7690–7704 Nucleic Acids Research, 2012, Vol. 40, No. 16 Published online 6 June 2012
doi:10.1093/nar/gks501

Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages

Xun Lan[1], Heather Witt[2,3], Koichi Katsumura[4], Zhenqing Ye[1], Qianben Wang[5], Emery H. Bresnick[4], Peggy J. Farnham[2,3,*] and Victor X. Jin[1,*]

[1]Department of Biomedical Informatics, 460 W 12th Avenue, 212 BRT, The Ohio State University, Columbus, OH 43210, [2]Department of Biochemistry and Molecular Biology, [3]Norris Comprehensive Cancer Center, 1450 Biggy Street, NRT 6503, University of Southern California, Los Angeles, CA 90089, [4]Department of Cell and Regenerative Biology, Wisconsin Institutes for Medical Research, University of Wisconsin Carbone Cancer Center, University of Wisconsin School of Medicine and Public Health and [5]Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA

Received March 16, 2012; Revised May 2, 2012; Accepted May 4, 2012

ABSTRACT

We have analyzed publicly available K562 Hi-C data, which enable genome-wide unbiased capturing of chromatin interactions, using a Mixture Poisson Regression Model and a power-law decay background to define a highly specific set of interacting genomic regions. We integrated multiple ENCODE Consortium resources with the Hi-C data, using DNase-seq data and ChIP-seq data for 45 transcription factors and 9 histone modifications. We classified 12 different sets (clusters) of interacting loci that can be distinguished by their chromatin modifications and which can be categorized into two types of chromatin linkages. The different clusters of loci display very different relationships with transcription factor-binding sites. As expected, many of the transcription factors show binding patterns specific to clusters composed of interacting loci that encompass promoters or enhancers. However, cluster 9, which is distinguished by marks of open chromatin but not by active enhancer or promoter marks, was not bound by most transcription factors but was highly enriched for three transcription factors (GATA1, GATA2 and c-Jun) and three chromatin modifiers (BRG1, INI1 and SIRT6). To investigate the impact of chromatin organization on gene regulation, we performed ribonucleicacid-seq analyses before and after knockdown of GATA1 or GATA2. We found that knockdown of the GATA factors not only alters the expression of genes having a nearby bound GATA but also affects expression of genes in interacting loci. Our work, in combination with previous studies linking regulation by GATA factors with c-Jun and BRG1, provides genome-wide evidence that Hi-C data identify sets of biologically relevant interacting loci.

INTRODUCTION

Transcriptional regulation involves a process by which different transcription factors bind to specific short deoxyribonucleic acid (DNA) sequences termed *cis*-regulatory elements (CREs), such as promoters, enhancers, silencers and insulators, and thus control the transcription of different genes. The accessibility of these CREs is often influenced by epigenetic modifications including histone acetylation and methylation, which can be associated with the activation or repression of genes. For example, H3K27ac is found at both active enhancers and promoters (1,2); H3K4 mono-, di- and tri-methylation is linked to gene activation (1,3,4), H3K27me3 is a mark of repressed regions (1,5–8), and H3K36me3 identifies transcribed regions (4,9).

Chromatin Immunoprecipitation sequencing (ChIP-seq) and DNase-seq are high throughput experimental technologies that have been shown to be effective in defining a detailed map of transcription factor-binding sites (TFBSs), histone modifications and open chromatin regions. Such techniques have been adopted by the Encyclopedia of DNA Elements (ENCODE) Consortium (http://encodeproject.org/ENCODE/) for the identification of many different TFBSs in various cell types, such as K562 (chronic myelogenous leukemia), GM12878 (lymphoblastoid cell), HepG2 (liver hepatocellular carcinoma) and HeLa (cervical cancer) (10); see http://encodeproject.org/ENCODE/cellTypes.html for a list of all ENCODE cell types. Many studies

*To whom correspondence should be addressed. Tel: +323 442 8015; Email: pfarnham@usc.edu
Correspondence may also be addressed to Dr. Victor X. Jin. Tel: +614 292 6931; Fax: +614 688 6600; Email: Victor.Jin@osumc.edu

have shown that certain transcription factors, such as MYC, the E2F family and YY1, usually bind to promoter regions, whereas many other factors such as GATA1, TCF7L2 (also called TCF4) and estrogen receptor α preferentially bind to distal regions that could be >20 kb away from a known transcription start site (TSS) (11–18). Although some distal binding sites may function as promoters for unannotated protein coding and/or non-coding genes, it is clear that binding sites can control gene regulation via specific three-dimensional (3D) conformations of the chromatin that bring them into close spatial contact with distant promoters (19–22).

The development of the chromosome conformation capture technique (23) has greatly facilitated our understanding of the effects of chromatin conformation on transcriptional regulation owing to greatly increased resolution over traditional co-localization techniques such as fluorescent *in situ* hybridization (24). Recently, by coupling with next generation sequencing technologies, Hi-C has, for the first time, enabled an unbiased genome-wide capturing of chromatin interactions (25). This study identified thousands of interacting loci in both K562 and GM06990 cells and identified nuclear substructures termed 'fractal globules'. A recent review (24) has proposed that there might be four types of genomic interactions, including contacts associated with nuclear lamina, nuclear pores and the nucleolus, as well as intra- and inter-chromosomal contacts. Although these recent studies provide great advances, there still remain many computational and biological challenges in organizing and deciphering Hi-C data. For example, the Hi-C data were initially modeled as a simple probability matrix and the identified interacting loci are thus at a 1 Mb scale. However, if the Hi-C data are modeled based on a statistical distribution of the real data, the interactions can not only be determined at finer scales but can also be differentiated into different types of interacting events (e.g. intra- versus inter-chromosomal interactions and random versus proximate ligation events). Also, the initial studies did not attempt to understand how epigenetic modifications correlate with the 3D chromatin interactions nor did they investigate how the binding of transcription factors might play a role in 3D genome organization. Although a recent study (26) correlated CCCTC-binding factor (CTCF)-binding sites with Hi-C data to investigate genome-wide CTCF-mediated interactions, it was purely an *in silico* computational analysis and did not comprehensively use other publically available transcription factor binding data.

In our study, we have integrated the available K562 Hi-C data with multiple data sets from the ENCODE Consortium, including ChIP-seq data for 45 Transcription Factors (TFs) and 9 histone modifications and DNase-seq data for open chromatin to dissect the underlying mechanisms of chromatin organization and its impact on genome regulation. We identified 12 distinct chromatin clusters that can be categorized into two different types. Our integrated analysis suggests that transcription factors and chromatin modifiers assemble to form functional complexes that bring distant elements in contact. To test this hypothesis, we used knockdown of transcription factors and ribonucleicacid (RNA)-seq analyses to provide genome-wide evidence that Hi-C data can identify sets of biologically relevant interacting loci.

MATERIALS AND METHODS

Overview of the integrated data analysis flow

In this study, we have comprehensively performed data modeling, analysis and integration to investigate the relationship of the spatial organization of the human genome with the local chromatin status and how it affects gene regulation (Figure 1). We began with analysis of K562 Hi-C data (25) using a Mixture Poisson Regression Model (MPRM) (27,28) and a power-law decay background to obtain a set of interacting genomic regions (composed of interacting loci with a pair of two ends) with a high level of specificity. We then associated the interacting partner loci with 9 histone modification

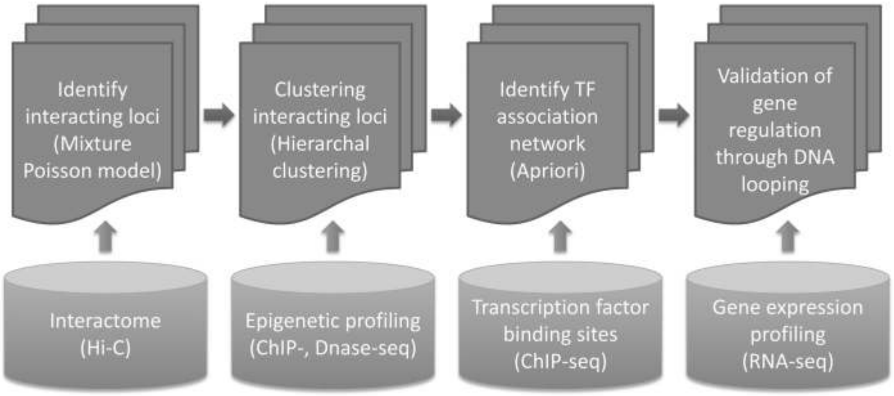
Identify interacting loci (Mixture Poisson model)
Clustering interacting loci (Hierarchal clustering)
Identify TF association network (Apriori)
Validation of gene regulation through DNA looping
Interactome (Hi-C)
Epigenetic profiling (ChIP-, Dnase-seq)
Transcription factor binding sites (ChIP-seq)
Gene expression profiling (RNA-seq)

Figure 1. Flow chart of data processing. The sequential analytical steps of this study rely on several types of experimental input. The process is as follows: (i) analyze Hi-C data using a MPRM and a power-law decay background; (ii) group interacting loci using hierarchical clustering based on the loci's epigenetic status as determined by ChIP-seq analysis of modified histones and regions of open chromatin; (iii) apply the Apriori algorithm to identify transcription factor (TF) association networks using ChIP-seq analysis of transcription factors and (iv) validate the effect of TF-induced DNA loops on gene regulation through comparison of gene expression profiling using RNA-seq before and after knockdown of a single transcription factor.

marks and open chromatin regions, followed by performing hierarchical clustering to classify the sets of partner loci into different groups. Next, we examined the distance between the TSS and the interacting loci in each cluster and applied the Apriori data mining algorithm (29) to identify which transcription factors may be involved in mediating the different sets of chromatin interactions. We also correlated the different sets of interacting loci with gene expression data such that each cluster was placed into one of two types: Type I is composed of active genes; Type II is composed of repressed genes. Finally, we performed a functional validation of one of the identified clusters of interacting loci using knockdown of transcription factors followed by RNA-seq analyses.

