Learning the human chromatin network from all ENCODE ChIP-seq data

Scott M. Lundberg[1], William B. Tu[2,3],
Brian Raught[2,3], Linda Z. Penn[2,3], Michael M. Hoffman[2,3,4], Su-In Lee[1,5]

[1] Department of Computer Science and Engineering, University of Washington
[2] Department of Medical Biophysics, University of Toronto
[3] Princess Margaret Cancer Centre
[4] Department of Computer Science, University of Toronto
[5] Department of Genome Sciences, University of Washington

Abstract

Introduction: A cell's epigenome arises from interactions among *chromatin factors* — transcription factors, histones, and other DNA-associated proteins — co-localized at particular genomic regions. Identifying the network of interactions among chromatin factors, the *chromatin network*, is of paramount importance in understanding epigenome regulation.

Methods: We developed a novel computational approach, ChromNet, to infer the chromatin network from a set of ChIP-seq datasets. ChromNet has three key features that enable its use on large collections of ChIP-seq data. First, rather than using pairwise co-localization of factors along the genome, ChromNet identifies *conditional dependence* relationships that better discriminate direct and indirect interactions. Second, our novel statistical technique, the *group graphical model*, improves inference of conditional dependence on tightly correlated datasets. These datasets include transcription factors that form a complex or the same transcription factor assayed in different laboratories. Third, ChromNet's computationally efficient method allows network learning among thousands of factors, and efficient relearning as new data is added.

Results: We applied ChromNet to all available ChIP-seq data from the ENCODE Project, consisting of 1,415 ChIP-seq datasets, which revealed previously known chromatin factor interactions better than alternative approaches. ChromNet also identified previously unreported chromatin factor interactions. We experimentally validated one of these interactions, between the MYC and HCFC1 transcription factors.

Discussion: ChromNet provides a useful tool for understanding the interactions among chromatin factors and identifying novel interactions. We have provided an interactive web-based visualization of the full ENCODE chromatin network and the ability to incorporate custom datasets at `http://chromnet.cs.washington.edu`.

Introduction

Chromatin factors — such as transcription factors, histones, and other DNA-associated proteins — interact with each other to regulate gene expression [10], the physical structure of the genome [7],

cell differentiation [4], and other cellular processes. Identifying this network of interactions among chromatin factors, the *chromatin network*, is important for understanding genome regulation and the function of each chromatin factor [49, 3]. To find interactions in the chromatin network we can use chromatin immunoprecipitation-sequencing (ChIP-seq) to measure genome-wide localization of chromatin factors, and then compare ChIP-seq datasets to find chromatin factors that co-localize [36, 8]. Co-localization may indicate that two factors interact directly, by forming a complex, or indirectly, such as by regulating similar DNA targets.

However, identifying pairwise co-localization alone fails to distinguish direct interactions from indirect interactions. Consider a simulated chromatin network among four factors, where factor C recruits A and B, and A in turn recruits D (Figure 1A, top). Because all pairs of ChIP-seq datasets are correlated to each other (Figure 1A, middle), a simple co-localization method would incorrectly infer indirect interactions among the factors (Figure 1A, bottom left). In a *conditional dependence network* (Figure 1A, bottom right), if two variables (here, factors) are *conditionally dependent*, then there is an edge between them. The *conditional dependence* between two factors measures their co-localization after accounting for information provided by other factors. If we infer a conditional dependence network, we eliminate indirect edges from the network, such as between factors A and B, because their co-localization at peaks 3 and 5 can be *explained away* by another factor C (C recruits A and B). Hence, incorporating more ChIP-seq datasets allows more indirect edges to be removed, resulting in a higher quality inferred network.

Here we present ChromNet, an approach that estimates the human chromatin network using a conditional dependence network among chromatin factors from 1,415 human ENCODE ChIP-seq datasets (Figure 2; Supplementary Table 1). Integrating all ENCODE datasets from many cell types into a single network provides several advantages. It enables extraction of global patterns in the relationships among chromatin factors. It also expands the chromatin network to include the union of all chromatin factors measured in any cell type. This is because a dataset for a chromatin factor in one cell type can serve as a proxy for a missing dataset on that factor in another cell type.

Learning this network involves two key challenges. First, learning a network among thousands of ChIP-seq datasets based on millions of samples (here, genomic positions) is computationally very intensive. To solve this challenge, we utilized an efficient approach that involves the computation of an *inverse correlation matrix* from binary data, which does not require an expensive iterative learning procedure. We show that our approach is as effective as more complex models in identifying validated protein-protein interactions. Second, identical or closely-related chromatin factors are often measured in different labs, conditions, or cell types. When some variables are highly correlated with each other, standard methods often learn edges only among these variables and disconnect them from the rest of the network (Figure 1B, middle) [2]. Incorporating more ChIP-seq datasets would exacerbate this problem. To solve this challenge, we developed the *Group Graphical Model* (GroupGM) that learns a network not only between individual chromatin factors, but also between groups of factors (Figure 1B, bottom). We show that GroupGM makes a conditional dependence network robust against data redundancy.

Previous work on learning interactions among chromatin factors from ChIP-seq data used much smaller data collections. ENCODE identified conditional dependence relationships among groups of up to approximately 100 datasets in specific genomic contexts [14]. Other authors used partial correlation on 21 datasets [25], Bayesian networks for 38 datasets [27], or partial correlation and penalized regression on 27 datasets [43]. Still other authors used a Markov random field with 73 chromatin factors in *Drosophila melanogaster* [55], or a Boltzmann machine with 116 human transcription factors [35].

ChromNet departs from previous approaches by enabling the inclusion of all 1,415 ENCODE ChIP-seq datasets into a single joint conditional dependence network. GroupGM and an efficient

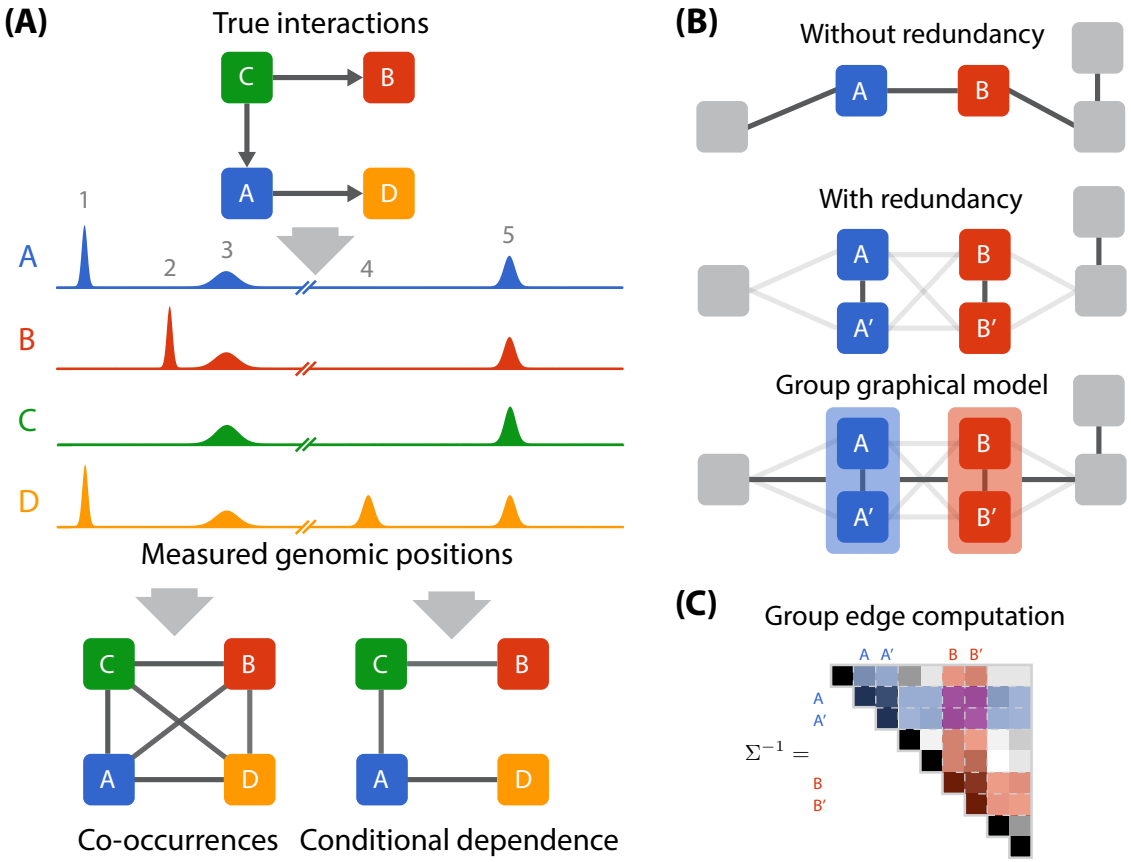

Figure 1: **(A)** *Top:* Interaction network among four simulated chromatin factors. *Middle:* Binding activity from simulated ChIP-seq datasets, where each peak represents a putative binding position of a protein. *Bottom:* Networks inferred from ChIP-seq datasets based on co-occurrence (left) or conditional dependence (right). **(B)** Redundant information obscures conditional dependence connections. *Top:* Without redundancy, standard methods robustly infer a conditional dependence network. *Middle:* Highly correlated variables (such as A and A') are strongly connected with each other and lose their connections with other variables. *Bottom:* A Group Graphical Model (GroupGM) represents the conditional dependence between groups of correlated variables, which restores the connection between A and B. **(C)** An inverse correlation matrix with highlighted columns for A, A', B, and B'. The strength of the group edge between (A, A') and (B, B') is the sum of the entries between the groups (purple).

