RESEARCH ARTICLE   **Open Access**

An intriguing characteristic of enhancer-promoter interactions

Amlan Talukder[1], Haiyan Hu[1*] and Xiaoman Li[2*]

Abstract

Background: It is still challenging to predict interacting enhancer-promoter pairs (IEPs), partially because of our limited understanding of their characteristics. To understand IEPs better, here we studied the IEPs in nine cell lines and nine primary cell types.

Results: By measuring the bipartite clustering coefficient of the graphs constructed from these experimentally supported IEPs, we observed that one enhancer is likely to interact with either none or all of the target genes of another enhancer. This observation implies that enhancers form clusters, and every enhancer in the same cluster synchronously interact with almost every member of a set of genes and only this set of genes. We perceived that an enhancer can be up to two megabase pairs away from other enhancers in the same cluster. We also noticed that although a fraction of these clusters of enhancers do overlap with super-enhancers, the majority of the enhancer clusters are different from the known super-enhancers.

Conclusions: Our study showed a new characteristic of IEPs, which may shed new light on distal gene regulation and the identification of IEPs.

Keywords: Enhancers, Promoters, Enhancer clusters, Super-enhancers

Background

Enhancers are short genomic regions that can boost the condition-specific transcription of their target genes [1, 2]. They directly interact with the promoters of their target genes via chromatin looping to control the temporal and spatial expression of target genes [3–7]. Enhancers can be several dozens to a couple of thousand base pairs (bps) long and can be located in the distal upstream or downstream of their target genes [1]. Although the longest distance between enhancers and their targets validated by low-throughput experiments is about one mega bps (Mbps) [3, 4], recent high-throughput experiments showed that the distance can be larger than two Mbps in many cases [8, 9]. Because of such a long distance, it is still challenging to identify interacting enhancer-promoter

pairs (IEPs). In this study, an IEP refers to an enhancer-promoter pair that physically interacts, although such an interaction may or may not have any functional effect observed yet.

Many methods are available to identify enhancers. Early experimental studies identify enhancers by "enhancer trap", which has established our rudimentary understanding of enhancers in spite of its low-throughput and time-consuming nature [10, 11]. Early computational methods predict enhancers through comparative genomics, which are cost-effective but may produce many false positives. With the next-generation sequencing (NGS) technologies, enhancers are identified through a variety of experimental methods such as chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq), DNase I hypersensitive sites sequencing (DNase-seq), global run-on sequencing (GRO-seq), cap analysis gene expression (CAGE), etc. [12–17]. In ChIP-seq experiments, genomic regions enriched with H3K4me1 and H3K27ac modifica-

*Correspondence: haihu@cs.ucf.edu; xiaoman@mail.ucf.edu
[1]Department of Computer Science, University of Central Florida, FL 32816 Orlando, USA
[2]Burnett School of Biomedical Science, College of Medicine, University of Central Florida, FL 32816 Orlando, USA

tions are widely considered as active enhancers, and those with H3K4me1 and H3K27me3 modifications are taken as repressed enhancers [14]. In DNase-seq, distal open chromatin regions are considered as potential enhancers for gene regulation studies [18–21]. In GRO-seq and CAGE experiments, bidirectional transcripts are employed to identify active enhancers [15, 22, 23]. Correspondingly, computational methods based on NGS data are developed to predict enhancers on the genome-wide scale [14, 24–26]. These methods range from the early ones that are based solely on H3K4me3 and H3K4me1 ChIP-seq experiments to the later ones that are based on various types of epigenomic and genomic signals and have been applied to predict enhancers in different cell lines.

A large number of enhancers have been discovered so far. For instance, about 2900 enhancers from comparative genomics were tested with mouse transgenic reporter assay and stored in the VISTA database [27]. The Functional Annotation of the Mouse/Mammalian Genome (FANTOM) project identified 32,693 enhancers from balanced bidirectional capped transcripts [15]. This set of enhancers is arguably the largest set of mammalian enhancers with supporting experimental evidence [28]. There are also hundreds of thousand computationally predicted human enhancers, such as those predicted by ChromHMM and Segway [24, 25]. This set of enhancers represents the most comprehensive set of computationally predicted human enhancers currently available although they are much less reliable. In addition to individual enhancers, super-enhancers are identified, each of which is a group of enhancers in a genomic region that collectively control the expression of genes involved in cell-identities [29, 30].

Despite this relatively effortless discovery of enhancers, the identification of IEPs is still nontrivial. Early experimental procedures to identify IEPs are expensive and time-consuming [31, 32]. Recent Hi-C experiments hold a great promise to identify IEPs on the genome-scale, while are still are not cost effective in order to generate high-resolution Hi-C interactions [8, 9, 33]. To date, these experiments have only been carried out on a few cell lines or cell types. Although computational methods, from the early ones defining the closest genes as target genes, to the later ones considering the correlation of epigenomic signals in enhancers and those in promoters, to the current ones based on more sophisticated approaches [15, 19, 34–40], have shown some success in predicting enhancer target genes, they either do not consider or have low–performance on cell-specific IEP prediction [36]. Through these experimental and computational studies, megabase size self-interacting genomic regions called topologically associated domains (TADs) are also discovered in mammalian genomes, where IEPs usually fall within the TADs instead of crossing different TADs [41].

All existing computational methods almost always consider one enhancer-promoter pair at a time to determine whether they interact. We hypothesized that when two enhancers interact with a common target gene, these two enhancers may be spatially close to each other and may thus interact with all target genes of both enhancers. In other words, if two enhancers share a target gene, they may share all of their target genes as well. If this hypothesis is true, we should consider the interactions of multiple enhancers and multiple target genes simultaneously to predict IEPs, which may improve the accuracy of the computational prediction of the IEPs, especially that of cell-specific IEPs.
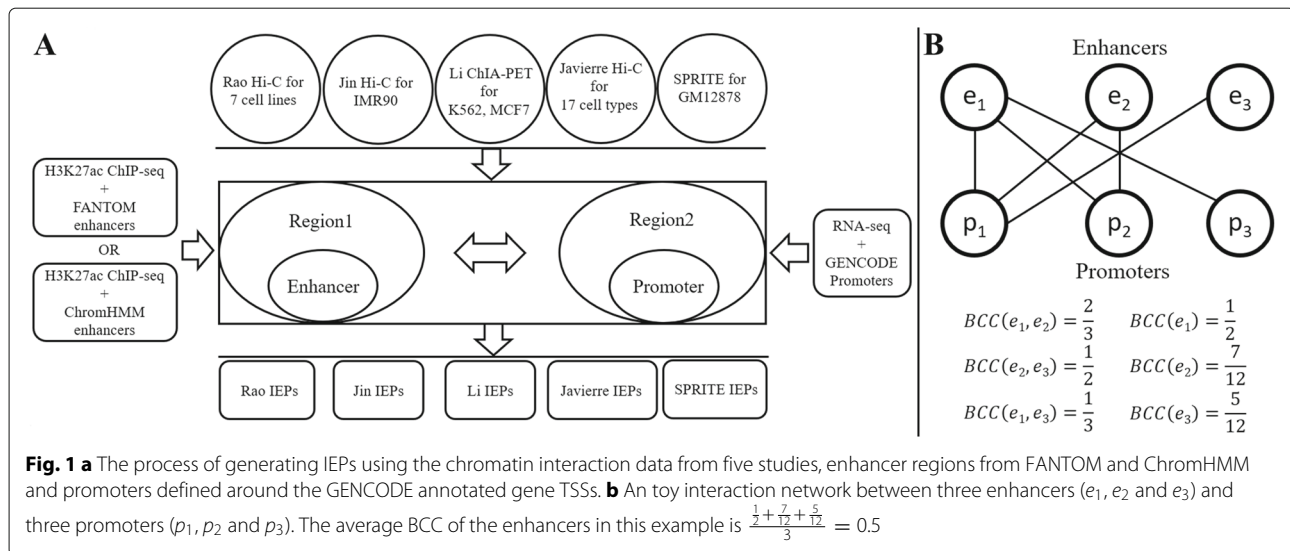
To test this hypothesis, we collected experimentally supported IEPs determined in five previous studies [6, 8, 9, 33, 42] and investigated how different enhancers may share their target genes in different cell lines and cell types (Methods). We considered both experimentally annotated enhancers from FANTOM and computationally predicted enhancers by ChromHMM in different samples [15, 24]. We observed that two enhancers are likely to either share almost all of their target genes or interact with two completely disjoint sets of target genes, in a cell line or a cell type. This observation implies an interesting characteristic of IEPs, which has not been considered by existing studies to predict IEPs. Our study may also shed new light on the underlying principles of chromatin interactions and facilitate the more accurate identification of IEPs.

