Unsupervised tensor decomposition-based method to extract candidate transcription factors as histone modification bookmarks in post-mitotic transcriptional reactivation

Y-h. Taguchi[1,*], Turki Turki[2]

1 Department of Physics, Chuo University, 112-8551 Tokyo, Japan
2 King Abdulaziz University, Department of Computer Science, Jeddah, 21589, Saudi Arabia

* tag@granular.com

Abstract

The histone group added to a gene sequence must be removed during mitosis to halt transcription during the DNA replication stage of the cell cycle. However, the detailed mechanism of this transcription regulation remains unclear. In particular, it is not realistic to reconstruct all appropriate histone modifications throughout the genome from scratch after mitosis. Thus, it is reasonable to assume that there might be a type of "bookmark" that retains the positions of histone modifications, which can be readily restored after mitosis. We developed a novel computational approach comprising tensor decomposition (TD)-based unsupervised feature extraction (FE) to identify transcription factors (TFs) that bind to genes associated with reactivated histone modifications as candidate histone bookmarks. To the best of our knowledge, this is the first application of TD-based unsupervised FE to the cell division context and phases pertaining to the cell cycle in general. The candidate TFs identified with this approach were functionally related to cell division, suggesting the suitability of this method and the potential of the identified TFs as bookmarks for histone modification during mitosis.

1 Introduction

During the cell division process, gene transcription must be initially terminated and then reactivated once cell division is complete. However, the specific mechanism and factors controlling this process of transcription regulation remain unclear. Since it would be highly time- and energy-consuming to mark all genes that need to be transcribed from scratch after each cycle of cell division, it has been proposed that genes that need to be transcribed are "bookmarked" to easily recover these positions for reactivation [Festuccia et al.(2017)Festuccia, Gonzalez, Owens, and Navarro, Bellec et al.(2018)Bellec, Radulescu, and Lagha, Zaidi et al.(2018)Zaidi, Nickerson, Imbalzano, Lian, Stein, and Stein, Teves et al.(2016)Teves, An, Hansen, Xie, Darzacq, and Tjian]. Despite several proposals, the actual mechanism and nature of these "bookmarks" have not yet been identified. [John and Workman(1998)] suggested that condensed mitotic chromosomes can act as bookmarks, some histone modifications were suggested to serve as these bookmarks [Wang and Higgins(2013), Kouskouti and Talianidis(2005), Chow et al.(2005)Chow, Georgiou, Szutorisz, Maia e Silva, Pombo, Barahona et al.], and some transcription factors (TFs) have also been identified as potential bookmarks [Dey et al.(2000)Dey, Ellenberg, Farina, Coleman, Maruyama, Sciortino et al., Kadauke

et al.(2012)Kadauke, Udugama, Pawlicki, Achtman, Jain, Cheng et al., Xing et al.(2005)Xing, Wilkerson, Mayhew, Lubert, Skaggs, Goodson et al., Christova and Oelgeschläger(2001), Festuccia et al.(2016)Festuccia, Dubois, Vandormael-Pournin, Tejeda, Mouren, Bessonnard et al.].

Recently, [Kang et al.(2020)Kang, Shokhirev, Xu, Chandran, Dixon, and Hetzer] suggested that histone 3 methylation or trimethylation at lysine 4 (H3K4me1 and H3K4me3, respectively) can act as a "bookmark" to identify genes to be transcribed, and that a limited number of TFs might act as bookmarks. However, there has been no comprehensive search of candidate "bookmark" TFs based on large-scale datasets.

We here propose a novel computational approach to search for TFs that might act as "bookmarks" during mitosis, which involves tensor decomposition (TD)-based unsupervised feature extraction (FE) (Fig. 1). In brief, after fragmenting the whole genome into DNA regions of 25,000 nucleotide, the histone modifications within each region were summed. In this context, each DNA region is considered a tensor and various singular-value vectors associated with either the DNA region or experimental conditions (e.g., histone modification, cell line, and cell division phase) are derived. After investigating singular-value vectors attributed to various experimental conditions, the DNA regions with significant associations of singular-value vectors attributed to various experimental conditions were selected as potentially biologically relevant regions. The genes included in the selected DNA regions were then identified and uploaded to the enrichment server Enrichr to identify TFs that target the genes. To our knowledge, this is the first method utilizing a TD-based unsupervised FE approach in a fully unsupervised fashion to comprehensively search for possible candidate bookmark TFs.

2 Materials and Methods

2.1 Histone modification

The whole-genome histone modification profile was downloaded from the Gene Expression Omnibus (GEO) GSE141081 dataset. Sixty individual files (with extension .bw) were extracted from the raw GEO file. After excluding six CCCTC-binding factor (CTCF) chromatin immunoprecipitation-sequencing files and six 3rd replicates of histone modification files, a total of 48 histone modification profiles were retained for analysis. The DNA sequences of each chromosome were divided into 25,000-bp regions. Note that the last DNA region of each chromosome may be shorter since the total nucleotide length does not always divide into equal regions of 25,000. Histone modifications were then summed in each DNA region, which was used as the input value for the analysis. In total, $N = 123,817$ DNA regions were available for analysis. Thus, with approximately $120,000$ regions of $25,000$ bp each, we covered the approximate human genome length of 3×10^9.

2.2 Tensor Data Representation

Histone modification profiles were formatted as a tensor, $x_{ijkms} \in \mathbb{R}^{N\times 2\times 4\times 3\times 2}$, which corresponds to the kth histone modification ($k = 1$: acetylation, H3K27ac; $k = 2$: H3K4me1; $k = 3$: H3K4me3; and $k = 4$:Input) at the ith DNA region of the jth cell line ($j = 1$: RPE1 and $j = 2$: USO2) at the mth phase of the cell cycle($m = 1$: interphase, $m = 2$: prometaphase, and $m = 3$: anaphase/telophase) of the sth replicate ($s = 1, 2$). x_{ijkms} was normalized as $\sum_i x_{ijkms} = 0$ and $\sum_i x^2_{ijkms} = N$ (Table 1). There are two biological replicates for each of the combinations of one of cell lines (either RPE1 or USO2), one of ChIP-seq (either acetylation or H3Kme1 or H3Kme4 or inout), and one of three cell cycle phases.

2.3 Tensor Decomposition

Higher-order singular value decomposition (HOSVD) [Taguchi(2020)] was applied to x_{ijkms} to obtain the decomposition

$$x_{ijkms} = \sum_{\ell_1=1}^{2} \sum_{\ell_2=1}^{4} \sum_{\ell_3=1}^{3} \sum_{\ell_4=1}^{2} \sum_{\ell_5=1}^{N} G(\ell_1 \ell_2 \ell_3 \ell_4 \ell_5) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 s} u_{\ell_5 i}, \tag{1}$$

where $G \in \mathbb{R}^{2\times4\times3\times2\times N}$ is the core tensor, and $u_{\ell_1 j} \in \mathbb{R}^{2\times2}, u_{\ell_2 k} \in \mathbb{R}^{4\times4}, u_{\ell_3 m} \in \mathbb{R}^{3\times3}, u_{\ell_4 s} \in \mathbb{R}^{2\times2}$, and $u_{\ell_5 i} \in \mathbb{R}^{N\times N}$ are singular-value vector matrices, which are all orthogonal matrices. The reason for using the complete representation instead of the truncated representation of TD is that we employed HOSVD to compute TD. In HOSVD, the truncated representation is equal to that of the complete representation; i.e., $u_{\ell_1 j}, u_{\ell_2 k}, u_{\ell_3 m}$, and $u_{\ell_4 s}$ are not altered between the truncated and the full representation. For more details, see [Taguchi(2020)].

Here is a summary on how to compute eq. (1) using the HOSVD algorithm, although it has been described in detail previously [Taguchi(2020)]. At first, x_{ijkms} is unfolded to a matrix, $x_{i(jkms)} \in \mathbb{R}^{N\times48}$. Then SVD is applied to get

$$x_{i(jkms)} = \sum_{\ell_5=1}^{N} u_{\ell_5 i} \lambda_{\ell_5} v_{\ell_5 jmks} \tag{2}$$

Then, only $u_{\ell_5 i}$ is retained, and $v_{\ell_5,jmks}$ is discarded. Similar procedures are applied to x_{ijkms} by replacing i with one of j, k, m, s in order to get $u_{\ell_1 j}, u_{\ell_2 k}, u_{\ell_3 m}, u_{\ell_4 s}$. Finally, G can be computed as

$$G(\ell_1 \ell_2 \ell_3 \ell_4 \ell_5) = \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{4} \sum_{m=1}^{3} \sum_{s=1}^{2} x_{ijmks} u_{\ell_5 i} u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 s} \tag{3}$$

2.4 TD-based unsupervised FE

Although the method was fully described in a recently published book [Taguchi(2020)], we summarize the process of selecting genes starting from the TD.

- To identify which singular value vectors attributed to samples (e.g., cell lines, type of histone modification, cell cycle phase, and replicates) are associated with the desired properties (e.g., "not dependent upon replicates or cell lines," "represents re-activation," and "distinct between input and histone modifications"), the number of singular value vectors selected are not decided in advance, since there is no way to know how singular value vectors behave in advance, because of the unsupervised nature of TD.

- To identify which singular value vectors attributed to genomic regions are associated with the desired properties described above, core tensor, G, is investigated. We select singular value vectors attributed to genomic regions that share G with larger absolute values with the singular value vectors selected in the process mentioned earlier, because these singular value vectors attributed to genomic regions are likely associated with the desired properties.