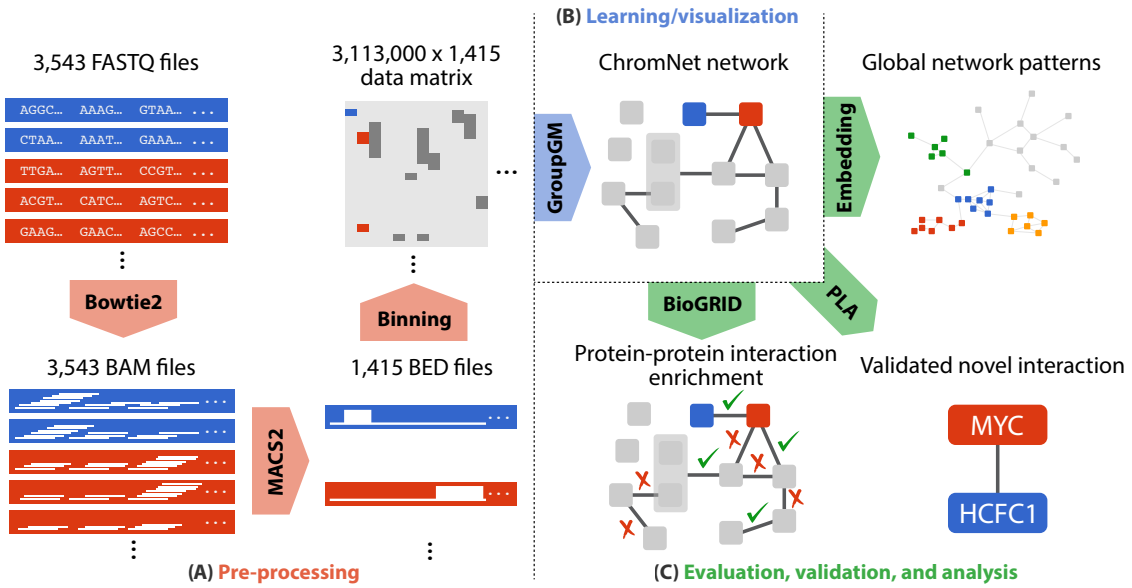

Figure 2: **(A)** Uniform processing pipeline includes aligning sequences using Bowtie2, calling peaks with MACS2, and then binning them into 1,000–base-pair regions. **(B)** We inferred the GroupGM from all 1,415 ChIP-seq datasets and integrated the learned model into a web interface to facilitate broad use. **(C)** We evaluated the learned model against known physical protein interactions (BioGRID), mapped global patterns through network embedding, and validated a novel predicted MYC-HCFC1 interaction with a Proximity Ligation Assay (PLA).

learning algorithm allow seamless integration of all datasets comprising 223 transcription factors and 14 histone marks from 105 cell types without requiring manual removal of potential redundancies (Supplementary Table 1). We show that this increases the power to detect previously known protein-protein interactions by approximately 50% (Supplementary Figure 1). We also demonstrate the potential of ChromNet to make new discoveries by experimentally validating a novel interaction.

Results

Uniformly processed binary data reduces noise when learning conditional dependence

To ensure comparable signals across all ChIP-seq datasets, we re-processed raw ENCODE sequence data with a uniform pipeline (Figure 2A). We downloaded raw FASTQ files from the ENCODE Data Coordination Center [8, 45, 11] and mapped them using Bowtie2 [24] to the human genome reference assembly (build GRCh38/hg38) [13]. Using MACS2 [54], we created a set of lenient peak calls that captured signal from both narrow and broad peaks. We binarized these peak calls within 1,000–base-pair bins across the entire genome, which results in a $3{,}113{,}000 \times 1{,}415$ binary data matrix X where genomic positions are viewed as samples (Figure 2A). Binarized data has less noise in long stretches of no-signal regions compared to transformed or raw read counts. We show that a network inferred from binarized peak calls better reveals previously known protein-protein interactions (Supplementary Figure 2; Supplementary Note 1).

A conditional dependence network can be efficiently learned from binary data

Learning a conditional dependence network among thousands of ChIP-seq datasets each containing millions of samples (genomic positions) requires an efficient algorithm (Figure 2B). It is well known that the nonzero pattern of the *inverse covariance matrix* of Gaussian random variables represents the conditional dependence network [26, 32]. The inverse correlation matrix, Σ^{-1}, is just a normalized version of the inverse covariance matrix and also represents conditional dependence. A zero element ($\{\Sigma^{-1}\}_{ij} = 0$) means that the ith and jth variables are conditionally independent of each other given all other variables—they are not connected by an edge. While it is common practice to learn the conditional dependence network among continuous-valued variables based on the estimation of Σ^{-1} [19], recently Loh et al. showed that Σ^{-1} also reveals conditional dependence in binary data for simple acyclic networks [30]. Through simulation, we found that Σ^{-1} can often reveal dependence from binary data for more complex networks as well (Supplementary Note 2).

ChromNet first computes the *inverse sample correlation matrix* $\hat{\Sigma}^{-1}$ from the binary data matrix X of 1,415 variables and 3,113,000 samples, and then uses elements of $\hat{\Sigma}^{-1}$ as weights of network edges (Figure 2B). We compared ChromNet with a conventional method for learning the conditional dependence network from binary data, the pairwise Markov random field model that has been applied to binary ChIP-chip data [55]. We evaluated both methods on fly modENCODE data ($73 \times 150{,}000$) [55] with and without L_1 regularization [19, 12]. To match the data processing used by the Markov random field implementation we did not normalize the data matrix, which results in an inverse covariance matrix instead of an inverse correlation matrix. We then examined how well these methods recovered known protein-protein interactions annotated in BioGRID [50]. ChromNet performed similarly to the Markov random field (Supplementary Figure 3; Supplementary Note 4), but runs much more efficiently. While learning a Markov random field is challenging for over 100 datasets, relying on $\hat{\Sigma}^{-1}$ allows ChromNet to learn a network from more than a thousand datasets across millions of samples in only a few minutes.

Group modeling mitigates the effects of redundancy

Many ENCODE ChIP-seq datasets contain redundant positional information. These redundancies arise from transcription factors in a complex or the same chromatin factor measured in different labs or different cell types. Conventional conditional dependence methods have a key limitation in modeling redundant data. If datasets A and A' are tightly correlated, a conventional method would connect A with A' but connect A to the rest of the network only weakly (Figure 1B). Arbitrarily removing or merging redundant datasets can hide or eliminate important information in the data (Figure 4B).

GroupGM overcomes challenges with redundant data in conditional dependence models. To do this, we define edges that connect groups of datasets (such as $[A,A']$ and $[B,B']$). A group edge weight represents the total dependence between the variables in the two groups that the edge connects, and is computed from Σ^{-1} as (Figure 1C, Methods):

$$G_{[A,A'][B,B']} = \Sigma^{-1}_{AB} + \Sigma^{-1}_{AB'} + \Sigma^{-1}_{A'B} + \Sigma^{-1}_{A'B'}.$$

An edge in a GroupGM model implies conditional dependence between the linked groups, but does not specify the involvement of individual factors in each group. We prove that GroupGM can reveal conditional dependencies in the presence of redundancy (Supplementary Note 3). In this paper, we learn the groups using hierarchical clustering applied to the correlation matrix among all ChIP-seq datasets, which results in only $2p-1$ groups, where p is the total number of ChIP-seq datasets (here, $p = 1{,}415$).