Results

Two enhancers are likely to interact with either exactly the same set or two completely different sets of genes

In order to study IEPs, we calculated the bipartite clustering coefficient (BCC) of enhancers in each cell line or cell type, with two sets of enhancers and five sets of experimentally supported IEPs (Methods, Fig. 1a). BCC is commonly used to measure how nodes share their neighboring nodes in a bipartite graph. Note that every set of IEPs corresponds to a bipartite graph, where the enhancer set and the gene set correspond to the two disjoint sets of nodes, and their interactions correspond to the edges (Fig. 1b). The neighboring nodes of an enhancer are the target genes of this enhancer. With the goal to investigate how different enhancers share their target genes, the BCC is a perfect measurement, which can measure the percentage of target genes pairs of enhancers may share in a given set of IEPs (Fig. 1b). We observed that the BCC of enhancers was usually larger than 0.90. This indicates that when any pair of enhancers interact with one common target gene, both enhancers are likely to interact with all target genes of these two enhancers.

First, we studied the IEPs based on the looplists from Rao et al. [9], with the annotated FANTOM enhancers [15] and the GENCODE promoters [43] (Fig. 1a). We

Fig. 1 a The process of generating IEPs using the chromatin interaction data from five studies, enhancer regions from FANTOM and ChromHMM and promoters defined around the GENCODE annotated gene TSSs. **b** An toy interaction network between three enhancers (e_1, e_2 and e_3) and three promoters (p_1, p_2 and p_3). The average BCC of the enhancers in this example is $\frac{\frac{1}{2} + \frac{7}{12} + \frac{5}{12}}{3} = 0.5$

noticed that the BCC of enhancers was no smaller than 0.97 in all cell lines with enough IEPs (Table 1 and Supplementary Table S1). We further calculated the average BCC of the enhancers interacting with more than one gene. We found that their average BCC was no smaller than 0.96 in all cell lines, suggesting that two enhancers are likely to interact with either the same set or two disjoint sets of target genes. In other words, the target genes of any pair of enhancers usually are either the same or completely different.

To assess the statistical significance of the above observation, we studied the BCC of enhancers in randomly generated IEPs (Supplementary Table S1). These random lEPs were constructed using the same set of enhancers and promoters but randomized interactions, where we randomly chose promoters to interact with an enhancer so that the same enhancer had the same number of interactions as it had originally. We generated five different sets of random IEPs in this way with five different random seeds. With these random IEPs in the eight cell lines, we barely had a handful of enhancers sharing promoters with other enhancers in any cell line, suggesting that it is not by chance that multiple enhancers interact with a common set of target genes in the Rao et al.'s looplists (Supplementary Tables S1). For all four cell lines we could calculate the BCC, the BCC of enhancers was 0.51, 0.37, 0.33 and 0, respectively, which was much smaller than the BCC of enhancers in the above sets of real IEPs (*p*-value = 0, Supplementary Table S1). When we considered the BCC of enhancers interacting with multiple genes, the BCC values were no larger than 0.34 for random IEPs, while it was no smaller than 0.96 for the real IEPs, also suggesting that the observation that the

BCC of enhancers being close to 1 was not by chance (Supplementary Table S1).

Second, we studied IEPs defined by different cutoffs in seven cell lines (Methods). Compared with the IEPs from the above Rao et al.'s looplists, these IEPs defined by cutoffs were likely to include many more bona fide IEPs and more false positives as well. Under the cutoffs 30, 50 and 100, the BCC of enhancers in all seven cell lines except GM12878 was no smaller than 0.85, 0.89 and 0.92, respectively (Supplementary Table S1). Since GM12878 had a much higher sequencing depth than other cell lines, it was understandable that a stringent cutoff for other cell lines was still loose for GM12878. We thus tried the cutoffs 150, 200, 300, and 400 for GM12878. We noticed that the BCC of enhancers was 0.97 in GM12878 with the cutoff 400. Coincidently, the number of IEPs in GM12878 defined at this cutoff was similar to that in other cell lines defined at the cutoff 100 (Supplementary Table S1). We thus considered the cutoff 400 in GM12878 and the cutoff 100 for other cell lines. Note that in HMEC, HUVEC, KBM7 and NHEK, the BCC of enhancers was no smaller than 0.92 even under the cutoff 100. Moreover, the BCC of enhancers was increasing with more stringently defined IEPs, suggesting that the BCC of enhancers is close to 1 if it is not 1 (Supplementary Table S1).

In order to assess the statistical significance of the observed BCC of enhancers in IEPs from different cutoffs, similarly, we compared the above BCC of enhancers with that from randomly generated IEPs (Supplementary Table S1). Again, for every cutoff in every cell line, the BCC of enhancers for random IEPs was much smaller than the BCC of enhancers for real IEPs (*p*-value = 0). For instance, under the cutoff 50, the BCC of enhancers was no larger

Table 1 The BCC of enhancers and that of promoters are likely to be 1 in a cell line

	Cell line	IEPs	BCC of enhancers		% of total enhancers with multiple promoters and BCC > 0(E_1)	% of E_1 with BCC ≥ 0.9	BCC of promoters		% of total promoters with multiple enhancers and BCC > 0(P_1)	% of P_1 with BCC ≥ 0.9
			all	multiple			all	multiple		
Rao	GM12878	294 (2384)	0.97 (0.99)	0.96 (0.96)	17.47 (18.44)	87.5 (88.95)	0.97 (0.95)	0.95 (0.93)	19.35 (27.15)	91.67 (87.98)
	HELA	11 (37)	1 (1)	NA (1)	NA (12.5)	NA (100)	1 (1)	NA (1)	NA (11.76)	NA (100)
	HMEC	260 (2558)	0.97 (0.99)	0.96 (0.98)	15.42 (26.9)	90.32 (95.91)	0.96 (0.97)	0.91 (0.96)	13.97 (37.41)	88 (93.82)
	HUVEC	9 (95)	1 (1)	NA (1)	NA (10.47)	NA (100)	NA (1)	NA (1)	NA (19.35)	NA (100)
	IMR90	144 (554)	1 (1)	1 (0.99)	4.8 (8.98)	100 (100)	1 (0.99)	1 (0.98)	6.25 (9.35)	100 (93.1)
	K562	47 (638)	1 (1)	1 (1)	10.81 (17.35)	100 (100)	1 (1)	1 (1)	12.82 (27.92)	100 (100)
	KBM7	8	NA	NA	NA	NA	1	NA	NA	NA
	NHEK	0 (0)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
Jin	IMR90	1167 (5303)	0.9 (0.93)	0.84 (0.87)	34.86 (34.97)	62.16 (70.75)	0.77 (0.68)	0.73 (0.66)	37.66 (49.11)	52.98 (40.92)
Li	K562	2916 (33449)	0.8 (0.86)	0.75 (0.78)	30.98 (41.9)	50.92 (53.62)	0.86 (0.67)	0.75 (0.65)	26.43 (57.26)	44.13 (38.73)
	MCF7	2190	0.89	0.83	25.15	66.76	0.86	0.75	22.59	57.41

In the head row, "multiple" means the enhancers (or promoters) with multiple interacting promoters (enhancers). "All" means all enhancers (or promoters). When two numbers are in an entry, the number in the parenthesis is from the ChromHMM enhancers. BCC value is labeled "NA", when the number of corresponding enhancers (or promoters) is zero

than 0.78 for random IEPs, while the corresponding number was no smaller than 0.89 for real IEPs. If we considered only enhancers interacting with multiple target genes, the BCC of enhancers for random IEPs was about two times smaller than that for real IEPs. For instance, under the cutoff 50, the largest BCC value was 0.40 for random IEPs, while the smallest BCC value for real IEPs was 0.69.