- Using the selected singular value vectors attributed to genomic regions, those associated with the components of singular value vectors with larger absolute values are selected, because such genomic regions are likely associated with the desired properties. Usually, singular value vectors attributed to genomic regions

are assumed to obey Gaussian distribution (null hypothesis), and P-values are attributed to individual genomic regions. P-values are corrected using multiple comparison correction, and the genomic regions associated with adjusted P-values less than the threshold value are selected.

- There are no definite ways to select singular value vectors. The evaluation can only be done using the selected genes. If the selected genes are not reasonable, alternative selection of singular value vectors should be attempted. When we cannot get any reasonable genes, we abort the procedure.

To select the DNA regions of interest (i.e., those associated with transcription reactivation), we first needed to specify the singular-value vectors that are attributed to the cell line, histone modification, phases of the cell cycle, and replicates with respect to the biological feature of interest, transcription reactivation. Consider selection of a specific index set $\ell_1, \ell_2, \ell_3, \ell_4$ as one that is associated with biological features of interest, we then select ℓ_5 that is associated with G with larger absolute values, since singular-value vectors $u_{\ell_5 i}$ with ℓ_5 represent the degree of association between individual DNA regions and reactivation. Using ℓ_5, we attribute P-values to the ith DNA region assuming that $u_{\ell_5 i}$ obeys a Gaussian distribution (null hypothesis) using the χ^2 distribution

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_5 i}}{\sigma_{\ell_5}}\right)^2\right], \tag{4}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution in which the argument is larger than x, and σ_{ℓ_5} is the standard deviation. P-values are then corrected by the BH criterion [Taguchi(2020)], and the ith DNA region associated with adjusted P-values less than 0.01 were selected as those significantly associated with transcription reactivation.

Algorithm displayed with mathematical formulas can be available in Fig. 2.

2.5 Enrichment analysis

Gene symbols included in the selected DNA regions were retrieved using the biomaRt package [Durinck et al.(2009)Durinck, Spellman, Birney, and Huber] of R [R Core Team(2019)] based on the hg19 reference genome. The selected gene symbols were then uploaded to Enrichr [Kuleshov et al.(2016)Kuleshov, Jones, Rouillard, Fernandez, Duan, Wang et al.] for functional annotation to identify their targeting TFs.

2.6 DESeq2

When DESeq2 [Love et al.(2014)Love, Huber, and Anders] was applied to the present data set, six samples within each cell lines measured for three cell cycles and associated with two replicates were considered. Three cell cycles were regarded to be categorical classes associated with no rank order since we would like to detect not monotonic change between cell cycles but re-activation during them. All other parameters are defaults. Counts less than 1.0 were truncated so as to have integer values (e.g., 1400.53 was converted to 1400).

2.7 csaw

Since csaw [Lun and Smyth(2015)] required bam files not available in GEO, we first mapped 60 fastq files to hg38 human genome using bowtie2 [Langmead and Salzberg(2012)] where 60 fastq files in GEO ID GSE141081 were downloaded from SRA. Sam files generated by bowtie2 were converted and indexed by samtools [Li et al.(2009)Li, Handsaker, Wysoker, Fennell, Ruan, Homer et al.] and sorted bam files

were generated. Generated bam files that correspond to individual combinations of cell lines and ChIP-seq were loaded into csaw in order to identify differential binding among three cell cycle phases.

2.8 Identification of overlapping regions between peak call

We retrieved 36 peak call data set (with extension peaks.txt.gz) that correspond to 48 Chip-Seq files with excluding 12 input files. Starting from these 48 peak call files, using `findOverlapsOfPeaks` function included in ChIPpeakAnno package in R, we selected overlap regions step by step as follows

- Identify overlap regions between two biological replicates; this results in 9 regions for U2OS cell lines and RPE1 cell lines, respectively, in total 18 peak calls.
- Identify overlap regions among three cell cycles; retrieve regions commonly expressed in three cell cycle phases for H3K4me1 and H3K4me3 whereas those expressed only in interphase and anaphase/telophase; this results in three regions, each of which was attributed to H3K4me1, H3K27ac, or H3K4me3, for U2OS cell lines and RPE1 cell lines, respectively, in total 6 peak calls.
- Identify overlap between 6 peak calls.

This process was illustrated in Fig. 3.

3 Results and Discussion

We first attempted to identify which singular-value vector is most strongly attributed to transcription reactivation among the vectors for cell line ($u_{\ell_1 j}$), histone modification ($u_{\ell_2 k}$), cell cycle phase ($u_{\ell_3 m}$), and replicate ($u_{\ell_4 s}$) (Fig. 4). First, we considered phase dependency. Fig. 5 shows the singular-value vectors $u_{\ell_3 m}$ attributed to cell cycle phases. In the case that there are a set of genes that share some dependence, singular value vectors reflect their mean behaviour. Specifically, singular value vectors act as some kind of pseudo representative genes. Thus, by investigating singular value vectors, we can find what kind of cell cycle dependence can appear in the group of genes. Since the reactivation means that being expressive in inter and ana/telophases whereas not expressive in prometapahse, singular value vectors supposed to be related to be reactivation take opposite signs between inter/ana/telophased and prometaphase. Thus, u_{3m} are most likely associated with reactivation. Although u_{2m} and u_{3m} were associated with reactivation, we further considered only u_{3m} since it showed a more pronounced reactivation profile. Next, we investigated singular-value vectors $u_{\ell_2 m}$ attributed to histone modification (Fig. 6). There was no clearly interpretable dependence on histone modification other than for u_{1k}, which represents the lack of histone modification, since the values for H3K27ac, H3K4me1, and H3K4me3 were equivalent to the Input value that corresponds to the control condition; thus, u_{2k}, u_{3k}, and u_{4k} were considered to have equal contributions for subsequent analyses. By contrast, since u_{1j} and u_{1s} showed no dependence on cell line and replicates, respectively, we selected these vectors for further downstream analyses (Fig. 7).

Finally, we evaluated which vector $u_{\ell_5 i}$ had a larger $\sum_{\ell_2=2}^{4} |G(1, \ell_2, 3, 1, \ell_5)|^{\alpha}, \alpha = 1, 2, 3$ (Fig. 8); in this case, we calculated the squared sum for $2 \leq \ell_2 \leq 4$ to consider them equally. Although we do not have any definite criterion to decide α uniquely, since $\ell_5 = 4$ always takes largest values for $\alpha \geq 1$, $\ell_5 = 4$ was further employed. The P-values attributed to the ith DNA regions were calculated using eq. (4), resulting in selection of 507 DNA regions associated with adjusted P-values less than 0.01.

We next checked whether histone modification in the selected DNA regions was associated with the following transcription reactivation properties:

1. H3K27ac should have larger values in interphase and anaphase/telophase than in prometaphase, as the definition of reactivation.

2. H3K4me1 and H3K4me3 should have constant values during all phases of the cell cycle, as the definition of a "bookmark" histone modification

3. H3K4me1 and H3K4me3 should have larger values than the Input; otherwise, they cannot be regarded to act as "bookmarks" since these histones must be significantly modified throughout these phases.

To check whether the above criteria are fulfilled, we applied six t tests to histone modifications in the 507 selected DNA regions (Table 2). The results clearly showed that histone modifications in the 507 selected DNA regions satisfied the requirements for transcription reactivation; thus, our strategy could successfully select DNA regions that demonstrate reactivation/bookmark functions of histone modification.

After confirming that selected DNA regions are associated with targeted reactivation/bookmark features, we queried all gene symbols contained within these 507 regions to the Enrichr server to identify TFs that significantly target these genes. These TFs were considered candidate bookmarks that remain bound to these DNA regions throughout the cell cycle and trigger reactivation in anaphase/telophase (i.e., after cell division is complete). Table 3 lists the TFs associated with the selected regions at adjusted P-values less than 0.05 in each of the seven categories of Enrichr.

Among the many TFs that emerged to be significantly likely to target genes included in the 507 DNA regions selected by TD-based unsupervised FE, we here focus on the biological functions of TFs that were also detected in the original study suggesting that TFs might function as histone modification bookmarks for transcription reactivation [Kang et al.(2020)Kang, Shokhirev, Xu, Chandran, Dixon, and Hetzer]. RUNX was identified as an essential TF for osteogenic cell fate, and has been associated with mitotic chromosomes in multiple cell lines, including Saos-2 osteosarcoma cells and HeLa cells (Young et al. 2007). Table 4 shows the detection of RUNX family TFs in seven TF-related categories of Enrichr; three RUNX TFs were detected in at least one of the seven TF-related categories. In addition, TEADs (Kegelman et al. 2018), JUNs [Wagner(2002)], FOXOs [Rached et al.(2010)Rached, Kode, Xu, Yoshikawa, Paik, DePinho et al.], and FosLs citepKang01072020 were reported to regulate osteoblast differentiation. Tables 5, 6,7, and 8 show that two TEAD TFs, three JUN TFs, four FOXO TFs, and two FOSL TFs were detected in at least one of the seven TF-related categories in Enrichr, respectively.

Other than these five TF families reported in the original study [Kang et al.(2020)Kang, Shokhirev, Xu, Chandran, Dixon, and Hetzer], the TFs detected most frequently within seven TF-related categories in Enrichr were as follows (Table 9): GATA2 [Kala et al.(2009)Kala, Haugas, Lilleväli, Guimera, Wurst, Salminen et al.], ESR1 [Kato and Ogawa(1994)], TCF21 [Kim et al.(2017)Kim, Pjanic, Nguyen, Miller, Iyer, Liu et al.], TP53 [Ha et al.(2007)Ha, Baek, Kim, Jeong, Kim, McKeon et al.], WT1 [Shandilya and Roberts(2015)], NFE2L2 (also known as NRF2 [Martin-Hurtado et al.(2019)Martin-Hurtado, Martin-Morales, Robledinos-Antón, Blanco, Palacios-Blanco, Lastres-Becker et al.]), GATA1 [Kadauke et al.(2012)Kadauke, Udugama, Pawlicki, Achtman, Jain, Cheng et al.], and GATA3 [Shafer et al.(2017)Shafer, Nguyen, Tremblay, Viala, Béland, Bertos et al.]. All of these TFs have been reported to be related to mitosis directly or indirectly, in addition to JUN and JUND, which are listed in Table 6. This further suggests the suitability of our search strategy to identify transcription reactivation bookmarks.

