chromstaR: Tracking combinatorial chromatin state dynamics in space and time

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

[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 Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[3]Institute for Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764, Neuherberg, Germany

[4]Department of Plant Sciences, Hans Eisenmann-Zentrum for Agricultural Sciences, Technical University Munich, Liesel-Beckmann-Str. 2, 85354, Freising, Germany

February 3, 2016

Abstract

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 compare genome-wide chromatin state maps across different conditions, such as experimental treatments, cell-types or developmental time points. 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 assign every genomic region to a discrete combinatorial chromatin state based on the presence/absence of each modification in every condition. This interpretation makes it easy to relate the inferred chromatin states back to the underlying histone modification patterns. Moreover, the algorithm computes the number of combinatorial chromatin states that are present in the genome without having to specify them *a priori*, thus providing an unbiased picture of their genome-wide frequencies. 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. chromstaR is a versatile computational tool that facilitates a deeper biological understanding of chromatin organization and dynamics. The algorithm is written in C++ and freely available as an R-package at `https://github.com/ataudt/chromstaR`.

*a.s.taudt@umcg.nl
†anhntm@gmail.com
‡matthias.heinig@helmholtz-muenchen.de
§frank@johanneslab.org
¶m.colome.tatche@umcg.nl

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 (Mikkelsen et al., 2007), maintenance of cellular identity (Barski et al., 2007; Koch et al., 2007) and silencing of transposable elements (Huda et al., 2010). 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 (Mikkelsen et al., 2007; Barski et al., 2007; Robertson et al., 2007), ChIP-seq technologies have been widely used to survey genome-wide patterns of histone modifications in a variety of organisms (Pokholok et al., 2005; Rintisch et al., 2014; Barski et al., 2007), cell lines (Bernstein et al., 2012) and tissues (Bernstein et al., 2010; Consortium et al., 2015).

The multitude of possible histone modifications has led to the idea of a "histone code" (Jenuwein and Allis, 2001), 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 (Hoffman et al., 2013) and Epigenomics Roadmap (Consortium et al., 2015). 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 (Luo et al., 2013; Wang

et al., 2008). 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). Several such methods have been proposed in recent years, all of which employ graphical probabilistic methods such as Hidden Markov Models (HMM) or dynamic Bayesian networks. ChromHMM (Ernst and Kellis, 2012) employs a multivariate HMM to classify the genome into a preselected number of chromatin states and was used to annotate the epigenome in the ENCODE (Hoffman et al., 2013) and Epigenomics Roadmap (Consortium et al., 2015) projects. Segway (Hoffman et al., 2012) is another tool based on dynamic Bayesian networks that classifies the genome into a preselected number of states. It requires, however, extensive computational resources and special cluster management, limiting its usability. TreeHMM (Biesinger et al., 2013) is an extension of ChromHMM which explicitly takes lineage information into account. Another tool, hiHMM (Sohn et al., 2015), was designed to share state definitions across different genomes. Finally, ChromDiff (Yen and Kellis, 2015) has been proposed for the group-wise comparison of chromatin states between two conditions.

A major drawback of these approaches is the need to specify the number of distinct chromatin states beforehand, which is usually not known *a priori*. Furthermore, the learned states are probabilistic, meaning that each state 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.

To adress some of these issues we 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. 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.

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 efficient computational solutions to these types of data problems, and facilitates deeper biological insights into the dynamic co-ordination of combinatorial chromatin states in genome regulation.

Results

Brief overview of analytical approach

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 can be summarized in three steps (Fig. 2): (1) 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) (Lawrence et al., 2013). (2) For every ChIP-seq experiment, we consider that the read count distribution is a two-component mixture of zero-inflated negative binomials (Rashid et al., 2011; Spyrou et al., 2009), 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 (van der Graaf et al., 2015). (3) 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 (2), are defined using a Gaussian copula (Sklar, 1959). A detailed description can be found in **Supplementary Materials**.

Application 1: Mapping combinatorial chromatin states in a reference tissue

Lara-Astiaso et al. (Lara-Astiaso et al., 2014) measured four histone modifications (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) and gene expression in 16 mouse hematopoietic cell lines and their progenitors (Fig. S1). 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. S2 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% (99.998%) 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. 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 93.10% of the genome. The high prevalence of this state reflects the fact that Lara-Astiaso et al. (Lara-Astiaso et al., 2014) focused on marks that had previously been shown to occur proximal to genic sequences (Bernstein et al., 2005; Barski et al., 2007; Koch et al., 2007). Indeed, only 36% of the empty state overlapped known genes while the remaining 64% mapped to non-genic regions throuhgout the genome, and probably tag other (unmeasured) histone modifications, such as repressive heterochromatin-associated marks. In order to explore this possibility we analyzed human Hippocampus tissue data from the Epigenomics Roadmap (Consortium et al., 2015), where seven histone modifications, both expressive and repressive, had been measured (**Supplementary Materials**). We found in this case 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 $\sim 32\%$ of the genome (Fig. S3).

Contrary to the empty state, on average 67.11% (range: 57.34-80.42%) 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 promoters (Heintzman et al., 2009). However, our analysis also uncovered a more subtle enrichment of state [me1] shouldering the TSS (Fig. 4a). We found that 42% of [me1] sites occur in regions directly flanking state [me1/2/3+ac] and 74% 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 ($p \approx 10^{-101}$, t-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 analyzed the same datasets using MACS2 (Zhang et al., 2008), one of the most widely used univariate peak callers, and ChromHMM (Ernst and Kellis, 2012). 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. S4). 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 (E2, E3) detected by ChromHMM. A third almost empty state E4 with very weak H3K27ac signal had slightly higher expression levels than the other two empty states and partially overlapped with chromstaR states [me1] and [me1+ac] (Fig. S4). 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 or if they are biologically different on some level.

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. As expected, MACS2 results were noisy: many of the combinatorial states detected by chromstaR showed very heterogenous state calls with MACS2 (Fig. S4). For instance, a considerable proportion (45%) 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 (**Supplementary Materials**).

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 dif-

ferential 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 (Lara-Astiaso et al., 2014), 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 (Leung et al., 2015; Andersson et al., 2014; Dixon et al., 2015; Amin et al., 2015).

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 (McLean et al., 2010) 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 (Zhang et al., 2008) and ChromHMM (Ernst and Kellis, 2012). 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. S5c 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. S5b) 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. S5b), 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 set up 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 (Lara-Astiaso et al., 2014), 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 brach, therefore assigning every bin in the genome to one of the 32 or 64 possible differential combinatorial 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. S6a).

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. S7). 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 S8 shows two examples of pluripotency genes that lose their open chromatin configuration in the differentiated stage.

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), respectively. 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. S9a). Among the constant transitions, regions with signature [me1/2/3+ac] mark constitutively expressed genes (Fig. S9a). 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 (Zhang et al., 2008) and ChromHMM (Ernst and Kellis, 2012). Strikingly, MACS2 found 34470 different chromatin state transitions with the most frequent 330 (with frequency $\geq 0.01\%$) covering only 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 2 cell types, but between 5 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 only 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. S9b).

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) The run-time of our algorithm scales linearly with the number of data points, and thus the easiest strategy is to decrease the resolution, e.g. halfing the run-time by doubling the bin size. (2) Calculations can be performed for each chromosome separately, allowing for easy parallelization of the task. (3) 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. S10). (4) 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. S6a) 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. S6b). 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 (similar to a univariate peak-calling method). The latter gives an accurate picture of the combinatorial chromatin landscape, while differences between cells might be overestimated.

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 ChIP-seq experiments together and cluster them into a finite number of chromatin states (Biesinger et al., 2013; Ernst and Kellis, 2012; Hoffman et al., 2012; Sohn et al., 2015; Zeng et al., 2013). Interestingly, these methods often demand that the user speci-

fies 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* 9 chromatin states have been reported (Kharchenko et al., 2011), while in *A. thaliana* 4 main states were found (Roudier et al., 2011). In human, Ernst et al. found 51 states in human T-cells (Ernst and Kellis, 2010). The Roadmap Consortium reported 15 to 18 states (Consortium et al., 2015). 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 or Segway provide possible computational solutions to such questions, these methods employ probablistic chromatin state definitions that are not always readily interpretatble. 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 defintion 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 modifications only 21 of the 128 possible combinatorial states already covered 99% of the total 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 transition 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 (Gentleman et al., 2004; Huber et al., 2015). chromstaR is freely available at `https://github.com/ataudt/chromstaR`.

Acknowledgements

Text for this section ...

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.

