Supplementary Materials for

Dynamic Trans-Acting Factor Co-localization in Human Cells

Supplementary Methods

Figures S1-S5

Tables S1

Supplementary Methods

TFs + conditions (described by appended underscores) included in the K562 SOM:

ARID3A ATF1 ATF3 BACH1 BCL3 BCLAF1 BDP1 BHLHE40 BRF1 BRF2 CBX3 CCNE1 CCNT2 CEBPB CHD1 CHD2 CTCF CTCFL E2F4 E2F6 EGR1 ELF1 ELK1 EP300 ESR1 ETS1 EZH2 FOS FOSL1 GABPB1 GATA1 GATA2 GTF2B GTF2F1 HDAC1 HDAC2 HDAC6 HDAC8 HMGN3 ILF2 IRF1_ifna30 IRF1_ifna6h IRF1_ifng30 IRF1_ifng6h JUNB JUND JUN_ifna30 JUN_ifna6h JUN_ifng30 JUN_ifng6h JUN KDM5B MAFF MAFK MAX MAZ MEF2A MLL5 MXI1 MYC_ifng30 MYC MYC_ifna30 MYC_ifna6h MYC_ifng6h NFE2 NFYA NFYB NR2C2 NR2F2 NR4A1 NRF1 PHF8 PML POLR2A_b POLR2A_48h POLR2A POLR2A_ifng30 POLR2A_phos POLR2A_ifna30 POLR2A_ifna6h POLR2A_ifng6h POLR3A POLR3G RAD21 RBBP5 RCOR1 RDBP REST RFX5 SAP30 SETDB1_mnased SETDB1 SIN3A SIRT6 SIX5 SMARCA4 SMARCB1 SMC3 SP1 SP2 SPI1 SRF STAT1_ifng30 STAT1_ifng6h STAT1 STAT1_ifna30 STAT1_ifna6h STAT2_ifna30 STAT2_ifna6h STAT5A TAF1 TAF7 TAL1 TBL1XR1 TBP TEAD4 THAP1 TRIM28 UBTF USF1 USF2 XRCC4 YY1 ZBTB33 ZBTB7A ZNF143 ZNF263 ZNF274

TFs included in the "deep comparison" of K562 and GM12878:

ATF3 BCL3 BCLAF1 BHLHE40 CEBPB CHD1 CHD2 CTCF E2F4 EGR1 ELF1 ELK1 EP300 ETS1 EZH2 FOS GABPB1 JUND MAX MAZ MEF2A MXI1 MYC NFE2 NFYA NFYB NR2C2 NRF1 PML POLR2A POLR2A_phos POLR3G RAD21 RCOR1 REST RFX5 SIN3A SIX5 SMC3 SP1 SPI1 SRF STAT1 STAT5A TAF1 TBL1XR1 TBP USF1 USF2 YY1 ZBTB33 ZNF143 ZNF274

TFs included in the "broad comparison" of K562, GM12878, HepG2, HeLaS3, and H1hESC:

CEBPB CHD2 CTCF EP300 EZH2 GABPB1 JUND MAX MXI1 MYC NRF1 POLR2A RAD21 REST RFX5 TAF1 TBP USF2

Significance test for co-localization patterns:

test for 1 cell type:

We used the p-value of a binomial test to represent the enrichment. For each window, $P(1|\theta)$ indicates the probability that the window contains the TF combination we are interested, $P(0|\theta)$ indicates the probability that the window does not have the TF combination. The parameter θ consists of $\{p_1, p_2, \dots, p_i\}$ whereiis the index of the TFs. p_i is the probability of TFi exists in the window, which is the ratio (#windowscontainTFi/#totalwindows).

$$\theta = \prod_{TF\ in\ the\ combination} p_i \times \prod_{TF\ not\ in\ the\ combination} (1 - p_i)$$

Let k be the number of windows in one neuron, then $B(k|n,\theta)$ follows a binomial distribution.

