Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps

CSH PRESS GENOME RESEARCH

Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps

Ali Mortazavi, Shirley Pepke, Camden Jansen, et al.

Genome Res. 2013 23: 2136-2148 originally published online October 29, 2013
Access the most recent version at doi:10.1101/gr.158261.113

Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps

Ali Mortazavi,[1,2,12,13] Shirley Pepke,[3,4,12] Camden Jansen,[1,2] Georgi K. Marinov,[4] Jason Ernst,[5] Manolis Kellis,[6,7] Ross C. Hardison,[8,9] Richard M. Myers,[10] and Barbara J. Wold[4,11,13]

[1]Department of Developmental and Cell Biology, University of California, Irvine, California 92697, USA; [2]Center for Complex Biological Systems, University of California, Irvine, California 92697, USA; [3]Center for Advanced Computing Research, California Institute of Technology, Pasadena, California 91125, USA; [4]Division of Biology, California Institute of Technology, Pasadena, California 91125, USA; [5]Department of Biological Chemistry, University of California, Los Angeles, California 90095, USA; [6]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA; [7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [8]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [9]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [10]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [11]Beckman Institute, California Institute of Technology, Pasadena, California 91125, USA

We tested whether self-organizing maps (SOMs) could be used to effectively integrate, visualize, and mine diverse genomics data types, including complex chromatin signatures. A fine-grained SOM was trained on 72 ChIP-seq histone modifications and DNase-seq data sets from six biologically diverse cell lines studied by The ENCODE Project Consortium. We mined the resulting SOM to identify chromatin signatures related to sequence-specific transcription factor occupancy, sequence motif enrichment, and biological functions. To highlight clusters enriched for specific functions such as transcriptional promoters or enhancers, we overlaid onto the map additional data sets not used during training, such as ChIP-seq, RNA-seq, CAGE, and information on *cis*-acting regulatory modules from the literature. We used the SOM to parse known transcriptional enhancers according to the cell-type-specific chromatin signature, and we further corroborated this pattern on the map by EP300 (also known as p300) occupancy. New candidate cell-type-specific enhancers were identified for multiple ENCODE cell types in this way, along with new candidates for ubiquitous enhancer activity. An interactive web interface was developed to allow users to visualize and custom-mine the ENCODE SOM. We conclude that large SOMs trained on chromatin data from multiple cell types provide a powerful way to identify complex relationships in genomic data at user-selected levels of granularity.

[Supplemental material is available for this article.]

Sequence-based functional genomics assays are generating vast amounts of data that map the occupancy of specific transcription factors, the chemical status (such as acetylation and methylation), and positions of chromatin components such as core histones, the loading of RNA polymerases, and domains of DNase I hypersensitivity across the human genome at high resolution (Barski et al. 2007; Johnson et al. 2007; Mortazavi et al. 2008; Hesselberth et al. 2009; for review, see Pepke et al. 2009). Such measurements are now being made for a myriad of cell types, states, and tissues by individual laboratories and by large consortia such as ENCODE and the Epigenome Roadmap (Bernstein et al. 2010; The ENCODE Project Consortium 2012). This wealth of data contains rich, complex, combinatoric information about the inputs and outputs of gene regulatory networks (GRNs) that define each cell type and state. However, it is not yet easy to extract and distill biologically meaningful relationships, especially not on the multiple scales that range from broad global relationships to fine-grained ones that affect small groups of similarly behaving genes or subgenic regulatory elements.

Numerous prior studies have focused on understanding the relationship between an increasingly complex histone modification "code" and the activity state of DNA elements, such as transcriptional enhancers, insulators, promoters, and more or less vigorously transcribed regions for a given cell type or tissue (for review, see Hon et al. 2009). Furthermore, apparent cross talk between context-dependent histone modifications suggests a complex grammar (for review, see Lee et al. 2010). Pioneering analyses focused on specific ad hoc combinations of modifications found in the proximity of transcription start sites (TSS) or in selected distal intergenic regions (Barski et al. 2007; Wang et al. 2008).

More recent approaches have been more general and agnostic, dividing the entire genome systematically, either at regular intervals or based on the data (i.e., "segmenting" the genome) and then classifying the resulting genome segments (regions) into five to 100 states of chromatin mark combinations (classes) by applying statistical or machine learning methods such as Hidden Markov

[12]These authors contributed equally to this work.
[13]Corresponding authors
E-mail woldb@caltech.edu
E-mail ali.mortazavi@uci.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.158261.113. Freely available online through the *Genome Research* Open Access option.

23:2136–2148 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

Models (HMMs) or Dynamic Bayesian Networks (e.g., Ernst and Kellis 2010; Hoffman et al. 2012). The resulting machine-derived "states" are then semi-manually annotated to relate them to functions such as gene activation or repression. However, it is not clear a priori if the limited numbers of states used in these analyses, partly for ease of interpretation, fully or optimally capture the biological richness in the data, especially for the much larger and more diverse collections of data sets now being generated by projects such as the ENCODE and NIH Roadmap Epigenomics Projects.

The self-organizing map (SOM) is an unsupervised machine-learning method that was developed to cluster and visualize high-dimensional data (for review, see Kohonen 2001). It projects high-dimensional data onto a two-dimensional map composed of many units, each of which can be regarded as a mini-cluster, defined by its associated prototype vector of component weights. SOMs capture similarity relationships present in the training data as map topology, such that individual neighboring hex-units can subsequently be clustered after training into "metaclusters" as appropriate. This is analogous to the way biologists typically interact with RNA expression patterns and subpatterns in a classic two-way hierarchical clustering (Eisen et al. 1998). Indeed, SOMs with modest map sizes of less than 100 units have been used for more than a decade for clustering gene expression data (Golub et al. 1999; Milone et al. 2010; Newman and Cooper 2010; Spencer et al. 2011) or modest numbers of other genomic data sets (Moorman et al. 2006; Suzuki et al. 2011). While SOMs with small map sizes produce results that are generally equivalent to K-means, SOMs with thousands of units on boundaryless maps can show emergent behavior (Ultsch 1999). We reasoned that large SOMs should be able to capture a greater variety of combined chromatin mark patterns compared with methods that find a relatively small number of chromatin states, and that the resulting organization could be more readily visualized and ultimately mined in an intuitive way. Specifically, we anticipated that a large SOM, constructed from multiple genome-wide data types, collected across biologically distinct ENCODE cell types, would begin to reveal patterns of active, cell-type-specific transcriptional control elements based on their associated chromatin signatures.

As a first test of these possibilities, the trained ENCODE chromatin SOM presented here displayed distinct spatial organization that reveals how combinations of histone marks, DNase I hypersensitivity, and RNA polymerase occupancy correlate with gene features and activity, such as a relatively large supercluster of transcription start sites (TSS) that are active in one or more cell types, or a cluster of genes repressed in another cell type or types. We show how additional ChIP-seq, RNA-seq, transcription factor binding motifs, and other functional data can be placed on the chromatin map to identify and interpret cell-type-specific regulatory elements and transcription start sites. We then hierarchically cluster the SOM hex-units to explore global relationships of the different data sets on the SOM. Gene Ontology (GO) analysis reveals distinct enrichments in individual, often neighboring, units on the map related to cell-type-specific gene regulation. Finally, we introduce an interactive web interface to facilitate further mining of the ENCODE SOM and apply it to the analysis of cell-type-specific EP300 (also known as p300)–enriched units.

