for which the new method compares favorably. The following section highlights benchmarking approaches based on co-occurrences of binding motifs, on experimental verification, on estimations of FDR from the primary data or from control data and on simulation.

**Co-occurrence of binding motifs**

The most prominent verification method is searching for instances of sequence motifs binding the target protein at the predicted binding loci [2, 5, 7, 8, 12, 13, 17, 18]. Two quantitative measures of quality can be derived from this test: the fraction of peaks that have a motif in their vicinity and the distance of the predicted site to the recognized motif. In general, the higher the content of binding motifs in comparable sets of inferred bindings sites, the better the performance of the corresponding peak-caller. The comparison of these quantitative measures should be unbiased provided the binding motif has been inferred from independent data. If peak-callers start using motif discovery to support identification of binding sites [12], there is a risk of circularity. Variants of this method include aligning against a database of motifs to verify the enrichment of the expected motif [7, 13, 19] or searching for motifs of known co-factors near the primary predicted binding loci [8]. The interpretation of measures based on binding motifs may however be impaired by limitations in the current understanding of protein–DNA interactions. For instance, reports of multiple divergent binding motifs for at least some of the DNA binding proteins [20] are not yet incorporated in these benchmarking methods.

**Experimental verification**

An independent and promising evaluation method is the determination of true positive and true negative binding sites based on experimental verification using
quantitative PCR (qPCR) or transfection assays [2]. While this approach has a straightforward readout with the area under the ROC curve, drawbacks include the limited throughput of experimental assays targeting single loci and the fact that the actual interaction of proteins with DNA may differ between different cell lineages, tissues and stages of development [21]. Therefore, experimental verifications from the literature performed on specific cell populations might not represent a universally true solution.

Spike-in samples
Benchmarking of ChIP-seq analysis tools is thus mainly impaired by the lack of a widely accepted gold standard to assess the performance of different analysis tools for a specific population of cells at a genome-wide scale. A corresponding problem has been addressed previously in a systematic evaluation of ChIP-chip experiments [22]. To alleviate the lack of verified solutions, the chromatin immunoprecipitation step was replaced by nearly 100 human sequences at various concentrations mixed into (‘spike-in’) human genomic DNA. Eight specialized research groups performed independent hybridizations of identical samples to four tiling array platforms in a blinded manner and made predictions of the spike-in locations. Given the enhanced positional resolution of ChIP-seq, realistic spike-in samples for ChIP-seq should ideally also reproduce the fuzziness of positions and lengths of ChIP fragments. Presently, unprocessed PCR products with a uniform length determined by the primer sequences would likely lead to an aberrant behavior in the peak-calling, especially for those algorithms which include the assessment of the positional distribution of reads within a peak.

Measures based on simulation
On the grounds that no spike-in data for ChIP-seq was available in public data repositories, ChIP-seq data sets from control experiments were supplemented with simulated reads from randomly distributed peaks [4, 6]. One such simulated spike-in approach was carried out as a community effort [23], but has not (yet) been formally summarized in a peer-reviewed publication. Furthermore, a study presenting a framework for simulating ChIP-seq reads [24] reports markedly non-uniform distributions difficult to model correctly. Biases missed by current simulation frameworks may include the frequently observed higher density of reads at regions of open chromatin, for instance, at the transcription start site (TSS), co-localized with an increased density of transcription factor binding sites. On the other hand, clusters in control data tend also to co-localize with open chromatin, putatively arising from higher probabilities of strand breaks due to an open chromatin structure [25, 26]. Current simulation models are furthermore unable to reproduce amplification biases in PCR resulting, for instance, from base composition effects [27]. Thus, current simulation-based approaches cannot realistically predict performance on real samples. Moreover, such assessment is biased in favor of peak-callers based on the same or closely related models used for simulation.

Measures based on consistency
Apart from the above measures, some studies propose more qualitative measures, which do not allow for a direct comparison of the performance of algorithms. Yet, they might still be valuable for assessing the reliability of individual methods. Some authors compare de novo motifs derived from regions around peaks to established motifs [2] or assess the presence of a newly discovered motif in all peak sets [8, 17]. However, the latter approach only attests a certain degree of similarity among binding sites and tends to give overconfident estimates of true binding motif content. Another study also examined the correlation of peak predictions with expression levels of downstream genes [2, 8]. Obviously, these approaches rely on a target protein with well-characterized binding motifs or acting as functional transcription enhancer/inhibitor, respectively. Other authors compare consistency among different approaches by examining the overlap between result sets from different peak-callers [7, 10] or results from ChIP-chip experiments [6]. Yet others consider the correlation between two ChIP-seq data sets obtained with different antibodies directed against the same target protein, assessing thereby the overlap of potentially variable specificity of different antibodies [8].

