Lundberg *et al. Genome Biology* (2016) 17:82
DOI 10.1186/s13059-016-0925-0

Genome Biology

METHOD Open Access

ChromNet: 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] and Su-In Lee[1,5*]

Abstract

A cell's epigenome arises from interactions among regulatory factors—transcription factors and histone modifications—co-localized at particular genomic regions. We developed a novel statistical method, ChromNet, to infer a network of these interactions, the *chromatin network*, by inferring conditional-dependence relationships among a large number of ChIP-seq data sets. We applied ChromNet to all available 1451 ChIP-seq data sets from the ENCODE Project, and showed that ChromNet revealed previously known physical interactions better than alternative approaches. We experimentally validated one of the previously unreported interactions, MYC–HCFC1. An interactive visualization tool is available at http://chromnet.cs.washington.edu.

Introduction

Regulatory factors—such as transcription factors, histone modifications, and other DNA-associated proteins—co-localize in the genome and interact with each other to regulate gene expression [14], the physical structure of the genome [10], cell differentiation [5], and other cellular processes. Identifying the genomic co-localization in this network among regulatory factors, which we termed the *chromatin network*, is important for understanding genome regulation and the function of each regulatory factor [4, 55]. To identify the chromatin network, we can use chromatin immunoprecipitation-sequencing (ChIP-seq) to measure the genome-wide localization of regulatory factors, and then compare ChIP-seq data sets to find regulatory factors that co-localize [11, 42]. Co-localization may indicate that two factors interact physically, by forming a complex, or functionally, by regulating similar DNA targets.

However, identifying pairwise co-localization alone fails to distinguish direct interactions from indirect interactions. A direct interaction represents physical contact or close functional coupling that requires spatial proximity. An indirect interaction is not from physical contact or direct functional coupling, but instead reflects the transitive effect of other direct interactions. Consider a simulated chromatin network among four factors, where factor C recruits A and B, and A in turn recruits D (Fig. 1a, top). Because all pairs of ChIP-seq data sets are correlated to each other (Fig. 1a, middle), a simple co-localization method would incorrectly infer interactions among all the factors (Fig. 1a, bottom left). In a *conditional-dependence network* (Fig. 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 data sets allows more indirect edges to be removed, resulting in a higher-quality network.

Here we present ChromNet, an approach that estimates the human chromatin network using a conditional-dependence network among regulatory factors from 1451 human ENCODE ChIP-seq data sets (Additional file 1: Table S1). Integrating all ENCODE data sets from many cell types into a single network provides several advantages. First, it enables the extraction of global patterns in the conditional-dependence relationships among regulatory factors in all cell types. Second, it provides a

*Correspondence: suinlee@uw.edu
[1] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[5] Department of Genome Sciences, University of Washington, Seattle, WA, USA
Full list of author information is available at the end of the article

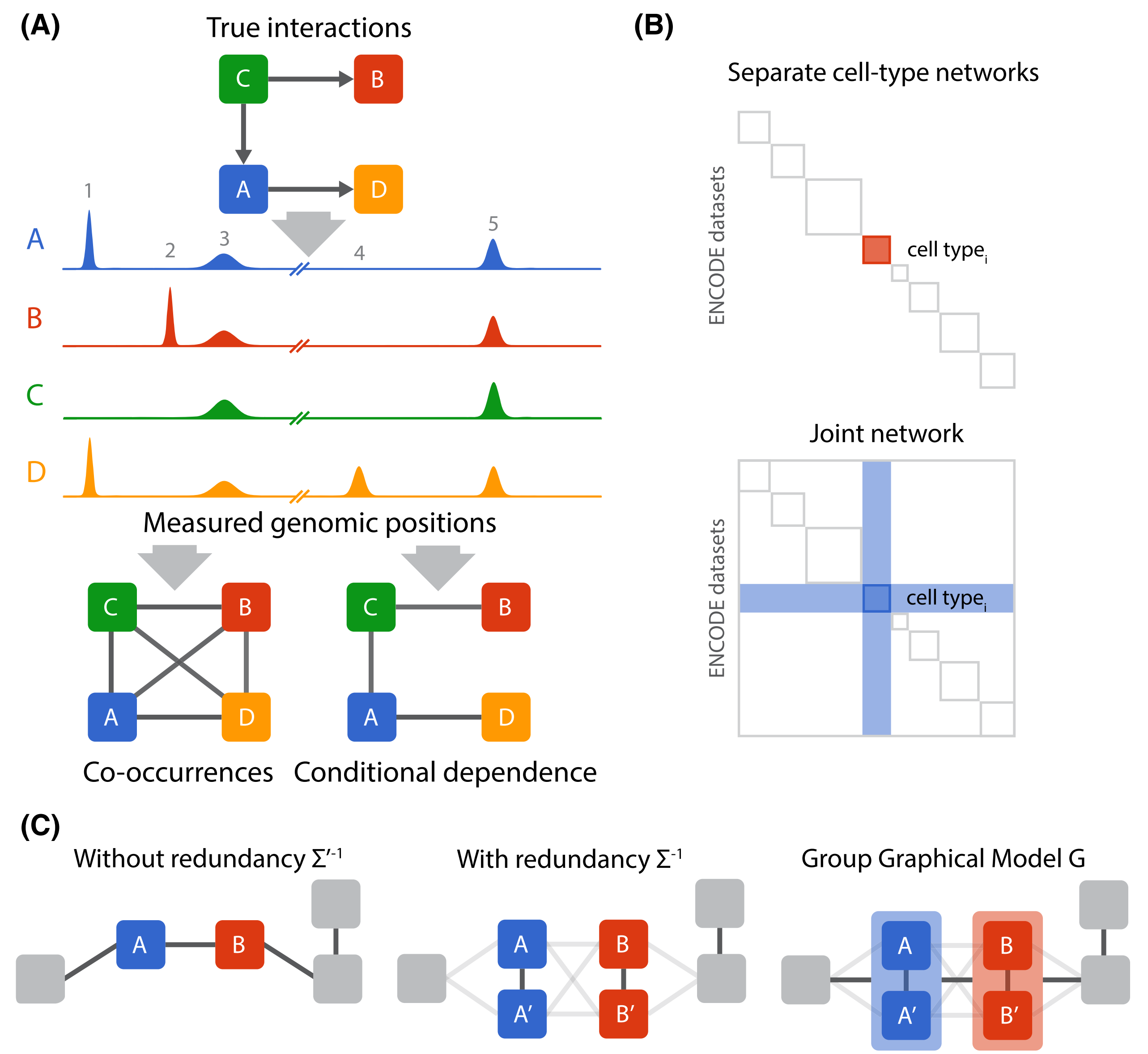

Fig. 1 **a** *Top:* Interaction network among four simulated regulatory factors. *Middle:* Binding activity from simulated ChIP-seq data sets, where each peak represents a putative binding position of a protein. *Bottom:* Networks inferred from ChIP-seq data sets based on co-occurrence (*left*) or conditional dependence (*right*). **b** Comparison of separate cell-type networks (*top*) with a single joint network (*bottom*). In a joint model, factors in each cell type have opportunities to be connected with new regulatory factors in other cell types, as highlighted by the *blue shaded region* (*bottom*). **c** Redundant information obscures conditional-dependence connections. *Left:* 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. *Right:* A group graphical model (GroupGM) represents the conditional dependence between groups of correlated variables, which restores the connection between *A* and *B*

flexible model that allows direct comparison of cell-type specific sub-networks because factors are conditioned on the same global set of ChIP-seq data sets across all cell types. Finally, it greatly increases the number of edges to consider by allowing edges connected to factors outside a single cell type (Fig. 1b). We show that this leads to a substantially increased fold enrichment for known protein interactions.

Learning a joint network among all available ENCODE ChIP-seq data sets involves three key challenges. First, learning a network among thousands of ChIP-seq data sets based on millions of genomic regions is highly computationally intensive. To solve this challenge, we utilized an efficient approach that involves the computation of an *inverse correlation matrix*, which does not require an expensive iterative learning procedure. This is in contrast to some other methods, such as Bayesian networks [3, 57] and Markov random fields [65], which face difficulties in scaling up, making it infeasible to run them on all 1451 ChIP-seq data sets (Additional file 1: Supplementary Note 1). Second, some regulatory factors are in the same complexes, and factors are often measured in different labs, conditions, or cell types, which creates significant correlations in the data. 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 (Fig. 1c, middle) [2]. Incorporating more ChIP-seq data sets exacerbates this problem. To solve this challenge, we present the *group graphical model* (GroupGM) representation of a conditional-dependence network that expresses conditional-dependence relationships among groups of regulatory factors as well as individual factors (Fig. 1c, right). We show that GroupGM improves the interpretation of a conditional-dependence network by allowing edges to connect groups of variables, which makes the edges robust against data redundancy. Third, network edges can be driven by interactions in specific genomic contexts. To help understand these contexts, we present an efficient method to estimate the impact of each genomic position on an inferred GroupGM edge.

Previous work on learning interactions among regulatory factors from ChIP-seq data used much smaller data collections. ENCODE identified conditional-dependence relationships among groups of up to approximately 100 data sets in specific genomic contexts [20]. Other authors used partial correlation on 21 data sets [32], Bayesian networks for 38 data sets [34], and partial correlation combined with penalized regression for 27 human data sets [49] and for 139 mouse embryonic stem cell data sets [25]. Still other authors used a Markov random field with 73 data sets in *D. melanogaster* [65], a Boltzmann machine with 116 human transcription factors [40], and bootstrapped Bayesian networks in 112 regulatory factors in *D. melanogaster* [3, 57]. Only other approaches also based on linear dependence models, such as the partial correlation used by Lasserre et al. [32], scale to all ENCODE data sets [Partial correlation and rank(Raw read pileup) in Additional file 1: Figure S1]. The ChromNet approach extends these methods in four distinct ways:

1. We show that linear dependence models can directly be applied to the genome-wide untransformed read count data (Additional file 1: Figure S1).
2. ChromNet addresses a fundamental challenge in network estimation when some of the variables are highly correlated with each other (collinearity) through a novel statistical method, the group graphical model.
3. ChromNet uses a novel method to identify genomic positions and genomic contexts that drive specific network edges.
4. Jointly modeling multiple cell types leads to a more informative network with a substantially higher enrichment for known protein interactions.

