METHOD

chromstaR: Tracking combinatorial chromatin state dynamics in space and time

Aaron Taudt[1], Minh Anh Nguyen[4], Matthias Heinig[3], Frank Johannes[2] and Maria Colomé-Tatché[1,3*]

Abstract

Background: Post-translational modifications of histone residue tails are an important component of genome regulation. It is becoming increasingly clear that the combinatorial presence and absence of various modifications define discrete chromatin states which determine the functional properties of a locus. An emerging experimental goal is to track changes in chromatin state maps across different conditions, such as experimental treatments, cell-types or developmental time points.

Results: Here we present chromstaR, an algorithm for the computational inference of combinatorial chromatin state dynamics across an arbitrary number of conditions. ChromstaR uses a multivariate Hidden Markov Model to determine the number of discrete combinatorial chromatin states using multiple ChIP-seq experiments as input and assigns every genomic region to a state based on the presence/absence of each modification in every condition. We demonstrate the advantages of chromstaR in the context of three common experimental data scenarios. First, we study how different histone modifications combine to form combinatorial chromatin states in a single tissue. Second, we infer genome-wide patterns of combinatorial state differences between two cell types or conditions. Finally, we study the dynamics of combinatorial chromatin states during tissue differentiation involving up to six differentiation points. Our findings reveal a striking sparcity in the combinatorial organization and temporal dynamics of chromatin state maps.

Conclusions: chromstaR is a versatile computational tool that facilitates a deeper biological understanding of chromatin organization and dynamics. The algorithm is implemented as an R-package and freely available from http://bioconductor.org/packages/chromstaR/.

Keywords: epigenetics; histone modification; chromatin state map; ChIP-seq; computational biology

*Correspondence:
maria.colome@helmholtz-muenchen.de
[1]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, A. Deusinglaan 1, NL-9713 AV Groningen, The Netherlands
Full list of author information is available at the end of the article

Introduction

Epigenetic marks such as DNA methylation or histone modifications play a central role in genome regulation. They are involved in a diversity of biological processes such as lineage commitment during development [1], maintenance of cellular identity [2, 3] and silencing of transposable elements [4]. The modification status of

many histone marks has been extensively studied in recent years, first with ChIP-chip and later with ChIP-seq, now the de-facto standard procedure for genome wide mapping of protein-DNA interactions and histone modifications. Since its advent in 2007 [1, 2, 5], ChIP-seq technologies have been widely used to survey genome-wide patterns of histone modifications in a variety of organisms [6, 7, 2], cell lines [8] and tissues [9, 10].

The multitude of possible histone modifications has led to the idea of a "histone code" [11, 12], a layer of epigenetic information that is encoded by combinatorial patterns of histone modification states (Fig. 1a). Major resources have been allocated in recent years to decipher this code, culminating in projects such as the ENCODE [13] and Epigenomics Roadmap [10]. Following their examples, most experiments nowadays are designed to probe several histone modifications at once, and often in various cell types, strains and at different developmental time points. These types of experiments pose new computational challenges, since initial solutions were designed to analyze one modification and condition at a time, therefore treating them as independent. Indeed, a commonly used strategy has been to perform peak calling for each experiment separately (univariate analysis) and to combine the peak calls post-hoc into combinatorial patterns [14, 15]. This approach is problematic for several reasons: Because of the noise associated with ChIP-seq experiments and peak calling, combining univariate peak calls will lead to the discovery of spurious combinatorial states that do not actually occur in the genome. Furthermore, different tools or parameter settings are often used for different modifications (e.g. peak calling for broad or narrow marks), making the outcome sensitive to parameter changes and control of the overall false discovery rate difficult. Lastly, this approach requires ample time and bioinformatic expertise, rendering it impractical for many experimentalists.

Accurate inferences regarding combinatorial histone modification patterns are necessary to be able to understand the basic principles of chromatin organization and its role in determining gene expression programs. One way forward is to develop computational algorithms that can analyze all measured histone modifications at once (i.e. combinatorial analysis) and across different conditions (i.e. differential

analysis). Some methods have been designed to integrate histone modifications into unified chromatin maps [16, 17, 18, 19, 20, 21, 22]. These methods can be classified into three different categories [23]: combinatorial (which define chromatin states based on the presence/absence of every histone modification) [19], continuous (which define chromatin states based on the shape of the ChIP-seq signal) [24, 25], and probabilistic (which have probabilities associated with finding each mark in a given state) [16, 17, 18, 20, 21, 22]. A major drawback of the majority of these approaches [16, 17, 18, 20, 21, 22] is the need to specify the number of distinct chromatin states beforehand, which is usually not known *a priori*. Moreover, in the probabilistic interpretation the inferred states can consist of multiple and overlapping combinatorial states (Fig. 1b). This probabilistic state definition is useful to reduce noise and to identify functionally similar genomic regions for the purpose of annotation, but at the same time it obscures a more direct interpretation of combinatorial states in terms of the presence/absence patterns of the underlying histone modifications.

Finally, none of these methods is designed for comparing chromatin maps across conditions. ChromDiff [26] and dPCA [27] are comparative methods that identify significant chromatin differences between groups of samples. ChromDiff discovers groups of epigenomic features which are the most discriminative in group-wise comparisons of samples, while dPCA uses a small number of differential principle components to explore differential chromatin patterns between two groups of samples. Both methods are useful for identifying defining features of each group, however they do not provide complete information about the genome-wide localization of all chromatin differences between all samples.

In order to overcome these problems we have developed chromstaR, a method for multivariate peak- and broad-region calling. chromstaR has the following conceptual advantages: 1) Every genomic region is assigned to a discrete, readily interpretable combinatorial chromatin state, based on presence/absence of every histone mark, providing a mechanistic interpretation of chromatin states which allows for better insights into how they regulate genome function. 2) The number of chromatin states does not have to be preselected but is a result of the analysis. 3) Histone modifications with narrow and broad profiles can be combined in a joint analysis along with an arbitrary number of conditions. 4) The same approach can be used for mapping

combinatorial chromatin states in one condition, or for identifying differentially enriched regions between several conditions, or for both situations combined. 5) Our formalism offers an elegant way to include replicates as separate experiments without prior merging.

The algorithm is implemented in C++ and available as R package to combine speed and ease-of-use, and offers functions to perform the most frequent downstream analysis tasks.

We demonstrate the advantages of chromstaR in the context of three common experimental scenarios (Fig. S1b). First, we consider that several histone modifications have been collected on a single tissue at a given time point (Fig. S1b, Application 1). The goal is to infer how these different modifications combine to form distinct combinatorial chromatin states and to describe their genome-wide distribution. Second, we consider that several histone modifications have been collected in two different cell types or conditions (Fig. S1b, Application 2). Here, the goal is to infer genome-wide patterns of combinatorial state differences between cell types or conditions. Third, we consider the more complex secenario where several histone modifications have been collected for multiple different time points or tissue types (Fig. S1b, Application 3). In this case, the goal is to infer how combinatorial chromatin states are modified during tissue differentiation or development. These three experimental scenarios broadly summarize many of the data problems that biologists and bioinformaticians currently face when analyzing epigenomic data. We show that chromstaR provides biologically meaningful results to these types of data problems, and facilitates deeper biological insights into the dynamic coordination of combinatorial chromatin states in genome regulation.

Results

Brief overview of analytical approach and validation of peak calls

Consider N ChIP-seq experiments: N histone modifications measured in one condition, or one histone modification measured in N conditions, or a combination of the two. After mapping the sequencing reads to the reference genome our method consists of two parts (Fig. 2), a univariate peak calling step to estimate the dis-

tribution parameters, and a multivariate peak calling step to integrate information from all experiments:

(1) Univariate peak calling (Fig. 2a): For each ChIP-seq experiment, we partition the genome into non-overlapping bins (default 1kb) and count the number of reads that map into each bin (i.e. the read count) [28]. We model the read count distribution as a two-component mixture of zero-inflated negative binomials [29, 30], with one component at low number of reads that describes the background noise and one component at high number of reads describing the signal. We use a univariate Hidden Markov Model (HMM) with two hidden states (i.e. unmodified, modified) to fit the parameters of these distributions [31].

(2) Multivariate peak calling (Fig. 2b): We consider all ChIP-seq experiments at once and assume that the multivariate vector of read counts is described by a multivariate distribution which is a mixture of 2^N components. We use a multivariate HMM to assign every bin in the genome to one of the multivariate components. The multivariate emission densities of the multivariate HMM, with marginals equal to the univariate distributions from step (1), are defined using a Gaussian copula [32]. A detailed description can be found in **Methods**.

The univariate part of our model (step 1 above), which serves as the basis for the construction of the multivariate model, provides high-quality peak calls that measure up against existing methods. We compared our method with two commonly used peak callers, Macs2 [33] and Sicer [34], using publicly available datasets of qPCR validated regions [35]. We compared the performance of the three methods on two datasets, one for H3K4me3 (narrow profile), and one for H3K27me3 (broad profile). The H3K4me3 dataset had 33 qPCR validated regions and the H3K27me3 dataset had 197 qPCR validated regions. The ChIP-seq datasets were analyzed with the standard settings of each peak caller (**SI text**), and each base pair was assigned a score by the algorithm. This output was used to compute receiver operator characteristic (ROC) curves and area-under-curve (AUC) values [36]. The performance of chromstaR for these datasets in terms of the AUC is equal or better than that of Macs2 and Sicer (Fig. S2). For the multivariate peak calling, in the absence of datasets with validated peak calls for multiple marks, we performed a simulation study based on parameters obtained from real data (**SI text**). ROC curves and AUC values show that multivariate peak calls are of high quality (Fig. S3).