One might wonder why we did not compare our methods with the other methods. As can be seen in Table 1, there are only two samples each in as many as 24 categories. Therefore, it is difficult to apply standard statistical tests for pairwise comparisons between two groups including only two samples. In addition, the number of features, N, which is the number of genomic regions in this study, is as many as 1,23,817, which drastically reduces the significance of each test if we consider multiple comparison criteria that increase P-values that reject the null hypothesis. Finally, only a limited number of pairwise comparisons are meaningful; for example, we are not willing to compare the amount of H3K4me1 in the RPE1 cell line at interphase with that of H3K27ac in the U2OS cell line at prometaphase. Therefore, usual procedures that deal with pairwise comparisons comprehensively, such as Tukey's test, cannot be applied to the present data set as it is. In conclusion, we could not find any suitable method applicable to the present data set that has a small number of samples within each of as many as 24 categories, whereas the number of features is as many as 1,23,817.

In order to demonstrate inferiority of other method compared with our method, we applied DESeq2 [Love et al.(2014)Love, Huber, and Anders] to the present data set, although DESeq2 was designed to not ChIP-seq but RNA-seq. The outcome is disappointing as expected (Table 10) if it is compared with Table 2. First of all, there are no coincidences between two cell lines. Although there are as many as 4227 regions within which H3K4me1 is distinct among three cell cycle phases when RPE1 is considered, there were no regions associated with distinct H3K4me1 when U2OS was considered. In addition to this, although only H3K27ac among three histone modifications measured is expected to be distinct during three cell cycle phases, other histone modifications are sometimes detected as distinct during three cell cycle phases. Finally, the number of genomic regions considered in each comparison varies, since DESeq2 automatically discarded regions associated with low variance among distinct classes. The reason why there are no regions associated with distinct histone modification for Input and H3K4me1 when RPE1 was considered is definitely because almost all genomic regions were considered for these two comparisons; too many comparisons increase the P-values because of multiple comparison corrections. On the other hand, our proposed TD based unsupervised FE can deal with all of the genomic regions, which resulted in more stable outcomes. Thus, it is obvious that DESeq2 was inferior to TD based unsupervised FE when it is applied to the present data set.

One might still wonder if it is because of usage of DESeq2 not designed specific to ChIP-seq data. In order to confirm this point, we sought integrated approaches designed specific to treatment of ChIP-seq data. In addition, we need some approaches that enable us not only pairwise comparison but also comparisons among more than two categories, since we have to compare among three cell cycle phases, i.e., terphase, prometaphase, and anaphase/telophase. There are not so many approaches satisfying these conditions [Wu et al.(2015)Wu, Bittencourt, Stallcup, and Siegmund, Steinhauser et al.(2016)Steinhauser, Kurzawa, Eils, and Herrmann, Tu and Shao(2017)]. For example, although DBChIP [Liang and Keleş(2011)] was designed to treat ChIP-seq data set, since it was designed to be specific to TF binding, it required to input single nucleotide positions where binding proteins bind, Thus, it is not applicable to histone modification measurements where not binding points but binding regions are provided. On the other hand, although DiffBind [Stark and Brown(2011)] was designed to deal with histone modification, it can accept only pairwise comparisions. SCIFER [Xu et al.(2014)Xu, Grullon, Ge, and Peng] can identify enrichment within single measurement compared with input experiment, MACS2 which is modified version of MACS [Zhang et al.(2008)Zhang, Liu, Meyer, Eeckhoute, Johnson, Bernstein et al.], can also accept only pairwise comaprisons, ODIN [Allhoff et al.(2014)Allhoff, Seré, Chauvistré, Lin, Zenke, and Costa] also can accept only pairwise comparisons,

RSEG [Song and Smith(2011)] also can accept only pairwise comparisons, MAnorm [Shao et al.(2012)Shao, Zhang, Yuan, Orkin, and Waxman] also can accept only pairwise comparisons, HOMER [Heinz et al.(2010)Heinz, Benner, Spann, Bertolino, Lin, Laslo et al.] also can accept only pairwise comparisons, QChIPat [Liu et al.(2013)Liu, Yi, SV, Lan, Ma, Huang et al.] also can accept only pairwise comparisons, diffReps [Shen et al.(2013)Shen, Shao, Liu, Maze, Feng, and Nestler] also can accept only pairwise comparisons, MMDiff [Schweikert et al.(2013)Schweikert, Cseke, Clouaire, Bird, and Sanguinetti] also can accept only pairwise comparisons, PePr [Zhang et al.(2014)Zhang, Lin, Johnson, Rozek, and Sartor] does not perform even pairwise comparison. ChIPComp [Chen et al.(2015)Chen, Wang, Qin, and Wu] was tested toward only pairwise comparisons when it was applied to real data set. Although MultiGPS [Mahony et al.(2014)Mahony, Edwards, Mazzoni, Sherwood, Kakumanu, Morrison et al.] can deal with multiple files, they must be composed of condition A and its corresponding input vs condition B and its corresponding input, it cannot be applied to the present case composed of three cell cycle phases and their corresponding inputs. Thus as far as we investigated there are no approaches designed to be applicable to three independent conditions, each of which is composed of a pair of treated and input experiments.

This difficulty is because of two kinds of distinct differential binding analyses required (Fig. 9), one of which is the comparison between treated and input experiments and another of which is the comparison between two experimental conditions (e.g., patients versus healthy control, two different tissues) whereas they are easily performed in tensor representation as shown in the above. Nevertheless, in order to emphasize the inferiority of ChIP-seq specific pipeline aiming differential binding analysis toward TD based unsupervised FE, we considered csaw [Lun and Smyth(2015)] as a representative since it accepts, at least, not pairwise but comparisons among multiple conditions as performed by DESeq2 (Table 10). Table 11 shows the results. It is very disappointing as expected. For example, although H3K27ac is expected to support reactivation, differential binding region among distinct cell cycle phases in U2OS cell line is almost none (only 0.1 % of whole tested regions). Although H3K4me3 should not distinctly bind to chromosome among thee cell cycles since it is expected to play a role of bookmark, it distinctly binds to chromosomes among three cell cycle phases for two cell lines. These behaviours are very contrast to those in Table 2 which exhibits the expected differential/undifferential binding to chromosome. Thus, in conclusion, even if we employ pipelines specifically designed to ChIP-Seq data analyses, they cannot outperform the results obtained by TD based unsupervised FE.

Since one might wonder why we have specifically used region length of 25,000 nucleic acid length, we discuss about it as follows.

- We have successfully used the region length [49, 50]. When started to employ this procedure, we tried multiple values and identified that it is most successful.
- Optimizing region length from studies to studies is not a good way to identify something biological. Region length should not be optimization parameters. If the optimal region length vary from studies to studies, we need to rationalize it. Nevertheless, the fact that employment region of 25,000 nucleic acid length was successful in three independent studies (including the present one) definitely suggest that this choice is reasonable.
- We expected that each region is coincident with one gene in average. Since the number of selected regions, 507, is almost equivalent to the number of gene symbols in these 507 regions, 525 (see supplementary materials), employment of region of 25,000 nucleic acid length seems to be reasonable.

- Since average gene length on human genome is $\sim 3 \times 10^4$, the selection of region of 25,000 nucleic acid length is supposed to be association of one gene in each region. As denoted in the above, this expectation was fulfilled.

Although the above discussion might be enough to rationalize the usage of region of 25,000 nucleic acid length, we tried an alternative strategy as described in Materials and Methods. We downloaded peak call data set from GEO and tried to identify overlaps between peak regions. As a result, we could find only 22 regions of mean length of 5000 nucleic acid, with which only 13 gene symbols were associated. This tells us two things. Smaller region length, 5,000, results in regions without gene symbols. Shorter region length reduces the number of commonly identified regions between multiple experiments. This prevents us from performing downstream analysis. This failure of an alternative approach definitely suggests the suitability of the selection of region of 25,000 nucleic acid length.

One might also wonder if TFs can also work in cell line specific ways; thus there might be no reasons to select TFs common between two cell lines. It is really true that TFs can work in cell line specific ways; nevertheless, what we are interested in is a more robust bookmark that can likely work in mitoic process universally. If we selected TFs that work in cell line specific manner, it reduces the possibility that selected TFs work universally in mitoic process. The reason why we validated the selected genes based upon Enrichr that might include the results for other cell lines than U2OS and RPE1 is similar; if the selected genes are coincident with data bases retrieved from other cell lines, results are more unlikely accidental and are more likely robust and universal.

In this study, reliability of selected genes was evaluated by enrichment analysis. Since we have selected very small amount of genes, as small as c.a. 500, it is very unlikely for them to be associated with numerous enrichment. In spite of that, since our selected genes are associated with so many TF activities, we can assume that our selection of genes are reasonable. In the case that we cannot find any enrichment, we regard that our selection of singular value vectors is failure and we try to check if other selections can work better or not. It is worth noting that because other methods are not designed to deal with the studied problem, applying these methods generate inferior outcomes.