Modeling of Hi-C data to identify interacting loci

K562 Hi-C data and gene expression data generated by Lieberman-Aiden (2009) were downloaded from the Gene Expression Omnibus (GEO) database. In Hi-C experiments, the chromatin is treated with formaldehyde to create protein–DNA and protein–protein interactions and, subsequently, digested with HindIII. The digested DNA (hereafter referred to as DNA segments) is ligated in the presence of biotin-labeled nucleotides in a diluted environment then treated with exonuclease to digest linear DNAs but leave DNA loops protected. The chromatin is sonicated; biotin-labeled hybrid DNA fragments (hereafter referred to as hybrid fragments to be distinguished from DNA segments) are precipitated using avidin-conjugated beads and subjected to paired-end sequencing; see Lieberman-Aiden et al. (25) for details. Each hybrid fragment identifies a potential interaction between two loci based on where the two ends of each fragment are mapped to the genome.

We have re-analyzed the K562 Hi-C data set, starting by separating the ligation events into two categories, proximate (defined as a ligation between two ends that are spatially adjacent to each other) and random ligation. Because the probability of a ligation event between two proximate regions is much higher than that between two random regions, the higher the number of hybrid fragments derived from the same two regions the higher the confidence that these two regions are spatially close. A MPRM (Figure 2A; Supplementary Figure S1A and B; see also Supplementary Methods) can be used to determine a threshold number of hybrid fragments beyond which the majority of the data corresponds to proximate ligation events. A previous study has shown that the background contact probability of two loci follows a power-law distribution within a certain distance interval (25). It is necessary to remove background ligations formed between two digested DNA segments that are within a short distance of each other in the genome (Supplementary Figure S1C; see also Supplementary Methods). In addition to random ligation, self loops can form when the two ends of a single digested DNA segment ligate to each other. Because the two free ends of one digested segment are also spatially close, self ligation events are not excluded from the set of identified proximate ligations. However, the length of a single DNA segment is limited because of the cutting site density of the endonuclease, and self-ligation can only produce hybrid fragments with two tags in a certain orientation. These features were used to remove self-ligation events from the proximate ligation set (Supplementary Methods). After removing the self-ligation and random ligation interactions, we identified 96 137 interacting loci (with a false discovery rate of 5.76%) from the starting total of 23 337 840. Although this may initially appear to be a severe filtering of the data, we note that most DNA fragments sequenced in genomic approaches are background noise (e.g. the sequenced tags under peaks represent <20% of most ChIP-seq data set). Therefore, proper biological interpretations of Hi-C data require stringent filtering.

Epigenetic and transcription factor data analysis

K562 open chromatin and DNA methylation data, as well as ChIP-seq data for histone modifications and TFBSs, were downloaded from the University of California at Santa Cruz (UCSC) genome browser database (Supplementary File S1). Nucleosome-depleted sites (open chromatin), sites of H3K4 mono-, di- and trimethylation, H3K9 trimethylation, H3K27 trimethylation, H3K27 acetylation, H3K36 trimethylation, H4K20 mono-methylation, H3K9 acetylation and TFBSs were identified using the web Bin-based Enrichment Level Threshold (wBELT) (for broad regions) or Bi-Asymmetric-Laplace Model (BALM) program (sharp peaks) (30,31). All data used in our analyses were highly reproducible, having >90% overlap when the top 40% of the sites from one replicate was compared with the set of all sites from the second replicate (ENCODE overlap rules).

RNA-seq analyses

K562 cells (3×10^6) were resuspended in 100 ml of Nucleofector solution V (Lonza) and transfected with 240 pmol of SMARTpool small interfering RNA (siRNA) targeting endogenous human GATA-1 or GATA-2 (Dharmacon/Thermo Fisher Scientific). siGenome nontargeting siRNA pool (Dharmacon/Thermo Fisher Scientific) was used as a control. Cells were transfected with the Nucleofector II (Lonza) using the T-016 program and were harvested 48-h post-transfection. The knockdown efficiency was quantitated by real-time polymerase chain reaction analysis of messenger RNA levels. The RNA libraries (two independent knockdowns for GATA1, two independent knockdowns for GATA2 and two control samples) were prepared in accordance with the Illumina RNA sample preparation protocol, using barcoded RNA-Seq adapters. Samples were sequenced using an Illumina Hi-Seq Genome Analyzer. All sequenced tags from each sample were aligned to the hg18 reference genome using the Burrows-Wheeler Aligner (BWA) aligner tool (32), and only uniquely mapped tags were used for the further analysis. To measure the differential gene expression between the GATA1 or GATA2 knockdown samples and the control samples, we applied the Cuffdiff program that uses the Cufflinks transcript

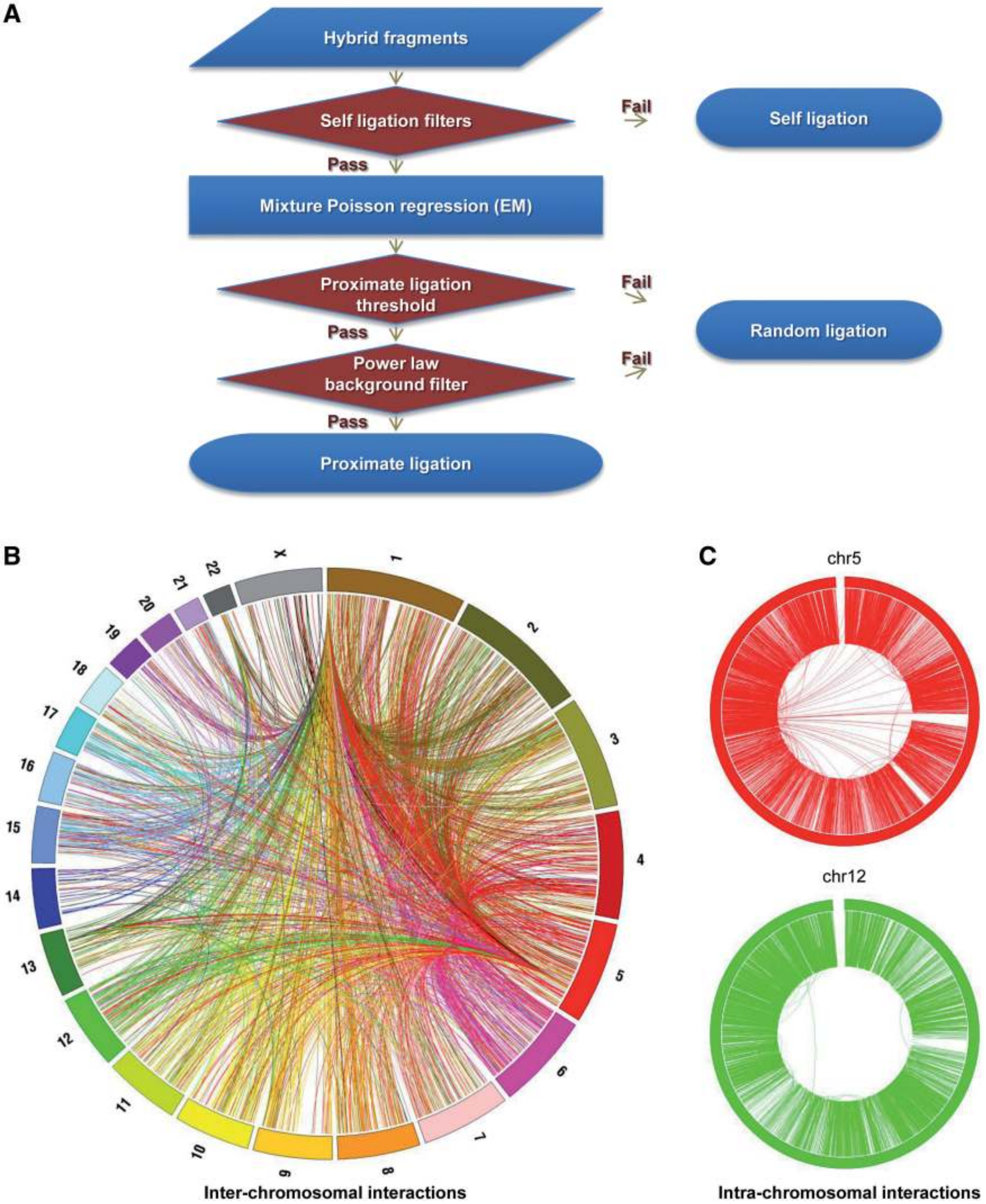

Figure 2. Hi-C analysis and genomic interactions in K562 cells. (**A**) Hi-C data analysis. Several steps were taken to select real interactions from the initial set of hybrid fragments. First, self-ligation was filtered based on its special properties (Supplementary Methods). Second, a MPRM was used to eliminate random loops. Next, the proximate ligation threshold was determined (Supplementary Figure 1C), and those interactions that pass the threshold were further filtered by a power-law decay background model. (**B**) Circos plot of the identified genome-wide inter-chromosomal interactions. The outer layer ring represents the human genome, with each chromosome labelled a different color. Each link between two genomic loci denotes one chromosomal interaction. (**C**) Circos plot of intra-chromosomal interactions for two representative chromosomes showing that thousands of interactions form within a relatively short distance on individual chromosomes; in fact, intra-chromosomal interactions (with the majority of the two loci being within 1 million bp) represent 95% of all interactions.

quantification engine to calculate gene and transcript expression levels in more than one condition and tests for significant expression differences (33). We then annotated the transcripts with UCSC RefSeq HG18. The expression value for each transcript was measured by a FRKM value (Fragments per kilobase of transcript per million mapped tags). After obtaining the FRKM values for each transcript in each sample, we performed a correlation analysis between the two biological replicates to measure the data reproducibility (Supplementary Figure S2).

RESULTS

Identifying interacting loci

Using the aforementioned MPRM and the power-law decay background (Supplementary Figure S1), 96 137 interacting loci with a FDR of 5.76% were selected from a total number of 23 337 840 hybrid fragments from K562 cells (Supplementary File S2); see the Hi-C data analysis section of Supplementary Methods for a more extensive description of the data modeling and analysis. Consistent with the previous study (25), most of the 96 137 interactions are intra-chromosomal (95%) and within 1 million bp distance (75%); see Supplementary Figure S3A. Circos plots illustrating the genome-wide inter-chromosomal interactions and two examples of intra-chromosomal interactions for two individual chromosomes are shown in Figure 2B and C. To demonstrate the applicability of the Hi-C analysis procedure, we also re-analyzed the Hi-C data from GM06990 cells. We identified 83 785 chromatin interactions, among which 82 683 are intra-chromosomal and only 1102 are inter-chromosomal. Similar to the analysis of K562 cells, >86% of the interactions in GM06990 cells are within 1 million bp distance, which indicates that the predominance of regional interactions is not a unique property of K562 cells (Supplementary Figure S3B).