Application 1: Mapping combinatorial chromatin states in a reference tissue

Lara-Astiaso et al. [37] measured four histone modifications (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) and gene expression in 16 mouse hematopoietic cell lines and their progenitors (Fig. S1). All four marks have a relatively narrow ChIP-seq profile. The authors' goal was to document the dynamic enhancer landscape during hematopoietic differentiation. With four measured histone modifications there are $2^4 = 16$ possible combinatorial states defined by the presence/absence of each of the modifications. In order to provide a snapshot of the genome-wide distribution of these combinatorial states in a given cell-type, we applied chromstaR to the ChIP-seq samples collected from monocytes (see Fig. S6 for the analysis of other cell types). In the following we introduce a shorthand notation where combinatorial states are denoted between brackets [ ] and each mark is abbreviated by its chemical modification. For example, the combination [H3K4me1+H3K4me2+H3K27ac] will be abbreviated as [me1/2+ac]. If we use the full name of a mark (e.g. "H3K4me1") we are referring to the mark in a classical, non-combinatorial, context. See Fig. 3d for all combinations with shorthands.

chromstaR found that many of the 16 possible combinatorial states were nearly absent at the genome-wide scale, with 7 of the 16 states accounting for nearly 100% (>99.99%) of the genome (Fig. 3a). This observation indicates that the "histone code" defined by these four histone modifications is much less complex than theoretically possible, perhaps as a result of biochemical constraints on the co-occurance of certain modifications on the same or neighboring aminoacid residues. However, some of the discovered chromatin states display "incompatible" combinations (the ones displaying more than two modifications on the same histone and residue, such as for example [me1/2/3]). Re-analysis of the data finds eight of the 16 states present in the genome, with a smaller frequency of incompatible states (Fig. S4). These results show that these states are in part due to having pooled data from several nucleosomes into the same bin, but are probably also caused by antibody cross-reactivity and residual cell heterogeneity.

The empty state [ ], which we here define as the simultaneous absence of all measured marks at a given genomic position, was the most frequent state, covering 94.8% of the genome. The high prevalence of this state reflects the fact that

Lara-Astiaso et al. [37] focused on marks with a narrow profile that had previously been shown to occur proximal to genic sequences [38, 2, 3]. Indeed, only 36% of the empty state overlapped known genes while the remaining 64% mapped to non-genic regions throughout the genome, and probably tag other (unmeasured) histone modifications, such as repressive heterochromatin-associated marks.

In order to evaluate chromatin state frequencies on a data set with a mixture of broad and narrow histone modifications, we analyzed human Hippocampus tissue data from the Epigenomics Roadmap with seven marks [10] and IMR90 cell line data from the ENCODE project with 26 marks [13] (see **Methods**). In the Hippocampus data we found that only 21 out of the 128 possible combinatorial states were necessary to explain more than 99% of the epigenome, and indeed the empty state covered only 32% of the genome (Fig. S7). Moreover, in the lung fibroblast cell line we found that from the possible 67 million states only 0.02% (∼12000) are needed to explain more than 95% of the genome, while the empty state covered only 16% of it, showing that when more marks are included the percentage of the genome in the empty state decreases [23].

Contrary to the empty state, on average 68% (range: 59-81%) of the genomic regions found to be in one of the 6 most frequent (non-empty) combinatorial states in mouse monocytes overlap known genes (Fig. 3b), thus suggesting an active role in the regulation of gene expression. To assess this, we examined the combinatorial state profiles of the 6 most frequent states relative to the transcription start site (TSS) of expressed and non-expressed genes (Fig. 4a). In contrast to non-expressed genes, expressed genes were clearly characterized by the presence of state [me1/2/3+ac] proximal to the TSS. This is consistent with previous reports that have used H3K4me3 together with H3K27ac to tag active promoters [39]. However, our analysis also uncovered a more subtle enrichment of state [me1] shouldering the TSS (Fig. 4a). We found that 18% of [me1] sites occur in regions directly flanking state [me1/2/3+ac] and 61% of all [me1] can be found within 10kb of [me1/2/3+ac] sites (see Fig. 5 for an example). These two states therefore constitute a single, broad chromatin signature that defines a subset of expressed genes. Interestingly, this subset of genes had significantly higher expression levels

($p \approx 10^{-101}$, Wilcoxon rank-sum test) and distinct GO terms compared with genes marked only by the active promoter state (i.e. [me1/2/3+ac] at the TSS and no [me1] in flanking regions, Fig. 6 and Table 1). This observation suggests that the co-occurance of [me1/2/3+ac] and [me1] in broad regions surrounding the TSS marks what may be called "enhanced" active promoters ([me1/2/3+ac]+[me1]).

To compare the results obtained with chromstaR to other computational approaches, we first re-analyzed replicate datasets from the mouse monocytes [37], the human Hippocampus [10] and the lung fibroblast cell line data [13] using several publicly available methods [16, 19, 21, 22] (**SI text**). chromstaR is the method that provides the best performance in assigning a consistent segmentation between replicates (Fig. S5c) and in detecting regions with high read count fold change as differential (Fig. S5d). Among the alternative methods, ChromHMM provides the best performance and flexibility [16], we will therefore use it in the following for comparison purposes. ChromHMM employs a multivariate HMM to classify the genome into a preselected number of probabilistic chromatin states, and was used to annotate the epigenome in the ENCODE [13] and Epigenomics Roadmap [10] projects. It therefore offers a method to segment the genome into a set of probabilistic chromatin states that can then *a posteriori* be interpreted at the biological level. We also compare the results obtained with chromstaR to the ones obtained using MACS2 [33], because it is one of the most widely used univariate peak callers.

When using a multivariate segmentation method like ChromHMM, the number of chromatin states needs to be decided beforehand, which is difficult as this number is rarely known *a priori*. In the absence of detailed guidelines we fitted a 16 state model to the mouse hematopoietic data. Our comparison uncovered substantial method-specific differences in state frequencies (Fig. 3). Both ChromHMM and MACS2 found all 16 states present in the genome with more than 0.01% genome coverage. To understand how state-calls compared between methods, we evaluated to which extent the states detected by one method coincided with those detected by the other method(s) (Fig. S8). Most notable, we found that genomic regions corresponding to chromstaR's active promoter state [me1/2/3+ac] were assigned to two alternative states (E7 and E9) by ChromHMM. These latter two states were very similar in terms of their emission densities, but significantly different at the

level of gene expression ($p \approx 10^{-90}$, t-test, Fig. 3c). Moreover, chromstaR's single empty state [ ] corresponded to two functionally similar (nearly) empty states (E3, E4) detected by ChromHMM. A third almost empty state E2 with weak H3K27ac signal had slightly higher expression levels than the other two empty states and also overlapped with chromstaR's empty state [ ] (Fig. S8b). These state redundancies highlight the difficulty in selecting the number of chromatin states for ChromHMM, for without extensive manual curation it is difficult to know if two states are truly redundant (likely E3 and E4) or if they are biologically different on some level (E7 and E9).

Although MACS2 is not designed for multivariate analysis, we constructed *ad hoc* combinatorial state calls from the univariate analyses obtained from each ChIP-seq experiment to illustrate the problems of this commonly used analysis technique. As expected, MACS2 results were noisy: many of the combinatorial states detected by chromstaR showed very heterogenous state calls with MACS2 (Fig. S8a). For instance, a considerable proportion (35%) of genomic regions detected by chromstaR as being in the active promoter state [me1/2/3+ac] were assigned to another promoter state (containing H3K4me3) by MACS2. We suspect that this is due to the limitations of MACS2 in calling broader marks (e.g. H3K4me1) or moderate enrichment with the default parameters, which results in frequent missed calls for individual modifications, and subsequently also in the limited detection of 'complex' combinatorial states such as [me1/2/3+ac] that are defined by the presence of all modifications.

To better understand the functional implications of the state frequency and state pattern differences between these methods, we evalute the chromatin state signatures of both ChromHMM and MACS2 around TSS of expressed and non-expressed genes (Fig. 4b,c). In contrast to chromstaR, chromatin signatures obtained by the other two methods did not as effectively distinguish these two classes of genes, suggesting that chromstaR has a higher sensitivity for detecting these signatures (Table 2).

Application 2: Differential analysis of combinatorial chromatin states

In order to understand combinatorial chromatin state signatures that are specific to a given cell type or disease state, it is necessary to compare at least two different tissues with each other, or a case and a control. In this context, the goal is to identify genomic regions showing differential (or non-differential) combinatorial state patterns. Such differential patterns are indicative of regions that underly the tissue differences and are therefore of substantial biological or clinical interest. chromstaR solves this problem by considering all 2^{2N} possible combinatorial/differential chromatin states (Fig. 1c), where N is the number of histone modifications measured in both conditions. Out of the 2^{2N} states, 2^N are non-differential and $2^{2N} - 2^N$ are differential.

We anlayzed two differentiated mouse hematopoietic cells (monocytes versus CD4 T-cells) from [37], with four histone marks each (H3K4me1, H3K4me2, H3K4me3 and H3K27ac). We found that 5.37% of the genome showed differences in combinatorial state patterns between the two cell types (Fig. 7a, example browser shot in Fig. 8). The most frequent differential regions involved the [me1] combination (2.37%) followed by regions with the [me1/2/3+ac] combination (0.92%). These differences are even more striking when viewed in relative numbers: 59% of the [me1/2/3+ac] sites were concordant between the two cell types, while only 8% of the [me1] sites were concordant. This is in line with previous findings showing that H3K4me1 is highly cell type specific [40, 41, 42, 43].

In order to determine if these differences in chromatin play a role in cellular identity, we explored gene expression differences for differential chromatin states. We found that loss of state [me1] as well as of state [me1/2/3+ac] is correlated with a decrease in expression levels (Fig. 7b). This is consistent with our previous observation (section **Application 1**) that [me1/2/3+ac] defines active promoters and [me1] together with [me1/2/3+ac] defines enhanced active promoters (Fig. 6). To investigate the function of the differential loci, we performed a GO term enrichment of these regions [44] and found an impressive confirmation of cell type identity in the GO terms (Table S1): While regions that are marked by [me1/2/3+ac] or [me1] in both cell types show enrichment for general immune cell differentiation

terms, regions that are marked with [me1] or [me1/2/3+ac] only in CD4 T-cells show terms such as "T-cell activation and differentiation". Vice versa, regions that are marked with those signatures in monocytes but not in T-cells show enrichment of terms such as "response to other organism" and "inflammatory response".

Again, we compared our results on the same dataset with MACS2 [33] and ChromHMM [16]. Neither method was specifically designed to deal with differences between combinatorial states, but both tools represent approaches that could have been chosen for that task in the absence of other suitable methods. For both methods, the percentage of the epigenome that was differentially modified was found to be 2.5 times higher than predicted by chromstaR, 13.02% for MACS2 and 13.59% for ChromHMM. MACS2 found most differences (3.90%) in state [me1], followed by the combination [me2+ac] (2.11%). None of these states yielded any significant enrichment in GO terms or showed correlation with expression data (Fig. S9c and Table S2). The third most frequent differential state was [me1+ac] (1.88%) and this state yielded GO term enrichments which reflect cellular identity. ChromHMM predicted two "enhancer-like states" E8 and E9 (Fig. S9b) as most differential between cell types (2.71% and 2.54%) which also showed cell type specific terms in the GO analysis (Table S3). However, expression analysis showed that ChromHMM's most frequent differential state (CD4:E12 and Mono:E14) corresponded to proximal genes that were transcriptionally nearly inactive (Fig. S9b), which raises the question if these differential chromatin states produce cell-specific functional differences.

