Published online 21 March 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 9* ***4835–4846***
doi:10.1093/nar/gkt188

High-throughput identification of long-range regulatory elements and their target promoters in the human genome

Yih-Chii Hwang[1], Qi Zheng[2,3], Brian D. Gregory[1,2,3,*] and Li-San Wang[1,3,4,*]

[1]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA, USA, [2]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA, [3]Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, PA, USA and [4]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA

Received October 11, 2012; Revised February 26, 2013; Accepted February 27, 2013

ABSTRACT

Enhancer elements are essential for tissue-specific gene regulation during mammalian development. Although these regulatory elements are often distant from their target genes, they affect gene expression by recruiting transcription factors to specific promoter regions. Because of this long-range action, the annotation of enhancer element–target promoter pairs remains elusive. Here, we developed a novel analysis methodology that takes advantage of Hi-C data to comprehensively identify these interactions throughout the human genome. To do this, we used a geometric distribution-based model to identify DNA–DNA interaction hotspots that contact gene promoters with high confidence. We observed that these promoter-interacting hotspots significantly overlap with known enhancer-associated histone modifications and DNase I hypersensitive sites. Thus, we defined thousands of candidate enhancer elements by incorporating these features, and found that they have a significant propensity to be bound by p300, an enhancer binding transcription factor. Furthermore, we revealed that their target genes are significantly bound by RNA Polymerase II and demonstrate tissue-specific expression. Finally, we uncovered that these elements are generally found within 1 Mb of their targets, and often regulate multiple genes. In total, our study presents a novel high-throughput workflow for confident, genome-wide discovery of enhancer–target promoter pairs, which will significantly improve our understanding of these regulatory interactions.

INTRODUCTION

While the Human Genome Project was declared complete in 2003, many regulatory elements still remain undefined. Enhancers are one such class of elements because true definition of an enhancer requires identification of both the regulatory sequence and its interacting promoter region(s). Enhancer–target identification is further complicated by the fact that they interact in an orientation-independent manner, can be millions of base pairs away from each other or even reside on different chromosomes (1–3). Enhancer elements also have dynamic regulatory activities under various developmental and environmental conditions. For instance, they can activate gene expression in a tissue- and temporal-specific manner. Thus, they affect different sets of genes in different tissues (4) and/or play variable regulatory roles during animal development (5,6). One well-studied example of this dynamic property is the locus-control region (LCR) that regulates the cluster of five human β-type globin genes on 11p15.4 (6). These globin genes are exclusively expressed in erythroid cells and are expressed differentially in fetal and adult cells mediated by the LCR that is located about 40 kb upstream.

Recent studies reveal that in eukaryotes, histone modifications such as histone 3 lysine 27 acetylation (H3K27ac), histone 3 lysine 4 mono-methylation (H3K4me1), di-methylation (H3K4me2) and tri-methylation (H3K4me3) can play crucial roles in the activation of enhancer elements under different environmental conditions, cell lineages, tissue types or developmental stages (4,7–10). These activating histone marks tend to be present in enhancer elements that are activated and absent when they are repressed. Additionally, activated regulatory elements are more likely to be located within the context of accessible (open) chromatin where they can be bound by transcription factors. The accessibility of specific DNA

*To whom correspondence should be addressed. Fax: +1 215 573 3111; Email: lswang@mail.med.upenn.edu
Correspondence may also be addressed to Brian D. Gregory. Fax: +1 215 898 8780; Email: bdgregor@sas.upenn.edu

sequences can be determined by their sensitivity to digestion by DNase I, with open chromatin being highly digested and vice versa. Recently, large-scale studies of activating (e.g. H3K27ac) histone marks and DNase I hypersensitive sites (DHSs) such as those from the Encyclopedia of DNA Elements (ENCODE) (11,12) have been used in various human cell types to predict enhancer elements (9,10). Additionally, other high-throughput studies assaying E1A binding protein p300 and CREB binding protein interaction sites have also been used to discover putative enhancers (4,13,14). Although these studies can predict enhancer elements on a large scale, they suffer from the inability to globally identify the target gene promoters of the identified enhancer elements.

Although the mechanism of enhancer–target promoter interaction formation is still not well understood, it is commonly accepted that enhancers and promoters interact with each other through a transcription factor protein complex (15). Based on this model, the chromosome conformation capture (3C) approach can be used to identify enhancer elements as well as their target genes simultaneously by detecting two linearly independent DNA segments that are bound to one another via a protein complex. One major drawback of the 3C approach is that it requires prior knowledge of the putative enhancer and promoter elements to allow design of specific PCR primers, which is often unknown. To address this limitation, a high-throughput version of 3C was developed (Hi-C) (16) to detect genome-wide DNA–DNA interaction events. This approach avoids multiple PCR steps by ligating interacting DNA elements followed by high-throughput sequencing to provide unbiased identification of DNA–DNA interacting pairs. Several variants have been developed by other groups to identify the chromosome organization and regulatory sites of the human, yeast and *Drosophila melanogaster* genomes (17–20). However, these original studies focused on determining large-scale chromosomal organization, and did not demonstrate whether the high-throughput sequencing variant of 3C is sensitive or specific enough for prediction of enhancer–promoter interactions.

More recently, Chepelev *et al*. (21) developed chromatin interaction analysis using paired-end tag sequencing (ChIA-PET), which is a strategy that combines 3C with ChIP-seq, for an enhancer associated histone modification (H3K4me2) to identify intra-chromosomal enhancer–promoter interactions (22). This led to the successful identification of only intra-chromosomal enhancer–promoter interactions that were associated with a specific histone modification (H3K4me2). Another recent study applied the variant 3C method [carbon-copy chromosome conformation capture (5C)] to identify ~100 enhancers and their specific target genes by designing ~6000 primers along the ENCODE pilot project regions (23). Although none of these previous studies were at the genome-wide scale, they have demonstrated that datasets produced by the 3C method can be used for genome-wide identification of enhancer–target promoter interactions.

Here, we revisit the original Hi-C experimental data with the goal of identifying enhancer–target gene interactions on a genome-wide scale for humans. To do this, we developed a new analysis framework for Hi-C experiments that integrates multiple genome-wide enhancer-defining datasets to identify enhancer–target gene pairs. Using this approach, we identified thousands of high-confidence enhancer–target promoter interactions in two different human cell types. We validated these interaction pairs by demonstrating our putative enhancer elements are highly correlated with known p300 binding sites, and their target gene promoters are enriched in RNA Polymerase II (Pol II) binding. Furthermore, we found that the predicted enhancer elements are conserved in the mammalian lineage, and their target genes are expressed in a highly cell type–specific manner. In total, our pipeline has allowed the first robust and genome-wide discovery of thousands of novel enhancer–promoter interactions in the human genome.