K562 cells are derived from the bone marrow of a patient who had chronic myelogenous leukemia and are characterized by the presence of the Philadelphia Chromosome, a specific chromosomal abnormality that is the result of a reciprocal translocation between chromosome 9 and 22, creating a fusion gene in which the 'c-abl oncogene 1 (ABL1)' gene on chromosome 9 (region q34) is juxtaposed to a part of the 'breakpoint cluster region' gene on chromosome 22 (region q11). Because 9q34 and 22q11 are fused in K562 cells, interactions between these two regions would be labeled as inter-chromosomal, although they are actually intra-chromosomal in the K562 genome. In fact, 780 of the 4806 inter-chromosomal interactions in K562 cells are between 9q34 and 22q11, suggesting that a portion of the inter-chromosomal interactions identified in cancer cells may be intra-chromosomal for that particular genome. In addition to the t(9;22)(q34;q11) translocation, a previous study (34) described at least four other chromosomal translocations in K562 cells, namely, der(10)t(3;10)(p21.3;q23), der(18)t(1;18)(p32;q21), der(21)t(1;21)(q23;p11) and der(12)t(12;21)(p12;q21). Interestingly, there are only two interactions between 3p21.3 and 10q23, and no interactions were found between the other two pairs of chromosomal fusions. These analyses suggest that the 9;22 translocation may have different properties than the other translocations. In fact, the region corresponding to the 9;22 translocation is highly amplified. To investigate a possible correlation between looping and amplification, we first identified all amplified regions in K562 cells using Sole-search, an integrated ChIP-seq peak-calling program that not only identifies binding sites but also performs an analysis of amplified and deleted regions of the input genome (35,36). We identified 102 amplified regions in K562 cells, with 6166 long-range interactions found within the amplicons. Overlapping these regions with the set of interacting loci in K562 cells, we found that only 41 of the amplified regions contained mapped interacting loci (Supplementary Table S1). Thus, not all amplified regions are involved in long-range interactions. There was no relationship between the fold amplification and the number of loops. For example, the top highest-amplified regions of K562 cells (> 16.5 fold) had no long-range interactions, whereas the 94th ranked amplicon (3.79 fold) had 1053 long-range interactions. Interestingly, the amplified regions encompassing the breast cancer (BRC) and ABL1 genes had 1037 and 1060 long-range interactions, respectively, comprising >34% of all interactions associated with amplified genomic regions of K562 cells (see Supplementary Figure S4 for an illustration of the number of loops in the amplified regions of chr 22). To further investigate potential issues in our analyses owing to the existence of genomic rearrangement in K562 cells, we searched for interactions around other previously identified fusion sites in K562 cells. Twenty-five fusion sites detected by FusionMap software (Amgen Inc.) were tested, and none of these sites has interactions within a 20-kb distance.

Clustering interacting loci

To determine the relationship between epigenomic modifications and the identified genomic interactions, we mapped 9 histone modification marks (H3K4me1/2/3, H3K36me3, H4K20me1, H3K9ac, H3K27ac, H3K9me3 and H3K27me3) and regions identified as open chromatin using DNase hypersensitivity onto the set of identified interacting loci. In the initial clustering, we also included CTCF, which is an insulator protein known to influence chromatin structure (37). Using our peak-finding programs, wBELT and BALM (30,31), we first identified genomic regions that are enriched for each mark (Supplementary Table S2, Supplementary File S3). We observed that the enriched regions of H3K9me3, H3K27me3, H3K36me3 or H4K20me1 showed broad peak patterns (>1000 bp) in contrast to regions marked by H3K4me1/2/3, H3K9ac or H3K27ac, which showed sharp peak patterns over a relatively small region (~200 bp), which is in line with previous studies (1,4) (Supplementary Figure S5). We defined an interacting locus as associated with a certain epigenetic mark if it was within a broad peak region of H3K9me3, H3K27me3, H3K36me3, H4K20me1 or if a sharp peak of open chromatin, CTCF, H3K4me1/2/3, H3K9ac or H3K27ac was in the interacting DNA segment. The peak score from the wBELT output was used to define the intensity of the epigenetic status of that locus. Because many interacting loci may be associated with several peak scores from different marks, the intensities were standardized among different marks. We then performed hierarchical clustering (38) on the interacting loci using Cluster 3.0 software (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm, Stanford University, 1998–99) with Pearson correlation as the distance measurement. We used epigenetic information from only one end of the hybrid fragments to cluster the 96 137 interacting loci pairs into 12 groups (Figure 3A loci 1, Supplementary File S2). Interestingly, the second loci

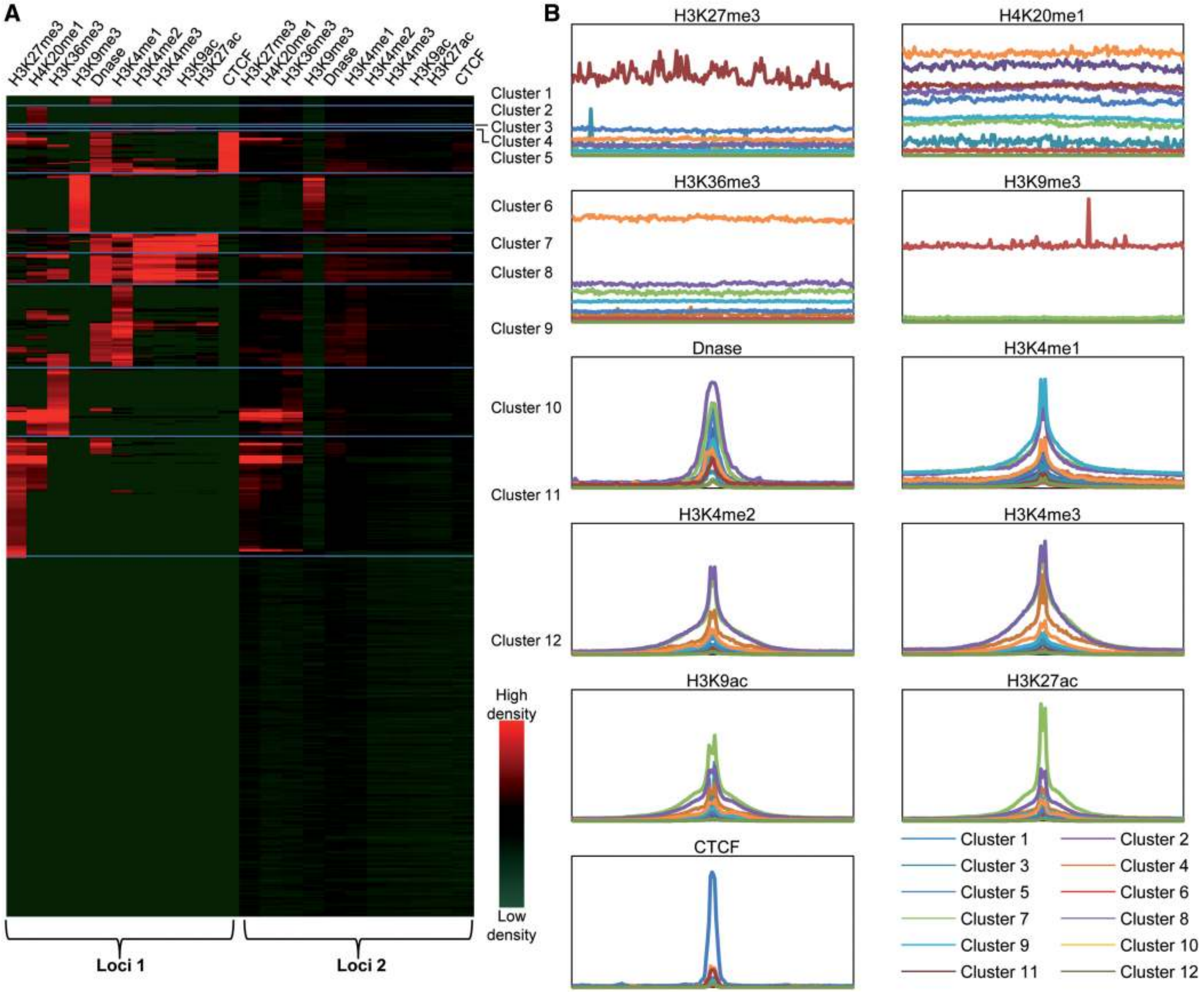

Figure 3. Clustering interacting loci. (**A**) Interacting loci were divided into 12 groups using hierarchical clustering based on the epigenetic status of loci 1. In many cases, the epigenetic status of interacting loci 2 shows a similar pattern to that of the partner loci 1. (**B**) Comparison of the distribution of the average intensities for each mark in each cluster. The plots of broad peak marks, including H3K9me3, H3K27me3, H4K20me1 and H3K36me3, are centered at the position of interacting loci, and the plots of sharp peak marks, including open chromatin, H3K4me1/2/3, H3K9ac and H3K27ac, are centered at the middle point of each peak. The *X*-axis is the relative distance to the central position of the analyzed mark, and the *Y*-axis is the tag density of that mark at each position.

often showed a very similar pattern of epigenetic status as the first loci (Figure 3A loci 2). All clusters can potentially interact with all the other clusters of chromatin; however, an interaction between two loci of the same cluster has a much higher formation rate (Chi-square test for every cluster, P value < 0.00001) (Supplementary Figure S6, Supplementary Table S3). For example, in cluster 6 (identified as having only H3K9me3 at loci 1), which constitutes 7.3% of the total interactions, 3307 out of 6980 (47.4%) of the loci 2 have a similar epigenetic status (H3K9me3 only) as loci 1. This formation rate of an interaction between two regions with H3K9me3 is much higher than the random formation rate of 7.3% ($P < 1$E-20, Chi-square test). Conversely, in cluster 6, only 163 out of 6980 (2.3%) of the loci 2 have an epigenetic status that is similar to that of cluster 9 (H3K4me1 and partially DNase). This formation rate of an interaction between cluster 6 and 9 is much lower than the random rate of 10.2% ($P < 1$E-20, Chi-square test). To compare the density of the individual epigenetic marks between clusters, we plotted the distribution of the average intensities for each mark in each cluster (Figure 3B). Each of the clusters, indeed, has distinct distribution of epigenetic marks. Clusters 7 and 8 appear similar in the clustergram (Figure 3A); however, cluster 7 has much higher level of H3K9ac and H3K27ac compared with cluster 8.

