Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation

Guoliang Li, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, Yufen Goh, Joanne Lim, Jingyao Zhang, Hui Shan Sim, Su Qin Peh, Fabianus Hendriyan Mulawadi, Chin Thing Ong, Yuriy L. Orlov, Shuzhen Hong, Zhizhuo Zhang, Steve Landt, Debasish Raha, Ghia Euskirchen, Chia-Lin Wei, Wei Hong Ge, Huaien Wang, Carrie Davis, Katherine I. Fisher-Aylor, Ali Mortazavi, Mark Gerstein, Thomas Gingeras, Barbara Wold, Yi Sun, Melissa J. Fullwood, Edwin Cheung, Edison Liu, Wing-Kin Sung, Michael Snyder, * and Yijun Ruan *

1Genome Institute of Singapore, Singapore 138672, Republic of Singapore
2Program in Computational Biology and Departments of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520, USA
3Department of Computer Science, School of Computing, National University of Singapore, Singapore 117417, Republic of Singapore
4Center for Genomics and Personalized Medicine, Department of Genetics, Stanford University, Stanford, CA 94305, USA
5Department of Molecular and Medical Pharmacology, UCLA, Los Angeles, CA 90095, USA
6Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11797, USA
7Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA
8School of Biological Sciences, Nanyang Technological University, Singapore 637551, Republic of Singapore
9College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan 430070, P.R. China
10These authors contributed equally to this work

*Correspondence: mpsnyder@stanford.edu (M.S.), ruanyj@gis.a-star.edu.sg (Y.R.)
DOI 10.1016/j.cell.2011.12.014

SUMMARY

Higher-order chromosomal organization for transcription regulation is poorly understood in eukaryotes. Using genome-wide Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET), we mapped long-range chromatin interactions associated with RNA polymerase II in human cells and uncovered widespread promoter-centered intragenic, extragenic, and intergenic interactions. These interactions further aggregated into higher-order clusters, wherein proximal and distal genes were engaged through promoter-promoter interactions. Most genes with promoter-promoter interactions were active and transcribed cooperatively, and some interacting promoters could influence each other implying combinatorial complexity of transcriptional controls. Comparative analyses of different cell lines showed that cell-specific chromatin interactions could provide structural frameworks for cell-specific transcription, and suggested significant enrichment of enhancer-promoter interactions for cell-specific functions. Furthermore, genetically-identified disease-associated noncoding elements were found to be spatially engaged with corresponding genes through long-range interactions. Overall, our study provides insights into transcription regulation by three-dimensional chromatin interactions for both housekeeping and cell-specific genes in human cells.

INTRODUCTION

A fundamental question in biology is how genes and regulatory regions are organized and coordinated for transcription regulation. While operons, in which one promoter transcribes multiple genes in a single unit, are common in bacteria (Jacob et al., 1960), and bicistronic transcript structures have been described in worms and flies (Pauli et al., 1988; Zorio et al., 1994), eukaryotic genes are thought to be individually transcribed from their own promoters. However, evidence from in situ fluorescence studies in the last decade suggests that transcription is not evenly distributed and is instead concentrated within large discrete foci in mammalian nuclei, raising the possibility that genes are organized into “transcription factories” (Cook, 1999) containing RNA polymerase II (RNAPII) and other components for transcription. However, this theory lacks evidence with molecular and structural details. Thus, the question of how the regulation of genes is coordinated for transcription in mammalian cells remains largely open.

Mammalian genomes are known to be organized intensively into higher-order conformation inside the micron-sized nuclear space. Consequently, three-dimensional (3D) organization must have a role in the mechanisms for transcription regulation and coordination (Cremer and Cremer, 2001). Chromosome Conformation Capture (3C) and similar techniques (van Steensel and Dekker, 2010) along with traditional in situ techniques have demonstrated that chromatin interactions can regulate
transcriptional and epigenetic states (Cope et al., 2010). However, such analyses are either limited to certain specific domains or of low resolution and lack functional details. Therefore, a global and high-resolution map of functional chromatin interactions is likely to uncover underlying principles of the higher-order genomic architectures regulating transcription.

Recently, we developed Chromatin Interaction Analysis by Paired-End-Tag sequencing (ChiA-PET) for genome-wide investigation of chromatin interactions bound by specific protein factors (Fullwood et al., 2009). By immunoprecipitation of a factor of interest along with associated DNA fragments and followed by diluted proximity ligation of distant DNA fragments tethered together within individual chromatin complexes, we elucidated the association of regulatory information through nonlinear arrangements. We demonstrated that long-range chromatin interactions occur between the transcription factor Estrogen Receptor α (ERα) bound regions and their target promoters. To globally investigate how all active promoters dynamically interact with their corresponding regulatory regions in vivo, we used ChIA-PET to analyze genome-wide chromatin interactions associated with RNAPII. Our results provide insights into the 3D interplay of active promoters as well as regulatory regions and suggest an architectural model in which related genes in mega-base range are organized for efficient and potentially cooperative transcription.

RESULTS
Organizational Complexity of RNAPII-Associated Chromatin Interactions
We analyzed five different human cell lines (MCF7, K562, HeLa, HCT116, and NB4) using ChIA-PET with a RNAPII antibody (8WG16) that recognizes the initiation form of the protein. The cell lines originated from a wide range of lineages, and provided a broad representation of human cells. In our pilot analysis, about 20 million uniquely mapped paired-end reads were generated for each of the ChIA-PET experiments (Table S1A available online), which resulted in two genome-wide datasets: the ChiP-enriched RNAPII binding sites and the RNAPII-bound long-range chromatin interactions. Both intrachromosomal and interchromosomal interaction data were obtained, and the vast majority of chromatin interactions identified by ChIA-PET were intrachromosomal (Table S1B). Twenty-five intrachromosomal and seven interchromosomal interactions were validated either by 3C, DNA-FISH, or both (Figure S1 and inset of Figure 1C).

To present an inclusive view of the RNAPII-associated human chromatin interactome, we combined the ChIA-PET sequence reads from the six pilot experiments into one dataset for analysis (Table S1). Using embedded nucleotide barcode controls and statistical analyses, we assessed the data quality, filtered out the technical noise, and identified high-confidence binding sites and interacting PET clusters (Experimental Procedures). From the combined pilot dataset, we identified 14,604 high-confidence (FDR < 0.05) RNAPII binding sites as well as 19,856 high-confidence intrachromosomal interaction PET clusters (Table S3). The majority (83%) of RNAPII binding sites in the combined dataset were proximal to 5′ Transcription Start Sites (TSS) of genes (Figure 1A). There were also distinct but relatively weaker enrichments of peaks at the 3′ Transcription End Sites (TES) of genes. Similar patterns were seen in all the individual experiments. Of the total RNAPII binding sites, 9,487 (65%) were involved in chromatin interactions and these sites showed higher RNAPII occupancy than those not involved in interactions (Figure 1B), indicating that most highly-enriched RNAPII binding sites are involved in looped chromatin conformations.

Three basic types of interactions were identified around gene promoters in the combined pilot dataset: intragenic (promoter to gene internal regions, 938, 5%), extragenic (promoter to distal regulatory elements such as enhancer, 6,530, 33%), and intergenic (promoter-promoter of different genes, 8,282, 42%). There was also a subcategory composed of intermediate enhancer-enhancer interactions (4,106, 20%). Some interactions (2,341, 12%) were standalone duplex interactions between two interacting anchor regions, whereas most (17,515, 88%) were further aggregated into 1,544 interaction complexes.

We speculated that the isolated RNAPII binding at promoter sites, which are not involved in interactions, may reflect the basal promoter function for gene transcription, and thus were termed “basal promoters.” By contrast, RNAPII-associated interactions might constitute a structural basis for complex regulatory mechanisms. These basic interactions further aggregated into complex architectures which we classified as “single-gene” or “multigene” complexes depending on the number of genes involved (Figure 1C). The single-gene models consisted of single or multiple enhancer interactions with only one gene promoter, whereas the multigene models included intergenic promoter-promoter interactions and could also include intragenic and extragenic enhancer-promoter interactions. Moreover, several such complexes, distantly separated on a chromosome or on different chromosomes, further converged to form higher-order multigene interaction complexes (Figures S1B, S1D, S1F, and S1G). Many chromatin complexes had genomic spans of 150 kb–200 kb, and a few complexes spanned several megabases. Although there were only 1,328 multigene complexes in this combined pilot dataset, 11,723 genes were engaged in these complexes for an average of 8.8 genes per interaction complex (Figure 1D), indicating that promoter-promoter interactions were widespread and may play a significant role in transcription regulation.

To understand how these looping structures influence transcription, we characterized these RNAPII-associated chromatin models (basal promoters, single-gene and multigene complexes) for structural features (genomic property), functional output (transcription activity), and epigenomic marks (chromatin state).

Distinct Genomic Properties of Single- and Multigene Interaction Models
To determine the genomic characteristics of RNAPII-associated chromatin structures, we mapped several genomic descriptors that were known to associate with the expressivity of the human genome (Versteeg et al., 2003), including GC content, gene density, SINE/LINE density, gene length, and the intron/exon ratio. In our analyses (Figure 2, Figure S2A), the multigene complexes were significantly enriched with higher GC content, higher gene and SINE density, and lower LINE density as
compared to the single-gene interaction complexes and the regions of basal promoters, suggesting that multigene complexes were located in open chromatin and highly transcribed regions. In addition, genes in the multigene complex regions were relatively shorter than other gene categories, which is yet another property of highly expressed genes (Eisenberg and Levanon, 2003). Conversely, genomic loci associated with the single-gene complexes lay in the regions with lower gene and SINE density. Moreover, the genes engaged in the single-gene complexes were significantly longer and had higher intron/exon ratios than the genes of other chromatin models (Figure 2B). These observations suggest that genes with enhancer-promoter interactions in single-gene complexes were more likely to be tissue-specific or developmentally regulated, in line with the previous findings that genes in gene-poor regions associated with several distant regulatory elements, tended to be longer and had a higher noncoding to coding ratio than housekeeping genes (Eisenberg and Levanon, 2003; Taylor, 2005).

Interacting Genes Show Correlated Expression

To investigate the functional output of genes involved in the different chromatin models, as defined by transcriptional activity, we focused our analyses on MCF7 cells, as it is a well-characterized human cancer cell model with complementary datasets including RNA-Seq (Experimental Procedures), time-course microarray gene expression (Fullwood et al., 2009), and GRO-Seq datasets (Hah et al., 2011).

Consistent with the combined pilot dataset, 90% binding sites in MCF7 cells were found proximal to known gene promoters and 97% genes with RNAPII present at their promoters had
basal promoter genes (PCC: 0.24). Moreover, we observed
coefficient: PCC: 0.46 and 0.45 respectively) as compared to
RNAPII binding and RNA-Seq signal (Pearson's correlation
and the multigene models showed higher correlation between
regulatory sites correlated well with the expression level of the correspond-
As shown in Figure 3B, in general, RNAPII binding at promoter
detectable transcriptional activity by RNA-Seq (Figure 3A). The
interactive RNAPII binding sites that were distal to gene promoters included intra- and extragenic regulatory elements
such as enhancers. Approximately 45% of the extragenic distal
regulatory sites had detectable RNA signals that could represent possible noncoding RNA (ncRNA) transcripts.

For genes associated with the three chromatin models, we
analyzed the transcription levels measured by RNA-Seq reads.
As shown in Figure 3B, in general, RNAPII binding at promoter
sites correlated well with the expression level of the correspond-
ing genes. Interestingly, the genes involved in the single-gene
and the multigene models showed higher correlation between
RNAPII binding and RNA-Seq signal (Pearson's correlation coefficient: PCC: 0.46 and 0.45 respectively) as compared to
basal promoter genes (PCC: 0.24). Moreover, we observed
that genes linked by complex chromatin interactions, especially
those in multigene complexes, had significantly higher expres-
sion levels than basal promoter genes (Figure 3C). This high expression appeared to be limited to genes interacting at the
RNAPII anchor sites, as compared to genes located in the inter-
vening chromatin loops. These data indicated that promoter-promoter interactions in multigene complexes were associated
with higher transcriptional activity, which is consistent with our
observations of their associated genomic features.