Conditional dependence and group modeling improve the recovery of known protein-protein interactions

To evaluate how conditional dependence and group modeling both contribute to the performance of ChromNet we estimated networks using the following methods, where each method produces a set of weighted edges:

1. **Correlation:** We learned a naive co-occurrence network, using pairwise Pearson's correlation between all pairs of datasets.

2. **Inverse correlation:** We learned a conditional dependence network, by computing the inverse of the correlation matrix.

3. **GroupGM:** We learned a group conditional dependence network, using our novel method that addresses tight correlation among datasets.

To assess the quality of the estimated networks, we identified the edges corresponding to published protein-protein interactions. As ground truth, we used the BioGRID database's assessment of physical interactions between human proteins from experiments deemed low throughput [50]. For evaluation, we excluded edges involving a histone mark because they do not exist in BioGRID. We also excluded edges connecting the same chromatin factor even when measured in different labs, cell types, or treatment conditions. These edges were excluded from evaluation to prevent them from artificially inflating the accuracy of the methods. When we measured the conditional dependence between a pair of ChIP-seq datasets in GroupGM, to avoid the inclusion of many redundant edges, we picked the maximum edge weight out of all network edges connecting groups that include the corresponding datasets. A large group would tend to contain a heterogeneous set of datasets; thus, we only considered the edges between groups that contain datasets corresponding to the same chromatin factor.

Dependencies learned within cell types

We compared performance of the three methods described above across a range of prediction thresholds. For each network, we varied a number N of evaluated edges from 1 to the total number of edges. For each value of N, we identified the set of N edges with the largest weights. We also randomly picked N edges without regard to weight rank as a background set. We then calculated how many edges in each set matched known protein-protein interactions from BioGRID. We computed fold enrichment by dividing the number of matched edges in the prediction set by the expectation of the number in the background set.

We first measured performance within all cell types, excluding edges between datasets in different cell types (Figure 3A). Since the limited number of annotations in BioGRID imperfectly represent the human chromatin network, one cannot draw too many inferences about absolute performance from this benchmark. Relative performance of the methods, however, is clear. Inverse correlation performed better than correlation, and GroupGM outperformed inverse correlation. This indicates that better resolution of direct versus indirect interactions contributes to improved performance of inverse correlation over correlation; and greater robustness against relationship-hiding redundancy contributes to improved performance of GroupGM over inverse correlation.

To assess the variability of the enrichment estimate, we performed bootstrap re-sampling of chromatin factor targets (Figure 3A and B, light curves). All datasets with the same factor are sampled together, leading to a conservative (high) estimated variability (Methods). GroupGM

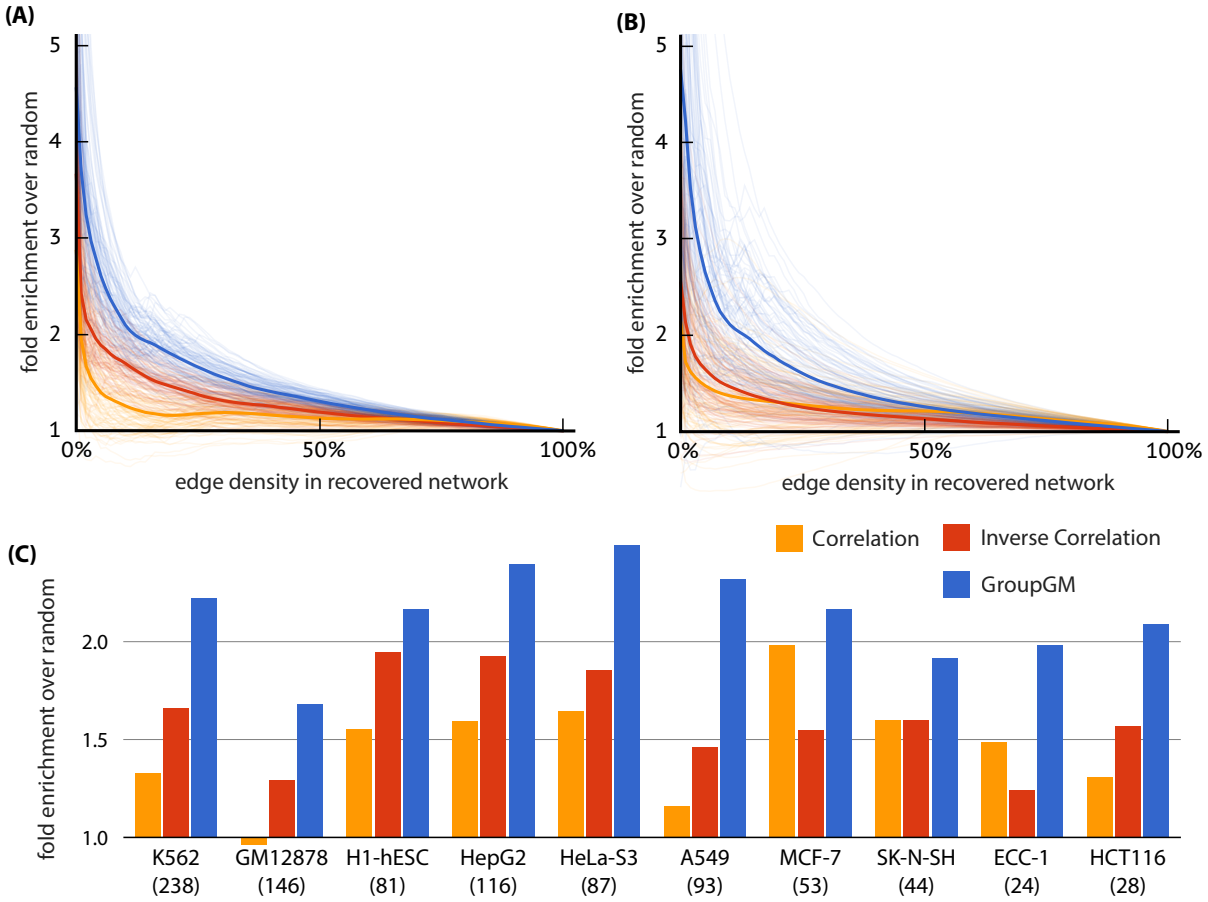

Figure 3: Enrichment of BioGRID-supported edges between transcription factors in networks estimated by correlation (yellow), inverse correlation (red), and GroupGM (blue). In the top plots, light lines represent bootstrap resampling variability, and dark lines represent average performance over all re-sampled networks. **(A)** Enrichment against evaluated proportion of network edges, excluding edges between different cell types. **(B)** Enrichment against evaluated proportion of network edges, only including edges between different cell types. **(C)** Enrichment within cell types that have 25 supported edges or more, where the network density was set to match number of BioGRID supported edges in each cell type. Beneath each cell type name is the number of datasets in that cell type.

showed a statistically significant improvement over both correlation ($P = 0.0055$) and inverse correlation ($P = 0.016$) for edges within cell types (Supplementary Figure 1).

To assess variability over cell types, we estimated enrichment separately for each cell type with 25 or more BioGRID-supported edges. In each cell type, we identified the number N of BioGRID-supported edges in that cell type. Then, we calculated the enrichment for BioGRID-supported edges among the top N edges in that cell type (Figure 3C). GroupGM performed consistently better than correlation or inverse correlation in individual cell types (Supplementary Figure 4).

Dependencies learned between cell types

We integrated data from all cell types into a single model for three reasons. First, a dataset for a chromatin factor in one cell type can serve as a proxy for a missing dataset for that factor in another cell type. This expands the chromatin network to include the union of chromatin factors measured in any cell type. This proves useful in analyzing data from ENCODE, which only measured a few chromatin factors in some cell types. Second, learning a single network enables comparisons of connections within different cell types, because the network in each cell type is conditioned on a common set of other datasets. Third, we can see high-level patterns in the joint chromatin network that would not otherwise be visible.

To assess how ChromNet leverages a joint model across all cell types, we checked edges between different cell types for enrichment in known protein-protein interactions. Our GroupGM network was enriched for BioGRID-supported edges when compared to a background set of random edges (Figure 3B). Bootstrap resampling showed this enrichment was significant ($P = 0.0095$; Supplementary Figure 5; Methods).