To examine how these different clusters of chromatin interactions may correlate with gene structure and transcriptional regulation, we determined the relative distance between TSS and interacting loci of each cluster (Figure 4,

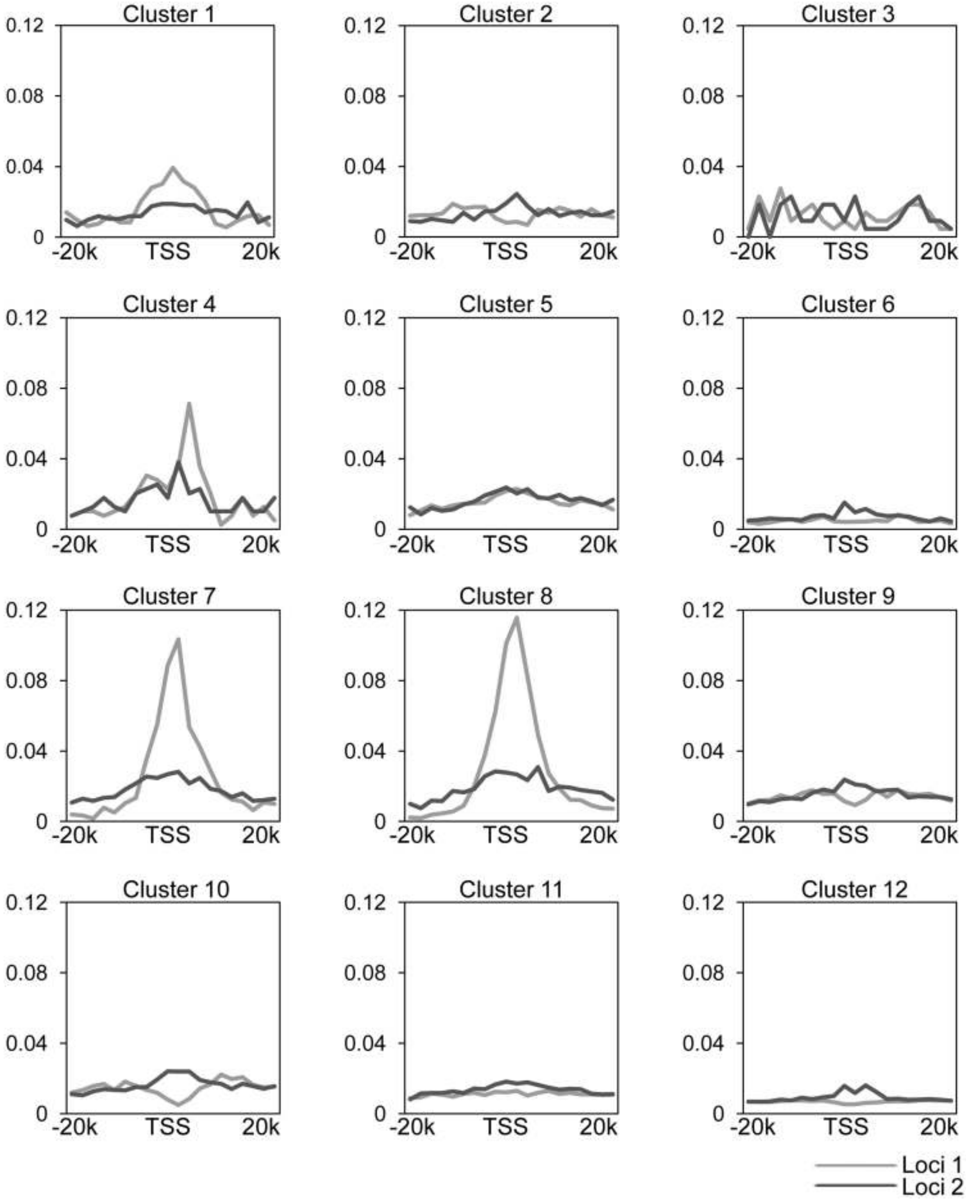

Figure 4. Relative distance of the interacting loci to a transcription start site. The *X*-axis is the relative distance between the nearest transcription start site (TSS) and the interacting loci, and the *Y*-axis is the occurrence frequency (i.e. the frequency that an interacting loci and the nearest TSS have a distance of x) in each cluster; 20 kb up- and downstream of the nearest TSS was analyzed, using a bin size of 2 kb.

Supplementary Figure S7). For example, (i) we found a higher promoter presence in the interacting loci of clusters 7 and 8, which is consistent with the abundance of H3K9ac and H3K27ac in these clusters. ii) cluster 10 loci are not at promoters, and loci 1 and loci 2 in this cluster are both enriched with the gene body marks H4K20me1 and H3K36me3, suggesting that cluster 10 may represent interactions between two regions that are both actively being transcribed. (iii) Cluster 9 had several interesting properties. First, loci 1 in this cluster are not directly at promoters (as determined by distance from a TSS and the lack of H3K9ac and H3K27ac marks) nor are they at active enhancers (they lack H3K27ac). However, these loci lack repressive marks but are marked by open chromatin and H3K4 monomethylation, suggesting that these regions are available for transcription factor binding. Further analyses of the TFs that bind to loci 9 are provided later in the text. (iv) Other clusters that were not enriched in promoter regions correspond to CTCF insulator-binding sites (cluster 5), had marks of epigenetic silencing (cluster 6, 11) or had none of the epigenetic modifications that was analyzed by the ENCODE Consortium (cluster 12).

The analysis presented earlier is representative of intra-chromosomal interactions because the majority of the interactions in K562 cells is in this category. To determine if the epigenetic patterns of the inter-chromosomal interactions are different from those of the intra-chromosomal interactions, we performed the clustering analysis using only the inter-chromosomal interactions. A large portion of these interactions is formed by regions that lack any of the epigenetic marks analyzed by the ENCODE Consortium, which is similar to cluster 12 in the combined analysis of intra- and inter-chromosomal interactions. Strikingly, the inter-chromosomal interactions consist of a significant portion of regions that is marked by H4K20me1 (active gene body regions), H3K36me3 (active gene body regions) and H3K27me3 (epigenetic silencing) (Supplementary Figure S8).

Associating TFBSs with the sets of clustered interacting chromatin loci

It is possible that the identified paired loci are brought together by interactions between a transcription factor bound to the loci 1 with another transcription factor bound to loci 2 in each paired set. CTCF has previously been shown to be highly correlated with looping (26). However, as shown in Figure 3, we found that a large portion of the chromatin interactions is not associated with a CTCF-binding site. In addition, CTCF can potentially interact with other types of chromatin (Supplementary Figure S6, cluster 5), suggesting that CTCF bound to loci 1 may interact with another TF at loci 2 to create a loop. Because certain transcription factor-mediated chromatin interactions have been associated with different types of gene regulation, it was possible that the different clusters may be regulated by distinct sets of transcription factors. The availability of a large set of ChIP-seq data from the ENCODE Consortium provided the opportunity to test this hypothesis.

We used the BALM program (31) to identify sets of genome-wide binding sites for 45 transcription factors using publicly available ChIP-seq data (Supplementary Table S4, Supplementary File S4). We found that a majority of the binding sites for the 45 factors is associated with DNase hypersensitive regions, which is an indication of nucleosome depletion (39) (Figure 5A). Furthermore, an inverse correlation between DNA methylation and DNase hypersensitivity was observed in these open chromatin regions (Supplementary Figure S9, Supplementary Methods). This result is consistent with our previous study (31) and many other studies, which have demonstrated that a majority of TF-DNA interactions requires the DNA fragment to be in an ‘open’ status (40), which is associated with nucleosome depletion (19), DNA hypomethylation (41) and specific side-chain modifications of histones (42). Open chromatin can reflect both promoter regions and other distal regulatory regions such as enhancer and repressor elements. To determine the preferential binding behavior of factors to promoters versus distal regulatory elements, we defined a promoter–distal ratio as the occurrence rate of a factor binding in a promoter region divided by the occurrence rate of that same factor binding in a non-promoter region. This analysis defined two distinct groups of transcription factors. For example, 14% of the non-promoter open chromatin regions contain GATA1-binding sites, whereas only 5% of the promoter open chromatin regions have GATA1-binding sites. Thus, the promoter–distal ratio is 0.36, which indicates that GATA1 preferably binds to non-promoter, open chromatin regions. Factors such as MYC and E2F4 are highly enriched in promoter, open chromatin regions, whereas factors such as GATA1 and GATA2 are over presented in non-promoter/open chromatin regions (Supplementary Figure S10).

We next correlated the identified TFBSs with the 12 sets of interacting loci to determine which sets of TFs are preferentially associated with each type of loci. We first defined a specific TF-associated interacting locus if a binding site was found in the interacting DNA segment. We then ranked the TFs according to the percentage of the binding sites that is associated with interacting loci (Table 1). Not surprisingly, CTCF, which is thought to be a major determinant of looping, ranked number 1 in this list (53% of CTCF sites were associated with interacting chromatin loci). However, many of the TFs showed a similar high percentage of binding sites associated with the identified interacting loci as did CTCF, suggesting that most TFs may be involved in looping. A hierarchical clustering (38) was performed to classify the factors (Figure 5A). Similar to the analysis of the epigenetic marks, the peak score from the BALM output was used to represent the binding affinity of the TF at that locus, and the scores were standardized among different TFs. The clustering result showed several major groups of transcription factors with distinct binding preferences for loci with a distinct epigenetic status. For example, one group of factors includes CTCF and RAD21, which can bind to insulators that are open (i.e. marked by DNAse hypersensitivity) (cluster 5). The majority of the transcription factors binds specifically to loci marked by H3K9ac and H3K27ac and which are likely to be active promoters (cluster 7 and 8) (1,4). However, three site-specific transcription factors (c-Jun, GATA1 and GATA2) and three chromatin regulators (BRG1, INI1 and SIRT6) bind specifically to loci in cluster 9 that are open chromatin marked by H3K4 mono-methylation but not by promoter (H3K9Ac) or active enhancer (H3K27Ac) modifications.