Network inference has also been applied to gene expression data, but the number of available samples in expression data is much lower than that in ChIP-seq data sets, which leads to different challenges.

ChromNet departs from previous approaches by enabling the inclusion of all 1451 ENCODE ChIP-seq data sets into a single joint conditional-dependence network. GroupGM and an efficient learning algorithm allow seamless integration of all data sets comprising 223 transcription factors and 14 histone marks from 105 cell types without requiring manual removal of potential redundancies (Additional file 1: Table S1). We show that this approach significantly increases the proportion of network relationships among ChIP-seq data sets supported by previously known protein–protein interactions compared to other scalable methods (see "Results"). We also demonstrate the potential of ChromNet to aid new discoveries by experimentally validating a novel interaction.

Results

Uniformly processed data reduces noise when learning conditional dependence

To ensure comparable signals across all ChIP-seq data sets, we reprocessed raw ENCODE sequence data with a uniform pipeline (Fig. 2a). We downloaded raw FASTQ files from the ENCODE Data Coordination Center [11, 15, 51] (Additional file 2) and mapped them using Bowtie2 [31] to the human genome reference assembly (build GRCh38/hg38) [19]. We binned mapped read start sites into 1000–base-pair (bp) bins across the entire genome, which results in a $3,209,287 \times 1451$ data matrix X where genomic positions are viewed as samples (Fig. 2a). We compared several different data preprocessing methods and chose binned read counts for three reasons:

1. They allow easy integration of external ChIP-seq experiments.
2. They do not require the determination of various cut-offs in a peak calling algorithm.
3. A network inferred from read counts performs well, revealing previously known protein–protein interactions (Additional file 1: Figure S1).

A conditional-dependence network can be efficiently learned from binned read count data

Learning a conditional-dependence network among thousands of ChIP-seq data sets each containing millions of samples (genomic positions) requires an efficient

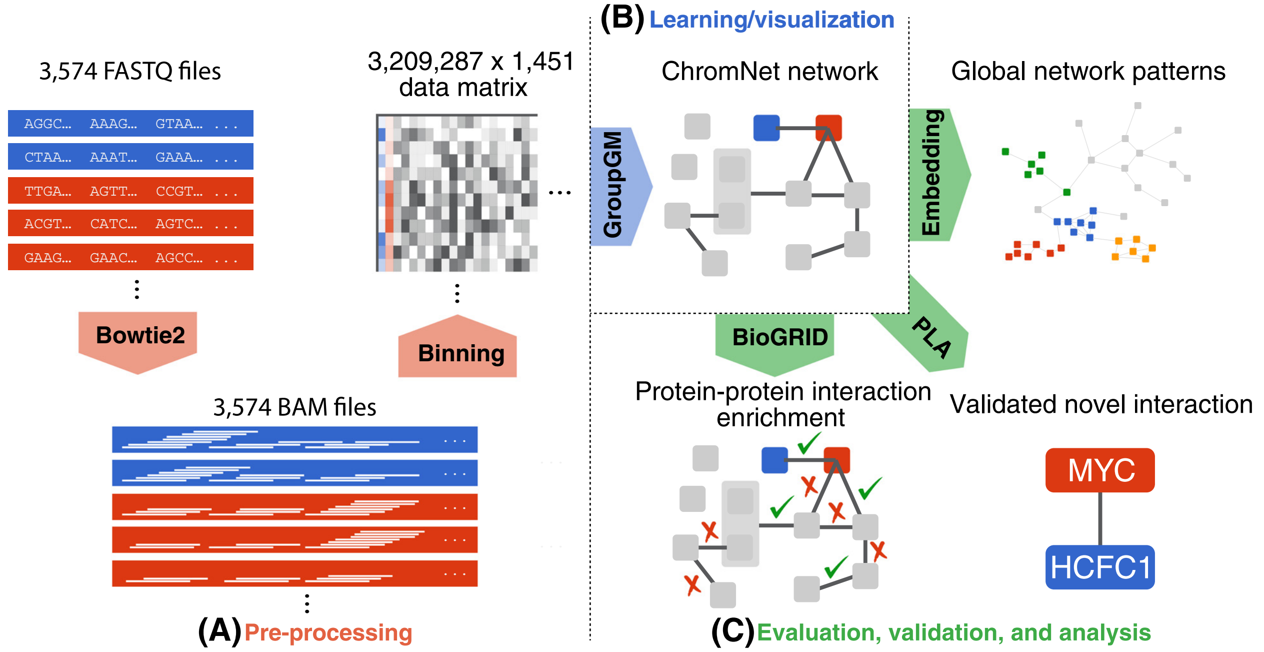
(B) Learning/visualization
3,574 FASTQ files
3,209,287 x 1,451 data matrix
ChromNet network
Global network patterns
AGGC... AAAG... GTAA... ...
CTAA... AAAT... GAAA... ...
TTGA... AGTT... CCGT... ...
ACGT... CATC... AGTC... ...
GAAG... GAAC... AGCC... ...
GroupGM
Embedding
Bowtie2
Binning
BioGRID
PLA
3,574 BAM files
Protein-protein interaction enrichment
Validated novel interaction
MYC
HCFC1
(A) Pre-processing
(C) Evaluation, validation, and analysis

Fig. 2 a The uniform processing pipeline includes aligning sequences using Bowtie2, and then binning them into 1000–base-pair regions. **b** We inferred the GroupGM from all 1451 ChIP-seq data sets and integrated the model learned into a web interface to facilitate broad use. **c** We evaluated the model learned against known physical protein interactions (BioGRID), mapped global patterns through network embedding, and validated a novel predicted MYC-HCFC1 interaction with the proximity ligation assay (PLA)

algorithm (Fig. 2b). It is well known that the nonzero pattern of the *inverse covariance matrix* of Gaussian random variables represents the conditional-dependence network [33, 37]. The inverse correlation matrix, Σ^{-1}, is 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} [23], count data requires more care. Distributions of counts in binned ChIP-seq reads are often clearly truncated at zero, and also increase in variance for high read counts. Multivariate distributions with count-valued marginal distributions are often very restrictive (for example only allowing positive correlations) or are infeasible to estimate for thousands of dimensions [62]. An often employed alternative is to use a multivariate Gaussian distribution after appropriately transforming the count data, such as with the sqrt or asinh function [9]. However, interestingly, our results show that applying a linear Gaussian model directly to the binned read counts of ENCODE ChIP-seq data better recovers known protein–protein interactions than when using standard normalizing data transforms ("Methods" and Additional file 1: Figures S1 and S2). This leads to an efficient and simple model formulation for ChromNet applied directly to the mapped read counts, which is relatively easier to obtain compared to other ChIP-seq data preprocessing methods and does not require any threshold.

ChromNet first computes the *inverse sample correlation matrix* $\hat{\Sigma}^{-1}$ from the data matrix X of 1451 variables and 3,209,287 samples, and then uses a GroupGM approach to interpret elements of $\hat{\Sigma}^{-1}$ as weights of network edges (Fig. 2b).

Group modeling mitigates the effects of redundancy

Many ENCODE ChIP-seq data sets contain redundant positional information. Conventional conditional-dependence methods have a key limitation in modeling redundant data. If data sets A and A' are highly correlated, a conventional method would connect A with A' but connect A to the rest of the network only weakly (Fig. 1c). Arbitrarily removing or merging redundant data sets can hide or eliminate important information in the data.

GroupGM overcomes challenges with redundant data in conditional-dependence models by allowing edges that connect groups of data sets (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 ("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 correctly reveals conditional dependencies in the presence of redundancy (Additional file 1: Supplementary Note 2).

A group is defined as a set of highly correlated variables whose individual conditional-dependence relationships with other variables are not likely to be captured, as illustrated in Fig. 1c. To obtain groups, we used complete linkage hierarchical clustering, and restricted groups to have a minimum pairwise correlation of ρ (= 0.8) within each group. The choice of complete-linkage clustering allows us to obtain groups where all the factors are highly correlated. Because the complete-linkage distance metric merges two clusters based on the minimum correlation between any two variables in the groups, we can stop merging when the minimum correlation becomes less than or equal to ρ before creating all $2p - 1$ groups, where $p = 1451$.

Each variable (a ChIP-seq data set) can be in multiple groups as long as it is highly correlated with at least one other data set. This multi-scale nature of groups is a unique feature of the group graphical model. It allows us to capture multiple ways each factor can be connected with other factors. Say that a data set for factor A forms a group with another data set for factor B. In the group graphical model, A can have connections specific to itself and connections shared with B, and their edge weight values would indicate which connections are statistically robust. This allows us to reveal multiple kinds of interactions A can have: specifically with itself and with A and B as a complex. The latter may not be captured by a conventional conditional-dependence network, such as inverse correlation or partial correlation, if A and B are highly correlated with one another.

The purpose of having a threshold for minimum pairwise correlation ρ is to identify sets of variables whose high within-group correlation is likely to prevent them from being connected to other variables in the network. The threshold used in this paper $\rho = 0.8$ captures 53 % of all the multi-factor groups formed by hierarchical clustering, and was chosen so as to include strong groups while still keeping the size of groups small enough to interpret (Additional file 1: Figure S3).

Conditional dependence and joint 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 three networks among ChIP-seq data sets using the following three methods, where each method produces a set of weighted edges:

1. Correlation: We learned a naive co-occurrence network, using a pairwise Pearson's correlation between all pairs of data sets.
2. Inverse correlation: We learned a conditional-dependence network, by computing the matrix inverse of the correlation matrix.
3. GroupGM: We learned a group conditional-dependence network, which addresses tight correlation among data sets by allowing edges between groups of variables.