We show that selected TFs are expressive in cell lines as follows: First of all, we evaluated TFs by not only binding to genome but also co-occurrence with selected genes (e.g. (III) and (V) in Table 3). Thus, it is very likely that some TFs are expressive in cell lines where the selected genes are expressive. Second, we seek GEO Profiles in order to see if these TFs are expressive in U2OS cell lines and RPE cells. Then, we have found that almost all TFs were expressive in both U2OS cell lines and RPE cells in GEO profiles (see Additional file 4). Thus, it is not unreasonable to expect the expression of these TFs in two cell lines used in this study.

4 Conclusions

We applied a novel TD-based unsupervised FE method to various histone modifications across the whole human genome, and the levels of these modifications were measured during mitotic cell division to identify genes that are significantly associated with histone modifications. Potential bookmark TFs were identified by searching for TFs that target the selected genes. The TFs identified were functionally related to the cell division cycle, suggesting their potential as bookmark TFs that warrant further exploration.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

YT planned and performed the study. YT and TT discussed the results and wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by KAKENHI 19H05270, 20K12067, 20H04848. This project was also funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. KEP-8- 611-38. The authors, therefore, acknowledge DSR with thanks for providing technical and financial support

Acknowledgments

This manuscript will be released as a pre-print at BioRxiv.

Supplemental Data

Additional file 1: Genes identified by TD-based unsupervised FE; Additional file 2: Potential TFs that target identified genes (in Additional file 1) identified by Enrichr; Additional file 3: Sample R code used in the analyses performed in this study. Additional file 4: Expression of TFs selected in this study in GEO profiles.

Data Availability Statement

All datasets analyzed in this study were obtained from GEO: GSE141139

References

1. Festuccia N, Gonzalez I, Owens N, Navarro P. Mitotic bookmarking in development and stem cells. Development. 2017;144(20):3633–3645. doi:10.1242/dev.146522.

2. Bellec M, Radulescu O, Lagha M. Remembering the past: Mitotic bookmarking in a developing embryo. Current Opinion in Systems Biology. 2018;11:41 – 49. doi:https://doi.org/10.1016/j.coisb.2018.08.003.

3. Zaidi SK, Nickerson JA, Imbalzano AN, Lian JB, Stein JL, Stein GS. Mitotic Gene Bookmarking: An Epigenetic Program to Maintain Normal and Cancer Phenotypes. Molecular Cancer Research. 2018;16(11):1617–1624. doi:10.1158/1541-7786.MCR-18-0415.

4. Teves SS, An L, Hansen AS, Xie L, Darzacq X, Tjian R. A dynamic mode of mitotic bookmarking by transcription factors. eLife. 2016;5:e22280. doi:10.7554/eLife.22280.

5. John S, Workman JL. Bookmarking genes for activation in condensed mitotic chromosomes. BioEssays. 1998;20(4):275–279. doi:10.1002/(SICI)1521-1878(199804)20:4¡275::AID-BIES1¿3.0.CO;2-P.

6. Wang F, Higgins JMG. Histone modifications and mitosis: countermarks, landmarks, and bookmarks. Trends in Cell Biology. 2013;23(4):175–184. doi:10.1016/j.tcb.2012.11.005.

7. Kouskouti A, Talianidis I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. The EMBO Journal. 2005;24(2):347–357. doi:10.1038/sj.emboj.7600516.

8. Chow CM, Georgiou A, Szutorisz H, Maia e Silva A, Pombo A, Barahona I, et al. Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. EMBO reports. 2005;6(4):354–360. doi:10.1038/sj.embor.7400366.

9. Dey A, Ellenberg J, Farina A, Coleman AE, Maruyama T, Sciortino S, et al. A Bromodomain Protein, MCAP, Associates with Mitotic Chromosomes and Affects G2-to-M Transition. Molecular and Cellular Biology. 2000;20(17):6537–6549. doi:10.1128/MCB.20.17.6537-6549.2000.

10. Kadauke S, Udugama MI, Pawlicki JM, Achtman JC, Jain DP, Cheng Y, et al. Tissue-Specific Mitotic Bookmarking by Hematopoietic Transcription Factor GATA1. Cell. 2012;150(4):725–737. doi:10.1016/j.cell.2012.06.038.

11. Xing H, Wilkerson DC, Mayhew CN, Lubert EJ, Skaggs HS, Goodson ML, et al. Mechanism of hsp70i Gene Bookmarking. Science. 2005;307(5708):421–423. doi:10.1126/science.1106478.

12. Christova R, Oelgeschläger T. Association of human TFIID–promoter complexes with silenced mitotic chromatin in vivo. Nature Cell Biology. 2001;4(1):79–82. doi:10.1038/ncb733.

13. Festuccia N, Dubois A, Vandormael-Pournin S, Tejeda EG, Mouren A, Bessonnard S, et al. Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network. Nature Cell Biology. 2016;18(11):1139–1148. doi:10.1038/ncb3418.

14. Kang H, Shokhirev MN, Xu Z, Chandran S, Dixon JR, Hetzer MW. Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. Genes & Development. 2020;34(13-14):913–930. doi:10.1101/gad.335794.119.

15. Taguchi YH. Unsupervised Feature Extraction Applied to Bioinformatics. Springer International Publishing; 2020. Available from: `https://doi.org/10.1007/978-3-030-22456-1`.

16. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols. 2009;4:1184–1191.

17. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: `https://www.R-project.org/`.

18. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016;44(W1):W90–W97. doi:10.1093/nar/gkw377.

19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12). doi:10.1186/s13059-014-0550-8.

20. Lun ATL, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Research. 2015;44(5):e45–e45. doi:10.1093/nar/gkv1191.

21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357–359. doi:10.1038/nmeth.1923.

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

23. Wagner EF. Functions of AP1 (Fos/Jun) in bone development. Annals of the Rheumatic Diseases. 2002;61(suppl 2):ii40–ii42. doi:10.1136/ard.61.suppl_2.ii40.

24. Rached MT, Kode A, Xu L, Yoshikawa Y, Paik JH, DePinho RA, et al. FoxO1 Is a Positive Regulator of Bone Formation by Favoring Protein Synthesis and Resistance to Oxidative Stress in Osteoblasts. Cell Metabolism. 2010;11(2):147–160. doi:10.1016/j.cmet.2010.01.001.

25. Kala K, Haugas M, Lilleväli K, Guimera J, Wurst W, Salminen M, et al. Gata2 is a tissue-specific post-mitotic selector gene for midbrain GABAergic neurons. Development. 2009;136(2):253–262. doi:10.1242/dev.029900.

26. Kato R, Ogawa H. An essential gene, ESR1, is required for mitotic growth, DNA repair and meiotic recombination Saccharomyces cerevisiae. Nucleic Acids Research. 1994;22(15):3104–3112. doi:10.1093/nar/22.15.3104.

27. Kim JB, Pjanic M, Nguyen T, Miller CL, Iyer D, Liu B, et al. TCF21 and the environmental sensor aryl-hydrocarbon receptor cooperate to activate a pro-inflammatory gene expression program in coronary artery smooth muscle cells. PLOS Genetics. 2017;13(5):1–29. doi:10.1371/journal.pgen.1006750.

28. Ha GH, Baek KH, Kim HS, Jeong SJ, Kim CM, McKeon F, et al. p53 Activation in Response to Mitotic Spindle Damage Requires Signaling via BubR1-Mediated Phosphorylation. Cancer Research. 2007;67(15):7155–7164. doi:10.1158/0008-5472.CAN-06-3392.

29. Shandilya J, Roberts SG. A role of WT1 in cell division and genomic stability. Cell Cycle. 2015;14(9):1358–1364. doi:10.1080/15384101.2015.1021525.

30. Martin-Hurtado A, Martin-Morales R, Robledinos-Antón N, Blanco R, Palacios-Blanco I, Lastres-Becker I, et al. NRF2-dependent gene expression promotes ciliogenesis and Hedgehog signaling. Scientific Reports. 2019;9(1). doi:10.1038/s41598-019-50356-0.

31. Shafer MER, Nguyen AHT, Tremblay M, Viala S, Béland M, Bertos NR, et al. Lineage Specification from Prostate Progenitor Cells Requires Gata3-Dependent Mitotic Spindle Orientation. Stem Cell Reports. 2017;8(4):1018–1031. doi:10.1016/j.stemcr.2017.02.004.

32. Wu DY, Bittencourt D, Stallcup MR, Siegmund KD. Identifying differential transcription factor binding in ChIP-seq. Frontiers in Genetics. 2015;6:169. doi:10.3389/fgene.2015.00169.

33. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in Bioinformatics. 2016;17(6):953–966. doi:10.1093/bib/bbv110.

34. Tu S, Shao Z. An introduction to computational tools for differential binding analysis with ChIP-seq data. Quantitative Biology. 2017;5(3):226–235. doi:10.1007/s40484-017-0111-8.

35. Liang K, Keleş S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics. 2011;28(1):121–122. doi:10.1093/bioinformatics/btr605.

36. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data; 2011.

37. Xu S, Grullon S, Ge K, Peng W. Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells. In: Methods in Molecular Biology. Springer New York; 2014. p. 97–111. Available from: `https://doi.org/10.1007/978-1-4939-0512-6_5`.

38. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137.

39. Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics. 2014;30(24):3467–3475. doi:10.1093/bioinformatics/btu722.

40. Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics. 2011;27(6):870–871. doi:10.1093/bioinformatics/btr030.

41. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biology. 2012;13(3):R16. doi:10.1186/gb-2012-13-3-r16.

42. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell. 2010;38(4):576–589. doi:10.1016/j.molcel.2010.05.004.

