Genome-wide analysis of human putative transcriptional target genes reveals significant functional enrichments

Naoki Osato[1*]

[1]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

[*]Whom correspondence should be addressed.

E-mail: naokiosato11@gmail.com

Keywords: transcriptional cascade; functional annotation; functional enrichment; open chromatin region; chromatin interaction; enhancer

Abbreviations: TF, transcription factors. TFBS, transcription factor binding sites. TSS, transcriptional start sites.

Short title: Functional enrichments of human putative transcriptional target genes

Abstract

Functional enrichments of putative transcriptional target genes have been utilized to understand the functions of transcription factors and cascades in a cell. To investigate their features, transcriptional target genes were predicted using open chromatin regions of human immune and ES cells, as well as known transcription factor binding sequences. Gene Ontology annotations showed four times larger numbers of functional enrichments in putative transcriptional target genes than gene expression information alone in the cell types. More than two times larger numbers of functional enrichments in putative target genes was observed using forward–reverse orientation of CTCF-binding sites than without them. These analyses would be useful to find genomic features involved in chromatin interaction and improve the prediction of transcriptional target genes.

Introduction

More than 400 types of cells have been found in the human body. Human development is accompanied by the differentiation of stem cells into various cell types, leading to a diversification of their phenotypes and functions. For example, the development of the immune system involves differentiation and diversification of stem cells into various types of mature immune cells. The functions of monocytes include phagocytosis and antigen presentation. $CD4^+$ T cells, however, play a central role in cell-mediated immunity and are involved in the activation of phagocytes and antigen-specific cytotoxic T-lymphocytes, and the release of various cytokines in response to an antigen. The $CD20^+$ B cells are involved in the production of antibodies against antigens.

Differentiation of cells is often triggered by the expression of transcription factors (TF) followed by the expression of their target genes, which results in the transformation of cells into other cell types. For example, the transcription factors PU.1 and CCAAT enhancer-binding protein α (C/EBPα) play a critical role in the expression of myeloid-specific genes and the generation of monocytes and macrophages [1, 2]. The transcription factor GATA-3 is essential

for early T cell development and the differentiation of naive $CD4^+$ T cells into Th2 effector cells [3]. E2A, EBF1, PAX5, and Ikaros are among the most important transcription factors that control early development in mice, thereby conditioning homeostatic B cell lymphopoiesis [4].

We previously examined the differentiation of monocytes and macrophages in mice, and discovered that the transcription factor IRF8 was essential for cellular differentiation [5]. An analysis of transcription factor-binding sites (TFBS) revealed that IRF8 regulated the expression of KLF4 through the IRF8 transcriptional cascade. Functional enrichment analyses revealed that the target genes of IRF8 showed functional enrichment for antigen presentation, whereas those of KLF4 showed functional enrichments for phagocytosis and locomotion. These results suggested that the transcriptional cascades of IRF8 and KLF4 included different functional modules of target genes.

Functional enrichments of transcriptional cascades of IRF8 and KLF4 appeared to be related to the cellular functions of monocytes and macrophages. Although several transcription factors were expressed in monocytes and macrophages, the number of these transcriptional target genes that resulted in functional enrichments remains unknown. Whether transcriptional target genes in other human cells showed functional enrichments also remain unclear. If the transcriptional target genes showed significant functional enrichment, analyzing transcriptional target genes and cascades would be useful in identifying genes involved in a specific cellular function. Using the budding yeast, previous studies examined the functional enrichments on a genome-scale genetic interaction map using the GeneMANIA algorithm [6, 7]. Using bacterial systems, the analyses of functional enrichments of predicted regulatory networks were performed using Gene Ontology annotations [8]. Various databases of functional annotations of genes and pathways exist. Analysis of functional enrichments is expected to be useful for understanding the association of genes involved in similar functions and same pathways, and for predicting unknown gene functions such as non-protein-coding RNAs. In addition, the extent of enhancer region contribution to functional enrichments of transcriptional target genes remains unknown.

In this study, transcriptional target genes were predicted using public databases of open chromatin regions of human monocytes, naive $CD4^+$ T and $CD20^+$ B cells, and known transcription factor binding sequences. Functional enrichment analyses of putative transcriptional target genes were conducted using 10 different annotation databases of functional annotations and pathways.

Materials and Methods

Searches for transcription factor binding sequences from open chromatin regions

To examine transcriptional regulatory target genes, bed files of hg19 narrow peaks of ENCODE DNase-DGF data for Monocytes-$CD14^+$_RO01746 (GSM1024791; UCSC Accession: wgEncodeEH001196), $CD4^+$_Naive_Wb11970640 (GSM1014537; UCSC Accession: wgEncodeEH003156), $CD20^+$_RO01778 (GSM1014525; UCSC Accession: wgEncodeEH002442), and H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556) from the ENCODE website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/) were used. For comparison with transcriptional target genes predicted using ChIP-seq data, bed files of hg19 narrow peaks of ENCODE ChIP-seq data for 19 transcription factors (TF) (BACH1, BRCA1, C/EBPbeta, CHD2, c-JUN, c-MYC, GTF2I, JUND, MAFK, MAX, MXI1, NRF1, RAD21, RFX5, SIN3A, SUZ12, TBP, USF2, ZNF143) in H1-hESC from the ENCODE website (https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform) were utilized.

To identify transcription factor binding sites (TFBS) from the DNase-DGF data, TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences of ENCODE ChIP-seq data were used [9-15]. Position weight matrices of transcription factor binding sequences were transformed into TRANSFAC matrices and then into MEME matrices using in-house Perl scripts and transfac2meme in MEME suite [16]. Transcription factor binding

sequences of transcription factors derived from vertebrates were used for further analyses. Searches were conducted for transcription factor binding sequences from the central 50-bp regions of each narrow peak using FIMO with *p*-value threshold of 10^{-5} [17]. Transcription factors corresponding to transcription factor binding sequences were searched computationally by comparing their names and gene symbols of HGNC (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known canonical transcripts (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as transcription factor binding sequences were not linked to transcript IDs such as UCSC, RefSeq, and Ensembl transcripts.