Next, we characterized the expression patterns of genes
present in the interacting regions using microarray data derived
from 84 human tissues (Su et al., 2002). We found distinct repre-
sentation of tissue-specific and housekeeping genes in the three
chromatin models (Figure 3D, Figures S3A and S3B). Most genes
in single-gene complexes with enhancer-promoter connectivity
were tissue-specific, consistent with growing evidence that the
expression levels of developmental and tissue-specific genes
are largely modulated through cis-remote regulatory elements
and trans-protein factors (Hou et al., 2010; Schoenfelder et al.,
2010), and consistent with their genomic features (less gene
density, longer gene body and higher intron/exon ratio) as
previously described. Conversely, genes involved in multigene
complexes as well as the basal promoter genes were character-
ized as both tissue-specific and housekeeping categories.
These observations were also supported by normalized CpG
content and GC-skew at their promoter regions (Figures S3C
and S3D).

As promoter-promoter interactions cluster multiple genes,
they could provide an ideal topological framework for potential
transcriptional coordination of both tissue-specific and house-
keeping genes. This observation agrees with the evidence that
"ridges," which are domains of highly transcribed genes, contain
both housekeeping and tissue-specific genes (Versteeg et al.,
2003). Since large numbers of genes are found in multigene
complexes, we propose that promoter-promoter interactions
could serve as a dominant mechanism for transcription regu-
lation of both housekeeping and tissue-specific genes in mamma-
lan genomes.

Next, we sought to determine whether genes with promoter-
promoter interactions were more likely to be transcriptionally
coordinated. RNA-Seq data showed that most of the paired
genes with promoter-promoter interactions were expressed
together at high levels (Figure 3E; Figure S3E). To further assess
the coordinated transcription of paired genes across different
conditions, we performed Pearson's correlation analysis using
estrogen-induced time course of GRO-Seq data (Hah et al.,
2011) that measured transcription initiation rates of estrogen
responsive genes, and observed significant transcriptional correlation (Figure 3F; p value < 2.2E-16). Interestingly, the corre-
alation was even greater for ERα-mediated gene pairs derived
from our earlier data (Fullwood et al., 2009), suggesting stronger
correlation of transcription for genes involved in multigene
complexes mediated by specific transcription factors. Similar
correlation was also observed from other gene expression data-
sets (Figures S3F–S3I). As expected, housekeeping genes and
genes belonging to the same GO classes showed even higher
correlation than the rest (Figures S3J and S3K). Altogether, our
analyses indicated that a significant proportion of gene pairs

Figure 2. Genomic Properties of Promoter-Centered Chromatin
Models
(A) Aggregation plots showing enrichment of genes, SINE and LINE elements
around the TSS of genes in different chromatin models. Unique RefSeq TSSs
were used for analyses. Red curve stands for multigene (MG) model, blue for
single-gene (SG) model, gray for basal promoter (BP) model, and black dotted
line for the rest of the genes (R).
(B) Box-plots showing distribution of percentage GC content of GC isochore
around different models, gene length, and intron/exon ratio of RefSeq genes
involved in the models. Triple asterisks (**) signifies p value < 2.2E-16. Red
box stands for MG, blue for SG, and gray for BP. Open box is for R (rest of
genic regions) as background.
See also Figure S2.
involved in promoter-promoter interactions tended to be transcribed cooperatively.

**Multigene Complexes Provide Structural Framework for Cotranscription**

Correlated expression of interacting genes suggests that the multigene interaction complex might provide a molecular basis for the postulated “transcription factory” (Cook, 1999). To elucidate the link between the multigene complexes revealed by ChIA-PET and transcription factories, we performed 3D DNA-FISH experiments using probes representing distinct multigene complexes in combination with RNAPII-IF staining in MCF7 nuclei (Experimental Procedures). All experiments on four genomic loci randomly chosen from multigene complexes revealed a significant association of the multigene complex loci with RNAPII foci (Figure 4A-B), adding further evidence to support our view that multigene complexes could provide a structural framework for cotranscription.

Furthermore, gene families were significantly over-represented (p value < 0.006) in the multigene complexes (Figure S3L), such as HIST, ZNF, KRT, HOXC, etc. (Table S4). Taking the HIST1H family as an example, the 58 genes of this family located on chromosome 6 formed three multigene complexes, and these three complexes converged into a higher-order supercomplex, suggesting that all HIST1H genes were organized in a single chromatin architecture for coordinated transcription (Figure 4C). All HIST1H genes were actively transcribed in both MCF7 and K562 cells, and were highly coregulated across different tissues and cellular conditions (Figure 4D). Interestingly, HFE, a gene was not a part of the HIST1H family but was located in the middle of the first HIST1H multigene complex, was not anchored at the interaction sites and was not expressed. Similarly, the genes located in the intervening loop regions between the three HIST1H interacting complexes were relatively less active and much less coordinated for coregulation across different tissues and cellular conditions. This case exemplifies the model where multigene complexes organize genes with similar functions across genomic space for coordinated expression.

**Multigene Complexes Support Synergistic Transcription Regulation**

To further investigate the likelihood that the multigene complex structure might provide a topological framework for transcriptional coregulation of interacting genes involved in such topology, we designed a set of perturbation experiments to...
test this. After comparing the RNAPII and ERα ChIA-PET data from MCF7 cells, we found that the RNAPII-bound multigene complex at the GREB1 locus partially overlaps with the ERα-bound chromatin loops, suggesting that this interaction complex, in part, is associated with ERα. Therefore, we performed siRNA experiments to knockdown the protein level of ERα in MCF7 cells, and monitored the alteration of chromatin interactions and gene transcription in the GREB1 multigene complex. Several chromatin interaction loops at this locus were disrupted by siERα transfection as tested by 3C experiments (Figure 4E). In addition to GREB1, which had a strong response to estrogen induction and reduction by siERα knockdown (Figures S4A–S4D), we observed that the other genes in this complex such as E2F6, KCNF1 and ATP6VC12 also had various levels of response to induction by estrogen and reduction by siERα knockdown (Figure 4F). Interestingly, these genes did not directly interact with ERα at their promoter regions, but indirectly associated with ERα through RNAPII-bound chromatin loops. As a control, this effect was not seen in the nearby genes such as NOL10 and HPCAL1 that were in other RNAPII interaction complexes and also did not interact with ERα (Figure 4G). Similar results were observed at another interaction locus centered on the GPRR68 and CCDC86C genes (Figure S4E). Thus, these results indicate that a specific stimulus (estrogen) could lead to coactivation of genes organized primarily through RNAPII-bound multigene complexes, and perturbation at one gene locus (loss of ERα binding in this case) in a multigene complex could alter the transcriptional states of other interacting genes within the same complex. Although genes in close genomic distances with each other had been reported to be correlated in expression levels (Singer et al., 2005), our data suggests that the conjoint expression can be mediated through chromatin interactions. The functional significance of such coregulation needs further investigation.

**Epigenomic Marks Associated with Chromatin Interaction Sites**

To study the association of transcription factors (TFs) with the RNAPII interactions, we examined the enrichment of 20 different TFs in K562 cells at the RNAPII interaction sites from the three chromatin models in our K562 ChIA-PET dataset (Figures 5A and 5B, Figures S5A–S5D). General TFs such as E2F4 and E2F6 (Figure 5A, Figure S5A) directly bound at TSS sites (Figure 5B for a specific example). By contrast, specific TFs such as JunD and Max preferentially bound to distal regulatory sites and marked potential enhancers (Figure S5B). Several chromatin remodeling factors and chromatin organization proteins such as IN1, BRG1, CTCF, and RAD21 associated primarily with non-TSS sites, suggesting that they may mediate long-range interactions with enhancer regions (Figure 5A, Figure S5C). This hypothesis is consistent with other observations that IN1 and BRG1, two subunits of the SWI/SNF complex, were involved in transcriptional looping (Euskirchen et al., 2011). A common observation among all the factors was that interaction sites in the multigene complexes consistently showed elevated levels of factor enrichment, suggesting that the cooperative binding of factors in gene-rich domains leads to higher transcriptional activity, or these transcriptionally active open chromatin domains might converge to distinct specialized transcription factories, each enriched with general and specific TFs.

We further explored the histone modification data available from the ENCODE Consortium. Collectively, we found high enrichment of active histone modification marks coupled with a lack of repressive marks in RNAPII interaction sites, confirming that the RNAPII interaction sites mapped by our ChIA-PET data were located in promoter and distal regulatory regions engaged and/or poised for high transcription levels (Figure S5D). Interestingly, the enrichment of active marks was highest in the multigene complexes, indicating that these might constitute transcriptional hubs. Our observations matched previous findings that the enrichment of active histone modifications positively correlated with RNAPII occupancy (Barski et al., 2007).

We observed similar histone modification profiles in MCF7 cells (Figure 5C) using data that we generated previously (Joseph et al., 2010). In particular, we applied the log ratio of H3K4me3/H3K4me1 signal as a quantitative measurement of the likelihood that a genomic locus can act as a promoter or enhancer. Most noninteracting RNAPII sites proximal to TSS in basal promoter model showed high log ratios (Figure 5D, plot 1; median = 2.4; > 90% of the binding regions have log ratios > 0), whereas most of the RNAPII interaction sites distal to TSS in the single-gene complex model and the multigene complex model (conventional enhancer sites) showed low H3K4me3/me1 log ratios (Figure 5D, plot 4 and 6; median < –0.72), confirming that this log ratio could reflect relative capacities of promoters and enhancers. Surprisingly, examination of RNAPII interaction sites proximal to known TSSs in the multigene complexes (Figure 5D plot 5) revealed two peaks in the histogram of the log ratios, suggesting a mixture of enhancer and promoter elements in the promoter regions. Detailed profiles of H3K4me3 and H3K4me1 marks around the center (±5 kb) of those RNAPII interaction sites showed distinct characteristics of promoter-like, enhancer-like sub-groups (Figure 5D, heat-map). Moreover, enhancer-like RNAPII interaction sites, on average, showed lower transcriptional activity than the promoter-like RNAPII sites (Figure S5J). Thus, a large portion of interacting promoters may also have potential enhancer functions. We observed the same inverse correlation of H3K4me3/me1 log ratio at the TSS proximal and TSS distal RNAPII sites for K562 (Figure 5A), indicating that this observation is a general phenomenon applicable to all cell types.

**Interacting Promoters Possess Combinatorial Regulatory Functions**

To examine potential enhancer activity of promoters, we performed luciferase reporter gene assays, a commonly used method for promoter and enhancer characterization (Pan et al., 2008). In these assays, approximately 500 bp fragments of the expected promoter regions were cloned upstream of a luciferase reporter gene construct either in a proximal position as the driving promoter or in a distal position as a presumed enhancer, and the constructs were transfected into MCF7 cells (Experimental Procedures, Figures S5E–S5I). As shown in Figure 5E, the two interacting loci INTS1 and MAFK were 26 kb apart, and our RNA-Seq data suggested that both genes were active
Figure 4. Transcriptional Coordination in Multigene Chromatin Complexes

(A) Colocalization of multigene loci with RNAPII foci. Shown are the nuclear images of RNAPII IF-staining with four randomly-selected multigene loci (MG1-4) and 2 control loci. Representative gene loci are MED20, SYVNN1, HIST1, and PLEC1.