43. Liu B, Yi J, SV A, Lan X, Ma Y, Huang TH, et al. QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. BMC Genomics. 2013;14(Suppl 8):S3. doi:10.1186/1471-2164-14-s8-s3.

44. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. PLOS ONE. 2013;8(6):1–13. doi:10.1371/journal.pone.0065598.

45. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC Genomics. 2013;14(1):826. doi:10.1186/1471-2164-14-826.

46. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. Bioinformatics. 2014;30(18):2568–2575. doi:10.1093/bioinformatics/btu372.

47. Chen L, Wang C, Qin ZS, Wu H. A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. Bioinformatics. 2015;31(12):1889–1896. doi:10.1093/bioinformatics/btv094.

48. Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, et al. An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding. PLOS Computational Biology. 2014;10(3):1–14. doi:10.1371/journal.pcbi.1003501.

49. Taguchi Y. One-class Differential Expression Analysis using Tensor Decomposition-based Unsupervised Feature Extraction Applied to Integrated Analysis of Multiple Omics Data from 26 Lung Adenocarcinoma Cell Lines. In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE); 2017. p. 131–138.

50. Taguchi Yh, Turki T. Tensor-Decomposition-Based Unsupervised Feature Extraction Applied to Prostate Cancer Multiomics Data. Genes. 2020;11(12). doi:10.3390/genes11121493.

Figure captions

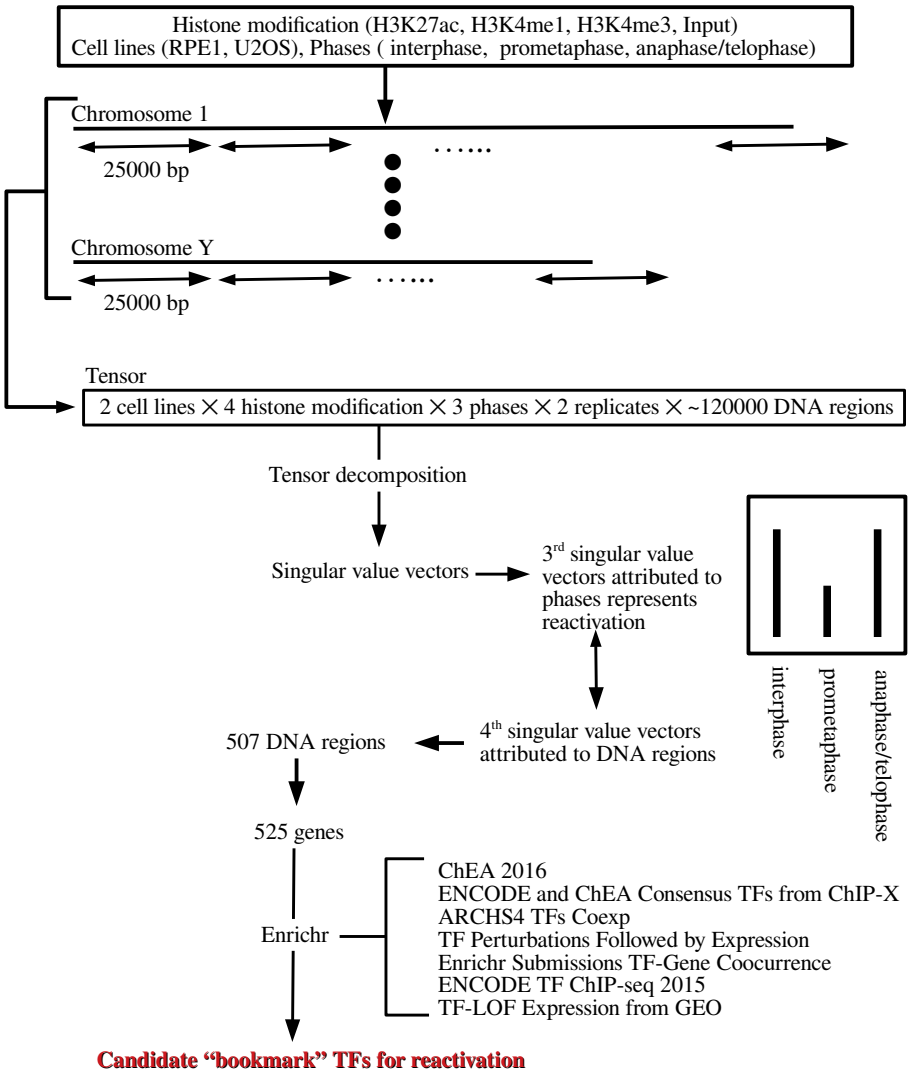

Fig 1. Flow chart of analyses performed in this study

References

Festuccia et al.(2017)Festuccia, Gonzalez, Owens, and Navarro. Festuccia N, Gonzalez I, Owens N, Navarro P. Mitotic bookmarking in development and stem cells. *Development* **144** (2017) 3633–3645. doi:10.1242/dev.146522.

Bellec et al.(2018)Bellec, Radulescu, and Lagha. Bellec M, Radulescu O, Lagha M. Remembering the past: Mitotic bookmarking in a developing embryo. *Current Opinion in Systems Biology* **11** (2018) 41 – 49. doi:https://doi.org/10.1016/j.coisb.2018.08.003.

Zaidi et al.(2018)Zaidi, Nickerson, Imbalzano, Lian, Stein, and Stein. Zaidi SK, Nickerson JA, Imbalzano AN, Lian JB, Stein JL, Stein GS. Mitotic gene bookmarking: An epigenetic program to maintain normal and cancer phenotypes.

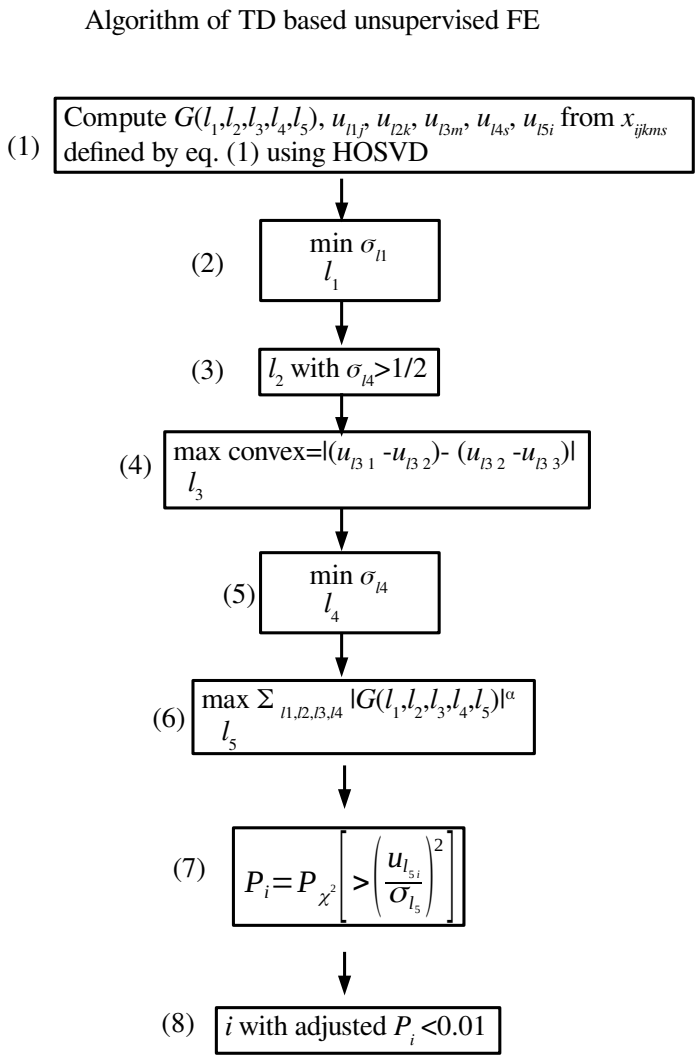

Fig 2. Algorithm of TD based unsupervised FE. (1) Perform TD to derive $G(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5)$. (2) Select $u_{\ell_1 j}$ that takes constant values between two cell lines as much as possible. (3) Select $u_{\ell_2 k}$ that has distinct values for Histone modification toward inputs. (4) Select $u_{\ell_3 m}$ that represents reactivation during three cell cycle phases as much as possible. (5) Select $u_{\ell_1 j}$ that takes constant values between two biological replicates as much as possible. (6) Select ℓ_5 associated with G having largest absolute values given $\ell_1, \ell_2, \ell_3, \ell_4$ (7) Attribute P-values to is with assuming that $u_{\ell_5 i}$ obeys Gaussian distribution (Null hypothesis). (8) Select is associated with adjusted P-values less than 0.01.

Molecular Cancer Research **16** (2018) 1617–1624. doi:10.1158/1541-7786.MCR-18-0415.

Teves et al.(2016)Teves, An, Hansen, Xie, Darzacq, and Tjian. Teves SS, An L, Hansen AS, Xie L, Darzacq X, Tjian R. A dynamic mode of mitotic bookmarking by transcription factors. *eLife* **5** (2016) e22280. doi:10.7554/eLife.22280.

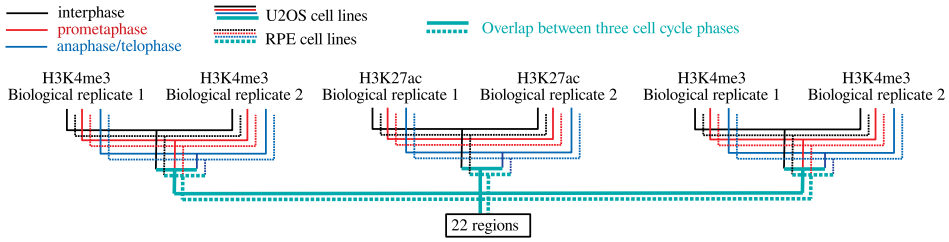

Fig 3. Flow chart on how we integrated peak call files.