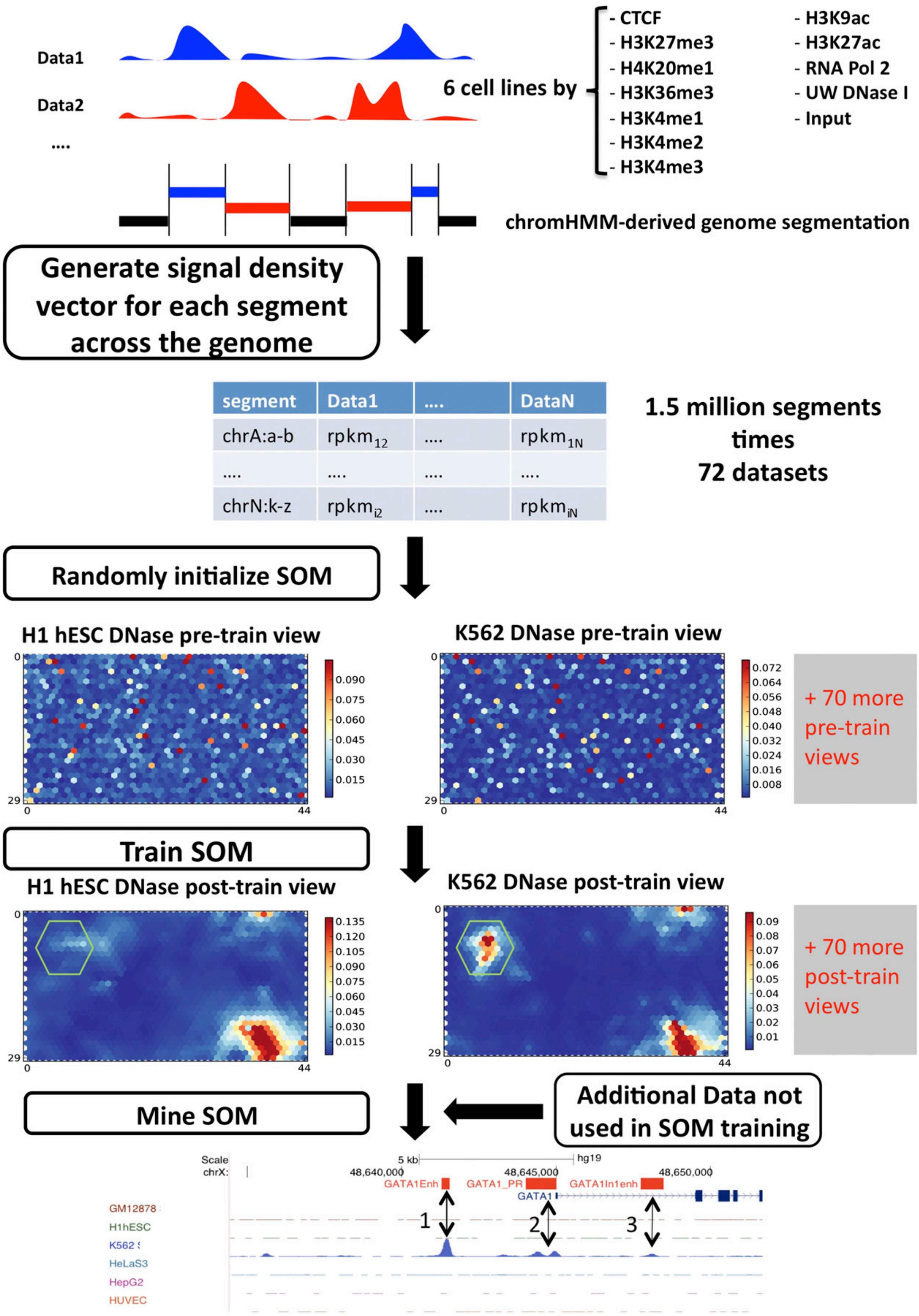
Results

Chromatin SOM construction and overall organization

The workflow for building a chromatin-based SOM begins with primary data mapping and genome segmentation and ends with visualization and data mining (Fig. 1). Briefly, the first step is to computationally break the genome into "segments" based on the data. The goal of segmentation is to define, across the entire genome, DNA segments that share the presence and absence of marks in the input data. To coordinate our results with other ENCODE Project Consortium work (The ENCODE Project Consortium 2012), we used a specific genome segmentation generated on 84 preselected data sets of eight histone modifications, RNA polymerase II, and CTCF from ChIP-seq, ChIP input control, and three open chromatin assays across six cell types using a "stacked" segmentation generated with ChromHMM (Ernst and Kellis 2010). We then constructed a training matrix consisting of the signal density for 72 of these data sets for each of the 1.5 million individual genome segments using only one of the DNase-seq assays to represent open chromatin. The Methods and Supplemental Figure S1 describe how the stacked segmentation differs from other segmentations of the same data.

We used the resulting matrix of 1.5 million 72-dimensional data vectors to train a SOM with map size of 30 rows of 45 columns (1350 units), and selected the best out of 10 maps based on the lowest quantization error (Methods) (Supplemental Fig. S2). The size of the map was selected to allow us to recover at least a thousand distinct states, if they were present in the data. In a uniformly distributed untrained map, we would expect 1170 segments/unit and 2.2 Mb/unit, on average. This map is a toroid, meaning that the top units on the map are seamlessly connected to the bottom units, and that the same applies to the leftmost and rightmost units (Supplemental Fig. S3). We chose the toroid form because it has no boundaries, which should prevent it from compressing clusters into map corners. To display a toroid map in two dimensions, we "slice it open," and some clusters are therefore visually split; that is, they "wrap around" the top edge to the bottom and from the left edge to the right, as indicated by the arrows (Supplemental Fig. S3). All assignments of segments to SOM hex-units are available for this SOM as a single bed file (Supplemental Table S1).

The distribution of DNA segments and nucleotides on the untrained map was without pattern and relatively even, while the trained map was much more uneven (Figs. 1, 2). This is expected because the segments on the trained map have been organized into clusters that contain differing segment numbers and nucleotide densities. For example, many of the larger DNA segments had little to no signal for any data set, and they were sequestered into a relatively small fraction of the SOM; on this 30-by-45 map, 48 contiguous units (3.5% of all units) captured 38% of the entire genome sequence, and is shown as high nucleotide density and segment count in Figure 2, A and B. The remainder of this map is dedicated to more finely parsing segments that have some signal in at least one of the training data sets. These overall organizational properties were not specific to this particular instance of the SOM nor to the ENCODE chromatin data. The top-scoring ENCODE SOM was very similar to the next nine best-scoring SOMs, each trained independently on the same input data, but from different random initializations. Specifically, we found that, for all of the units and regions of the SOM discussed below, segments within the same unit were clustered on the other nine maps within the same unit or adjoining units >80% of the time (Fig. 2C). We further analyzed the effect of leaving individual data sets out by retraining SOMs with 72 combinations of 71 data sets each and repeating the reproducibility analysis. We found that map reproducibility was robust to the removal of any one of 29 data sets (listed in Supplemental Table S2). While no single group of data sets was completely re-



segment	Data1	....	DataN
chrA:a-b	rpkm_{12}	....	rpkm_{1N}
....	....	....	....
chrN:k-z	rpkm_{i2}	....	rpkm_{iN}

Figure 1. Training the self-organizing map and general overview of data analysis. The genome is first segmented based on the signal density of input data sets. Any segmentation approach can be applied; in this case, the ChromHMM-derived segmentation in the primary publications by The ENCODE Project Consortium was used. The signal density is calculated for each segment and each data set, resulting in an input matrix of $M \times N$ dimensions, where M is the number of segments and N the number of data sets. The SOM is then initialized randomly from the input matrix, and trained. Additional data sets, not used for training, can then be mapped to the SOM, and these mappings and the distribution of segments on the trained SOM can be mined for interesting biological relationships.