Figures

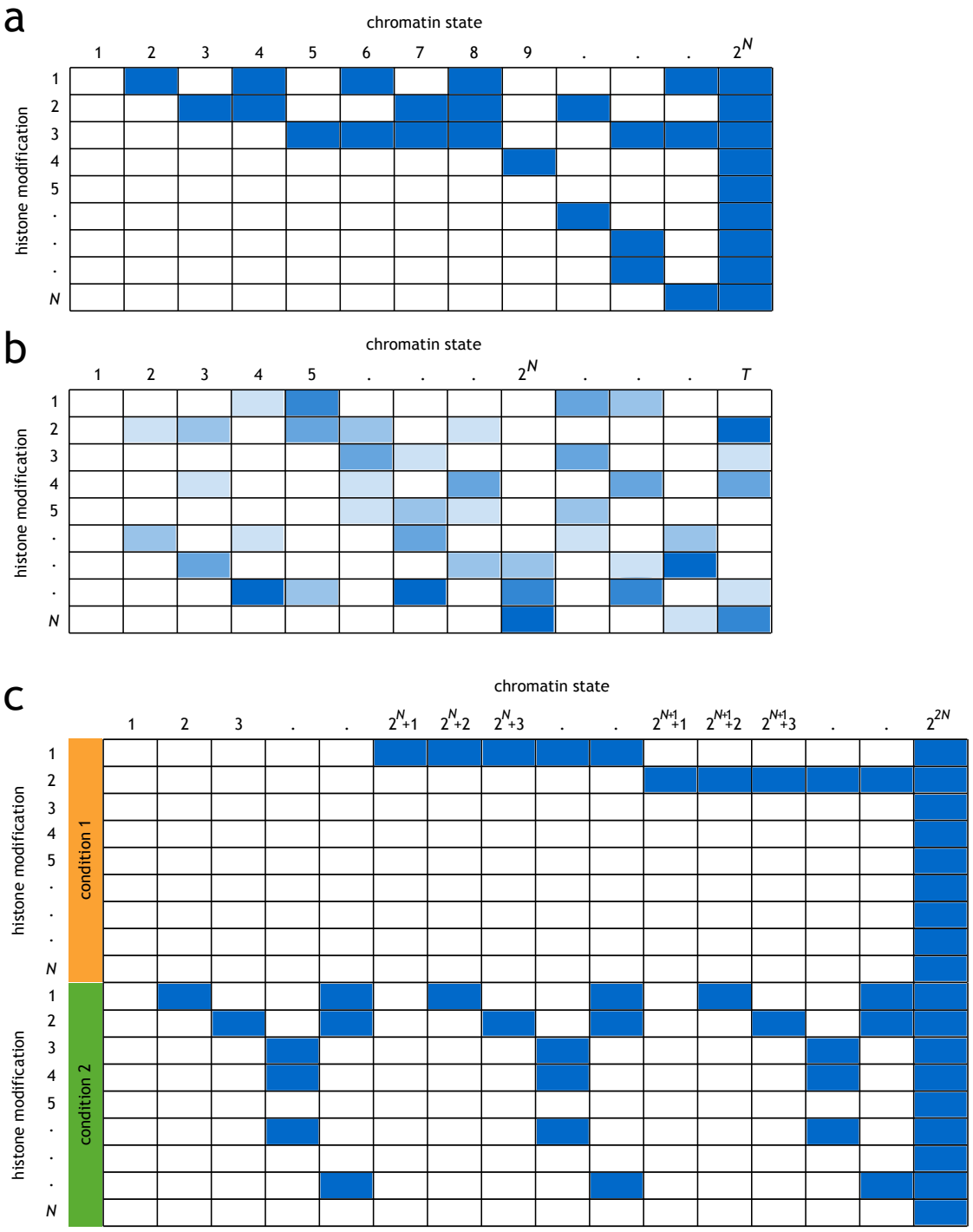

Figure 1: **Definition of chromatin states.** (a) Combinatorial chromatin state definition: Based on the presence (blue) or absence (white) of a histone modification, a chromatin state is the combination of the presence/absence calls at a given position. With N histone modifications there are 2^N different 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.

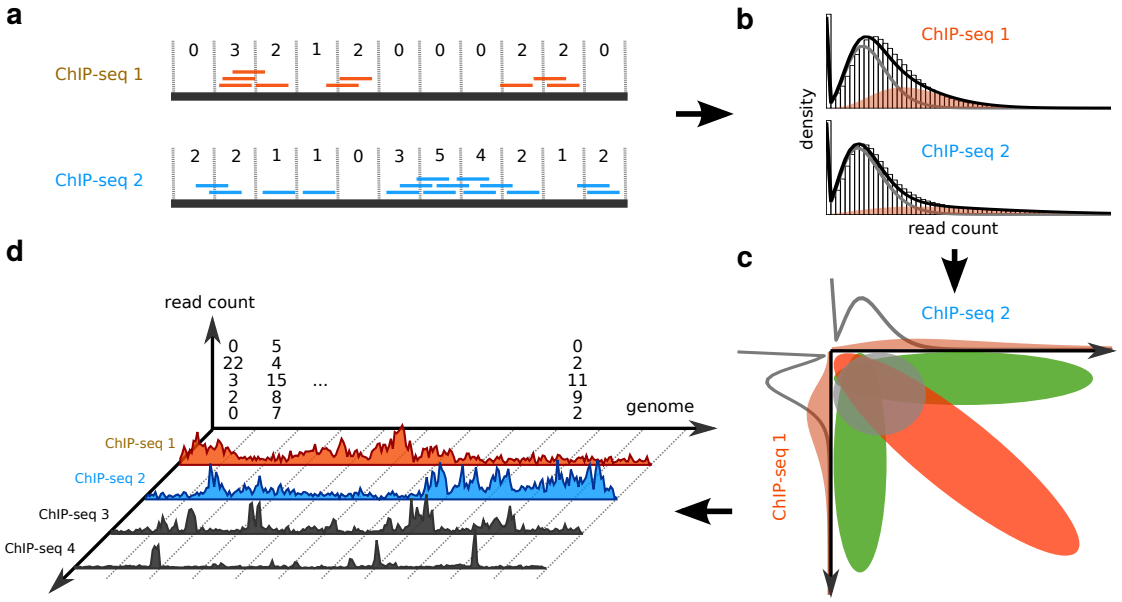

Figure 2: **Overview of analytical approach.** (a) Aligned reads are counted in equidistant, non-overlapping bins. (b) The resulting read count is used to fit a univariate Hidden Markov Model to each ChIP-seq experiment separately. (c) From the univariate emission densities, a multivariate emission density is constructed (shown here for two dimensions). (d) A multivariate Hidden Markov Model is employed to obtain peak-calls for all ChIP-seq experiments combined.

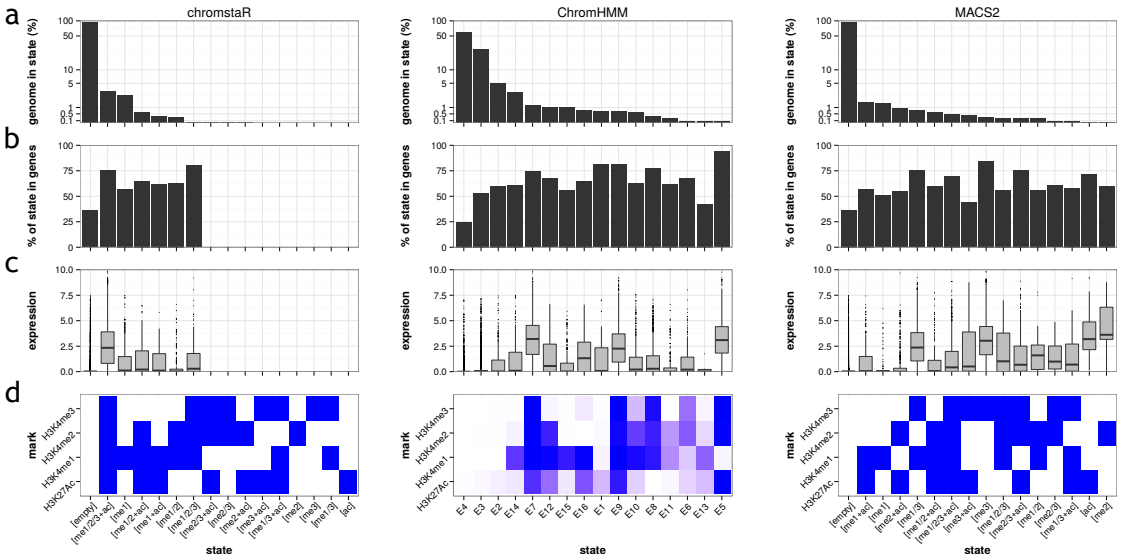

Figure 3: **Chromatin states in monocytes.** (a) Genomic frequency, i.e. the percentage of the genome that is covered by the chromatin state. The sum over all states equals 100%. (b) Overlap with known genes. (c) Expression levels of genes whose TSS overlaps the chromatin state. (d) Heatmap showing the 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).

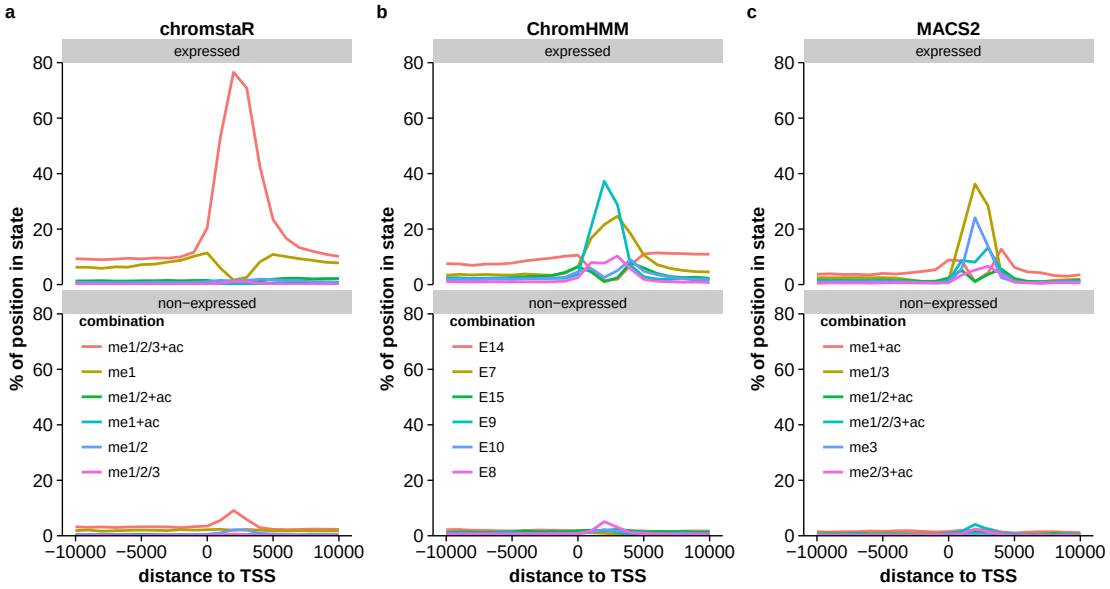

Figure 4: Enrichment of chromatin states around TSS of expressed and non-expressed genes. Shown are the enrichment profiles for the 6 states that are most enriched around TSS.

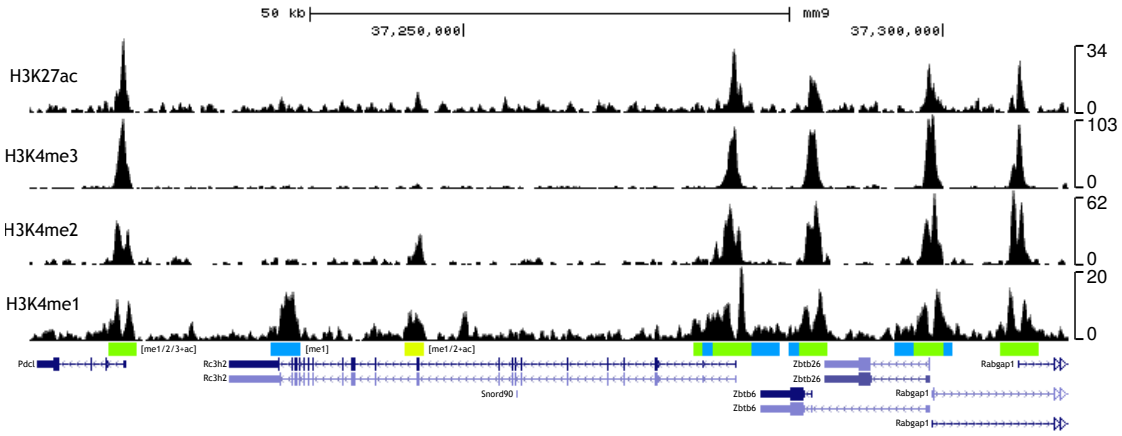

Figure 5: **Genome browser snapshot** showing an example of several active promoter signatures [me1/2/3+ac] flanked by the [me1] signature.