Application 3: Tracking combinatorial chromatin state dynamics in time

Arguably the most challenging experimental setup is when several histone modifications have been collected for a large number of conditions, such as different cell types along a differentiation tree or different terminally differentiated tissues (Fig. S1). We consider M conditions with N histone modifications measured in each of them. This leads to 2^N possible combinatorial states per condition, or alternatively to 2^M differential states per mark across all samples. Therefore, the number of possible dynamic combinatorial chromatin states is $2^{M \times N}$. For $M \times N \leq C$ the whole dynamic/combinatorial chromatin landscape is treatable computationally,

while for $M \times N > C$ the problem becomes intractable with current computational resources. The value of C is dependent on computational resources, genome length and bin size (see section **Limitations**).

We considered again the mouse hematopoietic data from [37], with four histone modifications (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) measured in 16 different cell types during hematopoietic differentiation (stem cells, progenitor and terminally differentiated cells). We explored the chromatin dynamics during the differentiation process for every hematopoietic branch (Fig. S1a): first, long term hematopoietic stem cells (LT-HSC) are transformed into short term hematopoietic stem cells (ST-HSC) and further into multipotent progenitors (MPP). The MPP cells differentiate into the several common lineage oligopotent progenitors, giving rise to the three different hematopoietic branches (myeloid, leukocyte and erythrocyte). Finally, after another one or two stages, cells become fully differentiated at the bottom of the tree. Every branch from root to leaf consists therefore of four histone marks in five or six time points, with $2^{M \times N} = 1048576$ or 16777216 possible dynamic combinatorial chromatin states, respectively. Because this number is computationally intractable, we implemented the following two-step approach for each branch: (1) for each of the four histone marks separately, we performed a multivariate differential analysis along the five or six cells in the branch, therefore assigning every bin in the genome to one of the 32 or 64 possible differential states; (2) We reconstructed the full combinatorial chromatin state dynamics by combining the differential calls of all four marks in step 1, bin by bin (Fig. S10a).

Using this two-step approach, we studied the dynamics of the inferred chromatin states over developmental time. We observed an initial increase in the frequency of the [me1] state from the LT-HSC to intermediate progenitor stages, followed by a decrease to the fully differentiated stages (Fig. S11). This decrease in [me1] was especially pronounced in the lymphoid and erythroid lineage. In the [me1/2/3+ac] signature we found a small but continuous decrease from LT-HSC to terminally differentiated stages. These observations are consistent with the view that chromatin transitions from an open configuration in multipotent cells to a closed configuration in differentiated cells. Figure S12 shows two examples of pluripotency genes, Gata2

and Cebpa, that lose their open chromatin configuration in differentiated CD4 T-cells.

We next explored the specific dynamic chromatin state transitions that occur in every region of the genome during the differentiation process. We found that the majority of all possible dynamic chromatin state transitions were not present in this system. For example, in the CD4 T-cell branch of the hematopoietic tree there are 5 developmental time points and at each stage 16 combinatorial states can be theoretically present. This leads to $16^5 = 1048576$ potential transitions between combinatorial states in this branch. However, we found only 1086 different chromatin transitions and the first most frequent 99 transitions (with frequency $\geq 0.01\%$) already involved 99.60% of the genome. To summarize these transitions further, we grouped them into 4 different classes: (1) "Empty" transitions, i.e. those regions that have no histone modification in any of the developmental stages. (2) "Constant" transitions, i.e. those regions that show the same (non-empty) combinatorial state in all stages of differentiation. (3) "Stage-specific" transitions, i.e. those regions that show a combinatorial state only in a subset of differentiation stages and are in the "empty" state otherwise. (4) All other transitions (see Fig. 9 for examples). In the CD4 T-cell branch, 85.98% of the genome has no measured chromatin signature in all 5 stages (class 1). The constant transitions (class 2) comprise 5.87% of the genome, stage-specific transitions 5.69% (class 3) and all other transitions 2.46% (class 4). Altogether, only 8.15% of the genome changes its chromatin state during differentiation and more than half of these changes are due to changes in the [me1] signature. This signature is highly cell type specific and gains and losses correspond to stage-specific terms in a GO analysis (Table S4) and to changes in gene expression (Fig. S13a). Among the constant transitions, regions with signature [me1/2/3+ac] mark constitutively expressed genes (Fig. S13a). Therefore we expect those regions to be enriched with housekeeping functions, which is confirmed by the GO analysis (Table S5).

We compared our results on the CD4 T-cell branch with MACS2 [33] and ChromHMM [16]. Strikingly, MACS2 found 34470 different chromatin state transitions, with the most frequent 330 (with frequency $\geq 0.01\%$) covering 94.47% of

the genome. This large number is expected since MACS2 is a univariate peak caller and not designed for differential analysis. Furthermore, this dataset represents a differential analysis not between two cell types, but between five different cell types and thus boundary effects (false positives, e.g. falsely detected differences) are extremely likely. This interpretation is supported by the expression data, which could not find clear expression differences for the most frequent differentially modified regions (Fig. S9c). Also the GO analysis could not identify any significant GO terms. ChromHMM found 38288 different state transitions of which the first 656 cover 91.21% of the genome. This large number of transitions is dependent on the number of states that are used to train ChromHMM, since extra states will artificially inflate the number of chromatin state transitions. However, consistent with the chromstaR predictions, ChromHMM predicts many stage-specific enhancer (state E15 and E16) and constant promoter (state E9) regions among the most frequent transitions. The expression profiles associated with those transitions show the expected behaviour (Fig. S13b).

Limitations and Solutions

The number of possible combinatorial states for N ChIP-seq experiments is 2^N, meaning that for each additional ChIP-seq experiment the number of combinatorial states doubles. Thus it soon becomes computationally prohibitive to consider all combinatorial states. We found that with current computational resources (Intel Xeon E5 2680v3, 24 cores @ 2.5 GHz, 128GB memory) a practical limit seems to be 256 states (= 8 experiments) with a run-time of several days for a mouse genome ($\approx 2.6 \cdot 10^9$bp) and a bin size of 1000bp ($\approx$ 2.6M datapoints). We investigated several possibilities to extend the usability of chromstaR beyond this limit: (1) Calculations can be performed for each chromosome separately, and chromstaR features an option to perform this calculation in parallel. (2) For the case of one cell type or tissue where the number of measured histone modifications N exceeds the upper limit, chromstaR provides a strategy to artificially restrict the number of combinatorial states to any number lower than 2^N. This strategy can yield proper results if the correct states are included, since our results have shown that the majority of combinatorial states are absent in the genome. In order to identify the states which are the most present in the genome, chromstaR ranks the combinatorial states based

on their presence according to the combination of univariate results from the first step of the chromstaR pipeline. This ranking is a good approximation of the true multivariate state-distribution (Fig. S14). (3) If there are multiple marks N in multiple tissues M, and 2^{N*M} is bigger than the maximum number of states that the algorithm can handle computationally, two strategies are possible: One can either perform a differential analysis for each mark and then reconstruct combinatorial states in a classical way (Fig. S10a) or one can perform a multivariate peak-calling of combinatorial states for each tissue and then obtain the differences by a simple comparison between tissues (Fig. S10b). Both strategies give a different perspective on the data: The former accurately identifies differences between marks, while the combinatorial states might be subject to boundary effects. The latter gives an accurate picture of the combinatorial chromatin landscape, while differences between cells might be overestimated. (4) The run-time of our algorithm scales linearly with the number of data points, and thus a strategy is to decrease the resolution, e.g. halfing the run-time by doubling the bin size.

Discussion

Understanding how various histone modifications interact to determine cis-regulatory gene expression states is a fundamental problem in chromatin biology. It is becoming increasingly clear that certain combinatorial patterns of these modifications define discrete chromatin states along the genome. These chromatin states "encode" cell-specific transcriptional programs, and constitute funtional units that are subject to dynamic changes in response to developmental and environmental cues.

Many experimental studies have recognized this and collected ChIP-seq data for a number of histone modifications on the same or different tissue(s) as well as for several developmental time points. Integrative analyses of such datasets often present formidable bioinformatic challenges. Only a few computational methods exist that can analyze multiple histone marks simultaneously in one sample and cluster them into a finite number of chromatin states [16, 17, 18, 19, 20, 21, 22]. Interestingly, these methods often demand that the user specifies the number of chromatin states beforehand. We find this problematic because this number is often a desired output

of the analysis rather than an input. Indeed, the true number of distinct chromatin states in the genomes of various species is subject to debate. In *D. melanogaster* nine chromatin states have been reported [45], while in *A. thaliana* four main states were found [46]. In human, Ernst et al. found 51 states in human T-cells [47]. The Roadmap Consortium reported 15 to 18 states [10]. It remains unclear whether these differences reflect species divergence at the level of chromatin organization, or whether they are due to differences in the assessed chromatin marks and bioinformatic treatment of the data. Without a formal computational framework for defining chromatin states these two possibilities cannot be confidently distinguished.

While multivariate methods such as ChromHMM provide possible computational solutions to such questions, these methods employ probablistic chromatin state definitions that are not always readily interpretable. A probalistic interpretation means that different combinatorial histone modification patterns can be simultaneously part of different underlying chromatin states. However, it is not immediately obvious whether the underlying chromatin state are biologically distinct or if they are only statistical entities that are otherwise biologically redundant. Identifying such redundancies is not easy, because of a lack of rules to decide whether two or more chromatin states can or cannot be considered to be equivalent. Such decisions require extensive manual curation of the output, and often presuppose the kind of biological knowledge that one wishes to obtain from the data in the first place.