To find the concurrence of these proteins at the two ends of the interacting loci, we applied the Apriori algorithm (29), which is a widely used data-mining method for searching for association rules in large data sets of transactions (Supplementary Methods). The Apriori algorithm revealed a protein interaction network through DNA looping (Figure 5B). First, our analysis showed a high concurrence of CTCF and RAD21 in the ends of interacting loci, which is consistent with their similar binding preference and with previous reports (43). Second, proteins in the polymerase (POL) III machinery, including POL3, TFIIIC, BRF1 and BDP1, also linked together through long-distance DNA looping. Third, E2F4 and RNA polymerase were highly linked, consistent with previous studies (13). Finally, we note that c-Jun, GATA1, GATA2, INI1 and BRG1 are closely linked. Interestingly, nearly all of the factor nodes have loops connected to themselves, which might be owing to the cross linking of the distant DNA elements to the protein during the ChIP-seq procedure. This may also explain the situation that some highly enriched peaks detected from a TF ChIP-seq experiments lack a consensus motif for that TF (11).

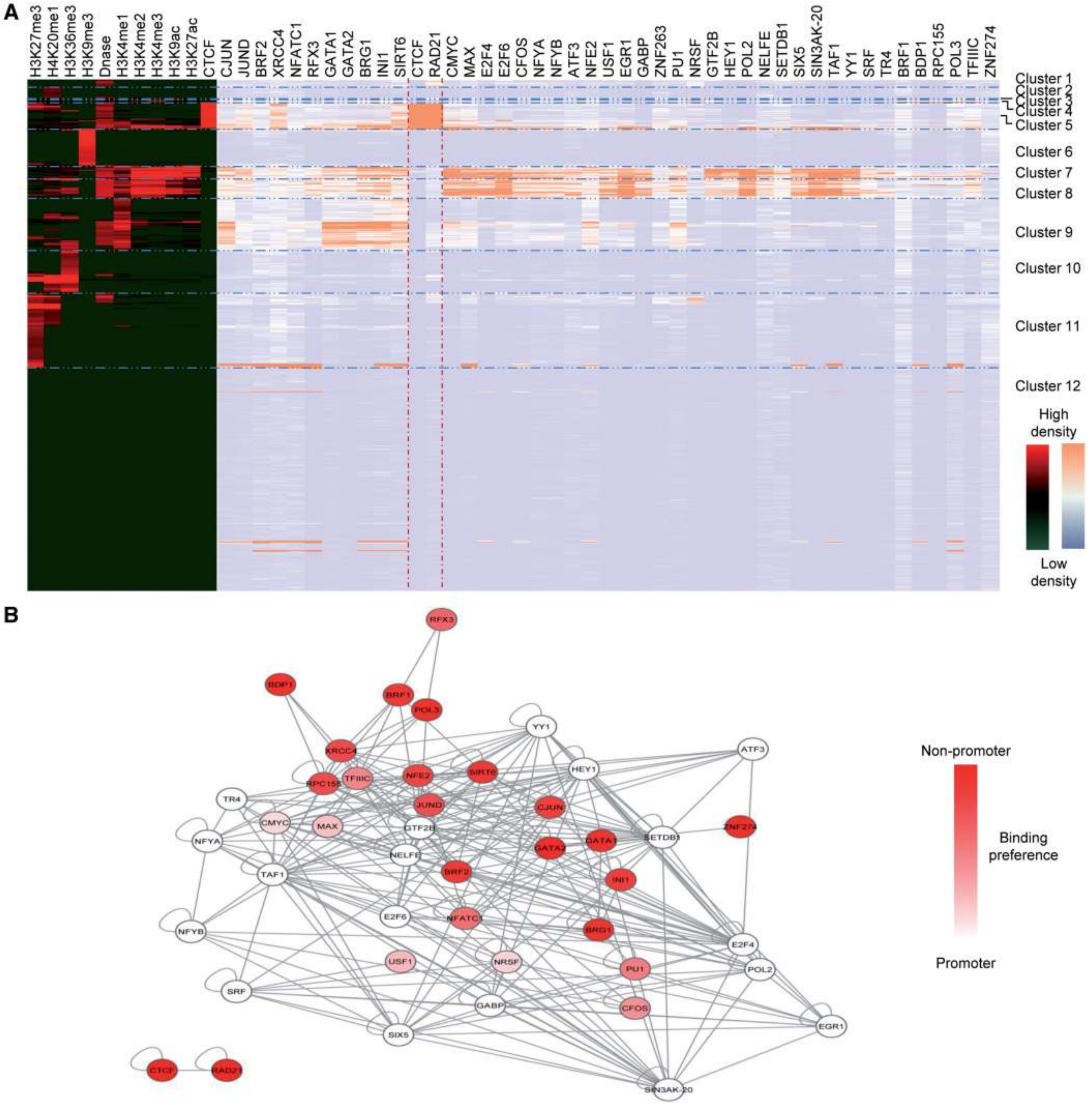

Figure 5. Interacting loci are bound by a network of transcription factors. (**A**) Hierarchical clustering suggests the existence of at least three discriminating groups of factors, namely, insulator-binding factors such as CTCF and RAD21 (bound to regions of open chromatin), promoter-binding factors such as c-Myc and E2F4 (bound to regions marked by open chromatin, H3K4 methylation and H3K9/27 acetylation) and non-promoter-binding factors such as GATA1 and GATA2 (bound to regions marked by open chromatin and H3K4 methylation). Promoter-binding TFs are mainly present in cluster 7 and 8, whereas non-promoter-binding factors showed less frequency in cluster 7 and 8 but higher frequency in cluster 9. This indicates that different transcription factors preferentially bind to regions with distinct epigenetic status. The left color bar corresponds to the color scale for the epigenetic marks, and the right color bar corresponds to the color scale for the TF-binding sites. (**B**) TF interaction network revealed by the Apriori algorithm. The color of each node represents the binding preference of the TF showed in Figure S9.

Correlating gene expression with different sets of interacting loci

To examine the influence of looping on gene expression, we assigned each interacting locus to a known gene if the locus is located within the gene body or is <10 kb from the transcription start site of the gene. The expression levels of the genes in loci 1 and paired loci 2 for each cluster versus a whole genome gene expression profile were then plotted (Figure 6). Interestingly, the interacting loci do not only

Table 1. Percentage of binding site associated with chromatin interaction

TF	Percentage (%)	TF	Percentage (%)	TF	Percentage (%)	TF	Percentage (%)	TF	Percentage (%)
CTCF	53.10	CMYC	45.47	JUND	41.98	NFYA	38.99	HEY1	34.98
USF1	50.19	CJUN	44.66	SRF	41.84	NFE2	38.84	RFX3	34.78
RAD21	49.85	GATA2	44.44	YY1	41.70	INI1	38.62	SETDB1	33.97
SIX5	48.26	MAX	44.11	ATF3	41.17	XRCC4	38.60	BDP1	32.99
PU1	47.72	GATA1	44.02	TR4	40.45	RPC155	38.41	ZNF274	32.51
CFOS	46.82	E2F4	43.99	NFYB	40.18	TFIIIC	37.50	POL3	32.08
NRSF	46.56	EGR1	43.75	SIN3AK-20	39.31	GTF2B	37.10	NFATC1	32.04
BRG1	45.80	GABP	43.71	E2F6	39.24	NELFE	36.85	BRF2	31.67
ZNF263	45.56	TAF1	43.48	POL2	39.04	SIRT6	36.44	BRF1	30.77

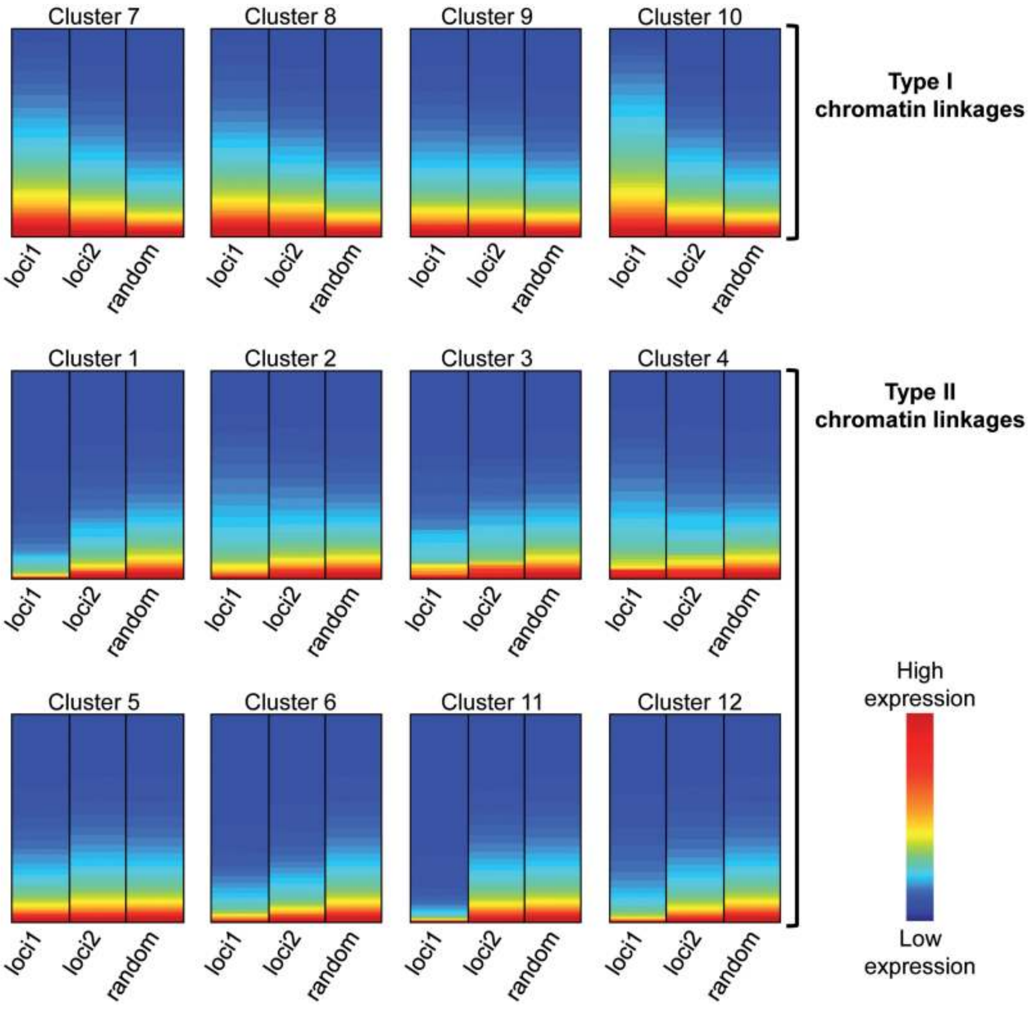

Figure 6. Gene expression analyses reveal two types of chromatin linkages. Gene expression heatmaps show the expression of genes in loci 1 and 2 versus a random set of genes in each cluster. This analysis revealed two types of chromatin linkages. Type I: genes associated with both interacting loci in each pair have a higher gene expression level than a random set of genes; the interacting genes are likely co-activated (permissive chromatin linkages). Type II: genes associated with both interacting loci in each pair have lower expression levels than a random set of genes (non-permissive chromatin linkages).