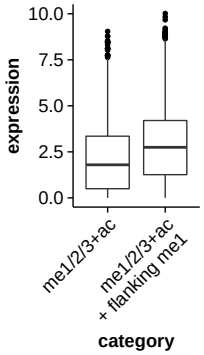

Figure 6: Expression levels of genes whose TSS shows either the [me1/2/3+ac] signature alone or the [me1/2/3+ac] signature flanked by [me1]. TSS flanked by [me1] show significantly higher expression levels ($p \approx 10^{-101}$, t-test).

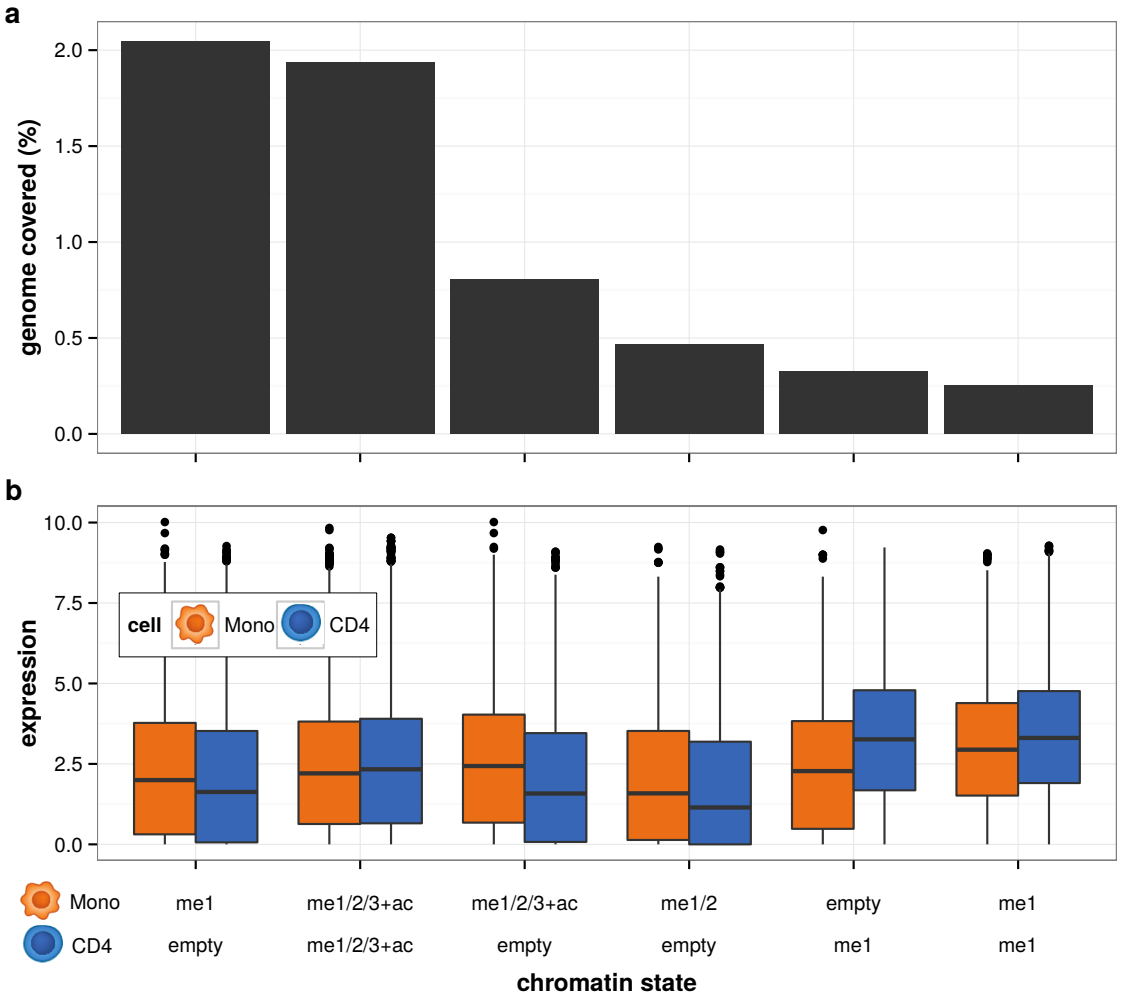

Figure 7: **Differential analysis of monocytes and CD4 T-cells.** (a) Genomic frequency of the 6 most frequent chromatin states. (b) Expression of genes which overlap the given chromatin state.

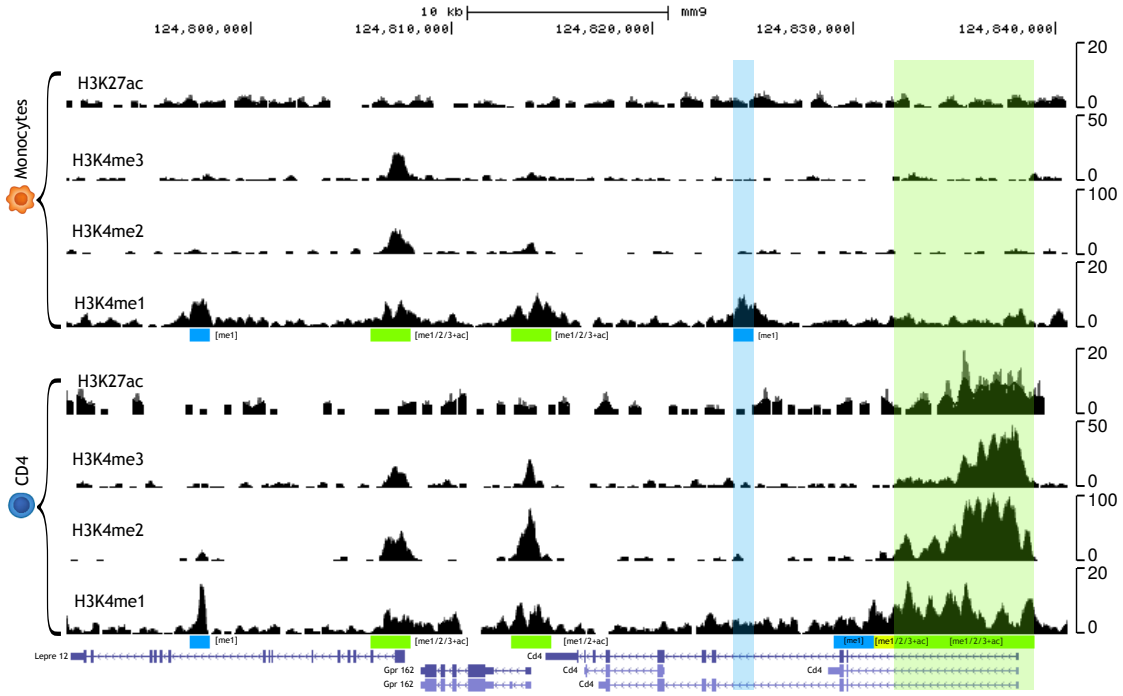

Figure 8: **Differential chromatin signature at the Cd4 locus.** Example of a differential promoter and enhancer signature at the Cd4 gene. The differential promoter signature [me1/2/3+ac] is only present in CD4 T-cells (shaded green), while the differential enhancer [me1] is present only in monocytes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

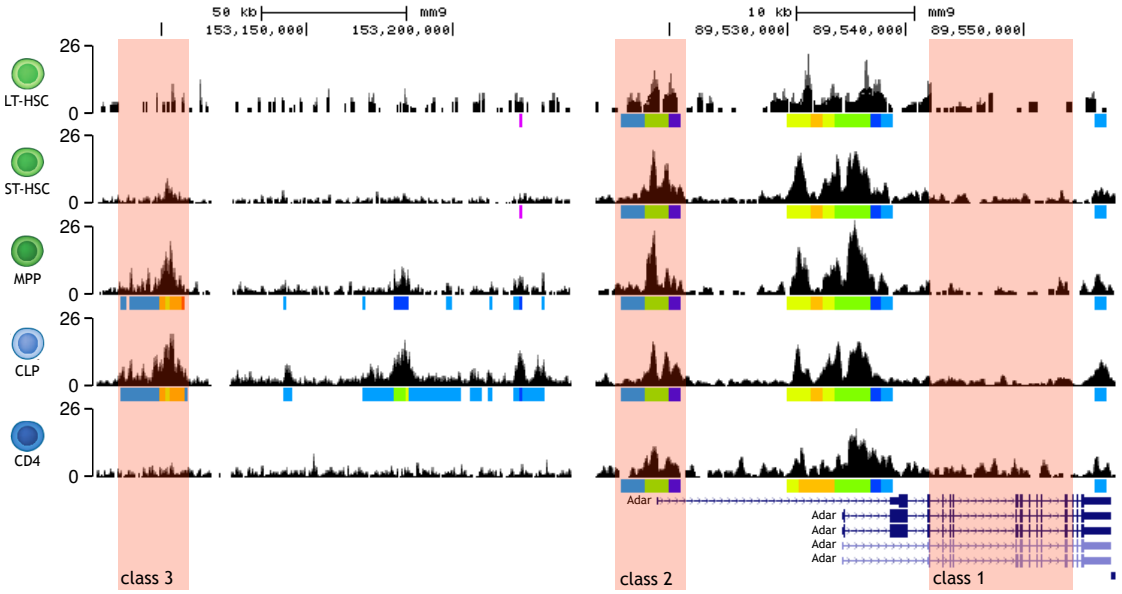

Figure 9: Examples of different classes of transitions (shaded in red): "Empty" (class 1), "constant" (class 2) and "stage-specific" transitions (class 3). Only the H3K4me1 tracks are shown. Combinatorial chromatin states as obtained by chromstaR are shown below the H3K4me1 tracks as colored bars.

Tables

	[me1/2/3+ac] + flanking [me1]	[me1/2/3+ac]
1	nucleobase-containing compound transport	ncRNA metabolic process
2	RNA localization	ncRNA processing
3	negative regulation of mRNA splicing, via spliceosome	tRNA metabolic process
4	RNA transport	protein folding
5	negative regulation of mRNA processing	DNA replication
6	mRNA transport	rRNA metabolic process
7	peptidyl-lysine modification	tRNA processing
8	response to misfolded protein	rRNA processing
9	purinergic nucleotide receptor signaling pathway	protein peptidyl-prolyl isomerization
10	regulation of gene expression, epigenetic	pseudouridine synthesis

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.

References

Amin, V., Harris, R. A., Onuchic, V., Jackson, A. R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C., and Milosavljevic, A., *et al.*, 2015. Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nature Communications*, **6**(May 2014):6370.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.*, 2014. An atlas of active enhancers across human cell types and tissues. *Nature*, **507**(7493):455–61.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K., 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4):823–37.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N., 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, **41**(1):164–171.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

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., *et al.*, 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**(2):169–81.