$$pvalue = \sum_{m=k}^{min(TFwindow)} B(m|n,\theta)$$

test for multiple cell types:

The null distribution is the sum of binomial distributions: $\sum B(k_i|n_i,\theta_i)$

as a good approximation, we could assume all the θ_i are close enough so we could use the same θ (#windowscontainTFiinallcelltypes/#totalwindowsinallcelltypes).

$$\sum B(k_i|n_i,\theta) = B(\sum k_i \,|\, \sum n_i, \theta)$$

RPKM and conservation scores:

Raw pair-end RNA-seq data generated from Cold Spring Harbor Laboratory for the ENCODE project were downloaded from UCSC genome browser server. The raw data were mapped to human genome hg19 using TopHat v1.4.1 with default parameters. The genome reference and gene annotation files are downloaded from TopHat website. The mapped reads are then processed using cufflinks v1.1.0 with default parameters to get RPKM for each gene.

Phastcon and PhyloP scores of 45 vertebrate genomes with human hg19 were downloaded from UCSC genome browser server. Customized software was written to extract average, maximum, and minimum conservation scores for each CRM.

DNase Overlap:

Narrowpeak files for DNaseI obtained for K562 from the UCSC ENCODE portal were used to overlap with ChIP-seq peak regions or CRMs. Peak regions or CRMs were considered overlap with DNaseI sites if there was more than 1bp overlap. This overlap requirement allowed for a very generous estimation of DNase I overlap with our CRMs.

GO and Pathway Analysis:

GO and pathway enrichment analysis for the genomic regions in each SOM neuron was conducted using R "ChIPpeakAnno" package. GO annotations were defined from GO.db version 2.7.1 and pathway enrichment was defined from reactome.db version 1.0.40. In general, the genes were assigned to each CRM according to the closest TSS. TSS annotations were obtained for human annotation GRCh37 from biomaRt. GO and pathway enrichment scores were computed based on the assigned genes in each neuron. Multiple testing adjustment was made using Benjamini and Yekutieli procedures. For multi-cell-type SOM, GO and pathway enrichment analysis was performed on all the genomic regions in each neuron and on genomic regions specific to each cell type in each neuron. The adjusted p-values for the enriched GO and pathway terms were clustered using hierarchical clustering method.

Protein identification by mass spectrometry

The immunoprecipitation samples were subjected to protein separation by NuPAGE Bis-Tris precast gel (Invitrogen) along with IgG control samples. Protein bands were visualized and destained by Pierce Silver Stain Kit for Mass Spectrometry (Thermo Scientific). Each protein lane was cut into 10 slices and digested with trypsin. Tryptic peptides were cleaned by C18 tip and analyzed by LC-MS/MS on an LTQ Orbitrap.

Mass spectrometry raw data were searched against a Uniprot human proteome database (last modified on May 18th, 2012), using the SEQUEST algorithm (Proteome Discoverer software, version 1.3, Thermo Scientific). Searches were performed using a 10 ppm mass tolerance for precursor ions and 0.8 Da for fragment ions, allowing up to two trypsin missed cleavages. Oxidation of methionine residues (+ 15.995 Da) was set as a variable modification; carbamidomethylation of cysteine residues (+57.021 Da) was set as static modifications. Peptides with high confidence (false positive rate < 1%) and rank 1st were selected.

Each sample was analyzed by LC-MS/MS with one to three replicates. Proteins identified with at least two high confidence peptides were considered as positive identification. After removing keratins, proteins with average spectral counts in the immunoprecipitation sample greater than that in the parallel control sample were selected for the following PPI network construction. A

total of 50 antibodies were used in this study. All of them can specifically pull down their targeted proteins according to the IP-MS data (Table S1).

Fig. S1. P-value of each co-localization pattern is shaded on the K562 SOM. Related to Figure 1, white neurons are significant co-localizations (small p-values).