In contrast to this probabilistic state definition, chromstaR outputs discrete chromatin states that are defined on the basis of the presence/absence of various histone modifications. That is, with N histone modifications, it infers all 2^N combinatorial chromatin states (Fig. 1a). This interpretation makes it easy to relate the inferred chromatin states back to the underlying histone modification patterns and thus fashions a direct mechanistic link between chromatin structure and function. Moreover, chromstaR's discrete state definition also provides an unbiased picture of the genome-wide frequency of various chromatin states and allows for easy genome-wide summary statistics. For instance, in our analysis of four histone modifications in mouse embryonic stem cells we found that only 7 of the 16 possible states covered almost 100% of the genome, and for the human Hippocampus with seven modifi-

cations only 21 of the 128 possible combinatorial states already covered 99% of the genome. Even more extreme, when analyzing 26 marks in a lung fibroblast cell line, we found that only 0.02% of all possible combinatorial states explain 95% of the genome. This striking sparsity in the combinatorial code is interesting and points at certain biochemical contraints that determine which histone modifications can or cannot co-occur at a genomic locus. Clearly, the genome-wide frequency of inferred combinatorial chromatin states depends on the number and the type of different histone modifications that are used in the analysis. Future studies should systematically investigate the dependency of the number of chromatin states on factors such as number and type of measured histone marks, resolution, organism etc.

By treating discrete combinatorial chromatin states as units of analysis chromstaR can also easily track chromatin state dynamics across cell types or developmental time points. In that respect chromstaR is unique as no other methdods exist to date that can peform a similar task. To illustrate this we have analyzed four different histone modification in 5 different cell types that are part of the mouse T-cell differentiation pathway. Of the 1048576 combinatorial state transitions, we find that only 99 comprise over 99.60% of the genome. Again, the sparsity in state transitions shows that a few key transitions define the developmental trajectory of T-cell differentiation. One notable transition is the gain or loss of state [me1] near promoters. We note that this state means that only H3K4me1 is present at a locus and no other marks. This is not the same as tracking H3K4me1 modification by itself as this latter mark can appear in a number of different, and often funtionally distinct, chromatin states such as [me1+ac], [me1/2+ac], [me1/2/3]. Hence, focusing on H3K4me1 alone would tag other chromatin state changes that may not be fully informative about T-cell differentiation.

Conclusions

chromstaR is a computational algorithm that can identify discrete chromatin states from multiple ChIP-seq experiments and detect combinatorial state differences betweeen cell-types and/or developmental time points. By defining chromatin states in terms of the presence and absence of combinatorial histone modification patterns,

it provides an intuitive way to understand genome regulation in terms of chromatin composition at a locus. chromstaR can be used for the annotation of reference epigenomes as well as for annotation of chromatin state transitions in well-described developmental systems. The algorithm is written in C++ and runs in the popular R computing environment. It therefore combines computational speed with the extensive bioinformatic toolsets available through Bioconductor [48, 49]. chromstaR is freely available at `http://bioconductor.org/packages/chromstaR/` and features a collection of downstream analysis functions.

Methods

Model Specification

The construction of the multivariate Hidden Markov Model can be divided in two steps (Fig. 2). In the first step, we fit a univariate Hidden Markov Model to each individual ChIP-seq sample. The obtained parameters of the mixture distributions are then used in the second step to construct the multivariate emission distributions. Finally, the multivariate Hidden Markov Model is fitted to the (combined) ChIP-seq samples. The following sections describe the two steps in detail.

Univariate Hidden Markov Model

For each individual ChIP-seq sample, we partition the genome into T non-overlapping, equally sized bins. We count the number of aligned reads (regardless of strand) that overlap any given bin t and denote this read count with x_t. Following others [29, 30], we model the distribution of the read counts x with a two-component mixture of (zero-inflated) negative binomial distributions. In our case, the first component describes the *unmodified* regions and is modeled by a zero-inflated negative binomial distribution. The second component describes the *modified* regions and is modeled by a negative binomial distribution. Furthermore, for computational efficiency, we split the first component into the zero-inflation and the negative binomial distribution [31]. Our univariate Hidden Markov Model has thus three states i: *zero-inflation*, *unmodified* and *modified*. We write the probability of observing a given read count as

$$P(x_t|\theta) = \gamma_1 f_1(x_t|\theta_1) + \gamma_2 f_2(x_t|\theta_2) + \gamma_3 f_3(x_t|\theta_3) \quad (1)$$

where γ_i are the mixing weights and θ_i are the component density parameters. The emission distribution of state 1 is defined as

$$f_1(x_t) = \begin{cases} 1 & \text{if } x_t = 0 \\ 0 & \text{if } x_t > 0 \end{cases} \tag{2}$$

and the emission distributions of state 2 and 3 are defined as

$$f(x_t|\theta = (n,p)) = \frac{\Gamma(n+x_t)}{\Gamma(n)x_t!}p^n(1-p)^{x_t} \tag{3}$$

where Γ denotes the Gamma function and p and n denote the probability and dispersion parameter of the negative binomial distribution, respectively.

We use the Baum-Welch algorithm [50] to obtain the best fit for the distribution parameter estimates, transition probabilities and posterior probabilities of being in a given state. We call a bin *modified* if the posterior probability of being in that state is > 0.5 and *unmodified* otherwise.

Multivariate Hidden Markov Model

Given N individual ChIP-sep samples with states *unmodified* and *modified*, the number of possible combinatorial states is 2^N. Let $\mathbf{x}_t$ be the vector of N read counts for the t-th bin. The probability of observing a random vector $\mathbf{x_t}$ can be written as a mixture distribution of 2^N components:

$$P(\mathbf{x}_t|\theta) = \sum_{i=1}^{2^N} \gamma_i f_i(\mathbf{x}_t, \theta_i) \tag{4}$$

Again, the γ_i denote the mixing weights and θ_i denote the component density parameters for each component i. We assume that the marginal densities of the multivariate count distributions f_i are given by the univariate distributions described in the previous section. A convenient way to construct a multivariate distribution from known marginal (univariate) distributions is copula theory [32, 51].

Under the assumption of a Gaussian copula, the multivariate emission density for combinatorial state i can be written as

$$f_i(\mathbf{x}_t) = \prod_{j=1}^{N} f_{i,j}(x_{j,t}) \times |\mathbf{\Sigma}_i|^{-1/2} \exp\left\{ -\frac{\mathbf{z}_{i,t}\ (\mathbf{\Sigma}_i^{-1} - \mathbf{I})\ \mathbf{z}_{i,t}^T}{2} \right\} , \tag{5}$$

$$\text{with} \quad \mathbf{z}_{i,t} = [\ \phi^{-1}(F_{i,1}(x_{1,t}))\ ,\ \phi^{-1}(F_{i,2}(x_{2,t}))\ ,\ \ldots\ ,\ \phi^{-1}(F_{i,N}(x_{N,t}))\ ]\ , \tag{6}$$

where $f_{i,j}$ are the marginal density functions for combinatorial state i and $\mathbf{\Sigma}_i$ is the correlation matrix between the transformed read counts $z_{i,t} = \phi^{-1}(F_i(x_t))$. The cumulative distribution function (CDF) of $f_{i,j}$ is denoted by $F_{i,j}$, while ϕ^{-1} denotes the inverse of the CDF of a standard normal [52].

The correlation matrix $\mathbf{\Sigma}_i$ for a given multivariate (combinatorial) state i is computed as follows: From the combination of univariate state calls (*unmodified* or *modified*) of all samples, we pick those bins that show combinatorial state i. The read counts $\mathbf{x}_{t\in i}$ in those bins are transformed to $\mathbf{z}_{t\in i}$ using equation 6 and the correlation matrix $\mathbf{\Sigma}_i$ is calculated from the transformed read counts.

Similarly to the univariate Hidden Markov Model, we use the Baum-Welch algorithm to obtain the best fit for the transition probabilities and posterior probabilities of being in a given state. However, the emission densities remain fixed in the multivariate case. We assign a combinatorial state to each bin by maximizing over the posterior probabilities. We found it useful to transform posterior probabilities for each combinatorial state ("posteriors-per-state") into posterior probabilites of peak calls for each experiment ("posteriors-per-mark"), such that a cutoff can be applied instead of maximizing over the posteriors. We found that out algorithm had a very high sensitivity for detecting peaks when maximizing over the "posteriors-per-state". To increase specificity, a strict cutoff (e.g. 0.99) can be applied on the "posteriors-per-mark".

Integration of chromatin input experiments

Chromatin input experiments serve as controls for bias in chromatin fragmentation and variations in sequencing efficiency [53]. We optionally integrate this information

by modifying the vector of read counts that serves as the observable in the Hidden Markov Model. Let $\mathbf{x}$ be the vector of read counts along the genome for the ChIP-seq experiment, and $\mathbf{y}$ be the vector of read counts for the input experiment. Let furthermore $\mathbf{y}_P$ be the vector $\mathbf{y}$ without zero read counts. In a first step, we null regions with artificially high read counts, e.g. repetitive regions around centromers, by setting $\mathbf{x}_t = 0$ for all bins t where $\mathbf{y}_t >= c_0$. c_0 is defined as the 99.9% quantile of $\mathbf{y}_P$. In a second step, we calculate a corrected read count $\mathbf{x}'$ as

$$\mathbf{x}' = \mathbf{x} \cdot min\left(\frac{mean(\mathbf{y}_P)}{runmean(\mathbf{y})}, 1.5\right) \tag{7}$$

where $runmean()$ calculates a running mean of 15 bins. This operation modifies the read count $\mathbf{x}$ in such a way that $\mathbf{x}'$ is decreased in bins which have more than average counts in the input and increased in bins that have less than average counts in the input.

Inclusion of replicates

The chromstaR formalism offers an elegant way to include replicates. For a single ChIP-seq experiment, there are two states - unmodified (background) and modified (peaks). For an arbitrary number of N experiments, there are thus 2^N combinatorial states. The same is true for an arbitrary number of replicates R, which would yield 2^R combinatorial states. However, in the case of replicates, the number of states can be fixed to 2, such that all replicates are forced to have the same state (e.g. either peak or background). Treating replicates in this way allows to find the most likely state for each position considering information from all replicates without prior merging.

Univariate approximation of multivariate state distribution

chromstaR offers the possibility to restrict the number of combinatorial states to any number lower than 2^N, where N is the number of ChIP-seq experiments. Because the first step of the chromstaR workflow is a univariate peak calling, we can combine those peak calls into combinatorial states and use their ranking to determine which states to use for the multivariate peak-calling. Because most systems seem to be sparse in their combinatorial patterns, i.e. do not utilize the full com-

binatorial state space, it is often not necessary to run the multivariate part with all 2^N combinations. For instance, for the human Hippocampus tissue with seven marks, running the multivariate with only 30 instead of 128 states recovers 98.2% of correct state assignments compared to the full 128 state model, and choosing 60 instead of 128 states recovers already 99.5% of correct state assignments compared to the full 128 state model (Fig. S14).