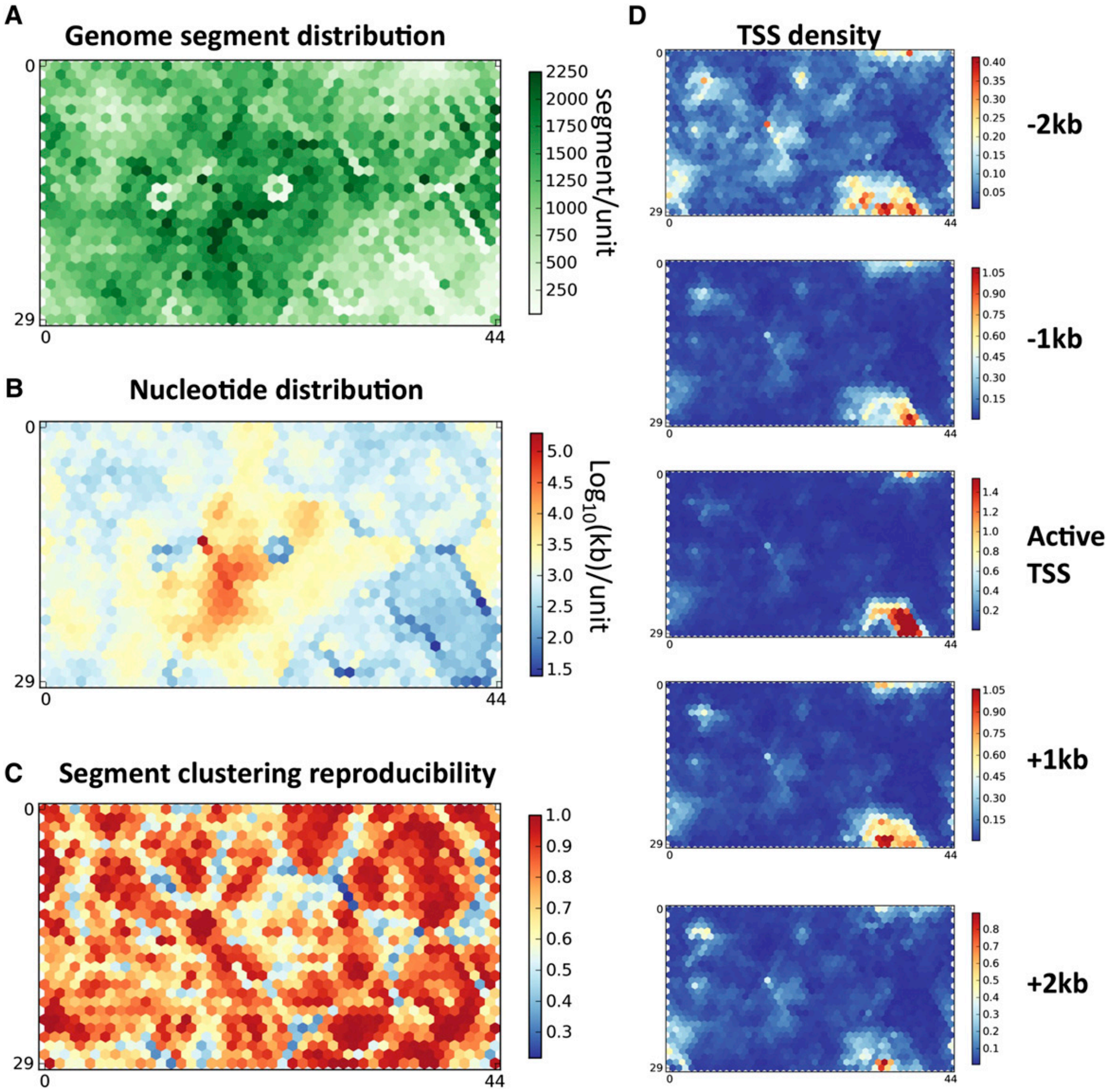


Figure 2. Map organization. (*A*) The segment count distribution over the map is uneven. While the average number of segments per unit is 1170, individual units range from 30 to 9334 segments. Note the distinct 1-unit-wide boundaries that contain very few segments separating denser regions. (*B*) The nucleotide distribution reflects the segment count, with the units with the most segments also containing the most nucleotides. These segments are also larger, thus accounting for the large portion of the genome that has little to no signal. (*C*) Reproducibility of clustering of two segments in the same unit or adjoining units as described in the text. (*D*) TSS-centric organization of active proximal promoters. The unit densities of points −2 kb, −1 kb, 0 bp, +1 kb, and +2 kb of GENCODE 7 TSS show the distinct organization of active promoters driven primarily by a common set of genes expressed in more than one cell type.

dundant, we found that three groups of data sets (H3K9ac, H3K36me3, and Control) were redundant in four out of six cell types, whereas another group of data sets (RNA Pol II, DNase I, and H3K4me3) was redundant in only one of six cell lines. Interestingly, the removal of these apparently redundant data sets still affected the reproducibility of a distinct subset of units, suggesting that they still contributed to the organization of the SOM in restricted regions of the map. These results argue that our SOM is robust and stable, and that segments with similar signatures are stably located near each other on the map, even though such segments do not always fall into a single hex-unit on independently trained SOMs. Local differences of the latter kind are expected for a nondeterministic method and can be discriminated from major differences, as shown below.

The SOM displayed several distinctive, very-low-segment-count "boundaries," usually just one unit wide and with as few as 30 segments/unit (Fig. 2A,B). These are, in effect, boundary units that separate clusters located on either side and that are characterized by distinct mark profiles. For example, H3K4me3-enriched segments are segregated from CTCF-associated ones in an adjacent map region (Supplemental Fig. S4).

We next explored where transcription start sites (TSS) map on the ENCODE SOM. No explicit information on annotated TSSs was used in building this SOM. Our expectations were that active TSSs would share a set of features present in the training data, including high DNase I hypersensitivity, RNA polymerase II occupancy (in varying intensities), H3K9ac, and H3K4me3 marks. This predicts that active TSSs would generally cluster together some-where on the SOM. In contrast, inactive TSSs were expected to lack these marks and, additionally, they might or might not show a repressive mark signature. We therefore expected inactive TSSs to occur elsewhere on the map, sequestered into one or a few clusters, depending on whether they have no other data from the training set or contain repressive mark data. A further expectation was that the SOM would detect and subcluster segments according to the intensity of their active-TSS signatures, since we had not reduced the data to simple present–absent calls for signal, but had retained all the quantitative information in the primary data. Finally, we expected that the SOM would subcluster active TSSs according to the cell type or combinations of types in which they were active.

All of the above expectations were met. A prominent region of the map, having relatively low segment and nucleotide density, showed the highest fractional enrichment in the number of GENCODE 7 (Harrow et al. 2012) TSS, with 27 units passing a threshold of 0.8 TSS/segment (Fig. 2D). Note that each TSS in this analysis was mapped as a single nucleotide, and was therefore assigned to only one DNA segment, even if there were several neighboring segments with very similar histone mark data. For this reason, we do not expect every DNA segment with an active TSS histone mark signature to score positive in this tally. As expected, the prominent TSS domain in the lower-right quadrant of the SOM corresponded with a domain of maximal DNase I hypersensitivity, as illustrated by comparing this with H1-hESC DNase-seq data (cf. Fig. 1 DNase I panels with Fig. 2D).

We next asked how DNA sequences located at varying distances from the nearest active TSS are organized on the map and found that 35 units are enriched in segments within 2 kb of these TSSs. We expected that near an active TSS, the chromatin signature would be very similar to the TSS point nucleotide for many segments, but that some segments would now display “mixed” chromatin signatures that retain some qualities of a pure TSS and add some characteristics of nearby chromatin. Such a “neighborhood” effect reflects properties of the original ChromHMM segmentation process as well as the biology of the histone mark pattern in each input cell type. As the distance from the TSS increases into the gene body or into the upstream promoter region, the histone signatures changed. On average, the distinct enrichments of single nucleotides that are located at −2 kb, −1 kb, +1 kb, and +2 kb from the TSSs in neighboring units demonstrates that the map has spatially clustered active promoters and their immediate upstream and downstream regions (Fig. 2D).

The prototype vectors for the units in the active-TSSs region revealed that most DNA segments at the center of this region possess signatures of expression in more than one cell type, although some adjacent clusters are cell-type-specific. When examined for RNA expression pattern and GO terms, the shared ones were housekeeping and other genes common to the cell types in this study, as expected. Investigating even more closely, we observed that individual units parse the levels of associated chromatin marks (e.g., high vs. medium vs. low H3K4me3) and the magnitude of the RNA polymerase signal, in different data sets and cell types. As discussed below, a user can drill even further down to select and extract DNA segments from hex-units with particular signature characteristics by using the SOM viewer and its associated DNA segment database.

Inspection of the SOM also reveals that multiple histone modification marks, previously shown to be associated with active transcription or active repression, drove the organization of the majority of the map (e.g., H3K4 mono-, di-, and tri-methylation, and H3K27me3 for activation and repression, respectively). This emphasis was expected, as several histone marks associated with active transcription tend to produce strong ChIP signals that are localized over relatively short DNA regions. The information-rich map regions typically show distinctive quantitative and qualitative combinations of marks. Most component planes, such as the ones shown for RNA polymerase II or H3K4me3 occupancy in the cell line GM12878 (Supplemental Fig. S4), form a single, internally connected cluster for their respective signal densities on the toroid. However, several other marks such as H3K4me2 and H3K27me3 have more than one distinct cluster on the map. This pattern suggests that they are found together with at least one other different additional chromatin profiles(s), or that regions rich in these marks are distinctive for individual cell types, or both (all component weights are displayed in Supplemental Figs. S5–S10). We return to dissecting the more complex patterns below.

Interactive SOM viewer for visualization and mining

We created an interactive JavaScript web-based SOM viewer with an associated map segment database to facilitate these explorations (http://woldlab.caltech.edu/ENCODESOM). It allows users to visualize and compare units on the map with respect to any input data set or to additional data types (see below), to find properties of different regions of the map, such as Gene Ontology enrichments, and to mine the segments in a given hex-unit or cluster. The interface for version 1.0 consists of five tabs: Training Data, TSS, GO, Other Data, and Clusters, which correspond to the results in this manuscript. A tool for highlighting groups of hex-units in one view and then seeing that outline on any subsequent view aids in evaluating the relatedness of one distribution (RNA polymerase II, for example) with another (TSS annotation or CAGE tags). Users can click on individual units and find the associated segments, genes, and GO-enriched genes. They can also select their own set of units and flag them across the different views of the data. This allows users, for example, to highlight a cluster of interest in the Cluster tab and see the clustering reproducibility of those highlighted units in the Other tab.