MATERIALS AND METHODS

Hi-C data from two human cell types

To comprehensively identify putative enhancer–target promoter interactions in the human genome, we first downloaded the original genomic alignments for the paired-end Hi-C sequencing data from two different human cell lines, a lymphoblastoid (GM06990) and a chronic myelogenous leukemia (K562) cell line (GEO accession number GSE18199). After an initial analysis, we found that the overlap of extended hotspots (defined below) between biological replicates of the GM/HindIII sample is 50.6% ($P < 2.2e{-}16$, chi-square test). Furthermore, the sequencing reads from these same samples were also combined in the original Hi-C study. Therefore, to more comprehensively identify DNA–DNA interacting pairs, mapped reads from biological replicates of the GM/HindIII sample were merged to single datasets. In total, we looked at the interacting patterns for three different sample sets from the original Hi-C study: GM06990 cells with HindIII (GM/HindIII), GM06990 with NcoI (GM/NcoI) and K562 with HindIII (K562/HindIII) digestion.

Identifying statistically significant DNA interacting hotspots from Hi-C datasets

We first identified significant clusters in the Hi-C data using a geometric distribution-based model (24). To do this, we assembled all mapped reads for a given dataset (GM/HindIII, GM/NcoI or K562/HindIII) into consecutive contigs (made up of overlapping reads) for each nuclear chromosome, without initially considering the read pairing information for these libraries. This approach allowed us to determine the gap regions between the identified contigs. These gap lengths should follow a geometric distribution:

$$P(X_i = k) = (1 - p_i)^{k-1} p_i$$

$$P(X_i \leq k) = 1 - (1 - p_i)^k$$

where X_i and p_i are the gap lengths and the probability of a position covered by any read on chromosome i,

respectively. Accordingly, we fit the gap lengths to a geometric distribution for each chromosome (Supplementary Figure S1) and estimated p_i based on the mean gap length. We then grouped contigs into clusters by merging nearby contigs based on the gap distances between them. Specifically, contigs were merged into significant clusters if they are closer to each other than the 5% quantile according to the fitted geometric distribution.

Next, we identified high-confidence DNA interacting hotspots by fitting cluster lengths to an additional geometric distribution for each nuclear chromosome (Supplementary Figure S2), where the X_i value is based specifically on cluster length and the p_i value is the emission probability based on the mean cluster length calculated for the Hi-C data for chromosome i. Only the significant clusters ($\geq$99% quantile) identified with this second geometric distribution-based test were retained and defined as DNA interacting hotspots. It is worth noting that we did not take into account the Hi-C interaction data for these hotspots during this analysis step, but only looked for interacting partners during our analysis to identify those hotspots that are putative enhancer elements (see below). We also analyzed DNA interacting hotpots identified using the quantiles of 98 and 99.9%, and the results of these analyses are presented in Supplementary Figures S4 and S5, respectively.

Compensating for restriction enzyme fragmentation bias and identifying bona fide DNA interacting hotspots

In Hi-C, the resulting sequencing reads are enriched for regions of the genome near or at the restriction enzyme (RE) sites used in the experiment (as also indicated by our motif analysis; see corresponding Results section) rather than the actual enhancer–target interaction site. Thus, the actual interaction site could be at any location between two restriction sites that are thousands of base pairs apart. To address this bias, we examined the distribution of the distances between adjacent restriction sites within the human genome, which was then fitted to a poisson distribution:

$$P(X = k) = \lambda^k e^{-\lambda}/k!$$

where X is based on the distances between adjacent restriction sites, and parameter λ is the mean of the distances along the genome (Supplementary Figure S3). Based on the estimated parameter λ, we extended the boundaries of DNA interacting hotspots in both directions from the midpoint of each hotspot to the length of the 95% quantile (~3 kb) of the fitted poisson distribution. This extension makes it much more likely that we are covering the actual DNA–DNA interaction sites, which are linked to the closest RE site by the nature of the Hi-C protocol. The resulting regions were referred to as extended hotspots.

Identification of candidate enhancer elements enriched in activating histone modifications

We began selecting for candidate enhancer elements (CEEs) by focusing on the extended hotspots that (i) had at least one of its interacting regions overlapping a protein-coding gene promoter and (ii) the particular CEE–promoter interaction was supported by >1 paired-end sequencing read in the corresponding Hi-C dataset. We then determined the overlap (see below for description of enrichment analyses) between these promoter-interacting extended hotspots and the four activating histone marks (H3K4me1, H3K4me2, H3K4me3 and H3K27ac) that are known to be associated with enhancer elements in the human genome (4,7–10). We also examined the overlap of these promoter-interacting extended hotspots with H3K27me3, a heterochromatic histone modification (23) that is not enriched at enhancer elements. To further select for CEEs that are likely bona fide enhancer elements, we only maintained promoter-interacting extended hotspots containing known enhancer-related histone modifications that are also enriched in DHSs. Thus, CEEs are defined as highly confident promoter-interacting extended hotspots enriched in known enhancer-related histone modifications and DHSs. For these enrichment analyses, the histone modification and DHS data was downloaded from the UCSC ENCODE production phase (hg18 assembly) (24,25). It is of note that we used the lymphoblastoid cell line (GM12878) and chronic myelogenous leukemia (K562) from the ENCODE project in our study, as they are the most closely related cell lines to those used in the original Hi-C study (GM06990 and K562). As a control, we generated 1000 sets of the same number of extended hotspots randomly selected from the human genome (random extended hotspots), and used them as a background to evaluate the significance of enrichment for all subsequent analyses.

Enrichment analyses

All enrichment analyses for CEEs and their target promoters (e.g. p300 binding) were performed by computing the enrichment index (*ERI*) as a ratio of the two proportions:

$$ERI(A) = C(A)/P(A)$$

where A is the set of intervals for a particular histone modification or other genomic feature (e.g. DHS, p300 binding or Pol II binding) determined using ENCODE ChIP-seq or DNase-seq experiments (26). $C(A)$ is the total length in base pairs of CEEs (or interacting promoter hotspots if we are examining target promoter characteristics) that overlap with A, *and* $P(A)$ is the mean of total lengths that overlap with A from 1000 random control sets (see above). It is worth noting that in enrichment analyses for CEEs, each permuted set is selected randomly from the collection of all extended hotspots with the additional constraints that they must have similar chromosomal and length distributions as the set of CEEs being analyzed (28). For enrichment analyses of CEE target promoters, each control set is selected randomly from the promoter regions of the 21 522 non-redundant protein-coding genes in the hg18 assembly. Thus, a high *ERI* for a set of CEEs or their target promoters indicates that they tend to be overlapping with a particular histone modification or binding feature when compared with all extended hotspots or non-redundant protein-coding gene promoters, respectively.