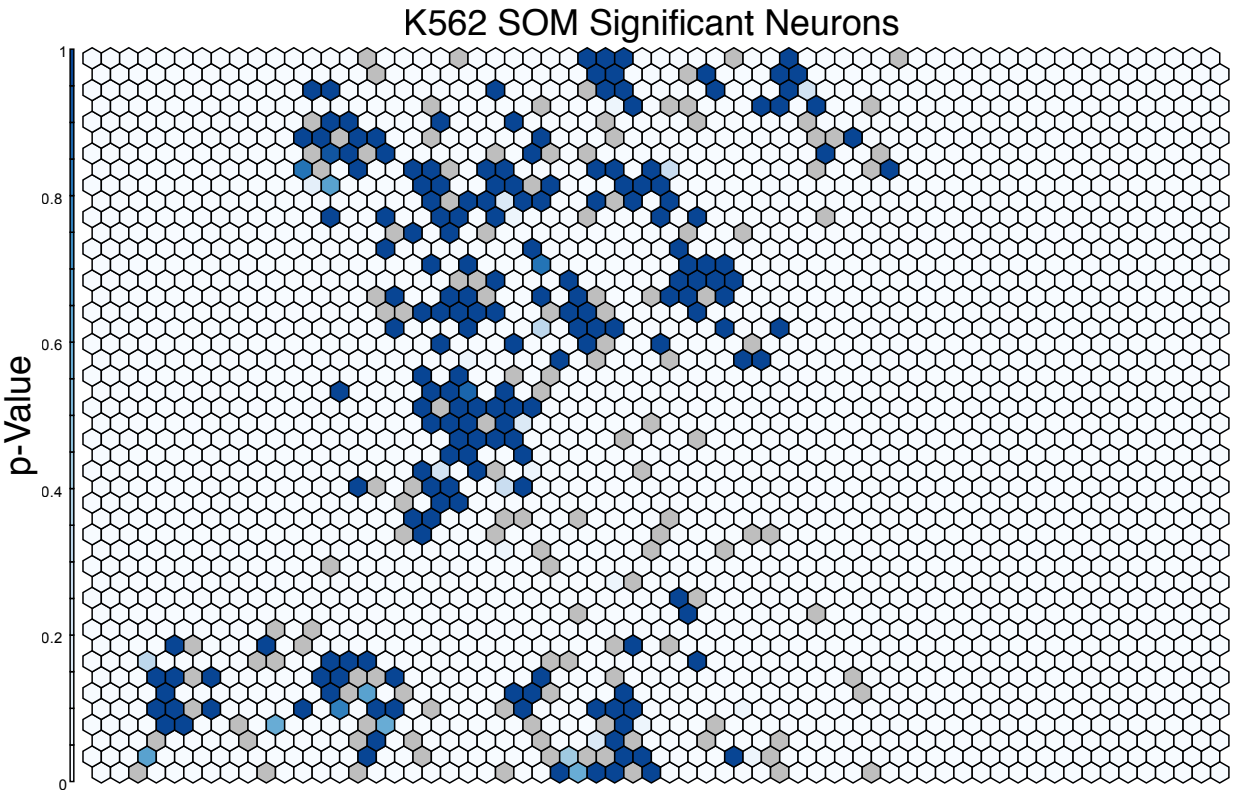


Fig. S2. (A) Co-immunoprecipitation of c-Fos and NFYB in K562. Related to Figure 2, Anti-c-Fos (sc-7202, Santa Cruz) IP samples were subjected to western blotting analysis using anti-NFYB antibody (sc-376546). Lane 1, input lysate of anti-c-Fos IP; lane 2, supernatant of anti-c-Fos IP; lane 3, precipitation of anti-c-Fos IP; lane 4, precipitation of control IgG. **(B)** Complexity network representing the TF co-localizations with the most context combinations. The nodes are TFs and the edges are "complexity" of the contexts that contain the TF pairs. The complexity is computed as the number of different co-localization patterns the TF pairs resided. Well-known co-localizations such as CTCF-RAD21-SMC3, EP300-TAL1, and MYC-MAX are included in these sets.

