Published online 2 August 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 20* ***e153***
doi: 10.1093/nar/gkw680

Differential peak calling of ChIP-seq signals with replicates with THOR

Manuel Allhoff[1,2,3], Kristin Seré[3,4], Juliana F. Pires[1,3,5], Martin Zenke[3,4] and Ivan G. Costa[1,2,3,*]

[1]IZKF Bioinformatics Research Group, RWTH Aachen University Medical School, Pauwelsstr. 19, 52074 Aachen, Germany, [2]Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Schinkelstr. 2, 52062 Aachen, Germany, [3]Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Pauwelsstr. 20, 52074 Aachen, Germany, [4]Institute for Biomedical Engineering, Department of Cell Biology, RWTH Aachen University Medical School, Pauwelstr. 30, 52074 Aachen, Germany and [5]Statistics Department, Federal University of Paraiba, Cidade Universitária, 58059-900, João Pessoa, PB, Brazil

Received March 14, 2016; Revised July 15, 2016; Accepted July 20, 2016

ABSTRACT

The study of changes in protein–DNA interactions measured by ChIP-seq on dynamic systems, such as cell differentiation, response to treatments or the comparison of healthy and diseased individuals, is still an open challenge. There are few computational methods comparing changes in ChIP-seq signals with replicates. Moreover, none of these previous approaches addresses ChIP-seq specific experimental artefacts arising from studies with biological replicates. We propose THOR, a Hidden Markov Model based approach, to detect differential peaks between pairs of biological conditions with replicates. THOR provides all pre- and post-processing steps required in ChIP-seq analyses. Moreover, we propose a novel normalization approach based on housekeeping genes to deal with cases where replicates have distinct signal-to-noise ratios. To evaluate differential peak calling methods, we delineate a methodology using both biological and simulated data. This includes an evaluation procedure that associates differential peaks with changes in gene expression as well as histone modifications close to these peaks. We evaluate THOR and seven competing methods on data sets with distinct characteristics from *in vitro* studies with technical replicates to clinical studies of cancer patients. Our evaluation analysis comprises of 13 comparisons between pairs of biological conditions. We show that THOR performs best in all scenarios.

INTRODUCTION

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments (1) have revolutionized our knowledge of chromatin structure and transcription factor regulation (2). Initial applications of ChIP-seq were based on measuring histone modification and transcription factors of unrelated cell lines (2). The next challenge is the study of protein–DNA interactions of dynamic systems, such as cell differentiation (3–5), response to treatments (6) or the comparison of healthy (3–5) and diseased individuals (4,7). ChIP-seq studies of clinical scenarios are of particular interest, as they allow detection of epigenetic markers and regulatory single nucleotide polymorphisms (SNPs) (8).

These applications require the joint analysis of ChIP-seq data with several technical or biological replicates. However, ChIP-seq is a multi-step experimental protocol, where each step introduces distinct sources of potential artefacts (9). Among others, these artefacts arise from bias of DNA fragmentation to open chromatin regions, variation of IP efficiency due to antibodies, as well as polymerase chain reaction (PCR) amplification and sequencing depth bias. These artefacts produce ChIP-seq experiments with distinct signal-to-noise ratios, even when they are produced in the same lab and follow the same protocols (9,10). Moreover, the clinical samples, where patients have a distinct genetic background and samples may arise from heterogeneous cell populations, introduce further variation to the ChIP-seq signals (11). All these artefacts impose great challenges to the computational analysis of ChIP-seq data.

The great majority of computational tools have concentrated on the single signal condition peak calling (SPC) problem, i.e. detection of genomic regions with putative protein–DNA interactions in individual ChIP-seq experiments (12–15). Later, detecting peaks in replicates of ChIP-seq data of a single condition has been investigated. For

[*]To whom correspondence should be addressed. Tel: +49 241 80 80270; Email: ivan.costa@rwth-aachen.de

example, the ENCODE project, which suggests ChIP-Seq experiments should contain at least two biological duplicates (16), proposed the use of irreproducible discovery rate (IDR) as a post-processing step to find common peaks after the application of SPC methods to individual replicates (16,17). Recently, Ibrahim *et al.* (18) proposed a method for the joint analysis of ChIP-seq replicates for the SPC problem. Their method detects peak boundaries with higher precision than identifying common peaks in replicates with IDR or pooling ChIP-seq reads of replicates.

Given the lack of appropriate tools for the analysis of distinct cell types, the initial strategy for the analysis of differentiation processes was the evaluation of peaks produced by SPC methods which were only detected in one of the cell types (19). Later, methods solving the differential peak calling (DPC) problem, i.e. the detection of genomic regions with changes in ChIP-seq profiles between two distinct samples, have been proposed (20–28). Differential peak calling approaches can be categorized in two broad classes: two-stage (21–23) and one-stage (20,24–28) approaches. Two-stage approaches are based on the use of candidate peak regions which are detected by SPC methods. These lists of candidate peaks are then applied to methods tailored for the differential expression analysis of RNA-seq data such as DESeq (29). While these methods can cope with technical and biological replicates in their second step, they have two main caveats. First, their DPs are restricted to their initial candidate regions. Therefore, they fail to detect subtle changes within these candidate regions (25,30). This is particularly crucial for ChIP-seq data of histone modifications, where differential peaks occur in small regions within larger genomic regions with high ChIP-Seq signal. This is not an issue for ChIP-seq from transcription factors that mostly happens in small isolated peaks associated to a single protein binding site. Second, two-stage DPC methods usually do not provide any preprocessing steps crucial for ChIP-seq analysis, such as fragment size estimation, GC-bias correction and input-DNA subtraction (31,32).

One-stage DPC methods are based on segmentation methods, such as Hidden Markov models (HMMs) (20,25,28) or sliding window based approaches (21,22,24,26,27). These methods solve most of the issues not addressed by two-stage DPC methods. However, sliding window approaches are sensitive to the window size. While large windows will fail to detect small changes in peaks on histone modifications, small windows can generate overly fragmented peaks. Methods based on HMMs, which intrinsically segment the signals on windows with varying size, are capable of detecting subtle changes in DNA–protein interactions (20,25,28), i.e. partial gains/losses of histone modifications within large chromatin domains. They have been successfully used by us (25,33) and others (28) in the analysis of cell differentiation and treatment response studies (30). Currently, only window based approaches support DPC with technical or biological replicates (24,26,27).

Furthermore, some aspects have been poorly addressed by current DPC methods. It is crucial to evaluate the strategy for the normalization of replicates. All DPC methods dealing with replicates (23,24,26,27) use a weighted trimmed mean of M-values (TMM) (34). This strategy was devised for gene expression experiments, which assumes that counts of most observations (genes or peaks) do not change. This is not necessarily the case for protein interactions, as two distinct cells can have distinct amounts of proteins or histone modifications bound to their DNA (9). Particularly problematic in this normalization approach is the effect of replicate specific background noise.

Moreover, evaluating DPs is still an open problem as there is no gold standard. We have previously developed a strategy to associate changes in activating histone modifications with fold changes in gene expression in the same cellular conditions that lie in the vicinity of DPs (25). A similar approach, based on a list of differentially expressed genes, has been proposed by Heinig et al. (28). However, these strategies do not explore the presence of replicates in the ChIP-seq or expression data.

Another strategy for the evaluation of DPs is the simulation of ChIP-seq experiments that currently is mostly based on the SPC problem (12,27,35). There is a clear need for methods to evaluate DPC methods by using both real and simulated data. In particular, it is crucial to have methodologies exploring the performance of DPC methods on data with distinct characteristics: samples with low variability and high signal-to-noise ratio (data from *in vitro* based studies with technical replicates) versus experiments with high variability and low signal to noise ratios (data from clinical studies with biological replicates and heterogeneous cell populations).

Our approach

We propose THOR a differential peak caller for comparison of two biological conditions with replicates. THOR expands our previous work on differential peak calling [ODIN (25)] by supporting replicates and providing two further approaches for normalization of ChIP-Seq libraries. For supporting replicates, THOR uses a Negative Binomial distribution that deals with the typical overdispersion in read count distributions with replicates (29). Concerning library normalization, THOR implements the commonly used TMM approach (34) and an novel approach based on housekeeping genes to normalize ChIP-Seq libraries of activating histone marks. These modifications are crucial for analysis of ChIP-seq data with high variance and distinct signal to noise ratios as in studies with biological replicates. This work also expands the evaluation procedure from (25) by using for the first time histone modifications with similar regulatory roles (activating marks) in the validation of differential peaks. Moreover, it extensively expands the methodology for simulation of ChIP-Seq reads from (25) that allow to simulate ChIP-Seq data with a distinct number of replicates and distinct variance within replicates.

THOR improves previous DPC methods supporting replicates (24,26,27), as it intrinsically detects peaks with variable size through the use of posterior decoding algorithms. This is of particular importance for the analysis of differential peaks in histone modifications that have subtle changes in signals within larger peaks. THOR provides all pre- and post-processing steps required in ChIP-seq analysis including fragment size estimation, input-DNA normalization, GC-bias correction, *P*-value estimation, multiple test correction and strand lag filtering of potential DPs. See Figure 1A for a schematic procedure of THOR.

We evaluate THOR on four biological studies measuring activating histone marks with distinct replicate characteristics. We generate H3K27ac ChIP-seq data with two technical replicates on an *in vitro* dendritic cell differentiation system (36). We also evaluate activating histone modifications during the response of mouse to cocaine intake with three biological replicates (6), monocyte to macrophage differentiation from up to eight individual replicates (37) and a case control study comparing B cells from patients with follicular lymphoma and healthy donors (7). While THOR is tailored for the analysis of histone modification data, it should also perform as well as other DPCs in the analysis of ChIP-seq with small peaks as with transcription factors. We therefore also include a case study with the analysis of Pol2 ChIP-seq data from (6). These data sets provide us altogether 13 pairs of conditions (differential peak problems) that are analysed for differential peaks (see Table 1).