Third, to see how this observation might change if we used the data from other labs or other experimental protocols, we studied the IEPs from four additional studies (Fig. 1a, Methods) [6, 8, 33, 42]. When we calculated the BCC of enhancers using the IEPs defined by Jin et al. themselves [33], it was 0.94. When considering the IEPs defined by Jin et al. based on the FANTOM enhancers and the annotated promoters by GENCODE, it was 0.90. In terms of the ChIA-PET datasets [6], it was 0.80 in K562 and 0.89 in MCF7 (Table 1). For the nine cell types from Javierre et al. [8], it was no smaller than 0.96 in all cell types. For the SPRITE data from Quinodoz et al. [42], it was 0.92, 0.92 and 1 for the cutoffs 30, 50 and 100, respectively (Supplementary Table S1). Although the IEPs were from different labs and from different experimental procedures, in all cases, the BCC of enhancers was larger than 0.80 and the majority of enhancers interacting with multiple promoters had their individual BCCs larger than 0.90, suggesting that the BCC of enhancers is likely to be 1 in these samples. Again, for the corresponding randomly generated IEPs for these datasets, on average, the BCC value was 0.48, much smaller than the corresponding ones from original IEPs, which was 0.96 (*p*-value = 0, Supplementary Table S1).

Finally, we repeated the above analyses with the ChromHMM enhancers instead of the FANTOM enhancers, because the number of the FANTOM enhancers was relatively small compared with the estimated number of enhancers and there were much more ChromHMM enhancers than FANTOM enhancers [24]. We had similar observations in all cases (Table 1, Supplementary Table S1). That is, the BCC of enhancers for IEPs in a cell line was close to 1. For instance, for IEPs based on the looplists, it was almost a perfect 1 in all cell lines. For the Hi-C data from Rao et al. under the cutoff 400 for GM12878 and 100 for other cell lines, it was no smaller than 0.93. For the Hi-C data from Jin et al. [33], it was 0.93. For the ChIA-PET data from Li et al. [6], it was 0.86. For the nine cell types from Javierre et al. [8], it was no smaller than 0.97. For the SPRITE data on GM12878 cell line [42], the BCC values were 0.9, 0.95 and 0.99 for the cutoffs 30, 50 and 100, respectively. In almost all cases, the majority of enhancers with multiple promoters had their individual BCCs larger than 0.90.

In summary, the BCC of enhancers was likely to be close to 1 for different sets of IEPs, data from different labs, different experimental protocols, different cell lines

and cell types, and different enhancer sets. The analyses based on IEPs from different cutoffs suggest that the BCC of enhancers is quite robust, although it is smaller when more loosely defined IEPs are used. It is close to 1 or becomes 1 when the IEPs are defined more and more stringently (with fewer false positive IEPs). These analyses suggest that what we observed may be an intrinsic property of enhancers. That is, if two enhancers interact with one common gene, they are likely to interact with each of their individual target genes.

Two target genes tend to interact with exactly the same set or two completely different sets of enhancers

We studied the BCC of promoters in each set of the aforementioned IEPs to see whether the similar hypothesis was true for the BCC of promoters. Our data showed that the BCC of promoters was likely to be 1 as well, although this was not so evident as the BCC of enhancers in certain cases.

First, we studied the BCC of promoters with IEPs based on the looplists [9]. It was close to 1 no matter whether we used the FANTOM enhancers or the ChromHMM enhancers (Table 1). We also calculated the BCC of promoters in randomly simulated IEP datasets, where we kept the same sets of enhancers and promoters but randomly selected enhancers to interact with promoters so that every promoter had the same number of interacting enhancers as it had in the original set of IEPs. The BCC of promoters was 0.52 at best in any cell line in these random datasets, suggesting that it was not by chance that the BCC of promoters was close to 1 in all cell lines (Supplementary Table S2).

Second, we studied the BCC of promoters based on lEPs defined with different cutoffs [9] (Supplementary Table S2). When we used the FANTOM enhancers, the BCC of promoters was often close to 1. For instance, with the cutoff 400 for GM12878 and the cutoff 100 for other cell lines, the BCC of promoters was no smaller than 0.91 in all the cell lines . For different cutoffs, it was usually no smaller than the BCC of enhancers, which was close to 1 in most cases. When we used the ChromHMM enhancers, however, it was not as large as those from the FANTOM enhancers. For instance, with the cutoff 400 for GM12878 and the cutoff 100 for other cell lines, the BCC of promoters varied from 0.64 to 0.91 in different cell lines. The BCC values got smaller with smaller cutoffs, which might be due to the much lower quality of the enhancers predicted by ChromHMM compared with the experimentally defined FANTOM ones.

Although the BCC of promoters was not as large as the BCC of enhancers when the ChromHMM enhancers were used, the actual BCC of promoters could also be close to 1. This was because the computationally predicted ChromHMM enhancers might result in predicting false

interactions and thus a low BCC of promoters. Moreover, the BCC of promoters was always increasing with more and more stringently defined IEPs. Although we did not observe that the BCC of promoters was close to 1 at the cutoff 100 we tried, it was indeed close to 1 when the looplists defined by Rao et al. were considered. In addition, the BCC of promoters for random IEPs in every cell line and under every cutoff was much smaller than that for the real IEPs, indicating that the observed much larger BCC of promoters was not by chance (Supplementary Table S2).

Third, we studied the BCC of promoters based on lEPs from other studies (Fig. 1a, Table 1 and Supplementary Table S2) [6, 8, 33, 42]. For the original IEPs from Jin et al., it was 0.11. However, when the IEPs were defined from the overlap of these original IEPs with the GENCODE promoters and the two types of enhancers, it was 0.77 and 0.68, respectively (Table 1). The low BCC of promoters for the original IEPs may be partially due to the promoters Jin et al. used, which had 11,313 promoters inferred by Jin et al., compared to the 57,820 promoters annotated by GENCODE [33]. In terms of the ChIA-PET data [6], when we used the FANTOM enhancers, the BCC of promoters was 0.86 in K562 and 0.86 in MCF7; when we used the ChromHMM enhancers [8], it was 0.67 in K562. ChromHMM did not have annotated enhancers in MCF7. For the nine cell types from Javierre et al., it was no smaller than 0.98 and 0.91 when the FANTOM enhancers and the ChromHMM enhancers were used, respectively. For the SPRITE data on the GM12878 cell line [42], the BCC values of promoters were no smaller than 0.89 and 0.71 in the IEPs defined with the FANTOM and ChromHMM enhancers, respectively. Overall, although it was not as large as the BCC of enhancers, because of the imperfectness of all these collected IEPs, and the fact that the majority of promoters interacting with multiple enhancers had their individual BCC larger than 0.90, and they were much larger than the corresponding BCC of promoters for random IEPs (Supplementary Table S2), the BCC of promoters was likely to be close to 1 as well. In other words, a gene usually interacts with all enhancers of another gene or interacts with a completely different set of enhancers from this second gene.