Characterizing p300 binding to CEEs and RNA Pol II binding to their target promoters

We downloaded the previously identified p300 and RNA Pol II binding sites from the UCSC ENCODE database for GM12878 and K562 cell lines (hg18 assembly) (25,26,29). We then calculated the *ERI* for p300 binding within CEEs as well as RNA Pol II binding to CEE interacting promoters as described above.

Determining cell type–specific expression of CEE target genes

We downloaded the previously published gene expression profiles for the nine ENCODE human cell lines (GSE26312) (30). Data were normalized by RMA (31–33) and $\log_2$-transformed. We aggregated probeset-level to gene-level expression values for each cell line as follows. For each probeset, we computed the average expression level across replicates. For each gene, we then computed the average expression across multiple probesets (if applicable). The gene expression profiles of the nine ENCODE cell lines were combined into a common gene set (13 436 genes), and between sample expression values were normalized again to eliminate any array-specific bias using quantile normalization (normalize.quantiles function in R/affy package) (31). To determine if a gene has strong tissue-specific expression in either GM12878 or K562 cells compared with the other seven cell types, we used an entropy-based metric (34) as follows. For each gene g, we computed $p_{c|g}$ as the expression level in cell type c divided by the sum of expression levels across all nine cell lines. The entropy (35) for g is defined as $H_g = -\sum_{1 \le c \le N} p_{c|g} \log_2(p_{c|g})$, where $N = 9$ is the total number of cell types in this study. H_g ranges between 0 (gene g is expressed in only one cell type) and $\log_2(N)$ (gene g is expressed uniformly in all cell types). To measure the specificity for a particular cell type c, we computed $Q_{g|c} = H_g - \log_2(p_{c|g})$. The quantity $-\log_2(p_{c|g})$ has a range between 0 (when gene g is only expressed in cell type c) and infinity (when gene g is not expressed in cell type c).

Sequence motifs in CEEs

We examined the sequence motifs of the CEEs using the HOMER software package (36), and only considered 8, 10 and 12 bp for the motif length in each sample. We used all extended hotspots as the background when searching for overrepresented motifs (-bg parameter in HOMER) in an effort to reduce potential biases introduced toward restriction sites owing to the original Hi-C protocol. Significance levels were set as $P < 0.05$.

RESULTS

Identifying candidate DNA interacting sites: hotspots and extended hotspots

We built an analysis workflow that extracts high-quality DNA interacting sites from Hi-C datasets. Figure 1 shows the overall workflow for identifying these DNA interacting hotspots, which we analyzed further to identify putative enhancer elements and their promoter partners. All three samples from the original Hi-C study (16) were used in our analyses [cell line GM06990 with REs HindIII and NcoI, as well as cell line K562 with HindIII (referred to as GM/HindIII, GM/NcoI and K562/HindIII, respectively)]. The original Hi-C study used a 1 Mb window size to uncover the 3D organization of human nuclear chromosomes. However, this resolution is far too coarse for studying regulatory elements, which requires single nucleotide resolution. To improve resolution for our purposes of identifying DNA interacting hotspots, we applied our genomic distribution-based analysis for identification of these specific genomic regions (24). Briefly, our algorithm first identifies clusters of Hi-C reads that are closer to each other than what the background geometric distribution dictates. We then labeled the resulting clusters as hotspots if their lengths on the chromosomes are longer than 99% of all clusters. We found that a hotspot is on average ~1 kb in length, and between 107 059 and 185 042 total hotspots were identified in each of the three samples.

The Hi-C method dictates that sequencing reads will start at or near the sites of the RE used in the experiment rather than the actual DNA–DNA interaction site. Therefore, the resolution of this method is limited to the distance between the genomic sites of the particular RE used for that study (Figure 2). To account for this shortcoming, we extended the length of the originally identified DNA interacting hotspots based on the estimated length between RE site positions on each human nuclear chromosome, while also allowing each nucleotide of an extended hotspot to represent the true interaction site. We found that on average an extended hotspot is 3–3.3 kb long (Table 1), indicating that our resolution has improved ~300-fold compared with the 1 Mb window size used in the original study.

Characterization of DNA interacting extended hotspots

We classified all extended hotspots based on human genome annotations and found that many of them are located within protein-coding genes, functional RNAs and tandem repeats, suggesting that some of the interaction hotspots may be involved in regulatory processes (Figure 3a–c). Interestingly, we observed that extended hotspots were located within 5–20% of total promoter regions (defined as the 500 bp upstream of protein-coding gene transcription start sites) of the human genome. This led us to speculate that some of the extended hotspots from our reanalysis of Hi-C data may actually reflect target promoters that are interacting with enhancer elements in the human genome.

Prediction of CEEs

To identify CEEs, we first considered extended hotspots that interact with a protein-coding gene promoter region(s) (defined as the 500 bp upstream of the annotated transcription start site). As shown in Table 2, 22–62% of the extended hotspots interact with a protein-coding gene promoter. The variation in promoter interactions is likely a consequence of the number of promoters that are covered by extended hotspots, which is influenced by

both the total sequencing depth in a particular sequencing library and the REs and cell types used in the Hi-C experiments. We next examined the enrichment of promoter-interacting extended hotspots in four activating histone modifications known to be associated with enhancer elements (H3K27ac, H3K4me1, H3K4me2 and H3K4me3), and a heterochromatic histone modification (H3K27me3) as a negative control (25,29). As expected, we found that promoter-interacting extended hotspots are enriched (permutation test, $P < 0.001$) in all four activating histone modifications but not with H3K27me3 (Figure 4a) when compared with the random background control. These results suggest that many of the promoter-interacting extended hotspots are human enhancer elements.

To further improve our confidence that we are detecting bona fide enhancer–target gene promoter interactions, we added an additional quality control step where we only retain promoter-interacting extended hotspots if their promoter interaction is supported by more than one read ($n > 1$) in the sequencing results (Supplementary Figure S6). This filtering step dramatically reduced the number of potential enhancer elements in all three samples. In fact, only 7.7–12.2% of the promoter-interacting extended hotspots were retained as potential enhancer elements (Table 2). This step likely reduced the number of false positives in our dataset, as we found it substantially increased the enrichment in the four enhancer-associated activating histone modifications (H3K27ac, H3K4me1, H3K4me2 and H3K4me3) in the remaining promoter-interacting extended hotspots (Figure 4b). Taken together, these results indicate that increased read support for the promoter-extended hotspot interactions is necessary for high-confidence identification putative enhancer elements and their targets from Hi-C experimental data.