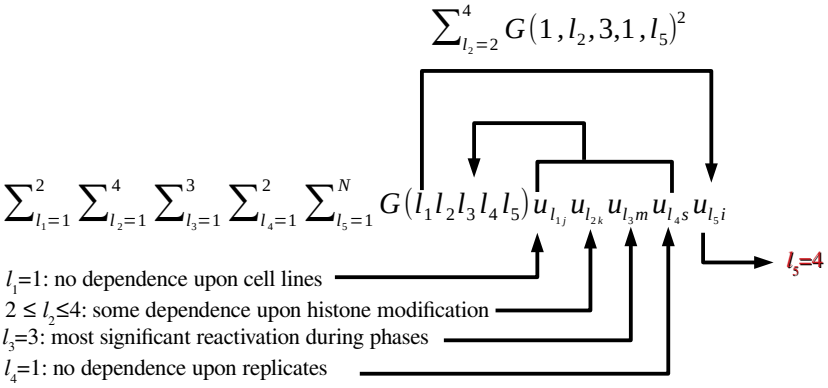

Fig 4. Schematic of the process for selecting u_{4i} to be used for DNA region selection.

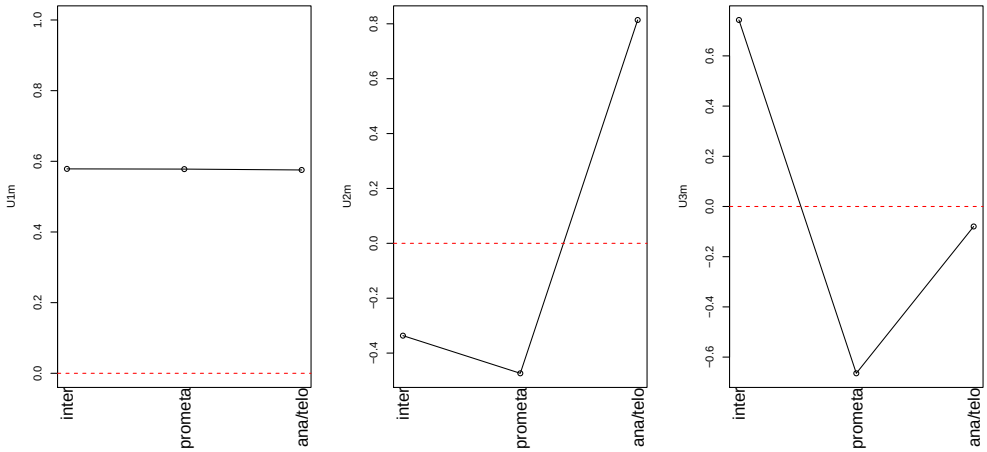

Fig 5. Singular-value vectors associated with cell cycle phase. Left: u_{1m}, middle: u_{2m}, right: u_{3m}

John and Workman(1998). John S, Workman JL. Bookmarking genes for activation in condensed mitotic chromosomes. *BioEssays* **20** (1998) 275–279. doi:10.1002/(SICI)1521-1878(199804)20:4⟨275::AID-BIES1⟩3.0.CO;2-P.

Wang and Higgins(2013). Wang F, Higgins JM. Histone modifications and mitosis: countermarks, landmarks, and bookmarks. *Trends in Cell Biology* **23** (2013) 175–184. doi:10.1016/j.tcb.2012.11.005.

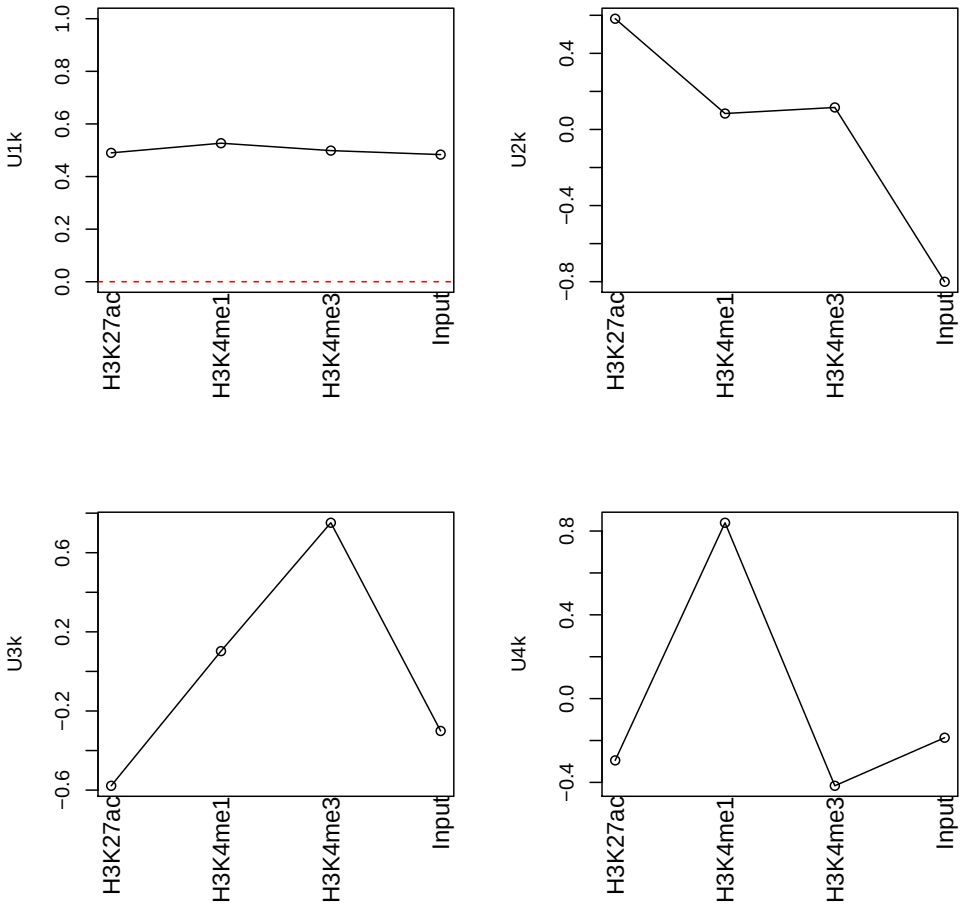

Fig 6. Singular-value vectors associated with histone modification. Upper left: u_{1k}, upper right: u_{2k}, lower left: u_{3k}, lower right: u_{4k}

Kouskouti and Talianidis(2005). Kouskouti A, Talianidis I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *The EMBO Journal* **24** (2005) 347–357. doi:10.1038/sj.emboj.7600516.

Chow et al.(2005)Chow, Georgiou, Szutorisz, Maia e Silva, Pombo, Barahona et al.. Chow CM, Georgiou A, Szutorisz H, Maia e Silva A, Pombo A, Barahona I, et al. Variant histone h3.3 marks promoters of transcriptionally active genes during mammalian cell division. *EMBO reports* **6** (2005) 354–360. doi:10.1038/sj.embor.7400366.

Dey et al.(2000)Dey, Ellenberg, Farina, Coleman, Maruyama, Sciortino et al.. Dey A, Ellenberg J, Farina A, Coleman AE, Maruyama T, Sciortino S, et al. A bromodomain protein, mcap, associates with mitotic chromosomes and affects g2-to-m transition. *Molecular and Cellular Biology* **20** (2000) 6537–6549. doi:10.1128/MCB.20.17.6537-6549.2000.

Kadauke et al.(2012)Kadauke, Udugama, Pawlicki, Achtman, Jain, Cheng et al.. Kadauke S, Udugama MI, Pawlicki JM, Achtman JC, Jain DP, Cheng Y, et al.

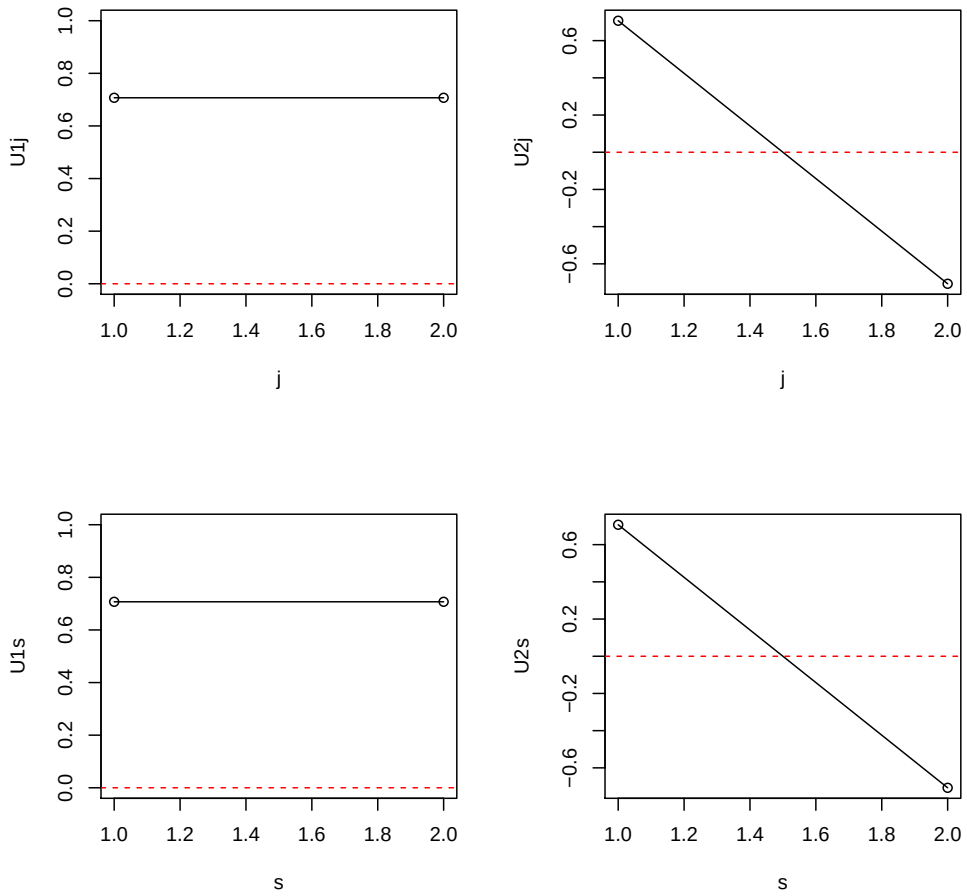

Fig 7. Dependence of vectors on cell line (j) and replicate (s). Top left: u_{1j}, top right: u_{2j}, bottom left: u_{1s}, bottom right: u_{2s}

Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. *Cell* **150** (2012) 725–737. doi:10.1016/j.cell.2012.06.038.