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., *et al.*,

2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10):1045–8.

Biesinger, J., Wang, Y., and Xie, X., 2013. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC bioinformatics*, **14 Suppl 5**:S4.

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.*, 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, **518**:317–330.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., *et al.*, 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539):331–336.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, **21**(16):3439–40.

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

Ernst, J. and Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, **28**(8):817–25.

Ernst, J. and Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3):215–216.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.*, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10):R80.

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**(1):60.

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., *et al.*, 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243):108–12.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, **9**(5):473–6.

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., *et al.*, 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2):827–41.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., *et al.*, 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**(2):115–121.

Huda, A., Mariño-Ramírez, L., and Jordan, I. K., 2010. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA*, **1**(1):2.

Jenuwein, T. and Allis, C. D., 2001. Translating the histone code. *Science (New York, N.Y.)*, **293**(5532):1074–80.

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., *et al.*, 2011. Com-

prehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**(7339):480–5.

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., *et al.*, 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research*, **17**(6):691–707.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4):357–9.

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. Chromatin state dynamics during blood formation. *Science (New York, N.Y.)*, **55**(233348):1–10.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J., 2013. Software for computing and annotating genomic ranges. *PLoS computational biology*, **9**(8):e1003118.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., and Yen, C.-a., 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, **518**:350–354.

Luo, C., Sidote, D. J., Zhang, Y., Kerstetter, R. A., Michael, T. P., and Lam, E., 2013. 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.

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. *Nature biotechnology*, **28**(5):495–501.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., *et al.*, 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**(7153):553–60.

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., *et al.*, 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**(4):517–27.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D., 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, **12**(7):R67.

Renard, B. and Lang, M., 2007. Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources*, **30**(4):897–912.

Rintisch, C., Heinig, M., Bauerfeind, A., Schafer, S., Mieth, C., Patone, G., Hummel, O., Chen, W., Cook, S., Cuppen, E., *et al.*, 2014. Natural variation of histone modification and its impact on gene expression in the rat genome. *Genome research*, **24**(6):942–53.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.*, 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, **4**(8):651–7.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., *et al.*, 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal*, **30**(10):1928–1938.

Sklar, M., 1959. Fonctions de répartition à n dimensions et leurs marges. .

Sohn, K.-A., Ho, J. W. K., Djordjevic, D., Jeong, H.-H., Park, P. J., and Kim, J. H., 2015. hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics (Oxford, England)*, :btv117–.

Spyrou, C., Stark, R., Lynch, A. G., and Tavaré, S., 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC bioinformatics*, **10**:299.

van der Graaf, A., Wardenaar, R., Neumann, D. A., Taudt, A., Shaw, R. G., Jansen, R. C., Schmitz, R. J., Colomé-Tatché, M., and Johannes, F., 2015. 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.

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. *Nature genetics*, **40**(7):897–903.

Yen, A. and Kellis, M., 2015. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nature Communications*, **6**:7973.

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., and Keleş, S., 2013. jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome biology*, **14**(4):R38.

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 biology*, **9**(9):R137.

Supplementary Materials

Model Specification

The construction of the multivariate Hidden Markov Model can be divided in two steps. 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 (Rashid et al., 2011; Spyrou et al., 2009), 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 (van der Graaf et al., 2015). 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

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

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 (Baum et al., 1970) to obtain a 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 (Sklar, 1959; Heinig et al., 2015).

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 (Renard and Lang, 2007).

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 $\mathbf{\Sigma}_i$ is calculated from the transformed read counts.

Similarly to the univariate Hidden Markov Model, we use the Baum-Welch algorithm to obtain a 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.

Data Acquisition

ChIP-seq data for the hematopoietic data (GSE60103) was downloaded from the Gene Expression Omnibus (GEO) and aligned to mouse reference mm9 following the procedure in (Lara-Astiaso et al., 2014) with bowtie2 (version 2.2.3) (Langmead and Salzberg, 2012), 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 arc-sinh transformation on the data.

Multivariate peak-calling

chromstaR was run with a bin size of 1000bp and convergence threshold of $eps = 0.01$ for both the univariate and multivariate part. Univariate fits were checked manually for proper convergence and rerun with different random initial parameter settings where necessary. For all analysis and comparisons, we excluded replicates SRR1521819, SRR1521851 and SRR1521852 (corresponding to CD8-H3K27ac-Rep1, MF-H3K4me1-Rep1, MF-H3K4me1-Rep2) because we could not obtain a proper fit with our method, regardless of initial parameter settings. Replicates were included in the chromstaR analysis as separate ChIP-seq experiments but forced to yield the same state calls see "Inclusion of replicates" below). Likewise, ChromHMM was run with a bin size of 1000bp, 16 states, parallel mode, assembly mm9 and default parameters otherwise. Signal input files for ChromHMM were produced by adding the read counts over replicates. MACS2 (version 2.1.0.20150731) was run with parameters "-g mm –keep-dup all" and default settings otherwise. Replicates were specified separately and handled by MACS2 internally. For the comparison with chromstaR and ChromHMM, MACS2 calls were transformed into a 1000bp-bin representation by simply extending each peak into its overlapping bin(s). chromstaR and ChromHMM were run on chromosomes 1-19 and X, MACS2 was run with all scaffolds but only chromosomes 1-19 and X retained for analysis.

Analysis

Genomic coordinates were downloaded with biomaRt (Durinck et al., 2005, 2009) (dataset=mmusculus_gene_ensembl, host=aug2010.archive.ensembl.org) and the first three basepairs of each gene were defined as coordinates for the transcription start site. For the overlap of chromation states with genes (Fig. 3b) we included the promoter region defined as 2kb upstream of each gene in the gene definition. Gene ontology enrichment was performed with GREAT (McLean et al., 2010) using the whole genome as background set. Significant terms were filtered out with the following thresholds: BinomFdrQ < 0.05, HyperFdrQ < 0.1, RegionFoldEnrich > 2. Presented terms in all tables are from category "GO Biological Process" and ordered by BinomFdrQ with the most significant results on top.

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).

	sensitivity	precision	F1-score
chromstaR	0.77	0.95	0.85
MACS2	0.60	0.98	0.75
ChromHMM	0.59	0.98	0.73

Table 2: Performance for detecting expressed TSS.

Analysis of human Hippocampus tissue

Bed-files for Hippocampus tissue were downloaded from "ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/sample-experiment/" for donors number 112 and 149. Histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3 were analyzed at bin size 1000bp with convergence threshold of $eps = 0.01$ and donors 112 and 149 included as replicates. We found 21 out of $2^7 = 128$ possible states (genomic frequency $\geq 0.1\%$) covering more than 99% of the genome (Fig. S3).

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 combinatorial 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 7 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. S10).

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 in all replicates (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.

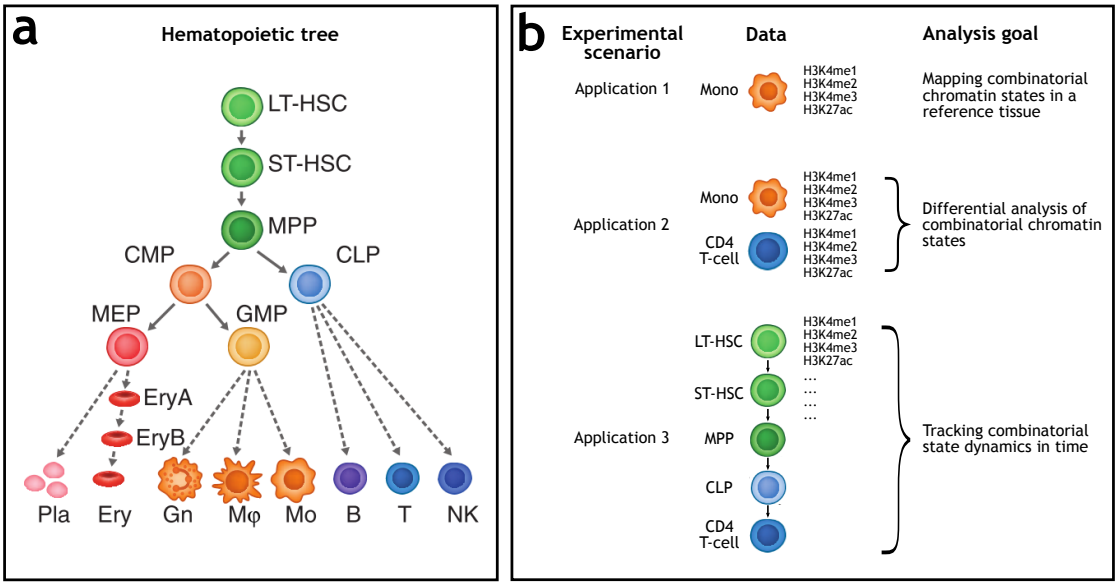

Figure S1: (a) Hematopoietic tree with cells that were probed by Lara-Astiaso et al. Lara-Astiaso et al. (2014). (b) Three experimental scenarios that are investigated in this manuscript.