The final filtering step in our pipeline to identify CEEs was to determine the enrichment of DHSs within the

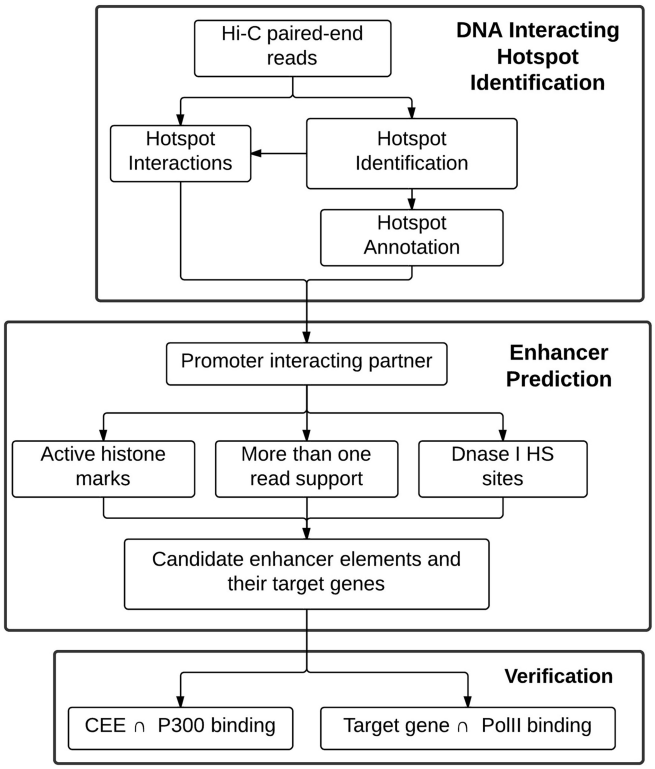

Figure 1. Genome-wide enhancer element identification workflow.

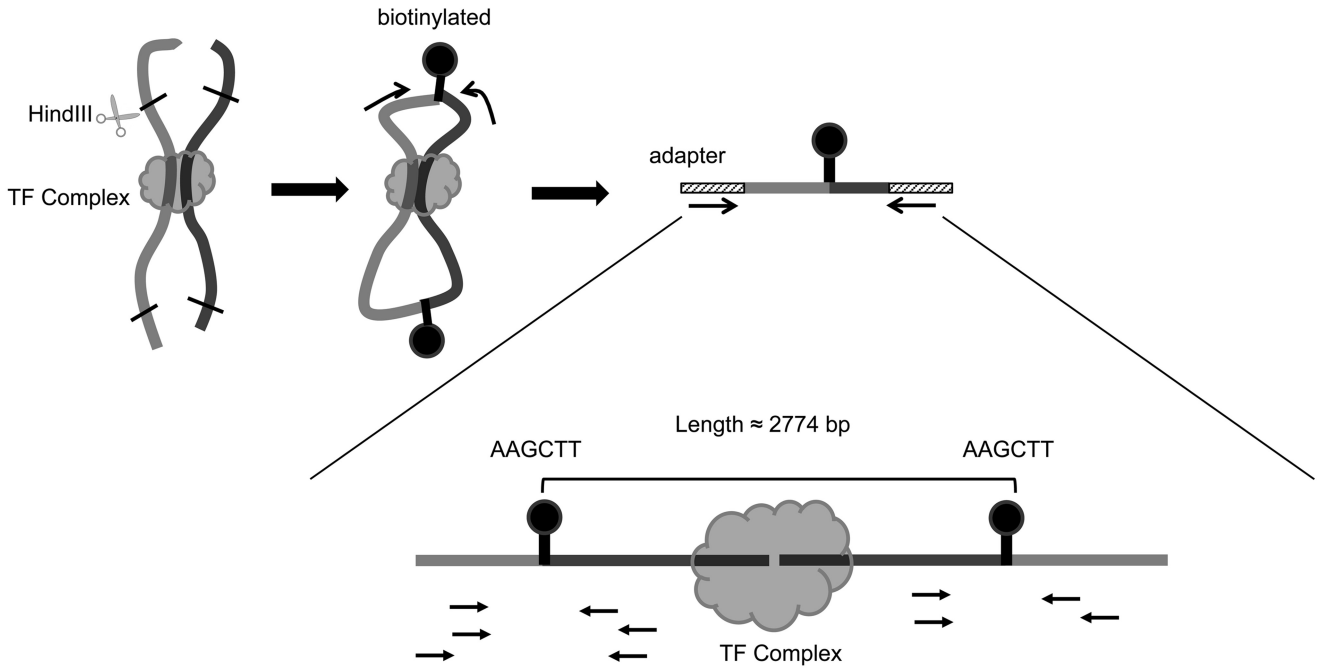

Figure 2. Identification of potential enhancer elements by our novel analysis pipeline requires an extension of regions that have been termed DNA interacting hotspots from the 3C data as depicted.

Table 1. Characterization of extended hotspots

Samples	GM/HindIII	GM/NcoI	K562/HindIII
# Raw reads (paired spots)	30 009 111	28 659 279	36 823 509
# Unique mapped pairs	18 728 707	18 891 283	21 744 849
Percentage of mapped to raw	62%	66%	59%
# Unique mapped single-end	37 457 414	37 782 566	43 489 698
# Clusters (merged by gap length)	4 973 281	5 076 539	6 247 694
Average cluster length (bp)	172.4	168.9	160.2
# Hotspots	107 059	166 990	185 042
Average hotspot length (bp)	1047.9	1007.8	964.1
# Extended hotspots	96 800	137 611	150 611
Average extended hotspot length (bp)	3065.8	3349.5	3282.1

GM = GM06990.