Prediction of transcriptional target genes

Target genes of a transcription factor were assigned when its TFBS was found in DNase-DGF narrow peaks in promoter or enhancer regions of genes. Promoter and enhancer regions were defined as follows: promoter regions were those that were within distances of ±5 kb from transcriptional start sites (TSS). Enhancer regions were defined as per the following four criteria, which are similar or same as those defined in a previous study [18]: (1) the basal plus extension association rule assigns a basal regulatory domain to each gene regardless of other nearby genes. The domain is then extended to the basal regulatory domain of the nearest upstream and downstream genes, and includes a 5 kb + 5 kb basal region and an extension up to 300 kb or the midpoint between the TSS of the gene and that of the nearest gene upstream and downstream; (2) 5 kb + 1 kb basal region and an extension up to 1 Mb; (3) the two nearest genes association rule, which extends the regulatory domain to the TSS of the nearest upstream and downstream genes without the limitation of extension length; and (4) the single nearest gene association rule, which extends the regulatory domain to the midpoint between the TSS of the gene and that of the nearest gene upstream and downstream without the limitation of extension length. Definition of criteria (1) was used in our previous study [5]. Definitions of criteria (2), (3), and (4) were the same as those in Figure 3 of the previous study [18], however,

definitions of criteria (3) and (4) did not have the limitation of extension length in this study. The genomic positions of genes were identified using knownGene.txt.gz in UCSC bioinformatics sites [19]. The website knownCanonical.txt.gz was also utilized for choosing representative transcripts among various alternate forms for assigning promoter and enhancer regions of the genes. From the list of transcription factor binding sequences and transcriptional target genes, redundant transcription factor binding sequences were removed by comparing the target genes of a transcription factor binding sequence and its corresponding transcription factor; if identical, one of the transcription factor binding sequences was used. When the number of transcriptional target genes predicted from a transcription factor binding sequence was less than five, the transcription factor binding sequence was omitted.

Gene expression analyses

For gene expression data, RNA-seq reads mapped onto human hg19 genome sequences were obtained, including ENCODE long RNA-seq reads with poly-A of monocytes $CD14^+$ cells, $CD20^+$ B cells, and H1-hESC (GSM984609, GSM981256, GSE26284, and GSM958733), and UCSF-UBC human reference epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells (GSM669617). Two replicates were present for monocytes $CD14^+$ cells, $CD20^+$ B cells, and H1-hESC and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using RSeQC [20]. For monocytes, Blueprint RNA-seq RPKM data (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was also used [21]. Based on RPKM, UCSC transcripts with expression levels among top 30% of all the transcripts were selected in each cell type.

Functional enrichment analyses

The functional enrichments of target genes of a TFBS and its corresponding transcription factor were examined using GO-Elite v1.2.5 with *p*-value threshold at 1, and after GO-Elite analyses a false discovery rate (FDR) test was performed with *q*-value threshold at 10^{-3} to

correct for multiple comparisons of thousands of groups of transcriptional target genes in each cell type and condition [22]. For examining functional enrichments of high or low expressed genes independent of transcriptional target genes, the *p*-value threshold was set to 0.01 or 0.05 to confirm that the results were not significantly changed. UCSC gene IDs were transformed into RefSeq IDs prior to GO-Elite analyses. GO-Elite uses 10 databases for identifying functional enrichments: (1) Gene Ontology, (2) Disease Ontology, (3) Pathway Commons, (4) GO Slim, (5) WikiPathways, (6) KEGG, (7) Transcription factor to target genes, (8) microRNA to target genes, (9) InterPro and UniProt functional regions (Domains), and (10) Cellular biomarkers (BioMarkers). To calculate the normalized numbers of functional enrichments of target genes, the numbers of functional enrichments were divided by the total number of target genes in each cell type and condition, and were multiplied by 10^5. In tables showing the numbers of functional enrichments in 10 databases, heat maps were plotted according to *Z*-scores calculated from the numbers of functional enrichments of each database using in-house Excel VBA scripts. In the comparisons of the normalized numbers of functional enrichments of target genes in cell types and conditions, if the number of a functional annotation in a cell type or condition was two times larger than that in the other cell type or condition, the functional annotation was recognized as more enriched than the other cell type or condition.

To investigate whether the normalized numbers of functional enrichments of transcriptional target genes correlate with the prediction of target genes, a part of target genes were changed with randomly selected genes with high expression level (top 30% expression level), and functional enrichments of the target genes were examined. First, 5%, 10%, 20%, 40%, and 60% of target genes were changed with randomly selected genes with high expression level in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells. Second, as another randomization of target genes, the same number of 5%, 10%, 20%, 40%, and 60% of target genes were selected randomly from highly expressed genes, then added them to the original target genes, and functional enrichments of the target genes were examined. All analyses were repeated three times to estimate standard errors (Figure 2A and B, and Supplementary Table S1). The same

analysis was performed using DNase-DGF data and ChIP-seq data of 19 TF in H1-hESC. Transcriptional target genes were predicted from promoter and combined promoter–enhancer regions (enhancer definition 4) (Supplementary Tables S2).

CTCF-binding sites

CTCF ChIP-seq data for monocytes $CD14^+$ cells (GSM1003508_hg19_wgEncodeBroadHistoneMonocd14ro1746CtcfPk.broadPeak.gz), $CD4^+$ T cells (SRR001460.bam), $CD20^+$ B cells (GSM1003474_hg19_wgEncodeBroadHistoneCd20CtcfPk.broadPeak.gz), and H1-hESC (wgEncodeAwgTfbsUtaH1hescCtcfUniPk.narrowPeak.gz) were used. SRR001460.bam was sorted and indexed by SAMtools and transformed into a bed file using bamToBed of BEDTools [23, 24]. ChIP-seq peaks were predicted by SICER-rb.sh of SICER with optional parameters 'hg19 1 200 150 0.74 200 100' [25]. Combined promoter–enhancer regions (enhancer definition 4) were shortened at the genomic locations of CTCF-binding sites that were the closest to a transcriptional start site, and transcriptional target genes were predicted from the shortened enhancer regions using TFBS. Furthermore, combined promoter–enhancer regions (enhancer definition 4) were shortened at the genomic locations of forward–reverse orientation of CTCF-binding sites. When forward or reverse orientation of CTCF-binding sites were continuously located in genome sequences several times, the most external forward–reverse orientation of CTCF-binding sites were selected.

Results

Prediction of transcriptional target genes

To examine functional enrichments of transcriptional target genes in a genome scale, Transcriptional target genes were predicted in human monocytes, $CD4^+$ T cells, and $CD20^+$ B cells. Searches for known transcription factor binding sequences, which were collected from various databases and papers, were conducted in open chromatin regions of the promoter

sequences of RefSeq transcripts (Figure 1, see Methods). Among 5,277 transcription factor binding sequences derived from vertebrates, 4,391 were linked to 971 TF transcripts computationally (see Methods). To maintain the sensitivity of the searches for transcription factor binding sites and as some transcription factors will recognize multiple distinctly different sequence motifs, transcription factor binding sequences that targeted the same genes were recognized as redundant, and one of the sequences was used [26] (see Methods). In total, 3,337 transcription factor binding sequences in human monocytes, 3,652 in $CD4^+$ T cells, and 3,187 in $CD20^+$ B cells were identified with their target genes, which were selected from highly expressed genes in a cell.