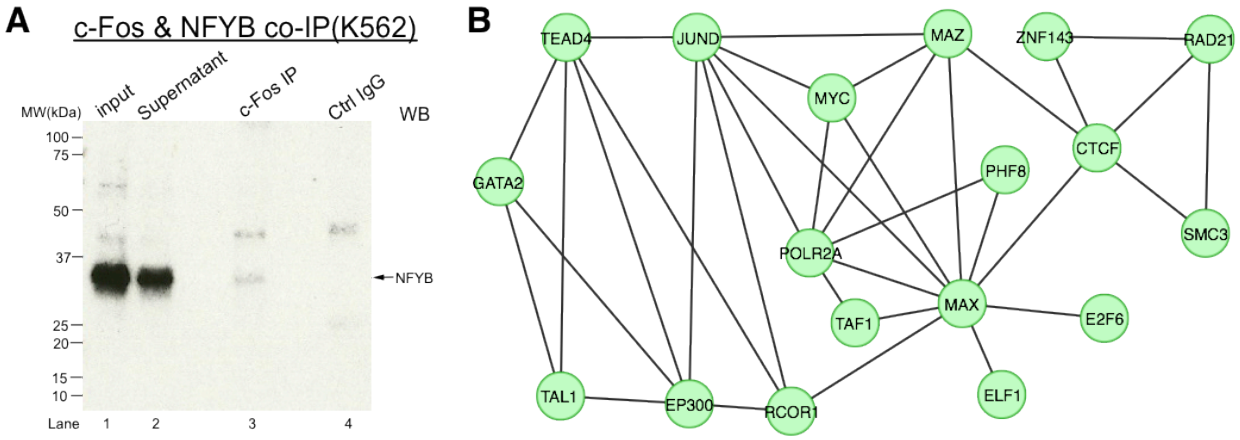


Fig. S3. Related to Figure 3, average PhyloP conservation scores for CLPs in the SOM reveal that HOT regions have a very different pattern of conservation that other genomic loci. These regions tend to have a slightly higher average conservation score (top panel), a higher average maximum conservation score (middle panel), and a lower average minimum conservation score (bottom panel).