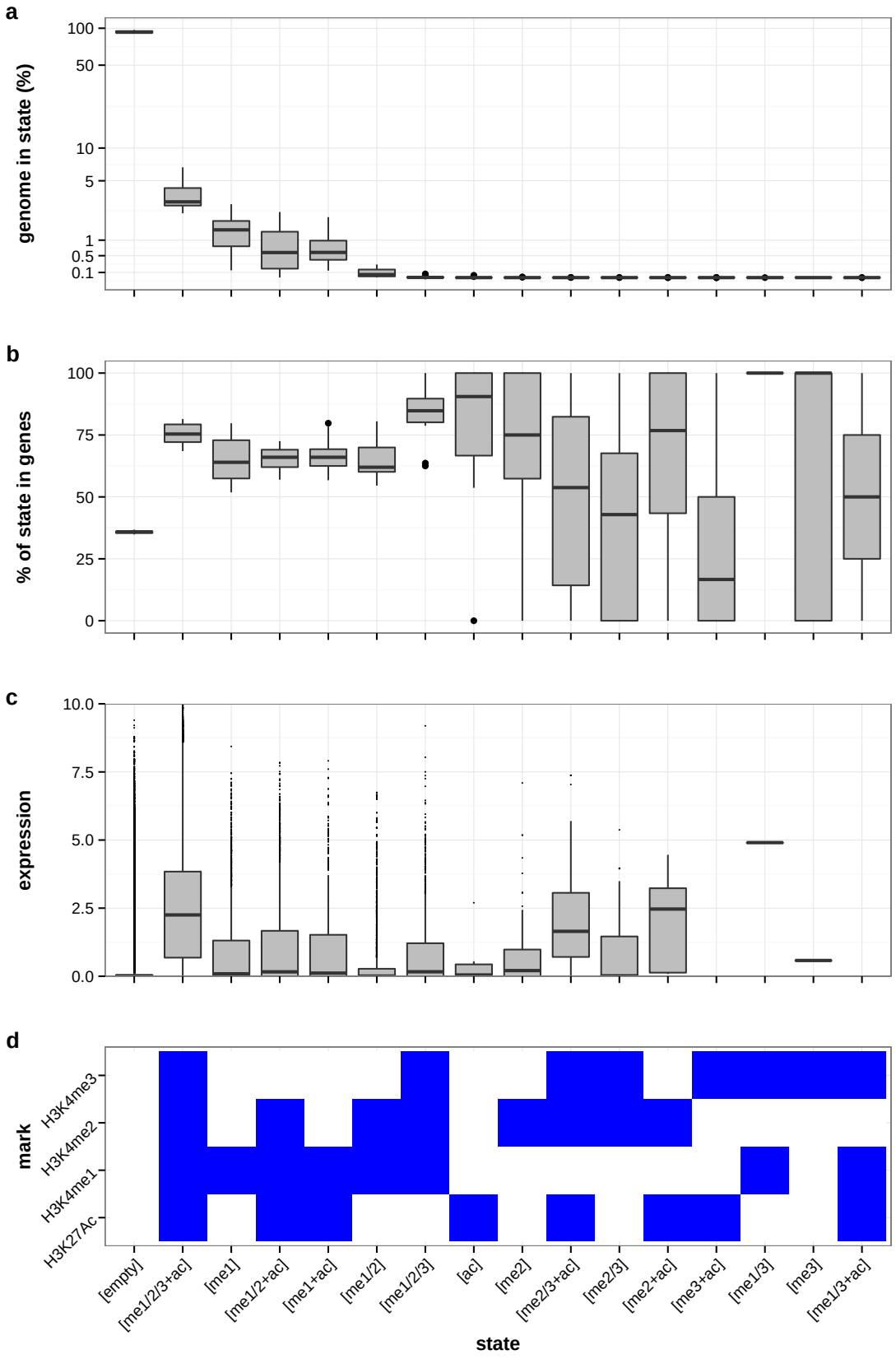

Figure S2: Boxplots depict values for all 16 measured hematopoietic cell types from Lara-Astiaso et al. (2014). (a) Genomic frequency, i.e. the percentage of the genome that is covered by the chromatin state. (b) Overlap with known genes. (c) Expression levels of genes whose TSS overlaps the chromatin state. (d) Heatmap showing the chromatin state definition (blue is present, white is absent).

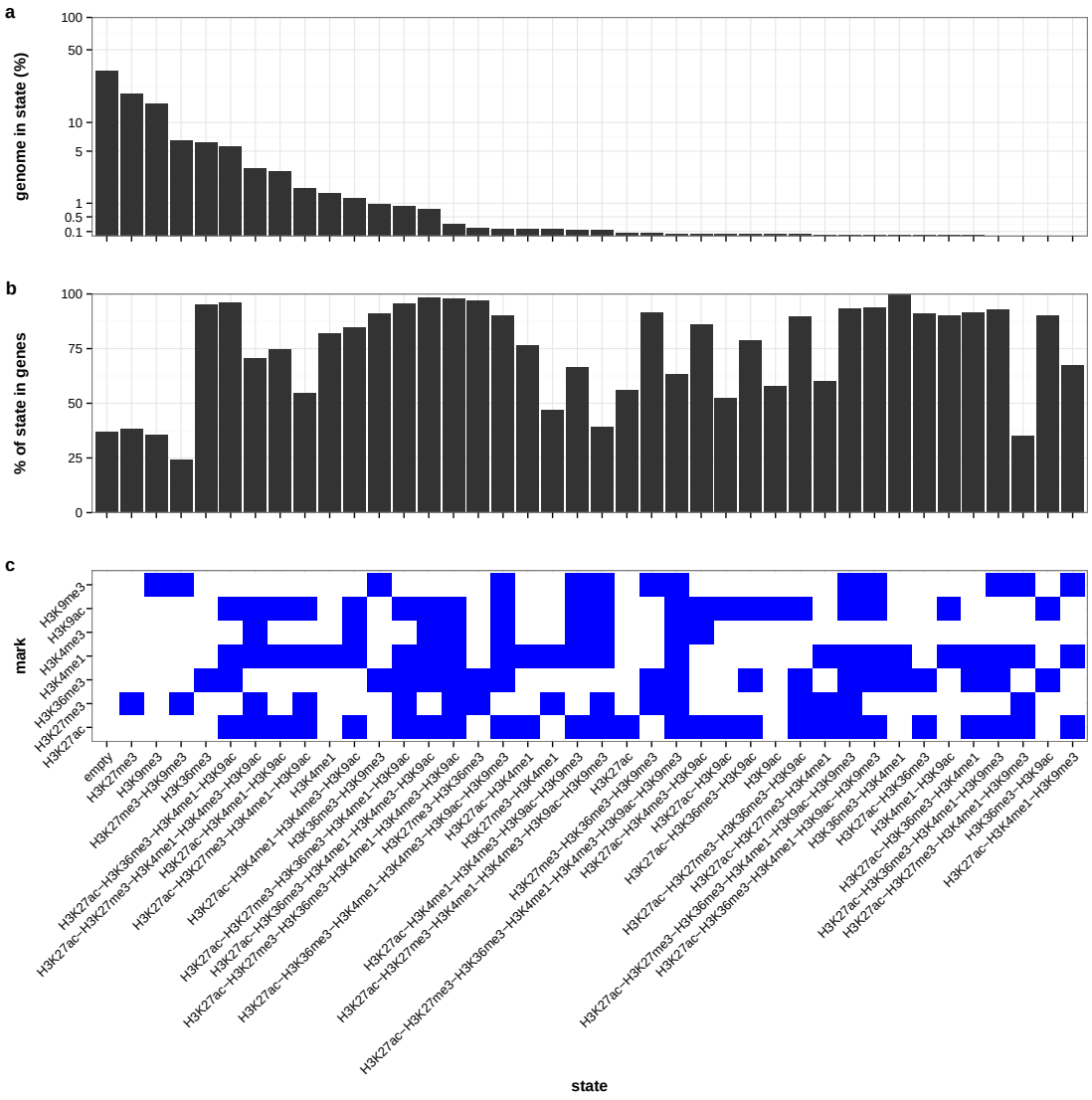

Figure S3: **Chromatin states in human Hippocampus tissue.** (a) Genomic frequency, i.e. the percentage of the genome that is covered by the chromatin state, for the 40 most frequent states. (b) Overlap with known genes. (c) Heatmap showing the chromatin state definition.

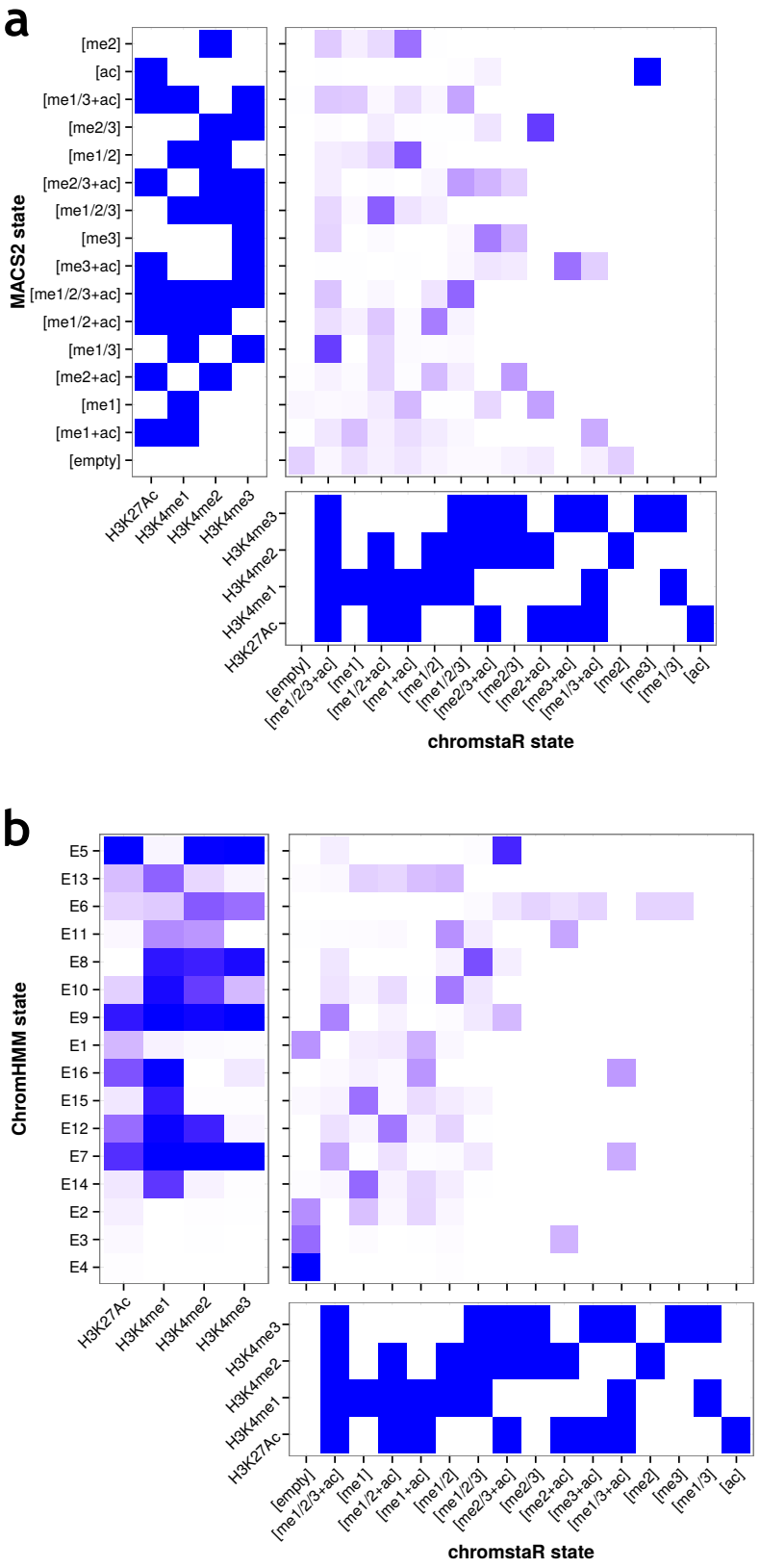

Figure S4: Confusion matrix for the comparison of chromstaR with (a) MACS2 and (b) ChromHMM. The confusion matrix shows the fold enrichment of states from both methods with each other, with darker tiles (blue) indicating a higher overlap.