Xing et al.(2005)Xing, Wilkerson, Mayhew, Lubert, Skaggs, Goodson et al.. Xing H, Wilkerson DC, Mayhew CN, Lubert EJ, Skaggs HS, Goodson ML, et al. Mechanism of hsp70i gene bookmarking. *Science* **307** (2005) 421–423. doi:10.1126/science.1106478.

Christova and Oelgeschläger(2001). Christova R, Oelgeschläger T. Association of human TFIID–promoter complexes with silenced mitotic chromatin in vivo. *Nature Cell Biology* **4** (2001) 79–82. doi:10.1038/ncb733.

Festuccia et al.(2016)Festuccia, Dubois, Vandormael-Pournin, Tejeda, Mouren, Bessonnard e Festuccia N, Dubois A, Vandormael-Pournin S, Tejeda EG, Mouren A, Bessonnard S, et al. Mitotic binding of esrrb marks key regulatory regions of the pluripotency network. *Nature Cell Biology* **18** (2016) 1139–1148. doi:10.1038/ncb3418.

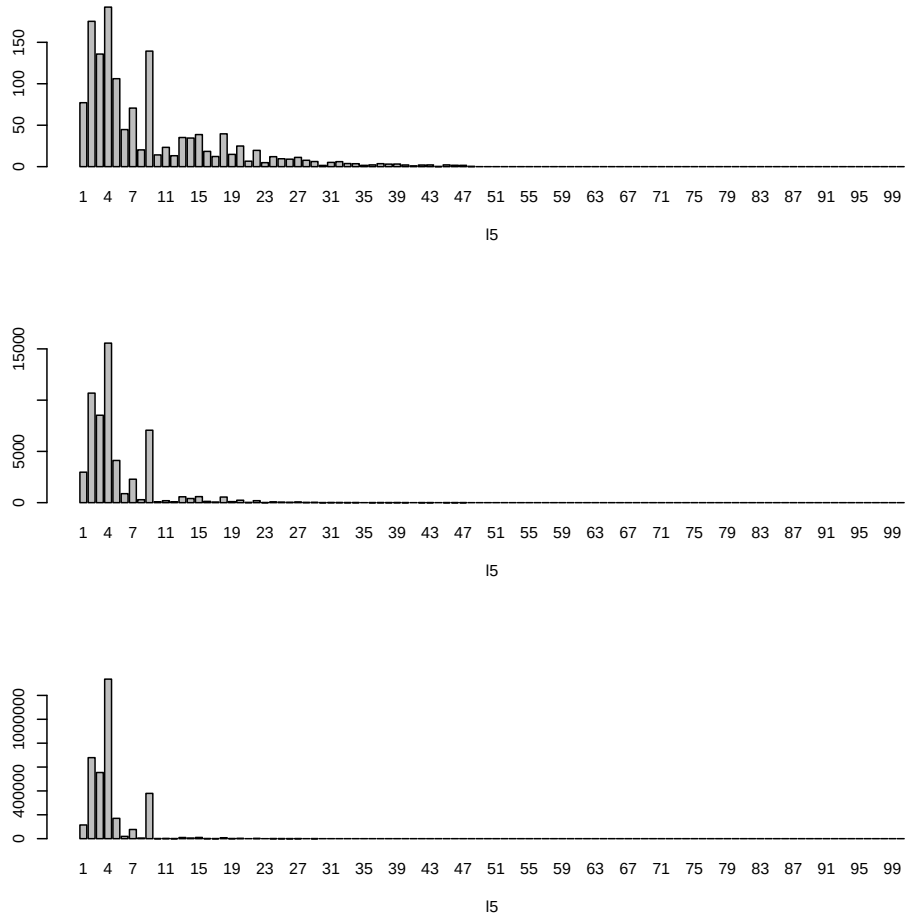

Fig 8. $\sum_{\ell_2=2}^{4} |G(1, \ell_2, 3, 1, \ell_5)|^{\alpha}, \ell_5 \leq 100$. Because of HOSVD algorithm, $G(\ell_1, \ell_2, \ell_3, \ell_4, \ell_5) = 0$ for $\ell_5 > 2 \times 4 \times 3 \times 2 = 48$. $\alpha = 1$ (Top), 2 (middle), and 3 (bottom).

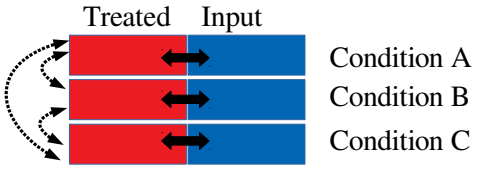

Fig 9. Schematics that illustrates the difficulty of differential binding analysis. In contrast to differential expression analysis that requires only inter conditions comparisons (displayed by broken bidirectional arrows), differential binding analysis requires additional intra conditions comparisons between treated and input experiment (displayed by bidirectional solid arrows). There are no pipelines that aim to identify differential binding considering simultaneously more than two conditions.

Kang et al.(2020)Kang, Shokhirev, Xu, Chandran, Dixon, and Hetzer. Kang H, Shokhirev MN, Xu Z, Chandran S, Dixon JR, Hetzer MW. Dynamic regulation of histone modifications and long-range chromosomal interactions during

Table 1. Combinations of experimental conditions. Individual conditions are associated with two replicates

Phases	Histone modifications / Cell lines							
	H3K27ac		H3K4me1		H3K4me3		Input	
	RPE1	U2OS	RPE1	U2OS	RPE1	U2OS	RPE1	U2OS
interphase	◯	◯	◯	◯	◯	◯	◯	◯
prometaphase	◯	◯	◯	◯	◯	◯	◯	◯
anaphase/telophase	◯	◯	◯	◯	◯	◯	◯	◯

Table 2. Hypotheses for t tests applied to histone modification in the selected 507 DNA regions. The null hypothesis was that the inequality relationship of the alternative hypothesis is replaced with an equality relationship. int: interphase, ana: anaphase, tel: telophase, pro: prometaphase.

Test	Alternative hypothesis	P-value	Description of desired relationships
1	$\{x_{ij1ms} \| m = 1, 3\} > \{x_{ij12s}\}$	3.30×10^{-3}	H3K27ac reactivation (int & ana/tel > pro)
2	$\{x_{ij2ms} \| m = 1, 3\} \neq \{x_{ij22s}\}$	0.60	H3K4me1 bookmark (int & ana/tel = pro)
3	$\{x_{ij3ms} \| m = 1, 3\} \neq \{x_{ij32s}\}$	0.72	H3K4me3 bookmark (int & ana/tel = pro)
4	$\{x_{ij4ms} \| m = 1, 3\} \neq \{x_{ij42s}\}$	0.86	Input as control (int & ana/tel = pro)
5	$\{x_{ij2ms}\} > \{x_{ij4ms}\}$	8.98×10^{-6}	H3K4me1 > Input
6	$\{x_{ij3ms}\} > \{x_{ij4ms}\}$	3.79×10^{-3}	H3K4me3 > Input

Table 3. Number of transcription factors (TFs) associated with adjusted P-values less than 0.05 in various TF-related Enrichr categories

	Terms	Adjusted P-values	
		> 0.05	< 0.05
(I)	ChEA 2016	537	97
(II)	ENCODE and ChEA Consensus TFs from ChIP-X	91	12
(III)	ARCHS4 TFs Coexp	1533	54
(IV)	TF Perturbations Followed by Expression	1577	346
(V)	Enrichr Submissions TF-Gene Coocurrence	587	1135
(VI)	ENCODE TF ChIP-seq 2015	788	28
(VII)	TF-LOF Expression from GEO	239	11

Table 4. Identification of RUNX transcription factor (TF) family members within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	RUNX1	◯			◯			
2	RUNX2	◯						
3	RUNX3					◯		

Table 5. Identification of TEAD transcription factor (TF) family members within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	TEAD4	◯					◯	
2	TEAD3			◯				

Table 6. Identification of JUN transcription factor (TF) family members within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	JUN	◯			◯	◯	◯	
2	JUND	◯			◯	◯	◯	
3	JUNB				◯	◯		

Table 7. Identification of FOXO transcription factor (TF) family members within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	FOXO1				○	○		
2	FOXO3	○						
3	FOXO4					○		
4	FOXO6					○		

Table 8. Identification of FosL transcription factor (TF) family members within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	FOSL2		○				○	
2	FOSL1				○		○	

Table 9. Top 10 most frequently listed transcription factor (TF) families (at least four, considered the majority) within seven TF-related categories in Enrichr. Roman numerals correspond to the first column in Table 3.