We perform a comparative analysis with all one-stage (PePr (24), DiffReps (24) and csaw (27)) and two-stage peak callers (DiffBind (23), MACS2 (unpublished) and the combination of DESeq with SPC peaks obtained from JAMM (18) and IDR (16)) that support replicates (Figure 1B). We also evaluate a version of THOR using the Poisson distribution, which is equivalent to ODIN (25) with support of replicates. The performance of methods are evaluated with expression based Differential Correlation Analysis (DCA), which is based on the correlation between the *P*-values of the top ranked DPs and the *P*-values of the expression values of the genes close to these DPs. This method is based on the assumption that activating histone marks correlates with gene expression (2,38,39). This assumption has been previously explored in the validation of differential peaks (25,28). We assume here that the best predictions (DPs called by the methods) will have a higher correlation with expression than poor predictions. Moreover, we propose for the first time the use of histone modifications marks instead of the gene expression in the DCA. This explores the fact that activating (or repressing) histone modifications correlate (40). Therefore, we can use a particular activating histone (validating mark) to evaluate peaks detected in another activating histone mark. Finally, we make use of a simulator for ChIP-Seq reads to evaluate the performance of methods on controlled scenarios. This allows us to evaluate the competing methods, when data have distinct characteristics such as the within condition variability and number of replicates (see Figure 1C).

MATERIALS AND METHODS

Basic notations and profile construction

We create a genomic profile to analyse ChIP-seq data by fragmenting the genome into bins and counting the reads falling in these bins. We divide the genome into a set $\{b_1, \dots, b_L\}$ by using a sliding window approach. Each bin b_j covers the genomic positions $[j \cdot s - 0.5 \cdot w, j \cdot s + 0.5 \cdot w]$, where s and w are the step size and the window size. The value of the genomic profile x_{ij} is the number of extended reads of ChIP-seq signal i aligned to regions overlapping bin b_j. Previous to this step, reads are extended to have a size f, which corresponds to the DNA fragment size of the ChIP-seq experiment. This parameter can be provided by the user or estimated from the data (see below).

Then, **X** is the matrix that represents the genomic signal

$$\mathbf{X} = \{x_{ij}\}^{D \times L}.$$

The ith genomic signal (experiment) is represented by the vector $x_i = (x_{i1}, \dots, x_{iL})$ and the genomic signals for bin j is represented by the vector $x_{\cdot j} = (x_{1j}, \dots, x_{Dj})$. Each experiment belongs to one of K biological conditions. More formally, we define the set of experiments associated to condition k as

$$G_k = \{i \mid i \in \{1, \dots, D\}, i \text{ belongs to } k\}$$

and experiment as

$$G = \{G_1, \dots, G_K\}.$$

Here, $x_{G_k j}$ represents the genomic signal restricted to experiments belonging to G_k and $\overline{x}_{G_k j}$ is the mean read count for all experiments in condition k, i.e.

$$\overline{x}_{G_k j} = \frac{\sum_{i \in G_k} x_{ij}}{|G_k|}.$$

Moreover, ChIP-seq experiments usually have reads from input-DNA for each cell type analysed. We will refer to input-DNA as $\mathbf{X}^{input}$.

Signal preprocessing. Previous to the application of the HMM, we perform several preprocessing steps to the raw ChIP-seq counts. These are in order of application: (i) estimation of fragment size, (ii) GC-content correction, (iii) Input-DNA normalization and subtraction and (iv) global ChIP-seq signal normalization.

Fragment extension estimation. As only the beginning of DNA fragments obtained by ChIP are sequenced, we have to reconstruct the fragments by estimating its extension size (31). Following Mammana *et al.* (41), we define a strand cross-correlation function

$$c(f) = \sum_{p \in F \cup R} h(p) \cdot h(f + p)$$

where F (R) gives all left-most (right-most) positions of reads aligned to the forward (reverse) strand and where h describes the overlap of these positions. Function c describes the correlation between read counts on the forward and reverse strand for a given fragmentation size f. The desired estimate $\hat{f}$ of the fragmentation size maximizes function c.

GC-content. To model and correct against bias introduced by GC-content, we use a histogram based approach (11). Considering signal **X**, we measure the average number of reads assigned to bins on a particular GC-content interval, compute the average for the GC-content interval and adjust the ChIP-seq profile if it differs from the expected GC-content.

Input-DNA subtraction. Control input-DNA corrects for bias associated to DNA shearing (31). The standard procedure is to subtract input DNA signal from the ChIP-seq

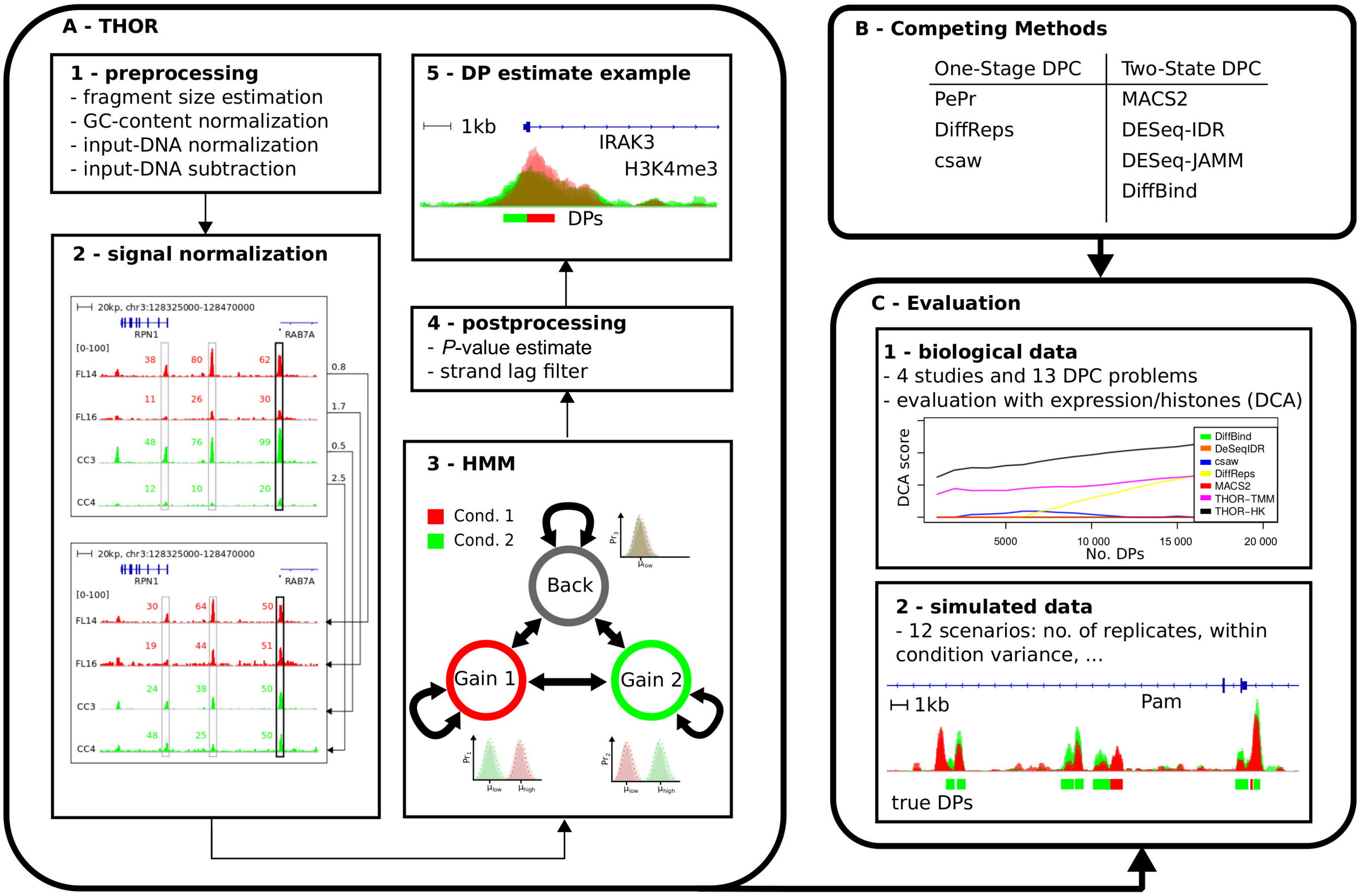

Figure 1. THOR's analysis workflow. (**A**) After pre-processing the ChIP-seq signal, normalization based on housekeeping genes (as shown) or TMM is performed. The normalized signal serves as input for the HMM which is used to estimate DPs. Post-processing includes the statistical assessment of DPs. (**B**) List of all competing methods categorized in one-stage and two-stage approaches. (**C**) Evaluation strategies. We evaluate DPC methods with biological data and the DCA statistic that is based on the association between DPs, gene expression and histone modifications. Moreover, simulated data are used to investigate the effect of distinct ChIP-seq signal characteristics on DPCs methods.

Table 1. Overview of DP experiments. We describe the experiment name, histone modification type, cellular conditions, number of replicates and data used for evaluation by the DCA strategy for each of the evaluated differential peak problems

Experiment	Histone	Cond. 1	Cond. 2	#rep	DCA Validation
DC-MPP-CDP	H3K27ac	MPP	CDP	2.2	RNA-seq, H3K4me1
DC-CDP-cDC	H3K27ac	CDP	cDC	2.2	RNA-seq, H3K4me1
DC-CDP-pDC	H3K27ac	CDP	pDC	2.2	RNA-seq, H3K4me1
DC-cDC-pDC	H3K27ac	cDC	pDC	2.2	RNA-seq, H3K4me1
CO-H3K36me3	H3K36me3	saline	cocaine	3.3	RNA-seq, H3K4me1, Pol2
CO-H3K4me1	H3K4me1	saline	cocaine	3.3	RNA-seq, H3K36me3, Pol2
CO-Pol2	Pol2	saline	cocaine	3.3	RNA-seq, H3K4me1, H3K36me3
MM-H3K27ac	H3K27ac	MONO	MAC	5.8	RNA-seq, H3K4me3, H3K4me1
MM-H3K4me1	H3K4me1	MONO	MAC	5.8	RNA-seq, H3K4me3, H3K27ac
MM-H3K4me3	H3K4me3	MONO	MAC	6.10	RNA-seq, H3K4me1, H3K27ac
LYMP-FL-CC	H3K27ac	FL	CC	8.5	Microarray, H3ac
LYMP-FL-PBBA	H3K27ac	FL	PBBA	8.3	Microarray, H3ac
LYMP-CC-PBBA	H3K27ac	CC	PBBA	5.3	Microarray, H3ac

signal (31). Due to different read concentration in the ChIP-seq profile and the input-DNA, one should first normalize the input-DNA in relation to the ChIP-seq signal (42). Here, we estimate a scaling factor α^{SES} based on the order statistics of binned count data (42). For a given input library x^{input}, we obtain an input subtracted signal for each ChIP-seq library as

$$x_{\cdot j}^{\mathrm{SES}} = x_{\cdot j} - \alpha \cdot x^{input}.$$

This step is only performed if matching input-DNA is available.