Partial correlation is similar to inverse correlation and performs nearly as well (Additional file 1: Figure S4). We did not include other previously described methods because they do not scale to the large data collection we used (Additional file 1: Supplementary Note 1 and Figure S5).

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 [56]. For evaluation, we excluded edges connecting the same regulatory 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. We also excluded edges involving a histone mark because they do not exist in BioGRID. For these edges, we ran a separate evaluation using the HIstome database [26] and showed that the group graphical model shows higher enrichment than the alternative methods (Additional file 1: Figure S6). When we measured the conditional dependence between a pair of ChIP-seq data sets in GroupGM, to avoid the inclusion of many redundant edges, for each pair of data sets, we picked the maximum edge weight out of all network edges connecting groups, each of which contains one of the corresponding data sets. This way, we consider exactly the same number of data set pairs for evaluation across all three methods. We only scored edges from groups containing a single type of factor (about half of the groups; see Additional file 1: Figure S3), because if a group contains more than one factor, there is no clear way to characterize such an edge as true or false from BioGRID, or match it with an edge from competing methods for comparison.

Group modeling improves the recovery of interactions within and between cell types

We compared the performance of the three methods described above across a range of prediction thresholds. For each network, we varied the number of evaluated edges N 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. Since 8.4 % of data set pairs in the same cell type are supported by a BioGRID physical interaction, an enrichment fold of 1 corresponds to 8.4 % of recovered edges matching prior knowledge. Enrichment fold captures the effect of both type I and type II error rates (see "Methods").

We first measured performance within all cell types, excluding edges between data sets in different cell types (Fig. 3a top). Since the limited number of annotations in BioGRID imperfectly represents the human chromatin network, one cannot draw strong conclusions about absolute performance from this benchmark. The relative performance of the methods, however, is clear; inverse correlation performs better than correlation, and GroupGM outperforms inverse correlation. This supports the idea that better resolution of direct versus indirect interactions contributes to improved performance of inverse correlation over correlation, while greater robustness against redundancy likely contributes to improved performance of GroupGM over inverse correlation. The value of conditional dependence and group modeling is also further supported by specific examples in the network (Fig. 4; Additional file 1: Figures S7 and S8), and by the fact that GroupGM still outperforms inverse correlation even after attempting to remove the strongest redundancies by merging data sets from different labs targeting the same factor in the same cell type/condition (Additional file 1: Figure S9).

To assess the variability of the enrichment estimate, we performed bootstrap resampling of regulatory factor targets (Fig. 3a, b, light curves). All data sets with the same factor are sampled together, leading to a conservative (high) estimated variability ("Methods"). GroupGM showed a statistically significant improvement over both correlation ($P = 0.0004$) and inverse correlation ($P = 0.0036$) for edges within cell types (Additional file 1: Figure S10).

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 (Fig. 3c). GroupGM performed consistently better than correlation or inverse correlation in individual cell types (Additional file 1: Figure S11).

We also generated a simulated data set meant to mirror the characteristics of real ChIP-seq data sets (Additional file 1: Figure S12). Using this simulated data, we found a similar relative performance of various methods, with GroupGM recovering the most true network edges (Additional file 1: Figure S13 and "Methods").

To assess how well a joint model can recover relationships between factors measured in different cell types, we checked edges between different cell types for enrichment in known protein–protein interactions (Fig. 3a bottom). The GroupGM network showed a clear enrichment for known interactions above random ($P = 0.0095$), and also outperformed inverse correlation ($P = 0.0174$) and correlation ($P = 0.0282$) (Additional file 1: Figure S14 and "Methods"). This implies that information about many physical protein interactions can be recovered even from data sets in different cell types.

Comparison between a joint model of all cell types and cell-type specific models

Integrating ChIP-seq data sets from multiple cell types into a single network model provides the following three advantages. First, we can capture high-level patterns in the joint chromatin network that would not otherwise be visible. Second, a joint model allows the direct comparison of cell-type specific sub-networks because factors are conditioned on the same global set of ChIP-seq data sets across all cell types. Finally, a data set for a regulatory factor in one cell type can serve as a proxy for a missing data set for that factor in another cell type, if the factor's localization in the genome is conserved between the cell types (Additional file 1: Figure S15). This greatly expands potential chromatin network edges to include the union of regulatory factors measured in any cell type. This global network contains both conserved and cell-type specific sub-networks, and proves useful in analyzing data from ENCODE, which only measures a few factors in some cell types.

To compare directly a joint model across all cell types with cell-type specific models for each cell type separately, we focused on the four best characterized ENCODE cell types and compared enrichment of BioGRID-supported edges (Fig. 3b). By varying the number of edges in the networks, we find that the joint model consistently identifies interactions with higher fold enrichment for known interactions. In addition, a joint model also identifies more unique BioGRID supported protein–protein interactions than cell-type specific models (Additional file 1: Figure S16).

We show as well that the large increase in potential edges from a joint model does not introduce spurious associations among edges within a cell type. When we excluded all cross-cell-type edges from the joint model, the joint model still marginally outperforms cell-type specific models ($P = 0.0672$; Additional file 1: Figure S17).

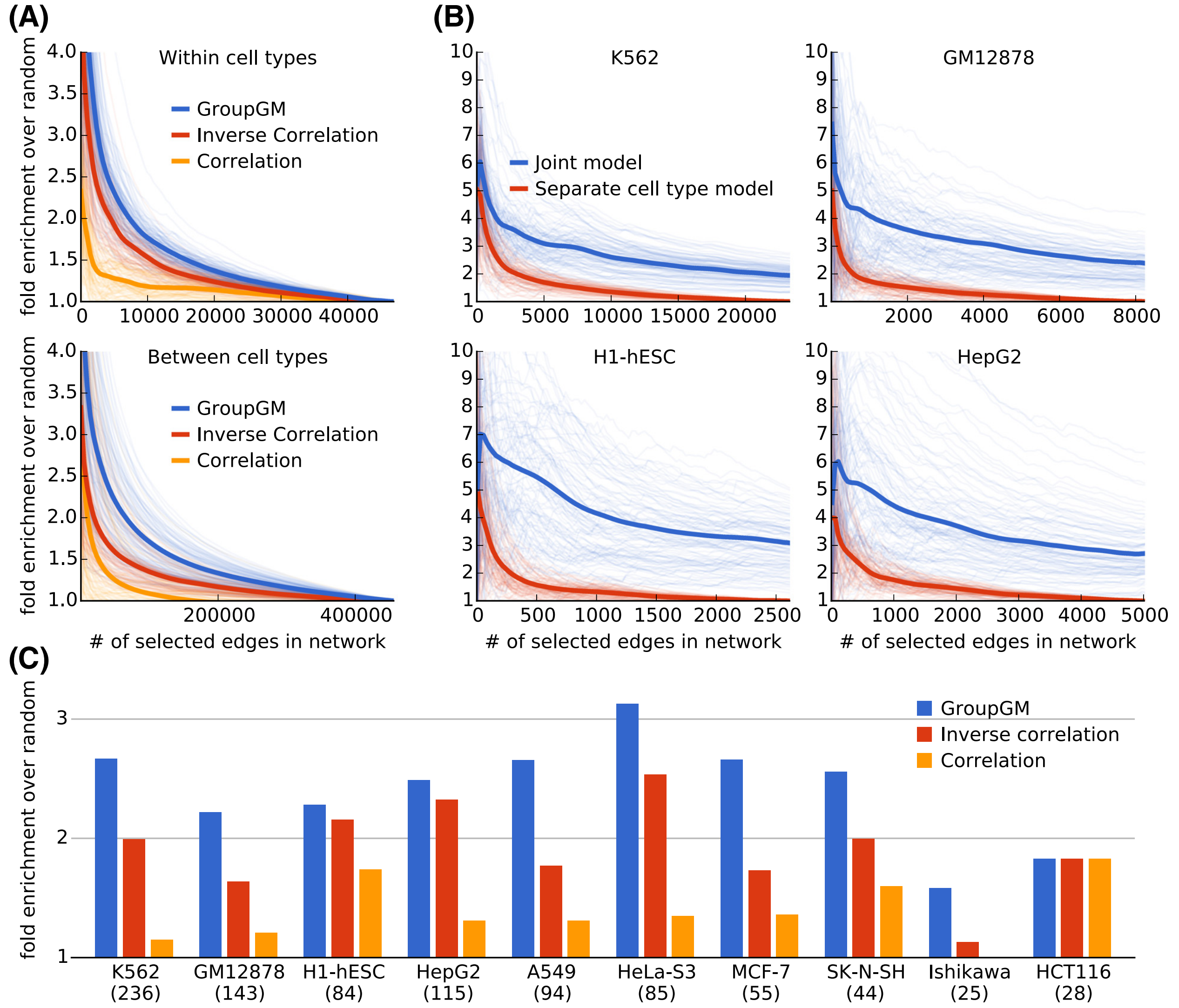

Fig. 3 Enrichment of BioGRID-supported edges between transcription factors in networks estimated by correlation (*yellow*), inverse correlation (*red*), and GroupGM (*blue*). In (**a**) and (**b**), *light lines* represent bootstrap resampling variability, and *dark lines* represent average performance over all resampled networks. **a** Fold enrichment for BioGRID-supported edges against a varying number of evaluated network edges. *Top:* Excluding edges between different cell types. *Bottom:* Only including edges between different cell types. **b** Enrichment against a varying number of evaluated network edges. Here, we compared between a joint model and a cell-type specific model in four different cell types. **c** Enrichment within cell types that have 25 supported edges or more, where the network density was set to match the number of BioGRID-supported edges in each cell type. Beneath each cell-type name is the number of data sets in that cell type

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 (Fig. 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 [21, 36]. 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 [44]. They did not identify a RAD21–MXI1 interaction, which implies that RAD21 may not directly interact with MXI1.