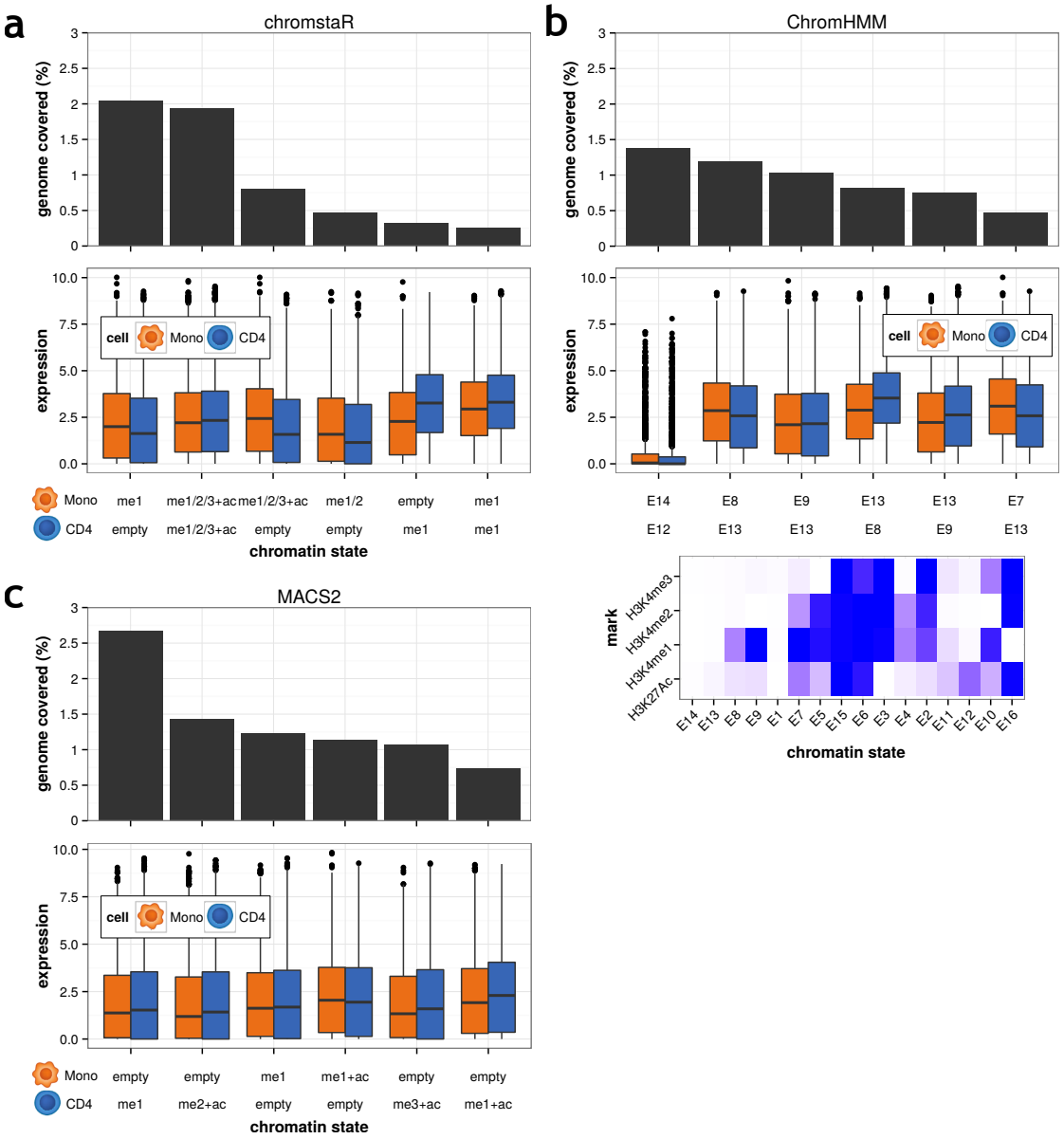

Figure S5: **Differential analysis of monocytes and CD4 T-cells.** Genomic frequency and expression levels for genes that overlap the 6 most frequent differential chromatin states for (a) chromstaR, (b) ChromHMM and (c) MACS2.

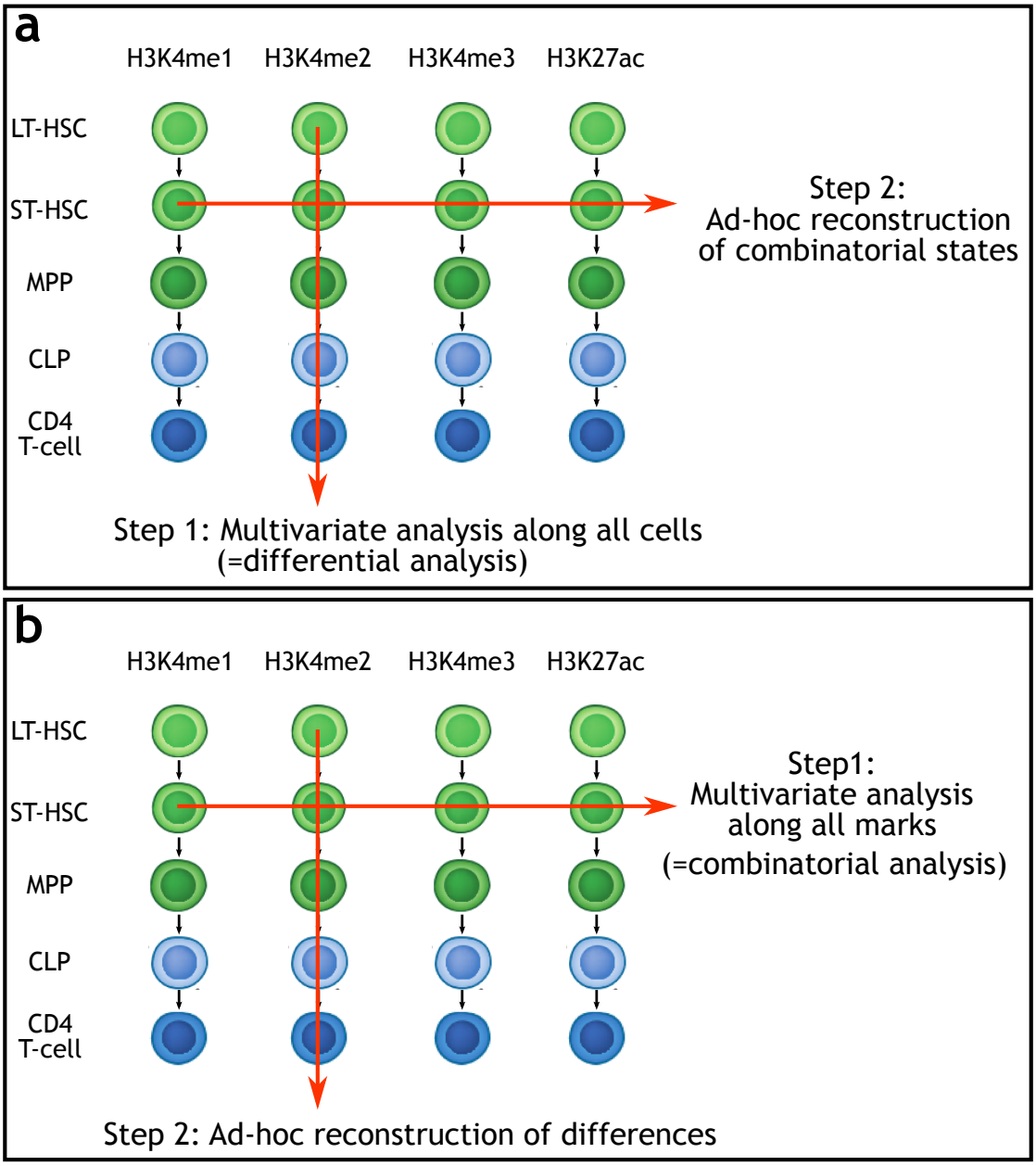

Figure S6: **Two-step approach for inferring combinatorial state differences.** (a) In a first step, multivariate peak-calls are obtained along all cells for each mark separately (differential analysis). Those calls are then combined, ad-hoc, into the combinatorial states. (b) In a first step, combinatorial states are obtained for each cell using the multivariate approach. Differences between those states are then obtained by a simple comparison between cells.

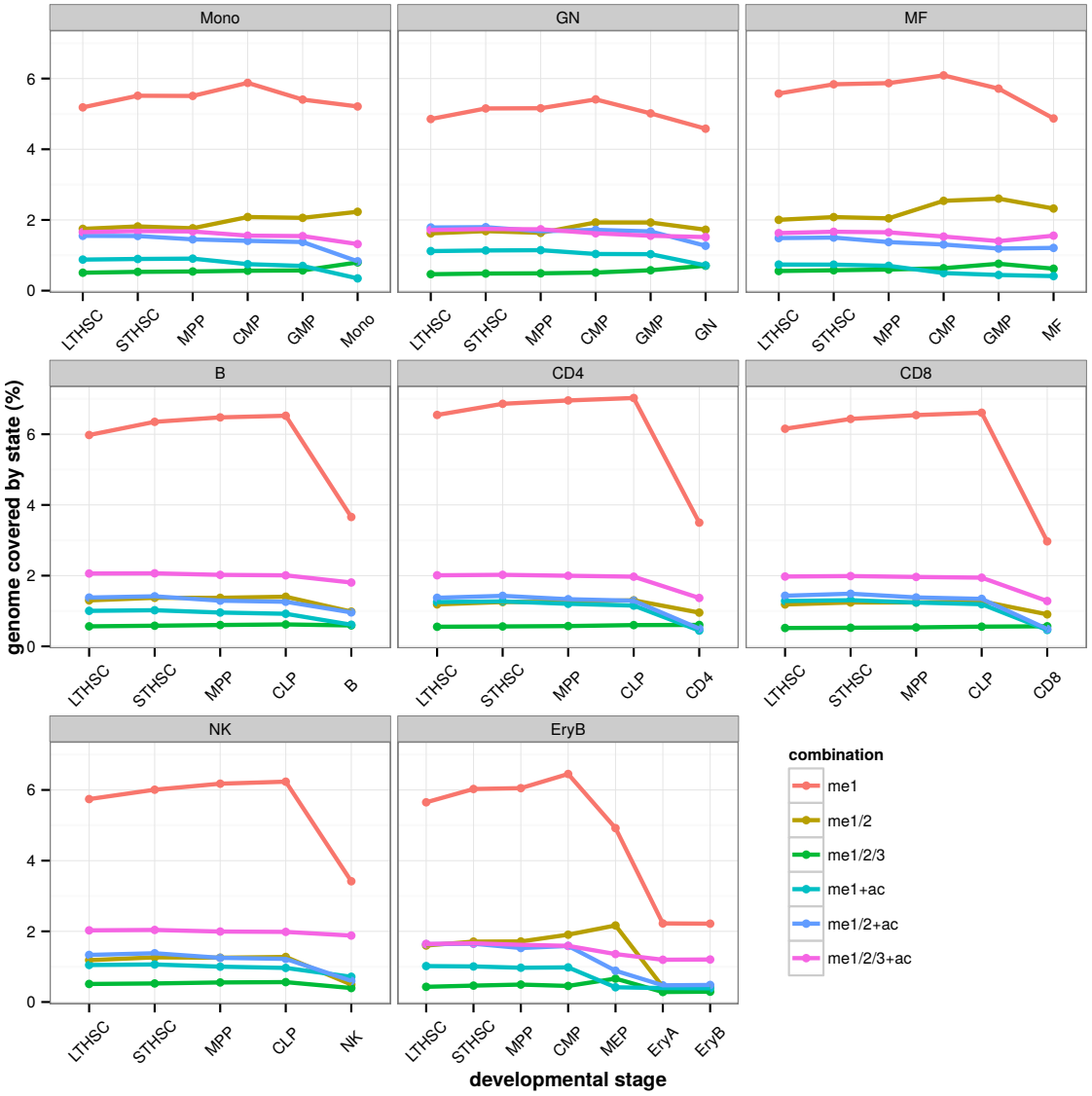

Figure S7: Genomic frequency of combinatorial states during differentiation for all branches of the hematopoietic tree (Fig. S1a).