ChIP-seq signal normalization. A crucial aspect in the analysis of multiple ChIP-seq samples is the strategy for global normalization of samples to bring them to a similar scale. Currently, most DPC methods use a weighted TMM (34). Our previous work ODIN (25), only supports normalization by library size.

We explored an alternative approach based on the use of control regions (9). Control regions can be obtained by ChIP-PCR on selected genomic regions. For the case of activating histone modification, we use the promoter of housekeeping genes (HK). Given that housekeeping genes do not change their expression, the abundance of activating histone marks in their promoter should also be constant (2,38,39).

More formally, we define a set of control genomic regions, $R = \{r_1, \ldots, r_{\mathrm{N}}\}$. The ChIP-seq signal of region r_{n} for sample i is

$$h_{\mathrm{in}} = \sum_{j} x_{\mathrm{ij}} \cdot \mathbf{1}(b_{\mathrm{j}} \text{ overlaps } r_{\mathrm{n}}).$$

First, for a given region n, we normalize the mean of each samples to the particular signal i

$$h'_{\mathrm{in}} = \frac{\overline{h}_{\cdot\mathrm{n}}}{h_{\mathrm{in}}}.$$

The normalization factor for sample i is

$$f_{\mathrm{i}} = \frac{\sum_{\mathrm{n}} h'_{\mathrm{in}}}{N}$$

where N is the number of HK genes. Final, ChIP-seq count estimates for sample i are given by, $x'_{\mathrm{i}\cdot} = f_{\mathrm{i}} \cdot x_{\mathrm{i}\cdot}$. We use regions 500 bps upstream of all housekeeping genes described by Eisenberg and Levanon (43) (C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29) as control regions for the human genome. These genes have been shown to have a stable expression pattern over several microarrays and RNA-seq expression experiments (43). For the mouse genome, we left out the human specific C1orf43 gene. Note that any list of genomic regions, where no changes in ChIP-seq signals are expected, can be used for this normalization approach. See Supplementary Figure S1 for an example of this normalization approach.

THOR also supports a TMM based normalization. In short, for a given signal $x_{\mathrm{i}\cdot}$ with $i \in G_{\mathrm{k}}$, we first estimate the mean signal $\overline{x}_{G_{\mathrm{k}}\cdot}$ of condition k. Then, the logarithmic ratio (M-values)

$$M_{\mathrm{j}} = \log\left(\frac{\overline{x}_{G_{\mathrm{k}}\mathrm{j}}}{x_{\mathrm{ij}}}\right)$$

and the logarithmic average (A-values)

$$A_{\mathrm{j}} = 0.5 \cdot \log(\overline{x}_{G_{\mathrm{k}}\mathrm{j}} x_{\mathrm{ij}})$$

are estimated for all bins j. The normalization factor f_{i} is the ratio of M- and A-values weighted by A-values

$$\log(f_{\mathrm{i}}) = \frac{\sum_{\mathrm{j}} A_{\mathrm{j}} \cdot M_{\mathrm{j}}}{\sum_{\mathrm{j}} A_{\mathrm{j}}}.$$

HMM for differential peak calling

The DPC problem is modelled by a three state HMM, which receives a D dimensional variable X as input. This first order HMM contains a state representing DPs gained in the first biological condition G_1 (`Gain 1`), a state for DPs gained in the second biological condition G_2 (`Gain 2`) and a background state (`Back`). Supplementary Figure S2 shows the HMM topology, where all states have transitions to other states and to themselves.

For a given state s, the emission distribution of the HMM is given by the product of probabilities for each biological condition G.

$$\Pr{}_{\mathrm{s}}(x_{\cdot\mathrm{j}}) = \prod\nolimits_{\mathrm{k} \leq |G|} \Pr{}_{\mathrm{sk}}(x_{G_{\mathrm{k}}\mathrm{j}})$$

The probability of observing $x_{G_{\mathrm{k}}\mathrm{j}}$ in state s and condition k is given by the product of the observation's probabilities associated to condition k

$$\Pr{}_{\mathrm{sk}}(x_{G_{\mathrm{k}}\mathrm{j}}) = \prod\nolimits_{\mathrm{i} \in G_{\mathrm{k}}} \Pr{}_{\mathrm{sk}}(x_{\mathrm{ij}})$$

The distribution for a given state s, condition k and sample i is modelled by a Negative Binomial distribution as follows

$$\Pr{}_{sk}(x_{\mathrm{ij}}) = g\left(x_{\mathrm{ij}} \mid a_{\mathrm{sk}}, \mu_{\mathrm{sk}}\right) =$$

$$\frac{\Gamma(x_{\mathrm{ij}} + a_{\mathrm{sk}}^{-1})}{\Gamma(x_{\mathrm{ij}} + 1) \cdot \Gamma(a_{\mathrm{sk}}^{-1})} \cdot \left(\frac{a_{\mathrm{sk}}^{-1}}{a_{\mathrm{sk}}^{-1} + \mu_{\mathrm{sk}}}\right)^{a_{\mathrm{sk}}^{-1}} \cdot \left(\frac{\mu_{\mathrm{sk}}}{a_{\mathrm{sk}}^{-1} + \mu_{\mathrm{sk}}}\right)^{a_{\mathrm{sk}}^{-1}}$$

where a_{sk} is the dispersion parameter, μ_{sk} the location parameter and Γ the gamma function. The parameters a_{sk} and μ_{sk} are fixed for samples of a same biological condition. The function g has mean $\mathbb{E}(x_{\mathrm{i}}) = \mu_{\mathrm{sk}}$ and variance

$$Var(x_{\mathrm{i}}) = \mu_{\mathrm{sk}}(1 + a_{\mathrm{sk}}\mu_{\mathrm{sk}}). \tag{1}$$

If $a_{\mathrm{sk}} = 0$, the mean equals the variance and the distribution results in a Poisson distribution. For $a_{\mathrm{sk}} > 0$, variance increases with mean as usual when dealing with NGS data containing replicates (29). This HMM model is the same as used by ODIN (25). ODIN is based on a Binomial distribution and only supports two signals ($D = 2$). The Binomial distribution approximates a Poisson distribution for large n, where n is the number of reads in the ChIP-seq libraries. Therefore, by fixing $a_{\mathrm{sk}} = 0$, we obtain a ODIN version supporting multivariate signals that cannot deal with over-dispersion. We refer to this method as Poisson-THOR.

HMM training. The HMM is estimated with the Baum–Welch algorithm (44). Initial state and transition probabilities are based on standard estimates (44). Concerning the emission distribution parameters μ_{sk} and a_{sk}, we estimate them based on a moment approach. We constrain location parameters of `Gain 1` (s = 1) and state `Gain 2` (s = 2) associated to enriched signals to be equal $\mu_{1G_1} = \mu_{2G_2} = \mu_{\mathrm{high}}$. We also constrain location parameters of low values and background states to be equal $\mu_{1G_2} = \mu_{2G_1} = \mu_{3G_1} = \mu_{3G_2} = \mu_{\mathrm{low}}$ (see Supplementary Figure S2). This avoids label switching problems in the HMM (44). It can be easily shown that the these constraints lead to the following estimates (Supplementary Section 1):

$$\hat{\mu}_{\mathrm{high}} = \frac{\sum_{i\in G_2}\sum_{\mathrm{j}=0}^{L} r_{2\mathrm{j}}x_{\mathrm{ij}} + \sum_{\mathrm{i}\in G_1}\sum_{\mathrm{j}=0}^{L} r_{1\mathrm{j}}x_{\mathrm{ij}}}{|G_2|\sum_{\mathrm{j}=0}^{L} r_{2\mathrm{j}} + |G_1|\sum_{\mathrm{j}=0}^{L} r_{1\mathrm{j}}} \text{ and}$$

$$\hat{\mu}_{\mathrm{low}} = \frac{\sum_{\mathrm{i}\in G_2}\sum_{\mathrm{j}=0}^{L} r_{1\mathrm{j}}x_{\mathrm{ij}} + \sum_{\mathrm{i}\in G_1}\sum_{\mathrm{j}=0}^{L} r_{2\mathrm{j}}x_{\mathrm{ij}} + \sum_{\mathrm{i}\in G_2}\sum_{\mathrm{j}=0}^{L} r_{3\mathrm{j}}x_{\mathrm{ij}} + \sum_{\mathrm{i}\in G_1}\sum_{\mathrm{j}=0}^{L} r_{3\mathrm{j}}x_{\mathrm{ij}}}{|G_2|\sum_{\mathrm{j}=0}^{L} r_{1\mathrm{j}} + |G_1|\sum_{\mathrm{j}=0}^{L} r_{2\mathrm{j}} + |G_2|\sum_{\mathrm{j}=0}^{L} r_{3\mathrm{j}} + |G_1|\sum_{\mathrm{j}=0}^{L} r_{3\mathrm{j}}}$$

To obtain reliable variance estimates on small sample sizes, we assume that the variance can be described by a smooth function based on the mean estimates similar as described by Anders and Huber (29). We use for this a quadratic function

$$v_{\mathrm{k}}(x) = c_{1\mathrm{k}} \cdot x^2 + x + c_{2\mathrm{k}}, \qquad (2)$$

which is estimated for the ChIP-seq data on samples of condition k previous to the Baum–Welch algorithm. The dispersion parameter a_{sk} is derived from Equation 1 and given by

$$a_{\mathrm{sk}} = \frac{v_{\mathrm{k}}(\mu_{\mathrm{sk}}) - \mu_{\mathrm{sk}}}{\mu_{\mathrm{sk}}^2}.$$

We apply the Viterbi algorithm to estimate a state sequence for the complete genomic signal. Finally, we merge consecutive bins associated to states `Gain 1` or `Gain 2` to obtain the candidate DPs.