To focus on the comparison between conditional dependence and correlation, we have not displayed the group that contains SMC3 and RAD21. This grouping reflects

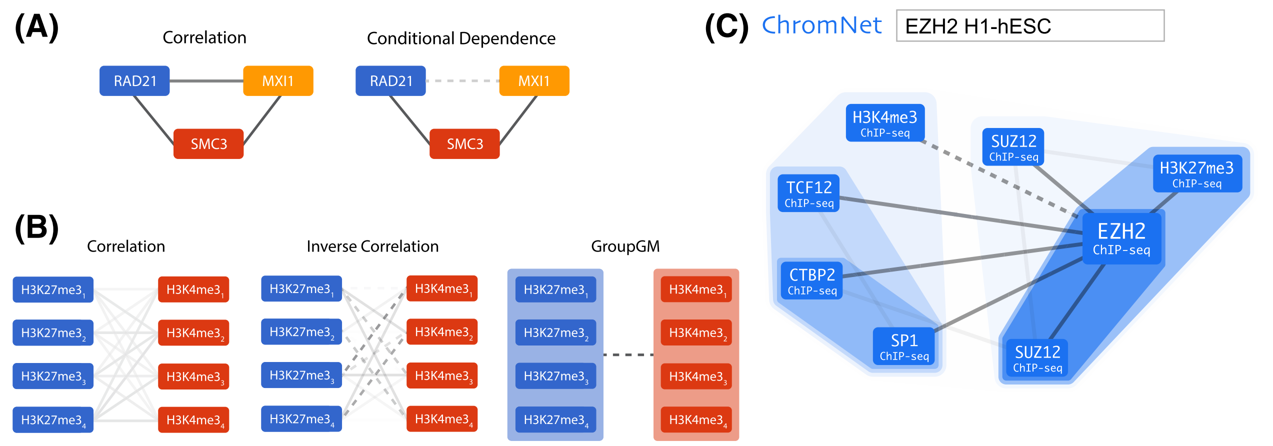

Fig. 4 a *Left:* RAD21, MXI1, and SMC3 co-localize with one another, suggesting that they may all interact with each other. *Right:* ChromNet reveals that the co-localization of RAD21 and MXI1 was largely mediated by the presence of SMC3. **b** GroupGM overcomes edge instability between tight clusters of H3K27me3 (*blue*) and H3K4me3 (*red*) data sets in H7-hESC measured at different differentiation time points. Edge darkness indicates the strength of the connection; *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 our web interface with a search for "EZH2 H1-hESC." Shaded regions represent the GroupGM groups learned by hierarchical clustering; darker regions represent tighter clusters. We set the edge threshold to capture the six strongest edges connected to EZH2

their common role in the cohesion complex and is present in many cell types. We also note that Fig. 4a is only a small part of the full ChromNet network and considering more factors reveals additional relationships that involve CTCF and ZNF143, which is consistent with prior knowledge [67] (Additional file 1: Figure S18).

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 data sets from H7-hESC embryonic stem cells, collected at different time points in differentiation [43]. H3K27me3 is a repressive mark and H3K4me3 is an activating mark. Since the data sets represent different portions of the differentiation process, one should not average them or pick a reference data set arbitrarily. However, the H3K27me3 data sets are correlated highly enough with one another to form a group, and so are the four H3K4me3 data sets. This implies that conventional conditional-dependence methods would identify edges between the two histone marks incorrectly.

Edges estimated using correlation indicate that the ChIP-seq data sets targeting H3K27me3 and those targeting H3K4me3 are positively correlated. However, H3K27me3 is associated with repressed genomic regions while H3K4me3 is associated with actively transcribed regions [66]. Since a minority of promoters in embryonic stem cells are bivalently marked, these two marks should not have an overall positive association [5, 66]. In fact, most ChIP-seq data sets are positively correlated with each other (Additional file 1: Figure S15), which is induced by mappability and many regions that are transcriptionally silent or active. Resolving this problem by removing some of these regions is unlikely to be successful, because it is not clear what criteria we need to exclude regions. In conditional-dependence models, such as inverse correlation and the group graphical model, by conditioning on many other variables, these global confounding effects are naturally removed. Edges estimated by inverse correlation account for these confounders but become weak and unstable showing a mixture of positive and negative associations (Fig. 4b, middle). By allowing group edges, GroupGM has power to recover the negative association between H3K27me3 and H3K4me3 (Fig. 4b, right), which is consistent with prior knowledge [5, 66].

An example of learning genomic context: ZNF143 mediates the conditional-dependence relationship between CTCF and SIX5

Many relationships between regulatory factors only occur in a particular genomic context. This raises the question of how, or whether, this context specificity is encoded in ChromNet. We can gain insight into this by considering what it means for one factor to mediate the relationship between two other factors, such as *A* mediating the relationship between *C* and *D* in Fig. 1a. When this occurs, it means that the connection between *C* and *D* can be explained by their co-occurrence with *A*. In other words, *A* is the context in which the relationship between *C* and *D* occurs.

A practical example of this is found in the relationships between SIX5, ZNF143, and CTCF in the K562 cell type. Simple correlation connects all three factors together with positive edges, but GroupGM shows that ZNF143 actually mediates the relationship between SIX5 and CTCF (Additional file 1: Figures S7 and S8). This means that the association of SIX5 with CTCF primarily occurs in the presence of ZNF143, the CTCF–SIX5 relationship is context-specific, and ZNF143 is the context. More generally, when an association between two factors, *C* and *D*, is specific to a certain genomic context and that context is well represented by a third factor, *A*, then *A* would mediate *C* and *D*. This gives the connections *C*–*A*–*D* in the conditional-dependence network; thus context-specific relationships, such as the relationship between CTCF and SIX5 in the presence of ZNF143, are captured in a GroupGM network, if all three factors are present.

It is important to understand the genomic context in which any given edge occurs regardless of whether that context is well represented by another factor in the network. Even if *A* is not observed, we want to be able to infer the genomic context of the interaction between *C* and *D*. To address this need, we designed an efficient method to label every genomic position with its influence on a group network edge ("Methods"). Using CTCF–ZNF143–SIX5 as an example, we removed all ZNF143 experiments from ChromNet and then computed the genomic context of the edge between CTCF and SIX5. To validate this genomic context, we took the top 1000 bins (1, 000, 000 bp) and intersected them with the top 1000 bins from all other experiments in K562, including ZNF143. Even though ZNF143 was not present in the model and ZNF143 data sets were not used when inferring the genomic context, it had the highest overlap of any experiment with the context driving the CTCF–SIX5 edge, even higher than the CTCF and SIX5 experiments themselves (Additional file 1: Figure S19).

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 regulatory factor, we examined a small portion of the network associated with the well-characterized protein EZH2 (Fig. 4c). We focused on the H1-hESC cell type because it had many strong EZH2 connections in ChromNet. Examining connections to EZH2 in H1-hESC highlighted several known interactions, which we discuss in decreasing order of edge strength. The strongest connection is from H3K27me3, and EZH2 is a methyltransferase involved in H3K27me3 maintenance [1]. The next strongest connections are with SUZ12, which is an essential part of the Polycomb repressive complex 2 (PRC2), and is required for EZH2's methyltransferase activity [8, 13]. The next connection to CTBP2 is supported by this co-repressor's possible role in deacetylation of H3K27 in preparation for PRC2-mediated methylation [28]. H3K4me3 is well known to be present in active regions of the genome, so a negative relationship with EZH2 (represented by a dashed line) that deposits the repressive H3K27me3 mark is expected. SP1 is a potentially novel interactor of EZH2, while TCF12 is known to co-immunoprecipitate with EZH2, which suggests that TCF12 interacts with PRC2 [35]. In summary, most of the strongest interactions with EZH2 have support in the literature. We found this mixture of interactions supported by the literature and potential novel connections in many parts of the network.

An example of cross-cell-type comparison: enhancer-associated regulatory factors

Learning a conditional-dependence network for all ENCODE cell types allows the comparison of within cell type connections across different cell types. Active enhancers are known to be flanked by a combination of the histone marks, H3K27ac and H3K4me1 [53]. To quantify how strongly different transcription factors associate with active enhancers in different cell types, we calculated the sum of the group edges between each regulatory factor (except histone marks) and H3K27ac and H3K4me1 measured in that cell type. This provides a score for each factor in each cell type. Seven ENCODE cell types with 20 or more data sets contain both H3K27ac and H3K4me1, while also containing EP300, which is known to bind active enhancers [53]. We focused on these seven cell types and ranked the factors in each cell type by their association with H3K27ac and H3K4me1. Additional file 1: Table S2 lists the top ten factors in each cell type most associated with active enhancers. EP300 can be considered a validation for the list and is highly ranked in all seven cell types ($P < 10^{-5}$). Interestingly, even more highly ranked than EP300 is POLR2A. This association is likely because active enhancers are in close proximity to active transcription start sites in promoters in 3D space, due to the looping mechanisms for enhancer–promoter communication. The influence that 3D conformation can have on measures of co-localization in the genome is important to bear in mind when analyzing the edges in ChromNet. Other factors that are consistently associated with enhancers across cell types are shown in red, while cell-type specific associations are in black (Additional file 1: Table S2).

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

[38, 47]. 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 [6, 38, 47]. 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 [22, 58]; however, only a few MYC interactors have been evaluated on a genome-wide level. Analysis of the large number of ENCODE ChIP-seq data sets can therefore further elucidate MYC interactions at the chromatin level.

ChromNet showed that MAX is the strongest interactor of MYC across multiple cell types (Additional file 1: Table S3), highlighting the ubiquitous nature of this interaction. Top-scoring ChromNet connections also included other known MYC interactors, for example, components of the RNA polymerase II complex such as POLR2A and chromatin-modifying proteins such as EP300 (Additional file 1: Table S3). This shows how ChromNet can help identify protein complexes and interactions.

