Spatial Clustering of Multivariate Genomic and Epigenomic Information

Rami Jaschek and Amos Tanay

Department of Computer Science and Applied Mathematics.
The Weizmann Institute of Science, Rehovot, 76100 Israel

Abstract. The combination of fully sequence genomes and new technologies for high density arrays and ultra-rapid sequencing enables the mapping of gene-regulatory and epigenetics marks on a global scale. This new experimental methodology was recently applied to map multiple histone marks and genomic factors, characterizing patterns of genome organization and discovering interactions among processes of epigenetic reprogramming during cellular differentiation. The new data poses a significant computational challenge in both size and statistical heterogeneity. Understanding it collectively and without bias remains an open problem. Here we introduce spatial clustering - a new unsupervised clustering methodology for dissection of large, multi-track genomic and epigenomic data sets into a spatially organized set of distinct combinatorial behaviors. We develop a probabilistic algorithm that finds spatial clustering solutions by learning an HMM model and inferring the most likely genomic layout of clusters. Application of our methods to meta-analysis of combined ChIP-seq and ChIP-chip epigenomic datasets in mouse and human reveals known and novel patterns of local co-occurrence among histone modification and related factors. Moreover, the model weaves together these local patterns into a coherent global model that reflects the higher level organization of the epigenome. Spatial clustering constitutes a powerful and scalable analysis methodology for dissecting even larger scale genomic dataset that will soon become available.

Introduction

The combination of fully sequenced genomes and new technologies for high density arrays (ChIP-chip) and ultra-rapid sequencing (ChIP-seq) enables the mapping of gene-regulatory and epigenetics marks on a global scale. Such mapping is being employed at an increasing pace to study genomic and epigenomic organization at different developmental stages [1-5]. The new massive experimental data has already revealed how genomes are programmed and reprogrammed by complex patterns of histone marks, transcription factors and regulatory complexes [6]. It was suggested that a mechanistic understanding of normal [7] and malignant [8, 9] differentiation processes can be facilitated through a comprehensive analysis of multiple marks and factors in multiple cell systems [10]. The great promise of the new datasets lies in their lack of bias and truly genomic scale. To fully exploit their potential, one must develop an adequate analysis methodology that can go beyond the study of a single

histone mark or transcription factor. The goal is to comprehensively identify complex genomic structures without an a-priori focus on known features (e.g., genes and promoters).

Current approaches for analyzing genomic information focus on the distributions of values relative to transcription start sites (TSS) or other genomic features (e.g., CTCF binding sites [11]). Computing such distributions is computationally easy and informative: one can depict the relations between TSSs and the profiled factors through graphs of the average factor/mark occupancy as a function of the distance to the nearest TSS. Comparing the profiles near active and inactive TSSs can highlight possible functional implications for the profiled distributions. Heat maps are used to show possible correlations between modification pairs [5]. While being easy to understand and very effective in mapping the organization around TSSs, the averaging methods provide little help when trying to understand the datasets as a whole. Focusing on the patterns around specific features does not enable the identification of novel genomic structures or higher level organization. The comprehensive and innovative nature of the experiments is therefore still unmatched analytically.

The problem of identifying patterns in large datasets is a hallmark of computational biology. Extensive literature is dealing with the analysis of gene expression data, dissecting it into clusters [12, 13] or biclusters [14] or in explaining the data by means of a complex regulatory model [15]. Clustering became the method of choice for analysis of gene expression, mostly due to its simple and unsupervised nature and since it allows effective visualization of the entire dataset by means of a few robust structures. The new generation of ChIP-chip and ChIP-seq datasets is however not easily approached using naïve clustering. At the most technical level, the datasets are huge and cannot be analyzed using the current algorithms' implementations. More fundamentally, the genomic datasets are spatially arranged over chromosomes and their analysis must account for this organization. Present clustering methodologies are inadequate for analysis of multi-track and heterogeneous ChIP-chip and ChIP-seq data.

In this paper we introduce the spatial clustering problem and describe our probabilistic algorithm for solving it. Our algorithm clusters a set of genomic profiles (tracks), representing epigenetic modifications, factor occupancy or other spatially distributed data. We model the data using an HMM which is being learned in an unsupervised fashion. The algorithm then infers the most likely coverage of the genome with contiguous spatial clusters. The HMM component of our model can flexibly express local structure (short genomic intervals with similar profiles) and global structure (groups of clusters that tend to co-occur and form larger domains). We demonstrate this by analyzing two combined whole genome datasets including epigenomic profiles of human cells and differentiating mouse stem cells. In both cases, our analysis provides a comprehensive map of epigenomic modes that extend beyond the reported patterns around TSSs. Spatial Clustering is a flexible and robust tool that is designed to meet the requirements of large genomic and epigenomic projects. It provides a powerful alternative to the limited and supervised analysis scheme which is currently in common use.