Initial HMM parameter. We initialize the HMM by using an initial set of candidate DPs based on two simple criteria. For state `Gain 1`, we select bins if there is a difference in counts between two signals (t_1)

$$\overline{x}_{G_1\mathrm{j}} - \overline{x}_{G_2\mathrm{j}} > t_1,$$

or if there is a high fold change t_2 and minimum signal support t_3

$$\overline{x}_{G_1\mathrm{j}}/\overline{x}_{G_2\mathrm{j}} > t_2 \quad \text{and} \quad \overline{x}_{G_1\mathrm{j}} + \overline{x}_{G_2\mathrm{j}} > t_3.$$

DPs associated to state `Gain 2` are defined accordingly. We use these DPs to obtain posterior probabilities and then perform a single M-Step of the Baum–Welch algorithm to obtain HMM's parameters.

Postprocessing. We developed a pipeline for post-processing candidate DPs. First, we ignore all DPs with a size smaller than the mean of estimated fragment sizes $\hat{f}$. Moreover, false positive DPs may be caused by a high strand lag (45). For each DP, we therefore count the forward and reverse reads, normalize the ratio by computing the z-scores and filter out all DPs that exhibit a high/low z-score. By default, we choose a z-scores threshold of 2 that corresponds to a 2-fold standard deviation from the normalized ratio distribution. Also, THOR allows filtering of DPs falling into blacklisted genomic regions. These are regions with unstructured, high signals in next generation experiments independently of cell lines and the type of experiment (2). Finally, we perform an exact statistical test to assign a *P*-value to each DP (25). We use a Negative Binomial distribution whose parameters are based on estimates from the `Back` states. Then, we merge significant DPs that have a distance less than the mean of all estimated fragment sizes $\hat{f}$. *P*-values are re-estimated after merging and corrected for controlling the False Discovery Rate (46).

Materials

Biological data sets dendritic cell (DC) differentiation. Dendritic cells (DC) are professional antigen presenting cells that develop from haematopoietic stem cells in bone marrow. We have developed a two-step culture system that recapitulates DC development *in vitro* (36). This system allows to differentiate *ex vivo* multipotent progenitors (MPP) to DC progenitors (CDP). CDP cells are further differentiated to either classical DC (cDC) or plasmacytoid DC (pDC). We performed ChIP-seq with two technical replicates for the histone modification H3K27ac. ChIP was performed with minor modifications (1,47).

Briefly, sorted cells were cross-linked at a concentration of 2 million cells/ml with 1% formaldehyde for 6 min at room temperature. Cross-linking was stopped with 0.125 M glycine. Chromatin sonication into fragments of 200-400 bp in size was done in Bioruptor with cooling device (Diagenode) at 4°C with 30 s pulse/pause cycles until a fragment size of 200 bp was obtained. Sheared lysates were clarified by centrifugation at 12 000g (10 min, 4°C). Dynabeads Protein A (Life Technologies) (10 μl) were preincubated with 1 μg anti-H3K27ac antibodies (Abcam). For immunoprecipitation, 10 μg sheared chromatin was added to the preincubated beads overnight at 4°C. Chromatin complexes were isolated by magnetic bead selection and washed with RIPA and TE buffer. Chromatin complexes were digested with 50 μg/ml RNase (Roche) at 37°C for 30 min. The ChIP procedure was repeated five times (50 μg chromatin in total) and immunoprecipitated DNA was purified using QIAquick PCR Purification Kit (Qiagen). DNA concentration of immunoprecipitated DNA was determined by using Qubit dsDNA HS Assay kit (Life Technologies). Libraries were prepared and subjected to deep-sequencing on the Illumina platform according to the manufacturer's protocols.

We performed total RNA sequencing for evaluation of DPs with DCA. RNA was isolated using RNeasy Plus Mini Kit with DNaseI digestion (Qiagen) (36). Libraries were prepared and subjected to strand-specific RNA-seq on the Illumina platform according to the manufacturer's protocols. DCA evaluation was complemented with H3K4me1 ChIP-seq without replicates previously generated by us (25,33). On this data set, we compare DPs following the lineage commitment steps (MPP to CDP, CDP to cDC, CDP to pDC) and DC subset specification (cDC and pDC) that lead to four DPs experiments. RNA-seq and ChIP-seq of dendritic cell differentiation samples are deposited in Gene Expression Omnibus with accession number GSE73143. This study represents a scenario with potentially very low variability within the biological conditions.

Epigenomics effects of cocaine (CO). Feng *et al.* (6) analyse epigenetics changes after cocaine intake on mouse nucleus accumbens. The study measures histone modifications of three biological replicates after treatment with a cocaine or saline solution (control). We use data from histone modifications H3K4me1 and H3K36me3 as well as Pol2, which lead to three DPC experiments. We also analyse RNA-seq data matching the samples for DCA evaluation (GEO accession number GSE42811 and GSE24850). No input DNA was provided. This study represents a scenario with biological replicates with similar genomic background. Therefore, we expect a low variance within biological conditions.

Monocyte and macrophages (MM). This study provides samples of monocytes (MONO) activated to macrophages (MAC) in up to 8 human samples (37). Here, we consider the histone modifications H3K4me1, H3K27ac and H3Kme3. For H3K4me1, there are 6 MONO and 10 MAC, for H3K27ac there are 5 MONO and 8 MAC samples and for H3K4me3 6 MONO and 10 MAC samples. We perform DP estimations between MONO and MAC for both histone modifications. Condition specific RNA-seq data (36 MAC and 25 MONO samples) are used for DCA evaluation. The study does not provide input-DNA data for the ChIP-seq experiments. The data are available with restricted access at EGA (EGAD00001001011). This study represent a scenario with human biological replicates with a moderate within group variability.

B cell lymphoma (LYMP). Koues et al.(7) performed the most comprehensive analysis of regulatory genomic features in lymphomas up to date. We use ChIP-seq data of the histone modification H3K27ac on follicular lymphoma cells (FLs), as well as distinct populations of B cells from healthy donors: proliferative centroblasts (CC) and peripheral blood B cells (PBBA). We only consider samples with a matching input-DNA, gene expression (measured with microarrays) and H3ac ChIP-seq: CC samples 1-5, FL samples 1,2,5,8,10,11,14,16 and PBBA sample 1-3 (GSE62246). We detect DPs by comparing: FL versus CC, FL versus PBBA and CC versus PBBA. Evaluation was performed with microarray gene expression and H3ac histone modifications. This data set contains human biological replicates and disease samples and is expected to have a high within group variability.

We use BWA (48) version 0.6.1 − r104 with default parameters for read mapping on either mouse genome (mm9) or human genome (hg19). Table 1 gives an overview of the DPC experiments.

Evaluation of biological data sets. Evaluating DPs is still an open problem as there is no gold standard for DPC. We have previously developed a strategy to associate changes in protein–DNA with fold changes in gene expression in regions (or genes) in the vicinity of the DPs (25). For this, we require gene expression (RNA-seq or microarrays) for the same cellular conditions. This idea is based on the fact that the level of histone modifications correlate with the expression of the surrounding genes (2,38,39). We expand this idea by using histone modifications, which were not used in the DPC problem itself, to evaluate the DPs. This is based on the fact that histones with similar code, i.e. activating histone marks, are known to correlate (40).

First, we associate each DP of an ‘evaluated histone’ with either gene expression or ‘validating histones’ for each condition. For RNA-seq or ChIP-seq data, we count the reads of these sequencing libraries into the DPs. We then perform a differential analysis with DESeq (29) on the RNA-seq and validating ChIP-seq values. This step, which provides us a *P*-value for each list of DPs, differs from our previous approach (25) that only considered the fold change ratio between gene expression data. For microarray data (only applies to the Lymphoma Study), DPs are assigned to genes if (i) they are located in the gene or close to the promoter of a gene, (1000 bp upstream) or (ii) if the peaks are located 50 Kbps away from the TSS without a TSS of another gene following (49). The average expression value of genes assigned to a peak is used. Peaks not assigned to genes are ignored. Finally, limma (50) is used to obtain *P*-values.

Next, we compare the DP *P*-values $p_{\mathrm{i}1}$ for each DPC algorithm A with the *P*-value $p_{\mathrm{i}2}$ based on the gene expression or ‘validating histone’ data. For this, we compute the Spearmann correlation between the list of both *P*-values $p_{\mathrm{i}1}$ and $p_{\mathrm{i}2}$ for the top k ranked DPs

$$e(k) = \mathrm{cor}(\langle p_{\mathrm{i}1}\rangle, \langle p_{\mathrm{i}2}\rangle)_{\mathrm{Spearmann}}$$

for all i < k. For increasing k, we obtain DCA curves that indicates the association of gene expression or ‘validating histones’ and DPs for distinct number of peaks called. Furthermore, we obtain a single score per method by estimating the area under the DCA curve

$$\mathrm{DCA} = \frac{\sum_{\mathrm{k=h}}^{H} \max\left(0, e(k)\right) \cdot h}{H}$$

for k ∈ $[h, 2 \cdot h, \ldots, H]$, where h is the step size and H the maximum number of DPs used.