In addition to the known interactors described above, ChromNet also revealed previously uncharacterized, high-scoring interactions, including the transcriptional regulator Host Cell Factor C1 (HCFC1) (Additional file 1: Table S3). HCFC1 binds largely to active promoters [39] and is involved in biological processes, such as cell cycle progression [46, 50] and oncogenesis [12, 45, 48]. 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 endogenous protein–protein interactions in intact cells [54] and has been used to validate novel interactors of MYC [18]. When two proteins that are probed with specific antibodies are within close proximity of each other, fluorescence signals are produced that are measured and quantified using fluorescence microscopy. We saw only background fluorescence when incubating with an antibody against MYC (Fig. 5a, top) or HCFC1 (Fig. 5a, middle) alone. Incubation with both MYC and HCFC1 antibodies yielded a significant increase in the fluorescence signal in the nuclear compartment (Fig. 5a, bottom, Fig. 5b and Additional file 1: Figure S20). This suggests that MYC and HCFC1 interact in the nucleus, and 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 illustrates 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 data sets from many cell types into a single network, ChromNet enables extraction of global patterns in the relationships among regulatory factors. We used multidimensional scaling [7] to embed the entire network into a 2D layout (Fig. 6 and "Methods"). In this embedding, the spatial proximity of two nodes is designed to reflect their distance in the network, where positive edges pull nodes closer together and negative edges push them father apart. Nodes for the same regulatory factor in different cell types form a cluster when that factor's genomic position is conserved across cell types. For example, CTCF forms a clear cluster in this manner (Fig. 6a). Relationships between regulatory factors are represented by their proximity in the embedding. For example, MYC and MAX nodes are located in the same region; so are CTCF and RAD21. In contrast to the joint network, relationships in individual cell-type-specific networks (Fig. 1b top) are much less distinct (Additional file 1: Figure S21).

The relative positions of regulatory factors in the embedded graph highlight important aspects of biology. This is especially apparent among histone marks, where there is a clear separation between activating marks such as H3K4me3 and H3K27ac on the lower right and repressive marks such as H3K27me3 and H3K9me3 on the upper left (Fig. 6a). H3K27me3 and H3K9me3 are both repressive marks, but form distinct clusters because they target distinct regions of the genome. H3K27me3 marks facultative heterochromatin, thought to regulate temporary repression of gene-rich regions [27]. H3K9me3 marks constitutive heterochromatin, and acts as a more permanent repressor [30]. Between the active and repressive marks, we find H3K36me3 and H3K79me2. H3K36me3 is closer to the inactive marks and is implicated in restricting the spread of H3K27me3 [63]. H3K79me2 varies with the cell cycle and is associated with replication initiation sites [17]. The relative position of histones and protein factors is also interesting. ZNF274 has been implicated in the recruitment of methyltransferases for H3K9me3 and is found nearby in the network [16]. EZH2 is involved in the deposition of H3K27me3 and is found between the H3K27me3 cluster and the rest of the network [61].

The positions of regulatory factor data sets reflect both their cell-type identities and association with chromatin states. Highlighting the three tier 1 ENCODE cell types shows a weak clustering of regulatory factor data sets by cell type (Fig. 6b). K562 and GM12878 are both derived from blood cell lines and overlap spatially with one another in the network more than with the H1-hESC human embryonic stem cells. Coloring the network by correlation with chromatin state also reveals spatial

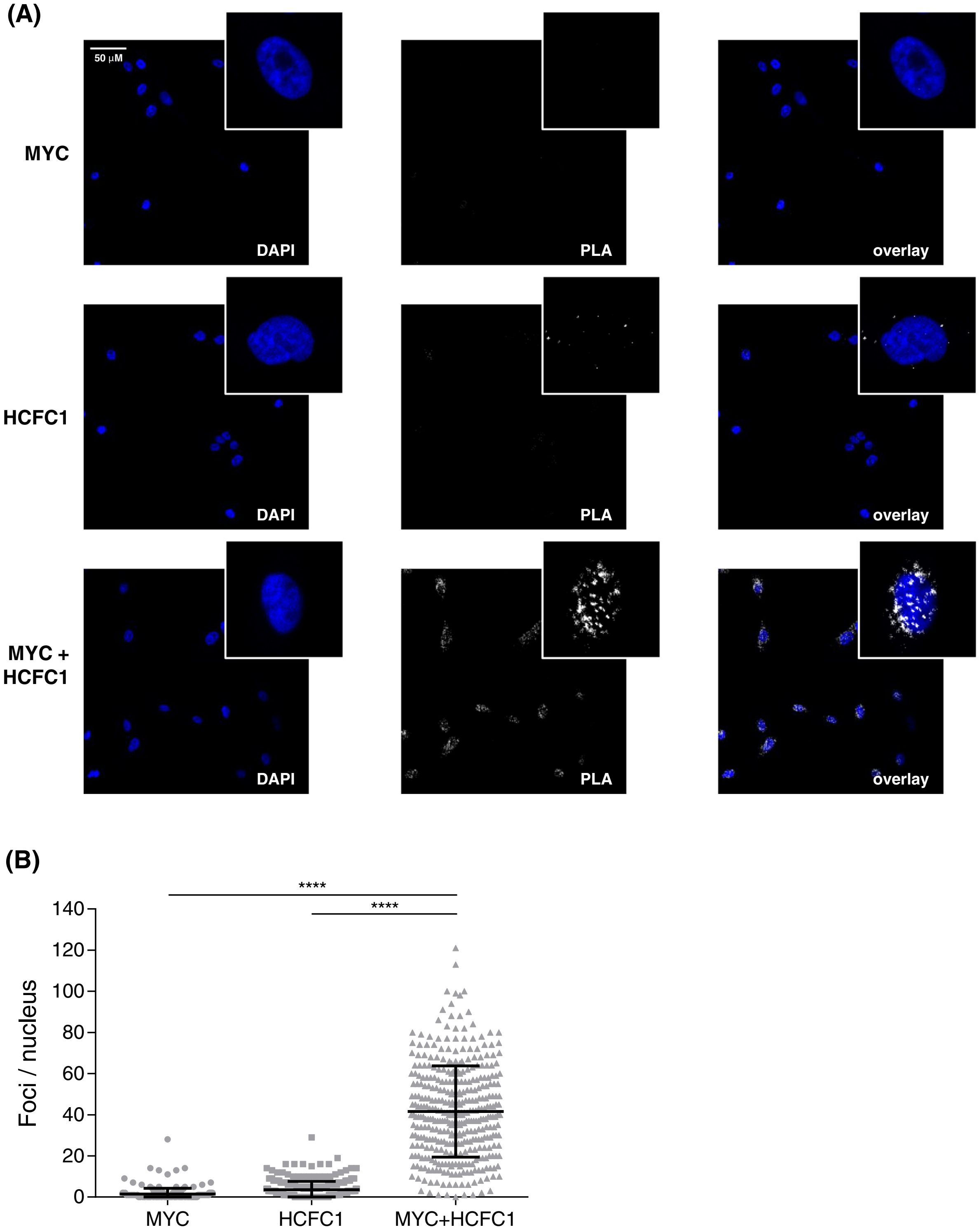

Fig. 5 a Proximity ligation assay showing MYC and HCFC1 interaction in the nucleus. Representative micrographs show DAPI nuclear staining (*left*), proximity ligation signal (*middle*), and overlay (*right*) at 20× magnification, with *insets* at 100× magnification. *Top:* Cells probed with MYC antibody 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 nucleus, with 254, 293, and 381 nuclei quantified for the MYC antibody alone, the HCFC1 antibody alone, and both antibodies together, respectively. Individual values (*gray dots*) and mean ± standard deviation *black bars* from three biological replicates are shown; **** $p < 0.0001$, one-way analysis of variance with Bonferroni post test. Quantifications for each independent replicate are shown in Additional file 1: Figure S20. *PLA* proximity ligation assay