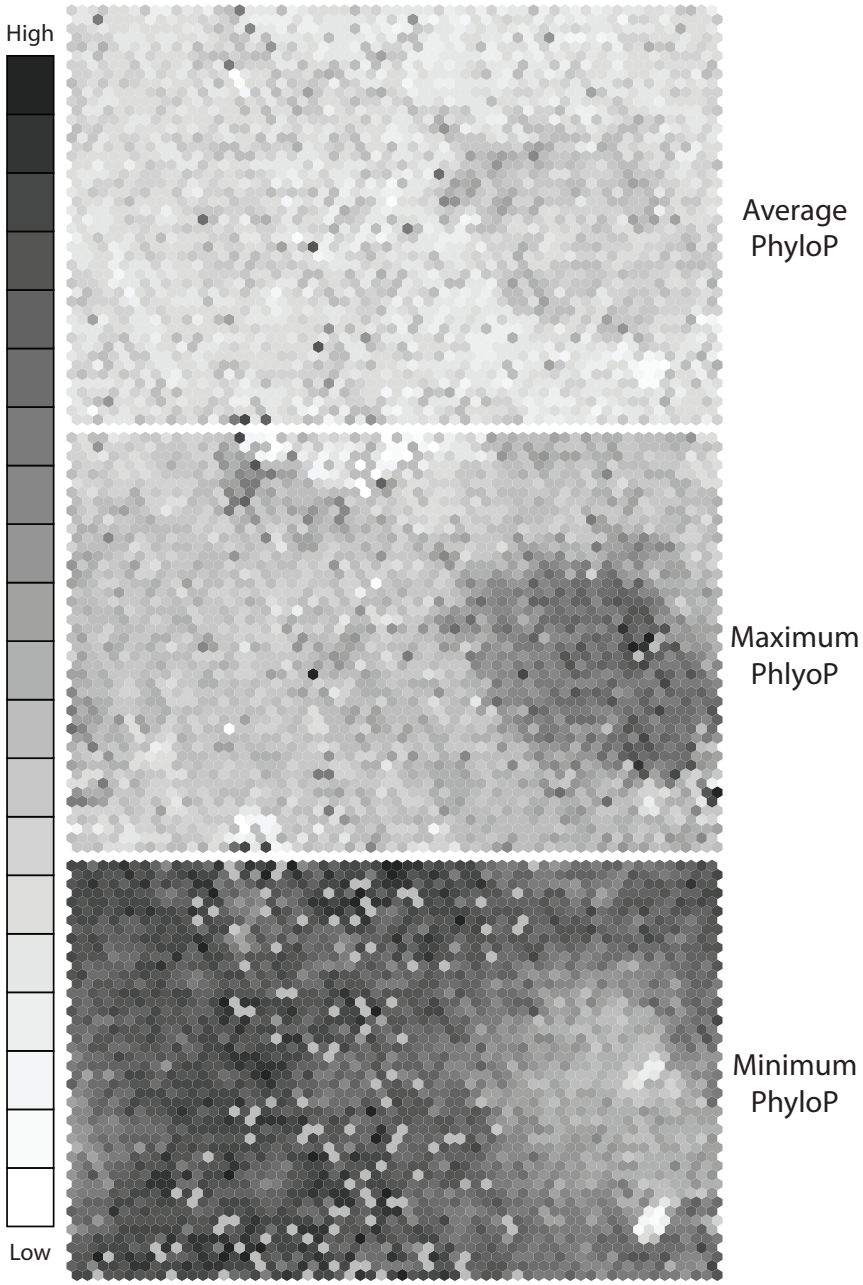


Fig. S4. Percentage of TF binding sites and CRMs overlap with DNaseI hypersensitive sites in K562. Related to Figure 3. **(A)** DNaseI sites overlap with more than 80% of binding sites for about 75% of the TFs assayed in K562 cell line. The binding sites of about 25% of the TFs overlap with DNaseI sites with significantly lower rates. The overlap rate is lower than 50% for 10 TFs. 60% of CRMs overlap with DNaseI sites. **(B)** Percentage differences of TF binding sites overlap with DNaseI between proximal (<5000bp to TSS) and distal (>10000bp to TSS) locations. The overlap rates are significantly lower at remote binding regions for almost all the TFs. The DNaseI hypersensitivity data produced using different methods from two ENCODE labs show different patterns of overlap with CLPs. **(C)** The data from the University of Washington shows partial overlap with more total CLPs (that is, not all members of a CLP have overlap with a DNaseI hypersensitive site). However, the overlap of each CLP is less complete. **(D)** The data from Duke University shows complete overlap of more CLPs (that is, all members of a CLP are covered by a DNaseI hypersensitive site), but fewer CLPs with partial overlap. **(E)** Histogram and cumulative distribution of the distances of CRMs to their nearest DNase I sites. **(F)** Bar plot showing the percentage of DNaseI overlapping and no DNaseI overlapping CRMs in each chromatin state.