Given that we only evaluate activating histone modifications, higher DCA scores indicates higher agreement of DPs with changes in expression and other activating histone modifications. We assume here that good DP predictions should have a higher DCA scores than poor DP predictions. We can therefore use DCA scores to rank DP predicted by evaluated methods. Note also that while this assumption does not hold for all transcription factors (51), the general transcription factor Pol2 is also known to correlate with gene expression and can therefore be evaluated with the DCA approach as well. See Table 1 for a listing of histone modification marks used in the DCA evaluation.

We use the Friedman–Nemenyi test (52) to rank the performance of evaluated DPCs methods by DCA values. This test allows the evaluation of several methods applied to multiple data sets. The Friedman–Nemenyi test indicates whether one of the methods has higher DCA values than others.

Simulated data sets. There is so far no publicly available framework for simulating ChIP-seq data in the scenario of DPC with replicates. We extensively expand a simple simulator previously proposed by us (25) for the simulation of DPs to account for replicates. In short, the algorithm works in five steps. (i) Initially, we define protein domains (regions with proteins) in the genome; and (ii) assign fragments to each protein in a domain. The number of reads follows a Negative Binomial distribution $NB(m, p)$. (iii) Next, we assign reads to biological conditions using a $B(\beta)$ distribution; and then assign reads to replicates within conditions using a Dirichlet distribution $Dir(\alpha)$. (iv) We include background noise to the data by assigning reads to randomly selected genomic regions. (v) Finally, true positive DPs are defined by a fold change criteria. Note that ODIN's simulator did not contain steps (iii) and (iv). Moreover, steps (i) and (ii) have been extensively refined. See Supplementary Section 2 for details about the simulation algorithm and Supplementary Figure S3 for an schematic of the algorithm.

We evaluate the simulated predictions by evaluating the ratio of true and false predicted DPs. Let T be the genomic region set of true positive DPs given by the simulation. Moreover, let $P_A = \{p_1, \ldots, p_m\}$ be the genomic region set of DPs (sorted by increasing P-value) that are predicted by algorithm A. Let

$$\hat{Y}_i = \frac{|p_i \cap T|}{|T|}$$

describe the ratio of the true called DPs and

$$\hat{X}_i = \frac{|p_i - T|}{|\text{genome} - T - P_A|}$$

describe the ratio of false called DPs. Note that operations $\cap$ and $-$ are based on interval algebra and give therefore a basepair overlap statistics for DPs predictions. See Supplementary Section 2 for details about the evaluation. We compute the cumulative sum $D(i) = \sum_{j \le i} \langle \hat{X}_j, \hat{Y}_j \rangle$ by element-wise addition of the list of tuples. We can then generate Area Under the Curve (AUC) statistics by the integral of the function D. The higher the value, the better is the DP prediction of algorithm A.

We are particularly interested how methods perform when the number of reads of each protein in a domain, the number of replicates and the variance within replicates changes. We therefore simulate the following parameter settings: $(m, p) \in \{(100, 200), (100, 400)\}$ to obtain peaks with distinct sizes and variance; $(r_1, r_2) \in \{(2, 2), (4, 4)\}$ to evaluate experiments with 2 and 4 ($r_1 = r_2$) replicates; and $\alpha \in \{5, 10, 15\}$ to obtain data with low (5) to high (15) variance within a biological condition. For each parameter setting, we repeat the simulation 25 times. We simulate proteins with histone like characteristics (147 bps contact with the DNA and an average protein spacing of 202 bps) in the DNA. DNA Fragments of 200 bps are sampled around the protein contact area. See Supplementary Figures S4 and S5 for examples of simulated data.

Reads are sampled from the chromosome 1 of mm9 and aligned with BWA with default parameters.

THOR parametrization and implementation. For the initialization of THOR, we use $t_2 = 1.6$ as fold change criteria, $t_1 = \langle x \rangle^{.95}$ as minimum difference between signals, where $\langle x \rangle^{.95}$ is the value in the 95% percentile of $\mathbf{X}$, and $t_3 = t_1/2$. If these parameters yield a training set smaller than $t^{\min} = 100$, we decrease t_2 by 15 and t_1 by 0.1, and repeat the training set construction procedure. To estimate the mean/variance function for each biological condition k, we randomly choose 20 000 bins, estimate mean and variance for each bin and fit the quadratic model described in Equation 2 using an non-linear least squares approach (53).

We evaluated different parameter settings for $t_2 \in \{1.3, 1.6\}$ and $t_1 \in \{\langle x \rangle^{.95}, \langle x \rangle^{.99}\}$ by predicting DPs for chromosome 1 for all 12 experiments. The Friedman–Nemenyi test on DCA statistics for $h = 100$ and $H = 1000$ shows no statistical significant difference, which indicates that THOR is robust against distinct initial parameter definitions (see Supplementary Table S1 and S2). We used the parametrization with highest ranking for further experiments. Chromosome 1 was left out of the comparative method analysis. See Supplementary Section 3 for details about the parametrization of competing methods and Supplementary Table S3 for an description of the methods' characteristics.

THOR is available at the Regulatory Genomics Toolbox www.regulatory-genomics.org/thor. Scripts for evaluation of differential peaks, ChIP-seq simulation as well as peak predictions are available at www.costalab.org/thor. For the data set with the largest number of ChIP-seq samples (MM-H3K4me3), THOR required 4 h and 16 GBs of memory on a 2.4GHz machine.

RESULTS AND DISCUSSION

Comparative analysis on simulated data

First, we evaluate 6 distinct DPCs [THOR, DiffReps, MACS2, DiffBind, DESeq-IDR and DESeq-JAMM (Csaw was not included in this analysis, as it failed to execute with simulated data)] with simulated data. Our simulator generates ChIP-seq reads from 'virtual proteins' positioned in the genome. First, proteins with histone-like characteristics (147 bps contact with DNA, occurring in domains with an average of 8 proteins and with an average spacing between these proteins of 202 bps) are placed in the DNA. DNA Fragments of 200 bps are sampled around the protein contact area. Next, fragments are distributed to one of the two biological conditions and then to the replicates of each condition. The first positions of the fragments define the ChIP-seq reads. The original positions of the proteins and the proportion of fragments assigned to each of the two conditions are used to define the true DPs. See Supplementary Section 2 for more details and Supplementary Figures S3, S4 and S5 for a schematic of the method and examples of the simulated data.

We investigate three characteristics that are important for the differential peak calling problem. The first characteristic is the number of replicates (2 or 4) per biological conditions. Experiments with 2 replicates are obtained by discarding 2 ChIP-seq experiments from each biological condition of the experiments with 4 replicates. Second, we evaluate distinct variance levels (moderate and high) of peak sizes for a given biological condition to model distinct types of histone modifications that have either uniform or varying peak sizes. Finally, we evaluate the level of variance within replicates in a given biological condition (high, medium and low variance). This parameter will control the consistency between replicates: higher variance will impose lower consistency and more difficult scenarios. For each of the 12 experimental combinations, we simulated 25 data sets.

Methods were evaluated with Area Under the Receiver Operating Characteristic Curves (AUC) statistics by estimating the overlap fraction of predicted and true DPs (see Section ***Simulated Data Sets***). Figure 2 shows the AUC values for all methods and experimental combinations. Concerning the number of replicates (red versus green lines), we observe that most methods have lower AUC values in experiments with 2 replicates (red line) then with 4 replicates (green line) (P-value < 0.05; one-sided Wilcoxon test). Exceptions are Poisson–THOR and DESeq-IDR. IDR returns very few peaks on cases with 4 replicates (green line), even when using an lenient threshold for the SPC method used as input for IDR. Poisson–THOR's poor performance on 4 replicates possibly stems from its simple distribution that does not cope with overdispersion.

The second characteristic is the variance of the peak sizes, where we evaluate scenarios with moderate and high variance (Figure 2). Two methods have higher AUC values in scenarios with moderate peak variance: THOR in case of low and medium within variance and 2 replicates (P-value $< 10^{-4}$, one-sided Wilcoxon test) and DiffBind in the case of low, medium and high within variance and 2 replicates (P-value $< 4.7 \times 10^{-8}$, one-sided Wilcoxon test). All other tools show no significant changes in AUC values. The last characteristic is the level of variance within the replicates. DESeq-JAMM, DiffReps and DESeq-IDR show decrease in AUC values for increasing variance. Interestingly, the performance of THOR, MACS2 and DiffBind shows increase in AUC values with increasing variance for respectively 5, 3 and 6 of the 8 cases on the comparison of low versus medium and medium versus high within condition variance (P-value < 0.05; one-sided Wilcoxon test).

Finally, we apply the Friedman–Neymeni test to evaluate the statistical significance in AUC value differences for distinct methods. Considering all data together, THOR has significantly higher AUC scores than all competing methods. MACS2 has significant higher AUC values than DiffReps, DESeq-IDR, DiffBind and Poisson–THOR; and DESeq-JAMM and DiffReps have significantly higher AUC values than DiffBind and Poisson–THOR (P-value < 0.05; Supplementary Table S4 and S5). Evaluating specific conditions, THOR has significantly higher AUC values than DiffReps, DiffBind and Poisson–THOR for all 12 cases (P-value < 0.05, Supplementary Table S6–S17). In the case with 2 replicates, THOR additionally has significantly higher AUC values than DESeq-JAMM (P-value < 0.05, Supplementary Table S6–S11) and in the case with 4 replicates significantly higher AUC values than DESeq-IDR (P-value < 0.05, Supplementary Table S12–S17). THOR is ranked top in all of the 12 cases.

Quality analysis on biological data sets

To better understand the characteristics of ChIP-seq experiments evaluated in our study, we first perform a quality check. For this we use the fractions of reads in peaks (FRIP) score from the ENCODE consortia (16) that gives an estimate of the signal-to-noise ratio of ChIP-seq experiments. We also propose the novel use of the quadratic coefficient (c_{1k} of the function modelling the mean versus variance relationship, see Equation 2) for a given biological condition as an indicator for 'overdispersion'. Example of overdispersion estimates and mean versus variance distributions of selected experiments are displayed in Figure 3A and B and complete statistics are provided in Supplementary File 1. As expected, overdispersion positively correlates with the number of replicates in the condition (R = 0.74, adjusted P-value = 0.0001; Spearman Correlation). Moreover, higher overdispersion is associated to lower FRIP scores (R = −0.78; adjusted P-value $= 2.9 \times 10^{-5}$).