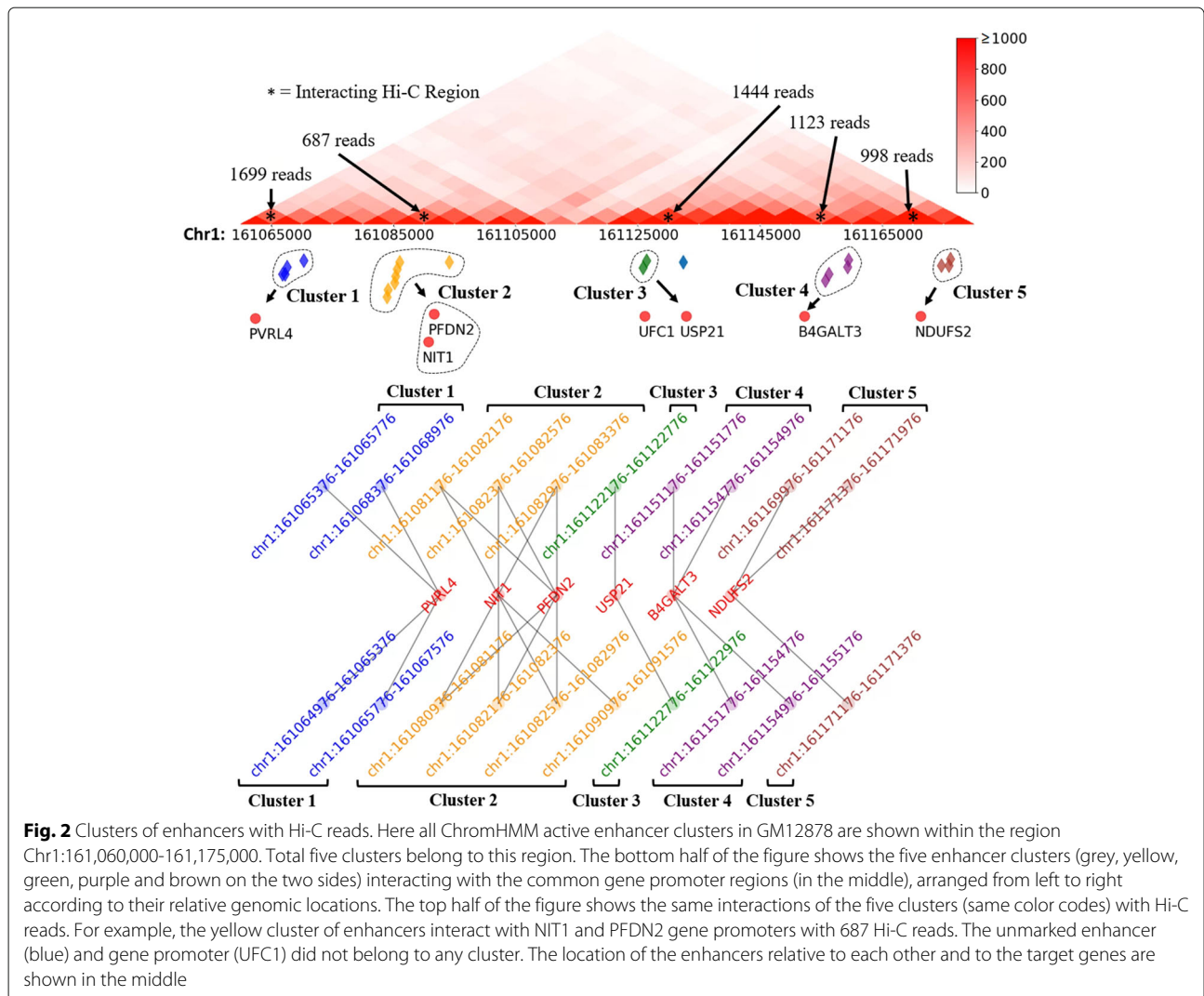### Enhancers form clusters that have special characteristics
Since the BCC of enhancers is close to 1, we can organize enhancers into clusters, where every enhancer in the same cluster is likely to interact wtih the same set of target genes. We thus built an enhancer graph by connecting enhancers that share at least one common target. We then grouped enhancers into clusters based on such a graph in each cell line (Methods, Fig. 2). Here we only considered the looplists and the IEPs obtained from the most stringent cutoff (400 in GM12878 and 100 in other cell lines) to

obtain enhancer clusters, as they were more reliable than other sets of IEPs.

We obtained 1 to 2134 clusters in different cell lines. The number of clusters in a cell line and across different cell lines varied dramatically, depending on the IEPs and the enhancers used (Supplementary Table S3). When the ChromHMM enhancers were used, there were many more clusters and 67% to 96% of all enhancers were included in clusters. When the FANTOM enhancers were used, fewer clusters were identified and about 16% to 67% of the total enhancers were in clusters. The average number of enhancers in a cluster varied from 2 to 5 in different cell lines. Enhancers in the majority of clusters interacted with only one gene, while on average, enhancers in 18.36% clusters interacted with at least two different genes.

We studied the distance between the consective enhancers in a cluster, the distance between their consecutive targets and the distance between enhancers and their target genes (Fig. 3 and Supplementary Table S4). On average, about 84% of the enhancers in a cluster were within 10 kbps. However, there was a small fraction of enhancers in a cluster that were more than 50 kbps away from each other. For instance, when the looplists and the FANTOM enhancers were considered, there were more than 8% enhancers in a cluster that were more than 50 kbps away from each other in GM12878, HMEC and IMR90. Although enhancers in a cluster were often close to each other, their distances to each other were not significantly smaller than the distances of random enhancer pairs (Supplementary Table S5, almost all p-values>0.2). In terms of the target genes, the majority of them were within 10 kbps, with a small fraction far from each other. For instance, when the looplists and the FANTOM enhancers were considered, we found 25.93%, 21.43% and 33.33% of the target genes of an enhancer cluster that were more than 50 kbps away from each other in GM12878, HMEC and IMR90, respectively. It was also worth pointing out that the enhancers in a cluster were normally consecutive and active enhancers while their target genes were normally not consecutive. In all cell lines, on average, more than 90% of the enhancers in a cluster were consecutive active enhancers while fewer than 17% of the target genes of an enhancer cluster were consecutive.

Since enhancers in a cluster were consecutive in the genome and the majority of enhancers in a cluster were close to each other, they seemed like the super-enhancers. We thus compared the enhancer clusters with known super-enhancers (Supplementary Table S6). On average, 29.77% of enhancer clusters overlapped with the corresponding super-enhancers in a cell line while the majority of enhancer clusters did not overlap with the known super-enhancers (Fig. 4a), which may represent new super-enhancers. On the other hand, a large proportion of