(B) Quantitative analysis of nuclei (n = 476) and alleles showing overlap of MG loci and RNAPII foci. Percentage overlaps from MG loci and those from control loci are significantly different.

(C) Super multigene complex of the histone gene family. Three distant clusters (C1, C2, C3) of HIST1H genes converge together in a super-MG complex. Shown are RNA-Seq, RNAPII and ChIA-PET tracks in MCF7 and K562 cells.

(D) Cotranscription of HIST1H genes in the super-MG complex in (C). Correlation matrix derived from publicly available microarray data of 4,787 samples (Supplemental Information). The rows and columns correspond to genes in each complex and the intervening regions.

(E) RNAPII-bound multigene complex at the GREB1 locus. Shown are the ERα- and RNAPII-bound chromatin interactions. Highlighted promoters are anchored by RNAPII, but not by ERα. The bottom panel shows relative interaction frequency by 3C-qPCR data for the perturbation experiments using siERα knockdown and estrogen induction.
in MCF7 cells. However, the normalized log ratio of H3K4me3/me1 was 0.36 for the INTS1 promoter and 1.13 for the MAFK promoter, suggesting that the INTS1 promoter may have enhancer properties. To test this, we cloned the INTS1 promoter fragment in both orientations upstream of the MAFK promoter flanking the luciferase gene. The luciferase reporter gene assay showed at least 7-fold enhancement of luciferase expression from the MAFK promoter activity by the INTS1 promoter fragment, indicating that a bona fide promoter can act as an enhancer to augment the activities of other promoters.

In another example (Figure 5F), the promoter of CALM1 interacts with an enhancer element 15 kb upstream and connects to the promoter of C14orf102 further upstream in 65 kb. Both RNA-Seq data and the H3K4me3/me1 log ratio indicated that the CALM1 promoter was strong, whereas the C14orf102 promoter was weak and enhancer-like. The luciferase reporter gene assay showed marginal enhancement to the CALM1 promoter reporter gene activity by the native CALM1 enhancer and the C14orf102 promoter individually. However, the combined CALM1 enhancer and the C14orf102 promoter together led to a significant ~3-fold enhancement of reporter expression from the CALM1 promoter. This result further validates the enhancer function by interacting promoters and elucidates a possibility of combinatorial effect among interacting elements in multigene interaction complexes for transcription regulation.

Next, we asked whether promoters with enhancer activity act specifically on their target genes. We swapped the promoter elements in the two examples of INTS1-to-MAFK and C14orf102-to-CALM1 for additional reporter genes assays (Figure 5G). Intriguingly, when placed upstream to the CALM1 promoter, the INTS1 promoter showed remarkable enhancement of CALM1 promoter activity. Similarly, the combined construct of C14orf102 promoter and CALM1 enhancer also increased MAFK promoter activity significantly. Meanwhile, a TATA box deleted promoter and other control promoters (either active or inactive), taken from the nearby genes that are not involved in a promoter-promoter relationship, did not show cooperative enhancement to MAFK and CALM1 promoter activities (Figures S5H and S5I). Thus, these results suggest a common property for promoters with enhancer capacity that could influence other promoters.

In addition, we also tested the combination of inserting the enhancer-like promoter fragment in the position proximal to luciferase gene and the strong promoter in the distal position in the reporter gene construct. Of the 20 such luciferase experiments, we observed that the weaker promoters conveyed significant enhancer function to their stronger interacting partners in luciferase activity rather than the reverse (Figure S5K). In the case of interacting pair INTS1 (enhancer-like promoter) and MAFK (strong promoter), the strong promoter MAFK did not demonstrate significant enhancer activity (Figure S5L). Thus, at promoter sites, there is an inverse relationship between enhancer and promoter functions.

**Cell-Line Specificity of Long-Range Chromatin Interactions**

To elucidate the cell-line specificity of chromatin interactions, we saturated the coverage of chromatin interactions through deep sequencing of more MCF7 and K562 ChIA-PET replicates (Experimental Procedures). The saturated libraries are highly reproducible for interactions, and thus highly reliable for intercell line comparative analysis. These libraries exhibit the same pattern of genomic descriptors as the pilot libraries (Figures S2B and S2C). With comprehensive ChIA-PET and RNA-Seq datasets, we performed comparative analysis between the two cell lines and identified cell-line specific genes and chromatin interactions (Figure 6A). Most of the genes specifically expressed in their respective cells also showed cell-specific interactions (Figure 6B), implying that cell-specific chromatin interactions provide the structural basis for cell-specific transcription. Gene Ontology (GO) analysis revealed significant enrichment of erythroid related GO terms such as response to stimulus and blood circulation for genes with specific expression and chromatin interactions in K562 cells, whereas GO terms such as ectoderm development and related biological process were enriched in MCF7 cells (Figure 6C, Figure S5A). As expected, the genes common in both cell lines showed enrichment of housekeeping functions like metabolism, cell-cycle and signal transduction (Figure S6B).

Among the chromatin interactions specific to K562 cells, we captured many previously characterized interactions including the x- and b-globin loci (Bau et al., 2011; Hou et al., 2010). Figure 6D shows extensive interactions identified by ChIA-PET data between the x-globin gene locus and the DNase hypersensitive (DHS) sites present in the gene body of the C16orf35 gene. Additionally, we found that the x-globin locus in K562 extended its interactions to the neighboring domains, which were constitutively active in both K562 and MCF7 cells, whereas the interactions to b-globin genes are K562-specific, suggesting a complex chromatin architecture for spatiotemporal regulation of both constitutive and cell-specific transcription. Similarly, the b-globin gene locus also displayed previously known K562-specific interactions with the nearby locus control region (Figure S6C).

**GREB1** is a well characterized MCF7-specific gene. As expected, we found abundant chromatin interactions associated with RNAPII at this locus in MCF7, but not in K562 cells (Figure 6E). In addition to recapitulating the previously identified ERx-associated interactions (Fullwood et al., 2009), RNAPII interaction data showed an additional interaction site on the far most upstream (left in Figure 6E) side of this complex. A strong H3K4me1 mark on this site suggested that this is potentially an enhancer site for a transcription factor other than ERx.

(F and G) Time course RT-qPCR following estrogen (E2) induction after siControl (solid) and siERx (dashed) transfections of MCF7 cells. Colors of the curves correspond to genes shown in (E). A secondary axis (red, right side) is used for GREB1 expression to accommodate its high expression level. Expression data of genes involved in the GREB1 multigene complex are in (F), and the data for genes outside of the complex are in (G). RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown.

See also Figure S4 and Table S2.
Intriguingly, a significant RNA-Seq peak was also identified at this site, indicating a possible enhancer RNA transcript, a new class of noncoding RNA species (Kim et al., 2010).

**Long-Range Enhancer-Promoter Interactions and Disease-Associated Noncoding Elements**

Our data showed that the enhancer-promoter interactions were significantly enriched over other types of interactions for cell-specific genes (Figure 7A) when compared to genes commonly expressed in both cell lines. This finding supported the general view that distant-acting enhancers tend to be specifically involved in tissue-specific genes, and was consistent with our analysis in Figure 3D. Although potential enhancer sites can be identified using high throughput approaches (Heintzman et al., 2009), it is still challenging to connect enhancers to their target genes that are hundreds of kilobases away. Moreover, many remote enhancers could be embedded in intronic regions of other distantly located genes (Visel et al., 2009), making it notoriously difficult to relate enhancers to their specific target genes. In this study, we identified tens of thousands enhancer-promoter interactions (Table S1C) including approximately 1000 ultra-long-distance (500 kb to megabases) events. We observed that ≥40% of enhancers do not interact with their nearest promoters and instead jump over to their target promoters, bypassing several intervening genes (Figure 7B, Figure S7).

An interesting example is the SHH gene that was expressed in MCF7 but not in K562 cells (Figure 7C). SHH is important in development and related to certain cancers (Lettice et al., 2002). Transcription of SHH is controlled by its enhancer which is located 1 Mb away and embedded in the intronic region of LMBR1; point mutation in this enhancer site is known to cause preaxial polydactyly, a common congenital limb malformation in mammals (Lettice et al., 2002). We found abundant interaction data between the SHH promoter and the previously characterized SHH enhancer site in the LMBR1 intronic region in MCF7 cells, but no interaction data in K562 cells (Figure 7C), which correlated well with their SHH transcription status. This is consistent with earlier observations (Amano et al., 2009).

In another interesting example, we identified two major interaction sites located ~600 kb and ~1 Mb downstream from the IRS1 gene promoter. IRS1 is known to participate in type-2 diabetes (T2D) mellitus, and is found specifically expressed in MCF7 cells (Figure 7D). A recent GWAS study uncovered a cluster of SNPs that is genetically associated with high risk to insulin resistance, T2D, and coronary artery heart disease (Kilpelainen et al., 2011). This high risk locus is found located in one of the IRS1 enhancer sites (Figure 7D). Thus, our data provides experimental evidence to suggest that this disease-risk locus could be physically connected with the IRS1 promoter, potentially serving as a critical long-range enhancer to regulate the expression of IRS1, in a similar manner as the SHH locus. Other examples of long-range and cell-specific enhancer-promoter interactions in MCF7 and K562 are shown in Figure S7. Taken together, these results suggest that ChIA-PET interaction data may better inform the association of a SNP with a gene involved in a disease process by providing evidence for direct physical interactions.

**DISCUSSION**

Through genome-wide mapping, we comprehensively analyzed RNAPII-associated long-range chromatin interactions. Our most interesting finding was the extensive promoter-promoter interactions among proximal and distant genes from 5 human cell-lines, which indicated that this mechanism is common in cells. Our work with reporter gene and siRNA knockdown assays provided experimental evidence that many promoters in the multigene complexes can cooperatively regulate the activity of other promoters with which they interact. Our observations thus blurred the conventional definition of promoter and regulatory elements for transcription. With such promoter-promoter interactions, we speculate that genetic error at one particular promoter might also propagate to other promoters and hence could lead to pleiotropic consequences depending on the interaction network within a cell type. Intriguingly, the multigene complexes illustrated in this study are, in principle, akin to the
Figure 6. Cell-Specific Chromatin Interactions

(A) Contour plots of RNA-Seq data (log RPKM, left) and chromatin interactions (log PET counts, right) in MCF7 and K562 cells, showing common and cell-specific gene expression and chromatin interactions.

(B) Contour plots of interaction data (log PET counts) for genes specifically and commonly expressed in MCF7 and K562 cells.

(C) Enrichment of cell-specific GO terms in genes and chromatin interactions specific in MCF7 and K562 cells. The p value of 0.01 is marked as dotted line.

(D) An example of K562-specific chromatin interactions. α-globin genes (in dotted line box) interact with distantly located (~20 kb) DHS sites (highlighted in yellow) which are known to interact with α-globin genes. In sharp contrast, the α-globin genes in MCF7 cells are not expressed and have no interactions with the DHS sites.
bacterial operon as a mechanism for coordinated transcriptional regulation of related genes, suggesting the possibility of a chromatin-based operon mechanism (chro-operon or chroperator) for spatiotemporal regulation of gene transcription in eukaryotic nuclei. However, the “chroperator” expression is not dependent on the linear arrangement of the genes, but is highly dynamic and can adopt a multitude of cassette configurations because of the combinatorics permitted by the looping interactions. Alternatively, these interactions could reflect stochastic movement of proximal and distant active genes to localized transcription factories.

An important question is how these multigene complexes are organized. A likely model is that a suite of protein factors for modulating gene expression in a functional regulatory cassette may result in optimal stoichiometry when aggregated in 3D space. This clustering also draws the regulated genes into a common spatial domain, similar to how the nucleolus is organized. The interacting regions can be established and/or maintained by potential chromatin bridging proteins such as cohesins (Merkenschlager, 2010) and CTCF (Handoko et al., 2011), and this process might be facilitated by chromatin remodeling proteins (Euskirchen et al., 2011), all of which are enriched at the interacting sites defined by RNAPII ChIA-PET data.