Data Acquisition

ChIP-seq data for the hematopoietic system (GSE60103) was downloaded from the Gene Expression Omnibus (GEO) and aligned to mouse reference mm9 following the procedure in [37] with bowtie2 (version 2.2.3) [54], keeping only reads that mapped to a unique location. The number of identical reads at each genomic position was restricted to 3. For the expression analysis, we used the provided RNA-seq data (GSE60101). We normalized the read counts by transcript length and scaled them to 1M reads. To reduce the effect of extreme expression values, we applied an arcsinh transformation on the data.

For the Hippocampus dataset, bed-files with aligned reads were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/Brain_Hippocampus_Middle/ for donors number 112 and 149 and seven histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3 and H3K9ac.

Bed-files with aligned reads for the IMR90 dataset were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/IMR90_Cell_Line for 26 histone marks (H2AK5ac, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me3, H3K36me3, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K56ac, H3K79me1, H3K79me2, H3K9ac, H3K9me1, H3K9me3, H4K20me1, H4K5ac, H4K8ac, H4K91ac).

Enrichment profiles around TSS

We calculated sensitivity (recall), precision and F1-score for the detection of expressed TSS based on the following assumptions: True positives are expressed TSS which are called into the promoter state ([me1/2/3+ac] for chromstaR, E7 and E9 for ChromHMM, [me1/3] and [me3] for MACS2, see Fig. 4). False negatives are expressed TSS which are not assigned into the promoter state. True negatives

are non-expressed TSS which are not assigned into the promoter state. False positives are non-expressed TSS which are assigned the promoter state. We found that chromstaR has a higher sensitivity than the other methods and a lower precision. The F1-score is highest for chromstaR (Table 2).

Availability of data and materials

ChIP-seq data for the hematopoietic system (GSE60103) was downloaded from the Gene Expression Omnibus. Bed-files for Hippocampus tissue were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/ for donors number 112 and 149. Bed-files for IMR90 cell line were downloaded from ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/IMR90_Cell_Line. *chromstaR* if available under the Artistic-2.0 license and can be downloaded from http://bioconductor.org/packages/chromstaR/.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

MCT and FJ designed the research. AT, MCT, MAN and MH developed the algorithm. AT analyzed the data. AT, MCT and FJ wrote the manuscript.

Acknowledgements

We thank G. de Haan and JJ. Schuringa for their comments. MCT acknowledges support from the Netherlands Organization for Scientific Research and from a University of Groningen Rosalind Franklin Fellowship. FJ was supported by the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement #291763. MH acknowledges support by the Federal Ministry of Education and Research (BMBF, Germany) in the project eMed:symAtrial (01ZX1408D).

Author details

[1]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, A. Deusinglaan 1, NL-9713 AV Groningen, The Netherlands. [2]Department of Plant Sciences, Hans Eisenmann-Zentrum for Agricultural Sciences, Technical University Munich, Liesel-Beckmann-Str. 2, 85354 Freising, Germany. [3]Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany. [4]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

References

1. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., Bernstein, B.E.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448**(7153), 553–60 (2007). doi:10.1038/nature06008
2. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K.: High-resolution profiling of histone methylations in the human genome. Cell **129**(4), 823–37 (2007). doi:10.1016/j.cell.2007.05.009
3. Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D., Dunham, I.: The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome research **17**(6), 691–707 (2007). doi:10.1101/gr.5704207
4. Huda, A., Mariño-Ramírez, L., Jordan, I.K.: Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mobile DNA **1**(1), 2 (2010). doi:10.1186/1759-8753-1-2

5. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A., Marra, M., Snyder, M., Jones, S.: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods **4**(8), 651–7 (2007). doi:10.1038/nmeth1068
6. Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D.K., Young, R.A.: Genome-wide map of nucleosome acetylation and methylation in yeast. Cell **122**(4), 517–27 (2005). doi:10.1016/j.cell.2005.06.026
7. Rintisch, C., Heinig, M., Bauerfeind, A., Schafer, S., Mieth, C., Patone, G., Hummel, O., Chen, W., Cook, S., Cuppen, E., Colomé-Tatché, M., Johannes, F., Jansen, R.C., Neil, H., Werner, M., Pravenec, M., Vingron, M., Hubner, N.: Natural variation of histone modification and its impact on gene expression in the rat genome. Genome research **24**(6), 942–53 (2014). doi:10.1101/gr.169029.113
8. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M.: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012). doi:10.1038/nature11247
9. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A.: The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology **28**(10), 1045–8 (2010). doi:10.1038/nbt1010-1045
10. Consortium, R.E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.-C., Pfenning, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.a., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.a., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.a., McManus, M.T., Sunyaev, S., Thomson, J.a., Tlsty, T.D., Tsai, L.-H., Wang, W., Waterland, R.a., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.a., Wang, T., Kellis, M., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.-C., Pfenning, A., Wang, X., ClaussnitzerYaping Liu, M., Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Scott Hansen, R., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.a., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Abdennur, N., Adli, M., Akerman, M., Barrera, L., Antosiewicz-Bourget, J., Ballinger, T., Barnes, M.J., Bates, D., Bell, R.J.a., Bennett, D.a., Bianco, K., Bock, C., Boyle, P., Brinchmann, J., Caballero-Campo, P., Camahort, R., Carrasco-Alfonso, M.J., Charnecki, T., Chen, H., Chen, Z., Cheng, J.B., Cho, S., Chu, A., Chung, W.-Y., Cowan, C., Athena Deng, Q., Deshpande, V., Diegel, M., Ding, B., Durham, T., Echipare, L., Edsall, L., Flowers, D., Genbacev-Krtolica, O., Gifford, C., Gillespie, S., Giste, E., Glass, I.a., Gnirke, A., Gormley, M., Gu, H., Gu, J., Hafler, D.a., Hangauer, M.J., Hariharan, M., Hatan, M., Haugen, E., He, Y., Heimfeld, S., Herlofsen, S., Hou, Z., Humbert, R., Issner, R., Jackson, A.R., Jia, H., Jiang, P., Johnson, A.K., Kadlecek, T., Kamoh, B., Kapidzic, M., Kent, J., Kim, A., Kleinewietfeld, M., Klugman, S., Krishnan, J., Kuan, S., Kutyavin, T., Lee, A.-Y., Lee, K., Li, J., Li, N., Li, Y., Ligon, K.L., Lin, S., Lin, Y., Liu, J., Liu, Y., Luckey, C.J., Ma, Y.P., Maire, C., Marson, A., Mattick, J.S., Mayo, M., McMaster, M., Metsky, H., Mikkelsen, T., Miller, D., Miri, M., Mukame, E., Nagarajan, R.P., Neri, F., Nery, J., Nguyen, T., O'Geen, H., Paithankar, S., Papayannopoulou, T., Pelizzola, M., Plettner, P., Propson, N.E., Raghuraman, S., Raney, B.J., Raubitschek, A., Reynolds, A.P., Richards, H., Riehle, K., Rinaudo, P., Robinson, J.F., Rockweiler, N.B., Rosen, E., Rynes, E., Schein, J., Sears, R., Sejnowski, T., Shafer, A., Shen, L., Shoemaker, R., Sigaroudinia,

M., Slukvin, I., Stehling-Sun, S., Stewart, R., Subramanian, S.L., Suknuntha, K., Swanson, S., Tian, S., Tilden, H., Tsai, L., Urich, M., Vaughn, I., Vierstra, J., Vong, S., Wagner, U., Wang, H., Wang, T., Wang, Y., Weiss, A., Whitton, H., Wildberg, A., Witt, H., Won, K.-J., Xie, M., Xing, X., Xu, I., Xuan, Z., Ye, Z., Yen, C.-a., Yu, P., Zhang, X., Zhang, X., Zhao, J., Zhou, Y., Zhu, J., Zhu, Y., Ziegler, S., Beaudet, A.E., Boyer, L.a., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.a., McManus, M.T., Sunyaev, S., Thomson, J.a., Tlsty, T.D., Tsai, L.-H., Wang, W., Waterland, R.a., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.a., Wang, T., Kellis, M.: Integrative analysis of 111 reference human epigenomes. Nature **518**, 317–330 (2015). doi:10.1038/nature14248

11. Strahl, B.D., Allis, C.D.: The language of covalent histone modifications. Nature **403**(6765), 41–5 (2000). doi:10.1038/47412
12. Jenuwein, T., Allis, C.D.: Translating the histone code. Science (New York, N.Y.) **293**(5532), 1074–80 (2001). doi:10.1126/science.1063127
13. Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.a., Birney, E., Hardison, R.C., Dunham, I., Kellis, M., Noble, W.S.: Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research **41**(2), 827–41 (2013). doi:10.1093/nar/gks1284
14. Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P., Lam, E.: Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. The Plant journal : for cell and molecular biology **73**(1), 77–90 (2013). doi:10.1111/tpj.12017
15. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., Zhao, K.: Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics **40**(7), 897–903 (2008). doi:10.1038/ng.154
16. Ernst, J., Kellis, M.: ChromHMM: automating chromatin-state discovery and characterization. Nature methods **9**(3), 215–216 (2012). doi:10.1038/nmeth.1906
17. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S.: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods **9**(5), 473–6 (2012). doi:10.1038/nmeth.1937
18. Biesinger, J., Wang, Y., Xie, X.: Discovering and mapping chromatin states using a tree hidden Markov model. BMC bioinformatics **14 Suppl 5**, 4 (2013). doi:10.1186/1471-2105-14-S5-S4
19. Zeng, X., Sanalkumar, R., Bresnick, E.H., Li, H., Chang, Q., Keleş, S.: jMOSAiCS: joint analysis of multiple ChIP-seq datasets. Genome biology **14**(4), 38 (2013). doi:10.1186/gb-2013-14-4-r38
20. Sohn, K.-A., Ho, J.W.K., Djordjevic, D., Jeong, H.-H., Park, P.J., Kim, J.H.: hiHMM: Bayesian non-parametric joint inference of chromatin state maps. Bioinformatics (Oxford, England), 117 (2015). doi:10.1093/bioinformatics/btv117
21. Song, J., Chen, K.C.: Spectacle: fast chromatin state annotation using spectral learning. Genome biology **16**(1), 33 (2015). doi:10.1186/s13059-015-0598-0
22. Mammana, A., Chung, H.-R.: Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. Genome biology **16**(1), 151 (2015). doi:10.1186/s13059-015-0708-z
23. Taudt, A., Colomé-Tatché, M., Johannes, F.: Genetic sources of population epigenomic variation. Nature Reviews Genetics **advance on** (2016). doi:10.1038/nrg.2016.45
24. Hon, G., Ren, B., Wang, W.: ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology **4**(10), 1000201 (2008). doi:10.1371/journal.pcbi.1000201
25. Won, K.-J., Zhang, X., Wang, T., Ding, B., Raha, D., Snyder, M., Ren, B., Wang, W.: Comparative annotation of functional regions in the human genome using epigenomic data. Nucleic acids research **41**(8), 4423–32 (2013). doi:10.1093/nar/gkt143
26. Yen, A., Kellis, M.: Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. Nature Communications **6**, 7973 (2015). doi:10.1038/ncomms8973
27. Ji, H., Li, X., Wang, Q.-f., Ning, Y.: Differential principal component analysis of ChIP-seq. Proceedings of the National Academy of Sciences **110**(17), 6789–6794 (2013). doi:10.1073/pnas.1204398110
28. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J.:

Software for computing and annotating genomic ranges. PLoS computational biology **9**(8), 1003118 (2013). doi:10.1371/journal.pcbi.1003118

29. Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., Lieb, J.D.: ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome biology **12**(7), 67 (2011). doi:10.1186/gb-2011-12-7-r67
30. Spyrou, C., Stark, R., Lynch, A.G., Tavaré, S.: BayesPeak: Bayesian analysis of ChIP-seq data. BMC bioinformatics **10**, 299 (2009). doi:10.1186/1471-2105-10-299
31. van der Graaf, A., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M., Johannes, F.: Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proceedings of the National Academy of Sciences of the United States of America **112**(21), 6676–81 (2015). doi:10.1073/pnas.1424254112
32. Sklar, M.: Fonctions de Répartition À N Dimensions et Leurs marges, (1959). http://books.google.nl/books/about/Fonctions_de_R%C3%A9partition_%C3%80_N_Dimension.html?id=nreSmAEACAAJ&pgis=1 http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Fonctions+de+Répartition+À+N+Dimensions+Et+Leurs+Marges#0
33. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based analysis of ChIP-Seq (MACS). Genome biology **9**(9), 137 (2008). doi:10.1186/gb-2008-9-9-r137
34. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., Peng, W.: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics (Oxford, England) **25**(15), 1952–8 (2009). doi:10.1093/bioinformatics/btp340
35. Micsinai, M., Parisi, F., Strino, F., Asp, P., Dynlacht, B.D., Kluger, Y.: Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. Nucleic acids research **40**(9), 70 (2012). doi:10.1093/nar/gks048
36. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Swets, J., Pepe, M., Sonego, P., Kocsor, A., Pongor, S., Fawcett, T., Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., Dougherty, E., Robin, X., Turck, N., Hainard, A., Lisacek, F., Sanchez, J., Müller, M., McClish, D., Jiang, Y., Metz, C., Nishikawa, R., Stephan, C., Wesseling, S., Schink, T., Jung, K., Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., Pepe, M., Longton, G., Janes, H., Hanley, J., McNeil, B., DeLong, E., DeLong, D., Clarke-Pearson, D., Bandos, A., Rockette, H., Gur, D., Braun, T., Alonzo, T., Venkatraman, E., Begg, C., Bandos, A., Rockette, H., Gur, D., Moise, A., Clement, B., Raissis, M., Venkatraman, E., Campbell, G., Carpenter, J., Bithell, J., Metz, C., Herman, B., Shen, J., Hanley, J., Zou, K., Hall, W., Shapiro, D., Venables, W., Ripley, B., Turck, N., Vutskits, L., Sanchez-Pena, P., Robin, X., Hainard, A., Gex-Fabry, M., Fouda, C., Bassem, H., Mueller, M., Lisacek, F., Ewens, W., Grant, G.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **12**(1), 77 (2011). doi:10.1186/1471-2105-12-77
37. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N., Amit, I.: Chromatin state dynamics during blood formation. Science (New York, N.Y.) **55**(233348), 1–10 (2014). doi:10.1126/science.1256271
38. Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., Gingeras, T.R., Schreiber, S.L., Lander, E.S.: Genomic maps and comparative analysis of histone modifications in human and mouse. Cell **120**(2), 169–81 (2005). doi:10.1016/j.cell.2005.01.001
39. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antosiewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M., Ren, B.: Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature **459**(7243), 108–12 (2009). doi:10.1038/nature07829
40. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.-a.: Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature **518**, 350–354 (2015). doi:10.1038/nature14217
41. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, a.M., Baillie, J.K.,

Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.a., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A.: An atlas of active enhancers across human cell types and tissues. Nature **507**(7493), 455–61 (2014). doi:10.1038/nature12787

42. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V.V., Ecker, J.R., Thomson, J.a., Ren, B.: Chromatin architecture reorganization during stem cell differentiation. Nature **518**(7539), 331–336 (2015). doi:10.1038/nature14222
43. Amin, V., Harris, R.A., Onuchic, V., Jackson, A.R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C., Milosavljevic, A.: Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. Nature Communications **6**(May 2014), 6370 (2015). doi:10.1038/ncomms7370
44. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G.: GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology **28**(5), 495–501 (2010). doi:10.1038/nbt.1630
45. Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M.Y., Luquette, L.J., Xi, R., Jung, Y.L., Park, R.W., Bishop, E.P., Canfield, T.K., Sandstrom, R., Thurman, R.E., MacAlpine, D.M., Stamatoyannopoulos, J.A., Kellis, M., Elgin, S.C.R., Kuroda, M.I., Pirrotta, V., Karpen, G.H., Park, P.J.: Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature **471**(7339), 480–5 (2011). doi:10.1038/nature09725
46. Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M., Colot, V.: Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. The EMBO Journal **30**(10), 1928–1938 (2011). doi:10.1038/emboj.2011.103
47. Ernst, J., Kellis, M.: Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology **28**(8), 817–25 (2010). doi:10.1038/nbt.1662
48. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome biology **5**(10), 80 (2004). doi:10.1186/gb-2004-5-10-r80
49. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.K., Pagès, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods **12**(2), 115–121 (2015). doi:10.1038/nmeth.3252
50. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics **41**(1), 164–171 (1970)
51. Heinig, M., Colomé-Tatché, M., Taudt, A., Rintisch, C., Schafer, S., Pravenec, M., Hubner, N., Vingron, M., Johannes, F.: histoneHMM: Differential analysis of histone modifications with broad genomic footprints. BMC Bioinformatics **16**(1), 60 (2015). doi:10.1186/s12859-015-0491-6
52. Renard, B., Lang, M.: Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. Advances in Water Resources **30**(4), 897–912 (2007)
53. Kidder, B.L., Hu, G., Zhao, K.: ChIP-Seq: technical considerations for obtaining high-quality data. Nature immunology **12**(10), 918–22 (2011). doi:10.1038/ni.2117
54. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4), 357–9 (2012). doi:10.1038/nmeth.1923
55. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W.: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics

(Oxford, England) **21**(16), 3439–40 (2005). doi:10.1093/bioinformatics/bti525

56. Durinck, S., Spellman, P.T., Birney, E., Huber, W.: Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols **4**(8), 1184–91 (2009). doi:10.1038/nprot.2009.97

Figures

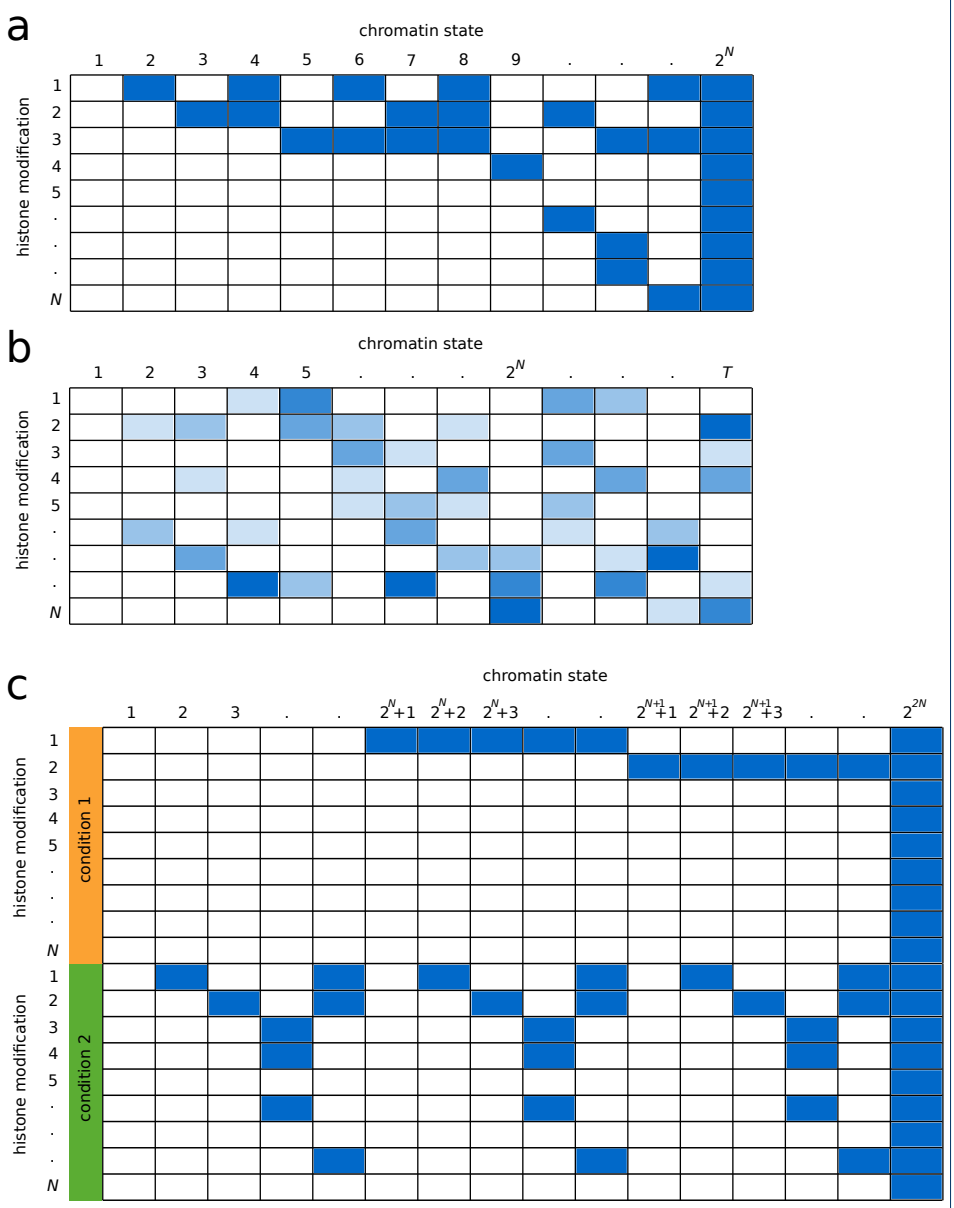

Figure 1 Definition of chromatin states. (a) Combinatorial chromatin state definition: Based on the presence (blue) or absence (white) of a histone modification, a chromatin state is the combination of the presence/absence calls at a given position. With N histone modifications there are 2^N different chromatin states. **(b)** Probabilistic chromatin state definition: Each chromatin state has a probability (shades of blue) of finding a histone modification at a given position. Note that a probabilistic state can consist of multiple combinatorial states and vice versa. There is in principle no upper limit for the number of possible probabilistic chromatin states (here, T). **(c)** Differential combinatorial chromatin states across two conditions: Based on the presence (blue) or absence (white) of a histone modification across different conditions. With N histone modifications and M conditions there are $2^{N \times M}$ different states.