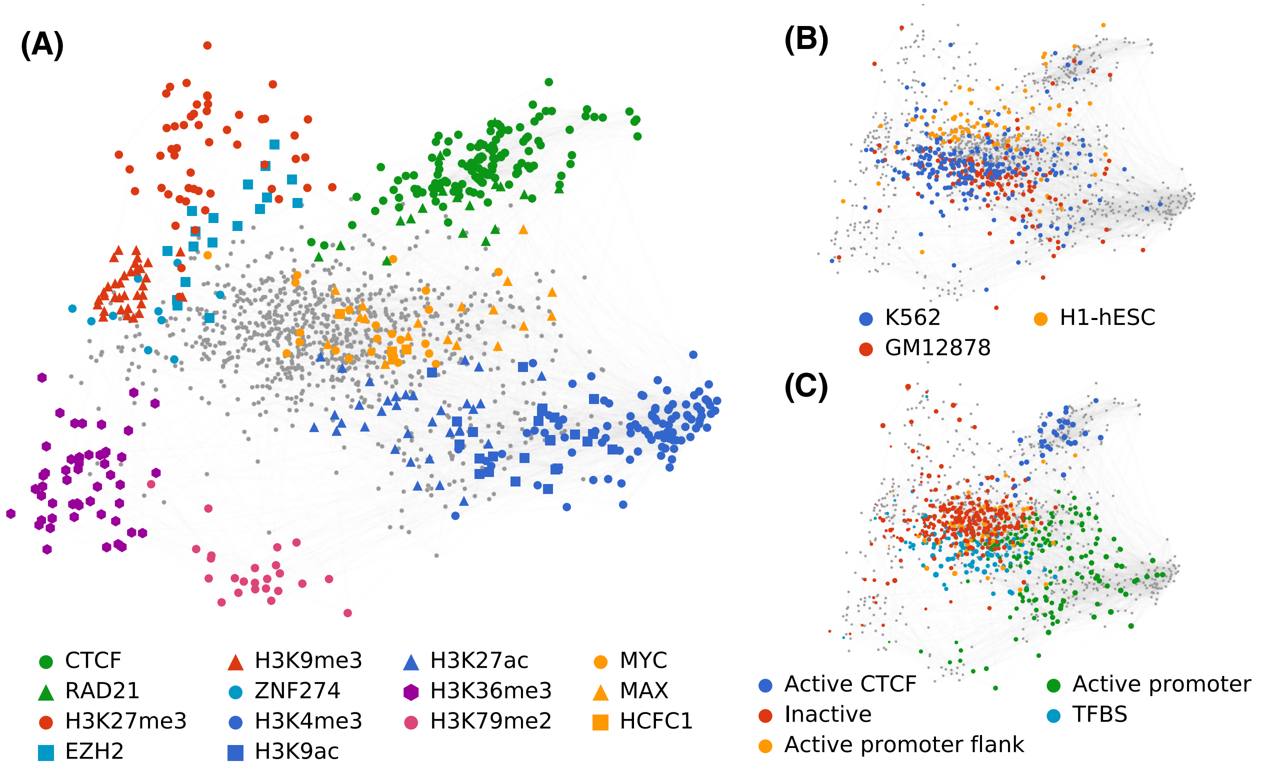

Fig. 6 2D embedding of the entire human chromatin network estimated by ChromNet. The spatial proximity of nodes and node groups reflects the strength of their inferred edge weights. In three views of this embedding, we have highlighted three different aspects: **a** Specific regulatory factors discussed in this article. **b** Data sets from the three ENCODE tier 1 cell types, showing a separation of data sets by cell type. **c** Correlation with five Segway genome annotation labels. We only colored the data sets from cell types where the Ensembl Regulatory Build had a corresponding Segway annotation. Node size represents correlation with a label, in comparison to all other nodes assigned to that label

patterns. We chose five (out of seven) Segway [24, 64] annotation labels that highlight distinct areas of the network (Fig. 6c), illustrating a clear separation between active and inactive regions of the genome, and that chromatin domains are reflected in the interactions of the chromatin network. Spatially embedding regulatory factor data sets using the ChromNet network simultaneously captures many important aspects of their function, such as chromatin state, cell lineage, and known factor–factor interactions.

Discussion

Characterizing the chromatin network, the network of interactions among regulatory 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 a joint conditional-dependence network from a large number of ChIP-seq data sets, such as the need to distinguish direct from indirect regulatory factor interactions while remaining robust to data redundancy. ChromNet also provides an efficient method to learn the genomic context driving an edge, which allows a more comprehensive understanding of the inferred interactions. We demonstrated that ChromNet's GroupGM network infers known protein–protein interactions in the joint chromatin network more accurately than other methods. Unlike many previous methods, ChromNet is also efficient enough to integrate thousands of genome-wide ChIP-seq data sets into a single joint network. To our knowledge, this study represents the first construction of an interaction network from all 1451 ENCODE ChIP-seq data sets. ChromNet already scales to the number of data sets necessary to represent all 1400–1900 human transcription factors [60], once such data is available.

ChromNet provides a general computational framework to identify a joint dependence network from many ChIP-seq data sets. It can build a custom joint dependence network by incorporating user-provided ChIP-seq data sets or a combination of the ENCODE ChIP-seq data sets and user-provided data sets. To allow easier exploration of regulatory 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 data sets (http://chromnet.cs.washington.edu). By building a large model and allowing easy inspection of small subnetworks, ChromNet combines a large-scale conditional-dependence model with practical accessibility.

To demonstrate ChromNet's ability to reveal novel regulatory 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. While this manuscript was under review, the MYC–HCFC1 interaction was independently described by Thomas et al. [59], further strengthening our validation of the interaction discovered through ChromNet and establishing HCFC1 as a bona fide interactor of MYC. Through ChromNet, we identified HCFC1 as a novel interactor of MYC that may be involved in regulating biological and oncogenic functions of MYC.

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 data sets could resolve regulatory 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 regulatory factors at genomic position offsets. Here, we considered only co-occurrence relationships within the same 1000-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 regulatory factors has been used to annotate the genome automatically, 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 regulatory factor interactions and the chromatin states in which they occur.

Methods

Data processing

ENCODE has the largest collection of high-quality ChIP-seq data sets [11], and continues depositing new data sets. ENCODE has processed many ChIP-seq data sets through a uniform pipeline. However, we reprocessed all the data sets from raw ChIP-seq reads (Fig. 2) for two reasons. First, this allowed us to incorporate data sets not available 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 ChIP-seq data sets that are not from the ENCODE project to the ChromNet network.

We aligned reads from 3574 FASTQ files to GRCh38/hg38 [19] using Bowtie2 [31]. We grouped BAM files by data set using metadata from the ENCODE web site [15]. Then, we pooled and processed BAM files using a custom binning method that counts the number of read starts in each of 3,209,287 1000-bp bins covering all contigs in GRCh38/hg38. Binning all count data sets yielded a $X \in \mathbb{Z}_*^{3,209,287\times 1451}$ count-valued *data matrix*. Each bin has a corresponding row in the matrix. We interpreted each of the 3,209,287 rows as a sample from a set $\mathcal{X} = \{X_1, \ldots, X_p\}$ of $p = 1451$ count-valued random variables representing occupancy of each regulatory 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 data sets 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 data sets (Fig. 1a, bottom left).

Generation of simulated data

A large-scale simulated data set was generated to validate the ability of ChromNet to recover interactions from raw count data. Representing the conditional dependence among large numbers of count variables for the purpose of simulation is not trivial. It is important that the model is not overly simplistic, but also still interpretable. Here we use a multivariate Gaussian distribution to represent the means of marginal Poisson distributions, threshold the values when they fall below zero, and add additional negative binomial distributed noise to represent random reads unrelated to regulatory factor localization.

In this model the count for a ChIP-seq data set j, c_j, at a given position is described by

$$c_j = \nu_j + \epsilon_j, \tag{1}$$

where $\nu_j \sim \text{Poisson}(\text{rate}_j), \text{rate}_j = \max(0, s_j)$, and $\epsilon_j \sim \text{NegativeBinomial}(r, p)$. The signal follows a thresholded normal distribution $s_j \sim \max(N(\mu, \Sigma), 0)$. The background noise ($r = 25, p = 0.9$) and the parameters of the normal distribution are all fixed during the course of the simulation. Σ^{-1} represents the structure of the underlying conditional-dependence network.

The inverse covariance matrix of the simulated data Σ^{-1} was randomly generated with a sparsity of 10 %. In addition, data sets were grouped into complexes of size one (60 %), two (20 %), or three (20 %) to represent the type of close coupling observed in real data among some factors. The correlation within complexes was set to be between 0.8 and 0.9 to match the magnitudes of high correlations observed in real data (Additional file 1: Figure S3). A total of 80 complexes were simulated across 200,000 positional samples. This results in 126 experiments and 200,000 samples. To model the complexities found in real data sets better, we added dependency between nearby samples by replacing each rate_j with

$$\frac{1}{20}\text{rate}_{j-1} + \frac{9}{10}\text{rate}_j + \frac{1}{20}\text{rate}_{j+1}.$$

This caused nearby bins to be more similar to each other and thus the samples are not independently and identically distributed. Since larger correlations between regions of the genome are also present due to batch effects or other confounding factors, we added one of eight different random genome-wide batch effects to each of the 126 data sets.

The resulting marginal count distributions from this model are visually similar to those observed in real data (Additional file 1: Figure S12). Because we based the correlations between data sets on a (largely transformed) multivariate normal distribution, we can treat data sets connected in the underlying generative model as true connections and seek to recover them using a variety of methods. The results of this analysis are shown in Additional file 1: Figure S13, which is consistent with Fig. 3, where the group graphical model performs better than alternative approaches including correlation, inverse correlation, and partial correlation.

Efficient estimation of conditional dependence from count data

Given data sets drawn from a set $\mathcal{X}$ of count-valued random variables, learning an exact joint model that captures the dependency structure of these data sets could be challenging. Although there are a variety of multivariate count distributions, all are either overly restrictive or challenging to estimate for large numbers of variables [62]. A common alternative is to use a multivariate Gaussian distribution and some type of transform on the marginals to make them more Gaussian, such as sqrt or asinh. Since count data are often heteroscedastic, where variance increases with higher counts, these transforms squash higher values, making the distribution more symmetric. This causes the least-squares error term to focus less on high-valued samples and proportionately more on lower values. Interestingly, for ChIP-seq data sets, this is not desirable because higher values are more likely to represent a strong signal while lower values are more likely driven by noise.

Because of its efficiency and interpretability, we used a multivariate Gaussian approximation to the count data for ChromNet. We also chose to use untransformed raw read counts in the model. This choice was based on observing a clear decrease in performance when using transforms designed to mitigate heteroscedasticity (Additional file 1: Figure S1). An additional benefit of using a multivariate Gaussian is that it can also serve as a reasonable approximation to a Markov random field distribution. This allows a comparison with other methods designed to work strictly with binary data (Additional file 1: Figures S22, S23 and Supplementary Note 3).

To create the *inverse correlation network* (Fig. 3), we began by inverting the sample correlation matrix $\hat{\Sigma}$ to get an inverse sample correlation matrix $\hat{\Sigma}^{-1}$ [33, 37]. We then set the weight of every edge between two data sets i and j equal to the corresponding entry $\{\hat{\Sigma}^{-1}\}_{i,j}$. This inverse correlation network captures the pairwise linear dependence between two data sets when conditioned on all other variables in the network.

Note that partial correlation is very similar to inverse correlation and has been used before by Lasserre et al. to model connections between histone marks from human ChIP-seq data effectively (using rank-transformed data from gene start sites) [32]. The matrix of partial correlations, P, is a renormalization of inverse correlation $P = -D^{-1/2} \times \Sigma^{-1} \times D^{-1/2}$ where D is the diagonal matrix of Σ^{-1}. A direct application of partial correlation to all ENCODE data suffers from the same issues as inverse correlation, performing slightly worse in the recovery of known protein–protein interactions (Additional file 1: Figure S4). We chose to use inverse correlation as the foundation of the group graphical model (GroupGM) because the proof that GroupGM recovers the correct edge weights in the presence of near perfect redundancy does not hold when applied to the partial correlation matrix (Additional file 1: Supplementary Note 2).