An example of the importance of conditional dependence: SMC3 separates RAD21 and MXI1

A specific example illustrates how conditional dependence reveals experimentally-supported direct interactions better than pairwise correlation (Figure 4A). In the correlation network among RAD21, SMC3, and MXI1, the three factors were tightly connected with one another in HeLa-S3 cervical carcinoma cells. The conditional dependence network, however, separated RAD21 and MXI1. This separation arose from the ability of SMC3 to explain away the correlation between RAD21 and MXI1. The factor pairs left connected in the conditional dependence network, RAD21–SMC3 and SMC3–MXI1, have physical interactions described in BioGRID [31, 16]. BioGRID lacks any direct connection between RAD21 and MXI1. Panigrahi et al. discovered more than 200 RAD21 interactors using yeast two-hybrid screening, immunoprecipitation–coupled mass spectrometry, and affinity pull-down assays [38]. They did not identify a RAD21–MXI1 interaction, which implies that RAD21 may not directly interact with MXI1.

An example of the importance of group dependency: recovering a connection between H3K27me3 and H3K4me3

Another specific example shows how GroupGM mitigates the effect of redundancy on conventional conditional dependence models. We examined edges between multiple H3K27me3 and H3K4me3 datasets from H7-hESC embryonic stem cells, collected at different time points in differentiation [37]. Since the datasets observe discrete portions of the differentiation process, one should not average them or pick a reference dataset arbitrarily. But both H3K27me3 datasets are very similar to one another, and so are the four H3K4me3 datasets. This means conventional conditional dependence methods identify edges between the two histone marks incorrectly. Edges estimated by

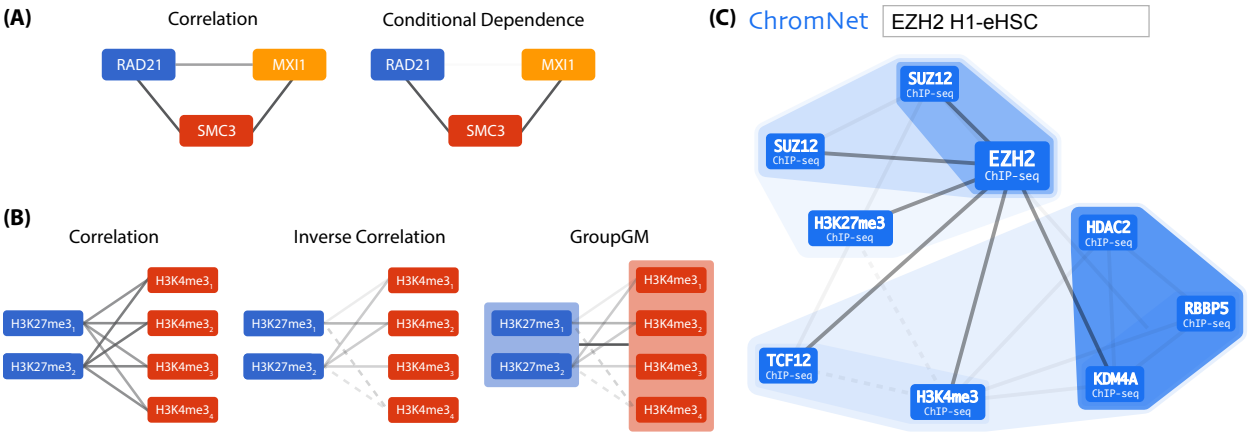

Figure 4: **(A)** *Left:* RAD21, MXI1, and SMC3 all co-localize with one another, suggesting they may all interact. *Right:* ChromNet reveals that the co-localization of RAD21 and MXI1 was largely mediated by the presence of SMC3. **(B)** GroupGM overcomes edge instability between tight clusters of H3K27me3 (blue) and H3K4me3 (red) modification datasets in H7-hESC at different differentiation time points. Edge darkness indicates connection strength. Dashed lines indicate negative interactions. We have removed within-group edges for clarity. *Left*: Correlation. *Middle*: Inverse correlation. *Right*: GroupGM. **(C)** The part of the ChromNet network that interacts with EZH2 in H1-hESC embryonic stem cells. This is a screen capture from the web interface with a search for "EZH2 H1-hESC". Shaded regions represent GroupGM hierarchical clustering. Darker regions represent tighter clusters. Dashed lines represent negative associations. We set the edge threshold to capture the six strongest edges connected to EZH2.

inverse correlation become weak and unstable, some even showing negative correlation (Figure 4B, middle). By allowing group edges, GroupGM has power to recover the positive correlation between H3K27me3 and H3K4me3 in developing embryonic stem cells (Figure 4B, right). This reflects the well-known bivalent co-occurrence of the repressive mark H3K27me3 and the activating mark H3K4me3 in embryonic stem cells [4].

An example of network accuracy: recovered interactions with EZH2 in H1-hESC recapitulate known functions

As an example illustrating the utility of ChromNet in revealing the potential interactors of a specific chromatin factor we examined a small portion of the network associated with the well-characterized protein EZH2 (Figure 4C). We focused on the H1-hESC cell type because it had the highest number of strong EZH2 connections in ChromNet.

Examining connections to EZH2 in H1-hESC highlighted many known interactions, which we discuss in decreasing order of edge strength. The strongest connection was from H3K27me3, a methyltransferase involved in H3K27me3 maintenance [1]. The next strongest connection was from SUZ12, an essential part of the Polycomb repressive complex required for EZH2's methyltransferase activity [6]. The connection from H3K4me3 is specific to H1-hESC (a negative association is observed in other non-embryonic cell types) and fits a model of "bivalent" chromatin marking of developmental genes in embryonic stem cells [4]. A connection from the histone demethylase KDM4A [15], may reflect a coordinated replacement of constitutive heterochromatin (H3K9me3) and active transcription (H3K36me3) states with the facultative heterochromatin (H3K27me3) state applied by EZH2. The final linked factor, TCF12, co-immunoprecipitates with EZH2 [28]. In

summary, most of the strongest interactions with EZH2 have support in the literature. We found this mixture of positive controls and potential novel connections in many parts of the network.

An example of a novel protein interaction: experimental validation of an interaction between MYC and HCFC1

The c-MYC (MYC) transcription factor is frequently deregulated in a large number and wide variety of cancers [33, 41]. It heterodimerizes with its partner protein MAX to bind an estimated 10-15% of the genome to regulate the gene expression programs of many biological processes, including cell growth, cell cycle progression, and oncogenesis [33, 41, 5]. The mechanisms by which MYC regulates these specific biological and oncogenic outcomes are not well understood. Interactions with additional co-regulators are thought to modulate MYC's binding specificity and transcriptional activity [18, 51]; however, only a few MYC interactors have been evaluated on a genome-wide level. Analysis of the large number of ENCODE ChIP-seq datasets can therefore further elucidate MYC interactions at the chromatin level.

ChromNet showed that MAX is the strongest interactor of MYC in multiple cell types (Supplementary Table 2), highlighting the ubiquitous nature of this interaction. Top-scoring ChromNet connections also included other known MYC interactors, like components of the RNA polymerase II complex such as POLR2A, TBP, and GTF2F1, and chromatin-modifying proteins such as EP300 (Supplementary Table 2). This shows how ChromNet analysis of transcription factors and co-regulators can help better understand transcription factor complexes.

In addition to the known interactors above, ChromNet also revealed previously uncharacterized, high-scoring interactions, including the transcriptional regulator Host Cell Factor C1 (HCFC1; Supplementary Table 2). HCFC1 binds largely to active promoters [34] and is involved in biological processes, such as cell cycle progression [40, 44] and oncogenesis [39, 9, 42]. This further supports its possible role as an interactor of MYC in regulating these activities. To validate the novel MYC-HCFC1 interaction, we performed a proximity ligation assay (PLA) in MCF10A mammary epithelial cells. This technique detects protein-protein interactions in intact cells [48]. When two proteins that are probed with specific antibodies are within close proximity of each other, fluorescence signals are produced and measured using fluorescence microscopy. We saw only background fluorescence when incubating with antibody against MYC (Figure 5A, top) or HCFC1 (Figure 5A, middle) alone. Incubation with both MYC and HCFC1 antibodies yielded a significant increase in fluorescence signal in the nuclear compartment (Figure 5A, bottom; Figure 5B). This suggests that MYC and HCFC1 interact in the nucleus.

We have shown that HCFC1 may be a novel co-regulator of MYC. Future investigation will reveal the importance of HCFC1 in regulating the biological functions of MYC, such as cell cycle progression and oncogenesis. This discovery demonstrates how ChromNet can suggest novel protein-protein interactions within chromatin complexes.