By using the viewer to ask how data from the input data sets are clustered and how those clusters relate to each other, one immediately sees the overlaps of units high in DNase I hypersensitivity, H3K9ac, H3K27ac, H3K4me2, and H3K4me3. Had we not known prior to this study that these chromatin signatures are affiliated with active promoters, the SOM would have allowed us to readily discover these relationships. Even knowing these general relationships, the SOM allows us to mine for fine structure that includes more complicated profiles of cell type specificity.

In contrast, we detected little overall change in H4K20me1 across the cell types and little affiliation of this mark with other signals, which leads segments high in those marks to cluster in a single location (upper-left quadrant of the map, Supplemental Fig. S11). Finally, we saw that the RNA Pol II component plane enrichments showed a gradient of RNA Pol II signal centered on a single unit that has the highest signal, which emphasizes that the SOM is clustering on the presence of the signal and also on its intensity. Units immediately around it have lower RNA Pol II intensity, and a user could then mine these, asking what additional information (possibly other marks and/or cell-type patterns) are distinguishing them from the single peak RNA Pol II unit.

Overlaying other ChIP-seq and functional data to find additional relationships

The SOM can also be used to test predictions, mine associations, and map relationships for data sets that were not used to train the

SOM. We began by exploring evidence for cell-type-specific *cis*-regulatory modules (CRMs) in the erythroid/monocyte lineage (K562) and in embryonic stem cells (H1-hESC) (Fig. 3). The transcription factors GATA2 and SPI1 (also known as PU.1) are important in erythroid differentiation, while POU5F1 (also known as OCT4) and NANOG are critical for defining embryonic stem cells. ENCODE ChIP-seq occupancy data for each factor was mapped onto the SOM (Fig. 3E–J). Occupancy for each factor was concentrated in two cell-type-specific clusters, one in the upper-left quadrant, and the other in the lower right (wrapping around to the top right, due to the continuous structure of the map). We then asked how these clusters relate to each other within each cell type, across cell types, and with underlying histone-mark signatures.

In K562 and H1-hESC cells, the upper-left quadrant of the SOM was prominent for the concentration of histone marks H3K27ac and H3K4me1, which have been affiliated with active enhancers and some promoters in previous studies. When H3K4me1 domains are outlined for K562 and H1-hESC (hexagon and triangle, respectively), prominent cell-type specificity is shown by the fact that they are largely separated (Fig. 3C,D). However, there is also a small domain of overlap, reflecting a few units in which similar chromatin signatures exist in both cell types.

We next asked how SOM domains of enhancer-associated histone marks are related to transcription factor occupancy data. We used well-studied factors that regulate hematopoetic target genes (GATA2 and SPI1) in K562 cells, and factors that regulate pluripotence target genes (NANOG and OCT4) in H1-hESC cells. When we overlaid the H3K4me1 chromatin outlines onto these individual factor ChIP-seq data views (Fig. 3E–H), the factors clearly coclustered with the enhancer histone marks in a cell-type-appropriate manner.

These transcription factors, plus PAX5 and SPI1 in the cell line GM12878 (Supplemental Fig. S12), also display some concentration of ChIP-seq signal in the lower-right portion of the map, where active TSS and their adjacent promoters are concentrated (Fig. 2D) and where H3K4me3, a mark of active and poised promoters, is strongly concentrated (Fig. 3A,B). This active TSS and peri-TSS domain of the SOM had especially prominent signals for SPI1 and NANOG, suggesting that these factors are associated by direct binding at or near promoters, or that they are otherwise physically engaged with promoter/TSS bound proteins (i.e., through protein:protein interactions that are recovered in ChIP). It is notable that there is a much weaker concentration of GATA2 in this SOM region. Taken at face value, this suggests that GATA2 is mainly associated with nonpromoter CRMs rather than with the peri-TSS domains, and that SPI1 has the opposite preference in K562.

Another expectation is that functionally active transcription factor occupancy will be marked with enhancer signatures (H3K4me1, H3K4me2, H3K27AC, and DNase I hypersensitivity). Active transcription factor occupancy is expected to be a subset of all sites of occupancy that should overlap with independently validated *cis*-regulatory modules (CRMs). We therefore asked where known CRMs are located on the SOM by taking advantage of a manually curated set of 118 erythroid CRMs. This set contains both distant enhancers and promoters. The CRMs localized prominently to the enhancer- and TSS-proximate zones of the map in K562 cells (Fig. 3K), with those in the enhancer area showing clear preference for the GATA2-enriched cluster of units (Fig. 3E). As would be predicted, the erythroid CRM map units are also enriched for K562-specific active enhancer histone marks and EP300 occupancy (Fig. 3C,I) that do not overlap with H1-hESC-specific enhancer marks and EP300 (Fig. 3D,J). A single hex-unit containing 979 genomic segments was most prominent for known erythroid CRMs, and we investigated it further (Fig. 3M,N). Remarkably, this single unit contained 11% of all high-confidence EP300 ChIP-seq peaks in the genome for K562 (P-value < 10^{-100}), and these overlapped strongly with segments also occupied by GATA2. The contents of this unit can now be further mined and tested to learn whether features lacking EP300 occupancy nevertheless contain active enhancers.

Functional CRMs are also expected to contain conserved sequence motifs that are targets for direct DNA binding. We used motifs curated from the literature for PAX5 and GATA2, along with closely related ones derived from ChIP-seq data, as defined by The ENCODE Project Consortium (The ENCODE Project Consortium 2012). We used phastCons conservation scores (Siepel et al. 2005) to compile a set of conserved motifs for each factor. We then mapped the locations of conserved instances of these motifs onto the SOM. As many transcription factor motifs in eukaryotes are short, they can occur within conserved domains for reasons other than being part of CRMs (i.e., being located with the coding portion of genes). Other instances of the motif are expected to be conserved on account of functioning in cell types or states other than this one. For these reasons, a dispersed map is expected. Nevertheless, NANOG motifs (Fig. 3L) and GATA motifs exhibited clear clustering, concentrated around the stem-cell-specific and erythroid-specific enhancer clusters of units.

Although we are herein primarily concerned with analyzing the ChromHMM-derived segmentation, we have also tested the behavior of the SOM using a naïve, 200-bp segmentation, as described in the Methods. We found that the map shows anisotropy, with enhancer-like and repressed regions more likely to cocluster, but with significant differences in some of the promoter regions. We conclude that the details of the segmentation do matter to a certain extent and that the particulars of each segmentation will interact differently in a way that depends on the data itself.

Taken together, these observations demonstrate the ability of a multi-cell chromatin SOM to concentrate and reveal cell-type-specific regulatory regions, and to allow users to visualize important patterns and relationships between transcription factor occupancy, candidate binding sites, chromatin signatures, and curated functional elements. Other relationships not shown in this set, but strongly visible in the data, include DNase I hypersensitivity and RNA Pol II occupancy. The ENCODE SOM-viewer allows users to explore these relationships by selecting views and marking the boundaries of one or more areas of interest based on more than 96 data sets.

SOM metaclusters capture regional and global properties of histone mark combinations

In addition to fine-grained unit-level clustering of relatively small numbers of segments into each unit done by the SOM itself, we can further cluster the unit prototype vectors across the entire map into metaclusters. We expect this level of analysis to be useful for further probing global genome-scale organization captured by the structure of the SOM. This clustering emphasizes more complex combinatoric chromatin signatures and thus augments the way we have already observed groups of units that cluster together based on the component plane of one training set (e.g., H3K4me1).