One additional concern when applying a Gaussian graphical model to ChIP-seq data is that the values at each 1000-bp bin in the genome are not independent of each other. Fortunately, while this may reduce the power of the model (i.e., it will need more samples), it does not bias the model. This is because the edges of a Gaussian graphical model can be interpreted in terms of linear regression coefficients. Standard linear regression coefficients are unbiased even when samples are not statistically independent when the data follows a linear relationship. To validate this on ChIP-seq data and to confirm that any loss of power is unimportant, we evenly subsampled the data at progressively larger intervals. We found that performance when recovering known protein–protein interactions does not degrade until we subsample 100-fold (Additional file 1: Figure S24).

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 data sets by considering a set $\mathcal{G}$ of q groups chosen by hierarchical clustering (see below). This effectively allows edges to express relationships between groups of variables. 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 (Fig. 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 Eq. 2 correctly maintains the original edge magnitude when there is redundancy (Additional file 1: Supplementary Note 2).

To select the set $\mathcal{G}$ of groups, we used complete-linkage hierarchical agglomerative clustering of the correlation matrix [23]. This clustering method starts by merging the two groups with the smallest maximum correlation distance between their data sets, then continues recursively until all groups have been merged. The use of hierarchical clustering eliminates the need to choose 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 Eq. 2, where $q = 2p - 1$ (the total number of internal and leaf nodes). This method avoids comparing all possible subsets of data sets, which would make calculating G prohibitively expensive. Since groups with low correlation are less likely to cause the collinearity problem (Fig. 1c), we only consider groups with a correlation greater than 0.8, which captures 53 % of all the multi-factor groups formed by the hierarchical clustering (Additional file 1: Figure S3).

Since GroupGM uses the cluster assignments to mitigate strong redundancy, clustering accuracy is most important for tightly correlated data sets. When two data sets are highly correlated, it is important to group them together to mitigate the outcome of correlated data sets in network inference. When two data sets are only mildly correlated, the effects of their redundancy will also be mild, so it is less important to group them together. Hierarchical clustering is an attractive choice because it starts by creating groups among the most correlated data sets.

Computing the genomic context that drives a network edge

The conditional-dependence relationships represented by an edge in ChromNet can occur primarily in certain genomic regions. Here we seek to identify what parts of the genome (i.e., samples) drove the creation of an edge in ChromNet. Understanding what positions in the genome caused ChromNet to estimate a network edge provides insight into the genomic regions driving the relationship.

The most natural way to define the influence of a genomic position (i.e., sample) on an edge is as the difference in edge value between when we observe a position and when we do not observe a position in the genome. If implemented directly, this could easily become computationally intractable since it involves relearning the entire model for every position in the genome. For a highly optimized implementation on 16 cores, computing the correlation matrix takes approximately 2 minutes, which would lead to a run time of over 12 years for 3,209,287 binned genomic positions. This can be sped up dramatically by using rank-1 matrix updates to avoid recalculating most of the correlation matrix. This results in a much faster method, where the slowest step is the inversion of the correlation matrix. However, computing this inversion for each genomic sample still leads to over 4 days of computation on recent high-performance servers. Pre-computing this information is also undesirable, since it would create 54 TB of largely incompressible data for all group edges. Below we show that for the ChromNet model, the calculation of a genomic position's impact on an edge can be made extremely efficient. The ideas are similar to those used in efficient leave-one-out cross-validation implementations for linear models.

Removing a genomic position and computing the new inverse correlation matrix can be written in terms of a rank-1 update and the inverse correlation matrix before the position (sample) is removed. This equation holds under the assumption that removing the sample does not change the mean of the data. Let Σ be the correlation matrix of all the data, and $\bar{\Sigma}$ be the correlation matrix with the sample removed. Let u be the column vector representing the sample to be removed (already mean centered). Letting D be a normalizing diagonal matrix $D_{i,i} = \sqrt{1 - u_i^2}$, we get:

$$\begin{aligned}
\bar{\Sigma} &= (D^{-1}(\Sigma - uu^{\mathrm{T}})D^{-1})^{-1} && (3)\\
&= D(\Sigma - uu^{\mathrm{T}})^{-1}D && (4)\\
&= D(\Sigma + uBu^{\mathrm{T}})^{-1}D \qquad (B = -1) && (5)\\
&= D(\Sigma^{-1} - \Sigma^{-1}u(B^{-1} + u^{\mathrm{T}}\Sigma^{-1}u)^{-1}u^{\mathrm{T}}\Sigma^{-1})D \\
&\qquad\qquad \text{(Woodbury formula)} && (6)\\
&= D(\Sigma^{-1} - \Sigma^{-1}u(-1 + u^{\mathrm{T}}\Sigma^{-1}u)^{-1}u^{\mathrm{T}}\Sigma^{-1})D && (7)\\
&= D(\Sigma^{-1} - v(-1 + u^{\mathrm{T}}v)^{-1}v^{\mathrm{T}})D \qquad (\Sigma^{-1}u = v) && (8)\\
&= D\left(\Sigma^{-1} - \frac{vv^{\mathrm{T}}}{u^{\mathrm{T}}v - 1}\right)D. && (9)
\end{aligned}$$

Included in the ChromNet software release is an optimized implementation utilizing the above inverse rank-1 update formulation. It can solve 40,000 model updates to the full joint chromatin network per second, which leads to a run time of just over 1 minute for a single-group edge over the human genome. The output is the effect each genomic position has on an edge when that position is added to the data set. This information can be used to examine the highest impact positions and determine the genomic context driving an edge (Additional file 1: Figure S19).

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 (Fig. 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. 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 0.2 to capture only edges enriched for known interactions (Additional file 1: Figure S25).

Fold enrichment reflects both type I and type II error rates

The fold enrichment is a single quantity that captures the effects of both type I and type II error rates. This can be seen from the definition of fold enrichment:

$$\text{fold enrichment} = \frac{\text{\# of correct edges}}{\text{\# of randomly correct edges}} \tag{10}$$

$$= \frac{\text{TP}}{\left(\text{\# network edge predictions}\right) \times (\text{\# BioGRID interactions})\,/N} \tag{11}$$

$$= \frac{\text{TP} \times N}{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN})} \tag{12}$$

where N is the total number of possible edges, and TP, FP, and FN refer to the number of true positives, false positives, and false negatives, respectively. The fold enrichment is inversely proportional to the number of false positives (type I error) and number of false negatives (type II error). The type I error rate is equal to (type I error) /(total number of BioGRID interactions), and the type II error rate is equal to (type II error)/(total number of interactions − number of BioGRID interactions). Since the denominators of the type I and type II error rates are fixed numbers, we can say that the fold enrichment is inversely proportional to the type I and type II error rates.

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 (Fig. 3) using bootstrap resampling over regulatory factors. We performed resampling over regulatory factors, and not over edges or individual data sets, because valid bootstrap resampling assumes independent and identically distributed samples. If we had resampled over the edges, we would have estimated a much smaller variability. This is because edges do not vary independently, and changes in a single data set can affect all edges connected to that data set. Variation specific to a single regulatory factor would affect all data sets measuring that factor. Those individual data sets, therefore, lack the independence assumed by the bootstrap sampling.

Under a regulatory factor bootstrap, we might sample a widely measured regulatory factor a number of times. For example, ChromNet contains 130 CTCF data sets. Every time we sample CTCF, we add all 130 of these columns (where a column represents a variable in the data matrix X) to the bootstrap data matrix. Adding many data sets in unison greatly increases variability in the resampled 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 (Fig. 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 (Additional file 1: Figures S10 and S14).

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 1 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). 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 the manufacturer's instructions (Sigma-Aldrich, St. Louis, MO). We imaged six fields of view per slide with a LSM700 confocal fluorescence microscope (Zeiss, Oberkochen, Germany). We unbiasedly quantified the PLA signal per nucleus (as defined by DAPI staining) using the software ImageJ [52].

Embedding the full chromatin network into a single plot

Embedding a graph into a space involves defining distances between all nodes in the graph. Because GroupGM is inherently multi-scale, we sought a distance metric that accurately represented forces between individual nodes, and between all possible node groupings. In GroupGM, the edge weight between two groups is the sum of the conditional-dependence weights between all the individual data sets of those groups.

A common method of computing graph distances that accounts for the total effect of all edges between two groups is the resistance distance [29]. The name is derived from an interpretation of the distance as the electrical resistance between two nodes in the graph where edges are viewed as wires. This can be computed as:

$$\Omega_{i,j} = \Gamma_{i,i} + \Gamma_{j,j} - \Gamma_{i,j} - \Gamma_{j,i},$$

where Γ is the inverse of the graph Laplacian. While at first glance the resistance distance may seem like an arbitrary metric to use for node distances, upon closer inspection we find striking parallels between it and Gaussian graphical models. First, note that the weighted graph Laplacian [41], L, is defined as:

$$L = W \otimes (D - A),$$

where D is a diagonal matrix of edge degrees, A is the binary adjacency matrix of the graph, W is a matrix of positive edge weights, and $\otimes$ represents element-wise multiplication. A general Gaussian graphical model has a complete graph, so A will be all ones, and D will be constant on the diagonal. The edge weights will be symmetric and can be positive or negative. Positive edge weights will lead to negative off-diagonal entries in L, just as positive connections in the GroupGM will lead to negative off-diagonal entries in $\Theta = \Sigma^{-1}$. So by allowing W to contain negative entries, we can view Θ as a type of graph Laplacian.

Viewing Θ as a type of graph Laplacian allows us to compute the resistance distance by setting $\Gamma = \Theta^{-1}$. Simplifying gives $\Omega_{i,j} = 1 - \Theta^{-1}_{i,j}$.