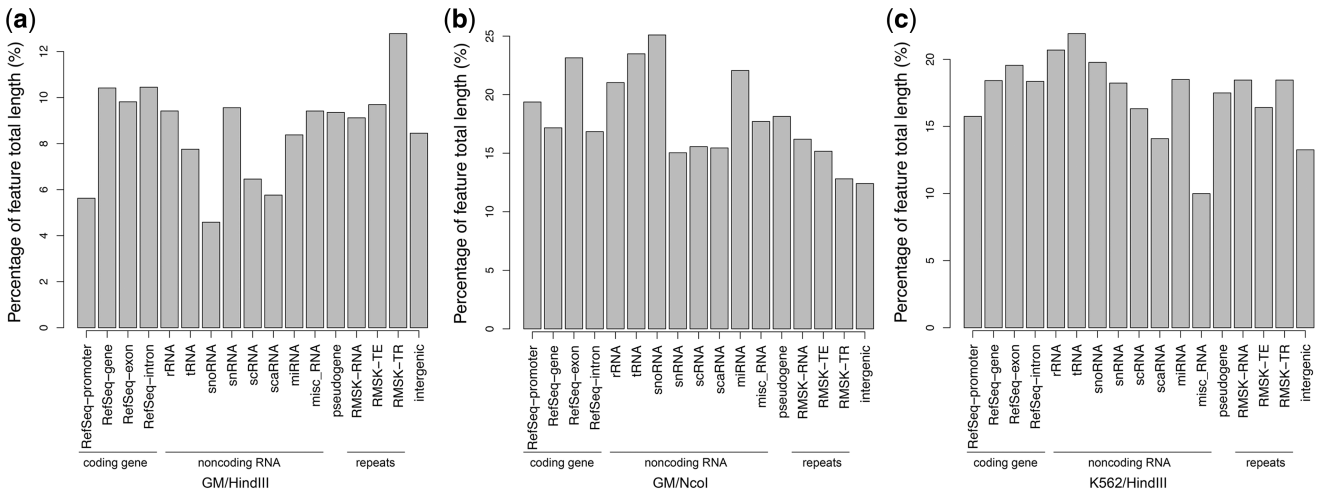

Figure 3. Functional annotation of extended hotspots for sample (**a**) GM/HindIII, (**b**) GM/NcoI and (**c**) K562/HindIII. Each bar (as labeled) represents the percent of total length for each genomic feature that overlaps with extended hotspots.

Table 2. Number of CEEs present after each filtering step

Filtering step	GM/ HindIII	GM/ NcoI	K562/ HindIII
Promoter partners	22 818	90 200	93 109
Strong interactions (>1 read)	1757	11 001	9955
Activating histone mark enrichment	928	5617	5814
DNase I HS sites	823	4809	5033

subset of high-confidence promoter-interacting extended hotspots (supported by >1 sequencing read) using previously published datasets (37,38). From this analysis, we found that the set of high-confidence promoter-interacting extended hotspots from all three original Hi-C experiments were enriched ($P < 0.001$) in DHSs (Figure 4c). The tendency of high-confidence promoter-interacting extended hotspots to co-localize with DHSs provides further evidence of the reliability of our analysis strategy to identify bona fide enhancer element–target promoter pairs in the human genome. In summary, the combination of these results has led us to incorporate all three of these analysis steps in our pipeline for genome-wide prediction of CEEs and their interacting target promoters in the human genome (see Supplementary Dataset 1 for the entire list).

CEEs and their target genes are enriched in binding activities associated with gene expression

To provide further evidence that our CEEs are bona fide enhancer elements, we examined the enrichment of p300 binding within these regions. We focused on p300 because it is a known enhancer-associated co-activator that mediates the regulation of target gene expression (39,40). We found that the CEEs from all three Hi-C experiments were enriched ($P < 0.001$) in p300 binding compared with a background control of all extended hotspots (Figure 5a). This enrichment in p300 binding within CEEs strongly suggests we have identified bona fide enhancers, and by using the Hi-C data in this analysis we also identify the gene promoter(s) that each element can target.

Enhancer elements generally activate gene expression through direct interaction with target promoters (40–42)

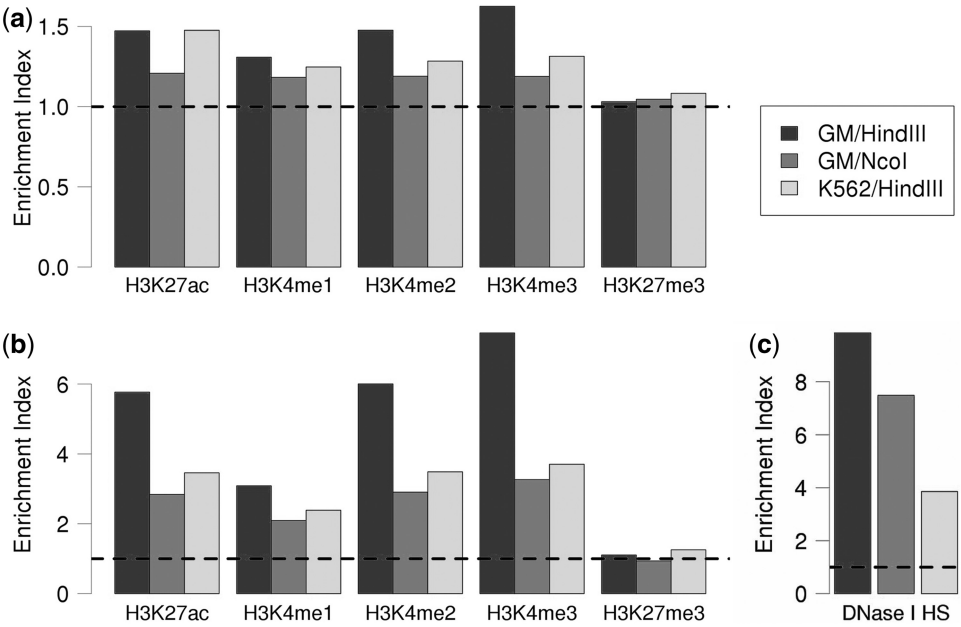

Figure 4. Potential enhancer elements are enriched for activating histone marks and DHSs. (**a** and **b**) Fold enrichment for activating (H3K27ac and H3K4me1–3) and repressive (H3K27me3) histone marks with (a) all CEEs that have a promoter partner, and (b) CEEs whose promoter partner is supported by >1 read. (**c**) Fold enrichment of DHSs in CEEs with a promoter interaction supported by >1 read and enriched in activating histone marks. The three samples are marked as follows: black bars, GM/HindIII; gray bars, GM/NcoI and light gray bars, K562/HindIII. Dashed line is expected value based on genomic control.

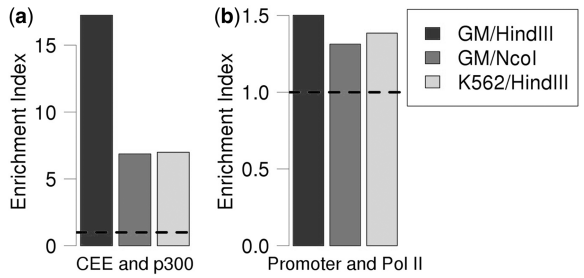

Figure 5. Potential enhancer elements are enriched for p300 binding, and their target genes are highly bound by Pol II. (**a**) P300 binding site enrichment in CEEs. (**b**) Pol II enrichment observed for the genes targeted by CEEs. Dashed line is the expected value based on a genomic control.

that results in increased RNA Pol II association. Therefore, we tested whether the CEE target gene promoters were enriched for Pol II binding. From this analysis, we found that CEE target promoter regions are >20% enriched ($P < 0.001$) for Pol II binding when compared with the promoters of all other protein-coding genes. These results suggest that transcription initiation is increased at promoters that are in contact with the CEEs compared with all other gene promoters in the human genome.