The full phylogenetic ordering of all units (Fig. 4A) is fine-grained, and it can be interpreted by a user visually in much the same manner as a phylogenetic ordering of genes. We also per-

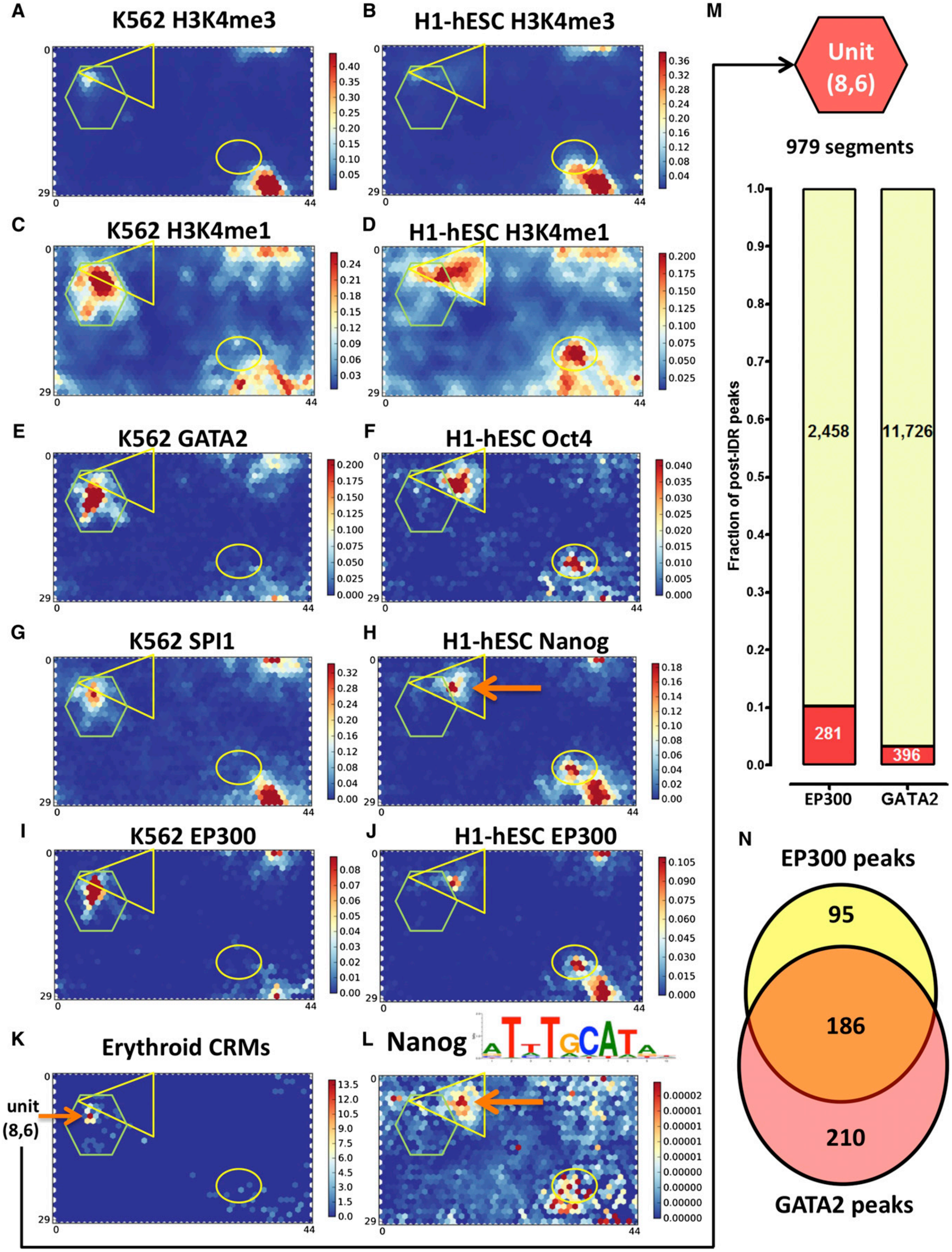


Figure 3. (Legend on next page)

formed an automated clustering to produce a nonsupervised set of boundaries for metaclusters of SOM units that are more similar to each other (based on their unit vector) than they are to other SOM units (see Methods). As with phylogenetic clustering of a single measurement, such as gene expression, we expect the phylogenetic ordering to be composed of graded similarity groups, rather than homogeneous and starkly bounded clusters. This is what we observed when we surveyed a stepped series of similarity thresholds versus metacluster number. The internal data structure identified several natural discontinuities as a function of clustering threshold, and we then selected three of these for full clusterings (Supplemental Fig. S13) to provide users with choices. Prominent driving relationships for the 126 cluster set that we found to be the most useful in our mining are shown in Figure 4B. Finally, we show the specific composition of each cluster for the 126-cluster instance (Supplemental Fig. S14).

The metaclusters showed enrichment patterns that are either cell-type-specific or common across multiple cell types. For example, cluster 1 contains 12 units that have high H3K36me3, RNA Pol II, and H4K20me1 in HUVEC cells (Fig. 4C,D). Different units within cluster 1 differ from each other based on which additional data sets are enriched in that unit. For example, two of the 12 units also show an additional enrichment for H3K36me3 and RNA Pol II in H1-hESC cells. The metaclustering captured features described in earlier sections, such as the active TSS region, and the K562-specific TSS with SPI1 region that corresponds to specific metaclusters, respectively.

Overall, the marks generally associated with active transcription, either at promoters or distant transcriptional enhancers, such as H4K4me1/2/3, H3K9ac, H3K27ac, and DNase I hypersensitivity, clustered in a cell-type-specific manner, whereas H3K36me3 and H4K20me1 clustered together by data type (Fig. 4E). The repressive mark H3K27me3 component planes also clustered together to form an outgroup. The SOM shows that while there is a strong common core of units shared by all six CTCF component planes, they each have more specific enriched units at the periphery. Whether these reflect cell-type-specific CTCF binding or have an alternative explanation such as changed chromatin marks near consistently CTCF-occupied sites is uncertain, and both could be at work. Interestingly, CTCF and RNA Pol II both displayed some clustering by cell type, and some that joined with other active marks from the same cell type.

Some Gene Ontology terms have distinctive chromatin mark signatures

We asked if any Gene Ontology (GO) functional terms are enriched in individual SOM units. Two hundred and twenty-eight GO terms displayed statistically significant enrichment following a Bonferroni correction (P-value < 10^{-10}) at the unit level (Supplemental Table S3). As might be expected, these included enrichments in GO terms that correspond to actively transcribed genes, or to actively repressed genes (for example, neuron-specific genes in non-neuronal cells). Most GO terms (164) were enriched in <1% of the map (13 units or less), and some of these are very specific. For example, "extracellular matrix" is enriched in five neighboring units (Fig. 5), and further inspection suggested that this enrichment is driven by genes that are much more highly expressed in HUVEC than in other cells. The regional GO enrichments typically correlated with metacluster boundaries of the SOM. In the case of "extracellular matrix" (Fig. 5A), four of the five units are part of cluster 1 (Fig. 4C). Another 30 GO terms were enriched in >5% of the map units, and these were typified by broad categories relating to the housekeeping functions of the cell such as "cell cycle." These GO terms are particularly associated with units that are high in H3K36me3 in one or more cell lines. Thirty-four GO terms were enriched in 1%–5% of the map, and these were typically much more specific, developmental terms in units with particular histone mark combinations. The enrichment in specific units for "GTPase activator activity," for example, is driven by gene families that show similar signal profiles across cell lines; the top two hex-units correspond to segments that have a high ratio of H3K4me1 over H3K4me2 in HUVECs that are candidate HUVEC-specific regulatory elements. Similarly, "sequence-specific transcription factor activity" (Fig. 5B) is enriched primarily in units that have cell-type-specific H3K27me3, whether in all cell types or in only some, such as H1-hESC cells and HUVEC. The two units with the most enrichment in Figure 5B have many additional associated developmental GO terms (Fig. 5C) and differ based on the presence of H3K27me3 signal in embryonic stem (ES) cells for segments in both units, but only H3K27me3 signal in HUVEC cells for one unit. This fine parsing by the SOM is nicely illustrated within the HOXD cluster, where the anterior and posterior parts of the cluster are split between these two units (Fig. 5D).

EP300 ChIP-seq overlay and cell-type-specific candidate enhancer segments

We extended our analysis of ENCODE EP300 data sets from K562 by including GM12878, H1-hESC, and HepG2 cells to identify 45 cell-type-specific and common EP300-high units, accounting for 1.4% of the genome and 1.9% of the segments. We found that each cell type had its own specific set of units with high EP300 occupancy, whereas only a few units showed EP300 signal in more than two cell types (Fig. 6). These common EP300 units correspond to the common TSS region, whereas the cell-type-specific clusters are primarily more than 2 kb from the TSSs (Fig. 2D). We showed earlier (Fig. 3) that we found K562 EP300 ChIP-seq signal in

Figure 3. Organization of genomic functional elements on the SOM. A triangle, hexagon, and ellipse are superimposed to allow comparison between maps. (*A*,*B*) H3K4me3 signal density in K562 and H1-hESC. (*C*) The hexagon encompasses the K562 units high in H3K4me1. (*D*) The triangle and hexagon capture the two disjoint regions that are high in H3K4me1 in H1-hESC. (*E*) GATA2 signal, which was not used in the training, is high in a subset of the H3K4me10high units in *C*. (*F*) Similarly, POU5F1 is primarily found overlapping the H3K4me1 high units. (*G*,*H*) In contrast to GATA2 and POU5F1, SPI1 and NANOG are found primarily in units that are high in H3K4me3 (to the *lower right* of the ellipse) with less signal found at H3K4me1 high units. (*I*,*J*) EP300 signal (also not used in the training) is found either primarily at enhancers in K562, but promoters in H1-hESC. (*K*) More than one-third of known erythroid CRMs cluster into a single unit with coordinates (8, 6). (*L*) Conserved NANOG motifs (motif derived from NANOG ChIP-seq data). ChIP-seq occupancy and motif occurrences were defined by the uniform ENCODE ChIP-seq binding site and motif calling pipelines. Conservation was assessed using the 46-way vertebrate phastCons scores for hg19 downloaded from the UCSC Genome Browser. The scores for each unit in the motif maps were normalized for the total number of base pairs in the unit to avoid the map being dominated by units with very high number of base pairs in them. (*M*) Ten percent of EP300 ChIP-seq calls and 3.2% of GATA2 calls in K562 fall within the top erythroid-CRM enriched unit (8, 6). (*N*) Sixty-six percent of the EP300 peaks in unit (8, 6) overlap a GATA2 peak.