So, the resistance distance is just a constant offset of the correlation matrix of the network. This means that if we are trying to compute distances between nodes in a graph represented by the inverse correlation matrix, correlation is a very natural distance measure. We note, however, that unlike the original data correlation matrix, this matrix is computed from the inverse of the edge weights matrix. This causes a difference because we threshold small edge values that are likely to represent only noise. We chose this threshold to maximize the visual clarity of the network, which led to a threshold of 0.01.

We overlaid chromatin state annotation on the graph embedding by computing the correlation between each data set and each Segway [24] region from the Ensembl Regulatory Build for GRCh38/hg38 [64]. We drew a separate network labeling for each region by sizing each data set 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 (Fig. 6c).

Availability of supporting data

ChromNet is freely available as a ready-to-use package under an Apache license at https://github.com/slundberg/ChromNet.jl. Supporting data including a preprocessed data matrix from all human ENCODE ChIP-seq data are linked from the code repository and at http://dx.doi.org/10.5281/zenodo.45900. The microscopy data that we used for validation of the MYC–HCFC1 interaction are available at http://dx.doi.org/10.5281/zenodo.45768.

Additional files

Additional file 1: Supplementary information. (PDF 4,962 kb)

Additional file 2: ENCODE data sets used in ChromNet (1451 total). (TXT 34 kb)

Ethical approval

No ethical approval was required.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SML and SIL designed the statistical methods and various analyses. SML developed the algorithm, performed the analyses, and developed the visualization tool. SML, WBT, LZP, MMH and SIL wrote the manuscript. WBT and LZP designed and performed the PLA. BR contributed to the data analysis of the experimental validation. All authors read and approved the final manuscript.

Acknowledgements

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

Funding

This work was supported by a National Science Foundation (NSF) Graduate Research Fellowship to SML (DGE-1256082), a NSF grant to SIL (DBI-1355899), a Natural Sciences and Engineering Research Council of Canada grant to MMH (RGPIN-2015-03948), a Canada Research Chair in Molecular Oncology to LZP, a Canadian Institute for Health Research grant to LZP and BR (MOP-275788), and a Canadian Breast Cancer Foundation Ontario Region Doctoral Fellowship to WBT. Cloud computing resources for this research were generously provided by Google.

Author details

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

Received: 14 January 2016 Accepted: 15 March 2016
Published online: 30 April 2016

References

1. Au SLK, et al. EZH2-mediated H3K27me3 is involved in epigenetic repression of deleted in liver cancer 1 in human cancers. PloS One. 2013;8:e68226.
2. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: identifying influential data and sources of collinearity. Vol. 571. Hoboken, NJ: John Wiley & Sons; 2005.
3. van Bemmel JG, et al. A network model of the molecular organization of chromatin in *Drosophila*. Mol Cell. 2013;49:759–71.
4. Berger SL. The complex language of chromatin regulation during transcription. Nature. 2007;447:407–12.
5. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125:315–26.
6. Blackwood EM, Eisenman RN. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science. 1991;251:1211–17.
7. Borg I, Groenen PJF. Modern multidimensional scaling: theory and applications. Berlin, Germany: Springer Science & Business Media; 2005.
8. Cao R, Zhang Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Mol Cell. 2004;15:57–67.
9. Carroll RJ, Ruppert D. Transformation and weighting in regression. Vol. 30. Boca Raton, FL: CRC Press; 1988.
10. Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. Annu Rev Biochem. 2009;78:273–304.
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
12. Dey A, et al. Loss of the tumor suppressor BAP1 causes myeloid transformation. Science. 2012;337:1541–6.
13. Di Croce L1, Helin K. Transcriptional regulation by Polycomb group proteins. Nat Struct Mol Biol. 2013;20:1147–55.
14. Diamond MI, et al. Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. Science. 1990;249: 1266–72.
15. Encode Project Data. encodeproject.org.
16. Frietze S, et al. ZNF274 recruits the histone methyltransferase SETDB1 to the **3′** ends of ZNF genes. PLoS One. 2010;5:e15082. Accessed 15 Jan 2016.
17. Fu H, et al. Methylation of histone H3 on lysine 79 associates with a group of replication origins and helps limit DNA replication once per cell cycle. PLoS Genet. 2013;9:e1003542.
18. Garcia-Sanz P, et al. Sin3b interacts with Myc and decreases Myc levels. J Biol Chem. 2014;289:22221–36.
19. Genome Reference Consortium. genomereference.org.
20. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489:91–100. Jan 2015.
21. Gupta K, et al. Mmip1: a novel leucine zipper protein that reverses the suppressive effects of Mad family members on c-Myc. Oncogene. 1998;16:1149–59.
22. Hann SR. MYC cofactors: molecular switches controlling diverse biological outcomes. Cold Spring Harb Perspect Med. 2014;4:a014399.
23. Hastie T, et al. The elements of statistical learning. Vol. 2.1. Berlin, Germany: Springer; 2009.
24. Hoffman MM, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6.
25. Juan D, et al. Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs. bioRxiv. 2015. doi:10.1101/008821.
26. Khare SP, et al. HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. Nucleic Acids Res. 2012;40: D337–42.
27. Kim J, Kim H. Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. ILAR J. 2012;53:232–9.
28. Kim TW, et al. Ctbp2 modulates NuRD-mediated deacetylation of H3K27 and facilitates PRC2-mediated H3K27me3 in active embryonic stem cell genes during exit from pluripotency. Stem Cells. 2015;33:2442–55.
29. Klein DJ, Randić M. Resistance distance. J Math Chem. 1993;12:81–95. doi:10.1007/BF01164627.
30. Lachner M, O'Sullivan RJ, Jenuwein T. An epigenetic road map for histone lysine methylation. J Cell Sci. 2003;116:2117–24.
31. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
32. Lasserre J, Chung HR, Vingron M. Finding associations among histone modifications using sparse partial correlation networks. PLoS Comput Biol. 2013;9:e1003168.
33. Lauritzen SL. Graphical models: Oxford University Press; 1996.
34. Le NT, et al. A nucleosomal approach to inferring causal relationships of histone modifications. BMC Genom. 2014;15(Suppl 1):S7.
35. Lee CC, et al. TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer. J Biol Chem. 2012;287:2798–809.
36. Losada A, Yokochi T, Hirano T. Functional contribution of Pds5 to cohesin-mediated cohesion in human cells and *Xenopus* egg extracts. J Cell Sci. 2005;118:2133–41.
37. Mardia KV, Kent JT, Bibby JM. Multivariate analysis: Academic Press; 1979.
38. Meyer N, Penn LZ. Reflecting on 25 years with MYC. Nat Rev Cancer. 2008;8:976–90.
39. Michaud J, et al. HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. Genome Res. 2013;23:907–16.
40. Min MR, et al. Interpretable sparse high-order Boltzmann machines. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics; 2014. p. 614–22.
41. Newman M. Networks: an introduction. Oxford, UK: Oxford University Press; 2010.
42. Niu W, et al. Diverse transcription factor binding features revealed by genome-wide ChIPseq in *C. elegans*. Genome Res. 2011;21:245–54.
43. Paige SL, et al. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell. 2012;151:221–32.
44. Panigrahi AK, et al. A cohesin-RAD21 interactome. Biochem J. 2012;442: 661–70.
45. Parker JB, et al. A transcriptional regulatory role of the THAP11-HCF-1 complex in colon cancer cell function. Mol Cell Biol. 2012;32:1654–70.
46. Parker JB, et al. Host Cell Factor-1 recruitment to E2F-bound and cell-cycle-control genes is mediated by THAP11 and ZNF143. Cell Rep. 2014;9:967–82.
47. Patel JH, et al. Analysis of genomic targets reveals complex functions of MYC. Nat Rev Cancer. 2004;4:562–8.
48. Peña-Llopis S, et al. BAP1 loss defines a new class of renal cell carcinoma. Nat Genet. 2012;44:751–9.
49. Perner J, et al. Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. Nucleic Acids Res. 2014;42:13689–95.
50. Piluso D, Bilan P, Capone JP. Host cell factor-1 interacts with and antagonizes transactivation by the cell cycle regulatory factor Miz-1. J Biol Chem. 2002;277:46799–808.
51. Rosenbloom KR, et al. ENCODE data in the UCSC genome browser: year 5 update. Nucleic Acids Res. 2013;41:D56–63.
52. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671–5.
53. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15:272–86.

54. Söderberg O, et al. Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. Methods. 2008;45: 227–32.
55. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.
56. Stark C, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.
57. van Steensel B, et al. Bayesian network analysis of targeting interactions in chromatin. Genome Res. 2010;20:190–200.
58. Thomas LR, et al. Interaction with WDR5 promotes target gene recognition and tumorigenesis by MYC. Mol Cell. 2015;58:440–52.
59. Thomas LR, Foshage AM, Weissmiller AM, Popay TM, Grieb BC, Qualls SJ, Ng V, Carboneau B, Lorey S, Eischen CM, et al. Interaction of MYC with host cell factor-1 is mediated by the evolutionarily conserved Myc box IV motif. Oncogene. 2015. Nature Publishing Group doi:10.1038/onc.2015.416.
60. Vaquerizas JM, et al. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10:252–63.
61. Viré E, et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature. 2006;439:871–4.
62. Winkelmann R. Econometric analysis of count data. Heidelberg: Springer Berlin; 2008, pp. 203–39. doi:10.1007/978-3-540-78389-3_7.
63. Yuan W, et al. H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. J Biol Chem. 2011;286:7983–9.
64. Zerbino DR, et al. The Ensembl Regulatory Build. Genome Biol. 2015;16:56.
65. Zhou J, Troyanskaya OG. Global quantitative modeling of chromatin factor interactions. PLoS Comput Biol. 2014;10:e1003525.
66. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011;12: 7–18.
67. Zlotorynski E. Chromatin: ZNF143 in the loop. Nat Rev Mol Cell Biol. 2015;16:127.