Methods

The K-Spatial Clustering problem. Given a multivariate genomic dataset, we seek a representation of the data using a limited repertoire of distinct behaviors that are spatially organized. We formalize this notion as the K-Spatial clustering problem. Assume that we are studying a linearly organized (possibly on more than one interval) set of genomic measurement loci (or probes). Loci can be ordered along complete chromosomes at fixed distances (as in binned ChIP-seq data), or may tile selected parts of the genome at variable distances (as in ChIP-chip data). For simplicity, we will look only at the sequential position of each locus inside its surrounding contiguous segment, breaking chromosomes into segments wherever there is large gap between two adjacent probes. Assume also that we are given a vector of experimental measurements at each of the loci. The experimental data can come from different types of experiments or different cell types and conditions, but we disregard any available information on a-priori connections between vector entries. A K-Spatial clustering is a partition of the underlying genomic region into disjoint and contiguous intervals and a tagging of these intervals with cluster numbers [1..K]. In the most general settings, we introduce a quality function that scores K-Spatial clustering instances given the genomic dataset and defines the K-Spatial clustering problem as an optimization problem that seeks the maximal scoring K-Spatial clustering. We note that in the degenerate case, when the scoring function ignores the genomic layout of the data and is based solely on similarity of measured values per loci, the K-Spatial clustering problem is equivalent to a standard clustering problem. Needless to say, in any reasonable application, the quality function will take advantage of the genomic organization of the data to derive better solutions.

Spatial clustering can be considerably more informative and powerful than simple clustering of the probes. First, in most practical cases, the experimental values for adjacent probes are highly correlative as important biological features would typically span several measurement loci. These effects are making the common assumptions underlying most clustering frameworks incorrect (i.e., the a-priori independence among samples is not holding). A good K-Spatial clustering solution would therefore maximize the number of adjacent loci that are part of the same cluster (reducing the overall number of intervals) while not compromising the integrity and specificity of the clusters. A second source of spatial information works at a higher level. In many genomic datasets we can observe coupling between related phenomena. For example, transcription start sites (TSS) would often be followed by a transcribed region. Expressing the couplings between clusters, and introducing it into the quality function can have an important contribution to the quality of the results, and to our ability to understand them.

Probabilistic K-Spatial clustering: Our approach for deriving K-Spatial clustering is based on a probabilistic formulation of the problem (**Fig. 1A**). This is analogous to the standard approach that estimates a mixture model (for example, a mixture of multivariate Gaussians) to cluster large data sets [16]. We extend the naïve mixture model using an HMM structure that expresses the tendency of adjacent loci to remain in the same cluster and the spatial coupling among clusters. Specifically, the model is defined using K multivariate distributions over the experimental tracks and an HMM

model connecting the clusters (with additional hidden states, see below). The topologies we shall consider for the HMM graph reflect constraints on higher level organization of clusters across the genome. Given a fixed topology, we apply the expectation maximization (EM) algorithm to learn the model and its parameters and then compute posterior probabilities for the association between measurement loci and HMM clusters. The set of contiguous intervals associated with the same HMM state (with high posterior probability) are then used as our spatial clustering. We can also study the parameters of the distributions defining each cluster and the transition probabilities between clusters to get a more global picture of the model and its implications.

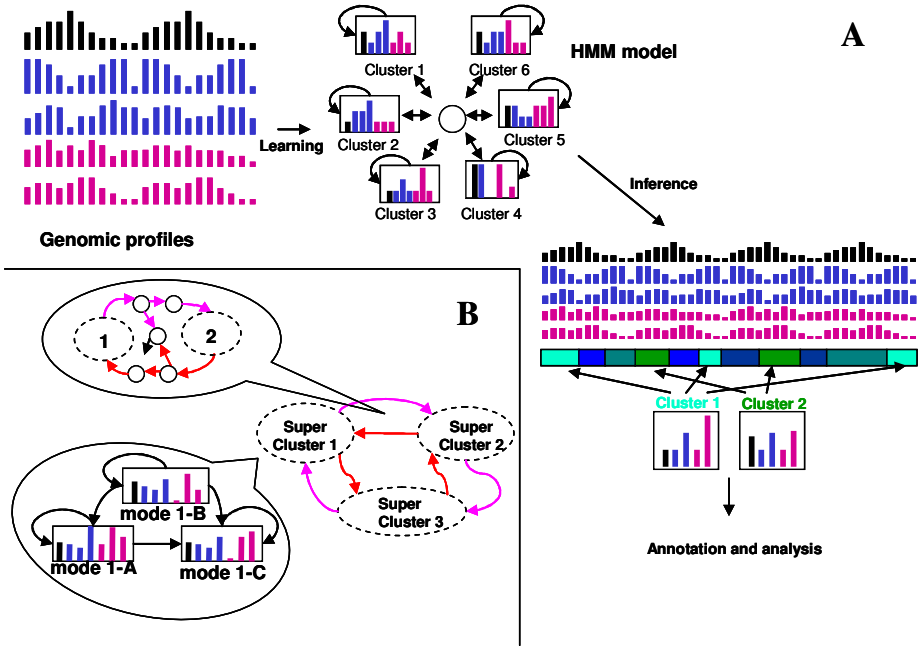

Figure 1. A) Spatial clustering of genomic profiles (tracks). Multi-dimensional genomic datasets, including heterogeneous ChIP-seq and ChIP-chip experiments as well as other genomic sources of information are analyzed together using a probabilistic hidden Markov model. The model is then used to infer the most likely partition of the genome into spatial clusters, each representing a specific genomic or epigenomic behavior which is determined based on the distribution of all data tracks. **B) Hierarchical spatial clustering**. Shown is a schematic view of the HMM topology we use for building hierarchical spatial clustering. The model consists of a set of small complete graphs (*super clusters*, here on 3 states) that are connected through dedicated connector state pairs. Transition probabilities inside each connector pair are fixed throughout the learning process and add a penalty for crossing super cluster boundaries.