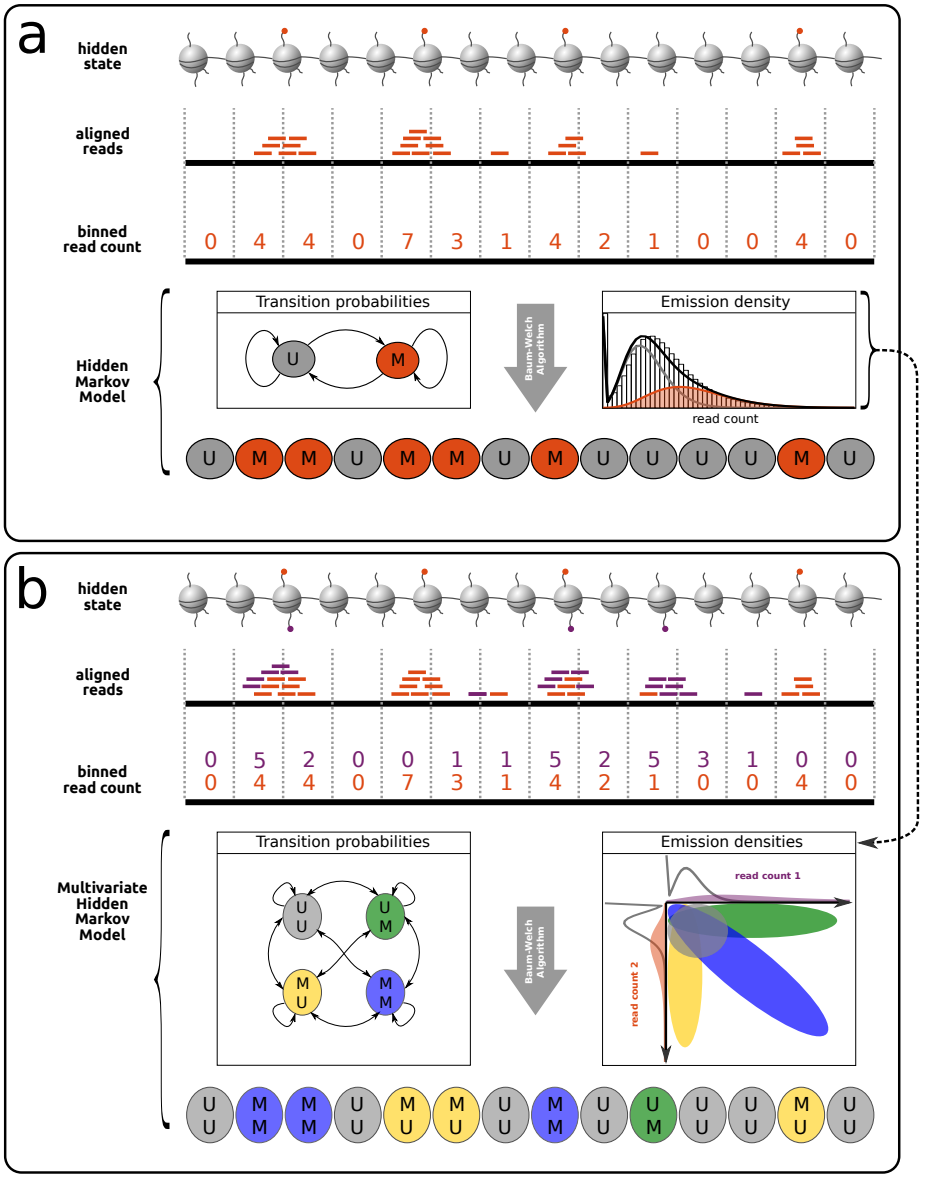

Figure 2 Overview of analytical approach. (a) Univariate peak calling: Aligned reads are counted in equidistant, non-overlapping bins. The resulting read count serves as observable for a Hidden Markov Model with components "unmodified" (U) and "modified" (M). Parameters are fitted with the Baum-Welch algorithm. **(b)** Multivariate peak calling: A multivariate emission density is constructed from the densities of step (a), illustrated here for two dimensions. With N ChIP-seq experiments, the resulting multivariate density has 2^N components. For the two illustrated dimensions (ChIP-seq 1 and 2), univariate densities are plotted on the x- and y-axis, and the 4 components of the multivariate density are indicated by shaded areas, corresponding to both unmodified (UU - gray), both modified (MM - blue) and unmodified/modified (MU - yellow, UM - green).

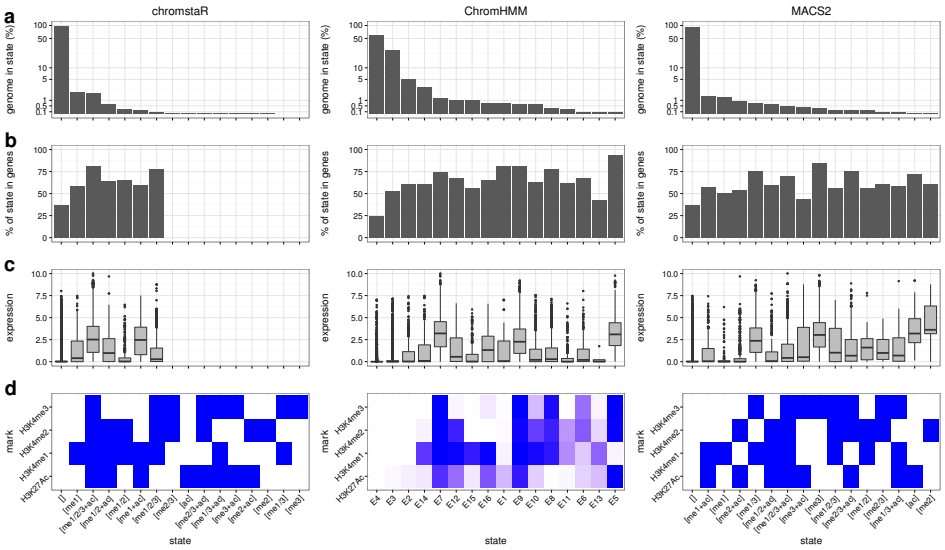

Figure 3 Chromatin states in monocytes. (a) Genomic frequency, i.e. the percentage of the genome that is covered by the chromatin state. The sum over all states equals 100%. **(b)** Overlap with known genes. **(c)** Expression levels of genes whose transcription start site (TSS) overlaps the chromatin state. **(d)** Heatmap showing the chromatin state definition. Histones in chromstaR and MACS2 states are either present (blue) or absent (white). ChromHMM states have a continuous emission probability from zero (white) to one (blue).

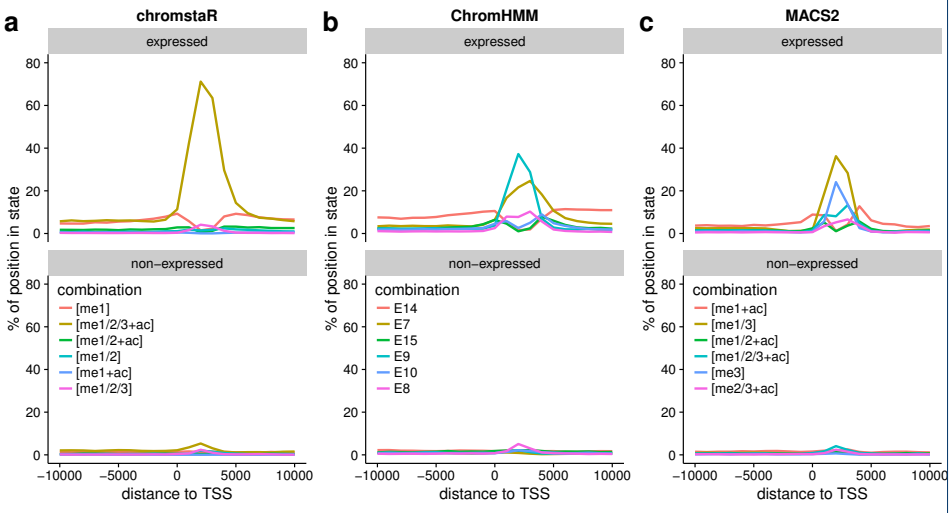

Figure 4 Fold enrichment of chromatin states around transcription start sites (TSS) of expressed (top) and non-expressed (bottom) genes. Shown are the enrichment profiles for the 6 states that are most enriched around TSS. chromstaR consistently assigns state [me1/2/3+ac] to expressed TSS and has a higher sensitivity for detecting those, while the other two methods assign two states with lower sensitivity.