Long-range chromatin interactions including enhancer-promoter interactions are increasingly being recognized as an important mechanism to regulate many important genes. However, methods to identify such long-range relationships have been technically challenging. High-throughput approaches such as ChIP-Seq and DNase-Seq are efficient in identifying such long-range relationships, but lack the ability to interrogate the connectivity between the prospective enhancers and their target gene promoters. In this study using RNAPII as the protein target for ChIA-PET analysis, we identified a comprehensive repertoire of distant regulatory elements directly interacting with gene promoters. Many of them act through ultra-long-range chromatin interactions. Such distal enhancer-promoter relationships are particularly difficult to be identified by other approaches. As demonstrated in the cases of SHH and IRS1, long range interactions derived from ChIA-PET data could provide the connectivity of GWAS-identified high-risk loci to their target genes, and thus offer possible mechanistic explanations to the function of disease-associated noncoding elements. Further examination of spatial architectures revealed in this study will enhance our understanding of transcription regulation in normal and diseased conditions of human cells.

**EXPERIMENTAL PROCEDURES**

**Cell Culture**

Five cell lines, namely MCF7 (ATCC# HTB-22), K562 (ATCC# CCL-243), HCT116 (ATCC# CCL-247), HeLa (ATCC# CCL-2.2), and NB4, were grown under standard culture conditions and harvested at log phase.

**ChIA-PET**

Harvested cells were cross-linked using 1% formaldehyde followed by neutralization with 0.2M glycine. Chromatin was isolated and subjected to the ChIA-PET procedure (Fullwood et al., 2009). The ChIA-PET sequence reads were analyzed using ChIA-PET Tool (Li et al., 2010). The data are available from NCBI/GEO (ID GSE32664). Control and reproducibility analyses are described in Figure S8.

**RNA-Seq Data**

MCF7 mRNA was isolated following the protocol described in Ruan et al. (Ruan et al., 2007) for strand-specific RNA-Seq analysis by SOLiD sequencing platform. The rest of the RNA-Seq datasets for other cell-lines were retrieved from the ENCODE data repository site (http://genome.ucsc.edu/ENCODE/).

**ChIP-Seq Data**

The ChIP-Seq data were retrieved from (Joseph et al., 2010), (Raha et al., 2010) and the ENCODE data repository site (http://genome.ucsc.edu/ENCODE/).

**RNPAP IF Stain and DNA-FISH**

MCF7 cells were fixed using 4% formaldehyde followed by permeabilization with 0.04% Triton-X. After blocking with donkey serum, cells were incubated with primary antibody (8WG16) overnight followed by Cy3 conjugated secondary antibody for 1 hr. IF-stained cells were post-fixed and subjected to dehydration by 70, 80, 100% ethanol series, rehydration with 2X SSC and denaturation in 2X SSC/50% formamide at 80°C for 40 min. Biotin-16-dUTP and digoxigenin-11-dUTP labeled DNA probes were hybridized to cells at 37°C overnight in a humid chamber. Slides were washed, stained with DAPI, mounted and visualized by a Carl Zeiss LSM confocal microscope.

**Quantitative Chromosome Conformation Capture Analysis**

Targeted 3C products were analyzed by qPCR. The 3C-qPCR protocol was adapted and modified from the previous publication (Fullwood et al., 2009).

**Luciferase Reporter Gene Assay**

Dual luciferase assays were performed as described (Pan et al., 2008). Testing fragments were cloned into pGL4.10-basic vector. Constructs were transfected into MCF7 cells, and luciferase activities were measured following standard protocols.

**Statistical Analysis**

All the statistical tests were executed using the R statistical package (http://www.r-project.org/).

More details are available in Extended Experimental Procedures.

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes Extended Experimental Procedures, five tables, and eight figures and can be found with this article online at doi:10.1016/j.cell.2011.12.014.

**ACKNOWLEDGMENTS**

We acknowledge the Genome Technology and Biology Group at the Genome Institute of Singapore for technical support. This work was supported by Singapore A*STAR and an NIH grant (HG004456) to Y.R.

Received: July 18, 2011

Revised: October 21, 2011

Accepted: December 12, 2011

Published: January 19, 2012

See also Figure S6 and Figure S7.
Figure 7. Long-Range Enhancers and Disease-Associated Noncoding Elements

(A) Percentage difference of enhancer-promoter (EP) and promoter-promoter (PP) interactions in cell-specific versus common genes from MCF7 and K562 cells. The representation of EP interactions is significantly increased in cell-specific interactions, while the representation of PP interactions is decreased, when compared to interactions that are common to both cell lines.
REFERENCES


Supplemental Information

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture and ChIP Preparation
Five cell lines were used in this study for RNAPII ChIA-PET analysis. These are MCF7 (ATCC# HTB-22), K562 (ATCC# CCL-243), HCT116 (ATCC# CCL-247), HeLa (ATCC# CCL-2.2), and NB4 (Roussel and Lanotte, 2001) (provided by Dr. Sherman Weissman, Yale University). The cells were grown under standard culture conditions and harvested at log phase. The cells were treated with 1% formaldehyde at room temperature for 10 min on a plate rotator, followed by neutralization using 0.2M glycine. The cross-linked chromatin was obtained from the fixed cells by cell lysis and nuclear lysis. The chromatin was then subjected to fragmentation using Branson digital sonifier S450D to an average size of 300 bp. The sonicated chromatin was precleared overnight using Protein G magnetic beads to remove nonspecific DNA. Simultaneously, RNAPII Monoclonal antibody 8WG16 (Covance, MMS-126R) was incubated with Protein G magnetic beads overnight to allow antibody coating to the beads. The precleared chromatin was then immunoprecipitated overnight with the antibody-coated beads to capture chromatin of interest. The beads that were coated with DNA of interest were washed several times to remove nonspecific binding. A portion of ChIP DNA was eluted off the beads for concentration quantification using Picogreen fluorimetry and enrichment checking using quantitative PCR.

ChIA-PET Library Construction
Immuno-precipitated chromatin fragments were subjected to ChIA-PET library construction following the protocol as previously described (Fullwood et al., 2010; Fullwood et al., 2009) with some modifications. Briefly, the chromatin DNA fragments bound to antibody beads were divided into two aliquots for DNA linker ligation by Linker A and Linker B, respectively. The two linkers have the same nucleotide sequences, except four nucleotides in the middle are different (Linker A with TAAG; Linker B with ATGT) as the nucleotide barcode. The linkers were in excess so as to saturate the ends of DNA fragments. After the linker ligation and removal of excess linkers by washing the beads, the two aliquots were combined together for proximity ligation in diluted conditions, in which DNA fragments in individual chromatin complexes would have the same specific linkers (either A or B). During proximity ligation, DNA fragments within the same chromatin complex with the same linker would ligate together, so as to generate ligation products with homo-dimer linker composition (AA or BB). However, if the ligation reactions took place between DNA fragments of different chromatin complexes, such nonspecific ligation products would have a 50% chance of resulting with hetero-dimer linker composition (AB or BA). Hence, the hetero-dimer linker composition is an indicator of nonspecific ligation, which can be used to assess the nonspecific ligation rate of each ChIA-PET library and such nonspecific ligation data can be removed from further analysis. After proximity ligation, Paired-End-Tag (PET) constructs were extracted from the ligation products, and the PET templates were subjected to Illumina GAIIx sequencing.

ChIA-PET Library Data Processing
The ChIA-PET sequence reads were processed by ChIA-PET Tool (Li et al., 2010), a software package designed for ChIA-PET data analysis, with some modifications. Briefly, nonredundant PET sequence reads were first analyzed for linker barcode composition and identified as sequences with hetero-dimer AB linker (barcode TAAG / ATGT) derived from nonspecific ligation products, or sequences with homo-dimer AA or BB linker (barcodes TAAG / TAAG or ATGT / ATGT) derived from specific ligation products. The linker composition information was used later for noise analysis. Then, the linker sequences were trimmed, and the PET tag sequences were mapped to the human reference genome (hg19). To further remove possible redundant PET sequences after genome mapping, the PETs with genomic locations from both head and tail tags within 2 bp were merged to further reduce the library sequence redundancy arising from clonal PCR amplification. This step takes into account any Single Nucleotide Polymorphisms (SNPs) between the reference and the test genome and sequencing errors that may have occurred and resulted in a 1 bp or 2 bp difference in the tag sequences.

Peak Calling of RNAPII Binding
The coverage of all self-ligation PET sequences across the genome reflects the enrichment by RNAPII ChIP on specific locations, similar to ChIP-Seq mapping for protein binding sites. Using a similar method as that of the ChIP-Seq peak calling program MACS (Zhang et al., 2008), we performed peak calling on the ChIA-PET data. The local summits of the sequence coverage were called as potential peaks. The significances of the potential peaks were estimated with p values from a Poisson distribution. The background parameters in the Poisson distribution were estimated from the maximum of the global tag density, tag density in...
a 10 kb window around the peak, and the tag density in a 20 kb window around the peak. The p value was corrected as false discovery rate (FDR) with the Benjamini-Hochberg (B-H) method (Benjamini and Hochberg, 1995) for multiple hypothesis testing. The criteria for our final peaks were that 1) the sequence coverage is at least 5 and 2) the FDR is smaller than 0.05.

**Interaction PET Clusters**

Interligation PETs potentially reflect long range chromatin interactions. However, inevitably, there are technical noises from various sources. To further distinguish true interaction signals from nonspecific interaction noise, we reasoned that for true interactions, multiple interaction PETs would be generated from the same interacting regions. To identify such chromatin interactions, mapping locations of the interligation PETs were extended 1.5kb downstream, and the PETs that overlapped at both ends formed interaction PET clusters. Overlapping PET clusters are used to distinguish detectable interaction signals over background noise represented by singleton PETs, which could also include weak interaction events that are not distinct from background noise. The PET count of a PET cluster is the frequency of the interaction between the two locations involved. The statistical significance of such interactions was evaluated with p values from a hyper-geometric distribution. The hyper-geometric model takes into consideration the tag counts from both anchor regions and the sequencing depth for p value calculation, thus normalizing the effects of random ligations between two highly-enriched regions that would give rise to potentially noisy interligation PETs. The p values were corrected as false discovery rate (FDR) with the B-H method (Benjamini and Hochberg, 1995) for multiple hypothesis testing and the FDR cutoff is 0.05.

**Transcription Models from RNAPII Peaks and Chromatin Interactions**

Each of these interactions identified by PET clusters is also termed as a duplex interaction because each of them involves a pair of interacting anchors. The duplex interactions are further collapsed based on the connectivity of overlapping anchors with other duplex interactions to form complex interactions. With the high-confidence RNAPII peaks and interaction PET clusters, we defined three transcription models based on how the genes were involved in the interaction regions: basal promoter (BP) models (gene promoters that overlapped with standalone RNAPII peaks, but did not overlap with interaction anchors), single gene (SG) interaction models and multigene (MG) interaction models.

**ChIA-PET Library Statistics Summary**

For our RNAPII ChIA-PET analysis, we generated the data in two stages. In the pilot stage, we generated 6 individual datasets, including two biological replicates from MCF7 cell-line (MCF7_pilot_rep1 and MCF7_pilot_rep2). All these 6 individual datasets were combined into one human combined pilot RNAPII ChIA-PET library. In the saturated stage, we sequenced much deeper for MCF7 and K562 RNAPII ChIA-PET libraries, and used the PETs with homo-dimer linkers for further analysis.