Local distribution parameterization: In our generative model, multi-track data is emitted from cluster states given a probability distribution associated with the cluster. To learn the model, we need to specify a family of distributions appropriate for the

tracks we analyze. A good selection of distributions family is one that will be able to generate, given the correct parameters, a distribution of values that is as close as possible to the one observed, and would also allow robust learning with limited data. The simplest class of distributions assumes tracks are generated independently (once the cluster is determined). This simplification, which we use in the present work, is economical in terms of parameters and allows for learning robustly, even when the number of tracks grows. More refined classes of distributions (e.g., multinormal distribution with arbitrary covariance matrices, [8]) are currently practical only for smaller number of tracks and should be further developed to allow robust learning in general.

The most common sources of comprehensive genomic data are ChIP-chip and ChIP-seq experiments. Results from ChIP-chip experiments are real valued binding ratios. ChIP-seq results are lists of sequenced reads which are normalized into some coverage statistics on genomic intervals. For ChIP-chip data, the combination of experimental and biological noise can be modeled using simple normal distributions. Chip-seq results are by nature discrete, and should be theoretically distributed as a combination of samples from two fragment pools (false positives and enriched IP fragments [4, 17]). However, according to our analysis, the empirical distribution of ChIP-seq tracks cannot be effectively approximated as a mixture of noise and a geometric distribution, featuring a very heavy tail (data not shown). We therefore model ChIP-seq tracks using an a-parametric discrete distribution on variable sized bins. We generate the bins by identifying, for each track, the values of the $(1-2^k)$ percentiles (for k=1 to 10). The distributions considered for the track are now defined using a discrete distribution over 10 bins, where we map physical measurements to bins with values in the percentile intervals $[(1-2^{k+1}): (1-2^k)]$.

The HMM structure. We use the HMM structure to impose constraints on clustering solutions, demanding clusters would occupy contiguous genomic intervals and coupling together clusters that are frequently occurring next to each other. We use several structure families with increasing degrees of details. In its most simple form, the model has a star structure, which is associating all cluster states through a central hidden (non-emitting) state. The structure is in essence a simple mixture model (transitions probabilities from the central states to cluster states correspond to the mixture coefficients), but uses the states self transitions to increase the likelihood of solutions with contiguous clusters. The star topology imposes no constraint on transitions between specific emitting clusters pairs (all transitions must go through the central hidden node). Such transitions can be inferred from the posterior state probabilities in post process. Alternatively, an HMM over a clique topology (connecting all pairs of cluster states) is by far more expressive, but may be too detailed to allow robust learning, especially when the number of clusters increase. We note that even the clique topology can capture only coupling between pairs of clusters and cannot be used to represent higher order structures, which are frequently observed in genomic data.

To try and model higher order genomic structure, we are using a hierarchical topology (**Fig. 1B**). In this scheme, we construct small clique HMMs (super-clusters) to represent specific genomic structures that tend to co-occur over larger domains. We then connect super clusters using dedicated connector state pairs that implicate a probabilistic payment for each transition between super clusters. In the current framework we have worked with two levels of hierarchy, but these can be naturally

generalized. More specific HMM topologies can be developed to model additional biological phenomena. For example, genomic structure is often polarized according to the direction of the nearby transcript, but our basic HMM implementation reads all data in the same order and can therefore learn only unidirectional couplings. To improve on that, we can form two copies of our model, each representing one genomic polarization, such that all transition probabilities are reverse symmetric between the two copies.

Model Learning: Given an HMM topology, we learn the model parameterization using a standard Expectation-Maximization (EM) algorithm [18]. The success of the local optimization performed by the EM algorithm depends, as is often the case, on careful selection of initial conditions. The most critical initialization for the spatial clustering model is proper selection of initial cluster state emission distributions. In our implementation, we first preprocess the data to identify standard clusters (ignoring spatial information). We then use a repertoire of detected clusters to initialize cluster states over some initial topology. The clustering algorithm in the preprocessing stage can be selected arbitrarily. The implementation we report here consist of three phases: a) We discretize all data to three levels (-1,0 and 1) using predefined Z-score thresholds for each track. b) We group together all probes with the same discretized multi-track vector and count how many probes are classified for each vector. c) For each discretized vector with a sufficient number of associated probes, we estimate from the original data the (non discrete) multivariate distribution over the tracks and add it to the set of potential initial cluster states (or seeds). In case too few heavy clusters are available, we adjust the Z-score threshold and return to step a. Given a set of seed states, we generate an initial model by randomly selecting initial cluster states from the pool of generated seeds. Our analysis suggests that testing few dozens of initial conditions is performing effectively, and is comparable in performance to a greedy scheme in which we construct the model state by state, at each step testing EM after addition of each of the seeds and choosing the seed with the maximal likelihood gain (data not shown).