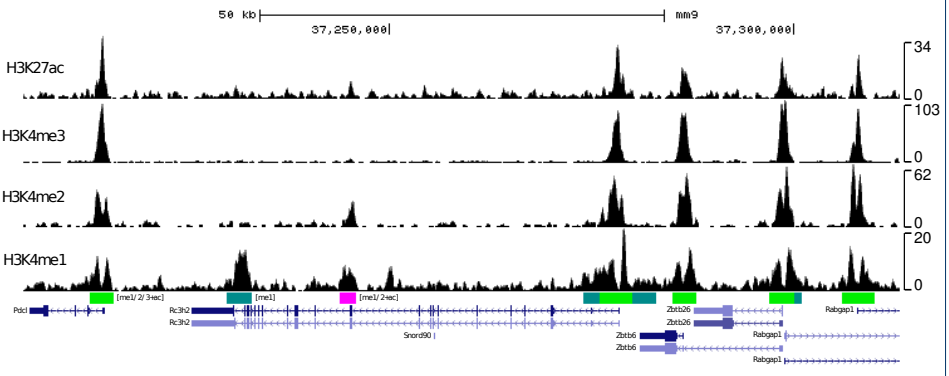

Figure 5 Genome browser snapshot showing a 100kb region on chromosome 2 with several active promoter signatures [me1/2/3+ac] (green) flanked by the [me1] signature (teal). The four black tracks show normalized signal coverage profiles for the four measured histone modifications. The colored track below shows combinatorial state calls, and the bottom track shows genes. Promoter signatures coincide with gene starts.

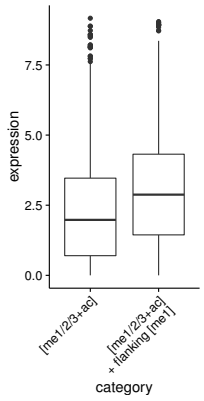

Figure 6 Expression levels of genes whose transcription start site (TSS) shows either the [me1/2/3+ac] signature alone or the [me1/2/3+ac] signature flanked by [me1]. TSS flanked by [me1] show significantly higher expression levels ($p \approx 10^{-101}$, Wilcoxon rank-sum test).

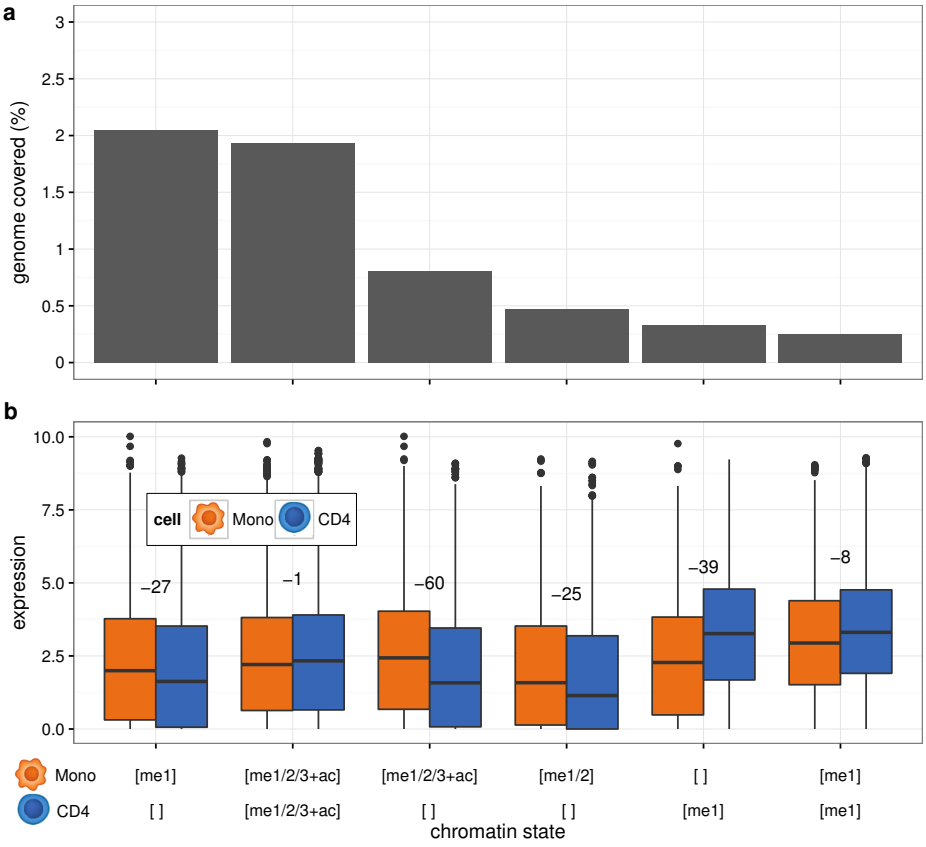

Figure 7 Differential analysis of monocytes and CD4 T-cells. (a) Genomic frequency of the six most frequent differential chromatin states. Differential chromatin states are shown on the x axes (top: combinatorial state in monocytes, bottom: combinatorial state in CD4). **(b)** Expression from monocytes and CD4 cells of genes which overlap the given differential chromatin state. Numbers give the base-10 logarithm of the multiple testing corrected p-value for the expression difference using a Wilcoxon rank-sum test. The more negative the number, the more significant the difference. Genes with differential chromatin signatures (e.g. [me1/2/3+ac] in monocytes but not in CD4 T-cells) show highly significant expression differences, whereas similar signatures ([me1/2/3+ac] in both cell types) are not significantly different.

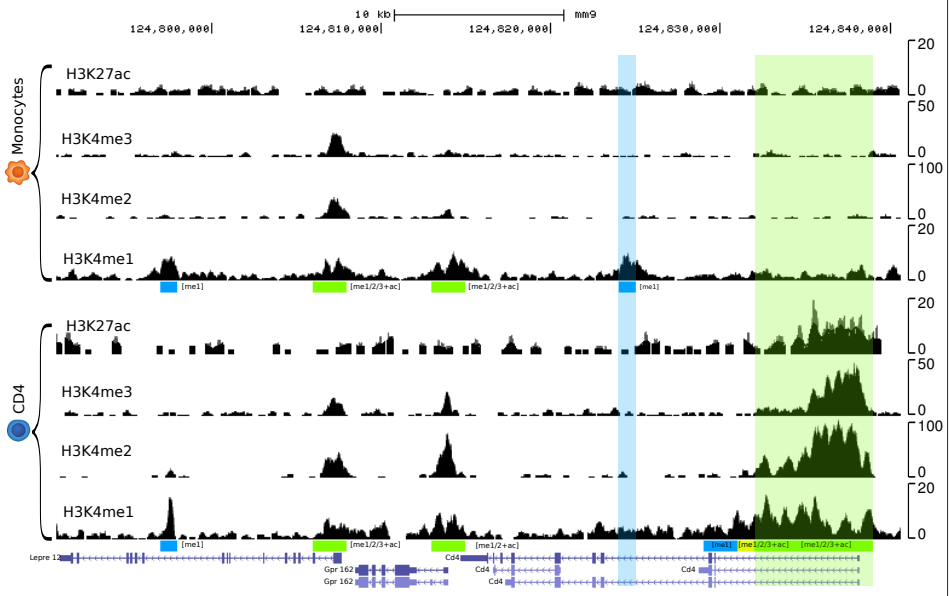

Figure 8 Differential chromatin signature at the Cd4 locus. Example of a differential promoter and enhancer signature at the Cd4 gene. The differential promoter signature [me1/2/3+ac] is only present in CD4 T-cells (shaded green), while the differential enhancer [me1] is present only in monocytes (shaded blue).

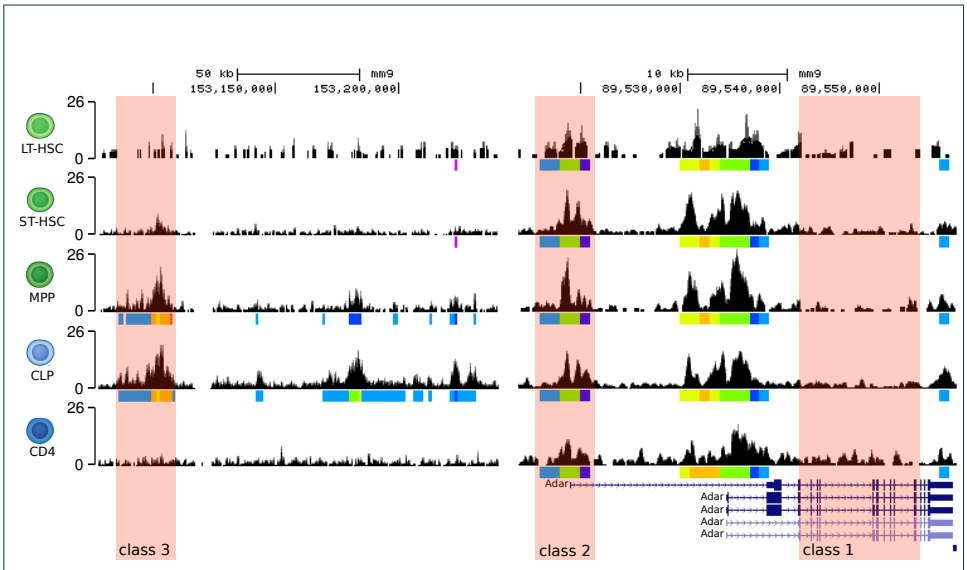

Figure 9 Examples of different classes of transitions (shaded in red): "Empty" (class 1), "constant" (class 2) and "stage-specific" transitions (class 3). Only the H3K4me1 tracks are shown. Combinatorial chromatin states as obtained by chromstaR are shown below the H3K4me1 tracks as colored bars.

Tables

	[me1/2/3+ac] + flanking [me1]		[me1/2/3+ac]	
1	posttranscriptional regulation of gene expression	5.75e-25	RNA processing	1.40e-36
2	regulation of translation	1.57e-21	ncRNA metabolic process	6.75e-34
3	peptidyl-lysine modification	1.69e-17	ncRNA processing	2.57e-29
4	microtubule nucleation	2.80e-13	DNA repair	1.14e-25
5	mRNA transport	1.23e-10	ribosome biogenesis	3.36e-17
6	RNA localization	3.61e-10	rRNA metabolic process	9.14e-17
7	RNA transport	7.01e-10	tRNA metabolic process	3.14e-16
8	GPI anchor biosynthetic process	1.01e-09	rRNA processing	1.85e-15
9	negative regulation of translation	1.16e-09	protein folding	3.74e-14
10	DNA replication	1.60e-09	tRNA processing	1.02e-11

Table 1 The first 10 significant gene ontology terms for TSS overlapping the [me1/2/3+ac] state with the [me1] state flanking it, versus the TSS overlapping the [me1/2/3+ac] state. Numbers indicate the binomial false discovery rate (BinomFdrQ) as reported by GREAT.

	Sensitivity	Precision	F1-score
chromstaR	0.71	0.97	0.82
Macs2	0.60	0.98	0.75
ChromHMM	0.59	0.98	0.73

Table 2 Performance for detecting expressed TSS.