Fig. 2 Clusters of enhancers with Hi-C reads. Here all ChromHMM active enhancer clusters in GM12878 are shown within the region Chr1:161,060,000-161,175,000. Total five clusters belong to this region. The bottom half of the figure shows the five enhancer clusters (grey, yellow, green, purple and brown on the two sides) interacting with the common gene promoter regions (in the middle), arranged from left to right according to their relative genomic locations. The top half of the figure shows the same interactions of the five clusters (same color codes) with Hi-C reads. For example, the yellow cluster of enhancers interact with NIT1 and PFDN2 gene promoters with 687 Hi-C reads. The unmarked enhancer (blue) and gene promoter (UFC1) did not belong to any cluster. The location of the enhancers relative to each other and to the target genes are shown in the middle
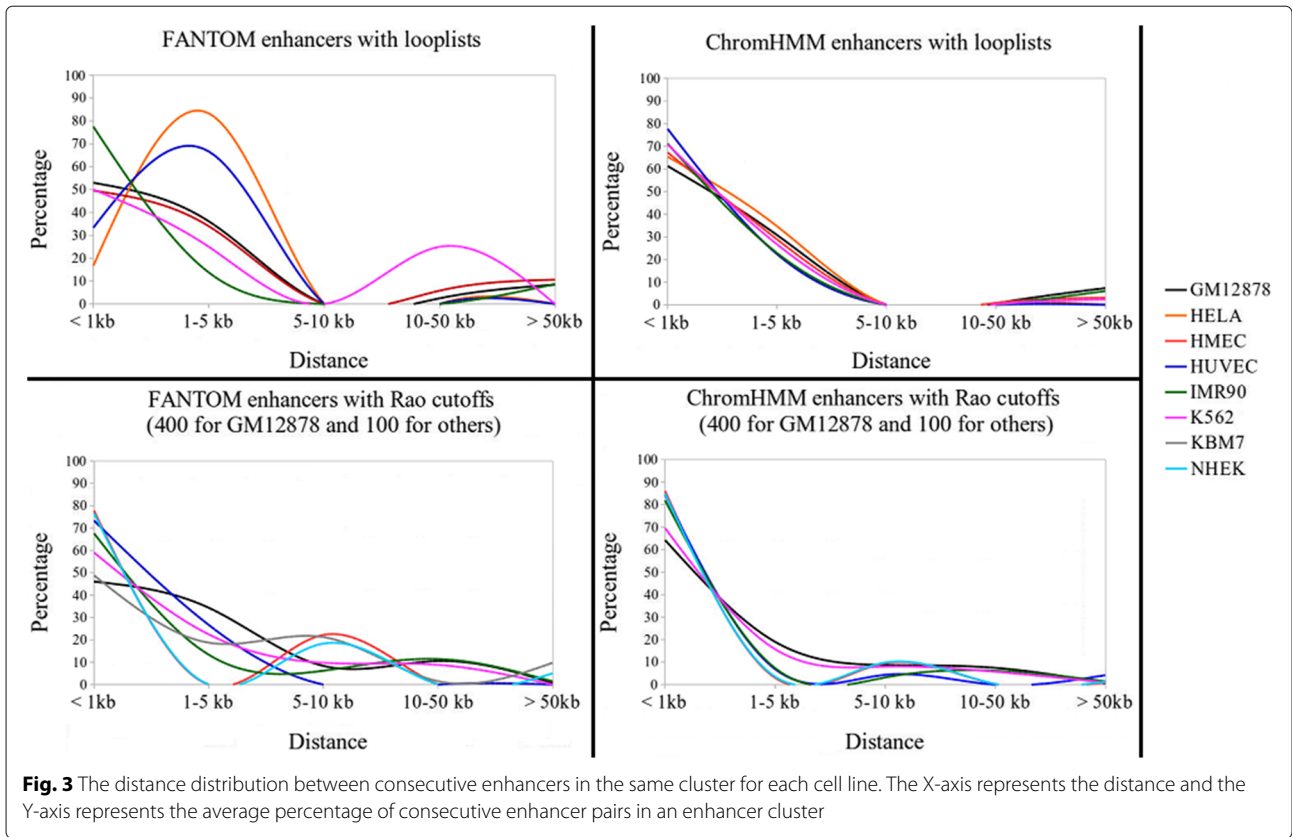
known super-enhancers did not overlap with the enhancer clusters in the corresponding cell lines (Fig. 4b). Interestingly, when a super-enhancer overlapped an enhancer cluster, more than 80% of the genomic regions that contain all enhancers in this enhancer cluster were within this super-enhancer.

We also studied how the enhancers in a cluster located relative to a TAD (Supplementary Table S7). The enhancers in a cluster were usually within the same TAD, with no smaller than 98.08% of enhancers in a cluster within a TAD in every cell line, independent of IEPs and enhancers used. In most cell lines, for all clusters, all enhancers in a cluster were within a TAD. The slight deviation from the 100% in certain cases may be due to the imperfectness of the IEPs, enhancers, and TADs, mostly due to the computationally predicted enhancers, as the percentage was always 100% in almost all the cell lines when the FANTOM enhancers were used.

We studied how the enhancer clusters were shared by different cell lines as well (Supplementary Table S8). That is, for an enhancer cluster in a cell line, how likely was the same cluster identified in another cell line. We found that on average no more than 12% enhancer clusters were identified in two cell lines. Moreover, the percentage was smaller when the looplists were used than when the stringent cutoffs were used to define IEPs, implying that the looplists were too strict to include many bona fide IEPs. The small percentage of shared enhancer clusters suggested that most enhancer clusters were cell-specific, which is consistent with the properties of super-enhancers [29, 30].
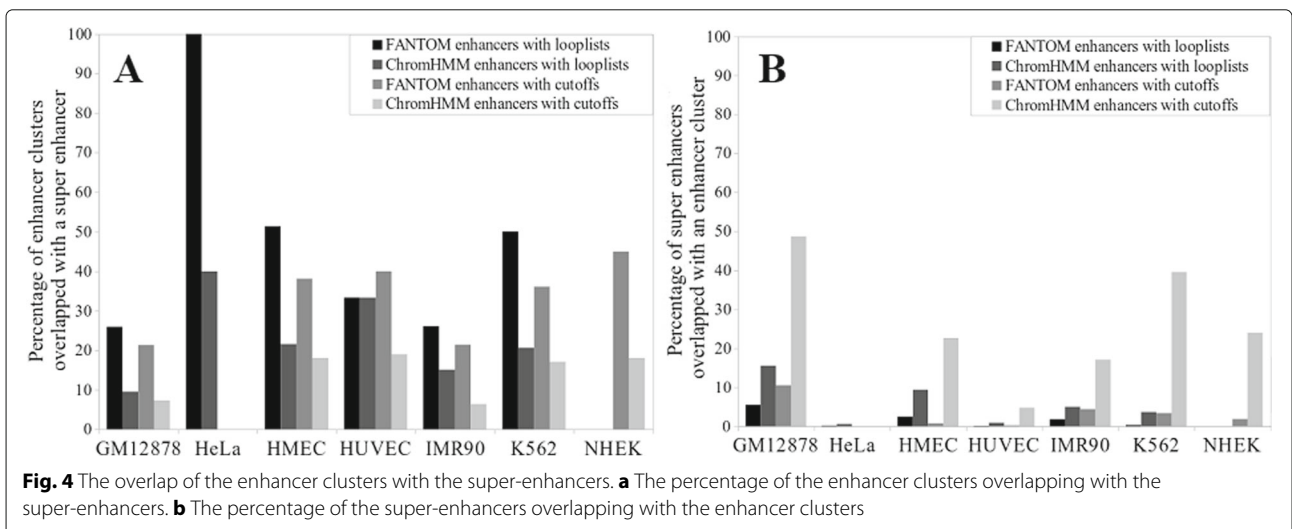
Discussion

We observed that two enhancers either do not share any target gene or share almost all of their target genes. This observation was true when different sets of IEPs, two

Fig. 3 The distance distribution between consecutive enhancers in the same cluster for each cell line. The X-axis represents the distance and the Y-axis represents the average percentage of consecutive enhancer pairs in an enhancer cluster

sets of enhancers, and a variety of cell lines and cell types were considered. Moreover, the BCC of enhancers became closer and closer to 1 when the criteria to define IEPs became more and more stringent. In addition, the same observation did not hold to be true for randomly generated IEPs. These analyses suggested that the BCC of enhancers in a cell line or a cell type was likely to be close to 1 if it is not 1.