	TF	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
1	GATA2	○	○		○	○	○	
2	ESR1	○	○		○	○	○	
3	TCF21	○		○	○	○		
4	TP53	○	○		○	○		
5	JUN	○			○	○	○	
6	JUND	○			○	○	○	
7	WT1	○			○	○		○
8	NFE2L2	○	○		○	○		
9	GATA1	○	○		○	○		
10	GATA3				○	○	○	○

Table 10. The performances achieved by DESeq2 applied to the present data set. Adjp: adjusted P-values computed by DESeq2

	RPE1		U2OS	
	Adjp > 0.01	Adjp < 0.01	Adjp > 0.01	Adjp < 0.01
H3K27ac	30649	1829	28849	1425
H3K4me1	113784	0	52323	4227
H3K4me3	26420	8259	24359	1559
Input	112976	0	5995	196

Table 11. The performances achieved by csaw applied to the present data set. Adjp: adjusted P-values computed by csaw

	RPE1		U2OS	
	Adjp > 0.01	Adjp < 0.01	Adjp > 0.01	Adjp < 0.01
H3K27ac	4127704	113803	4477318	6126
H3K4me1	5552148	0	6060553	5
H3K4me3	3054309	140962	2197717	27570
Input	3310106	0	5040796	0

postmitotic transcriptional reactivation. *Genes & Development* **34** (2020) 913–930. doi:10.1101/gad.335794.119.

Taguchi(2020). Taguchi YH. *Unsupervised Feature Extraction Applied to*

Bioinformatics (Springer International Publishing) (2020). doi:10.1007/978-3-030-22456-1.

Durinck et al.(2009)Durinck, Spellman, Birney, and Huber. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols* **4** (2009) 1184–1191.

R Core Team(2019). R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).

Kuleshov et al.(2016)Kuleshov, Jones, Rouillard, Fernandez, Duan, Wang et al.. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44** (2016) W90–W97. doi:10.1093/nar/gkw377.

Love et al.(2014)Love, Huber, and Anders. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15** (2014). doi:10.1186/s13059-014-0550-8.

Lun and Smyth(2015). Lun AT, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44** (2015) e45–e45. doi:10.1093/nar/gkv1191.

Langmead and Salzberg(2012). Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9** (2012) 357–359. doi:10.1038/nmeth.1923.

Li et al.(2009)Li, Handsaker, Wysoker, Fennell, Ruan, Homer et al.. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** (2009) 2078–2079. doi:10.1093/bioinformatics/btp352.

Wagner(2002). Wagner EF. Functions of ap1 (fos/jun) in bone development. *Annals of the Rheumatic Diseases* **61** (2002) ii40–ii42. doi:10.1136/ard.61.suppl_2.ii40.

Rached et al.(2010)Rached, Kode, Xu, Yoshikawa, Paik, DePinho et al.. Rached MT, Kode A, Xu L, Yoshikawa Y, Paik JH, DePinho RA, et al. FoxO1 is a positive regulator of bone formation by favoring protein synthesis and resistance to oxidative stress in osteoblasts. *Cell Metabolism* **11** (2010) 147–160. doi:10.1016/j.cmet.2010.01.001.

Kala et al.(2009)Kala, Haugas, Lilleväli, Guimera, Wurst, Salminen et al.. Kala K, Haugas M, Lilleväli K, Guimera J, Wurst W, Salminen M, et al. Gata2 is a tissue-specific post-mitotic selector gene for midbrain gabaergic neurons. *Development* **136** (2009) 253–262. doi:10.1242/dev.029900.

Kato and Ogawa(1994). Kato R, Ogawa H. An essential gene, ESR1, is required for mitotic growth, DNA repair and meiotic recombination Saccharomyces cerevisiae. *Nucleic Acids Research* **22** (1994) 3104–3112. doi:10.1093/nar/22.15.3104.

Kim et al.(2017)Kim, Pjanic, Nguyen, Miller, Iyer, Liu et al.. Kim JB, Pjanic M, Nguyen T, Miller CL, Iyer D, Liu B, et al. TCF21 and the environmental sensor aryl-hydrocarbon receptor cooperate to activate a pro-inflammatory gene expression program in coronary artery smooth muscle cells. *PLOS Genetics* **13** (2017) 1–29. doi:10.1371/journal.pgen.1006750.

Ha et al.(2007)Ha, Baek, Kim, Jeong, Kim, McKeon et al.. Ha GH, Baek KH, Kim HS, Jeong SJ, Kim CM, McKeon F, et al. p53 activation in response to mitotic spindle damage requires signaling via BubR1-mediated phosphorylation. *Cancer Research* **67** (2007) 7155–7164. doi:10.1158/0008-5472.CAN-06-3392.

Shandilya and Roberts(2015). Shandilya J, Roberts SG. A role of WT1 in cell division and genomic stability. *Cell Cycle* **14** (2015) 1358–1364. doi:10.1080/15384101.2015.1021525. PMID: 25789599.

Martin-Hurtado et al.(2019)Martin-Hurtado, Martin-Morales, Robledinos-Antón, Blanco, Pa Martin-Hurtado A, Martin-Morales R, Robledinos-Antón N, Blanco R, Palacios-Blanco I, Lastres-Becker I, et al. NRF2-dependent gene expression promotes ciliogenesis and hedgehog signaling. *Scientific Reports* **9** (2019). doi:10.1038/s41598-019-50356-0.

Shafer et al.(2017)Shafer, Nguyen, Tremblay, Viala, Béland, Bertos et al.. Shafer ME, Nguyen AH, Tremblay M, Viala S, Béland M, Bertos NR, et al. Lineage specification from prostate progenitor cells requires Gata3-dependent mitotic spindle orientation. *Stem Cell Reports* **8** (2017) 1018–1031. doi:10.1016/j.stemcr.2017.02.004.

Wu et al.(2015)Wu, Bittencourt, Stallcup, and Siegmund. Wu DY, Bittencourt D, Stallcup MR, Siegmund KD. Identifying differential transcription factor binding in chip-seq. *Frontiers in Genetics* **6** (2015) 169. doi:10.3389/fgene.2015.00169.

Steinhauser et al.(2016)Steinhauser, Kurzawa, Eils, and Herrmann. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics* **17** (2016) 953–966. doi:10.1093/bib/bbv110.

Tu and Shao(2017). Tu S, Shao Z. An introduction to computational tools for differential binding analysis with ChIP-seq data. *Quantitative Biology* **5** (2017) 226–235. doi:10.1007/s40484-017-0111-8.

Liang and Keleş(2011). Liang K, Keleş S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28** (2011) 121–122. doi:10.1093/bioinformatics/btr605.

Stark and Brown(2011). Stark R, Brown G. *DiffBind: differential binding analysis of ChIP-Seq peak data* (2011). Bioconductor.

Xu et al.(2014)Xu, Grullon, Ge, and Peng. Xu S, Grullon S, Ge K, Peng W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods in Molecular Biology* (Springer New York) (2014), 97–111. doi:10.1007/978-1-4939-0512-6_5.

Zhang et al.(2008)Zhang, Liu, Meyer, Eeckhoute, Johnson, Bernstein et al.. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS). *Genome Biology* **9** (2008) R137. doi:10.1186/gb-2008-9-9-r137.

Allhoff et al.(2014)Allhoff, Seré, Chauvistré, Lin, Zenke, and Costa. Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics* **30** (2014) 3467–3475. doi:10.1093/bioinformatics/btu722.

Song and Smith(2011). Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* **27** (2011) 870–871. doi:10.1093/bioinformatics/btr030.

Shao et al.(2012)Shao, Zhang, Yuan, Orkin, and Waxman. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-seq data sets. *Genome Biology* **13** (2012) R16. doi:10.1186/gb-2012-13-3-r16.

Heinz et al.(2010)Heinz, Benner, Spann, Bertolino, Lin, Laslo et al.. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell* **38** (2010) 576–589. doi:10.1016/j.molcel.2010.05.004.

Liu et al.(2013)Liu, Yi, SV, Lan, Ma, Huang et al.. Liu B, Yi J, SV A, Lan X, Ma Y, Huang TH, et al. QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. *BMC Genomics* **14** (2013) S3. doi:10.1186/1471-2164-14-s8-s3.

Shen et al.(2013)Shen, Shao, Liu, Maze, Feng, and Nestler. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffreps: Detecting differential chromatin modification sites from chip-seq data with biological replicates. *PLOS ONE* **8** (2013) 1–13. doi:10.1371/journal.pone.0065598.

Schweikert et al.(2013)Schweikert, Cseke, Clouaire, Bird, and Sanguinetti. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-seq data sets. *BMC Genomics* **14** (2013) 826. doi:10.1186/1471-2164-14-826.

Zhang et al.(2014)Zhang, Lin, Johnson, Rozek, and Sartor. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics* **30** (2014) 2568–2575. doi:10.1093/bioinformatics/btu372.

Chen et al.(2015)Chen, Wang, Qin, and Wu. Chen L, Wang C, Qin ZS, Wu H. A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics* **31** (2015) 1889–1896. doi:10.1093/bioinformatics/btv094.

Mahony et al.(2014)Mahony, Edwards, Mazzoni, Sherwood, Kakumanu, Morrison et al.. Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, et al. An integrated model of multiple-condition chip-seq data reveals predeterminants of cdx2 binding. *PLOS Computational Biology* **10** (2014) 1–14. doi:10.1371/journal.pcbi.1003501.