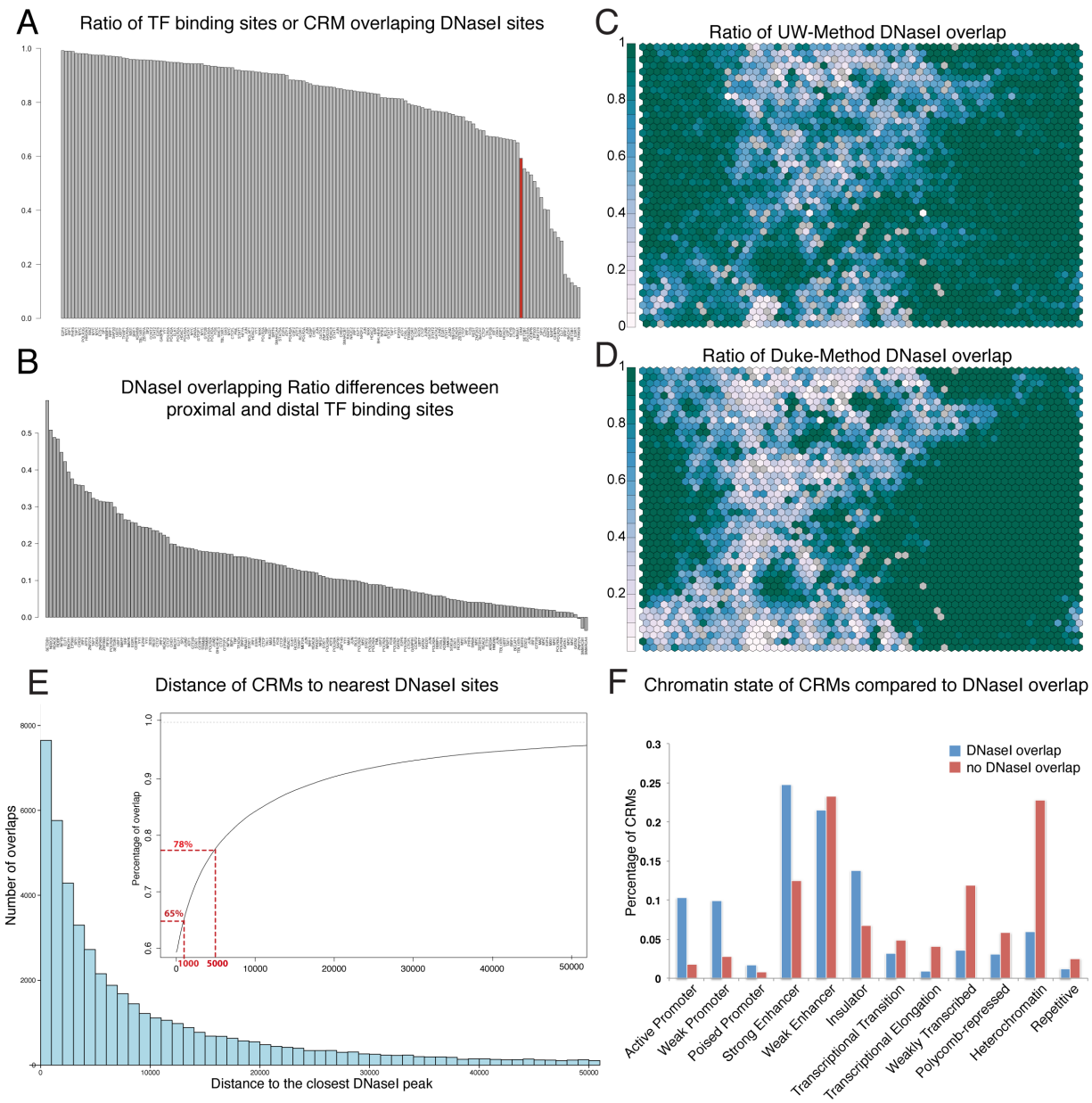
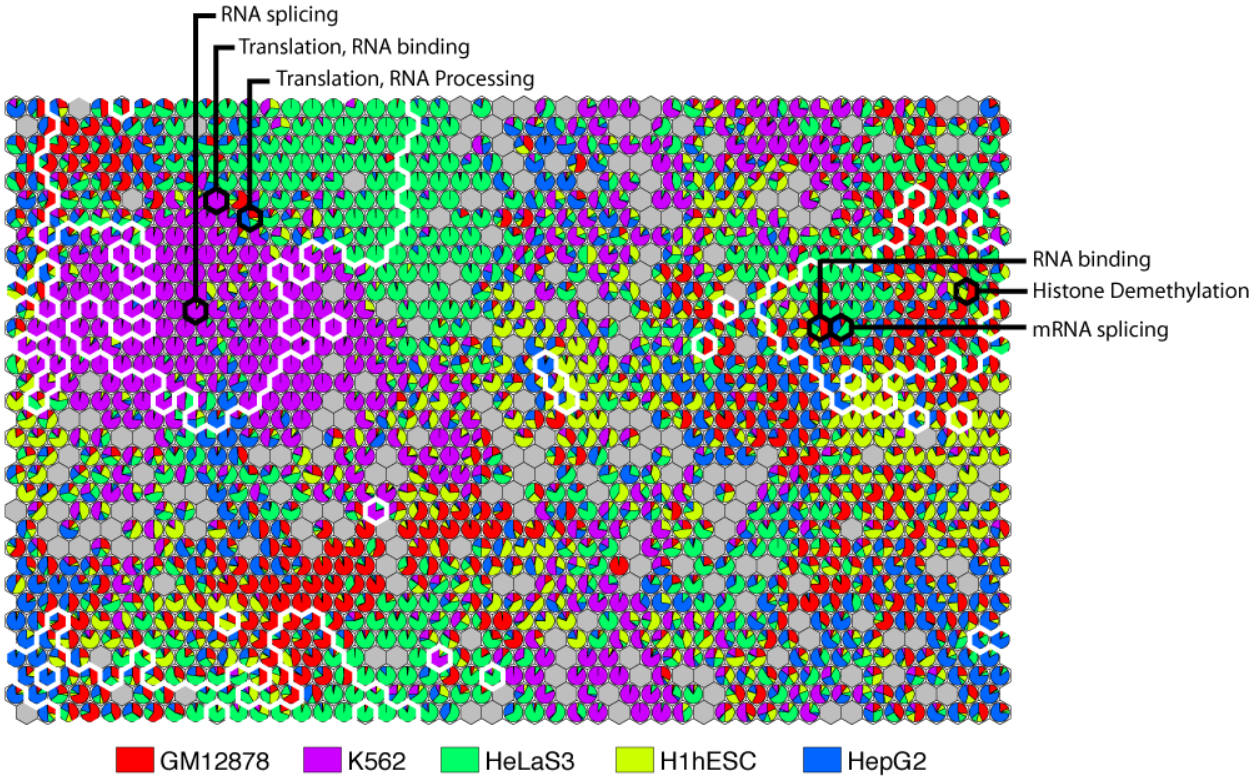