Benchmarking studies
A practical comparison of peak-callers [28] exemplifies possible approaches and also difficulties in the benchmarking. This study quantitatively evaluates the performance of nine peak-callers on a set of 10 pre-processed ChIP-seq data sets derived from the Illumina sequencing platform, including biological
replicates and control samples. Performance measures involve the reproducibility between different methods or among several biological replicates of a ChiP-seq experiment, the content of binding motifs and experimental verification using qPCR. Regrettably, the study yields in very unequal and seemingly inconsistent performance measures for each algorithm. Factors accounting for varying performance measures likely include limitations in the design of the study as well as in the available data sets. The performance measures used in this study may insufficiently consider effects arising from significantly differing numbers of predicted peaks as well as dissimilar sizes of the peak regions.

A second benchmarking study [18] evaluates the performance of eleven peak-callers based on extensive pairwise comparisons, on co-occurrence of binding motifs and on corresponding data sets derived from experimental verification experiments. The extensive pairwise comparison of result sets does not provide definitive conclusions on the sensitivity and the specificity of peak-calling results, given the trivial fact that sets with small peak numbers and width are prone to display a better relative overlap within larger sets of wider regions. Similarly, the authors admit that there is not sufficient qPCR data available to reliably estimate the number of false positives. Instead, they use qPCR results and high confidence motif occurrences to analyze sensitivity on equally sized sets of highest-ranking peaks according to peak significance. Specificity and spatial precision is only measured by occurrence of canonical binding motifs within the peak region. Unfortunately, these measures do not clearly stratify the peak-callers, and the study does not give a universal recommendation on a specific tool. Possibly related to presently unknown biases, the study reports a striking positional offset of the closest binding motif for only two of the eleven assessed tools. It also concludes on an enhanced positional precision of tools exploiting the directional information of the mapped reads.

Regrettably, this study focuses on ChiP-seq data sets with available control data and well-established canonical binding motifs, and thus considers only three ‘classical’ data sets. It is conceivable that at least some of the algorithms and default parameters were tailored to one of the few data sets publicly available in the early phase of ChiP-seq, which may explain variable performance on diverse data sets. Thus, a shortage of data in a benchmarking study might give an advantage to software developed using the same data sets for evaluation.

Both benchmarking studies have in common that they use only the default settings of parameters of each algorithm, which might not lead to optimal results on each data set. Yet, the comprehensive exploration of the parameter space is for practical reasons currently not possible.

A striking observation in some of these comparisons is the effect of the inclusion of control data in the peak-calling. The use of input DNA has its origins in ChiP-chip experiments, and early approaches [2] introduced the concept in the ChiP-seq data analysis. With the exception of FindPeaks [29] and GeneTrack [30], all peak-callers in the aforementioned comparisons can use control data in some manner. Interestingly enough, FindPeaks is not among the ‘worst’ performers. Also among the peak-callers with a conditional use of control data, the direct comparison displays a very limited, if any, advantage of the inclusion of control data, if assessing the reproducibility or the content of consensus binding motifs. These observations are supported by a recent study assessing the content of STAT1 consensus binding sites in sets derived from different peak-callers [31]. As a hypothesis, peaks from true binding sites located, for instance, in open chromatin close to the TSS might obtain statistical scores below the significance threshold due to increased densities of reads related to strand breaks in open chromatin in the control set. A critical aspect might also represent the early cycles of PCR-based amplification steps with low DNA concentrations. It appears therefore not granted that independent control experiments exactly reproduce the background distribution observed in a specific sample experiment. Independent benchmarking approaches are thus also required to qualify an appropriate use of control data sets.

Summarizing the benchmarking of peak-calling, ChiP-seq is a broadly embraced method with enhanced signal-to-noise ratio and positional resolution as compared to hybridization-based methods. Given the still incomplete characterization of putative biases in the methodology or in the underlying biology, some of the current analysis methods may misleadingly suggest the determination of exact solutions applying precise values for statistical thresholds. The grey area within the solution space is typically rather large and therefore subtle changes in parameters and in thresholds can conceivably have a considerable impact on the results.
CONCLUSION
As pointed out previously [18, 28], most peak-callers exhibit very unequal performances on different data sets. While ChIP-seq data may display a very clear-cut picture at certain individual loci, the biological conclusions at a whole genome scale can differ severely depending on the adopted peak-calling algorithm.

The review of previous approaches may lead to the following general recommendations for benchmarking studies. To allow for generalized conclusions, a first prerequisite is large and diverse ChIP-seq data sets comprising different types of controls including whole-cell extracts (input DNA), ChIP with antibodies lacking a specificity for DNA-binding epitopes, and samples with deficient DNA-binding activity of the target protein (abolished activation pathway, mutants, etc.). If the DNA binding activity can be induced, for instance, by an activation signal, this provides an excellent negative control sample with a very comparable population of cells and thus DNA preparations differing only in the DNA binding of the target protein. In practice, though, an incomplete abolishment of the DNA-binding activity or secondary effects caused by the deficiency in the DNA-binding might slightly blur the expected differences. Despite these limitations, a negative control involving a ChIP step might be preferable to input DNA. Having such a panel of controls for each data set would allow for a systematic evaluation of most peak-calling algorithms and their divergent treatment of controls. Comparing data sets generated in different labs and on multiple platforms in such a benchmarking might in addition reveal platform-specific biases.