To learn a hierarchical model we work from top down. We first learn an HMM model on K states using a star topology. We then use inference with the derived model to partition the data into K sets, one for each cluster. We apply the initialization procedure described above separately for each of the K probes sets, and construct an initial model by combining together K clique models (each with L states) using the connector states described above (**Fig. 1B**).

Our unsupervised approach to the learning problem provides us with a robust and unbiased way to analyze the data. The approach can however lead to models that are optimized toward better representation of the experimental noise in the genomic background rather than models outlining meaningful biological effects. A large percentage of the genome may be showing little or no meaningful signal. Modeling these parts of the genome accurately yields a high likelihood gain – pushing the learning process to use additional clusters for refined background models. We can somewhat control this problem by increasing the total number of clusters in the model. More heuristically, we can also employ constraints to the learning process, limiting the variance of the emission distributions so that states cannot be too tuned for a specific

background behavior (by having small variance), or too broad to absorb several biological phenomena (by having large variance).

We note that the datasets our algorithm is handling are imposing new kinds of technical challenges on design and implementation. For example, the largest gene expression datasets analyzed thus far are limited to ~10^4 genes and ~10^3 conditions. The dataset we have analyzed here have ~10^8 probes (tiling the human genome in 25bp-50bp resolution) and ~10^2 tracks, making it 2-3 order of magnitudes larger. These numbers are expected to increase further. We are therefore forced to utilize considerable computational resources even if the algorithm is extremely efficient. In our implementation, the EM inference (Forward-Backward) step is massively parallelized over a computer cluster, allowing learning on large genomes to be completed quickly.

Visualization. The most direct visualization scheme for genomic data is as data tracks in genome browsers [19, 20]. When the data includes many tracks this scheme may prove difficult to follow, and at any rate can only provide local information on a specific locus and not global understanding on how the data is organized. Another common approach is to average the available genomic profiles with respect to a genomic feature (usually TSSs), but as argued above, this approach is strongly biased to a specific phenomenon and may miss important genomic structures. Our spatial cluster model opens the way to new visualization schemes of complex genomic data. We use a learned model to infer the most probable cluster state associated with each locus. We can then color code the genome according to the associated clusters in a way that summarizes all available experimental profiles in one color per locus. With appropriate selection of colors, this can be an effective way to identify both global and local behavior (see below). A complementary approach is attempting to visualize the entire data set in one figure. To do this we identify contiguous intervals associated with a cluster by looking at ranges of probes with consecutive high posterior probabilities for the same HMM state. We then pool together groups of intervals that were associated with the same cluster and plot the genomic profiles inside and around them in standard cluster-gram (each row represents an interval and its margins). From our experience, although the color coding approach is somewhat qualitative, it is currently the best way to rapidly understand the entire dataset in one view, providing a good starting point for further analysis.

Results

Spatial clustering model for human T cells epigenetics. Barksi et al. [1] have used ChIP-seq and a collection of antibodies for 20 histone methylation marks, RNAP, the Histone variant H2A.Z and CTCF to globally map the epigenome of human T-cells. This constitutes the most comprehensive epigenomic dataset on a single mammalian cell type to date. TSS averaging analysis of the data confirmed and extended known principles of chromatin organization around active and repressed TSSs [5]. Recently, work from the Van Steensel group has put forward a genome wide map of chromatin interactions with the nuclear lamina [21], characterizing large lamina-associated domains that are very gene sparse and flanked by transcriptional units, insulators and

H3K27 trimethylation. In an attempt to gain an unbiased view on the interactions among the profiled histone marks and other factors, we have combined these two datasets and applied spatial clustering analysis to them. We note that the two sets of experiments we used were derived from different cell types (T-cells vs. fibroblasts) and different technologies. To allow common analysis, we transformed all ChIP-seq data to coverage statistics on 50bp bins, assuming fragment length of 300bp (as described in [4]). We also assumed that each Lamin B1 tiling array probe represents the occupancy at all 50bp bins in the range of 1000bp around the probe.