The numbers of unique PETs, peaks and interactions from the individual libraries and the combined libraries are summarized in Table S1A. RNAPII binding peaks and high-confidence interaction PET clusters from the combined MCF7 pilot library, human combined pilot library, and two saturated libraries are in Table S3 (in separate sheets of the Excel file). Interaction categories and transcription models are summarized in Table S1C.

**Noise Analysis**

Noise level in ChIA-PET data was measured with two different methods: one method based on barcode linkers and the other method based on randomly rewired PETs. In the first method, two different barcode linkers have been designed in the ChIA-PET protocol.

Using the linker nucleotide barcode embedded in the PET constructs (Li et al., 2010), we estimated that up to 94% PETs with homo-dimer linker barcodes in a ChIA-PET library are specific intrachromatin fragment ligation products. Second, we used the hetero-dimer PET (nonspecific ligation) dataset to estimate the distribution of noisy data in terms of PET counts and the genomic spans covered by PET mapping. The heatmaps in Figures S8A and S8B show the frequencies of the interdimer linker barcodes in a ChIA-PET library are specific intrachromatin fragment ligation products.

The other peak in the span density of the homo-dimer PETs is from 10 kb to 1 Mb. Therefore, we chose a 1 Mb genomic span as a cutoff to define high-confidence interacting PET data. We used this cutoff to reduce the number of false positives, although this cutoff may increase the number of false negatives by excluding some interactions that are beyond the 1 Mb limit.

To further test the PET interaction cluster data, we randomly rewired the tags from specific homo-dimer ligation PETs to break up the pairing relationship of each PET. The pairing information in the real interligation PETs was removed first and the tags were randomly paired again to generate random rewired PETs. As shown, the span distribution from the rewired PETs (Figure S8D) is similar to the span distribution from the hetero-dimer linker PETs, but much different from the homo-dimer linker PETs. This means that our homo-dimer linker PETs are not random, but specific.
interligation PET clusters: (a) PET count ≥ 2 for each PET cluster from the pilot libraries, and PET count ≥ 3 for each PET cluster from the saturated libraries; (b) FDR < 0.05; and (c) Genomic span from 8 kb to 1 Mb. Such interaction clusters were mainly used for single-gene and multiple-gene complex study. Super long range intrachromosomal interactions (beyond 1 Mb span) were analyzed separately.

**Reproducibility Analysis of ChIA-PET Libraries**

To evaluate the robustness of the ChIA-PET method, we analyzed biological replicates of RNAPII ChIA-PET libraries from MCF7 and K562 cells at several different resolutions. There are two replicates from MCF7 in the pilot stage (MCF7_pilot_rep1 & MCF7_pilot_rep2), and two replicates from MCF7 (MCF7_saturated_rep1 & MCF7_saturated_rep2) and two replicates from K562 (K562_saturated_rep1 & K562_saturated_rep2) in the saturated stage. We assessed the reproducibility of the replicates at four different levels: reproducibility of library sequence reads, reproducibility of the RNAPII peaks, reproducibility of the individual interactions and reproducibility of RNAPII interaction regions. Some of the analysis results are presented in Figure S8.

**Reproducibility of Library Reads**

To evaluate the reproducibility of library sequence reads, we divided the genome into bins with bin size 10 kb and counted the numbers of tags in the individual bins for each library. For two replicates from the same cell-line, we used the tag counts from the same bins as pairs of data values to generate scatter plots. The result (Figures S8E–S8H) shows that the correlation of library reads from different replicates of the same cell line is high (correlation coefficients > 0.97). In contrast, the library reads from MCF7 and K562 RNAPII ChIA-PET libraries showed lower correlation (correlation coefficient 0.52). This data suggests that at the library sequence level, ChIA-PET sequencing is highly reproducible, and can reflect cell specificity.

**Reproducibility of RNAPII Peaks**

Next, we evaluated the reproducibility of RNAPII peaks identified by ChIA-PET libraries. First, we examined the distribution of the distances from the RNAPII peaks in one replicate to the nearest peaks in another replicate in order to decide the distance cutoff to be used for the overlap of RNAPII peaks from different biological replicates. The maximum distance allowed was 10 kb. The results in Figure S8I show the distribution of distances between RNAPII peaks from K562 saturated biological replicates, with the peaks from the first biological replicate at the center. We saw that the peaks from the second biological replicate are symmetrically distributed around the peaks from the first biological replicate and noticed that most peaks from one biological replicate have a peak from another biological replicate within 200 bp. Accordingly, 200 bp was used as the distance cutoff to overlap peaks from different biological replicates.

Using 200 bp as the cutoff for RNAPII peak overlap, we generated a Venn diagram (Figure S8J) and a scatter plot (Figure S8K) of the overlapped RNAPII peaks from biological replicates. The Venn diagram shows that the majority of the RNAPII peaks from two biological replicates are overlapped, and the scatter plot shows that the peak intensities of the overlapped peaks from two biological replicates are highly correlated.

In addition, we examined the peak intensities of the overlapped and nonoverlapped RNAPII peaks from different biological replicates. The box plots in Figure S8L show the peak intensities from overlapped (or common) peaks and the nonoverlapping (or unique) peaks for one of the biological replicates. The common peaks have higher peak intensities and the nonoverlapped peaks have lower peak intensities. This suggests that most peaks with significant binding signals are reproducible and reliable. The similar results of peak reproducibility were observed from MCF7 pilot replicates and saturated replicates. In summary, the RNAPII peak overlap analysis shows that the RNAPII peaks from different ChIA-PET biological replicates are highly reproducible.

**Reproducibility of Chromatin Interaction PET Clusters**

Third, we evaluated the reproducibility of interaction PET clusters from the replicates. Two interaction PET clusters are considered as overlapped if both anchors from one cluster have at least one base pair overlap with the anchors from another interaction PET cluster. Analysis result shows the scatter plot (Figure S8M) of the overlapped clusters from different biological replicates and the x axis and y axis show the PET counts from the overlapped clusters. If an interaction cluster from one replicate does not overlap with any clusters from another replicate, a dummy PET count 1 is assigned as the PET count from the counterpart replicate. The scatter plot shows that the interaction clusters from different replicates have a high overlap. The Venn diagram (Figure S8N) shows that the overlap ratios of the interactions from different replicates are up to 53%. The violin plot (Figure S8O) shows that the common interactions have higher PET counts and are more reliable, and the replicate-unique interactions have lower PET counts and are more dynamic. Similar results of interaction reproducibility were observed from MCF7 pilot replicates and saturated replicates.

**Reproducibility of Chromatin Interaction Regions**

Lastly, we analyzed the reproducibility of interaction regions from the biological replicates. Interaction regions are genomic regions covered by chromatin interactions, which include regions covered by standalone duplex interactions or complex interaction regions that are connected by a chain of multiple individual interactions. High-confidence intrachromosomal interaction regions were used for this analysis. If two regions have at least one base overlap, the two regions are deemed to be overlapped. The overlap ratios of the different replicates are 46.5% (MCF7 pilot), 75% (MCF7 saturated), and 66% (K562 saturated, Figure S8Q). If we sort the interaction regions by PET counts, we found that most of the interaction regions with high PET count are overlapped in the two replicates, as shown the Venn diagram (Figure S8R), where \( I_1, I_2, \) and \( I_3 \) are used to indicate the top 25%, 50% and 75% of the interaction regions from one replicate. This data suggests that most strong interaction regions identified in this study are reliable. The interactions with low PET counts are weak interactions and may vary between different replicates.
Saturation Estimates of RNAPII ChIA-PET Libraries

We assessed the saturation level of RNAPII interaction clusters identified by ChIA-PET based on the current sequencing depth and the overlap of the interaction clusters from the replicates. We assume that the total RNAPII interaction clusters in each cell line were sampled randomly and independently by the replicates. Given that most of the RNAPII interactions from the pairs of replicates are common to each other, the proportion of nonobserved interaction clusters from the replicates will be the product of the proportions of nonobserved interaction clusters from each replicate. Based on the assumption, we have the following formula:

$$\frac{N - \bigcup_{i=1,2} N_i}{N} = \frac{N - N_1}{N} \times \frac{N - N_2}{N}$$

Where $N$ is the total number of RNAPII interaction clusters in the RNAPII interactome of the specific cell line, $N_i$ is the number of RNAPII clusters found from replicate $i$, and $\bigcup_{i=1,2} N_i$ is the union of the interaction clusters from two replicates. This estimation is similar to the method called “capture-recapture” for population size estimation (Baillargeon and Rivest, 2007). The estimated numbers of interaction clusters at the same significance level (FDR < 0.05) and the similar sequence depth of the replicates are 4196 (MCF7 pilot), 48751 (MCF7_saturated), and 122244 (K562_saturated). Relative to these estimated numbers of the interaction clusters, the nonobserved interactions are about 44% (MCF7 pilot), 26% (MCF7_saturated) and 22% (K562_saturated). Such estimates suggest that the number of clusters from the union of two different biological replicates is nearly saturated for MCF7_saturated and K562_saturated libraries, when compared to MCF7 (unsaturated) library, at a given significance level.

Another way to study the saturation is to generate the interaction clusters from the randomly-sampled PETs from one replicate and overlap them with the interaction clusters from another replicate. We used two saturated K562 replicates for testing. Different proportions (10%, 20% ..., and 90%) of the PETs were randomly sampled from K562 saturated replicate 1, and interaction clusters were generated with the same pipeline. The overlap of the interaction clusters from different proportions of PETs with K562 saturated replicate 2 is near saturated to the overlap of the two replicates as in Figure S8P. This means that deeper sequencing from one replicate can’t improve the overlap ratio between different replicates substantially. To increase the total coverage of the chromatin interactions, it is better to generate more biological replicates.

RNAPII Immunofluorescence Staining Combined with DNA FISH

MCF7 cells were grown to 70% confluency in hybridization chambers and fixed with 4% paraformaldehyde (PFA) (Sigma-Aldrich) for 15 min at room temperature (RT) followed by permeabilization with 0.04% Triton X (Promega) for 30 min at RT. Prior to staining, cells were blocked with 10% normal donkey serum (Millipore) for 1 hr at RT and incubated with mouse RNA polymerase II 8WG16 monoclonal antibody (Covance, 1:1000) overnight at 4 °C. Cells were subsequently incubated with Cy3-conjugated donkey anti-mouse IgG polyclonal antibody (Millipore, 1:1000) for 1 hr at RT, after which slides were mounted with ProLong Gold antifade reagent containing DAPI (Invitrogen).

For combined RNAPII staining-DNA FISH, RNAPII-stained cells were post-fixed in 4% PFA for 10 min at RT after secondary antibody incubation and permeabilized in 0.5% Triton X for 10 min at RT. Cells are subsequently dehydrated through a 70%-80%-100% ethanol series, rehydrated with 2x SSC and denatured in 2x SSC/50% formamide at 80 °C for 40 min. Prior to FISH probe hybridization, cells were incubated in 2x SSC for 5 min at 4 °C and permeabilized in 2x SSC/0.5% Triton X for 5 min at 4 °C. Probe preparation and hybridization for FISH were performed as described for DNA-FISH. Carl Zeiss Meta LSM confocal microscope was used to analyze the association of RNAPII foci with promoter-promoter interaction loci. 63x optical lens and interslice distance (Z-axis) of 0.4 μm was used. Data was analyzed using LSM image browser and % overlap was determined manually by counting 476 interphase nuclei.