The total numbers of unique highly expressed target genes of transcription factor binding sequences (top 30% expression level) were 4,481, 7,558, and 4,753 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells respectively using promoter regions. The mean target genes of a transcription factor were 124, 164, and 144 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively, with the corresponding medians being 24, 33, and 24, respectively. With regard to the genomic localizations of TFBS, 51%, 65%, and 61% of TFBS were located within promoter regions (±5 kb of TSS) of target genes in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively (according to enhancer definition 1, see Methods).

Functional enrichments of putative transcriptional target genes

Functional enrichments of the putative target genes were examined. The distribution of functional enrichments in transcriptional target genes was predicted using genome sequences of promoter regions in the three cell types (Figure 1 and Table 1, see Methods). Furthermore, the effect of transcriptional target genes including randomly selected genes on functional enrichments was investigated using DNase-DGF data of monocytes, $CD4^+$ T cells and $CD20^+$ B cells and ChIP-seq data of H1-hESC (Figure 2A and B, see Methods). The native putative transcriptional target genes not including randomly selected genes showed the highest functional enrichments using Gene Ontology, GO Slim, KEGG, Pathway Commons,

WikiPathways, InterPro and UniProt functional regions (Domains) in both DNase-DGF and ChIP-seq data of the four types of cells. Of the 10 databases used in this analysis, the Gene Ontology database consists of three types of functional annotations, i.e., 20,836 biological processes, 9,020 molecular functions, and 2,847 cellular components. The numbers of functional enrichments of Gene Ontology annotations in target genes of a transcription factor were 2,902, 4,077, and 2,778 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively. An examination of functional enrichments of highly expressed genes (top 30% expression level) independent of the transcriptional target genes revealed 237, 301, and 239 unique Gene Ontology annotations in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively (Table 1). Further, the examination of functional enrichments of highly expressed target genes (top 30% expression level) in target genes revealed 1,271, 1,654, and 1,192 unique Gene Ontology annotations in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively i.e., These numbers were four times larger than functional enrichments identified by gene expression information alone, suggesting that transcriptional target genes were frequently associated with similar functions or pathways (Supplementary Table S3 and S4).

Functional enrichments of transcriptional target genes from other databases were also examined (Table 1). KEGG, Target genes of transcription factors, Disease Ontology, GO Slim, Pathway Commons, Cellular biomarkers, Target genes of microRNAs, Protein domains, and WikiPathways had 95, 16, 127, 12, 242, 17, 97, 303, and 105 unique functional annotations, respectively. The numbers of functional enrichments of transcriptional target genes in the other annotation databases except for microRNAs and Protein domains were significantly higher than gene expression information alone as well as Gene Ontology annotations (Table 1). The functional enrichments of transcriptional target genes from Pathway Commons for monocytes, $CD4^+$ T cells, and $CD20^+$ B cells are shown in Table 2 and Supplementary Table S5. Functional enrichments were found to be related to cellular functions, e.g., interferon signaling, GMCSF (Granulocyte-macrophage colony-stimulating factor, a kind of cytokine)-mediated signaling events, antigen processing-cross presentation in monocytes; TCR (T-cell receptor) signaling in

naive $CD4^+$ T cells, IL-12 (Interleukin-12, a kind of cytokine)-mediated signaling events, and downstream signaling in naive $CD8^+$ T cells in $CD4^+$ T cells; interferon alpha/beta signaling, IL8- and CXCR2 (Chemokine receptor type 2, a kind of cytokine)-mediated signaling events, and BCR (B cell antigen receptor) signaling pathway in $CD20^+$ B cells. WikiPathways also revealed that functional enrichments were associated with cellular functions (Supplementary Table S6).

Effect of enhancer regions on functional enrichments

To understand the effect of enhancer regions on the functional enrichments of target genes, the definition of enhancer regions was modified according to four criteria (Figure 3A and see Methods) [18], and functional enrichments were investigated.

According to the definition of enhancer (1), the means of target genes were 177, 217, and 175 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively, whereas the corresponding medians were 55, 58, and 37, respectively (Supplementary Table S7). The numbers of functional enrichments of Pathway Commons annotations using promoter regions were 1,005, 1,806, and 821 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively (Supplementary Table S8). With the use of combined promoter–enhancer regions (enhancer definition 1), the numbers of functional enrichments of Pathway Commons annotations were 3,087, 7,216, and 3,900, representing 3.07-, 4.00-, and 4.75-fold increases, respectively, in the three cells types. Additionally, the numbers of unique Pathway Commons annotations with promoter regions were 321, 415, and 329 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively; the corresponding numbers with the use of combined promoter–enhancer regions (enhancer definition 1) were 364, 437, and 364, representing 1.13-, 1.05-, and 1.11-fold increases, respectively, in the three cell types. The normalized numbers of functional enrichments of Pathway Commons annotations were 44.75, 84.51, and 59.32, representing 1.84-, 2.80-, and 3.32-fold increases, respectively, in the three cell types (enhancer definition 1, Table 3).

The normalized numbers of the functional enrichments of transcriptional target genes

showed enhancer definition (4) as the highest number, followed by enhancer definition (1) and (2) in the three cell types. Although enhancer definition (3) was the longest among the four criteria, it showed the lowest number of functional enrichments in the three cell types (Figure 3A and Table 3). ChIP-seq data of 19 TF in H1-hESC (Human embryonic stem cells) also showed the same or similar tendencies (Supplementary Table S9, see Supplementary information).