To further test if transcription is generally higher from target genes of the CEEs, we also investigated the enrichment of Pol II Ser2 phosphorylation, which marks elongating Pol II, within these loci for the K562 dataset using previously published data (GSM935547). Interestingly, we observed a 1.47-fold enrichment ($P < 0.001$) in Pol II Ser2 phosphorylation within target genes of our CEEs compared with all other protein-coding genes. These results indicate that the CEE–target gene interaction not only increases Pol II promoter binding, but also effects transcription elongation.

In summary, the consistent enrichment of the CEEs in p300 binding as well as their target genes with initiating and elongating Pol II strongly suggests that we have identified bona fide enhancer–target gene pairs by reanalyzing the previous Hi-C results. We also compared our CEEs with the enhancers predicted in the recently published study using 5C with primers designed to the ENCODE pilot project regions covering only ~1% of the human genome (23). We found that our CEEs overlap significantly with these enhancer elements compared with all extended hotspots lying within the ENCODE pilot region as a background control (Table 3). Thus, our method provides an important improvement over previous approaches for identifying human enhancer elements because we not only identify enhancers, but we also uncover their specific regulatory targets on a genome-wide scale.

Characterization of CEE-target gene interactions

We found that CEEs interact with 1.17–1.62 target gene promoters on average (Table 4), which is consistent with recent results (23) and suggest that human enhancer elements can interact pleiotropically. Additionally, we found that most target gene promoters interacted with multiple (1.17–2.36) CEEs (Table 4), suggesting the existence of enhancer redundancy in the human genome.

We also determined the interaction characteristics of CEEs and their target genes. We found that the vast majority of these interactions are intra-chromosomal (on the same chromosome), while fewer than 13% are inter-chromosomal (Figure 6a) with little read support for these latter associations (Supplementary Table S1). Interestingly, we found that >95% of the intra-chromosomal interactions occur within a range of 1 Mb (Figure 6b). In total, these results indicate that the majority of the CEEs that we have identified from the Hi-C data are in relatively close proximity to their target promoters.

CEEs and target genes are enriched in enhancer-associated motifs

To identify specific sequence motifs in the CEEs and their target promoters, we further searched for overrepresented sequences using HOMER (36). Not surprisingly, a quick search using a random genomic background yielded the recognition site of HindIII (AAGCTT) as top motif in the CEEs identified in by the original GM/HindIII Hi-C experiment (Table 5). These results suggest that as expected the Hi-C experimental approach identifies DNA interaction sites that are localized near restriction sites in the human genome (Figure 2). To minimize this bias for RE sites, we performed the motif searches with a background of all extended hotspots. As a result, we identified 38, 54 and 39 motifs from each experiment, including the binding motifs of known enhancer-associated transcription factor families such as Sp1, NRF1, E2F, GATA and ETS (Supplementary Dataset S2 and Table 6). Remarkably, we found that in all three CEE datasets, there was significant enrichment for the binding sequence of the E26 transformation-specific (ETS) family binding domain-containing proteins. These proteins act as transcription factors that bind to specific enhancers and promoters, and facilitate the assembly of transcription machinery to initiate gene expression (43,44). Thus, our CEEs are enriched in sequences known to bind enhancer-specific proteins.

Table 3. Comparison of the CEEs predicted using Hi-C and enhancer predictions in 5C (27)

Samples	Hi-C # CEEs	5C # Enhancers	# Intersects	*P* value of intersects
GM[a]/HindIII	19	87	1	0.1316
GM[a]/NcoI	37	87[b]	5	0.0001
K562/HindIII	137	119	9	<0.0001

CEEs shows the number of CEEs that is overlapped with the 5C primers along the ENCODE pilot regions.
P-value is calculated by permutation tests using the extended hotspots that overlap the ENCODE primer sets as background.
[a]5C study uses GM12878; Hi-C study uses GM06990.
[b]5C study uses only HindIII as the RE; here we are comparing using the GM/HindIII dataset.

CEEs are conserved within vertebrates

Functional elements are often under evolutionary selection because of their cellular function(s) (45). To study if the CEEs are under evolutionary selection, we investigated the conservation score in these elements across the mammalian clade [cons44way conservation score (46)] compared with their upstream and downstream flanking sequences. We found that the CEEs tend to be more conserved than their flanking regions ($P < 0.05$ for all datasets) (Figure 7a). In total, these results revealed that the CEEs that we have identified are under purifying selection in the human genome, suggesting that they are functional enhancer elements.

CEE target genes tend to display tissue-specific expression

Enhancer elements are known to function in a cell type–specific manner (10), so the expression profiles of their target genes are likely to display a similar pattern. To determine whether genes targeted by CEEs exhibit this cell type–specific tendency, we computed the Q statistic (34) for every human gene expressed in nine ENCODE cell types (see Methods for descriptions), and then compared CEE target genes with all other loci in the cell types (GM12878 or K562) most closely corresponding to those used in the original Hi-C experiment (GM06990 or K562). We found that genes targeted by the CEEs have significantly lower Q values, indicating that these loci are expressed in a cell type–specific manner. This is true for CEEs identified using all three Hi-C experiments ($P = 0.01$, 2.11e–05, 4.24e–16 for GM/HindIII, GM/NcoI and K562/HindIII, respectively). In total, all of our results suggest that we have identified thousands of bona fide enhancer–target gene interactions. A significant amount of future attention can now be focused on determining the biological functions and significance of these newly identified interactions in human cells.

DISCUSSION

The original Hi-C article suggested that this experimental approach could identify regulatory elements with better sequencing throughput, although this analysis was never performed. In this study, we show that with careful

Table 4. Characteristics of enhancer–target interactions

Samples	# CEEs	Average CEE interactions	# of target promoters	Average target promoter interactions	# of enhancer–promoter interactions
GM/HindIII	823	1.17	820	1.17	953
GM/NcoI	4809	1.62	3444	2.27	7757
K562/HindIII	5033	1.42	3032	2.36	7104

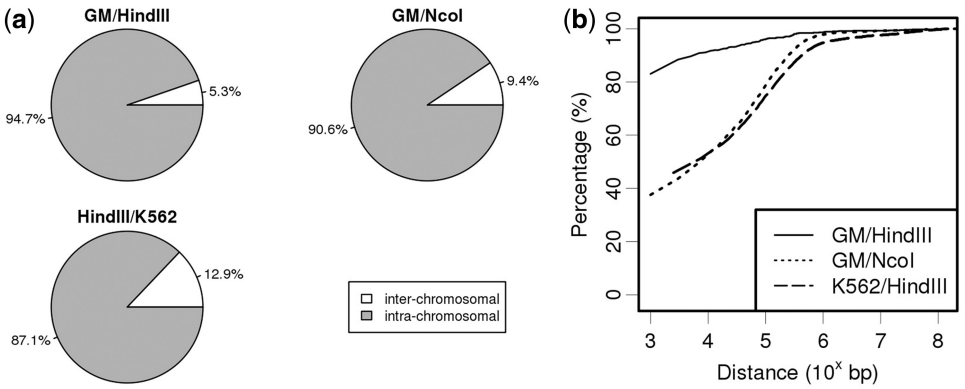

Figure 6. Characterization of CEE–target gene interaction distance. (**a**) The portion of inter- and intra-chromosomal CEE–target interactions for the three different Hi-C samples as denoted. (**b**) The distance distribution of intra-chromosomal CEE–target interactions.