possess a similar epigenetic status, the expression of the associated genes is also co-regulated. We defined four co-active clusters (24.8% of total interactions) as Type I linkages and eight co-repressive clusters (75.2% of total interactions) as Type II linkages. Specifically, clusters 7, 8, 9 and 10 were composed of paired loci in which the nearest genes to both loci were more active than a set of randomly selected genes, whereas the rest of the clusters (1–6, 11–12) were composed of paired loci in which the nearest genes to both loci showed lower expression than a randomly selected set of genes. Distinct gene expression patterns were observed in clusters having a different epigenetic status. Type I chromatin linkages are associated with active marks, such as H3K4me3, H3K9ac, H3K27ac, H3K36me3 and H4K20me1, whereas the majority of Type II chromatin linkages showed low level of these marks. Cluster 5, 6, 11 and 12, which composed >96% of Type II chromatin linkages, are bound by insulator protein (cluster 5), by the heterochromatin mark H3K9me3 (cluster 6), by the repressive histone modification H3K27me3 (cluster 11) or have none of the tested marks (cluster 12). Ingenuity Pathway Analysis (www.ingenuity.com) showed that Type I chromatin linkages were specifically associated with genes involved in cell cycle and chronic myeloid leukemia signaling, which is appropriate for genes actively expressed in K562 myeloid leukemia cells (Supplementary Figure S11A and B).

Repetitive elements have been shown to play an important role in chromatin organization. Therefore, we examined the relationship between repetitive elements with different types of interactions. A few associations are notable, including depletion of Medium Reiteration frequency interspersed repetitive elements 1 (MER1) in cluster 6 (which is high in H3K9me3) and cluster 12 (no epigenetic marks) and depletion of Mammalian apparent LTR Retrotransposon (MaLR) in cluster 10 (which is high in H3K36me3). Other types of repetitive elements are evenly distributed in the different clusters (L1, L2, Alu element and mammalian interspersed repeats) or not present in any cluster (retrotransposable element, AcHobo, Tip100, chicken repeat 1, MER2, endogenous retrovirus 1 and endogenous retrovirus like); see (Supplementary Figure S12).

Combining Hi-C analyses, ChIP-seq and RNA-seq to classify GATA target genes

One of the most surprising findings in our analyses is the unique relationship of three site-specific factors, GATA1, GATA2 and c-Jun, with the clusters of interacting loci. Another factor that co-localizes with the GATA factors in cluster 9 is BRG1, which has been implicated in the formation of a GATA-associated DNA loop structure (44). These results suggested that looping may be involved in regulating at least some GATA target genes. To investigate this possibility, we performed RNA-seq analyses before and after knockdown of GATA1 or GATA2 in K562 cells. We sequenced two biological replicates each of RNA samples from cells treated with siRNAs to GATA1 or GATA2 and two control cell populations; see Supplementary Table S5. A correlation analysis of the two replicates for each sample showed high reproducibility (Supplementary Figure S2). Therefore, we combined the two replicates for each sample to determine the gene expression levels using a Fragments Per Kilobase of exon per Million fragments mapped (FPKM) value (see MATERIAL AND METHODS and Supplementary Files S5, S6). Not surprisingly, profound changes of gene expression occurred in K562 cells after knockdown of GATA1 or GATA2. We found that the expression of 7,497 genes (21.0% of Refseq genes) was altered upon knockdown of GATA1. Of these, 2769 (36.9%) genes were downregulated and 4728 (63.1%) genes were upregulated. Similarly, the expression of 2512 (7.0% of Refseq genes) genes was altered on knockdown of GATA2. Of these, 1183 (47.1%) genes were down regulated and 1329 (52.9%) genes were upregulated.

One of the major problems in studying the function of transcription factors is understanding which genes are directly regulated by a factor and which genes are indirectly regulated by a factor because they are in a downstream signaling pathway. The typical approach is to assign direct targets as those deregulated genes that have a nearby TF-binding site (as determined by ChIP-seq) and indirect targets as those deregulated genes that are not near to a ChIP-seq peak. However, by incorporating Hi-C data, we can now identify genes that are far from a TF-binding site on the linear genome but closely linked in 3-dimensional space. Using BALM software (31), we identified 10 828 GATA1-binding sites and 8284 GATA2-binding sites in the K562 ChIP-seq data. We then classified the genes showing altered expression in the GATA1 or GATA2 knockdown cells into three categories: genes directly regulated by a GATA factor binding near the promoter of that gene, genes directly regulated by a GATA factor binding to an interacting loci and 'downstream' genes regulated indirectly by reduction of the GATA factor (Figure 7). Overall, 48% for GATA1 and 39.0% for GATA2 of the genes showing altered regulation can be linked directly or indirectly (through DNA looping) to a GATA-binding site. Among the 7497 genes that are affected by GATA1 gene knockdown, 2684 (35.8%) genes have at least one GATA1-binding site within the gene body or within 10 kb up- or downstream of the transcribed region. Although not directly bound by GATA1 protein, an additional 924 (12.3%) genes are regulated via DNA looping, with GATA1 binding to an interacting locus (i.e. within a pair of interacting loci, one locus has at least one GATA1 site and the other locus in the paired set is near the regulated gene). Similarly, among the 2512 genes that are affected by GATA2 gene knockdown, 749 (29.8%) genes have at least one GATA2-binding site within the gene body or within 10 kb up- or downstream of the gene and an additional 233 (9.28%) genes are regulated via DNA looping, with GATA2 binding to an interacting locus. Thus, by including the information about DNA loops, the set of genes directly regulated by GATA1 and GATA2 in K562 cells could be greatly expanded. We further performed Gene Ontology (GO) functional analysis using Ingenuity Pathway Analysis software (Ingenuity Systems, Inc., www.ingenuity.com) on these different sets of GATA-regulated genes (Supplementary Figure S13). We found that overall the genes directly and indirectly regulated by the GATA factors are enriched in similar functional categories. However, both GATA1 and GATA2 impact the expression of cancer-related genes more through DNA looping rather than through proximal regulation. This indicates the importance of studying chromatin organization in understanding disease development and progression.

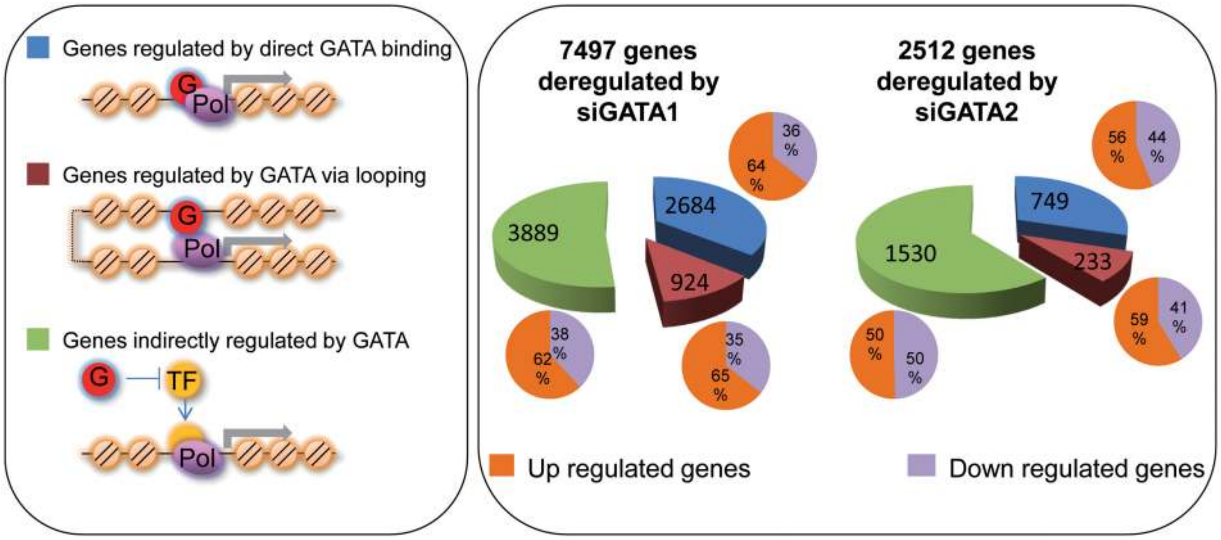

Figure 7. Gene expression changes induced by GATA1 and GATA2 knockdown. RNA-seq was performed using control cells and cells treated with siRNAs to either GATA1 or GATA2, and genes whose expression was altered upon knockdown were identified. By intersection of the list of deregulated genes with genes having nearby bound GATA factors, direct target genes were identified (blue pie segments). A second set of direct target genes were identified by intersection of the list of deregulated genes with genes linked to a bound GATA factor via chromatin looping (maroon pie segments). All other genes were classified as indirectly regulated genes (green pie segments). The colors in the pie chart correspond to these different categories of GATA-regulated genes, as labelled in left panel. For each category, the percentage of genes that were upregulated versus downregulated in the knockdown cells is shown by the purple and orange closed pie graphs. (G: GATA factor; Pol: RNA Polymerase II, TF: a transcription factor whose expression is regulated by a GATA factor).