Differences in functional enrichments obtained using promoter *versus* combined promoter–enhancer regions (enhancer definition 1) were examined using the functional enrichments obtained using Pathway Commons (Supplementary Table S10). A comparison of 321 and 364 functional enrichments using the promoter and combined promoter–enhancer regions, respectively, in monocytes revealed that 152 (47% in promoter, 42% in promoter–enhancer) of them were common. For example, IFN-gamma (Interferon gamma) pathway, GMCSF (Granulocyte-macrophage colony-stimulating factor, a kind of cytokine)-mediated signaling events, and PDGF (Platelet-derived growth factor) receptor signaling network were enriched using enhancer regions (definition 1) as opposed to promoter regions (Supplementary Table S32). The comparison of 415 (promoter) and 437 (combined promoter–enhancer) functional enrichments in $CD4^+$ T cells revealed that 163 of them (39% in promoter, 37% in promoter–enhancer) were common. IFN-gamma pathway, TCR (T-cell receptor) signaling in naive $CD4^+$ T cells, and IL3 (Interleukin-3, a kind of cytokine)-mediated signaling events were enriched using enhancer regions definition (1). The comparison of 329 (promoter) and 364 (combined promoter–enhancer) functional enrichments in $CD20^+$ B cells revealed that 171 of them (52% in promoter, 47% in promoter–enhancer) were common. IL5-mediated signaling events, IL4-mediated signaling events, and cytokine signaling in immune system were enriched in $CD20^+$ B cells using enhancer regions according to definition (1). These results showed that new functional enrichments related to cellular functions could be identified using enhancer regions.

Effect of CTCF-binding sites on functional enrichments

CTCF have the activity of insulators to block the interaction between enhancers and promoters [27]. Recent studies identified a correlation between the orientation of CTCF-binding sites and chromatin loops (Figure 3B) [28]. Forward–reverse (FR) orientation of CTCF-binding sites are frequently found in chromatin loops. To examine the effect of forward–reverse orientation of CTCF-binding sites on functional enrichments of target genes, combined promoter–enhancer regions were shortened at the genomic locations of forward–reverse orientation of CTCF-binding sites, and transcriptional target genes were predicted from the shortened enhancer regions using TFBS (see Methods). The numbers of functional enrichments of target genes were investigated. According to combined promoter–enhancer regions shortened at genomic locations of forward–reverse orientation of CTCF-binding sites (enhancer definition 4), the means of target genes were 58, 56, and 64 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively, whereas the corresponding medians were 21, 19, and 17, respectively (Supplementary Table S11). The normalized numbers of functional enrichments of Pathway Commons annotations using combined promoter–enhancer regions (enhancer definition 4) were 71.42, 108.08, and 90.99 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively (Table 4). With the use of combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4), the normalized numbers of functional enrichments of Pathway Commons annotations were 196.58, 220.54, and 220.77, representing 2.75-, 2.04-, and 2.43-fold increases, respectively, in the three cells types. Additionally, the normalized numbers of functional enrichments of unique Pathway Commons annotations with combined promoter–enhancer regions (enhancer definition 4) were 5.09, 5.34, and 6.00 in monocytes, $CD4^+$ T cells, and $CD20^+$ B cells, respectively; the corresponding normalized numbers with the use of combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) were 9.88, 10.72, and 9.10, representing 1.94-, 2.01-, and 1.52-fold increases, respectively, in the three cell types (Supplementary Table S12). The normalized numbers of were significantly increased between

combined promoter–enhancer regions and combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites in Gene Ontology, Disease Ontology, Pathway Commons, GO Slim, WikiPathways, KEGG, InterPro and UniProt functional regions (Domains) annotations. These increases were also significant, compared with combined promoter–enhancer regions shortened at CTCF-binding sites without the consideration of their orientation.

Differences in functional enrichments obtained using combined promoter–enhancer regions *versus* combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) were examined using the functional enrichments of Pathway Commons (Supplementary information). These results showed that new functional enrichments related to cellular functions could be identified using forward–reverse orientation of CTCF-binding sites.

Comparison of expression levels of putative transcriptional target genes

To examine the relationship between functional enrichments and expression levels of target genes, the expression levels of target genes predicted from promoter regions and three types of combined promoter–enhancer regions were investigated in monocytes, $CD4^+$ T cells and H1-hESC (Figure 4). Median expression levels of the target genes of the same transcription factor binding sequences were compared between promoter regions and three types of combined promoter–enhancer regions. Red and blue dots in Figure 4 show statistically significant difference of the distribution of expression levels of target genes between promoter regions and combined promoter–enhancer regions. Additionally, red dots show the median expression level of target genes of a TFBS was higher in combined promoter–enhancer regions than promoter regions, and blue dots show the median expression level of target genes of a TFBS was lower in combined promoter–enhancer regions than promoter regions. The ratios of red dots were higher in combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) *versus* promoter regions than combined promoter–

enhancer regions (enhancer definition 4) *versus* promoter regions in monocytes and $CD4^+$ T cells. The ratios of blue dots were higher in combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) *versus* promoter regions than combined promoter–enhancer regions (enhancer definition 4) *versus* promoter regions in H1-hESC. Moreover, the ratio of the sum of the median expression levels between the three types of combined promoter–enhancer regions and promoter regions in monocytes and $CD4^+$ T cells was the highest in combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) (Supplementary Table S14). The ratio of the sum of median expression levels between the three types of combined promoter–enhancer regions and promoter regions in H1-hESC was the lowest in combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4).

Combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) tended to change (increase or decrease) the expression levels of target genes more than the other types of combined promoter–enhancer regions. This implied that gene expression tended to be activated in monocytes and $CD4^+$ T cells, but repressed in H1-hESC by enhancers. Combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites (enhancer definition 4) also showed the highest normalized number of functional enrichments of transcriptional target genes, as shown in the previous paragraphs.

Discussion

Genome-wide functional enrichments of putative target genes of human transcription factors were investigated. This is the first report to show human genome-wide functional enrichment analyses of putative transcriptional target genes using various functional annotation databases. Functional enrichments common to the three cell types included immunological terms representing their common features; in addition, immunological terms related to each cell

type were also found. This result suggested that predicted transcriptional target genes included cell-specific and common target genes, which were related to cellular functions.

In the analysis of combined promoter–enhancer regions, only about 40% of functional enrichments of Pathway Commons annotations were unchanged between promoter and combined promoter–enhancer regions. Combined promoter–enhancer regions significantly affect the functional enrichments of transcriptional target genes. Of the various definitions of combined promoter–enhancer regions, definition (3) specifies the longest region and supports the highest means and medians of transcriptional target genes. However, the ratios of the numbers of functional enrichments in target genes were not highest as per definition (3).