Table 5. Top 3 most enriched motifs for all CEEs using the whole-genome as the background sequence in the K562/HindIII library

Motif	*P*-value	% of Targets	% of Background
	<1e–300	87.9	55.2
	<1e–258	84.5	80.7
	<1e–228	60.2	55.6

Table 6. Top 10 most enriched motifs in CEEs from the GM/NcoI library using extended hotspots as the background

Transcription factor (DNA binding domain)	Motif	*P*-value	% of Targets	% of Background
Sp1(Zf)		1e–134	38.5	23.6
NRF1(NRF)		1e–111	15.0	6.4
ETS(ETS)		1e–69	41.6	30.3
ELF1(ETS)		1e–64	58.2	46.8
GFY-Staf		1e–54	7.5	3.1
NRF1		1e–51	19.2	12.0
YY1(Zf)		1e–45	9.1	4.61
E2F		1e–45	30.2	22.0
GFY		1e–38	10.0	5.6
GABPA(ETS)		1e–32	77.3	70.1

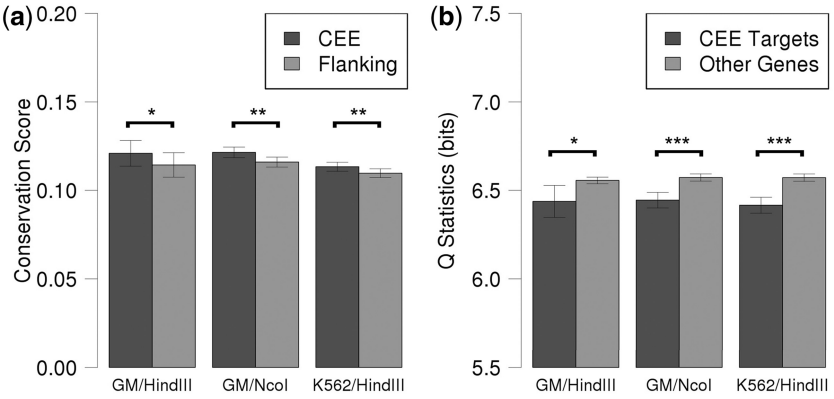

Figure 7. Potential enhancer elements are evolutionarily conserved, and their target genes are expressed in a cell type–specific manner. (**a**) The conservation score of CEEs (black bars) compared with similarly-sized flanking regions (gray bars) from the three different Hi-C experiments (as specified). (**b**) The Q statistic values for CEE target (black bars) compared with non-target (gray bars) genes from the three different Hi-C experiments (as specified). Error bars indicate s.e.m. Differences are statistically significant ($*P < 0.05$, $**P < 0.01$ and $***P < 0.001$, Wilcoxon rank-sum test).

analyses and comprehensive integration of publicly available functional genomic datasets, Hi-C data can be used to comprehensively identify enhancer–target gene interactions genome-wide. We first used a geometric distribution model to identify DNA interacting hotspots instead of using a sliding window to probe 1 Mb segments of the human genome. This change in analysis methods significantly improved our genomic resolution (~3.3 kb or 300-fold improvement). This increase in resolution is necessary for identifying the actual sequences of regulatory elements in the human genome. From this initial list of DNA interacting hotspots, we focused on intergenic sequences that interact with protein-coding gene promoters, and found these elements overlap significantly with enhancer-associated chromatin marks such as H3K27ac and H3K4me1 that have been previously used to identify enhancer elements (9,10). Interestingly, a recent study by Chepelev *et al*. used this property by combining Hi-C with H3K4me2 immunoprecipitation to identify enhancer–promoter interactions (22). However, this study focused solely on *cis*- interactions and did not examine other enhancer-associated epigenetic marks. Here, we used multiple chromatin marks as well as DHS datasets to identify thousands of CEEs in two human cell types with high confidence. We also uncovered that not all epigenetic marks are equal for these purposes. Specifically, we found that all four activating histone marks are enriched on the putative enhancers, but they demonstrate distinct levels of enrichment (Figure 4). In total, our analysis pipeline incorporates data for multiple histone modifications and DHSs, which increases confidence that bona fide enhancer elements are truly being identified.

In addition, our analysis is unique when compared with three other studies that were recently published. Specifically, Lan *et al*. (47) integrated histone modification data that overlapped sites enriched with reads from Hi-C experiments for the K562 cell line, and found 12 clusters of Hi-C sites with different combinations of histone modifications. However, their analyses were limited only to these overlapping regions and did not also interrogate all of the other relevant datasets as we have done here. Furthermore, their study only examined enhancer–promoter interactions on a specific subset of the human genome (GATA1/GATA2 target genes). In another recent study, 5C experiments were performed to study enhancer–promoter interactions. However, they focused entirely on the 44 ENCODE pilot genomic regions instead of performing a genome-wide analysis (23). This is because a global study of enhancer–promoter interactions is not feasible with the 5C protocol, as this approach requires the design of specific primers for a select group of targeted regions. ChIA-PET, another recently developed protocol that detects chromosomal interactions using high-throughput sequencing (22) was also used to study enhancer–promoter pairs. However, as pointed out by the developers of this method, their approach is different from unbiased approaches like Hi-C because it requires an antibody to a specific histone modification, protein, etc. Thus, this method will not detect any enhancers that are not in close proximity to the histone modification, protein, etc. being immunoprecipitated.

Our analyses revealed that unannotated long-range and inter-chromosomal enhancer–target gene interactions can be detected using Hi-C data. This is in strong contrast to previous studies of short-range enhancer–target gene interactions, namely predicting *cis*-targeted genes within a small fixed window (13) or by defining a variable but local transcriptional domain (48) around the identified enhancer elements. We found inter-chromosomal interaction to be much less frequent than both *cis* and *trans* intra-chromosomal interactions (Figures 6a and b). This may be because the inter-chromosomal and long-range interactions are underestimated due to the limited sequencing depth of the initial Hi-C experiments, or to these being less stable and/or transient interactions. Thus, we may identify more of these interactions with future Hi-C experiments with much greater sequencing depth.