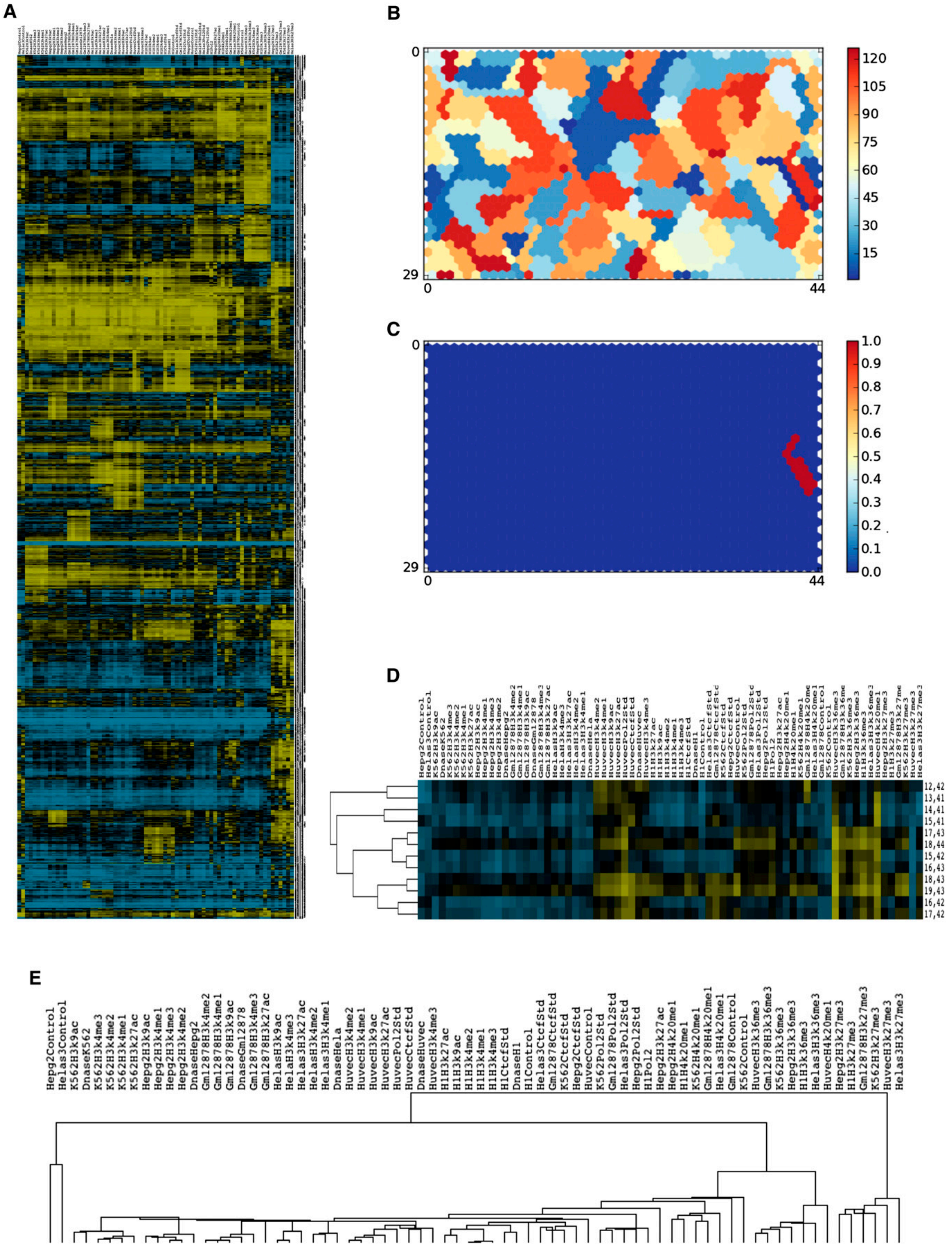


Figure 4. Metaclustering of the SOM. (*A*) Hierarchical clustering of the ranked unit weights (rows) and components (columns) shows both the large-scale and fine structure of the SOM unit ranked weights (yellow, high enrichment rank; blue, low enrichment rank). (*B*) Metaclustering of the SOM into ~120 clusters based on a consistency threshold of 2.6. (*C*) Twelve units make up metacluster 1. (*D*) Ranked component weights of metacluster 1. All 12 units share enrichment in HUVEC RNA Pol II, H3K36me3, and H4K20me1. Individual units show additional distinct enrichments, which distinguish them from one another. (*E*) Clustering of the component columns of Figure 5A, showing the relationships of the data sets to one another.

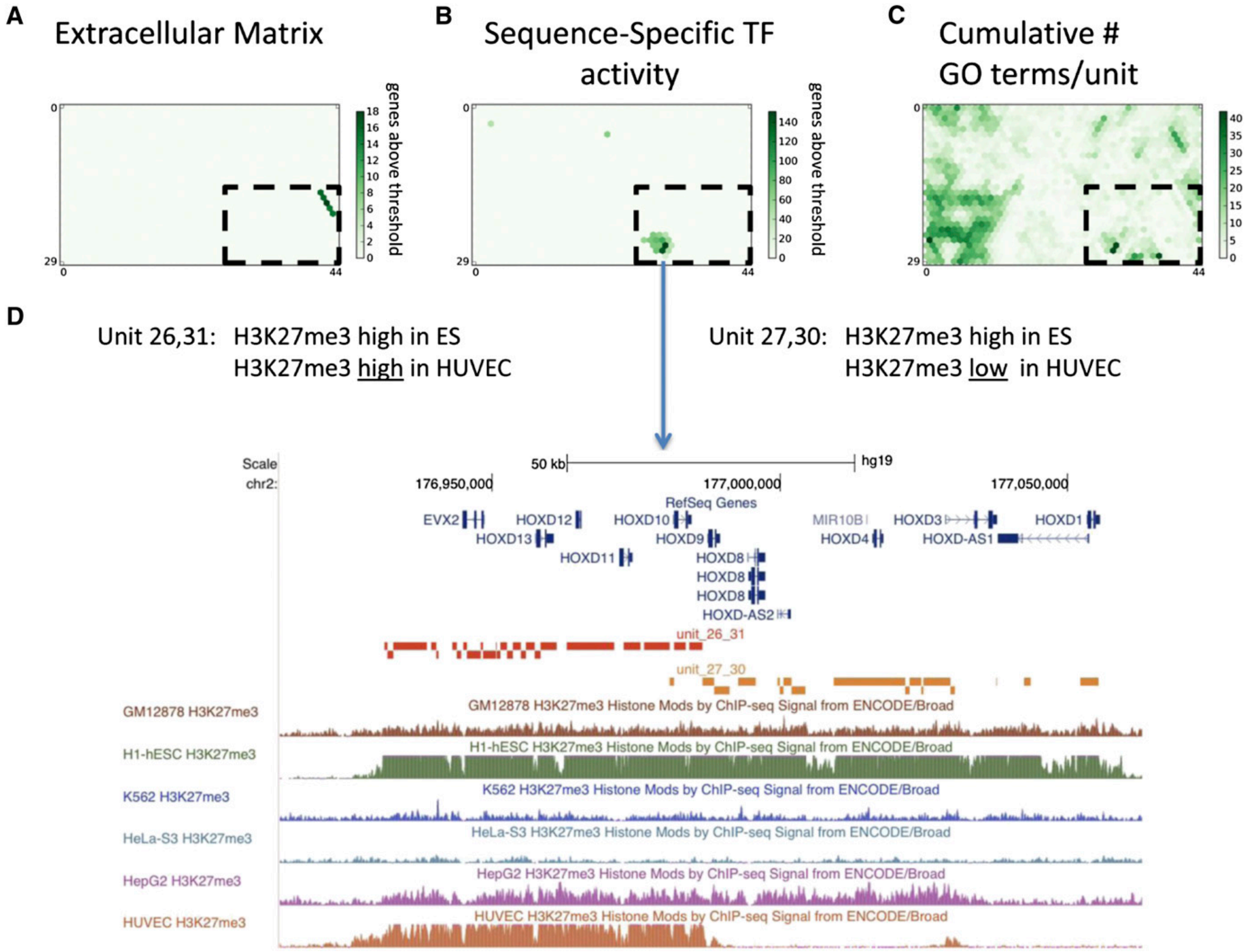


Figure 5. Specific patterns of GO enrichment over the SOM. (*A*) Specific GO terms such as "extracellular matrix" are highly enriched in portions of the map because of activity in one or more cell types. (*B*) Other GO terms are enriched because of their pattern of repression over the map. (*C*) The map has overall highly uneven distribution of GO enrichments away from the regions with the highest nucleotide density. (*D*) An example of the different patterns of H3K27me3 distribution across cell lines captured by neighboring units in the map in the HOXD cluster.