Forward–reverse orientation of CTCF-binding sites also increased the normalized numbers of functional enrichments of Pathway Commons annotations two times more than combined promoter–enhancer regions (Enhancer 4). Transcriptional target genes predicted from genomic regions shortened at the CTCF-binding sites tended to include the similar function of genes. About 40 – 80% of functional enrichments were unchanged between promoter and combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites, and the functional enrichments observed in combined promoter–enhancer regions shortened at forward–reverse orientation of CTCF-binding sites as opposed to promoter regions included various immunological terms. Among 33,939 RefSeq transcripts, 7,202 (21%), 4,404 (13%), and 6,921 (20%) (p-value < 10^{-5} in the search for CTCF-binding motifs using FIMO) to 9,608 (28%), 5,806 (17%), and 9,137 (27%) (p-value < 10^{-4}) of transcripts had forward–reverse orientation of CTCF-binding sites within 1Mb from transcriptional start sites in the three cell types, respectively. Other insulator sites and mechanisms may also play role in enhancer-promoter interactions.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16K00387. This research was partially supported by the Platform Project for Supporting in Drug Discovery and Life

Science Research（Platform for Dynamic Approaches to Living System）from Japan Agency for Medical Research and Development (AMED). This research was partially supported by Development of Fundamental Technologies for Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for Medical Research and Development (AMED). The supercomputing resource was provided by Human Genome Center of the Institute of Medical Science at the University of Tokyo. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

References

1. Valledor, A. F., Borras, F. E., Cullell-Young, M. & Celada, A. (1998) Transcription factors that regulate monocyte/macrophage differentiation, *Journal of leukocyte biology.* **63**, 405-17.

2. Nagamura-Inoue, T., Tamura, T. & Ozato, K. (2001) Transcription factors that regulate growth and differentiation of myeloid cells, *International reviews of immunology.* **20**, 83-105.

3. Ho, I. C., Tai, T. S. & Pai, S. Y. (2009) GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation, *Nature reviews Immunology.* **9**, 125-35.

4. Rothenberg, E. V. (2010) B cell specification from the genome up, *Nature immunology.* **11**, 572-4.

5. Kurotaki, D., Osato, N., Nishiyama, A., Yamamoto, M., Ban, T., Sato, H., Nakabayashi, J., Umehara, M., Miyake, N., Matsumoto, N., Nakazawa, M., Ozato, K. & Tamura, T. (2013) Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation, *Blood.* **121**, 1839-49.

6. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z. Y., Liang, W., Marback, M., Paw, J., San Luis, B. J., Shuteriqi, E., Tong, A. H., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F. P., Giaever, G., Nislow, C.,

Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A. C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J. & Boone, C. (2010) The genetic landscape of a cell, *Science (New York, NY)*. **327**, 425-31.

7. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D. & Morris, Q. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function, *Nucleic acids research.* **38**, W214-20.

8. Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J. & Stolovitzky, G. (2012) Wisdom of crowds for robust gene network inference, *Nature methods.* **9**, 796-804.

9. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites, *Nucleic acids research.* **24**, 238-41.

10. Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. & Sandelin, A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic acids research.* **38**, D105-10.

11. Newburger, D. E. & Bulyk, M. L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions, *Nucleic acids research.* **37**, D77-82.

12. Zhao, Y. & Stormo, G. D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity, *Nature biotechnology.* **29**, 480-3.

13. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. (2013) DNA-binding specificities of human transcription factors, *Cell.* **152**, 327-39.

14. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. & Qian, J. (2010) hPDI: a database of experimental human protein-DNA interactions, *Bioinformatics (Oxford, England)*. **26**, 287-9.

15. Kheradpour, P. & Kellis, M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments, *Nucleic acids research.* **42**, 2976-87.

16. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching, *Nucleic acids research.* **37**, W202-8.

17. Grant, C. E., Bailey, T. L. & Noble, W. S. (2011) FIMO: scanning for occurrences of a given motif, *Bioinformatics (Oxford, England).* **27**, 1017-8.

18. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions, *Nature biotechnology.* **28**, 495-501.

19. Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M. & Kent, W. J. (2014) The UCSC Genome Browser database: 2014 update, *Nucleic acids research.* **42**, D764-70.

20. Wang, L., Wang, S. & Li, W. (2012) RSeQC: quality control of RNA-seq experiments, *Bioinformatics (Oxford, England).* **28**, 2184-5.

21. Saeed, S., Quintin, J., Kerstens, H. H., Rao, N. A., Aghajanirefah, A., Matarese, F., Cheng, S. C., Ratter, J., Berentsen, K., van der Ent, M. A., Sharifi, N., Janssen-Megens, E. M., Ter Huurne, M., Mandoli, A., van Schaik, T., Ng, A., Burden, F., Downes, K., Frontini, M., Kumar, V., Giamarellos-Bourboulis, E. J., Ouwehand, W. H., van der Meer, J. W., Joosten, L. A., Wijmenga, C., Martens, J. H., Xavier, R. J., Logie, C., Netea, M. G. & Stunnenberg, H. G. (2014) Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity, *Science (New York, NY).* **345**, 1251086.

22. Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., Conklin, B. R., Pico, A. R. & Salomonis, N. (2012) GO-Elite: a flexible solution for pathway and ontology over-representation, *Bioinformatics (Oxford, England).* **28**, 2209-10.

23. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics (Oxford, England)*. **25**, 2078-9.

24. Quinlan, A. R. & Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics (Oxford, England)*. **26**, 841-2.

25. Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K. & Peng, W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics (Oxford, England)*. **25**, 1952-8.

26. Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C. F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R. & Bulyk, M. L. (2009) Diversity and complexity in DNA recognition by transcription factors, *Science (New York, NY)*. **324**, 1720-3.

27. Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K. & Peters, J. M. (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor, *Nature.* **451**, 796-801.

28. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M. Q., Ren, B., Krainer, A. R., Maniatis, T. & Wu, Q. (2015) CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function, *Cell.* **162**, 900-10.

29. de Wit, E., Vos, E. S., Holwerda, S. J., Valdes-Quezada, C., Verstegen, M. J., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. & de Laat, W. (2015) CTCF Binding Polarity Determines Chromatin Looping, *Molecular cell.* **60**, 676-84.