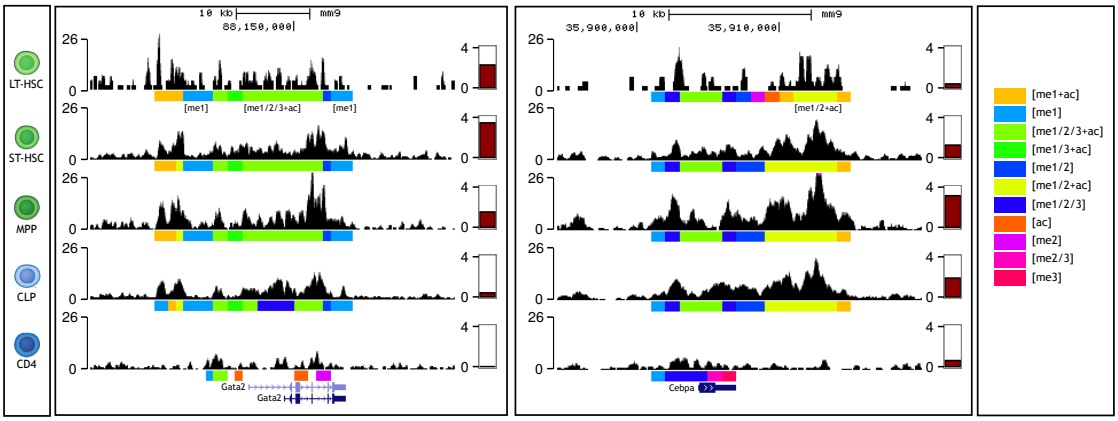

Figure S8: **Chromatin state transitions at (a) Gata2 and (b) Cebpa.** Black genome browser tracks show H3K4me1 levels and combinatorial chromatin states as determined by chromstaR below. Red bars on the right of each track indicate normalized expression levels of Gata2 and Cebpa, respectively. Both Gata2 and Cebpa are involved in maintenance of pluripotency in stem cells and transition from an open into a closed chromatin configuration during differentiation.

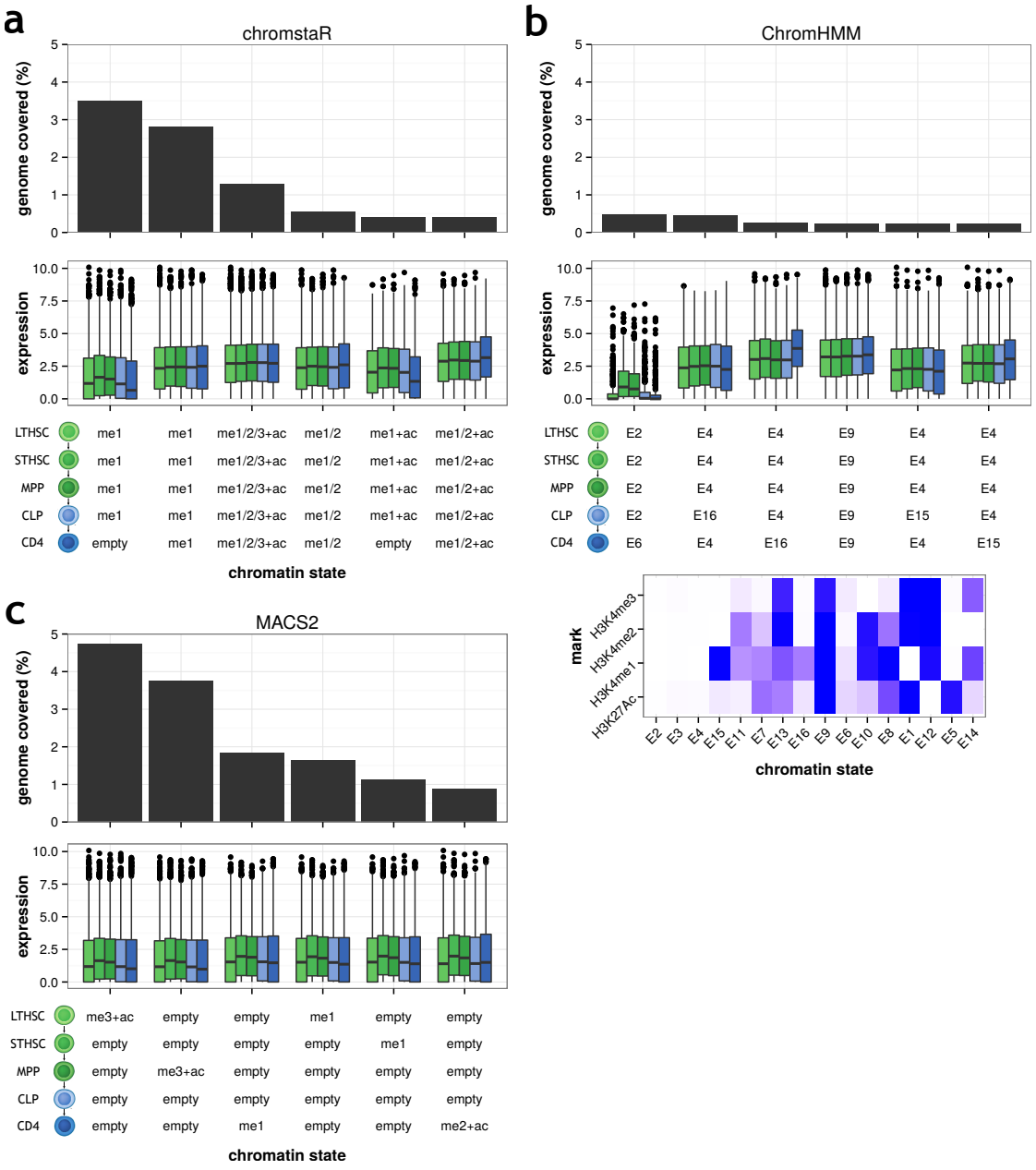

Figure S9: **Chromatin state transitions for the CD4 branch.** Genomic frequency and expression levels for genes that overlap the 6 most frequent chromatin state transitions for (a) chromstaR, (b) ChromHMM and (c) MACS2.

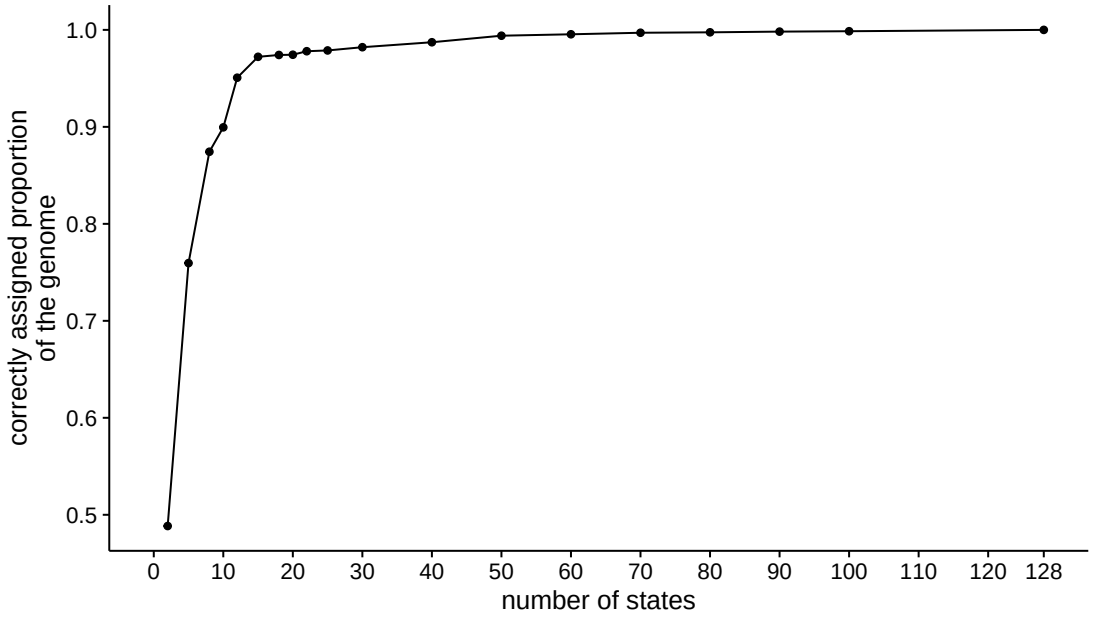

Figure S10: **Approximation of multivariate state distribution with less than 2^N states.** For the Hippocampus data with 7 marks there are $2^7 = 128$ possible combinatorial states (last data point). The figure shows the proportion of the genome that is correctly assigned compared to the full 128-state model (y-axis) if the multivariate is run with fewer than 128 states (x-axis).