Spatial embedding reveals global patterns in the human chromatin network

By integrating all ENCODE datasets from many cell types into a single network, ChromNet enables extraction of global patterns in the relationships among chromatin factors. Using Cytoscape's additive force model [47], we visualized the entire network (Figure 6; Methods). In this embedding, the spatial proximity of two nodes correlates with the edge weight between those nodes. Nodes for the same chromatin factor in different cell types form a tight cluster when that factor's genomic position is conserved across cell types. For example, CTCF forms a tight cluster in this manner (Figure 6A). Relationships between chromatin factors are represented by their proximity in the

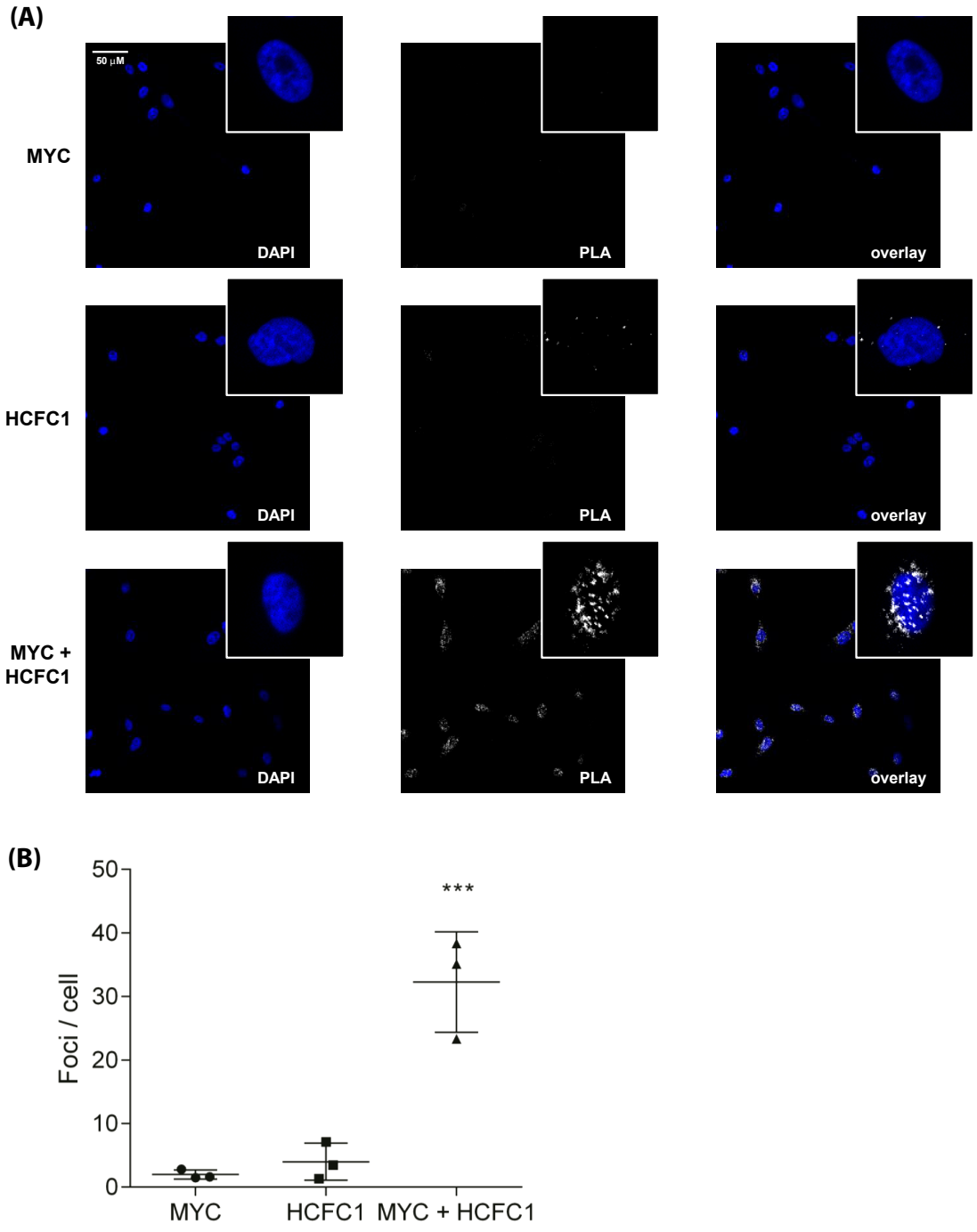

Figure 5: **(A)** Proximity ligation assay showing MYC and HCFC1 interaction in the nucleus. Representative micrographs show DAPI nuclear staining (left), proximity ligation signal (middle), and overlay (right) at 20× magnification, with insets at 100× magnification. *Top:* Cells probed with MYC antibody alone. *Middle:* Cells probed with HCFC1 antibody alone. *Bottom:* Cells probed with both antibodies. **(B)** Proximity ligation assay signal quantified as number of foci per cell. Mean ± standard deviation from three independent experiments is shown; *** $p < 0.001$, one-way analysis of variance with Tukey post test.

embedding. For example, MYC and MAX nodes are closely located. So are CTCF and RAD21. In contrast to the GroupGM network, proximity between related factors in a correlation network embedding is much less distinct (Supplementary Figure 8)

Relative positions of chromatin factors in the embedded graph highlight important aspects of biology. This is especially apparent among histone marks, where a clear pattern of two repressive marks are separated by a continuum of different activating marks. H3K9me3 and H3K27me3 are both repressive marks, but form very distinct clusters because they target distinct regions of the genome. H3K27me3 marks facultative heterochromatin, thought to regulate temporary repression of gene-rich regions [21]. H3K9me3 marks constitutive heterochromatin, and acts as a more permanent repressor [23]. Next to H3K9me3 is the H3K36me3 activating mark. H3K36me3 specifically marks zinc finger proteins when combined with H3K9me3 [17]. Next to H3K36me3 is H3K79me2, and both activating marks are found on gene bodies [8]. After H3K79me2 is a core set of activating marks centered around H3K4me3. The activating marks are separated from the H3K27me3 by a group of EZH2 nodes. This is consistent with prior knowledge that EZH2 deposits H3K27me3.

Positions of chromatin factor datasets reflect both their cell type identities and association with chromatin states. Highlighting the three Tier 1 ENCODE cell types shows a clustering of chromatin factor datasets by cell type (Figure 6B). K562 and GM12878 overlap spatially in the network and are both derived from blood cell lines, while H1-hESC (embryonic cells) is more distinctly separated. Coloring the network by correlation with chromatin state also reveals spatial patterns. We chose three (out of seven) Segway [20, 53] annotation labels that highlight distinct areas of the network (Figure 6C). Spatially embedding chromatin factor datasets using the ChromNet network simultaneously captures many important aspects of their function, such as chromatin state, cell lineage, and known factor-factor interactions.

Discussion

ChromNet enables understanding the chromatin network

Characterizing the chromatin network, the network of interactions among chromatin factors, is a key part of understanding gene regulation. ChromNet provides a new way to learn the chromatin network from ChIP-seq data. ChromNet addresses key problems encountered when learning protein-protein interactions from ChIP-seq correlations, such as the need to distinguish direct from indirect interactions and be robust to dataset redundancy. We demonstrated that ChromNet's GroupGM network infers known protein-protein interactions in the chromatin network more accurately than simpler methods. We also showed that ChromNet outperforms other approaches for several individual analyses.

Unlike most previous methods, ChromNet is also efficient enough to integrate thousands of genome-wide ChIP-seq datasets into a single joint network. To our knowledge, this study represents the first construction of an interaction network from all 1,415 ENCODE ChIP-seq datasets. ChromNet already scales to the number of datasets necessary to represent all 1,400–1,900 human transcription factors [52], once such data is available.