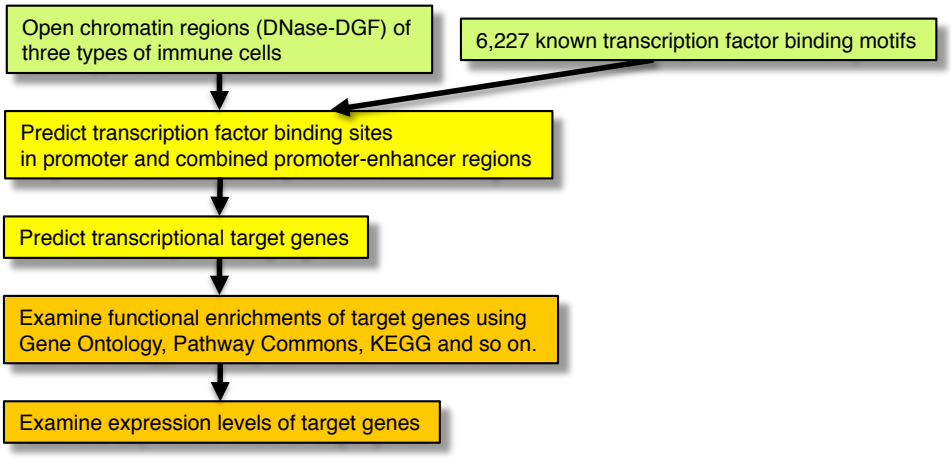

Figure 1. Analyses of functional enrichments of putative transcriptional target genes. Transcriptional target genes were predicted using open chromatin regions (DNase-DGF) and known transcription factor binding sequences. Functional enrichments of target genes were analyzed using 10 annotation databases, and were changed based on the criteria of enhancer regions. To compare with the tendency of the normalized numbers of functional enrichments, the median expression levels of target genes were examined using enhancer and promoter regions.

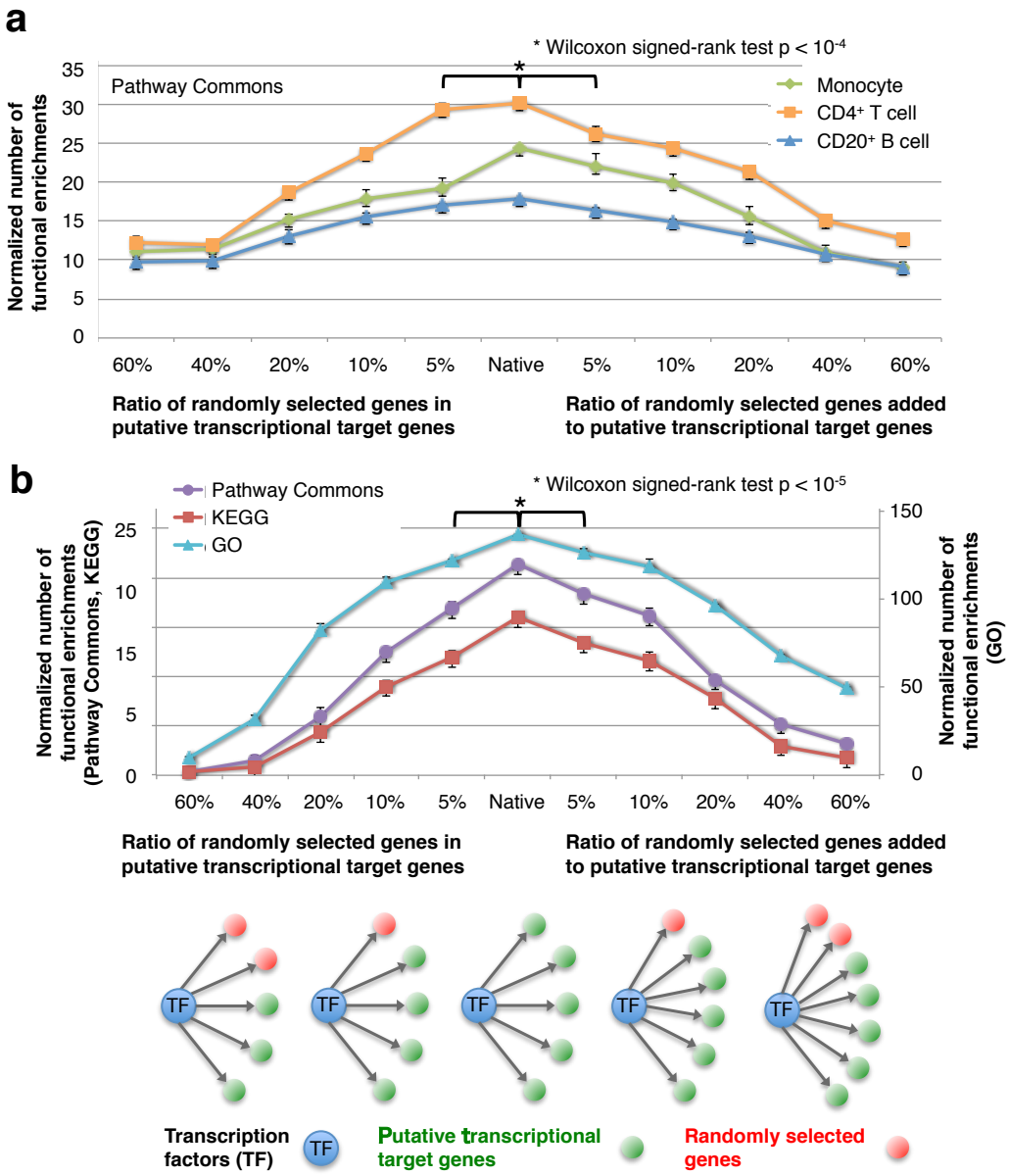

Figure 2. Effect of randomly selected genes on functional enrichments. a, Effect of randomly selected genes on functional enrichments using DNase-DGF data. Transcriptional target genes were predicted using DNase-DGF data in human monocytes, $CD4^+$ T cells, and $CD20^+$ B cells. The ratio of randomly selected genes in the target genes of each TF was changed between 5% and 60%. Native target genes showed the most functional enrichments. **b,** Effect of randomly selected genes on functional enrichments using ChIP-seq data. Transcriptional target genes were predicted using ChIP-seq data of 19 TF in H1-hESC. Native target genes showed the most functional enrichments.