As depicted in Figure 3C, average FRIP versus overdispersion space separates the experiments by their expected complexity. The DC differentiation experiments, which were obtained by *in vitro* differentiation of cells with technical replicates, have highest FRIP values and lowest overdispersion values. The follicular LYMP, which arise from patients/donors with distinct genetic background and with potential tissue heterogeneity, have both highest overdispersion scores and lowest average FRIP. This indicates that the experiments evaluated here covers a large spectra of peak size variance within biological conditions. Moreover, it demonstrates the potential value of the overdispersion metric as a statistic for quality analysis of ChIP-seq data with replicates.

Comparative analysis on biological data sets

We evaluate THOR and six competing methods [csaw, MACS2, DiffReps, PePr, DiffBind, DeSeq-IDR, Poisson-THOR (we were not able to execute JAMM on these data sets that was therefore left out for this analysis)] on 13 DPC problems using data from the cocaine intake on mouse (CO), dendritic cell differentiation (DC), B cell follicular lymphoma (LYMP) and monocyte differentiation (MM) study (see Section ***Biological Data Sets*** and Supplementary Section 3 for details). We also evaluate the application of THOR with either the TMM (THOR-TMM) or the housekeeping genes (THOR-HK) normalization approach.

The performance of methods was evaluated by comparing DPs with differential expression and activating histone modifications of neighboring genomic regions. This evaluation methodology is justified by previous work indicating the correlation of activating histone marks with gene expression (2,38,39) and other activating histone modifications (40). As a sanity check, we measured the correlation of all pairs of histones and RNA-seq signals from a single sample of each of the 11 individual cell types (Supplementary File 1). We observed significantly positive Spearman correlations (average R = 0.52) between all 39 pairs of sequencing libraries measured in the same conditions. The correlation between activating histone modification marks are higher (average R = 0.66) than the correlation between histones and RNA-seq (average R = 0.38).

In short, the DCA is based on the Spearman correlation between the P-values of top k DPs (estimated by the evaluated DPC) and P-values of differential expression of

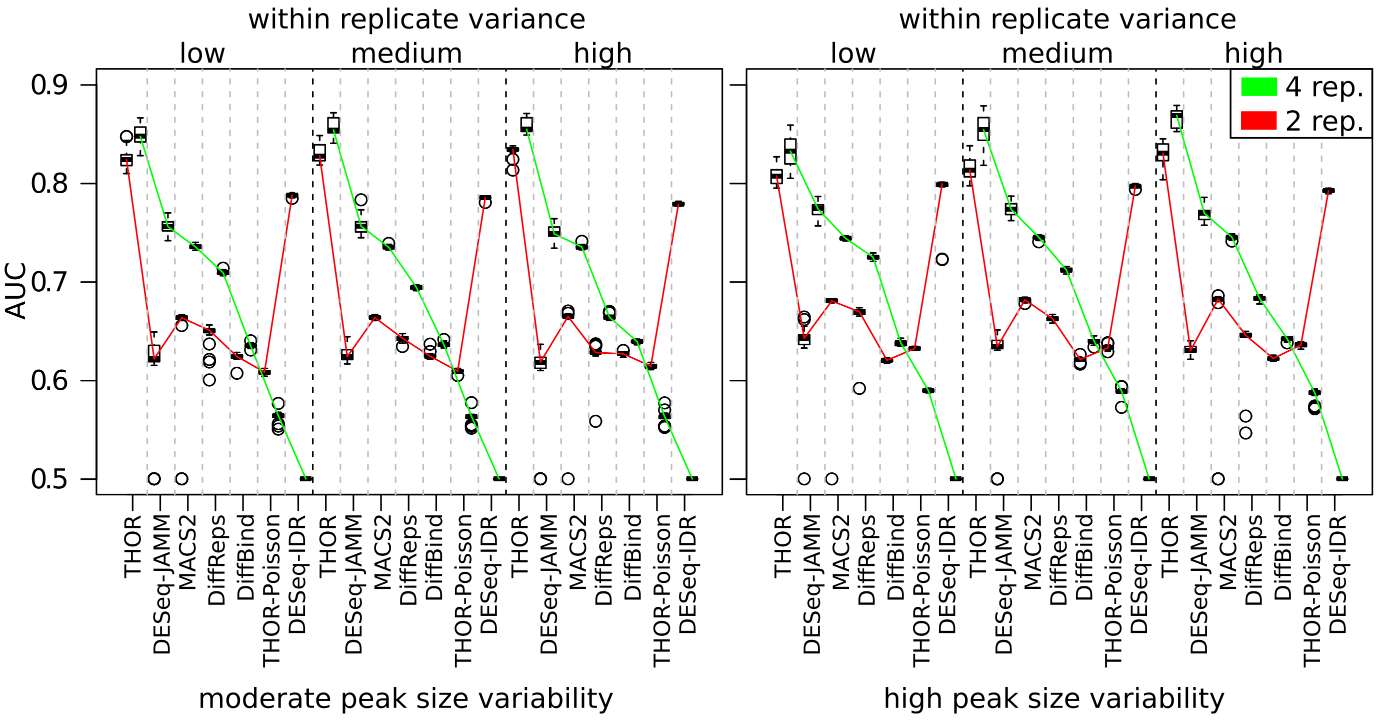

Figure 2. Results for simulated data. We show the AUC distribution for 25 repetitions of each scenario. Simulated data were based on (**A**) moderate and (**B**) high condition peak size variability and 2 (red lines) and 4 (green lines) replicates. Each boxplot is divided by the level of within condition variance (low, medium and high). Methods (x-axis) are ordered by decreasing median AUC values for the cases with 4 replicates.

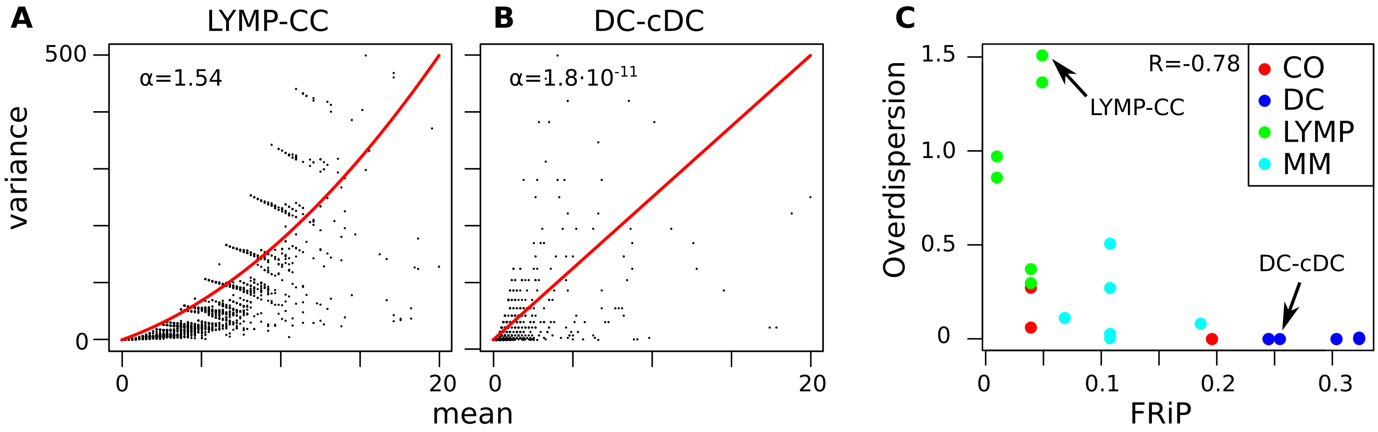

Figure 3. Association between average FRIP and overdispersion scores α. (**A–B**) We show the relation between mean and variance of replicates on two selected experimental conditions (LYMP-CC) and (DC-cDC). (**C**) FRIP and overdispersion scores for the 26 biological conditions analysed: cocaine intake (CO), monocyte differentiation (MM), lymphoid cancer (LYMP) and dendritic cell differentiation (DC). Higher FRIP indicates higher signal-to-noise ratio and better ChIP-seq experiments. Higher overdispersion scores indicates higher within condition variability.

regions neighboring the top k DPs (estimated by DESeq (29) or limma (50)). Higher values indicate higher association between the histone modifications and gene expression and therefore better performance of the DPC method. The DCA curves are obtained by evaluating the DCA for increasing k (see Section ***Evaluation of Biological Data Sets***). Selected DCA curves are seen in Figure 4 and the complete results are shown in Supplementary Figures S6–S13.