Similarly, we observed that two promoters were likely to interact with either the same set of enhancers or two disjoint sets of enhancers. This observation about promoters was not as evident as that about enhancers. However, it was pervasive in all cases when the FANTOM enhancers were used. It was also evident when the looplists and the IEPs defined by the most stringent cutoffs were used. Although it seemed not compelling

Fig. 4 The overlap of the enhancer clusters with the super-enhancers. **a** The percentage of the enhancer clusters overlapping with the super-enhancers. **b** The percentage of the super-enhancers overlapping with the enhancer clusters

Talukder *et al. BMC Genomics*   (2021) 22:163

Page 9 of 13

when the ChromHMM enhancers and the sets of IEPs that were defined with loose criteria were used, this might be due to the imperfectness of enhancers and IEPs we had. More importantly, the fact that the BCC of enhancers was close to 1 implied that the BCC of the promoters should be close to 1 as well based on the definition of the BCC.

The BCC of enhancers being close to 1 suggested that enhancers form clusters to interact with the target genes. As shown above, these clusters are different from the known enhancer clusters such as super-enhancers, although they do overlap in certain regions. Enhancers in the clusters here were likely to interact with the same set of genes, while enhancers in a super-enhancer do not necessarily interact with multiple target genes. Moreover, the enhancers in a cluster here could be far from each other while the enhancers in a super-enhancer are quite close to each other.

The BCC of enhancers was not 1 sometimes, which implied that when a group of enhancers interacts with a set of target genes, the majority of target genes interact with each enhancer in this group while the rest interact with only a subset of enhancers in this group. We called the former the fully shared target genes and the latter the partially shared target genes. The percentage of the partially shared target genes by a group of enhancers varied from 0% to 6.57%. We compared these two types of target genes in terms of TAD, tissue specificity, and correlations with the enhancers, with the IEPs from the looplists and the IEPs from the most stringent cutoff (400 in GM12878 and 100 in other cell lines) (Methods). We did not observe any difference between the two types of target genes.

In practice, several aspects may prevent the BCC of enhancers and the BCC of promoters from being 1. First, the resolution of the interaction data prevents from obtaining accurate IEPs. The two interacting regions in the interaction data are often long, which is around 5 kbps in most of the cases we studied. We defined IEPs by overlapping enhancers and promoters with pairs of interacting regions, which might be prone to errors, given the fact that many known enhancers were much shorter [44, 45]. Second, the IEPs defined imperfectly might have produced "false" interactions and thus decreased the BCCs. Third, the enhancers were not perfectly defined either. The FANTOM enhancers are still far from complete while the computationally predicted ChromHMM enhancers may contain many "false" enhancers.

We also studied the functional similarities between the targets of enhancers in the same clusters. With the GREAT tool [46], we found the cluster targets associated with DNA packaging complex, DNA binding, nucleosome, immune response etc. (Supplementary Table S9, *p*-value<1e-5). We measured the sequence similiarity of enhancers within clusters in a cell line as well (Methods). We found that the pairs of enhancers in the same clusters

did not share more sequence similarity compared with enhancer pairs randomly chosen in the same cell lines (Supplementary Table S10, *p*-value>0.5).

There are other measurements to study bipartite graphs. We chose BCC because we intended to investigate how enhancers (promoters) shared their target genes (enhancers). In this sense, the BCC value perfectly reflected what we hoped to measure. In the future, we may explore other measurements to study other characterisitcs of IEPs. Moreover, we focused on enhancers interacting with multiple targets. There is no doubt that a proportion of enhancers only interacting with individual target genes. These enhancers and their target genes were not considered here, as they did not share target genes with each other. In the future, the characteristics of these enhancers may be worth studing as well.

Conclusion
In a cell line or cell type, both active enhancers and active promoters form their own clusters. When an enhancer interacts with a promoter, consistent with the transcriptional factories proposed previously [47, 48], almost all enhancers in the same enhancer cluster interact with almost all promoters in the corresponding promoter cluster. It is thus important to consider the relationship among enhancers and among promoters when studying their interactions, which may help improve our understanding of the distal gene regulation and the chromatin structures.

Methods
Enhancers and promoters
To study IEPs, we considered two sets of enhancers (Fig. 1a). The first set contained the 32,693 enhancers annotated by FANTOM, which had been obtained from the balanced bidirectional capped transcripts [15]. The second set was the computationally predicted enhancers by ChromHMM [24] in the following seven cell lines: GM12878, HMEC, HUVEC, K562, NHEK, IMR90 and HeLa. ChromHMM is widely used to partition genomes into different functional units including enhancers.

The FANTOM enhancers are not cell-specific, while the ChromHMM predicted enhancers are specific for the seven different cell lines mentioned. We thus defined cell-specific "active" FANTOM enhancers, by overlapping the enhancers with the H3K27ac ChIP-seq peaks in the corresponding cell lines obtained from the Encyclopedia of DNA Elements (ENCODE) project [49]. When there was no H3K27ac ChIP-seq data available for a cell line such as KBM7, we considered the enhancers that overlapped with the chromatin interacting anchors in this cell line as "active" enhancers [9].

We used the gene transcriptional start sites annotated in the GENCODE V19 [43] to define promoters. The upstream region of 1 kbps to the downstream region of

100 bps of each transcriptional start site was considered as a promoter. In total, we obtained 57,820 promoters in the human genome. To obtain cell-specific active promoters, we considered the available RNA-Seq data in different cell lines (GM12878, HeLa, HUVEC, IMR90, K562 and NHEK) [43]. In a cell line, a promoter was considered "active" if the corresponding gene had at least 0.30 reads per kbps of transcript per million mapped reads with the irreproducible discovery rate of 0.1, similarly as previously [37, 50]. For samples without RNA-Seq data (HMEC and KBM7), all promoters were considered as active promoters [50].

IEPs from five studies

To learn new characteristics of IEPs, we collected experimentally supported IEPs from five previous studies [6, 8, 9, 33, 42] (Fig. 1a). These IEPs arguably represent the intra-chromosomal chromatin interactions defined with the highest resolutions by the corresponding techniques. The first set of IEPs was from the Hi-C dataset GSE63525 in the Gene Expression Omnibus (GEO) database, where Rao et al. extracted significant intra-chromosomal chromatin interactions called "looplists" in the following eight cell lines: GM12878, HeLa, HMEC, HUVEC, IMR90, K562, KBM7 and NHEK [9]. These looplists were defined with stringent criteria and were most likely to be true pairs of interacting genomic regions, each of which was about 5 kbps long (Supplementary Table S1). In every cell line, we overlapped each looplist with the aforementioned two sets of active enhancers and with the annotated active promoters to obtain IEPs. In other words, an obtained IEP consisted of an enhancer and a promoter, where the enhancer overlapped with one region specified in a looplist and the promoter overlapped with the other region specified in the same looplist. Since we had two sets of enhancers, we obtained two sets of IEPs for each of the eight cell lines (Fig. 1a). Note that the enhancer and promoter in an IEP are always from the same chromosome.

The number of IEPs obtained from the above looplists was small, especially when we considered the FANTOM enhancers (Supplementary Table S1). The reason might be, the criteria Rao et al. used to define looplists was quite stringent and many true interacting genomic regions might therefore be missed [50]. To capture more IEPs in these cell lines, we also defined alternative sets of IEPs with three cutoffs: 30, 50, and 100, from the normalized Hi-C contact matrices with 5 kbps resolution defined by Rao et al. [9]. Given a normalized read cutoff, say x, if an enhancer-promoter pair overlapped with a pair of interacting genomic regions that were supported by at least x normalized Hi-C reads, we considered this EP-pair as an IEP. The cutoff 30 was used since this cutoff was likely to include of almost all known IEPs in K562 and IMR90 from other studies [6, 33] without allowing too many false

positives [50]. The two other cutoffs were used to see how the observed enhancer characteristics may change with more stringent cutoffs. Intuitively, the larger the cutoff was, the more likely that the two genomic regions interacted. Based on our previous studies [38, 51], we believed that the majority of the IEPs defined by these cutoffs in the eight cell lines except HeLa and GM12878 were likely to be bona fide IEPs and considered the IEPs defined by the cutoff 100 as highly reliable. We could not define IEPs in HeLa by cutoffs because Rao et al. did not provide a Hi-C contact matrix in HeLa. Since the sequencing depth was much higher in case of GM12878 than that in other seven cell lines, we considered the IEPs defined by the cutoff 400 in GM12878 as highly reliable after testing different cutoffs.