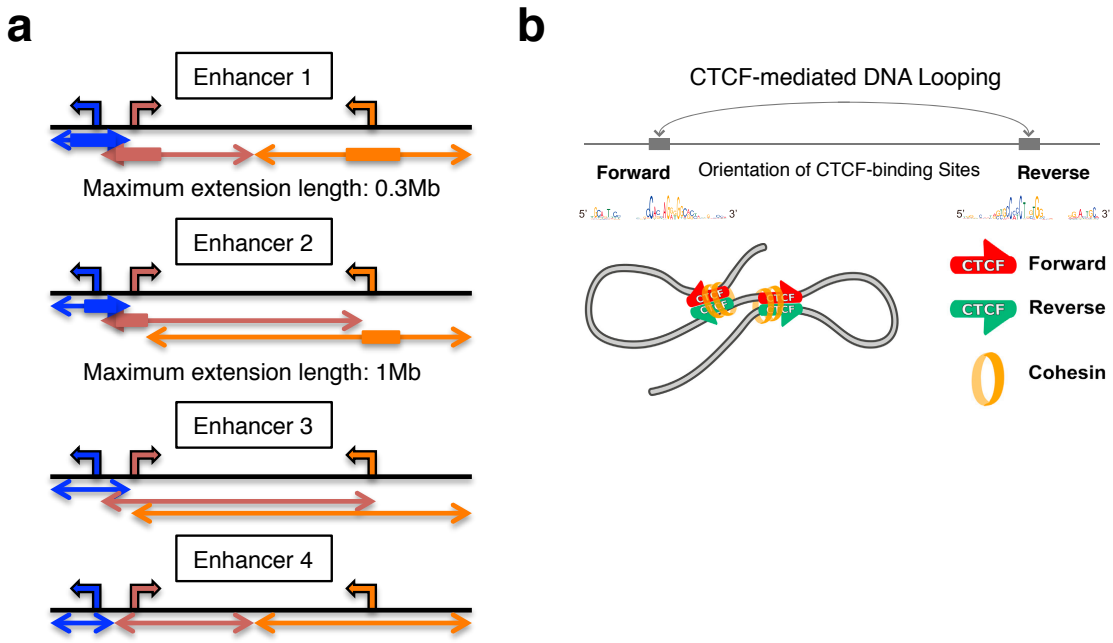

Figure 3. Criteria of enhancer regions and features of chromatin interactions. a, Computationally-defined regulatory domains [18]. The transcription start site (TSS) of each gene is indicated as an arrow. The corresponding regulatory domain for each gene is shown in a matching color as an arrowed line. A basal regulatory domain to each gene was assigned regardless of nearby genes (thick line). (see Methods). **b,** forward–reverse orientation of CTCF-binding sites are frequently found in chromatin interactions. Figures adapted from [28], [29].

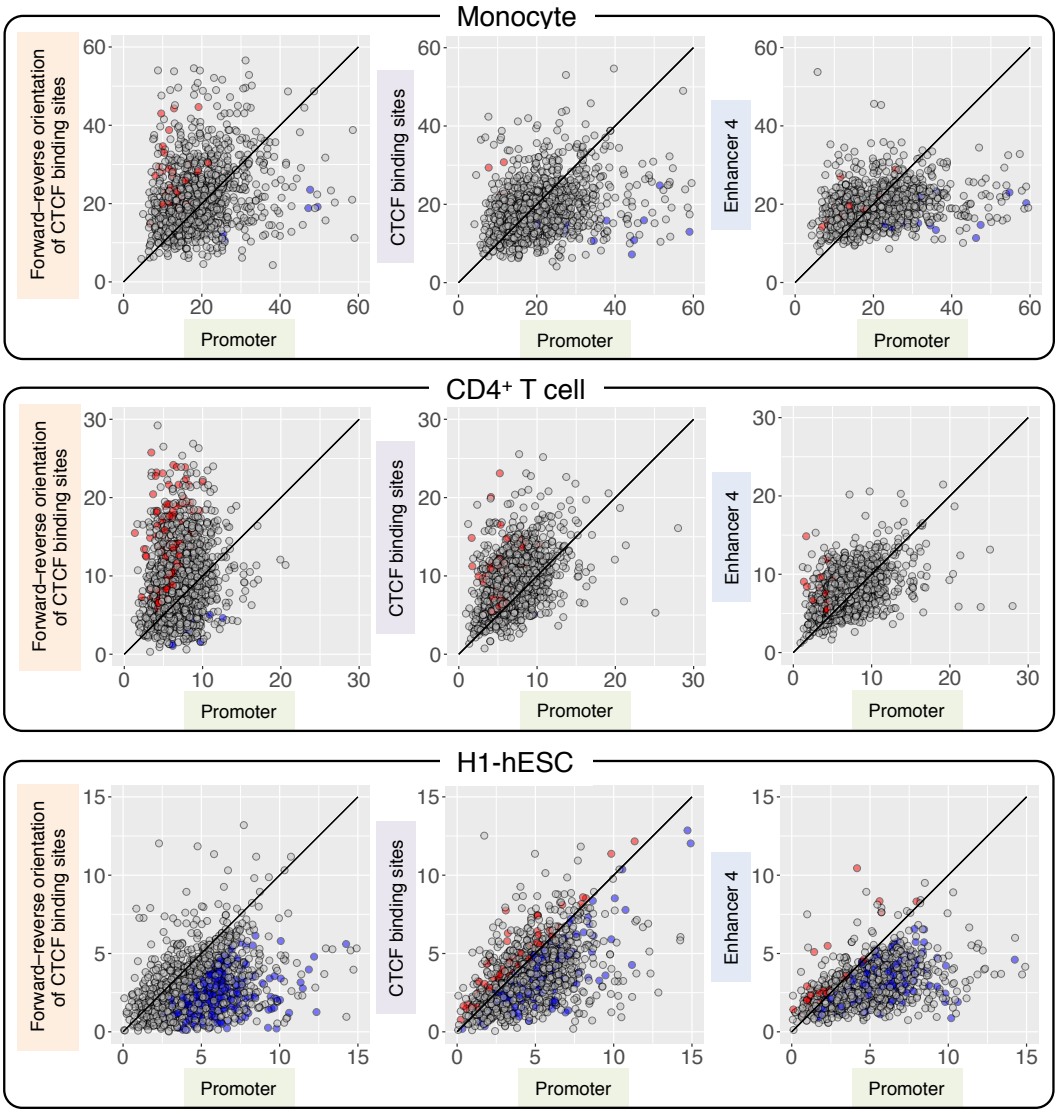

Figure 4. Comparison of the median expression levels of transcriptional target genes predicted from enhancer and promoter regions. The median expression levels of the target genes of the same transcription factor binding sequences were compared between promoter and three types of combined promoter–enhancer regions. Red and blue dots show statistically significant difference of the distribution of expression levels of target genes between promoter and combined promoter–enhancer regions. Red dots show the median expression level of target genes was higher in combined promoter–enhancer regions than promoter regions, and blue dots show the median expression level of target genes was lower in combined promoter–enhancer regions than promoter regions.

Table 1. Number of functional enrichments and unique functional enrichemnts of putative transcriptional target genes.