We have also uncovered both one-to-one and multiple-to-multiple CEE-target interactions (Table 4).

These results reveal the extreme complexity of enhancer–target promoter relationships in the human genome. Interestingly, genes targeted by the same enhancer element could be co-regulated, competing or activated in different developmental stages or tissue types. Similarly, enhancer elements that target the same gene could also be cooperative or competing in maintaining gene expression homeostasis or for altering expression activities of the target gene. Comprehensive time-course studies with high read coverage (for better sensitivity) will be necessary to further elucidate the regulatory mechanisms behind each enhancer–target promoter interaction.

The Hi-C protocol has an inherent limitation for enhancer discovery as we have described (Figure 2). Specifically, we have revealed that the data from the Hi-C protocol actually detects RE sites around the bona fide DNA–DNA interaction regions. While increasing the read coverage is still essential for obtaining high-quality enhancer–promoter interaction data, it does not solve this particular limitation. To accommodate this lapse in resolution, we had to extend the identified DNA interacting hotspots in both directions, as we could not predict which direction to extend based on the Hi-C sequencing data alone. Thus, our analysis workflow is the first to allow confident prediction of enhancer–target promoter interactions from Hi-C data, and provides the framework for future studies that will use this approach for these same purposes.

Applying our analysis workflow to identify DNA interaction information from Hi-C, allows identification of candidate enhancers and their associated target genes. In the future, comparing datasets similar to the ones provided here with findings from GWAS and eQTL studies is likely to provide mechanistic insights into how many intergenic SNPs can be associated with a certain disease. In total, a comprehensive list of enhancer–promoter interactions is likely to significantly improve the resources available to future genetic studies focused on human disease.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–6 and Supplementary Datasets 1–2.

ACKNOWLEDGEMENTS

The authors thank all members of the Gregory and Wang lab for helpful comments on the manuscript.

FUNDING

Funding for open access charge: National Institute of General Medical Sciences [R01-GM099962 to L.S.W. and B.D.G.]; National Institute on Aging [U24-AG041689 to L.S.W., P30-AG010124 to L.S.W. and B.D.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Geyer,P.K., Green,M.M. and Corces,V.G. (1990) Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in *Drosophila*. *EMBO J.*, **9**, 2247–2256.
2. Lomvardas,S., Barnea,G., Pisapia,D.J., Mendelsohn,M., Kirkland,J. and Axel,R. (2006) Interchromosomal interactions and olfactory receptor choice. *Cell*, **126**, 403–413.
3. Banerji,J., Rusconi,S. and Schaffner,W. (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**, 299–308.
4. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
5. Levine,M. (2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**, R754–R763.
6. Wilber,A., Nienhuis,A.W. and Persons,D.A. (2011) Transcriptional regulation of fetal to adult hemoglobin switching: new therapeutic opportunities. *Blood*, **117**, 3945–3953.
7. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
8. Roh,T., Wei,G., Farrell,C.M. and Zhao,K. (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.*, **17**, 74–81.
9. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA*, **107**, 21931–21936.
10. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
11. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
12. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
13. Rödelsperger,C., Guo,G., Kolanczyk,M., Pletschacher,A., Köhler,S., Bauer,S., Schulz,M.H. and Robinson,P.N. (2011) Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Res.*, **39**, 2492–2502.
14. Lee,D., Karchin,R. and Beer,M.A. (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, **21**, 2167–2180.
15. Schoenfelder,S., Clay,I. and Fraser,P. (2010) The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.*, **20**, 127–133.
16. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
17. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
18. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
19. Tanizawa,H., Iwasaki,O., Tanaka,A., Capizzi,J.R., Wickramasinghe,P., Lee,M., Fu,Z. and Noma,K.I. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.*, **38**, 8164–8177.

20. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
21. Fullwood,M. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell Biochem.*, **107**, 30–39.
22. Chepelev,I., Wei,G., Wangsa,D., Tang,Q. and Zhao,K. (2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.*, **22**, 490–503.
23. Young,M.D., Willson,T.A., Wakefield,M.J., Trounson,E., Hilton,D.J., Blewitt,M.E., Oshlack,A. and Majewski,I.J. (2011) ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, **39**, 1–13.
24. Zheng,Q., Ryvkin,P., Li,F., Dragomir,I., Valladares,O., Yang,J., Cao,K., Wang,L.S. and Gregory,B.D. (2010) Genome-wide double-stranded rna sequencing reveals the functional significance of base-paired rnas in arabidopsis. *PLoS Genet.*, **6**, e1001141.
25. Raney,B.J., Cline,M.S., Rosenbloom,K.R., Dreszer,T.R., Learned,K., Barber,G.P., Meyer,L.R., Sloan,C.A., Malladi,V.S., Roskin,K.M. *et al.* (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.
26. ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of dna elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
27. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
28. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
29. Rosenbloom,K.R., Dreszer,T.R., Long,J.C., Malladi,V.S., Sloan,C.A., Raney,B.J., Cline,M.S., Karolchik,D., Barber,G.P., Clawson,H. *et al.* (2011) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, **40**, D912–D917.
30. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
31. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
32. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**, e15.
33. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
34. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
35. Shannon,C. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
36. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
37. Wu,C., Wong,Y.C. and Elgin,S.C. (1979) The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell, **16**, 807–814.
38. Gross,D.S. and Garrard,W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159–197.
39. Eckner,R., Ewen,M.E., Newsome,D., Gerdes,M., DeCaprio,J.A., Lawrence,J.B. and Livingston,D.M. (1994) Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.*, **8**, 869–884.
40. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
41. McKnight,S. and Kingsbury,R. (1982) Transcriptional control signals of a eukaryotic protein-coding gene. *Science*, **217**, 316–324.
42. Nolis,I.K., McKay,D.J., Mantouvalou,E., Lomvardas,S., Merika,M. and Thanos,D. (2009) Transcription factors mediate long-range enhancer-promoter interactions. *Proc. Natl Acad. Sci. USA*, **106**, 20222–20227.
43. Gutierrez-Hartmann,A., Duval,D.L. and Bradford,A.P. (2007) ETS transcription factors in endocrine systems. *Trends Endocrinol. Metab.*, **18**, 150–158.
44. Hollenhorst,P.C., McIntosh,L.P. and Graves,B.J. (2011) Genomic and biochemical insights into the specificity of ETS transcription factors. *Annu. Rev. Biochem.*, **80**, 437–471.
45. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
46. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
47. Lan,X., Witt,H., Katsumura,K., Ye,Z., Wang,Q., Bresnick,E.H., Farnham,P.J. and Jin,V.X. (2012) Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res.*, **40**, 7690–7704.
48. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 1–5.