Mono	me1/2/3+ac	me1/2/3+ac	empty
CD4	me1/2/3+ac	empty	me1/2/3+ac
1	mature B cell differentiation	response to other organism	leukocyte activation
2	apoptotic mitochondrial changes	response to biotic stimulus	T cell activation
3	regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	immune response	T cell differentiation
4	antigen processing and presentation of exogenous peptide antigen	inflammatory response	regulation of immune system process
5	positive regulation of mRNA catabolic process	response to bacterium	cell activation
6	regulation of RNA stability	positive regulation of immune system process	lymphocyte activation
7	nuclear-transcribed mRNA catabolic process	leukocyte activation	lymphocyte differentiation
8	regulation of mRNA stability	regulation of immune response	alpha-beta T cell activation
9	positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	regulation of cytokine production	immune system process
10	GPI anchor metabolic process	response to molecule of bacterial origin	leukocyte differentiation

Mono	me1	me1	empty
CD4	me1	empty	me1
1	regulation of immune system process	response to other organism	regulation of immune system process
2	leukocyte activation	response to bacterium	leukocyte activation
3	cell activation	inflammatory response	lymphocyte activation
4	immune system development	response to molecule of bacterial origin	cell activation
5	lymphocyte activation	immune effector process	T cell activation
6	hematopoietic or lymphoid organ development	response to lipopolysaccharide	lymphocyte differentiation
7	hemopoiesis	negative regulation of transferase activity	immune system process
8	positive regulation of immune system process	myeloid cell differentiation	leukocyte differentiation
9	lymphocyte differentiation	negative regulation of kinase activity	T cell differentiation
10	regulation of lymphocyte activation	homeostasis of number of cells	alpha-beta T cell activation

Table S1: The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with chromstaR. Only significant terms are shown.

Mono CD4	empty me1	empty me2+ac	me1 empty	me1+ac empty	empty me3+ac	empty me1+ac
1				myeloid cell differentiation		T cell activation
2				homeostasis of number of cells		T cell differentiation
3				cytokine-mediated signaling pathway		alpha-beta T cell activation
4				myeloid cell homeostasis		alpha-beta T cell differentiation
5				erythrocyte homeostasis		T cell activation involved in immune response
6				erythrocyte differentiation		leukocyte activation involved in immune response
7				B cell differentiation		lymphocyte activation involved in immune response
8				Ras protein signal transduction		T cell selection
9				myeloid leukocyte differentiation		T cell proliferation
10				myeloid leukocyte activation		positive regulation of myeloid leukocyte differentiation

Table S2: The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with MACS2. Only significant terms are shown.

Mono	E14	E8	E9
CD4	E12	E13	E13
1	homophilic cell adhesion	immune system process	intrinsic apoptotic signaling pathway
2	neuron recognition	response to biotic stimulus	apoptotic mitochondrial changes
3	axon choice point recognition	response to other organism	myeloid cell homeostasis
4	axon midline choice point recognition	leukocyte activation	nuclear export
5	negative chemotaxis	immune response	B cell differentiation
6	startle response	cell activation	regulation of intrinsic apoptotic signaling pathway
7	olfactory bulb interneuron differentiation	immune system development	response to starvation
8	innervation	positive regulation of immune system process	regulation of mitochondrial membrane permeability
9	gamma-aminobutyric acid signaling pathway	response to bacterium	fatty acid biosynthetic process
10	corticospinal tract morphogenesis	immune effector process	erythrocyte homeostasis

Mono	E13	E13	E7
CD4	E8	E9	E13
1	leukocyte activation	T cell activation	immune system process
2	cell activation	intra-Golgi vesicle-mediated transport	response to biotic stimulus
3	hematopoietic or lymphoid organ development	peptidyl-lysine modification	response to other organism
4	chromatin modification	macromolecule methylation	regulation of immune system process
5	protein modification by small protein conjugation or removal	protein methylation	leukocyte activation
6	immune system development	protein acylation	multi-organism process
7	lymphocyte activation	protein acetylation	cell activation
8	protein modification by small protein conjugation	regulation of B cell activation	immune response
9	hemopoiesis	internal protein amino acid acetylation	response to bacterium
10	protein ubiquitination	ncRNA processing	myeloid leukocyte activation

Table S3: The first 10 gene ontology terms for selected differential regions between monocytes and CD4 T-cells after analysis with chromHMM. Only significant terms are shown.

LTHSC	me1	me1	empty
STHSC	me1	me1	me1
MPP	me1	me1	me1
CLP	me1	me1	me1
CD4	empty	me1	empty
1	negative regulation of MAP kinase activity	apoptotic signaling pathway	cell chemotaxis
2	myeloid leukocyte activation	lymphocyte differentiation	leukocyte migration
3	phagocytosis	T cell activation	mesodermal cell differentiation
4	regulation of lipid kinase activity	intrinsic apoptotic signaling pathway	positive regulation of peptidyl-tyrosine phosphorylation
5	cellular response to vascular endothelial growth factor stimulus	immune effector process	leukocyte chemotaxis
6	regulation of phosphatidylinositol 3-kinase activity	regulation of T cell activation	platelet-derived growth factor receptor signaling pathway
7	regulation of myeloid leukocyte mediated immunity	negative regulation of protein kinase activity	regulation of vascular endothelial growth factor receptor signaling pathway
8	negative regulation of multi-organism process	negative regulation of kinase activity	odontogenesis of dentin-containing tooth
9	regulation of lipopolysaccharide-mediated signaling pathway	negative regulation of transferase activity	positive regulation of cell migration involved in sprouting angiogenesis
10	regulation of leukocyte degranulation	regulation of B cell activation	glomerulus vasculature development

LTHSC	empty	me1
STHSC	empty	me1
MPP	empty	empty
CLP	empty	empty
CD4	me1	empty
1	leukocyte activation	regulation of mRNA stability
2	alpha-beta T cell activation	regulation of small GTPase mediated signal transduction
3	T cell activation	retinoic acid receptor signaling pathway
4	regulation of immune system process	myeloid cell differentiation
5	leukocyte differentiation	microtubule-based transport
6	lymphocyte differentiation	negative regulation of lymphocyte proliferation
7	lymphocyte activation	cytoskeleton-dependent intracellular transport
8	cell activation	otolith development
9	T cell differentiation	negative regulation of leukocyte proliferation
10	T cell selection	otolith morphogenesis

Table S4: The first 10 gene ontology terms for selected regions in the CD4 T-cells lineage after analysis with chromstaR. Only significant terms are shown.

LTHSC	me1/2/3+ac	me1/2/3+ac	empty	me1/2/3+ac	me1/2/3+ac
STHSC	me1/2/3+ac	me1/2/3+ac	empty	me1/2/3+ac	me1/2/3+ac
MPP	me1/2/3+ac	me1/2/3+ac	empty	empty	empty
CLP	me1/2/3+ac	me1/2/3+ac	empty	empty	empty
CD4	me1/2/3+ac	empty	me1/2/3+ac	empty	me1/2/3+ac
1	protein folding	immune system process	T cell differentiation		
2	GPI anchor metabolic process	immune response	lymphocyte differentiation		
3	GPI anchor biosynthetic process	homeostasis of number of cells	lymphocyte activation		
4	apoptotic mitochondrial changes	immune system development	immune system process		
5	nuclear export	regulation of immune response	T cell activation		
6	intrinsic apoptotic signaling pathway in response to DNA damage	hematopoietic or lymphoid organ development	T cell selection		
7	protein lipidation	B cell activation	T cell receptor V(D)J recombination		
8	positive regulation of mRNA catabolic process	leukocyte activation	leukocyte activation		
9	vacuole organization	hemopoiesis	positive T cell selection		
10	regulation of nuclear-transcribed mRNA catabolic process	myeloid leukocyte differentiation	leukocyte differentiation		

Table S5: The first 10 gene ontology terms for selected regions in the CD4 T-cells lineage after analysis with chromstaR. Only significant terms are shown.

References

Amin, V., Harris, R. A., Onuchic, V., Jackson, A. R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C., and Milosavljevic, A., *et al.*, 2015. Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nature Communications*, **6**(May 2014):6370.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.*, 2014. An atlas of active enhancers across human cell types and tissues. *Nature*, **507**(7493):455–61.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K., 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4):823–37.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N., 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, **41**(1):164–171.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

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., *et al.*, 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**(2):169–81.

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., *et al.*, 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10):1045–8.

Biesinger, J., Wang, Y., and Xie, X., 2013. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC bioinformatics*, **14 Suppl 5**:S4.

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kherad-

pour, P., Zhang, Z., Wang, J., *et al.*, 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, **518**:317–330.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., *et al.*, 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539):331–336.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, **21**(16):3439–40.

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

Ernst, J. and Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, **28**(8):817–25.

Ernst, J. and Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3):215–216.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.*, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10):R80.

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**(1):60.

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., *et al.*, 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243):108–12.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, **9**(5):473–6.

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., *et al.*, 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2):827–41.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., *et al.*, 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**(2):115–121.

Huda, A., Mariño-Ramírez, L., and Jordan, I. K., 2010. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA*, **1**(1):2.

Jenuwein, T. and Allis, C. D., 2001. Translating the histone code. *Science (New York, N.Y.)*, **293**(5532):1074–80.

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., *et al.*, 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**(7339):480–5.

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., *et al.*, 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research*, **17**(6):691–707.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4):357–9.

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. Chromatin state dynamics during blood formation. *Science (New York, N.Y.)*, **55**(233348):1–10.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J., 2013. Software for computing and annotating genomic ranges. *PLoS computational biology*, **9**(8):e1003118.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., and Yen, C.-a., 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, **518**:350–354.

Luo, C., Sidote, D. J., Zhang, Y., Kerstetter, R. A., Michael, T. P., and Lam, E., 2013. 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.

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. *Nature biotechnology*, **28**(5):495–501.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., *et al.*, 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**(7153):553–60.

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., *et al.*, 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**(4):517–27.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D., 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, **12**(7):R67.

Renard, B. and Lang, M., 2007. Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources*, **30**(4):897–912.

Rintisch, C., Heinig, M., Bauerfeind, A., Schafer, S., Mieth, C., Patone, G., Hummel, O., Chen, W., Cook, S., Cuppen, E., *et al.*, 2014. Natural variation of histone modification and its impact on gene expression in the rat genome. *Genome research*, **24**(6):942–53.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol,

R., Delaney, A., *et al.*, 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, **4**(8):651–7.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., *et al.*, 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal*, **30**(10):1928–1938.

Sklar, M., 1959. Fonctions de répartition à n dimensions et leurs marges. .

Sohn, K.-A., Ho, J. W. K., Djordjevic, D., Jeong, H.-H., Park, P. J., and Kim, J. H., 2015. hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics (Oxford, England)*, :btv117–.

Spyrou, C., Stark, R., Lynch, A. G., and Tavaré, S., 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC bioinformatics*, **10**:299.

van der Graaf, A., Wardenaar, R., Neumann, D. A., Taudt, A., Shaw, R. G., Jansen, R. C., Schmitz, R. J., Colomé-Tatché, M., and Johannes, F., 2015. 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.

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. *Nature genetics*, **40**(7):897–903.

Yen, A. and Kellis, M., 2015. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nature Communications*, **6**:7973.

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., and Keleş, S., 2013. jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome biology*, **14**(4):R38.

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 biology*, **9**(9):R137.