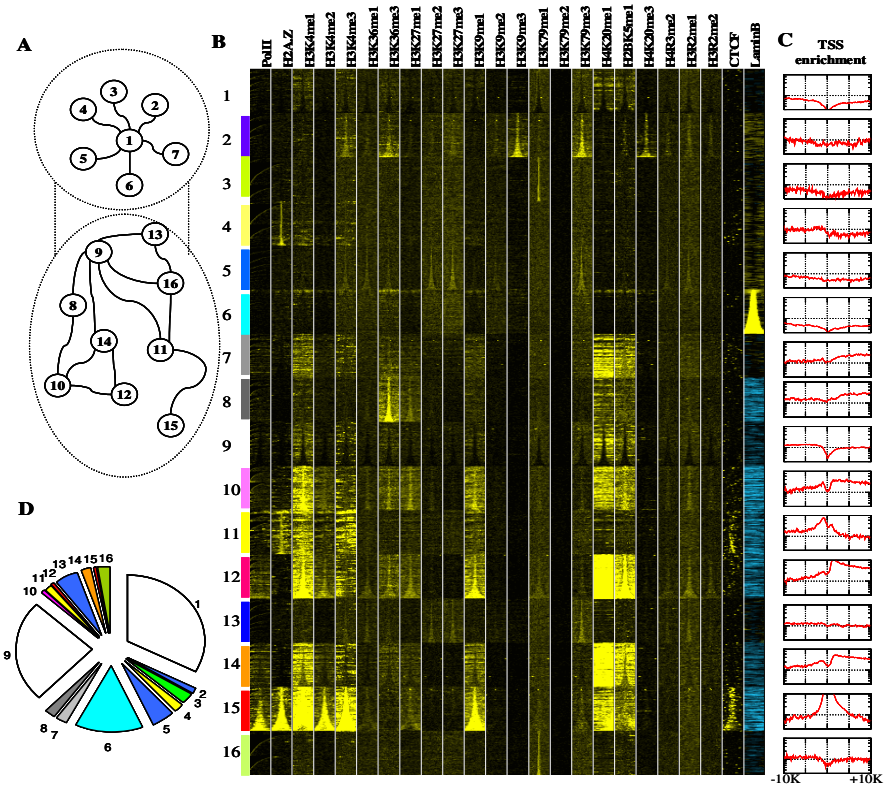

Figure 2. Spatial clustering of the human epigenome. A) **HMM basic topology**. Shown are the model states, where we connect pairs of states that frequently follow each other. B) **Color coded visualization of clusters and their flanking region**. Each row segment represents the occupancy of one mark or factor over one spatial cluster and its flanking region, where we stack together loci that were associated to the same cluster (blocks). The color coding is providing a quick visualization of the complex model, for a more quantitative view consult http://www.wisdom.weizmann.ac.il/~ramij/recomb2009.html. C) **TSS enrichments**. The graph shown indicates for each cluster and offset from the TSS (X axis) the ratio between the fraction of loci associated with the cluster at that TSS offset to the overall fraction of genomic loci at that TSS offset (Y axis, 0.2 to 6, log scale). D) **Cluster coverage**. Shown are the fractions of genomic loci covered by each of the clusters. Color coding is similar to panel B**.**

In **Fig. 2A** we show the major topological interactions in the spatial cluster model we derived and in **Fig. 2B** we present a color coded clustergram of the inferred clustering, depicting the profiles in and around the clusters (Each row represent one spatial cluster at one genomic locus, rows are grouped according to their associated cluster state, see Methods). A more quantitative summary of **Fig. 2** is shown in http://www.wisdom.weizmann.ac.il/~ramij/recomb2009.html. We used the hierarchical learning process described above, with two clusters at the first phase and eight clusters per super-cluster in the second phase. The algorithm chose to first partition the genome according to the strength of interaction with the nuclear lamina (upper clusters – strong interaction, lower clusters – no interaction) and then constructed a detailed model to describe different combinations of histone modification and factor occupancy and their couplings. To further illustrate the possible relationships between the clusters and genes we computed the enrichment of cluster associations as a function of the distance from the TSSs (**Fig. 2C**).

Several clusters we detected represent known structure around TSSs which the algorithm rediscovered in a completely unsupervised fashion and without using information on the TSSs locations. Cluster 15 is the only one with high levels of RNA PolII, and is further associated with very high H3K4me3 levels and significant enrichments of H2A.Z. The cluster is also enriched with H3K9me1, and is sometime observed with CTCF binding. CTCF binding was reported before to occur at alternative promoters sites [22], and this may explain the partial co-occurrence between RNAP and CTCF at cluster 15. Cluster 12 and cluster 10 represent a combination of mono methylation at multiple positions (H3K4me1, H3K36me1, H3K9me), all of which were associated before with the regions flanking the TSS. For cluster 12, there are very strong enrichments of H4K20me1 and H2BK5me1 and a detected preference for the regions downstream the TSS, while for cluster 10, little H4K20me1 and H2BK5me1 enrichment is detected. Each of the clusters are covering about 2% of the genome (**Fig. 2D**). The five monomethylation marks (at H3 K4, K9, K36, H4K20 and H2BK5) were associated before with active chromatin, but here we see that there are at least two modes of correlation among them. We note that since we observe such complex pattern, general monomethylation antibody specificity is unlikely to explain these patterns. The monomethylation marks may still share a mechanism that will explain their high degree of correlation. Cluster 8 is representing enriched trimethylation at H3K36, which is largely free of other marks and is found in the longer range downstream of transcribed genes. This modification was previously associated with tran-

scription elongation. Cluster 11 shows insulator patterns, including hotspots of CTCF and H2A.Z, with a preference to the region upstream of the TSS.