DISCUSSION

In this study, we present an integrated analytical method to identify and characterize different types of chromatin linkages. By integrating multiple genome-wide data sets from K562 cells, including Hi-C data, ChIP-seq for 45 transcription factors and 9 histone modifications, DNase hypersensitivity assays and RNA-seq data, we have identified 12 distinct sets of chromatin linkages comprising a total of 96 137 sets of two spatially separated interacting loci, shown that each cluster has distinct epigenetic markings, is composed of paired sets of genes with distinct expression patterns and is differently correlated with TFBSs. To validate the biological importance of the identified interacting loci, we investigated genes regulated by GATA1 and GATA2. Our analyses are consistent with the hypothesis that the GATA factors regulate a subset of target genes via looping.

The Hi-C data analysis demonstrated that the regions involved in creating loops between interacting loci were preferentially in or near open chromatin, suggesting that bound transcription factors may play a crucial role in creating the genomic interactome. Many of the transcription factors we analyzed bind near promoter regions (as defined by specifically modified histones), suggesting that a factor bound near a start site may interact with another factor bound to a distal region of open chromatin. In support of this concept, many of the paired interactions showed evidence for a promoter region at one locus but not at the other. However, our analyses also identified a subset of interacting loci (cluster 9) that has unique properties. These loci show evidence of open chromatin and H3K4me1 but do not resemble active promoters or enhancers. Analysis of the ChIP-seq data identified a set of three transcription factors (GATA1, GATA2 and c-Jun) and three chromatin modifiers (SIRT6, BRG1 and INI1) that were specifically enriched at sites having only these two chromatin marks. BRG1 (also called SMARCA4) and INI1 (also called SMARCB1) are both components of the SWI/SNF chromatin-remodeling complex. The presence of SIRT6, a histone deacetylase, perhaps explains the absence of H3K27Ac at the regions of open chromatin bound by this complex. Interestingly, analysis of the 45 ChIP-seq data sets using the Apriori algorithm also showed that GATA1, GATA2, c-Jun, BRG1 and INI1 were closely linked. Two of the factors, GATA1 and GATA2, are members of the same gene family, have several similar DNA-binding motifs and bind to many of the same sites in K562 cells (17). Also, we have previously shown that GATA2 co-localizes and regulates gene expression in concert with c-Jun in human endothelial cells (45), providing support that GATA factors cooperate with c-Jun to regulate expression of genes in cluster 9. Finally, BRG1 is reported as a cofactor of GATA1 (46,47). We, and others, have previously shown that BRG1 functions cooperatively with GATA1 at certain genes through chromatin loop structure (44,48). However, the overall involvement of GATA factors in chromatin looping has not been previously investigated. Thus, taken together, the unbiased clustering of Hi-C interacting loci, the unbiased clustering of ChIP-seq data and the correlation of transcription factor binding with histone modifications, in combination with previous reports of linkage between GATAs, c-Jun, and BRG1, suggest that a subset of GATA targets may be regulated via interacting loci. Accordingly, we tested this prediction by introduction of siRNAs to GATA1 or

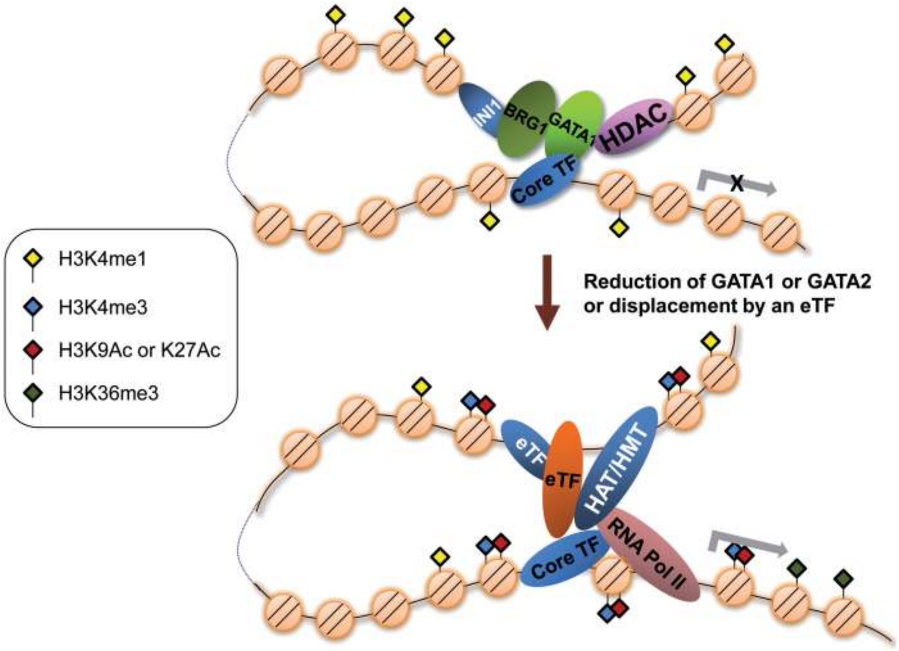

Figure 8. Schematic model of GATA-regulated chromatin linkages. Dynamic chromatin interactions form globule structures, which may function to initiate or stabilize three-dimensional gene regulatory structures. Multiple CREs, including enhancers and promoters, are bridged by different groups of transcription factors (TFs) and mediators under different conditions. In the example shown, a pair of interacting loci from cluster 9 is represented. A loop is formed between a promoter and a distal region via interactions of GATA1, BRG1, INI1 and SIRT6 (a histone deacetylase), all bound to a region having H3K4me1 but no marks of an active enhancer or promoter; the nearby gene is repressed. On loss of GATA1 (via reduction of levels of the protein by treatment with siRNAs or on normal physiological changes or owing to displacement of GATA1 by another DNA-binding factor), a different set of enhancer-binding factors are recruited to the distal open chromatin region (which now gains the H3K27Ac mark), the loop changes from a repressive to an activating structure, active histone modifications are placed on the promoter region, RNA Polymerase II is recruited and begins transcription, and the transcribed region gains H3K36me3. We note that GATA1 can also activate transcription, and thus other functions of GATA1-mediated loops can be envisioned.

GATA2 followed by RNA-seq analyses. We found that 7497 genes were deregulated on knockdown of GATA1 of which ~36% were regulated by a nearby bound GATA1, 12% by a GATA1 bound to an interacting loci and 52% were indirectly regulated. Similarly, 2512 genes were deregulated on knockdown of GATA2, of which ~30% were regulated by a nearby bound GATA2, 9% by a GATA2 bound to an interacting loci and 61% were indirectly regulated. Our experimental validations support the concept that GATA factors indeed regulate gene expression through interaction with distal loci. We note that GATA1 and GATA2 bind to many of the same sites in K562 cells (17) and thus most likely regulate some of the same genes. These factors bind independently (and not at the same time) to GATA sites; knockdown of GATA1 would still allow binding of GATA2 (and vice versa). Therefore, it is likely that more robust changes in gene expression may have been observed if both factors could be knocked down at the same time.

GATA factors have previously been shown to be both activators and repressors, and our data demonstrate that genes regulated by looping can be either upregulated or downregulated on loss of GATA1 or GATA2. One possible model by which loss of GATA1 could result in activation of a distal gene is shown in Figure 8. In this model, a loop is shown between a promoter and a distal region that is created by interactions of GATA1, BRG1, INI1 and SIRT6 (a histone deacetylase), all bound to a region having H3K4me1 but no marks of an active enhancer or promoter and the nearby gene is repressed (consistent with the histone marks enriched in cluster 9). On reduction of GATA1 levels, a different set of enhancer-binding factors is recruited to the distal open chromatin region, the loop changes from a repressive to an activating structure, and transcription is initiated. However, we recognize that there are many other mechanistic possibilities for how the GATAs and BRG1 could regulate transcription, such as the complex serving as an activator. For example, a recent study showed that 58% of the GATA1 sites identified in Cluster of Differentiation (CD) 36+ erythrocyte precursor cells were also bound by BRG1 (49); the authors suggest that recruitment of BRG1 by GATA1 allows binding of T-cell acute lymphocytic leukemia protein 1 (TAL1) to the enhancer region and results in transcriptional activation of certain GATA1 target genes.

In conclusion, we demonstrate that when combined with in-depth analysis of histone modifications and transcription factor binding, Hi-C data can serve as a powerful tool for exploring the complex underlying mechanisms of chromatin organization. Previous studies have shown that environmental changes such as estrogen treatment can cause intensive looping and de-looping events (50–52), providing evidence that chromatin-bound TFs may induce dynamic changes in genome organization. Our analyses show that most TFs have thousands of binding sites that are associated with chromatin interaction sites and that distinct clusters of interacting loci can be bound by subsets of TFs provide genome-wide evidence in support of the concept that a set of TFs may create distinct types of chromatin linkages, where co-regulated genes are brought into close proximity from different chromosomal locations. We also note that our identification of a GATA-enriched set of physically interacting loci was obtained using unbiased clustering of Hi-C and ChIP-seq data from K562 cells. Given the documented role of GATA factors in controlling hematopoiesis and erythroid differentiation (53–55), the identification of a GATA-enriched set of chromatin linkages provides evidence that the clustering analysis can identify master regulators of the transcriptome. With the rapid development of sequencing technologies, Hi-C data collection is becoming more readily available for a variety of cell types. As other cell type-specific Hi-C data are obtained and the set of factors analyzed by ChIP-seq increases, our analyses can be repeated using data from these additional cell types. We predict that clusters defined by open chromatin and specific histone marks (such as cluster 9 in K562 cells) will show co-association with different sets of transcription factors in different cell types. We suggest that an integrative analysis of the Hi-C data with histone modifications and transcription factor ChIP-seq data sets will identify different biologically-relevant clusters in different cell types and help to identify cell type-specific master regulators.