DNA Fluorescence In Situ Hybridization (DNA-FISH)

MCF7 cells were harvested by trypsinization and treated with 0.75M KCl for 15 min at 37 °C. Subsequently cells were fixed with Methanol/Acetic acid (3/1) and dropped onto slides for FISH. Following overnight culture in LB media, BAC DNA was extracted with a Nucleobond PC500 Plasmid Purification Kit (Macherey-Nagel) and labeled by nick translation in the presence of biotin-16-dUTP or digoxigenin-11-dUTP using a Nick Translation Labeling Kit (ENZO Life Sciences). In the presence of 1 μg/ml of Human Cot1 and Salmon sperm DNA (Invitrogen), labeled BAC clones were resuspended to 5 ng/μl in hybridization buffer (2xSSC, 10% dextran sulfate, 1x PBS, 50% formamide). Prior to hybridization, MCF7 nuclei on slides were digested with 0.005% pepsin (Sigma)/0.01M HCl at 37 °C for 3min followed by fixation with 1% formaldehyde (Merck - Calbiochem) and dehydrated through a 70%-80%-100% ethanol series. Labeled probes were denatured at 75 °C for 5min and hybridized to pretreated slides at 37 °C overnight. Post-hybridization washes were performed twice at 45 °C in 2xSSC/50% formamide for 7min each followed by 2 washes in 2xSSC at 45 °C for 7min each. Slides were revealed with avidin-conjugated fluorescein isothiocyanate (FITC) (Vector Laboratories, CA) for biotinylated probes and anti-digoxigenin- Rhodamine for digoxigenin-labeled probes (Roche). After washing, slides were mounted with vectashield (Vector Laboratories, CA) and observed under an epifluorescence microscope (ZEISS, Image.Z2) and under 63× lens magnification. Between 300-800 interphase nuclei were analyzed for each probe mix. Images were analyzed using metafer4. Center-to-center distance of 1μm was taken as cut-off for colocalization analysis. The BACs used in the experiments are listed in Table S2A.
Quantitative Analysis of Chromosome Conformation Capture Assays (3C-qPCR)

The 3C-qPCR protocol used for this experiment was developed and modified from the previous publications (Fullwood et al., 2009; Hagege et al., 2007; Pan et al., 2008). Briefly, 1 x 10^7 cells were fixed with 1% formaldehyde for 10 min at room temperature and lysed, and then nuclei were digested with EcoRI or HindIII (New England Biolabs) before ligation. All primers were designed within 150 bp from EcoRI or HindIII digestion site in unidirectional side. The specificity and amplification efficiency of each primer were tested by performing quantitative PCR on serial dilution of the BACs which were mixed in equal molar before digestion and ligation, after which the standard curve was obtained. The linear range of the 3C template was determined by a serial dilution of 3C sample with looping primers and nonlooping primers, and the optimal concentration of 3C sample for qPCR experiments was determined. Digestion efficiency, ligation efficiency, and sample purity were all checked as per established protocols (Hagege et al., 2007). To obtain the “cross-linking frequency,” Ct values of the experimental primer with the 3C template were first normalized with Ct values of the GAPDH control primer with the 3C template to ensure data between different cell types such as K562 and MCF7 were comparable. Then, we calculated each primer value of (Ct-b)/a according to a (slope) and b (intercept) based on the standard curve of the BACs sample. Finally, we transformed the values as 10^((Ct-b)/a) for each primer, and then divided the primer values (10^((Ct-b)/a)) by either the loading control (“Inner Ctrl,” meaning this control was designed against a genomic region shared with the BAC clone, where there were no restriction enzyme digestion sites) or an internal interaction control (inside the GAPDH region of the genome). The 3C quantitative results are presented as the mean ± s.e.m from two to four independent preparations of 3C sample with duplicate qPCR data. The BACs used in the 3C-qPCR experiments are listed in Table S2B.

Histone Modification and Transcription Factor ChIP-Seq Data

To characterize the peaks and interaction anchors from RNAPII ChIA-PET libraries, we used histone modification and transcription factor ChIP-Seq data to measure their profile around the peaks and interaction anchors. Histones can undergo a variety of post-translational modifications that can be identified by chromatin immunoprecipitating DNA with an antibody specific to a particular modification of interest. Several histone marks have been shown to mark sites associated with transcription such as promoters of active and inactive genes, exons of transcribed genes, etc. For K562, we used publicly available histone modification ChIP-Seq libraries from the Broad Institute and the ENCODE Project targeting the following histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K20me1, H3K27ac, H3K27me3, and H3K36me3 (Broad Institute and UCSC Genome Browser), and transcription factor ChIP-Seq libraries from Raha et al. (Raha et al., 2010). Histone modification ChIP-Seq libraries in MCF7 were from Joseph et al. (Joseph et al., 2010).

To determine whether the unique locations in RNAPII ChIA-PET libraries arose from promoters and enhancers, we used the log ratio of H3K4me3 and H3K4me1 tag counts from the unique locations. H3K4me3 is known to be a general mark for promoter regions, and H3K4me1 is known to be a mark for promoters and enhancers (Barski et al., 2007). For each location (RNAPII peaks or interaction anchors), the number of tags within ± 2 kb from the center of the locations were counted. The log ratio of the tag counts from H3K4me3 and H3K4me1 was normalized with the sequencing depths of both libraries. The locations with positive log ratios were considered to be potential promoters and the locations with negative log ratios were considered to be potential enhancers. In the pairs of interaction sites from promoter-promoter interactions, the interaction sites with lower H3K4me3/me1 log ratio in the individual pairs are treated as enhancer-like sites, and the interaction sites with higher H3K4me3/me1 log ratio in the individual pairs are treated as promoter-like sites. Generally, enhancer-like RNAPII interaction sites showed lower transcriptional activity, and the promoter-like RNAPII sites showed higher transcriptional activity (Figure S5J).

From multigene models in MCF7 pilot library, we generated the heatmap (Figure 5D) with H3K4me3 and H3K4me1 profiles from the peaks and interaction anchor centers. For each RNAPII peak or interaction anchor (if no RNAPII peaks), 10 bins with bin size 500 bp at each side of the centers were generated to count the histone modification ChIP-Seq reads. The profiles were normalized to the number of reads per bin per million reads. The signals are log-transformed. The rows in the heatmap correspond to the interaction anchors, and the columns correspond to the bins. The rows were clustered with kmeans method (k = 7) from R package. From the heatmap, we can clearly see that H3K4me3 is enriched at more than half of the centers of the interaction anchors, while H3K4me1 is enriched in smaller number of interaction anchors. For some interaction anchors, both H3K4me3 and H3K4me1 are enriched at heterogeneity regions.

Analysis of Histone Modification and Transcription Factor Binding Profiles in Chromatin Interacting Sites

To further characterize the transcription models from RNAPII ChIA-PET libraries, we used histone modification and transcription factor ChIP-Seq data to mark the locations from different transcription models. Signal tracks for all K562 ChIP-Seq experiments involving histone marks and transcription factors were downloaded from the ENCODE Project hosted by the UCSC Genome Browser (Kuhn et al., 2009). The signal levels for all tracks are given in terms of fold-change with regard to an equivalent distribution of reads and were generated with respect to human genome build hg19. All signal tracks were also corrected for local mappability and sequence depth.

Using these signal tracks and our region lists for TSS-associated and non-TSS-associated ChIA-PET peaks within each anchor, we used the ACT tool (Jee et al., 2011) to perform the signal aggregation over each list. The ACT tool operates on three distinct regions for each feature. Regions immediately upstream and downstream of each feature are subdivided into bins and the average signal level of
Total RNA was extracted from MCF7 cells and then mRNA was isolated following the protocol previously described in Ruan et al. (Ruan et al., 2007). Approximately one microgram of mRNA was used for fragmentation with RNase III, followed by gel-selection to obtain desired fragment range at 100-150 nt. The randomly fragmented mRNA was then hybridized and ligated to a mixture of linkers and adaptors obtained from SOLiD for reverse transcription (RT) to generate cDNA. The resulting cDNA template was amplified by PCR and analyzed by RNA-Seq using ABI SOLiD platform with single direction reads of 50 bp in length. Analysis was performed following the protocol and reagent kit provided from Life Technologies Inc., which generates strand-specific RNA reads.

**RNA-Seq Data Processing and Mapping**

After individual RNA reads were generated, initial filtering was undertaken to remove any noise sequences such as rRNA, tRNA, mitochondria RNAs and repeat sequences etc. The strand-specific reads were mapped on human reference genome hg19 with SOLiD whole transcriptome alignment pipeline and analyzed by ABI SOLID Bioscope (version 1.0.1) analysis pipeline. Two mismatches were allowed in each 25 bp seed with progressive alignment to find full mapping location. Mapping score was computed for each read and any location that was scored < 26 was filtered out and not processed further. A split-read mapping approach was also applied to map reads which spanned two exons. K562 RNA-Seq data was generated using the Illumina platform. Uniquely-mapped reads were mapped to human genome build hg19 and expression values obtained for all UCSC and RefSeq genes using RSeqTools (Habegger et al., 2010). Expression values were determined in terms of reads per kilobase per million mapped reads (RPKM) for coding sequences.

**Reproducibility Analysis of RNA-Seq Data**

The reproducibility of RNA-Seq data was evaluated with scatter plots and correlation of the reads. The genome was divided into bins with bin size 10 kb and the number of reads in each bin was counted. The scatter plot of sequence reads per 10 kb in biological replicates (Figure S8S) shows good reproducibility of MCF7 RNA-Seq data (Pearson’s correlation coefficient = 0.99). For K562 RNA-Seq data, we compared our in-house RNA-Seq library (as replicate 1) with that from Stanford (ENCODE data, as replicate 2). The scatter plot (Figure S8T) shows reasonable correlation between the two K562 RNA-Seq replicates with Pearson’s correlation coefficient = 0.72.

**Computational Characterization of Chromatin Interactions**

**Genomic Span Analysis of Chromatin Interactions**

To analyze the genomic spans of chromatin interactions identified by ChIA-PET, we calculated the distance between interacting loci at different PET count cut-offs. As a control, we randomly rewired the connections of the tags in the mapped PETs. Our results show the distribution of genomic spans of intra chromosomal interactions identified in MCF7 and K562 cell-lines. Difference between real and randomly rewired chromatin interactions can be seen. In particular, real chromatin interaction data shows a peculiar hierarchy of chromatin interactions. Short range interactions (~10 kb), long range interactions (~100 kb) and super long range interactions (>1 Mb) suggesting short range enhancer loops at local level, long range enhancer-promoter and promoter-promoter interactions at the middle level and super long range enhancer-promoter and promoter-promoter interactions at the top level respectively.

**Analysis of Genomic Descriptors**

GC isochores were taken from Costantini et al., 2006 (Costantini et al., 2006) and Hg17 coordinates were converted to Hg19 using the liftover utility of the UCSC genome browser. Gene, SINE and LINE densities (per 200 kb) are mapped around unique RefSeq TSSs engaged in each model. Gene length was calculated by subtracting TSS from TES coordinates of RefSeq genes present in the three models. To collect the genes present in the three models, an arbitrary distance of ± 1500 bp from anchor boundary to TSS site is considered. Total gene-span covered by introns is divided by total span covered by exons to calculated intron/exon ratio. Chromosome-wide distribution of the genomic descriptors examined in this study is exemplified in six different chromosomes (Figure S2A). This analysis was done on the combined pilot cell-line data and validated at the individual cell-line level (Figures S2B and S2C).
**Promoter CpG Analysis**
For HCG/LCG promoter analysis, the criterion adapted by Saxonov et al. (Saxonov et al., 2006) was used. Briefly, all the promoters (±1500 bp to the nonredundant TSSs) were pulled down from UCSC and the following equation was used to normalize the CpG content:

\[
\text{CpG}^{(0)} = \frac{P_{CG}}{(P_C + P_G)^2}
\]

Where, \(P_{CG}\), \(P_C\) and \(P_G\) are the frequencies of CG, C and G residues in the sequence respectively.