Other clusters are not tightly associated to TSSs and represent weaker enrichments than the clear preferences discussed above. They still show a high degree of correlation between marks and can provide insights into the organization of chromatin out of the TSS context. Cluster 6 represents cores of Lamin B1 interaction and is not linked to any of the histone modifications. All clusters associated with the lamina are generally void of activation mark, as noted before, and even though the datasets compared two different cell types. Clusters 5 and 13 are characterized by the polycomb modifications, H3K27me3 and H3K27me2 and are occupying over 10% of genome, mostly in intergenic regions. Polycomb marks were characterized extensively in embryonic stem cells, and domains with strong H3K27me3 enrichment were linked to gene repression in genes poised for later activation upon differentiation. Cluster 5 and 13 may represent a broader and weaker pattern of polycomb marks, reminiscent of recent evidence on large H3K27me3 domains or large scale polycomb domains in flies [23]. A different type of repressive mark is H3K9me3 which is the main distinguishing feature of cluster 2. Interestingly, cluster 2 is also associated with H4K20me3 which was observed before to correlated with H3K9me3 [24] but was not suggested to co-occur with H3K9me3 in the original TSS-centric analysis of the data [1]. Here we also observe association of the H3K9me3 heterochromatic mark with trimethylation at H3K79, in concordance with recent evidence on H3K79me3 pericentromeric localization in mice [25] and with results derived from the same dataset using a local pattern search approach [26]. Finally, two clusters (3 and 16) are characterized by cryptic enrichment of H3K79me1 – with unclear functional or organizational specificity.

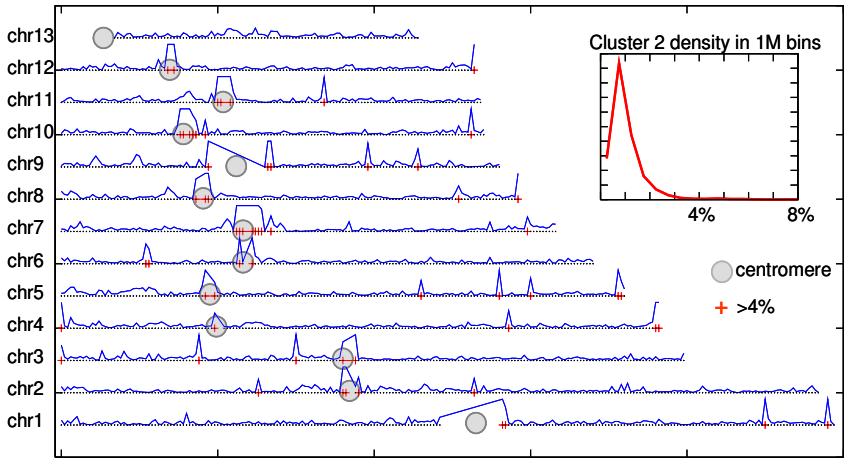

Figure 1: Hotspots of H3K9me3/H3K79me3/H4K20me3 (cluster 2) enrichment. We computed the fraction the genome covered by cluster 2 in bins of 1MB. Shown is the overall distribution of these fractions (upper right) and how it is spatially organized in chromosome 1-13 (main figure, blue curves). Hotspots (fraction over 4%) are marked in red and are preferentially occurring in pericentromeric regions

Beyond the set of clusters and their properties, the spatial cluster model also reflects higher level genomic organization. As shown in **Fig. 3**, cluster 2 occupancy (Representing trimethylation in H3K9, H3K79 and H4K20) is distributed in a highly non uniform fashion across the genome, with mean occupancy of less than 1%, but a significant number of 1MB genomic bins with more than four fold that number. As expected, the strongest cluster 2 hotspots are observed in pericentromeric regions, but many additional hotspots are apparent. **Fig. 3** provides a general reference for the organization of H3K9me3 heterochromatin in CD4+ T-cells.

In **Fig. 4** we illustrate the clustering of a specific chromosomal region around the HOXA gene family. The higher level view shows domains of "open" chromatin (red and gray bars) packed between larger domains of repressed ("closed") chromatin (blue bars: clusters 5,13 and 2, yellow bars: cluster 11). "Closed" clusters are frequently co-occuring with lamina clusters (light blue: cluster 6). The 100KB around the HOXA genes are unique in showing both active (red) and repressive (blue) marks. A higher resolution view (lower panel) shows activation domains at key HOXA gene promoters, and insulator/lamina domains flanking the entire region.

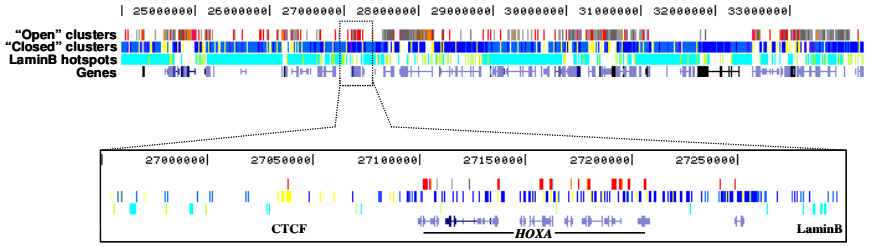

Figure 2: Spatial clustering in the HOXA region. Shown is the clustering of some 10MB in the human chromosome 7, color coded according to the clustering shown in **Fig 2**. Genes are marked in the lower track.

Spatial clustering model for differentiation of mouse embryonic stem cells. The genomes of embryonic stem cells (ESC) are uniquely organized to ensure pluripotency and maximize flexibility upon a differentiating signal. Genome wide maps of key histone marks, involving the trithorax- (H3K4) and Polycomb- (H3K27) associated modifications have revolutionized our understanding of the pluripotent epigenetic state [4, 27, 28]. To analyze the mouse ESC epigenomic state we applied spatial clustering to a combined ChIP-seq dataset [3, 4]. The data set included only few of the histone marks that were analyzed above for human T-cells, but allowed comparison between different cell types. To apply our algorithm, we used ChIP-seq coverage statistics in bins of 50bp as described [4] and derived a model including 3 superclusters and 6 sub clusters in each of them. We also used DNA methylation data from

the same experiments, but since their overall genomic coverage was very low, we omit them from the discussion.