ChromNet provides a general computational framework to identify a joint dependence network from many ChIP-seq datasets. It can build a custom joint dependence network by incorporating user-provided ChIP-seq datasets or a combination of the ENCODE ChIP-seq datasets and user-provided datasets. To allow easier exploration of chromatin factor interactions and to facilitate generation of novel hypotheses, we have created a dynamic search and visualization web interface for both the ENCODE network and networks built from custom datasets (`http://chromnet.cs.`

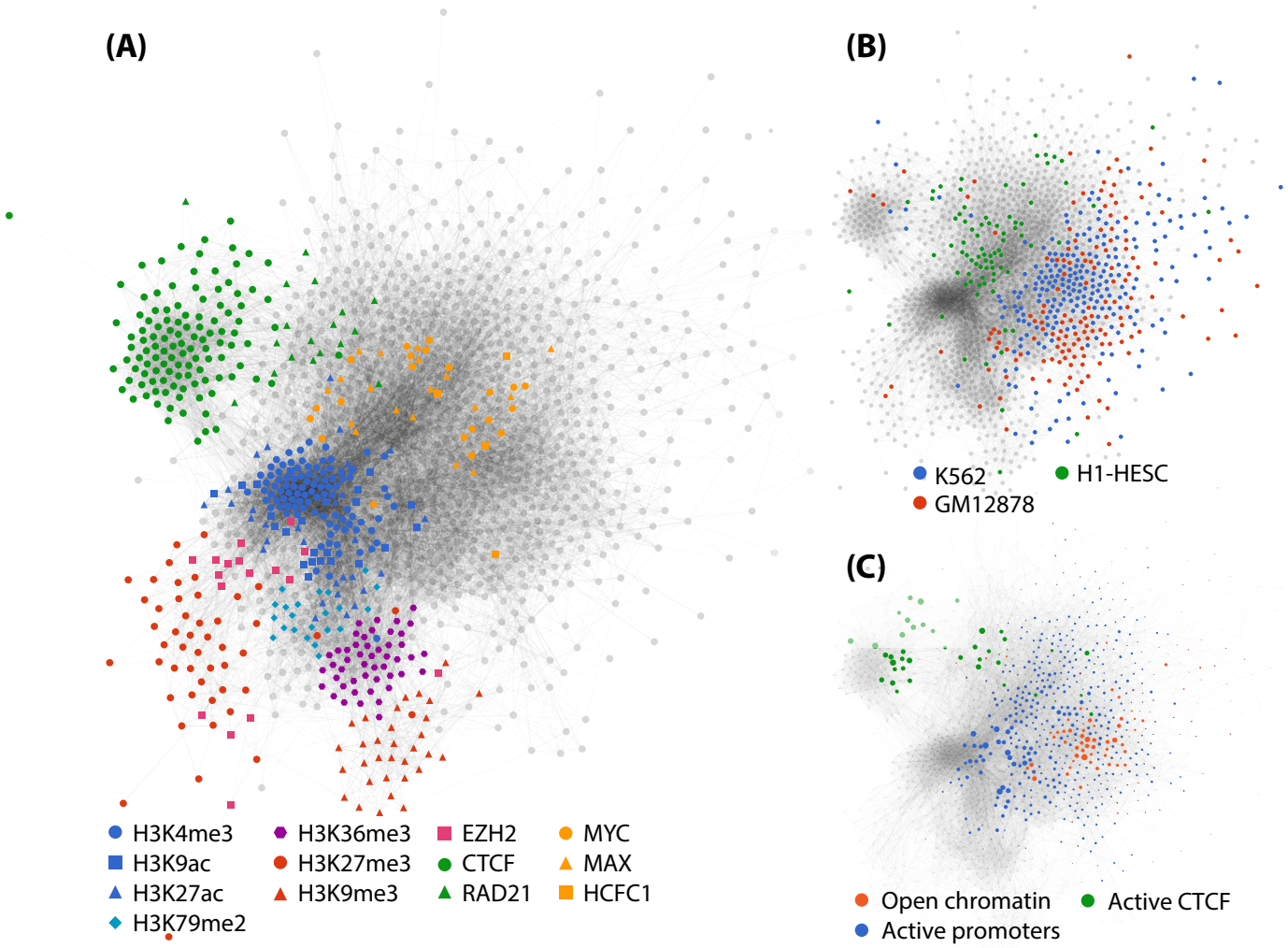

Figure 6: Force-directed 2D embedding of the entire human chromatin network, estimated by ChromNet. Spatial proximity of nodes and node groups reflects the strength of their inferred connection. In three views of this embedding, we have highlighted three different aspects: **(A)** Specific chromatin factors discussed in this article. **(B)** Datasets from the three ENCODE Tier 1 cell types, showing a separation of chromatin factors by cell type. **(C)** Correlation with three Segway genome annotation labels. We only colored the datasets from cell types where the Ensembl Regulatory Build had a corresponding Segway annotation. Node size represents correlation to a label, in comparison to all other nodes assigned to that label.

washington.edu). By building a large model and allowing easy inspection of small sub-networks, ChromNet combines a large-scale conditional dependence model with practical accessibility.

To demonstrate ChromNet's ability to reveal novel chromatin factor interactions, we experimentally validated the interaction between the MYC and HCFC1 proteins. The biological functions of the MYC oncoprotein are complex and dependent on its protein-protein interactions. Uncovering these interactions will provide insights into MYC transcriptional complexes involved in the oncogenic process and may also reveal potential targets for anti-cancer therapies. Through ChromNet, we identified HCFC1 as a novel interactor of MYC that may be involved in regulating biological and oncogenic functions of MYC.

Future directions

We envision several future extensions to the approach described in this article. First, while we have demonstrated the utility of applying ChromNet to ChIP-seq data alone, we plan to incorporate other data types into the network. RNA-seq expression datasets could resolve chromatin factor relationships that occur as a consequence of mutual involvement in gene expression. Incorporating feature annotations such as gene models could highlight direct interactions between factors and genomic regions of interest. The human genome's billions of base pairs provide a large sample size that allows joint comparisons of many genome-wide signals in a single model. Robust conditional dependence networks provide a benefit that is likely not limited to ChIP-seq data.

Second, we plan to consider relationships between chromatin factors at genomic position offsets. Here, we considered only co-occurrence relationships within the same 1,000 bp region. To model positional ordering constraints, we can also consider relationships between a factor in one region and another factor in an adjacent or nearby region. This would allow us to learn phenomena such as promoter-associated factors preceding gene-body–associated factors.

Third, just as the co-occurrence of different chromatin factors has been used to automatically annotate the genome, variations in the chromatin network at different positions may also prove useful to annotate functional genomic regions. This would also provide insight into the biological mechanisms behind specific chromatin factor interactions.

Methods

Data processing

ENCODE has the largest collection of high-quality ChIP-seq datasets [8], and continues depositing new datasets. ENCODE has processed many ChIP-seq datasets through a uniform pipeline. However, we reprocessed all the datasets from raw ChIP-seq reads (Figure 2) for two reasons. First, this allowed us to incorporate datasets not yet through ENCODE's uniform pipeline. Second, specifying our own pipeline makes it easier to process external users' data in an identical way. This facilitates adding non-ENCODE data to the ChromNet network.

We aligned reads from 3,543 FASTQ files to GRCh38/hg38 [13] using Bowtie2 [24]. To reduce noise before processing we filtered resulting Binary Alignment/Map (BAM) files to quality level 13 ($P \approx 0.05$) using SAMtools [29]. We grouped BAM files by dataset and matched them with controls using metadata from the ENCODE web site [11]. Then, we pooled and processed target and control BAM files using MACS2 [54]. We ran MACS2 without peak shift adjustments and with a P-value peak threshold of 0.05. This lenient threshold was chosen in an effort to capture both broad and narrow genomic binding signals across the genome. A more stringent threshold of 0.001 did not significantly change enrichment for known protein-protein interactions (Supplementary Figure 6).

We binned peak data from MACS2 into 1,000 bp windows by labeling a window 1 if any peak overlapped the window and 0 otherwise. Binning all datasets in this way yielded a $X \in \{0,1\}^{3{,}113{,}000 \times 1{,}415}$ binary *data matrix*. Each bin has a corresponding row in the matrix. We interpreted each of the 3,113,000 rows as a sample from a set $\mathcal{X} = \{X_1 \dots X_p\}$ of $p = 1415$ binary random variables representing presence or absence of each chromatin factor at a given position. Using this interpretation, we computed a sample correlation matrix $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ among the standardized variables in $\mathcal{X}$.

To create the *correlation network*, we set the weight of every edge between two datasets i and j equal to the corresponding entry $\hat{\Sigma}_{i,j}$ in the sample correlation matrix. This captures the pairwise linear dependence between two datasets (Figure 1A; bottom left).

Efficient estimation of conditional dependence from binary data

Given datasets drawn from a set $\mathcal{X}$ of binary random variables, we can represent a joint pairwise model of these datasets without loss of generality as a pairwise Markov random field:

$$P(x) = \frac{1}{Z} \exp\left(- \sum_{X_i \in \mathcal{X}, X_j \in \mathcal{X}} \Phi_{i,j} X_i X_j \right) \tag{1}$$

where Φ is a matrix of pairwise interaction terms and Z is a normalizing constant. Previous work on estimating a smaller subset of the chromatin network from binary data used Markov random fields and higher-order extensions [55, 35]. These works employ iterative or approximate methods, as exact inference on their models with many variables is computationally intractable. For certain graph classes, however, the sparsity structure of Φ and of the inverse correlation matrix Σ^{-1} are equivalent [30].