We also obtained 57,578 IEPs in IMR90 from another Hi-C study [33]. To our knowledge, this was the only Hi-C dataset for human samples with a comparable sequencing depth as that in GSE63525. In this study, Jin et al. defined active enhancers with H3K4me1 and H3K27ac ChIP-seq peaks and active promoters with H3K4me3 ChIP-seq peaks together with the known genes from the University of California, Santa Cruz genome browser. In addition to using the original IEP dataset which was provided in the hg18 version [33], we also converted it into the hg19 version and overlapped with the aforementioned enhancers and promoters used in this study to define IEPs.

We used the IEPs defined by the ChIA-PET experiments in K562 and MCF7 as well for this study [6]. ChIA-PET datasets for other cell lines have much lower sequencing depth than the two cell lines used here. Using the interacting regions in these datasets we found 2,923 and 2,190 IEPs with the FANTOM enhancers for K562 and MCF7, respectively. When we considered the ChromHMM enhancers, there were 33,598 IEPs for K562. There were no ChromHMM enhancers available in MCF7.

We obtained additional IEPs based on active enhancer and promoter links defined by Javierre et al. from promoter capture Hi-C experiments in nine cell types (Supplementary Table S1 in [8]). Javierre et al. did the experiments on seventeen primary cell types while the active enhancer and promoter links were provided for nine cell types. Each link defined a pair of interacting regions, with the average length of 5,709 and 8,599 bps, respectively. Since Javierre et al. did not explicitly specify the enhancers and promoters, we overlapped these links with the two sets of enhancers and the GENCODE promoters to define two sets of IEPs. In total, we obtained 20,764 and 607,274 IEPs with FANTOM and ChromHMM enhancers, respectively.

Finally, we considered the chromatin interaction data detected using a newly developed method named "SPRITE" by Guttman lab [42]. This dataset was

downloaded from GEO database of NCBI with the accession number GSE114242. Among the available SPRITE datasets, the only human dataset was on GM12878 cell line with the lowest resolution of 25 kbps. We considered this dataset with three different read cutoffs; 30, 50 and 100 to obtain IEPs.

We applied a distance filter on all the IEP sets found above. For every IEP, if the distance between the corresponding enhancer and promoter is less than 2.5 kbps, we filtered that IEP out from our analysis. The number of filtered IEPs for all the datasets are shown in Supplementary Table S1.

Other data used
Rao et al. annotated chromatin contact domains in each of the eight cell lines [9]. We downloaded these domains from GSE63525 and considered them as the topologically associating domains (TAD)s. We also downloaded the annotated TADs in IMR90 by Dixon et al., which were generated by the same lab that generated the Jin et al. data [41].

We downloaded the super-enhancers in GM12878, HeLa, HMEC, HUVEC, K562 and NHEK from dpSuper database. We could not find the known super-enhancers in KBM7. The super-enhancers in a cell line were compared with the clusters of enhancers that interact with the same set of target genes in the same cell line identified in this study.

BCC (Bipartite clustering coefficient)
All IEPs in a cell line form a bipartite graph, where the enhancers on one side connect with the target genes on the other side. We thus applied the BCC [52] to characterize how enhancers share their target genes and how genes share their enhancers (Fig. 1b).

For a pair of enhancers (or a pair of genes), say u and v, their BCC is defined as,

$$BCC(u, v) = \frac{|n(u) \cap n(v)|}{|n(u) \cup n(v)|}$$

Here $n(u)$ and $n(v)$ are the set of genes (or enhancers) interacting with u and v, respectively. Intuitively, if u and v are a pair of enhancers, $BCC(u, v)$ measures the percentage of target genes both u and v interact with among all of their target genes. Similarly, if u and v are a pair of genes, $BCC(u, v)$ measures the percentage of enhancers both u and v interact with among all the enhancers they interact with. Correspondingly, the BCC of an individual enhancer (or gene), say u, is defined as,

$$BCC(u) = \frac{\sum_{v \in n(n(u)), v \neq u} BCC(u, v)}{|n(n(u))| - 1}$$

Here $n(n(u))$ is the set of enhancers (or genes) that share at least one target gene (or enhancer) with u. Under a

given condition, for all the enhancers (or target genes) sharing at least one target gene (or an enhancer) with the other enhancers (or target genes), we averaged their individual BCCs to obtain the BCC of enhancers (or target genes) under this condition.

Clusters of enhancers that interact with the common set of genes
We built an enhancer graph, with nodes representing enhancers and edges representing pairs of enhancers interacting with at least one common target gene. We applied the Bron-Kerbosch algorithm [53] to this graph to find all maximal cliques. Enhancers in a clique formed a cluster of enhancers that interact with the same set of genes (Fig. 2). Different clusters may share enhancers.

Statistical tests
To assess the statistical significance of the observed BCC values in a given set of IEPs, we compared the observed BCC of enhancers (promoters) in this original set of IEPs with that in random IEPs, respectively. In every comparison, the BCC values of individual enhancers (promoters) that interacted with multiple promoters (enhancers) from the original IEPs were pooled together and compared with those from the random IEPs, respectively. In brief, we calculated the frequency of the enhancers (promoters) with BCC >0.9 in the random IEPs as the Binomial probability parameter (p). Assume there are n such individual enhancers in the original IEPs and k of them have their BCC >0.9, the p-value is calculated using the following formula.

$$p - \text{value} = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1 - p)^{n-i}$$

Additional analyses
To assess whether enhancers within a cluster have more sequence similarity to each other, we aligned every pair of enhancers within a cluster for every cluster in a cell line. We then measured the similarity of a pair of enhancers as the percentage of identities in the corresponding alignment [54]. In this way, we obtained the similarity scores of pairs of enhancers with clusters in a cell line. Similarly, we obtained the similarity scores of pairs of enhancers that are randomly selected in the same cell line. We then compared the two sets of similarity scores by the Mann-Whitney test [55]. To assess whether enhancers in a cluster tend to be close to each other in a cell line, We also compared the relative distance of pairs of enhancers within clusters in a cell line with pairs of enhancers randomly chosen in the same cell line by the Mann-Whitney test. In addition, we measured the function similarity of enhancers in clusters by the GREAT tool [46], with the

Talukder *et al. BMC Genomics*   (2021) 22:163

Page 12 of 13

input of the targets genes of enhancers in clusters in a cell line. This gave us the significant functional terms (*p*-value<1e-05) associated with the target genes of the enhancer clusters.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07440-5.

Additional file 1: Table S1. BCC statistics for enhancers.
Table S2. BCC statistics for promoters.
Table S3. Clusters of enhancers. An enhancer shares at least one promoter target with all the other enhancers in its cluster.
Table S4. The distance between each consecutive enhancer pairs in an enhancer clusters are shown in the left columns of the table. The right columns show the distance between each consecutive target pairs of the targets of the enhancers in an enhancer cluster.
Table S5. The average distance between each enhancer pairs in enhancer clusters and in randomly chosen enhancers that do not belong to any cluster.
Table S6. Overlap between enhancer clusters and super-enhancers.
Table S7. Average percentage of enhancers in the same clusters mapped in a common TAD or TAD gap. On average, almost all of the enhancers in a cluster were found to be located within the same TAD or TAD gap.
Table S8. Percentages of common enhancer clusters between two cell lines.
Table S9. Gene ontology enrichment analysis for the target genes of the enhancer clusters.
Table S10. The average sequence similarity score among the enhancers in a cluster and in randomly chosen enhancers that do not belong to any cluster.