a cluster of units in the upper-left quadrant of the map that did not correspond to TSSs, but that did overlap with validated erythroid CRMs. These units are high in H3K4me1 and H3K27ac that are specific to each cell type. We then asked whether the segments within these units show functional enrichment. For example, three of the GM12878-specific units are enriched with the GO term "immune response." We can easily extend the analysis of the SOM by pooling segments from multiple units and analyzing them using tools such as GREAT (McLean et al. 2010) that associate *cis*-regulatory regions with genes for enrichment in many functional annotations besides GO. Applying GREAT to pooled segments from the cell-specific enriched EP300 units returned a wealth of enriched functional annotations that are predictably associated with the cell-type tissue of origin (Fig. 6). We illustrate this by showing enrichments in Pathway annotations for each cell type. Whereas the units with EP300 signal in more than two cell types are enriched in housekeeping pathways, the GM12878 units show the most enrichment in "immune system" and "interferon signaling," which nicely captures the biology of the cells. This functional enrichment of neighboring units on the map suggests richness of the SOM.

Discussion

Rapidly growing bodies of functional genomics data require methods to integrate and mine large numbers of data sets of multiple kinds. We constructed a self-organized map (SOM) of ENCODE chromatin data from 72 ChIP-seq and DNase-seq data sets from six ENCODE cell lines. Subsequent analyses and mining were facilitated by an interactive web-based SOM-viewer (http://woldlab.caltech.edu/ENCODESOM), which allows users to extend the analysis and extract groups of DNA segments that have characteristics of interest for further computational or wet-bench analysis. While most prior studies of global chromatin data have focused on a specific cell type or tissue, the ENCODE collection allowed us to explore relationships among multiple cell types in a single coherent analysis. By projecting high-dimensional chromatin data onto the two-dimensional SOM, we identified clusters of units

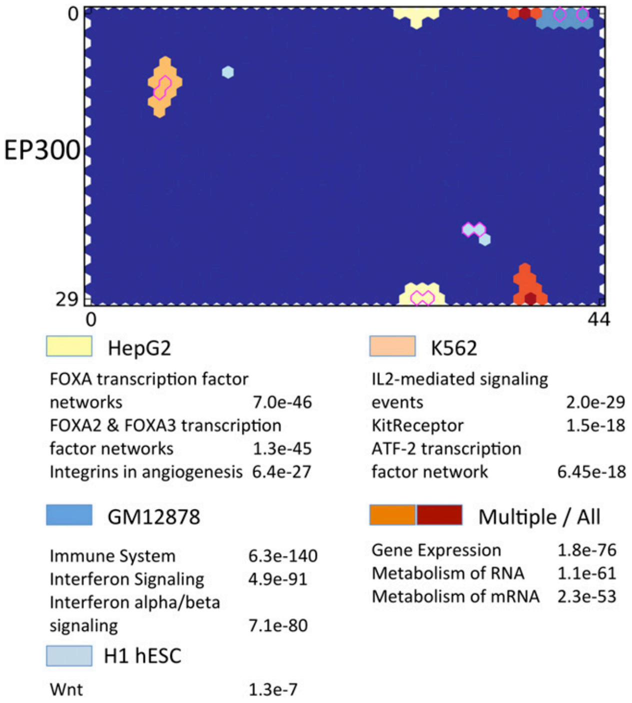


Figure 6. EP300 enrichment highlights cell-type-specific enrichments. ChIP-seq signals of the transcriptional coactivator EP300 in four ENCODE cell types were overlaid on the SOM. While some of the signal is common to multiple/all cell types (orange/brown), each EP300 ChIP-seq data set highlights a different set of adjoining units on the map that is specifically enriched based on the cell type. These cell-type-specific units are also high in H3K4me1 and H3K27ac, which suggest that they hold cell-type-specific enhancers. Segments from each of the colored clusters were pooled and analyzed for functional enrichment with GREAT such as pathways (*top* three terms per cluster shown). While the units common to multiple cell types are enriched in genes involved in housekeeping pathways, those in the cell-type-specific regions are enriched in pathways that are known to be relevant to the biology of those cells.

with chromatin mark combinations corresponding to promoter activity and transcriptional enhancer activity. These were further parsed into smaller clusters that were either cell-type-specific or more ubiquitous. By overlaying data for specific transcription factor binding, enhancer activity, and transcription start sites onto the SOM, we show that the user can discover relationships and mine corresponding genome segments of interest. This was demonstrated for known and candidate erythroid CRMs (Fig. 3). To our knowledge, this is the first use of self-organizing maps for multicell data integration and mining. Although we used a specific, "stacked" genome segmentation generated by ChromHMM, the overall approach can be applied to any segmentation. As discussed below, we expect that the choice of segmentation strategy and the mixture and quality of data sets used in training will affect the resulting SOM.

We mined the SOM to address specific classes of questions. First, individual training data sets revealed clusters that are cell-type-specific or shared for individual marks. The same was true for certain shared sets of marks. Second, units of the SOM were hierarchically clustered based on their prototype vectors, to investigate how multiple mark densities interact with each other. Third, additional data not used in training were projected onto the SOM to map their enrichment in one or more areas, and to relate the underlying chromatin characteristics to map units and clusters where other specific data features are concentrated. In this way, we investigated how individual sequence-specific regulatory factor occupancy for GATA2, SPI1, OCT4, and NANOG, their DNA binding motifs, and the EP300 coactivator are related to each other and to underlying chromatin signatures. Fourth, we mined the SOM for specific functional classes using transcription start sites (TSSs) as the best-defined test case, followed by a curated set of CRMs. The SOM segregated TSSs that are commonly expressed in multiple cell types from the TSSs with cell-type-specific activity into subclusters. Finally, we found that some individual GO terms are preferentially affiliated with different chromatin signatures. To facilitate exploration of the ENCODE SOM by users, we provide a web interface SOM viewer that allows users to explore all the data sets mapped here and to mine out the DNA segment coordinates in any hex-cell or group of cells. We expect this web interface to be the primary means by which users interact with the SOM results.

At the highest level, most observations agreed with conclusions of previous studies using other methods to integrate chromatin data such as hidden Markov models, which were applied to these ENCODE data (The ENCODE Project Consortium 2012). The SOM, however, provided an additional level of granularity that is not accommodated by a relatively small number of states. The SOM also lent itself well to visualizing relationships between the chromatin data and additional data of any type that can be mapped to specific points or intervals on the genome (and hence to the DNA segments in the map). The fine structure of the SOM allowed us to identify distinct combinations of marks and mark intensities shared by only a small number of genomic regions, and did so without any a priori decision about the number of states. For example, the SOM easily separated the variety of different types of TSS into a major cluster of active TSSs versus inactive ones. The active TSSs were internally more finely parsed, based on levels of H3K4me3, as well as distinct cell-type-specific units.

A summary analysis of new candidate transcriptional enhancers is shown in Figure 6. This aggregate analysis is the same one performed for K562 cells (Fig. 3) and uses EP300 signal from each cell type to further concentrate and focus on units active in individual cell types, as well as units that correspond to activity in multiple cell types. Just two units displayed activity in all participating cell types, while a surrounding set of units is variously multitype. Analysis of these units by GREAT showed that those active in all cell types are enriched for well-known housekeeping functions such as protein synthesis. The cell-type-specific units were enriched according to cell type (B lymphocyte, hepatocyte, embryonic stem cell), just as K562 showed erythroid and monocyte categories.