**Promoter Strand Bias Analysis**
Promoter GC skew is calculated in the sliding window of 100 bp across ±1500 bp around unique TSSs using following equation:

\[
\text{GC skew} = \frac{P_C - P_G}{P_C + P_G}
\]

Where, \(P_C\) = frequency of Cytosine and \(P_G\) = frequency of Guanine.

**Expression Breadth Analysis**
Gene expression data for 84 human tissues was downloaded from GNF Gene Atlas (http://biogps.gnf.org/) (Su et al., 2004). Genes showing at least 100 arbitrary units of probe difference value were considered as expressed. Probes not showing probe value ≥ 100 in any of the tissues were not considered in the analysis. A more stringent cut-off of 150 does not alter our conclusions (data not shown). Expression breadth is defined as the number of tissues a gene is expressed in. Tissue specificity of genes in the enhancer model was further validated by following two descriptors (Figures S3A and S3B):

a. \(TS(1) = \frac{\max(x_i)}{\text{median}(x_i)}\)

b. \(TS(2) = \frac{\sum (1 - \frac{x_i}{\max(x)})}{n - 1}\)

Where, \(x_i\) is the expression of a gene in a tissue \(i\), while \(i\) ranges from 1..n when \(n\) is the total number of tissues (84 here).

**PCC Analysis for Coexpression**
Gene expression data for 4,787 human microarray samples is downloaded from Boolean network dataset (http://gourd.stanford.edu/BooleanNet/), which covers wide range of expression differences like gender, tissues, development and differentiation stages etc. (Sahoo et al., 2008). Probes mapping to MG complexes are paired as per their specific interactions in the MG units and Pearson Correlation Coefficient (PCC) is calculated. Supporting analysis based on one probe per gene suggests that multiple probes mapping to single gene does not alter our observations (data not shown). The genes are then randomly rewired to compile a random control with the same gene background, but different pairing than the original MG pairs. Proximal gene pairs (up to 1 Mb) from the genomic spans over BP, SG and MG models are also taken as additional controls (Figure S3G). The same strategy was implemented for analysis (Figure S3F) on Estrogen induced time course Affymatrix data (6 time points) generated from MCF7 cell-line (Fullwood et al., 2009). GRO-Seq data used in Figure 3F was downloaded from (Hah et al., 2011). Similarly RNA-Seq data used in Figures S3H and S3I was downloaded from ENCODE UCSC page (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/). Controls taken in the GRO-Seq and RNA-Seq PCC analysis were: (1) randomly rewired MG gene pairs and (2) randomly picked proximal (upto 1 Mb) gene pairs from a control dataset having similar distribution of genomic span and gene density as MG regions.

**Gene Ontology Analysis**
For Gene Ontology (GO) analysis, loci were mapped to Entrez Genes and GO semantic similarities were calculated by the method defined by Fröhlich et al. (Fröhlich et al., 2007). To evaluate whether the higher PCC of paired genes in MG units is contributed by the genes in the same GO class, the PCC data described before was further split into two parts, one for the paired genes from the same GO class (GO similarity = 1) and the other for the pairs of genes belonging to different GO classes (GO similarity = 0) (Figure S3J). Gene pairs from the same GO classes have higher PCC in average.

**Gene Family Analysis**
Gene family information was downloaded from http://www.genenames.org/genefamily.html and curated manually. The gene family information from the MG models is listed in Table S4. To calculate the probability of finding two proximal genes (within 1Mb) from the same gene family, we performed > 1 million random simulations. This results to a distribution plot shown in Figure S3L, which suggests that gene families are significantly over-represented in MG complexes. To further scrutinize our observations, we compiled a random control dataset with the same genomic span and gene density distribution as of MG units. The dataset shows 3.4% of regions with at least two genes from the same gene family. Incidentally, the value overlaps precisely with the mean probability of finding 2 proximal genes form the same gene family.
**An Example of a Large Multigene Complex**

The largest multigene structure found in the human combined pilot RNAPII ChIA-PET dataset is on chromosome 11, spanning a 7.8 Mb segment (chr1:60927774-68723689) and covering 317 genes (247 genes with promoters proximal to multigene anchors and 70 genes in intervening loop regions) (Figure S1G). The large number of genes involved in multigene structures suggests that the multigene complexes are a major mechanism to organize multiple genes into distinct foci for efficient and coordinated transcription. It should also be noticed that there are 5,090 genes that are included in multigene loop regions, but far from any interaction anchors. As loop regions may be far from high local concentrations of transcription-related proteins (Fullwood et al., 2009), genes resided in loop regions may potentially be displaced into transcriptionally inactive zones (Figure S1G).

**Cell-Line Specificity Analysis of Chromatin Interactions and Genes Therein**

To identify the cell-line specific genes, we studied the gene expression from MCF7 and K562 cell lines. Each gene had a pair of expression levels from two cell lines, which were used to generate a scatter plot. The gene expression levels from the RNA-Seq library were measured as reads per kilobase per million reads (RPKM) and normalized with the median expression level from the library and log-transformed (the expression level of the genes without any reads were replaced with the minimum expression level from the library). The smoothed scatter plot in Figure 6A in the main text clearly shows that some genes only expressed in one cell line, but not the other.

We applied a stringent criterion to identify the cell-line specific genes: if a gene has reads from one cell line, but does not have reads from another cell line, this gene is specific to the first cell line. With this criterion, we identified 2025 MCF7 specific genes and 2486 K562 specific genes.

We then examined their chromatin interaction patterns. For each gene, we measured the PET counts from the interactions linked to the gene promoter region. Figure 6B in the main text shows the smoothed scatterplot of the PET counts at the same genes from two cell lines. Clearly, cell-line specific genes have more interactions in their corresponding cell line, suggesting that the expression of cell-line specific genes is associated with their cell-line specific chromatin interactions. The cell-line specific genes can be classified into different categories based on the chromatin interaction models they are involved in: genes with enhancer-promoter (EP) interactions, genes with promoter-promoter (PP) interactions, or none of above categories. Figure 7A shows the percentage difference in the representation of genes in EP and PP models in cell line specific versus common interactions. It is apparent that genes involved in enhancer-promoter interactions are significantly enriched in cell-type specific interactions, while genes engaged in PP interactions are enriched in chromatin interactions common in both cell-lines. Cell-line specific enhancer-promoter interactions from MCF7 and K562 cells are summarized in Table S8. Figure 6D-E shows two examples of the well known chromatin interactions: one for K562-specific chromatin interactions around HBA locus and one for MCF7-specific interactions around GREB1 gene.

**Gene Ontology Analysis of Cell-Line Specific Genes**

For all the cell-line specific genes engaged in distinct interaction categories, we examined the enrichment of GO terms (biological process) using Panther (Thomas et al., 2003). The highlighted functions are shown as Figure 6C in the main text. The complete list of significant GO terms is shown in Figure S6A. The genesets common in both cell-lines show enrichment of housekeeping functions as shown in Figure S6B. Two cell-line specific interactions are shown in Figure S6C (around β-globin gene locus, K562 specific) and Figure S6D (around GATA3 gene locus, MCF7 specific).

**Promoter Reporter Gene Assay**

To explore the potential regulatory relationship between interacting chromatin loci, we performed the dual luciferase assay for various combinations of promoter and enhancer regions taken from RNAPII ChIA-PET data of MCF7 cells. In these analyses, approximately 500 bp fragments of promoter and enhancer regions identified by RNAPII ChIA-PET data are isolated from MCF7 genomic DNA by PCR. All the primers used for plasmid construction are listed in Table S2C.

These fragments are cloned into pGL4.10-basic vector (Promega, Madison, WI). For the promoter activity test, the fragments are cloned flanking the 5' of the luciferase gene, whereas for enhancer test the fragments are cloned in the distal location approximately 2 kb to the 5' of the luciferase gene, in either orientations. For luciferase assay of reporter constructs, a constitutive renilla luciferase expressing vector pRL-SV40 (Promega, Madison, WI) was cotransfected as a control for transfection efficiency. MCF7 cells were transiently transfected using Lipofectamine 2000 (Invitrogen, Carlsbad, CA) and Opti-MEM® I Reduced Serum Media (Invitrogen, Carlsbad, CA). The luciferase activity was measured following the protocol as previously described (Pan et al., 2008), with minimal three replicates. We calibrated this reporter system using known examples of enhancers and promoters as positive and negative controls and tested a number of promoter-promoter and enhancer-promoter-promoter examples identified in MCF7 cells. The main results are presented in Figure 5 in the main paper. Additional results are in Figure S5.

RNA-PET analysis revealed that the CALM1 locus has a classic promoter and enhancer structure. RNA-Seq data showed that this gene is actively transcribed, and the H3K4me3 and H3K4me1 marks also characterized the expected promoter and enhancer properties. The normalized luciferase activities from the constructs (Figure S5F) suggested that the construct with promoter alone had significant signal above the background (basic construct and enhancer alone construct). The combination of enhancer and promoter construct showed significant increase of luciferase activity, suggesting that this reporter system is adequate for promoter and enhancer activity analysis.
As a negative control, we deleted the TATA box in the CALM1 promoter and cloned this TATA-less promoter fragment into the reporter construct. The luciferase assay results (Figure S5H) showed that in all constructs, the TATA-less CALM1 promoter had no activity whereas the native CALM1 promoter showed appropriate activities.

We have shown that certain promoters in multigene complexes have enhancer functions to other promoters in multigene complexes (Figures 5E–5G). We asked if nonmultigene promoters have the same property. We cloned the DDHD1 (an active promoter not in a multigene complex) fragment into the CALM1 promoter construct (Figure S5G), the ELNFI (an inactive promoter not in a multigene complex) and the DLD (an active promoter not in a multigene complex) into the MAFK promoter constructs (Figure S5I). In all cases, these non-MG promoters did not show enhancements in CALM1 and MAFK promoter activities.

Perturbation Experiments by siRNA Knockdown and Estrogen Induction

We used MCF7 cells for siRNA knockdown experiments. Following standard log phase growth in T75 cell culture flasks, on the first day of the siERx knockdown experiment, we passaged the cells into 6-well cell culture plates, and maintained the cells in phenol red-free DMEM/F12 containing 5% charcoal-dextran-striped fetal bovine serum to starve the cells of hormones. The next day, we changed the medium to DMEM/F12 without fetal bovine serum, and transfected the cells with siERx (on-targetplus SMARTpool, Dharmacon) or control siRNA (Dharmacon) with DharmaFECT 1 Transfection Reagent (Dharmacon). Successful siRNA knockdown was confirmed using Western Blot against ERα protein as well as RT-qPCR of ESR1 gene. The next day, we changed the medium to phenol red-free DMEM/F12 containing 5% charcoal-dextran-striped fetal bovine serum, left the cells to grow for another day, and in the following day, we induced the cells with 100 nM estradiol (Sigma Aldrich) or control vehicle, ethanol.

The quantitative Chromosome Conformation Capture (3C-qPCR) protocol used for this experiment was performed as described in the section “Quantitative analysis of chromosome conformation capture assays (3C-qPCR)” of this Extended Experimental Procedure. The 3C quantitative results are presented as the mean ± s.e.m from two to four independent preparations of 3C sample with duplicate qPCR data.

The cells were induced for 0h, 3h, or 6h, following which the cells were harvested and RNA was extracted using an RNAeasy kit (QIAGEN). The RNA was then reverse-transcribed to cDNA using a Superscript III cDNA synthesis kit (Invitrogen). RT-qPCR was then performed using primers designed with Roche Universal ProbeLibrary Assay Design Center against the genes of interest, using a Roche SYBR Green I master mix on a Roche Lightcycler 480 machine. At each time point, the RT-qPCR data were double normalized against GAPDH, and against the siControl, ethanol-treated sample as a baseline. The RT-qPCR data are presented as mean ± s.d. from two independent preparations of RNA sample with duplicate to triplicate qPCR data. The primer sequences used for RT-qPCR are listed in Table S2D.