Number of functional enrichments of putative trancriptional target genes

	KEGG	TF Targets	CTD Ontology	GO Slim	GO	Pathway Commons	BioMarkers	MicroRNA	Domains	WikiPathways
Monocyte	349	107	209	114	2,902	1,005	42	451	1,202	242
CD4⁺ T cell	317	135	278	77	4,077	1,806	47	754	1,401	405
CD20⁺ B cell	323	103	170	88	2,778	821	39	948	950	288

Number of unique functional enrichments of gene expression information alone and putative transcriptional target genes.
*** Wilcoxon signed-rank test p < 0.01 (except for MicroRNA and Domains)**

Gene expression information alone										
Monocyte	43	0	35	11	237	101	7	314	404	58
CD4⁺ T cell	47	0	19	9	301	165	9	136	397	81
CD20⁺ B cell	42	0	27	12	239	247	6	370	409	65
Putative transcriptional target genes										
Monocyte	95	16	127	12	1,271	242	17	97	303	105
CD4⁺ T cell	105	26	146	23	1,654	415	24	224	585	133
CD20⁺ B cell	93	23	96	23	1,192	329	16	231	397	106

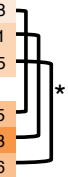

Table 2. Functional enrichments of putative transcriptional target genes using Pathway Commons

Monocyte - Pathway Commons	No. of TFs
Proteoglycan syndecan-mediated signaling events	18
Regulation of CDC42 activity	15
LKB1 signaling events	14
Glypican pathway	13
Interferon Signaling	13
Sphingosine 1-phosphate (S1P) pathway	13
IL5-mediated signaling events	12
Syndecan-1-mediated signaling events	12
IL3-mediated signaling events	11
Mitotic Prophase	11
Golgi Cisternae Pericentriolar Stack Reorganization	11
Interferon alpha/beta signaling	11
IFN-gamma pathway	10
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	10
Recruitment of mitotic centrosome proteins and complexes	10

CD4⁺ T cell - Pathway Commons	No. of TFs
TCR signaling in naive CD8⁺ T cells	36
IL12-mediated signaling events	24
Downstream signaling in naive CD8⁺ T cells	21
TCR signaling in naive CD4⁺ T cells	21
IL12 signaling mediated by STAT4	20
Validated transcriptional targets of AP1 family members Fra1 and Fra2	19
CXCR4-mediated signaling events	17
ATF-2 transcription factor network	15
Thrombin/protease-activated receptor (PAR) pathway	14
TCR signaling	14
PAR1-mediated thrombin signaling events	14
Downstream TCR signaling	14
Internalization of ErbB1	13
Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	13
ErbB receptor signaling network	13

CD20⁺ B cell - Pathway Commons	No. of TFs
Interferon alpha/beta signaling	12
Alpha6Beta4Integrin	11
Validated targets of C-MYC transcriptional activation	11
IL8- and CXCR2-mediated signaling events	10
Antigen processing-Cross presentation	9
BCR signaling pathway	9
IL6-mediated signaling events	9
Cell junction organization	9
ER-Phagosome pathway	8
Regulation of CDC42 activity	8
CXCR4-mediated signaling events	8
CDC42 signaling events	8
Syndecan-4-mediated signaling events	8
Noncanonical Wnt signaling pathway	7
Class I MHC mediated antigen processing & presentation	7

Table 3. Normalized number of functional enrichments of putative transcriptional target genes using combined promoter–enhancer regions.
*** Wilcoxon signed-rank test $p < 0.05$**

	KEGG	CTD Ontology	GO Slim	GO	Pathway Commons	Domains	Wiki Pathways
Monocyte							
Promoter	8.46	5.07	2.76	70.34	24.36	29.13	5.87
Enhancer 1	10.44	6.48	2.01	133.54	44.75	42.19	11.10
Enhancer 2	9.03	6.29	1.45	125.13	38.69	40.17	8.30
Enhancer 3	8.06	5.29	1.37	106.60	24.25	38.96	7.62
Enhancer 4	11.47	8.22	2.46	164.18	71.42	47.85	12.78
CD4⁺ T cell							
Promoter	5.30	4.64	1.29	68.11	30.17	23.41	6.77
Enhancer 1	13.60	7.07	2.74	142.40	84.51	43.78	13.65
Enhancer 2	13.57	6.69	3.05	141.15	86.36	46.33	12.02
Enhancer 3	12.40	5.89	2.50	115.76	68.84	41.85	10.00
Enhancer 4	16.40	7.86	4.03	177.55	108.08	53.86	16.72
CD20⁺ B cell							
Promoter	7.02	3.70	1.91	60.39	17.85	20.65	6.26
Enhancer 1	8.88	6.21	2.59	104.55	59.32	34.34	8.29
Enhancer 2	8.60	5.32	1.55	105.34	57.31	38.05	9.95
Enhancer 3	9.01	5.28	1.42	88.85	26.49	35.17	8.26
Enhancer 4	9.95	6.62	3.07	134.46	90.99	41.30	10.67

Table. 4 Normalized number of functional enrichments of putative transcriptional target genes using CTCF binding sites. * Wilcoxon signed-rank test $p < 0.05$

Monocytes	KEGG	CTD Ontology	GO Slim	GO	Pathway Commons	Domains	Wiki Pathways
Enhancer 4	11.47	8.22	2.46	164.18	71.42	47.85	12.78
CTCF (FR+RF+FF+RR)	13.19	10.39	2.74	134.26	34.37	44.46	12.96
CTCF (FR)	42.92	19.53	5.66	509.86	196.58	112.14	35.11
CD4⁺ T cells							
Enhancer 4	16.40	7.86	4.03	177.55	108.08	53.86	16.72
CTCF (FR+RF+FF+RR)	26.33	8.73	5.11	206.05	130.71	57.53	23.56
CTCF (FR)	69.39	14.66	24.91	560.44	220.54	133.26	46.54
CD20⁺ B cells							
Enhancer 4	9.95	6.62	3.07	134.46	90.99	41.30	10.67
CTCF (FR+RF+FF+RR)	8.78	4.22	2.78	94.13	27.70	28.55	7.08
CTCF (FR)	28.86	9.72	6.61	304.01	220.77	99.89	22.68
H1-hESC							
Enhancer 4	8.22	4.07	3.46	133.30	34.04	43.09	4.67
CTCF (FR+RF+FF+RR)	13.06	5.03	2.85	130.99	23.41	45.89	7.52
CTCF (FR)	28.08	10.62	10.19	334.14	70.10	160.05	17.15