We use the Friedman–Nemenyi test to check for significant differences in the area under the DCA curves. Concerning the gene expression based DCA, THOR variants are the best ranked methods and have significantly higher DCA values than DESeqIDR, csaw and Poisson-THOR (Supplementary Table S18–S19, adjusted P-value < 0.05). THOR-HK has also significantly higher DCA scores than DiffBind and DiffReps (adjusted P-value < 0.1). As PePr requires input-DNA data and therefore cannot be executed for MM and CO, we repeated the Friedman–Nemenyi test on DCA values from DC and LYMP only. In this case, THOR variants have significantly higher DCA score than csaw and Poisson–THOR (P-value < 0.05, Supplementary Table S20 and S21). There is no significant differences for

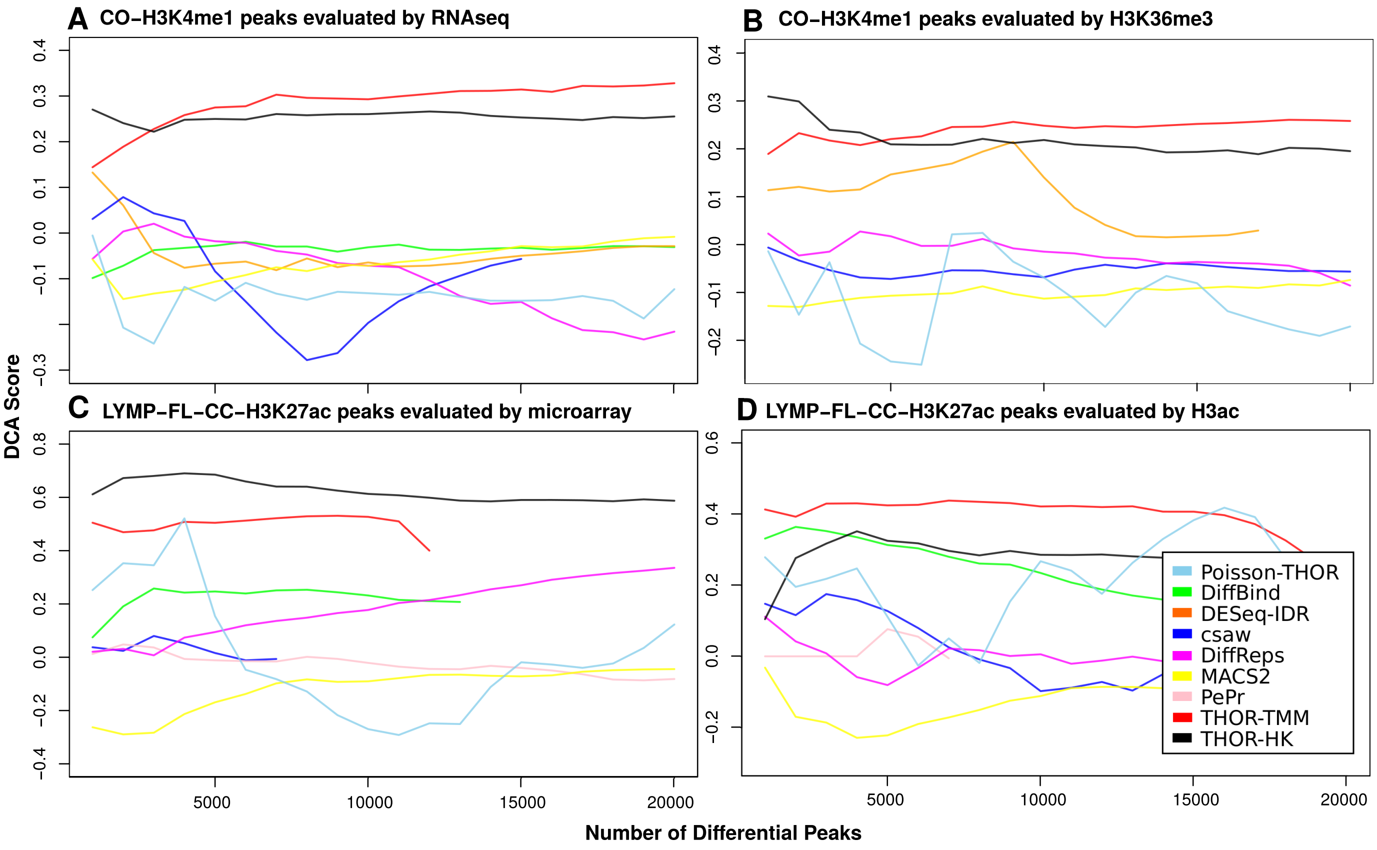

Figure 4. DCA curves for selected DP problems. We show DCA curves for peaks detected in H3K4me1 of the Cocaine Response study (CO-H3K4me1), which were evaluated by (**A**) RNA-seq and (**B**) H3K36me3; or H3K27ac peaks of the Lymphoma study (LYMP-FL-CC-H3K27ac), which were evaluated by (**C**) microarrays and (**D**) H3ac histone. Higher DCA values indicate higher association between differential peaks and differential expression or validating histones.

all other competing methods. For the histone based DCA scores, THOR variants are the best ranked methods and have significantly higher DCA values than csaw, DESeqIDR, macs2, DiffReps and Poisson–THOR (Supplementary Table S22 and S23, adjusted *P*-value < 0.05). In the scenario containing PePr (and therefore no MM and CO data sets), only THOR-TMM has significantly higher DCA scores than PePr (Supplementary Table S24 and S25, adjusted *P*-value < 0.05).

Impact of overdispersion on differential peak calling

As previously described, the overdispersion score indicates LYMP experiments has higher within peak variability, while the dendritic cell differentiation study has the lowest (Figure 3). Interestingly, the DCA scores support the notion that THOR has better relative performance than competing tools in data with high overdispersion such as LYMP-FL-CC (Figure 4C), while it performs comparatively well with other competing methods in differential peak problems from the DC data set (Supplementary Figure S7). Indeed, we observe a moderate association between Δ DCA and the overdispersion score (R = 0.38 for expression based DCA and R = 0.30 for histones based DCA; adjusted *P*-value < 0.1; Supplementary Figure S14C).

Another important question is the performance of the two normalization approaches supported by THOR. The difference in ranks between the TMM and HK approaches based on gene expression or histone modification DCA are not statistically significant (Supplementary Table S18–S25). Considering the difference in expression based DCA scores for THOR-HK and THOR-TMM (Supplementary Figure S14A and B and Supplementary File 1), we observe that both methods perform similarly in most data sets. However, expression based DCA of THOR-HK are higher than THOR-TMM in four conditions from LYMP (LYMP-CC-PBBA gain, LYMP-CC-PBBA loose, LYMP-FL-CC loose and LYMP-FL-PBBA loose Supplementary Figure S8 and Supplementary File 1). These experiments have the worst quality scores among all analysed data sets, i.e. FRIP < 0.05 and large overdispersion > 0.03 as shown in Figure 3C. This suggests the use of house keeping gene normalization for data sets with lower quality.

Example of differential peaks

As an example, we show DPs in the vicinity of regions discussed in the original publications providing the ChIP-seq data (6,7,37). In Figure 5A, we display PDK2 and IRAK3 that show respectively an increase (decrease) in expression during monocyte to macrophage differentiation. THOR calls a combination of gain (green) and lost (red) DPs in H3K4me3 levels in the promoter of IRAK3 and PDK2. MACS2 only detects a small peak in the IRAK3 promoter

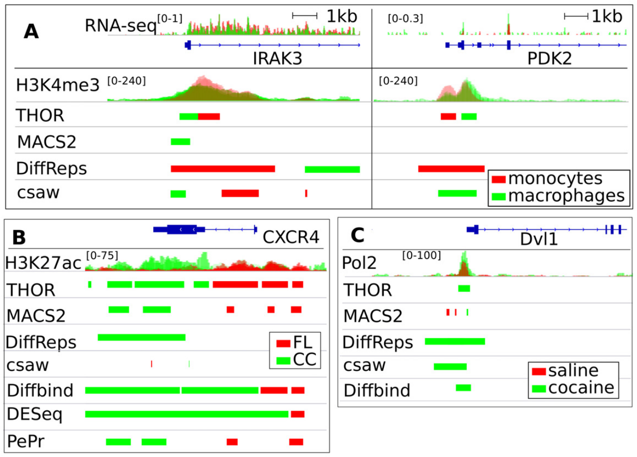

Figure 5. Example of differential peaks. We depict H3K4me3 and RNA-seq signals for monocytes (red) and macrophages (green) around the promoter of (**A**) IRAK3 and PDK2; (**B**) H3K27ac signals around CRCX4 for Follicular Lymphoma (FL) and control (CC) individuals; and (**C**) Pol2 signal around Dvl1 for mouse treated with cocaine (green) and saline (red); Below the ChIP-seq signals, we depict differential peaks of all evaluated methods. Methods that do not detect peaks for a given experiment are not listed.

and DiffReps identifies rather large lost peaks (red) in both promoters. Csaw misses regions with largest histone losses in both genes. DESeq-IDR and PePr do not call any DPs in this region. In Figure 5B, we show H3K27ac signal around the chemokine receptor CXCR4, which is a marker for definition of B cell sub-populations. CXCR4 was discussed in (7), as it has increased expression and H3K27ac marks in patients with Follicular Lymphoma (FL). THOR detected three FL gain peaks (red) in the promoter of CXCR4 and four FL lost peaks (green) in the gene body. MACS2 and PePr detected similar peaks with sizes smaller than supported by the ChIP-seq signals. DiffReps and csaw fail to detect FL gain peaks in the promoter of CXCR4, while DESeq-IDR and DiffBind detected rather large lose peaks (green).

Finally, Figure 5C depicts Pol2 ChIP-seq signals in chronic cocaine study around the Dvd1 gene that is shown to have increased expression in response to cocaine (6). We observe a typical DNA binding protein ChIP-seq profile for Pol2, i.e. a small well defined peak. THOR and DiffBind detects a peak gained (green) in cocaine treated mice that nicely fits the ChIP-seq profile. DiffReps and csaw call larger and MACS2 calls smaller gain peaks than the ChIP-seq signal supports. PePr and DESeq-IDR do not detect any peak in this region. Altogether, we observe that THOR peaks nicely delineate changes in ChIP-seq profiles. In all cases, we observed a tendency of MACS2 to detect small peaks, and of DiffReps to call large peaks. Indeed, the average peak size of all biological data supports the fact that DiffReps tends to call larger (1893 bps) and MACS2 smaller DPs (296 bps) than the average peak size of all other tools (1133 bps) (Supplementary Figure S15).

CONCLUSION

There are very few methods dealing with the integrative analysis of multiple ChIP-seq signals. In particular, no study has previously evaluated DPC methods with replicates of histone modifications ChIP-seq experiments with distinct data characteristics. Our evaluation analysis based on simulated and read data sets evaluates all DPC methods natively dealing with replicates. Overall, THOR outperformed competing methods for most simulated and biological data sets. This was particularly the case for data with high overdispersion and low quality. The best performance of THOR is justified from methodological aspects of THOR, as it is the only method that intrinsically analyses windows of varying size during the detection of DPs. Other competing methods are based on fixed window searches (PePr, DiffReps, csaw) or pre-defined peaks (DESeq-JAMM, DESeq-IDR, DiffBind). This makes THOR unique in the analysis of ChIP-seq data of histone modifications that usually occurs in domains with small changes of ChIP-seq signal. Moreover, as indicated in Supplementary Table S3, which lists features of all evaluated tools, THOR is the most complete differential peak caller concerning support of typical computational steps for ChIP-seq analyses (31). For example, some methods do not provide input-DNA normalization (csaw), while other methods are only able to run on data without input-DNA (PePr).