SUPPLEMENTAL REFERENCES


Figure S1. Validation of Chromatin Interactions in MCF7 and K562 Cells by DNA-FISH and 3C-qPCR, Related to Figure 1.

(A) Quantitative DNA FISH data for positive (interaction) and negative (no interaction) hits randomly selected from MCF7 interchromosomal ChIA-PET data. (B) An example of a chromatin interaction between chr11 and chr17 in MCF7 cells. This exemplifies that multigene complexes from different chromosomes could further converge to a common active nuclear compartment. (C–G) Detailed 3C-qPCR validations for several long-range (up to \( >17 \) Mb) intrachromosomal interactions and an interchromosomal interaction (D). Most of the intrachromosomal interactions are tested in both MCF7 and K562 cell-lines. P values are calculated using binomial test. Panel D, F and G represent local interactions at distant genomic loci converging to each other via long range \( \text{cis} \) or \( \text{trans} \) interactions. 3C-qPCR mean values and standard error of means (SEM) from three independent experiments are shown. See also Table S2.
Figure S2. Detailed Genomic Features of Distinct Chromatin Models, Related to Figure 2

(A) Detailed examples from 6 different chromosomes illustrating the association of distinct chromatin architectures with genomic descriptors. Density of each descriptor (except %GC, which is measured in isochores) and the interacting anchors in each of our chromatin architectures is measured in each 1 Mb domain across chromosomes and running mean over 5 values are plotted. Certain gene rich domains enriched in multigene (MG) models and depleted in single-gene (SG) models are highlighted in red, while relatively gene-poor domains enriched in SG and depleted in MG are marked in blue.

(B and C) Genomic features of BP, SG, MG models in MCF7 (B) and K562 (C) saturated libraries. The plots validate our observations on the combined pilot data presented in Figure 2 of the main text.
Figure S3. Promoter Properties and Functional Output (Transcription) of Different Categories of Chromatin Models, Related to Figure 3

(A and B) Tissue specificity measured by descriptor-1 (A) and descriptor-2 (B). Equations for tissue specificity descriptors are given in the Extended Experimental Procedures.

(C and D) Normalized CpG content (C) of promoters (±1500 bp to TSS) and strand bias (CG-skew) (D) at promoters of genes in different models. Difference in the representation of High CpG (HCG) promoters (associated with housekeeping genes) and Low CpG (LCG) promoters (associated with tissue specific genes) is found to be significant between SG and MG complexes, while BP model has relatively negligible representation of LCG promoters suggesting their association primarily with housekeeping function. Similarly, CG-skew in (D) shows greater bias (associated with high and housekeeping expression) at promoter sites for BP and MG models, while lower bias (associated with lower and tissue specific expression) for SG model. These predictive measures support our observation in Figure 3D in the main text.

(E) Coexpression of interacting genes in K562 cells.

(F–K) Density plots for Pearson’s Correlation Coefficient (PCC) values of gene pairs in MG complexes (red), rewired pairs and random gene pairs selected from a control dataset of the same distribution of genomic spans and gene density as MG pairs with an upper limit of 1 Mb. The gene expression datasets analyzed are: (F) E2 induced time course microarray at 6 time points (Fullwood et al., 2009); (G) microarray dataset of 4,787 human samples covering a wide range of diversity in gene expression, like distinct tissues, gender, developmental and differentiation stages etc. (Sahoo et al., 2008). Different controls are selected over genomic spans of BP, SG and MG genes; (H–I) ENCODE RNA-Seq datasets for 5 different cell-lines (K562, MCF7, HeLa, HCT116 and GM12878) for MCF7 and K562 interactions; (J) PCC distribution for MG gene pairs belonging to the same and different functions (GO process); (K) PCC distribution for housekeeping (HK) and tissue specific (TS) gene pairs in MG units.

(L) Representation of gene families in random (#3383) and MG complexes (#1487) datasets with respect to expected probability “p” of finding 2 proximal genes (<1Mb) belonging to the same gene family. The method to compile random control and to calculate the probability of finding two proximal genes from the same gene family is given in the Extended Experimental Procedure. The plot suggests greater enrichment of gene families in multigene complexes. See also Table S4.
Figure S4. Chromatin Interactions and Gene Expression Following siERα Transfection in MCF7 Cells, Related to Figure 4

(A) Overlap of RNAPII loops with ERα loops at the GREB1 locus. P1, P2, and P3 are RNAPII interacting sites; E1, E2, and E3 are ERα interacting sites.

(B) ERα knockdown by siERα as tested by Western blot and RT-qPCR.

(C) GREB1 expression following 0, 3 and 6 hr of ethanol (ET) and estrogen (E2) treatment after siControl and siERα transfections. RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown.

(D) 3C-qPCR data for chromatin interactions at GREB1 locus following ET and E2 treatment after siControl and siERα transfections. 3C-qPCR mean values and standard error of means (SEM) from three independent experiments are shown.

(E) Estrogen induction and siERα knockdown led to correlated changes in the expression of interacting genes (CCDC88C and GPR68). ChiA-PET tracks clearly show that interaction between promoters of GPR68 and CCDC88C is associated with RNAPII, while ERα binds only at promoter and gene-body of CCDC88C. Color codes of the bars are shown in Figure S4D. RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown.
Figure S5. Enrichment Profiles of Transcription Factors and Histone Modification Marks Centered at the Interaction Anchor Regions of RNAPII-Bound Chromatin Interaction Structures in K562 Cells, and Reporter Gene Assays in MCF7 Cells, Related to Figure 5

(A) Aggregation plots of TFs enrichments centered at the RNAPII interaction sites, proximal to TSS (TSS) or distal to TSS (non-TSS). RNAPIII, as a negative control, shows negligible enrichment at the RNAPII interacting sites. y axis: sliding median for ChIP-Seq enrichment in the region. x axis: distance (bp) from RNAPII sites.

(B) TFs enriched at non-TSS (potential enhancer sites).

(C) Enrichment profile of chromatin remodeling and chromatin architectural factors.

(D) Enrichment profile of open chromatin and histone marks around RNAPII interacting sites. Clearly, the open chromatin mark DHS and active histone marks are substantially enriched at the RNAPII interacting sites, while the repressive histone marks show little enrichment.

(E) Map of pGL4.10 vector and cloning sites of promoters and enhancers for luciferase assays.

(F) Standard promoter and enhancer reporter assay for elements around the CALM1 locus. The enhancer upstream of CALM1 significantly, but modestly, enhanced the luciferase activity of the CALM1 promoter, which was involved in the typical enhancer-promoter interaction.

(G) The DDH1 promoter, which is located on the same chromosome as C14orf102-CALM1 locus and had no interaction with CALM1, showed no significant enhancement to the CALM1 promoter activity in luciferase assays.

(H) Deletion promoter reporter assay around CALM1 locus. The TATA box of the CALM1 promoter was deleted (from −133 bp to +100 bp, black arrow). The reporter construct containing this deletion promoter did not show any promoter activity in luciferase assays by itself or in any combinations with C14orf102.
Non-MG promoters do not possess enhancer functions. The ELFN1 and DLD promoters, which are located on the same chromosome as for the INTS1-MAFK locus and had no interaction with MAFK, did not enhance the promoter activity of MAFK in luciferase assays.

Box plots of RNA-Seq data in log2 RPKM for the genes with low and high log ratio of H3K4me3/H3K4me1 in the pairs of interaction sites. The genes with higher log ratio in a pairing relationship have higher RNA-Seq counts on average than the interacting partner with lower log ratio.

Box plots of normalized luciferase activities when the promoters with low log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs, or when the promoters with high log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs. The promoters with low log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs have higher enhancing effects in general.

Swap of INTS1 and MAFK promoters in positions in reporter gene construct for luciferase assays. The promoter sequence from INTS1 (with lower log ratio of H3K4me3/me1 signals) enhanced the luciferase activity of MAFK (with higher log ratio of H3K4me3/me1 signals). On the reverse, the MAFK promoter showed no enhancer function.

The mean values and standard deviations (SD) of luciferase activities from at least three independent experiments are shown.
Figure S6. Cell-Specific Interaction Analysis, Related to Figure 6

(A) All the Gene Ontology (GO) terms over-represented in the gene sets engaging cell-specific expression and interactions.
(B) Overrepresented GO terms in the gene set engaged in chromatin interactions common to both MCF7 and K562 cells. The abundance of housekeeping terms is apparent.
(C) K562-specific interactions around β-globin gene locus on chromosome 11. ChIA-PET loop tracks clearly show that there are chromatin interactions from β-globin genes to the locus control region (LCR) in K562 cells, but not in MCF7 cells. Correspondingly, the RNAPII and RNA-Seq show higher expression of β-globin genes in K562 cells, but not in MCF7 cells.
(D) MCF7-specific interactions around GATA3 gene locus on chromosome 10. In contrast to the β-globin gene locus, the ChIA-PET loop tracks clearly show that there are chromatin interactions from GATA3 gene locus to multiple enhancer sites in MCF7, but not in K562 cells. Correspondingly, the RNAPII binding and RNA-Seq showed high activity in MCF7 and low activity in K562 cells. Especially, one super-long-distance enhancer is about 1.2 Mb away from GATA3 promoter.
Figure S7. Examples of Cell-Type-Specific Long-Range Enhancer-Promoter Interactions, Related to Figure 6 and Figure 7
(A–E) specific to MCF7 cells, and (F–K) specific to K562 cells.
Most of regulatory sites in these long distance interaction examples are bypassing the nearest promoters and linking to other gene promoters.
(L) Several distant enhancers converging to MYC gene promoter. Cell-specific alternative usage of certain enhancers can be seen from the interaction loop views from MCF7 and K562 cells.
Figure S8. Assessment of Technical Noise, Library Reproducibility and Saturation Analysis, Related to Experimental Procedures

(A and B) Heatmaps of PET sequence counts versus genomic span for interactions identified from homo-dimer and hetero-dimer PETs from the combined pilot dataset.

(C) Densities of genomic spans of interactions from homo-dimer PETs and hetero-dimer PETs of the combined pilot data.

(D) Densities of genomic spans of interactions from rewired PETs.

(E–G) Scatter plot of sequence reads per 10 kb from RNAPII ChIA-PET replicates: (E) MCF7 pilot datasets, (F) MCF7 saturated and (G) K562 saturated datasets.

(H) Scatter plot of sequence reads per 10 kb from K562 saturated and MCF7 saturated RNAPII ChIA-PET datasets.

(I–L) RNAPII binding site reproducibility of K562 saturated replicates. (I) Histogram of genomic distances between RNAPII peaks from replicates. (J) Venn diagram of RNAPII peak overlap between replicates. (K) Scatter plot of RNAPII peak intensities of replicates. (L) Box plot of peak intensities of RNAPII peaks common and unique in replicate 1.

(M–O) RNAPII interaction reproducibility of K562 saturated replicates. (M) Scatter plot of interaction PET counts between replicates. (N) Venn diagram of interaction region overlaps from replicates. (O) Violin plot of interaction PET counts from common and unique interactions from replicate 1.

(P) Saturation assessment of chromatin interactions from K562 saturated RNAPII ChIA-PET replicates. The overlap ratio between replicates against the proportion of PETs sampled from K562 saturated replicate 1 (more details in the Extended Experimental Procedure; under saturation analysis).

(Q and R) RNAPII interaction region reproducibility of K562 saturated replicates. (Q) Scatter plot of interaction region PET counts between replicates. (R) Venn diagram of interaction region overlaps from replicates. I, II, and III for the top 25%, 50% and 75% interaction regions from K562 saturated replicate 1.

(S and T) Scatter plots of RNA-Seq reads per 10 kb in replicates from MCF7 (S) and K562 (T).