To justify using Σ^{-1} to approximate arbitrary networks, we compared with estimates of conditional dependence from a pairwise Markov random field of binary data [55] (Supplementary Note 2). The Markov random field implementation was based on unnormalized binary data, so for comparison we used the inverse covariance matrix which results from an unnormalized dataset. We used the original processed data kindly provided by the authors of [55] (J. Zhou, personal communication). These data are from 73 modENCODE ChIP-chip datasets on *Drosophila melanogaster* S2-DRSC cells. We calculated precision for the inverse covariance of binary data using the same bootstrap procedure as the authors, and compared against Markov random field precision numbers from their published precision-recall plot [55].

The ordering of coefficient magnitudes in $\hat{\Phi}$ and the inverse covariance matrix were similar across a wide range of Markov random fields (Supplementary Figure 3). This near-equivalence between the orderings means we can use an inverse covariance (or correlation) matrix to estimate edge strength in a pairwise Markov random field. This dramatically increases computational efficiency, but maintains results similar to a full binary pairwise Markov random field.

To create the *inverse correlation network* (Figure 3), we began by inverting $\hat{\Sigma}$ to get an inverse sample correlation matrix $\hat{\Sigma}^{-1}$ [32, 26]. We then set the weight of every edge between two datasets i and j equal to the corresponding entry $\{\hat{\Sigma}^{-1}\}_{i,j}$. This inverse correlation network captures the pairwise linear dependence between two datasets when conditioned on all other variables in the network.

Group graphical model

To create the *Group Graphical Model (GroupGM) network*, we began with the inverse correlation matrix created above. We extended the idea of pairwise relationships to groups of datasets by

considering a set $\mathcal{G}$ of q groups chosen by hierarchical clustering (see below). We let $\hat{G} \in \mathbb{R}^{q \times q}$ represent pairwise interaction strengths between all groups in the model. For any two groups i and j in the model their weight is given by the sum of entries between them in the inverse correlation matrix (Figure 1C).

$$\hat{G}_{i,j} = \sum_{k \in \mathcal{G}_i, l \in \mathcal{G}_j} \hat{\Sigma}^{-1}_{k,l} \tag{2}$$

We prove that Equation 2 correctly maintains the original edge magnitude in the case of redundancy (Supplementary Note 3).

To select the set $\mathcal{G}$ of groups, we used complete-linkage hierarchical agglomerative clustering of the correlation matrix [19]. This clustering methods starts by merging the two groups with the smallest maximum correlation distance between their datasets, then continues recursively until all groups have been merged. The use of hierarchical clustering eliminates the need to chose a fixed arbitrary number of clusters in advance. From the clustering results we chose all the leaf and internal nodes from the clustering algorithm as groups $\mathcal{G}$. Then, G became a $q \times q$ matrix filled according to Equation 2, where $q = 2p - 1$ (the total number of internal and leaf nodes). This method avoids comparing all possible subsets of datasets, which would make calculating G prohibitively expensive.

Since GroupGM uses the cluster assignments to mitigate strong redundancy, clustering accuracy is most important for tightly correlated datasets. When two datasets are highly correlated, it is important to group them together to mitigate the outcome of correlated datasets in network inference. When two datasets are only mildly correlated, the effects of their redundancy will also be mild, so it is not important to group them together. Hierarchical clustering is an attractive choice because it starts by creating groups among the most correlated datasets.

Visualization of the hierarchical chromatin network

To enable exploration of the chromatin network, we built an interactive visualization tool (`http://chromnet.cs.washington.edu`). This tool displays the nodes and edges of the chromatin network using a real time force model (Figure 4C). The tool's responsive interface lets users control which nodes and edges it displays. It immediately changes its display after a user types a search term to restrict displayed nodes. It also immediately changes its display when a user moves a slider that controls the minimum strength of a displayed edge. Our visualization tool facilitates exploring the chromatin network without excessive visual distraction.

The ChromNet visualization tool displays hierarchical groups from GroupGM by shading areas that enclose a group's members. It shades these areas with some amount of transparency. It displays the strongest groups with the highest opacity. The parents of two connected groups in the GroupGM hierarchy are themselves very likely connected. Therefore, for clarity we hide redundant parental edges.

To find a reasonable lower bound for the user-defined strength threshold, we examined the relationship between edge magnitude and known physical interactions. Within cell type edges from all cell types were sorted by magnitude and then binned into 200 groups. For each bin we computed the number of edges matching low throughput physical interactions in BioGRID and plotted how this varied over the bins. This enrichment curve suggested a lower bound of 0.03 to capture only edges enriched for known interactions (Supplementary Figure 7).

A conservative bootstrap estimate of protein-protein interaction enrichment variability

We estimated the variability of enrichment for known protein-protein interactions in the chromatin network (Figure 3) using bootstrap re-sampling over chromatin factors. We performed re-sampling over chromatin factors, and not over edges or individual datasets, because valid bootstrap re-sampling assumes independent and identically distributed samples. If we had re-sampled over the edges, we would have estimated a much smaller variability. This is because edges do not vary independently, and changes in a single dataset can affect all edges connected to that dataset. Variation specific to a single chromatin factor would affect all datasets measuring that factor. Those individual datasets, therefore, lack the independence assumed by the bootstrap sampling.

Under a chromatin factor bootstrap, we might sample a widely-measured chromatin factor a number of times. For example, ChromNet contains 150 CTCF datasets. Every time we sample CTCF, we add all 150 of these columns (where a column represents a variable in the data matrix X) to the bootstrap data matrix. Adding many datasets in unison greatly increases variability in the re-sampled data matrix. This yields conservative high variability estimates, ensuring that enrichment performance is not solely due to a few commonly measured factors. Using these bootstrap samples we compared the area under the enrichment rank curves (Figure 3A,B) between methods. The statistical significance of GroupGM's improvement was quantified as the fraction of bootstrap samples where GroupGM outperformed the other methods (Supplementary Figure 1; Supplementary Figure 5).

Proximity ligation assay

We seeded 2.5×10^4 MCF10A cells (a kind gift from S. Muthuswamy, Princess Margaret Cancer Centre) onto glass cover slips. After one day, we fixed cells in 2% paraformaldehyde, permeabilized the cells, and blocked them with bovine serum albumin. We then incubated the cells overnight with a mouse monoclonal antibody against MYC (1:25; C-33, Santa Cruz Biotechnology, Dallas, TX) and a rabbit polyclonal antibody against HCFC1 (1:50; A301-400, Bethyl Laboratories, Montgomery, TX) overnight. Then, we incubated cells with Duolink In Situ PLA anti-mouse MINUS and anti-rabbit PLUS probes (Sigma-Aldrich, St. Louis, MO). We processed cells using Duolink In Situ Detection Reagents Red following manufacturer's instructions (Sigma-Aldrich, St. Louis, MO). We imaged slides with a LSM700 confocal fluorescence microscope (Zeiss, Oberkochen, Germany). We quantified proximity ligation assay signal in at least 100 nuclei per condition using ImageJ [46].

Embedding the full chromatin network into a single plot

We embedded the full ChromNet network into a two-dimensional plane using force-directed layout (Figure 6). Because the GroupGM is inherently multi-scale, we sought a force model that accurately represented forces between individual nodes, and between all possible node groupings. In GroupGM, the edge weights between two groups are sums of the conditional dependence weights between all the individual datasets of those groups. This means we can correctly model the cumulative forces between groups using an additive force model with edge weights equal to the magnitude of the non-group edges between datasets. Cytoscape's standard spring layout [47], driven by additive spring coupling forces, implements this model. Using the same layout with weights derived from correlation instead of conditional dependence results in much tighter clustering and much weaker relationships between chromatin factors (Supplementary Figure 8).

We overlaid chromatin state annotation on the graph embedding by computing the correlation between each dataset and each Segway [20] region from the Ensembl Regulatory Build for

GRCh38/hg38 [53]. We drew a separate network labeling for each region by sizing each dataset node by its correlation with that Segway region. We normalized the size of the largest node in each network to a constant value and overlaid three of these network colorings (Figure 6C). For clarity, we showed only the largest node from each of the three networks for each dataset.