A
Ratio of TF binding sites or CRM overlaping DNaseI sites
B
DNaseI overlapping Ratio differences between proximal and distal TF binding sites
C
Ratio of UW-Method DNaseI overlap
D
Ratio of Duke-Method DNaseI overlap
E
Distance of CRMs to nearest DNaseI sites
Number of overlaps
Distance to the closest DNaseI peak
Percentage of overlap
78%
65%
1000
5000
F
Chromatin state of CRMs compared to DNaseI overlap
DNaseI overlap
no DNaseI overlap
Percentage of CRMs
Active Promoter
Weak Promoter
Poised Promoter
Strong Enhancer
Weak Enhancer
Insulator
Transcriptional Transition
Transcriptional Elongation
Weakly Transcribed
Polycomb-repressed
Heterochromatin
Repetitive

Fig. S5. A SOM trained from 18 shared TFs for GM12878, K562, HeLaS3, H1hESC, and HepG3 cell types. Related to Figure 5, the proportion of the cell types where a co-localization pattern bind is represented by a pie chart in the neuron. The five cell types are represented with five different colors. The HOT regions are circled by white lines. Housekeeping GO terms are enriched in the HOT regions and some are shared by cell types.



Supplemental Tables:

Table S1. The results of protein-protein interaction network derived from IP-MS data showing spectral counts of control and TF IP in all replicates. Related to Figure 6.