We largely discounted the mapping of sequencing reads in this review, assuming only minor consequences on the performance of peak-callers. Yet, different mapping algorithms display at times astonishingly diverging results, and thus the uniform remapping of raw ChIP-seq reads is preferable to pre-processed data sets potentially incorporating yet unknown mapping biases.

Another unexplored area concerns the evaluation of ChIP-seq of proteins with a diffuse binding pattern, such as epigenetic histone modifications on nucleosomes distributed over larger genomic regions. Algorithms determining wide genomic ranges with increased tag densities are conceivably better suited to analyze this type of data distributions than algorithms which define peak positions based on local maxima in the tag density. Given the presumed absence of conserved binding motifs, an evaluation of a simple motif content is not applicable either in this case. It might be possible to determine an alternative encoding of binding preferences (nucleotide composition, DNA accessibility, etc.) and evaluate correlation with these measures.

Measures evaluating the performance of peak-callers should consider the notion that the current cellular state can influence the interactions of DNA-associated proteins at specific genomic loci. Thus measures of reproducibility among different ChIP-seq data sets targeting the same protein in replicate cell populations might tell more about the robustness of the biological system than the performance of the peak-callers. As an alternative approach, we suggest to assess robustness on replicate data sets created by computational re-sampling from a ChIP-seq data set in a bootstrap or jackknife style [32].

The content of sequence motifs within the sets of inferred binding sites represents a robust evaluation measure. Still, existing limitations, including unknown (secondary) binding preferences, are improved by expanding databases [33] and may be supplemented by data derived from alternative experimental approaches including protein binding arrays [20]. Evidently, the sequence motif must be defined independently from any of the data sets under investigation. The latter must contain identical numbers of top-ranking inferred binding sites. Larger sets likely contain a larger proportion of lower affinity binding sites, e.g. with a lower number of ChIP-seq reads. Low affinity sites may not be recognized by sequence motifs and thus the proportional content of sequence motifs might differ for larger sets. Given that the occupation of potential binding sites likely depends on factors additional to the nucleotide sequence, e.g. DNA accessibility related to chromatin modifications; the absolute content of sequence motifs in independent ChIP-seq experiments can not be used as performance measure.

Spike-in experiments with precisely defined true positive and negative solutions would obviously represent a significant advance for the benchmarking of peak-callers. A tantalizing option for benchmarking purposes would be the development of procedures modifying, for instance, PCR products with a variability in fragment length and positions resembling that of ‘real’ immunoprecipitated DNA fragments. Other options include the further
characterization of ChIP-seq distributions and eventually a computational simulation closer to real data.

Given the magnitude of such a benchmarking study, a community effort possibly in the context of the popular DREAM challenges [34] appear to represent an optimal framework. The organizers of DREAM challenges typically set up the contest based on anonymized data sets and define the evaluation criteria. By inviting, for instance, the developers of several peak-calling software to participate in the prediction, the choice of optimal parameter settings for each tool is delegated to specialists. Such a contest might lead to the development of a comprehensive benchmark suite for peak-callers. The design of novel challenges will certainly benefit from the experience of previous benchmarking contests.

On another topic, the availability of genome-wide protein-DNA interaction data linked to corresponding expression data is spurring the development of a novel generation of motif discovery algorithms. Past benchmarking studies of motif discovery have demonstrated a limited performance of the tools available at the time [35, 36]. Thus with faster development cycles, frameworks established in the context of benchmarking studies might have an extended lifespan and can be re-applied to novel tools.

In conclusion, benchmarking approaches in peak-calling in ChIP-seq have to date not been performed in a sufficiently stringent and comprehensive manner. Thus, they did not always yield the conclusive outcomes expected from them. Given the interest and the impact in the scientific community, systematic benchmarking studies of ChIP-seq peak-calling software are needed to guide current and future uses and foster future improvements.

Acknowledgements
We thank Christophe Dessimoz from ETH Zurich as well as several lab members for their feedback on the draft. This article started as an assignment for the graduate course “Reviews in Computational Biology” (263-5151-00L) at ETH Zurich.

FUNDING
This work was supported by the Swiss Tropical & Public Health Institute and A.M.S. was supported by the Swiss National Science Foundation grant 31003A_127325 to M. Anisimova.

References

Key points
- We survey recent developments in ChIP-seq peak-calling, which is currently advancing at a staggering pace.
- We review common methods for verification of ChIP-seq peak-calling results.
- ChIP-seq peak-calling is not often accompanied by adequate verification of results.
- New systematic and comprehensive benchmarking studies are needed.