As shown in **Fig. 5** the spatial clustering reveals both cell-type conserved and cell type specific patterns (compare also http://www.wisdom.weizmann.ac.il/~ramij/recomb2009.html for a better quantitative resolution),. As expected, and similarly to the pattern observed in somatic human cell (**Fig. 2**), the algorithm generated a cluster (#16) that is defined by strong H4Kme3 presence and is observed almost exclusively at annotated TSSs. Cluster 15, which is coupled by the HMM to cluster 16, represents a region with dominant H3K4me1 levels, which are conserved between cell types (see http://www.wisdom.weizmann.ac.il/~ramij/recomb2009.html) and clusters 18,14 (and to a lesser extent 2) are based on cell-type conserved H3K36me3 elongation marks. Clusters 17 and 12 represent H3K4me1 domains that are specific to mESC or NP cells. As noted before [29], the active chromatin state in embryonic stem cells is frequently co-occurring with high levels of H3K27me3, and this is reflected in cluster 16. On the other hand, cluster 11 represent domains in which H3K27me3 is the only significantly enriched mark, with very mild bias to promoters and conserved intensity between stem cells and derived lineages (in contrast to the decreasing H3K27me3 intensity in cluster 17)

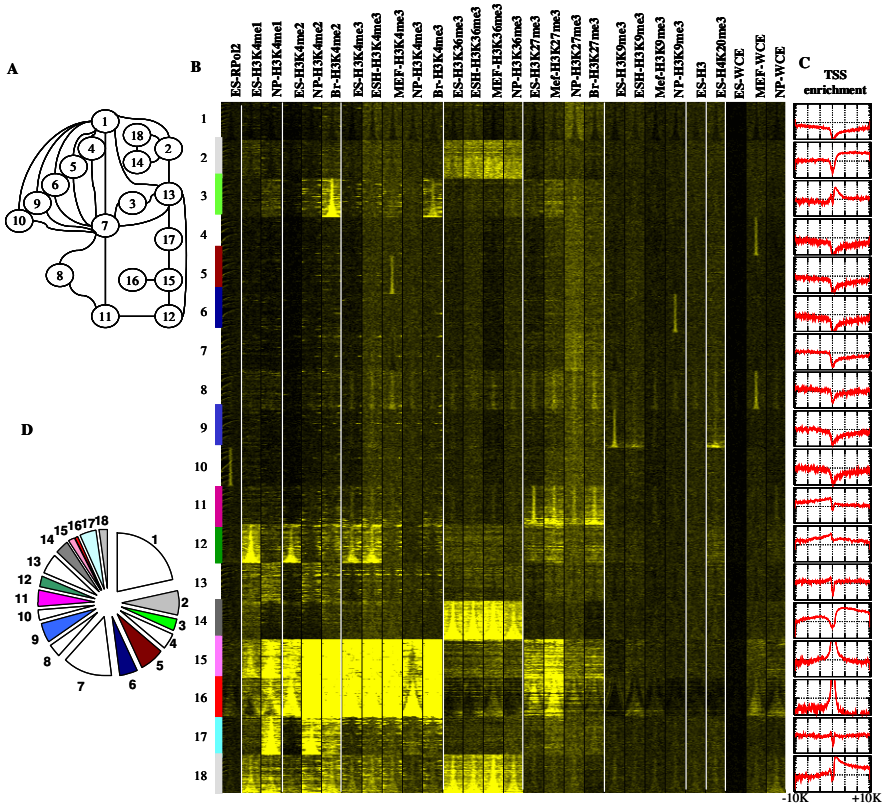

Figure 3: Spatial clustering of the mouse ESC epigenome. A) **HMM topology.** Shown are the model states, where we connect pairs of states that frequently follow each other. B) **Global view**. Color coded visualization of clusters and their flanking region. Each row segment represents the occupancy of one mark or factor over one spatial cluster and its flanking region, where we stack together loci that were associated to the same cluster (left color coded blocks). The color coding is providing a quick visualization of the complex model, but may become saturated (e.g., in clusters 15 and 16). For a more quantitative view consult http://www.wisdom.weizmann.ac.il/~ramij/recomb2009.html. C) **TSS enrichments**. The graph shown indicates for each cluster and offset from the TSS (X axis) the ratio between the fraction of loci associated with the cluster at that TSS offset to the overall fraction of genomic loci at that TSS offset (Y axis, 0.2 to 6, log scale). D) **Cluster coverage**. Shown are the fractions of genomic loci covered by each of the clusters. Color coding is similar to panel B.