Availability of data and materials
The FANTOM enhancers were downloaded from FANTOM5 Human Enhancer Selector (http://slidebase.binf.ku.dk/human_enhancers/results). The ChromHMM enhancers for GM12878, HMEC, HUVEC, K562 and NHEK cell lines were downloaded from the ENCODE composite track (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeBroadHmm/) of UCSC Genome Browser. The ChromHMM enhancers for HeLa and IMR90 cell lines were downloaded respectively from the ENCODE Genome Segmentation track of UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/) and Chromatin state model based on imputed data (25 state, 12 marks, 127 epigenomes) (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/E017_25_imputed12marks_mnemonics.bed.gz). The gene transcripts to define promoters were downloaded from the "Comprehensive Gene Annotation" GTF file Gencode version 19 (GRCh37.p13). The Hi-C data of Rao et al. [9] and the SPRITE dataset of Guttman lab [42] were downloaded respectively from GSE63525 and GSE114242 in the Gene Expression Omnibus (GEO) database. The other chromatin contact data [6, 8, 33] were downloaded from the supplementary materials of the corresponding publications.

References
1. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14(4):288–95. https://doi.org/10.1038/nrg3458.
2. Mossing M, Record M. Upstream operators enhance repression of the lac promoter. Science. 1986;233(4766):889–92. https://doi.org/10.1126/science.3090685.
3. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. Science. 2018;361(6409):1341–5. https://doi.org/10.1126/science.aau0320.
4. Lettice LA, Horikoshi T, Heaney SJH, et al. Disruption of a long-range cis-acting regulator for shh causes preaxial polydactyly. Proc Natl Acad Sci. 2002;99(11):7548–53. https://doi.org/10.1073/pnas.112212199.
5. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93. https://doi.org/10.1126/science.1181369.
6. Li G, Ruan X, Auerbach RK, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148(1-2):84–98. https://doi.org/10.1016/j.cell.2011.12.014.
7. Wang S, Hu H, Li X. Shared distal regulatory regions may contribute to the coordinated expression of human ribosomal protein genes. Genomics. 2020;112:2886–93.
8. Javierre BM, Burren OS, Wilder SP, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016;167(5):1369–138419. https://doi.org/10.1016/j.cell.2016.09.037.
9. Rao SSP, Huntley MH, Durand NC, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
10. Weber F, de Villiers J, Schaffner W. An SV40 "enhancer trap" incorporates exogenous enhancers or generates enhancers from its own sequences. Cell. 1984;36(4):983–92. https://doi.org/10.1016/0092-8674(84)90048-5.
11. Bellen HJ, Kane CJO, Wilson C, Grossniklaus U, Pearson RK, Gehring WJ. P-element-mediated enhancer detection: a versatile method to study development in drosophila. Genes Dev. 1989;3(9):1288–300. https://doi.org/10.1101/gad.3.9.1288.
12. Wang Y, Li X, Hu H. H3k4me2 reliably defines transcription factor binding regions in different cells. Genomics. 2014;103(2-3):222–8. https://doi.org/10.1016/j.ygeno.2014.02.002.
13. Crawford GE. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2005;16(1):123–31. https://doi.org/10.1101/gr.4074106.
14. Heintzman ND, Stuart RK, Hon G, Fu Y, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311–8. https://doi.org/10.1038/ng1966.
15. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455–61. https://doi.org/10.1038/nature12787.
16. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–502. https://doi.org/10.1126/science.1141319.
17. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4(8):651–7. https://doi.org/10.1038/nmeth1068.
18. Malin J, Aniba MR, Hannenhalli S. Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers. Nucleic Acids Res. 2013;41(14):6828–38. https://doi.org/10.1093/nar/gkt374.

19. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82. https://doi.org/10.1038/nature11232.

20. Zheng Y, Li X, Hu H. Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new features of motifs. Nucleic Acids Res. 2014;43(1):74–83. https://doi.org/10.1093/nar/gku1261.

21. Zheng Y, Li X, Hu H. PreDREM: a database of predicted DNA regulatory motifs from 349 human cell and tissue samples. Database. 2015;2015. https://doi.org/10.1093/database/bav007.

22. Daniel B, Nagy G, Hah N, et al. The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. Genes Dev. 2014;28(14):1562–77. https://doi.org/10.1101/gad.242685.114.

23. Danko CG, Hyland SL, Core LJ, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015;12(5):433–8. https://doi.org/10.1038/nmeth.3329.

24. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6. https://doi.org/10.1038/nmeth.1906.

25. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9(5):473–6. https://doi.org/10.1038/nmeth.1937.

26. Won K-J, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol. 2010;11(1):7. https://doi.org/10.1186/gb-2010-11-1-r7.

27. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA enhancer browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(Database):88–92. https://doi.org/10.1093/nar/gkl822.

28. Chen H, Li C, Peng X, et al. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. Cell. 2018;173(2):386–39912. https://doi.org/10.1016/j.cell.2018.03.027.

29. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307–19.

30. Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2015;47(1):8–12.

31. Dekker J. Capturing chromosome conformation. Science. 2002;295(5558):1306–11. https://doi.org/10.1126/science.1067799.

32. Dostie J, Richmond TA, Arnaout RA, et al. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16(10):1299–309. https://doi.org/10.1101/gr.5571506.

33. Jin F, Li Y, Dixon JR, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4. https://doi.org/10.1038/nature12644.

34. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–49. https://doi.org/10.1038/nature09906.

35. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci. 2014;111(21):2191–9. https://doi.org/10.1073/pnas.1320308111.

36. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res. 2015;43(18):8694–712. https://doi.org/10.1093/nar/gkv865.

37. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48(5):488–96. https://doi.org/10.1038/ng.3539.

38. Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep. 2016;6(1):. https://doi.org/10.1038/srep30043.

39. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, dot*brace*lari RC-S, Lupien M, Markowitz S, Scacheri PC. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2013;24(1):1–13. https://doi.org/10.1101/gr.164079.113.

40. Rödelsperger C, Guo G, Kolanczyk M, Pletschacher A, et al. Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. Nucleic Acids Res. 2010;39(7):2492–502. https://doi.org/10.1093/nar/gkq1081.

41. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80. https://doi.org/10.1038/nature11082.

42. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y, et al. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. Cell. 2018;174(3):744–57.

43. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;22(9):1760–74. https://doi.org/10.1101/gr.135350.111.

44. Blanchette M. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 2006;16(5):656–68. https://doi.org/10.1101/gr.4866006.

45. Cai X, Hou L, Su N, Hu H, Deng M, Li X. Systematic identification of conserved motif modules in the human genome. BMC Genomics. 2010;11(1):567. https://doi.org/10.1186/1471-2164-11-567.

46. McLean YC, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495–501.

47. Edelman LB, Fraser P. Transcription factories: genetic programming in three dimensions. Curr Opin Genet Dev. 2012;22(2):110–4.

48. Papantonis A, Cook PR. Transcription factories: genome organization and gene regulation. Chem Rev. 2013;113(11):8683–705.

49. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.

50. Talukder A, Saadat S, Li X, Hu H. EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz641.

51. Li X, Zheng Y, Hu H, Li X. Integrative analyses shed new light on human ribosomal protein gene regulation. Sci Rep. 2016;6(1):. https://doi.org/10.1038/srep28619.

52. Latapy M, Magnien C, Vecchio ND. Basic notions for the analysis of large two-mode networks. Soc Netw. 2008;30(1):31–48. https://doi.org/10.1016/j.socnet.2007.04.006.

53. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Commun ACM. 1973;16(9):575–7. https://doi.org/10.1145/362342.362367.

54. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

55. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18(1):50–60.

Publisher's Note