While much of the map organization was driven by histone marks associated with active promoters and enhancers, we point out that this is partly the result of the histone marks used in the ENCODE study for genome segmentation and SOM training. Our input histone marks to the ENCODE SOM clearly favored a fine parsing of active regions over passive ones, and important repressive marks such as H3K9me3 were not included. This makes the ability of this SOM to parse differences in H3K27me3 in different cell lines quite remarkable. Overall, the ENCODE integration efforts showed that a relatively small number of HMM-derived states can capture the broad landscape of active and repressed regions in the ENCODE cell lines (The ENCODE Project Consortium 2012), while the SOM detailed here does this and also gives the biologist access to a wealth of increased resolution and specificity that we coupled with visualization and mining tools. We anticipate that this kind of analysis will be even more useful as the number of cell types and diversity of chromatin marks increase in future studies, making the challenge of combinatoric signatures and their functional correlates greater. In a similar way, as transcription factor location data for many more factors accumulates, the SOM approach and tools developed here will enable end users to better identify and stratify the functionally important and interesting minority of occupied sites that are active in various subsets of cell types.

Methods

Rationale for training matrix design

The joint analysis of multiple cell types presents additional challenges beyond the analysis of multiple data sets in a single cell line. If each cell line is analyzed separately, one is left with the difficult task of trying to reconcile the states found for each with different definitions, before proceeding to analyze state changes between cell lines. Alternatively, one can "concatenate" the data from multiple cell lines (Ernst et al. 2011). Concatenation has the great advantage that the states defined will be consistent across cell lines, but this approach still requires intensive post-processing to extract the segments that change states across cell lines; assuming that a concatenated HMM had seven states in six cell lines, any given genomic segment could be in one of $7^6 = 117{,}649$ combinations of states. Another solution, which we implement here, is to train on all data jointly as a "stack" to learn a single set of states with a single set of genomic boundaries. In this case, one is then left only with the problem of how to interpret the states, whose definitions are virtually certain to involve nonintuitive, complex combinations of marks in one or more cell types and requires additional methods to mine the results in a systematic and intuitive way.

"Stacked" training matrix implementation

To train the SOM, we first built a training matrix composed of signal densities of all 72 data sets (columns) over all segments (rows). The segments were taken from a ChromHMM segmentation of a "stacked" training set of 84 data sets (ChIP-seq for eight histone modifications, RNA Pol II, and CTCF; and three open chromatin data sets for each of six cell lines) using 25 states. We set aside two of the open chromatin data sets to avoid overtraining on open chromatin, and only used the UW DNase-seq data to represent open chromatin as the three experiments are effectively redundant. We converted uniformly processed signal densities of the remaining 72 data sets used for the SOM training into RPKM (reads per kilobase per million reads) for every segment on each training data set using the ERANGE 3.3 getDensity.py script. The training matrix was built using the ERANGE 3.3 buildMatrix.sh script, with a maximum threshold of 100 RPKM and the rescale option.

Training the SOM

The self-organizing maps were trained and analyzed using ERANGE v3.3. For every SOM instance, we shuffled the training set, randomly initialized the toroid map of hexagonal units from the training set, and incrementally trained a SOM with map size 30 by 45 using 5 million iterations, which is equivalent to going through the entire data set 3.3 times, starting with an update bubble radius of 15 and a learning rate of 0.2, both of which decreased exponentially over the course of training. Each segment was assigned to its best matching unit based on the Euclidean distance. We selected for analysis the best of 10 trials based on the lowest quantization error, which is defined as the average Euclidean distance of all segments to the prototype vector of their assigned unit. The other nine instances were used to evaluate the reproducibility of the map by analyzing the fraction of segments from each unit of our best map that resided in the same unit or adjoining units in the other nine map instances.

While we decided to use the entire training matrix for training for the SOM discussed in the main text, the software supports training on the training set and scoring on a distinct test set. In particular, we trained 10 SOMs with half of the segments from the 200-bp naïve segmentation (i.e., half of 1.5 million segments) for 25 million iterations, selected the best one based on the scoring of the other half of the segments, and rescored the best SOM with the ChromHMM segmentation to provide directly comparable genomic coordinates.

There are no theoretical limits to the number of data sets, segments, or map size that could be analyzed with the SOM. However, the ERANGE implementation of the SOM was designed for compatibility with the rest of the package rather than for scalability or performance and will be significantly slower on much larger data sets or number of training iterations. The final training run for the main ENCODE SOM above took a couple of hours, while the naïve segmentation run took 1 d. The per-unit gene-level analysis took significantly longer.

Gene-level analysis

We recovered the identity of the nearest gene within 20 kb of each segment within a unit using the NCBI gene annotation, which is conservative and means that in lower gene-density areas of the genome, many segments were not affiliated with any gene. We then analyzed every unit for Gene Ontology (GO) enrichment as previously described (Mortazavi et al. 2006), adjusting for multiple-hypotheses testing by applying a Bonferroni correction for both the number of tested Gene Ontology terms and the map size.

Metaclustering methods

The unit prototype vectors were automatically aggregated into the larger clusters using standard hierarchical clustering, subject to the constraint that only adjacent clusters on the SOM could be aggregated. A centered correlation distance and centroid linkage were used. Prior to the hierarchical clustering, the prototype vector values along each dimension were replaced with rank values normalized to range between -1 and 1. Heat map visualizations of the hierarchical clustering were rendered using Java Treeview (Saldanha 2004). The clustering itself and the SOM visualizations of it were done using custom C++ and Python code (available at http://woldlab.caltech.edu/~spepke/somclustering/).

Partitionings of the hierarchical clustering at varying levels of detail were generated using the branch length inconsistency criterion implemented in SciPy (depth = 6). The inconsistency of a branch is the ratio of its length to the average length of branches to clusters less then a specified depth below it. For a specified threshold value *t*, the hierarchical clustering is cut at branches that exhibit an inconsistency coefficient greater than *t*. Partitioning of the unit vectors was performed over a broad range of values of *t* up to that for which no branch's inconsistency criterion exceeded *t*, i.e., only one cluster resulted. Sharp drops in the number of clusters as a function of the threshold value occur and are typically followed by plateaus that show little or no change in cluster number. Such behavior suggests partitionings that are relatively robust with respect to the threshold value (see Supplemental Fig. S13).

Acknowledgments

We gratefully acknowledge Ewan Birney, Ian Dunham, Eric Mjolsness, and Paul Sternberg for general discussion of SOMs and their applications to functional genomics; and Diane Trout, Henry Amrhein, and Anna Abelin for computational assistance. At Caltech this work was supported by grants to B.J.W. from the Beckman Foundation, the Donald Bren Endowment, the Gordon Moore Cell Center at Caltech, NIH U54HG004576, NIH U54HG006998, and RC2HG005573; at Hudson Alpha Institute by grant NIH U54HG004576 and NIH U54HG006998 to R.M.M.; and at Penn State University by NIH R01DK065806, U54HG006998, and RC2HG005573 to R.C.H. A.M. was partly supported as a Caltech Beckman Fellow and a Moore Cell Center Fellow while at Caltech. The Mortazavi laboratory at UC Irvine was supported by grant NIH U54HG006998 as well as EU-FP7 project STATegra (306000) to A.M.

References

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28:** 1045–1048.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis of genome-wide expression patterns. *Proc Natl Acad Sci* **95:** 14863–14868.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for the ENCODE Project. *Genome Res* **22:** 1760–1774.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6:** 283–289.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9:** 473–476.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18:** R195–R201.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316:** 1497–1502.

Kohonen T. 2001. *Self-organizing maps*, 3rd ed. Springer, New York.

Lee JS, Smith E, Shilatifard A. 2010. The language of histone crosstalk. *Cell* **142:** 682–685.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.

Milone DH, Stegmayer GS, Kamenetzky L, Lopez M, Lee JM, Giovannoni JJ, Carrari F. 2010. *omeSOM: A software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics* **11:** 438.

Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu X, White KP, Bussemaker HJ, et al. 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster. Proc Natl Acad Sci* **103:** 12027–12032.

Mortazavi A, Leeper Thompson EC, Garcia ST, Myers RM, Wold B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire. *Genome Res* **16:** 1208–1221.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* **5:** 621–628.

Newman AM, Cooper JB. 2010. AutoSOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinformatics* **11:** 117.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6:** S22–S32.

Saldanha AJ. 2004. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20:** 3246–3248.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. 2011. A spatial and temporal map of *C. elegans* gene expression. *Genome Res* **21:** 325–341.

Suzuki M, Oda M, Ramos MP, Pascual M, Lau K, Stasiek E, Agyiri F, Thompson RF, Glass JL, Jing Q, et al. 2011. Late-replicating heterochromatin is characterized by decreased cytosine methylation in the human genome. *Genome Res* **21:** 1833–1840.

Ultsch A. 1999. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen maps* (ed. Oja E), pp. 33–46. Elsevier Science, New York.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903.

Received March 29, 2013; accepted in revised form October 7, 2013.