As observed in the human datasets, we observe clusters of H3K9me3 activity. Cluster 9 shows H3Kme9 in mESC, with some matching H4K20me3 co-occurrence. Perhaps surprisingly, this pattern is not conserved in NP cells, which have their own H3K9me3 cluster (#6). It is unclear if this represents real plasticity of these heterochromatic marks, or experimental limitations (for example, cluster 8 and 4 isolate experimental artifacts by detecting clusters that are defined by elevated coverage in whole cell extract controls). These questions should be further addressed experimentally.

Discussion

We have presented spatial clustering as a new analysis methodology for dissecting large ChIP-chip and ChIP-seq datasets into defined clusters of common genomic or epigenomic behavior. The new method is allowing unsupervised and global modeling of the data, in a way that matches the unbiased and comprehensive nature of the experiments. It represents the entire genome as an organized set of contiguous clusters, and is capable of capturing both the nature of each cluster and the relations between them, something that is not available in local views [26]. Spatial clustering does not assume any gene structure or information on TSSs for defining clusters, and can therefore be used to study both TSS-related and TSS-unrelated genomic phenomenon. The model is constructing patterns that are based on a combined behavior over all experimental tracks and therefore scales well with increasing number of experiments. Spatial clustering can be the first line of analysis for genomic data, serving as a starting point for more careful hypothesis testing in a way similar to that by which standard clustering is used for gene expression analysis.

The data we analyzed here is providing us with a comprehensive view of the epigenomic structure of human T-cells and differentiating mouse ESCs. The analysis concisely summarizes known chromatin modes and reveals some new testable associations between histone marks (e.g., H3K9me3 with H3K79me3). The results emphasize the need for an integrative and coherent model that can broadly combine multiple epigenomic datasets and derive architectural insights. Ultimately, such a model should be expanded to include more regulatory factors, in an attempt to explain how genomes organization is regulated. It should also take into account higher order chromatin structure, which may be critical for the understanding of genomic organiza-

tion as already suggested by the remarkable nuclear lamina interaction data we used here. The Spatial clustering framework we introduced provides an immediate answer to problems with analysis of current data, as well as foundations for the development of more holistic models for genome organization.

Acknowledgments

This work was supported by an Asher and Jeannette Alhadeff Research Award and by the Israeli science foundation converging technologies program. AT is an Alon fellow.

References

1. Barski, A., et al., High-resolution profiling of histone methylations in the human genome. Cell, 2007. 129(4): p. 823-37.
2. Li, X.Y., et al., Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol, 2008. 6(2): p. e27.
3. Meissner, A., et al., Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature, 2008. 454(7205): p. 766-70.
4. Mikkelsen, T.S., et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 2007. 448(7153): p. 553-60.
5. Wang, Z., et al., Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet, 2008. 40(7): p. 897-903.
6. Liu, C.L., et al., Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol, 2005. 3(10): p. e328.
7. Guenther, M.G., et al., A chromatin landmark and transcription initiation at most promoters in human cells. Cell, 2007. 130(1): p. 77-88.
8. Gal-Yam, E.N., et al., Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc Natl Acad Sci U S A, 2008. 105(35): p. 12979-84.
9. Kondo, Y., et al., Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. Nat Genet, 2008. 40(6): p. 741-50.
10. Moving AHEAD with an international human epigenome project. Nature, 2008. 454(7205): p. 711-5.
11. Fu, Y., et al., The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet, 2008. 4(7): p. e1000138.
12. Ben-Dor, A., R. Shamir, and Z. Yakhini, Clustering gene expression patterns. J Comput Biol, 1999. 6(3-4): p. 281-97.
13. Eisen, M.B., et al., Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A, 1998. 95(25): p. 14863-8.
14. Tanay, A., et al., Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci U S A, 2004. 101(9): p. 2981-6.
15. Segal, E., et al., Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003. 34(2): p. 166-76.
16. Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer Series in Statistics. 2001: Springer. 533.

17.Kuznetsov, V.A., et al., Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. Genome Inform, 2007. 19: p. 83-94.
18.Durbin, R., Eddy, S., Krogh, A., Mitchison, G. , Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1998, Cambridge, UK: Cambridge University Press.
19.Hubbard, T., et al., The Ensembl genome database project. Nucleic Acids Res, 2002. 30(1): p. 38-41.
20.Kent, W.J., et al., The human genome browser at UCSC. Genome Res, 2002. 12(6): p. 996-1006.
21.Guelen, L., et al., Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature, 2008. 453(7197): p. 948-51.
22.Kim, T.H., et al., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell, 2007. 128(6): p. 1231-45.
23.Schuettengruber, B., et al., Genome regulation by Polycomb and trithorax proteins. Cell, 2007. 128(4): p. 735-45.
24.Vakoc, C.R., et al., Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol, 2006. 26(24): p. 9185-95.
25.Ooga, M., et al., Changes in H3K79 methylation during preimplantation development in mice. Biol Reprod, 2008. 78(3): p. 413-24.
26.Hon, G., B. Ren, and W. Wang, ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol, 2008. 4(10): p. e1000201
27.Boyer, L.A., et al., Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature, 2006. 441(7091): p. 349-53.
28.Lee, T.I., et al., Control of developmental regulators by Polycomb in human embryonic stem cells. Cell, 2006. 125(2): p. 301-13.
29.Bernstein, B.E., et al., A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell, 2006. 125(2): p. 315-26.