DATA ACCESS

All data are publicly available via the UCSC Genome Preview Browser (http://genome-preview.ucsc.edu/) except for the RNA-seq data, which have been submitted to GEO, GSE32213.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figures 1–13, Supplementary Methods and Supplementary Files 1–6.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Fournier-Viger for providing the source code and the binary of the latest version of the SPMF software. The authors thank the ENCODE Consortium and the Data Coordination Center at UCSC for providing access to the DNA methylation, DNAse-seq and ChIP-seq data; the RNA-seq libraries were sequenced at the USC Epigenome Center Next Generation Sequencing Core.

FUNDING

Public Health Service National Institutes of Health (NIH) [CA45250 and 1U54HG004558 to P.J.F.]; Public Health Service NIH [DK68634 to E.H.B.]; HHMI Med Into Grad scholarship (to X.L.) and by funds from PhRMA foundation and the Department of Biomedical Informatics at the Ohio State University (to V.X.J.). Funding for open access charge: PhRMA foundation; the Department of Biomedical Informatics at the Ohio State University (to V.X.J.).

Conflict of interest statement. None declared.

REFERENCES

1. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.
2. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA*, **107**, 21931–21936.
3. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
4. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
5. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
6. Lee,T.I., Jenner,R.G., Boyer,L.A., Guenther,M.G., Levine,S.S., Kumar,R.M., Chevalier,B., Johnstone,S.E., Cole,M.F., Isono,K. *et al.* (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, **125**, 301–313.
7. Roh,T.Y., Cuddapah,S., Cui,K. and Zhao,K. (2006) The genomic landscape of histone modifications in human T cells. *Proc. Natl Acad. Sci. USA*, **103**, 15782–15787.
8. Rosenfeld,J.A., Wang,Z., Schones,D.E., Zhao,K., DeSalle,R. and Zhang,M.Q. (2009) Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, **10**, 143.
9. Bannister,A.J., Schneider,R., Myers,F.A., Thorne,A.W., Crane-Robinson,C. and Kouzarides,T. (2005) Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.*, **280**, 17732–17736.
10. Rosenbloom,K.R., Dreszer,T.R., Pheasant,M., Barber,G.P., Meyer,L.R., Pohl,A., Raney,B.J., Wang,T., Hinrichs,A.S., Zweig,A.S. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.
11. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
12. Adhikary,S. and Eilers,M. (2005) Transcriptional regulation and transformation by Myc proteins. *Nat. Rev. Mol. Cell Biol.*, **6**, 635–645.
13. Bieda,M., Xu,X., Singer,M.A., Green,R. and Farnham,P.J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome. Res.*, **16**, 595–605.
14. Xu,X., Bieda,M., Jin,V.X., Rabinovich,A., Oberley,M.J., Green,R. and Farnham,P.J. (2007) A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome. Res.*, **17**, 1550–1561.
15. Rabinovich,A., Jin,V.X., Rabinovich,R., Xu,X. and Farnham,P.J. (2008) E2F *in vivo* binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome. Res.*, **18**, 1763–1777.
16. Hatzis,P., van der Flier,L.G., van Driel,M.A., Guryev,V., Nielsen,F., Denissov,S., Nijman,I.J., Koster,J., Santo,E.E., Welboren,W. *et al.* (2008) Genome-wide pattern of TCF7L2/ TCF4 chromatin occupancy in colorectal cancer cells. *Mol. Cell Biol.*, **28**, 2732–2744.

17. Fujiwara,T., O'Geen,H., Keles,S., Blahnik,K., Linnemann,A.K., Kang,Y.A., Choi,K., Farnham,P.J. and Bresnick,E.H. (2009) Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol. Cell*, **36**, 667–681.
18. Carroll,J.S., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.
19. Gondor,A. and Ohlsson,R. (2009) Chromosome crosstalk in three dimensions. *Nature*, **461**, 212–217.
20. Ling,J.Q., Li,T., Hu,J.F., Vu,T.H., Chen,H.L., Qiu,X.W., Cherry,A.M. and Hoffman,A.R. (2006) CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science*, **312**, 269–272.
21. Osborne,C.S. (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol.*, **5**, e192.
22. Sutherland,H. and Bickmore,W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.
23. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
24. van Steensel,B. and Dekker,J. (2010) Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.*, **28**, 1089–1095.
25. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
26. Botta,M., Haider,S., Leung,I.X., Lio,P. and Mozziconacci,J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.*, **6**, 426.
27. Wedel,M., Desarbo,W.S., Bult,J.R. and Ramaswamy,V. (1993) A latent class Poisson regression model for heterogeneous count data. *J. Appl. Econom.*, **8**, 397–411.
28. Yang,M. and Lai,C. (2005) Mixture Poisson regression models for heterogeneous count data based on latent and fuzzy class analysis. *Soft. Comput.*, **9**, 512–524.
29. Agrawal,R. and Srikant,R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. *The 20th International Conference on Very Large Data Bases*. Morgan Kaufmann, Los Altos, CA, pp. 487–499.
30. Lan,X., Bonneville,R., Apostolos,J., Wu,W. and Jin,V.X. (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics*, **27**, 428–430.
31. Lan,X., Adams,C., Landers,M., Dudas,M., Krissinger,D., Marnellos,G., Bonneville,R., Xu,M., Wang,J., Huang,T.H. *et al.* (2011) High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology. *PLoS One*, **6**, e22226.
32. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
33. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
34. Gribble,S.M., Roberts,I., Grace,C., Andrews,K.M., Green,A.R. and Nacheva,E.P. (2000) Cytogenetics of the chronic myeloid leukemia-derived cell line K562: karyotype clarification by multicolor fluorescence in situ hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet. Cytogenet.*, **118**, 1–8.
35. Blahnik,K.R., Dou,L., O'Geen,H., McPhillips,T., Xu,X., Cao,A.R., Iyengar,S., Nicolet,C.M., Ludascher,B., Korf,I. *et al.* (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, **38**, e13.
36. Blahnik,K.R., Dou,L., Echipare,L., Iyengar,S., O'Geen,H., Sanchez,E., Zhao,Y., Marra,M.A., Hirst,M., Costello,J.F. *et al.* (2011) Characterization of the contradictory chromatin signatures at the 3' exons of zinc finger genes. *PLoS. One*, **6**, e17121.
37. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
38. Ward,J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*, **58**, 236–244.
39. Audit,B., Zaghloul,L., Vaillant,C., Chevereau,G., d'Aubenton-Carafa,Y., Thermes,C. and Arneodo,A. (2009) Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res.*, **37**, 6064–6075.
40. Ohyama,T. (2005) *DNA conformation and transcription*. Landes Bioscience, Springer Science Business Media, Georgetown, TX, New York, NY.
41. Vanselow,J., Pohland,R. and Furbass,R. (2005) Promoter-2-derived Cyp19 expression in bovine granulosa cells coincides with gene-specific DNA hypo-methylation. *Mol. Cell Endocrinol.*, **233**, 57–64.
42. Bessler,J.B., Andersen,E.C. and Villeneuve,A.M. (2010) Differential localization and independent acquisition of the H3K9me2 and H3K9me3 chromatin modifications in the *Caenorhabditis elegans* adult germ line. *PLoS. Genet.*, **6**, e1000830.
43. Wada,Y., Ohta,Y., Xu,M., Tsutsumi,S., Minami,T., Inoue,K., Komura,D., Kitakami,J., Oshida,N., Papantonis,A. *et al.* (2009) A wave of nascent transcription on activated human genes. *Proc. Natl Acad. Sci. USA*, **106**, 18357–18361.
44. Kim,S.I., Bultman,S.J., Kiefer,C.M., Dean,A. and Bresnick,E.H. (2009) BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc. Natl Acad. Sci. USA*, **106**, 2259–2264.
45. Linnemann,A.K., O'Geen,H., Keles,S., Farnham,P.J. and Bresnick,E.H. (2011) Genetic framework for GATA factor function in vascular biology. *Proc. Natl Acad. Sci. USA*, **108**, 13641–13646.
46. Kim,S.I., Bultman,S.J., Jing,H., Blobel,G.A. and Bresnick,E.H. (2007) Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. *Mol. Cell Biol.*, **27**, 4551–4565.
47. Im,H., Grass,J.A., Johnson,K.D., Kim,S.I., Boyer,M.E., Imbalzano,A.N., Bieker,J.J. and Bresnick,E.H. (2005) Chromatin domain activation via GATA-1 utilization of a small subset of dispersed GATA motifs within a broad chromosomal region. *Proc. Natl Acad. Sci. USA*, **102**, 17065–17070.
48. Kim,S.I., Bresnick,E.H. and Bultman,S.J. (2009) BRG1 directly regulates nucleosome structure and chromatin looping of the alpha globin locus to activate transcription. *Nucleic Acids Res*, **37**, 6019–6027.
49. Hu,G., Schones,D.E., Cui,K., Ybarra,R., Northrup,D., Tang,Q., Gattinoni,L., Restifo,N.P., Huang,S. and Zhao,K. (2011) Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome. Res.*, **21**, 1650–1658.
50. Hsu,P.Y., Hsu,H.K., Singer,G.A., Yan,P.S., Rodriguez,B.A., Liu,J.C., Weng,Y.I., Deatherage,D.E., Chen,Z., Pereira,J.S. *et al.* (2010) Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping. *Genome. Res.*, **20**, 733–744.
51. Krebs,A. and Tora,L. (2009) Keys to open chromatin for transcription activation: FACT and Asf1. *Mol. Cell*, **34**, 397–399.
52. Gaulton,K.J., Nammo,T., Pasquali,L., Simon,J.M., Giresi,P.G., Fogarty,M.P., Panhuis,T.M., Mieczkowski,P., Secchi,A., Bosco,D. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.
53. Pevny,L., Simon,M.C., Robertson,E., Klein,W.H., Tsai,S.F., D'Agati,V., Orkin,S.H. and Costantini,F. (1991) Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*, **349**, 257–260.
54. Bresnick,E.H., Katsumura,K.R., Lee,H.Y., Johnson,K.D. and Perkins,A.S. (2012) Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Res.*, **40**, 5819–5831.
55. Tsai,F.Y., Keller,G., Kuo,F.C., Weiss,M., Chen,J., Rosenblatt,M., Alt,F.W. and Orkin,S.H. (1994) An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*, **371**, 221–226.