Acknowledgements

We would like to acknowledge William S. Noble, Zhiping Weng, R. David Hawkins, and Maxwell W. Libbrecht for their helpful feedback during the development of ChromNet.

This work was supported by a National Science Foundation (NSF) Graduate Research Fellowship (DGE-1256082) to S.M.L.; NSF (DBI-1355899) to S.I.L.; Natural Sciences and Engineering Council of Canada (RGPIN-2015-03948) to M.M.H.; Canada Research Chair in Molecular Oncology to L.Z.P., Canadian Institute for Health Research (MOP-275788) to L.Z.P. and B.R; and a Canadian Breast Cancer Foundation Ontario Region Doctoral Fellowship to W.B.T. Cloud computing resources for this research were generously provided by Google.

References

[1] Sandy Leung-Kuen Au et al. "EZH2-mediated H3K27me3 is involved in epigenetic repression of deleted in liver cancer 1 in human cancers". In: *PloS One* 8.6 (2013), e68226.

[2] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Vol. 571. John Wiley & Sons, 2005.

[3] Shelley L Berger. "The complex language of chromatin regulation during transcription". In: *Nature* 447.7143 (2007), pp. 407–412.

[4] Bradley E Bernstein et al. "A bivalent chromatin structure marks key developmental genes in embryonic stem cells". In: *Cell* 125.2 (2006), pp. 315–326.

[5] Elizabeth M Blackwood and Robert N Eisenman. "Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc". In: *Science* 251.4998 (1991), pp. 1211–1217.

[6] Ru Cao and YI Zhang. "SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex". In: *Molecular Cell* 15.1 (2004), pp. 57–67.

[7] Cedric R Clapier and Bradley R Cairns. "The biology of chromatin remodeling complexes". In: *Annual Review of Biochemistry* 78 (2009), pp. 273–304.

[8] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[9] Anwesha Dey et al. "Loss of the tumor suppressor BAP1 causes myeloid transformation". In: *Science* 337.6101 (2012), pp. 1541–1546.

[10] Marc I Diamond et al. "Transcription factor interactions: selectors of positive or negative regulation from a single DNA element". In: *Science* 249.4974 (1990), pp. 1266–1272.

[11] *Encode Project Data*. encodeproject.org.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[13] *Genome Reference Consortium*. genomereference.org.

[14] Mark B Gerstein et al. "Architecture of the human regulatory network derived from ENCODE data". In: *Nature* 489.7414 (2012), pp. 91–100.

[15] Lissania Guerra-Calderas et al. "The role of the histone demethylase KDM4A in cancer". In: *Cancer Genetics* 208.8 (2014), 215–224.

[16] Kalpana Gupta et al. "Mmip1: a novel leucine zipper protein that reverses the suppressive effects of Mad family members on c-Myc." In: *Oncogene* 16.9 (1998), pp. 1149–1159.

[17] Maria A Hahn et al. "Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks". In: *PLoS One* 6.4 (2011), e18844.

[18] Stephen R Hann. "MYC Cofactors: Molecular Switches Controlling Diverse Biological Outcomes". In: *Cold Spring Harbor Perspectives in Medicine* 17.4 (2014), p. 9.

[19] Trevor Hastie et al. *The Elements of Statistical Learning*. Vol. 2. 1. Springer, 2009.

[20] Michael M Hoffman et al. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation". In: *Nature Methods* 9.5 (2012), pp. 473–476.

[21] Joomyeong Kim and Hana Kim. "Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3". In: *ILAR Journal* 53.3-4 (2012), pp. 232–239.

[22] Clarence CY Kwan. "A Regression-Based Interpretation of the Inverse of the Sample Covariance Matrix". In: *Spreadsheets in Education (eJSiE)* 7.1 (2014), p. 3.

[23] Monika Lachner, Roderick J O'Sullivan, and Thomas Jenuwein. "An epigenetic road map for histone lysine methylation". In: *Journal of Cell Science* 116.11 (2003), pp. 2117–2124.

[24] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (2012), pp. 357–359.

[25] Julia Lasserre, Ho-Ryun Chung, and Martin Vingron. "Finding associations among histone modifications using sparse partial correlation networks". In: *PLoS Computational Biology* 9.9 (2013), e1003168.

[26] Steffen L Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[27] Ngoc T Le et al. "A nucleosomal approach to inferring causal relationships of histone modifications". In: *BMC Genomics* 15.Suppl 1 (2014), S7.

[28] Chun-Chung Lee et al. "TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer". In: *Journal of Biological Chemistry* 287.4 (2012), pp. 2798–2809.

[29] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[30] Po-Ling Loh, Martin J Wainwright, et al. "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses". In: *The Annals of Statistics* 41.6 (2013), pp. 3022–3049.

[31] Ana Losada, Tomoki Yokochi, and Tatsuya Hirano. "Functional contribution of Pds5 to cohesin-mediated cohesion in human cells and *Xenopus* egg extracts". In: *Journal of Cell Science* 118.10 (2005), pp. 2133–2141.

[32] Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate Analysis*. Academic Press, 1979.

[33] Natalie Meyer and Linda Z. Penn. "Reflecting on 25 years with MYC". In: *Nat Rev Cancer* 8.12 (2008), pp. 976–990.

[34] Joëlle Michaud et al. "HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy". In: *Genome Research* 23.6 (2013), pp. 907–916.

[35] Martin Renqiang Min et al. "Interpretable sparse high-order boltzmann machines". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. 2014, pp. 614–622.

[36] Wei Niu et al. "Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*". In: *Genome Research* 21.2 (2011), pp. 245–254.

[37] Sharon L. Paige et al. "A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development". In: *Cell* 151.1 (2012), pp. 221–232.

[38] Aswini K. Panigrahi et al. "A cohesin-RAD21 interactome". In: *Biochemical Journal* 442.3 (2012), pp. 661–670.

[39] Brandon J Parker et al. "A transcriptional regulatory role of the THAP11–HCF-1 complex in colon cancer cell function". In: *Molecular and Cellular Biology* 32.9 (2012), pp. 1654–1670.

[40] Brandon J Parker et al. "Host Cell Factor-1 recruitment to E2F-bound and cell-cycle-control genes is mediated by THAP11 and ZNF143". In: *Cell Reports* 9.3 (2014), pp. 967–982.

[41] Jagruti H. Patel et al. "Analysis of genomic targets reveals complex functions of MYC". In: *Nat Rev Cancer* 4.7 (2004), pp. 562–568.

[42] Samuel Peña-Llopis et al. "BAP1 loss defines a new class of renal cell carcinoma". In: *Nature Genetics* 44.7 (2012), pp. 751–759.

[43] Juliane Perner et al. "Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling". In: *Nucleic Acids Research* 42.22 (2014), pp. 13689–13695.

[44] David Piluso, Patricia Bilan, and John P Capone. "Host cell factor-1 interacts with and antagonizes transactivation by the cell cycle regulatory factor Miz-1". In: *Journal of Biological Chemistry* 277.48 (2002), pp. 46799–46808.

[45] Kate R Rosenbloom et al. "ENCODE data in the UCSC Genome Browser: year 5 update". In: *Nucleic Acids Research* 41.D1 (2013), pp. D56–D63.

[46] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9.7 (2012), pp. 671–675.

[47] Paul Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11 (2003), pp. 2498–2504.

[48] Ola Söderberg et al. "Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay". In: *Methods* 45.3 (2008), pp. 227–232.

[49] François Spitz and Eileen EM Furlong. "Transcription factors: from enhancer binding to developmental control". In: *Nature Reviews Genetics* 13.9 (2012), pp. 613–626.

[50] Chris Stark et al. "BioGRID: a general repository for interaction datasets". In: *Nucleic Acids Research* 34.suppl 1 (2006), pp. D535–D539.

[51] Lance R Thomas et al. "Interaction with WDR5 promotes target gene recognition and tumorigenesis by MYC". In: *Molecular Cell* 58.3 (2015), pp. 440–452.

[52] Juan M Vaquerizas et al. "A census of human transcription factors: function, expression and evolution". In: *Nature Reviews Genetics* 10.4 (2009), pp. 252–263.

[53] Daniel R Zerbino et al. "The Ensembl Regulatory Build". In: *Genome Biology* 16.1 (2015), p. 56.

[54] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137.

[55] Jian Zhou and Olga G Troyanskaya. "Global quantitative modeling of chromatin factor interactions". In: *PLoS Computational Biology* 10.3 (2014), e1003525.