The performance of competing tools varied across distinct experiments. While DESeq-IDR performed well on simulated data cases with low within condition variance and low number of replicates, it failed to call peaks on data with large variance. This is expected as IDR was conceived for a conservative peak detection on technical replicates. JAMM (with DESeq) had good performance on simulated data (2nd rank) and is the only framework performing integrative analysis of SPC problems with replicates. Some methods, such as PePr and DiffReps, had a tendency to call peaks larger than other tools and the observed histone changes. This explains the below average performance of these methods in our evaluation. MACS2 was ranked third on simulated data and second on expression based DCA. However, visual inspection indicates that its peaks are covering smaller regions than indicated by the ChIP-seq signal. Finally, Poisson–THOR, which can be seen as a version of ODIN supporting replicates with a distribution not coping with overdispersion, has poor results in most evaluated scenarios. This reinforces the importance of support to overdispersion on the presence of replicates.

ChIP-seq experiments still requires protocol improvements for allowing comparable signals (54). Here, we explore computational approaches for ChIP-seq normalization and alternatives to the TMM approach. This normalization strategy, which is usually employed in gene expression analysis, is based on the assumption that the number of reads in most peaks does not change over conditions. We propose here the use of housekeeping genes as a strategy to normalize ChIP-seq signals for activating histone marks. As indicated by our analysis, this approach leads to the best ranked method on the expression based DCA. In particular, DCA values for THOR with housekeeping genes normalization were higher for experimental conditions from the follicular lymphoma study. This study has overall lowest quality statistics (FRIP) and highest within condition variance scores (overdispersion). Indeed, THOR framework includes the estimation of overdispersion quality measures that can be used to guide the choice of normalization strategy.

Despite a great number of methods for the detection of differential peaks in ChIP-seq experiments, there has been few efforts on benchmarking strategies or studies (9,30). This work contains the most comprehensive evaluation study on differential peak calling with replicates with focus on histone modifications. We evaluate 10 differential peak calling methods using three evaluation strategies: DCA with gene expression, DCA with histone modifications and simulated data. These methods are evaluated on 13 differential peak calling problems based on 'real' ChIP-seq data and 150 problems on simulated data. The code for evaluation of methods and resulting statistics are available for the research community. This provides a useful resource for future work proposing differential peak calling methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Interdisciplinary Center for Clinical Research (IZKF Aachen), RWTH Aachen University Medical School, Aachen, Germany; Excellence Initiative of the German Federal and State Governments and the German Research Foundation [GSC 111]; START-program (AZ 22/13) of the Faculty of Medicine, RWTH Aachen; Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq/Brazil) through the Science Without Borders program; Donation by U. Lehmann and funds of German Research Foundation [DFG ZE432/5-2 to M.Z. in part]; this study makes use of data generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available at www.blueprint-epigenome.eu; European Union's Seventh Framework Programme (FP7/2007-2013) [282510 – BLUEPRINT]. Funding for open access charge: IZKF Aachen and German Research Foundation (GSC 111).
Conflict of interest statement. None declared.

REFERENCES

1. Dahl,J.A. and Collas,P. (2008) MicroChIP–a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res.*, **36**, e15.
2. Dunham,I., Kundaje,A., Aldred,S. F., Collins,P. J., Davis,C. A, Doyle,F. and Lochovsky,L. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
3. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A. and Kellis,M. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
4. Martens,J.H.A. and Stunnenberg,H.G. (2013) BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, **98**, 1487–1489.
5. Lara-Astiaso,D., Weiner,A., Lorenzo-Vivas,E., Zaretsky,I., Jaitin,D.A., David,E., Keren-Shaul,H., Mildner,A., Winter,D., Jung,S. *et al.* (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science*, **345**, 943–949.
6. Feng,J., Wilkinson,M., Liu,X., Purushothaman,I., Ferguson,D., Vialou,V., Maze,I., Shao,N., Kennedy,P., Koo,J. *et al.* (2014) Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biol.*, **15**, R65.
7. Koues,O.I., Kowalewski,R.A., Chang,L.-W., Pyfrom,S.C., Schmidt,J.A., Luo,H., Sandoval,L.E., Hughes,T.B., Bednarski,J.J., Cashen,E.F. *et al.* (2015) Enhancersequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in B-Cell lymphoma. *Immunity*, **42**, 186–198.
8. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
9. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.
10. Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.
11. Ashoor,H., Hérault,A., Kamoun,A., Radvanyi,F., Bajic,V.B., Barillot,E. and Boeva,V. (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, **29**, 2979–2986.
12. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
13. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
14. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
15. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PloS One*, **5**, e11471.
16. Landt,S.G., Marinov,G.K., Kundaje,K., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
17. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
18. Ibrahim,M.M., Lacadie,S.A. and Ohler,U. (2015) JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, **31**, 48–55.
19. 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**, 576–89.
20. Xu,H., Wei,C.-L., Lin,F. and Sung,W.-K. (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.
21. Liang,K. and Keleş,S. (2012) Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, **28**, 121–122.
22. Shao,Z., Zhang,Y., Yuan,G.-C.C., Orkin,S.H. and Waxman,D.J. (2012) MAnorm: A robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.
23. Stark,R. (2012) Research in Computational Molecular Biology: 16th Annual International Conference, RECOMB 2012, Barcelona, Spain, April 21-24, 2012. *Proceedings chapter Differential Oestrogen Receptor Binding is Associated with Clinical Outcome in Breast Cancer*, Springer, Berlin Heidelberg; Heidelberg, pp. 286–286.
24. Shen,L., Shao,N.-Y., Liu,X., Maze,I., Feng,J. and Nestler,E.J. (2013) diffReps: Detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One*, **8**, e65598.
25. Allhoff,M., Sere,K., Chauvistré,H., Lin,Q., Zenke,M. and Costa,I.G. (2014) Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics*, **30**, 3467–3475.
26. Zhang,Y., Lin,Y.-H., Johnson,T.D., Rozek,L.S. and Sartor,M.A. (2014) PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, **30**, 2568–2575.

27. Lun,A. T.L. and Smyth,G.K. (2014) De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: Controlling error rates correctly. *Nucleic Acids Res.*, **42**, e95.
28. Heinig,M., Colomé-Tatché,M., Taudt,A., Rintisch,C., Schafer,S., Pravenec,M., Hubner,N., Vingron,M. and Johannes,F. (2015) histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics*, **16**, R40.
29. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
30. Maze,I., Shen,L., Zhang,B., Garcia,B.A., Shao,N., Mitchell,A., Sun,H., Akbarian,S., Allis,C.D. and Nestler,E.J. (2014) Analytical tools and current challenges in the modern era of neuroepigenomics. *Nat. Neurosci.*, **17**, 1476–1490.
31. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
32. Cheung,M.-S., Down,T.A., Latorre,I. and Ahringer,J. (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.*, **39**, e103.
33. Lin,Q., Chauvistre,H., Costa,I.G., Gusmao,E.G., Mitzka,S., Hanzelmann,S., Baying,B., Klisch,T., Moriggl,R., Hennuy,B. *et al.* (2015) Epigenetic program and transcription factor circuitry of dendritic cell development. *Nucleic Acids Res.*, **43**, 9680–9693.
34. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
35. Humburg,P., Helliwell,C. A., Bulger,D. and Stone,G. (2011) ChIPseqR: analysis of ChIP-seq experiments. *BMC Bioinformatics*, **12**, R39.
36. Felker,P., Seré,K., Lin,Q., Becker,C., Hristov,M., Hieronymus,T. and Zenke,M. (2010) TGF-beta1 accelerates dendritic cell differentiation from common dendritic cell progenitors and directs subset specification toward conventional dendritic cells. *J. Immunol.*, **185**, 5326–5335.
37. Saeed,S., Quintin,J., Kerstens,H.H.D., Rao,N.A., Aghajanirefah,A., Matarese,F., Cheng,S.-C., Ratter,J., Berentsen,K., van der Ent,M.A. *et al.* (2014) Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, **345**, 1251086.
38. Karlić,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2926–2931.
39. Costa,I.G., Roider,H.G., do Rego,T.G. and de Carvalho,F. d. A.T. (2011) Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, **12**(Suppl 1), S29.
40. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Peng,W., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.
41. Mammana,A., Vingron,M. and Chung,H.-R. (2013) Inferring nucleosome positions with their histone mark annotation from ChIP data. *Bioinformatics*, **29**, 2547–2554.
42. Diaz,A., Park,K., Lim,D.A. and Song,J.S. (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol.*, **11**, R9.
43. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
44. Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
45. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**(Suppl. 11), 22–32.
46. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B*, **57**, 289–300.
47. Chauvistré,H., Kustermann,C., Rehage,N., Klisch,T., Mitzka,S., Felker,P., Rose-John,S., Zenke,M. and Sere,K.M. (2014) Dendritic cell development requires histone deacetylase activity. *Eur. J. Immunol.*, **44**, 2478–2488.
48. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.
49. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
50. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
51. Cheng,C., Alexander,R., Min,R., Leng,J., Yip,K. Y., Rozowsky,J., Snyder,M. and Gerstein,M. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.
52. Demšar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.
53. Levenberg,K. (1944) A method for the solution of certain problems in least squares. *Quart. Applied Math.*, **2**, 164–168.
54. Arrigoni,L., Richter,A.S., Betancourt,E., Bruder,K., Diehl,S., Manke,T. and Bönisch,U. (2016) Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res.*, **44**, e67.