www.sciencemag.org/cgi/content/full/science.aab1601/DC1

Supplementary Materials for

Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing

Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, Jay Shendure*

*Corresponding author. E-mail: shendure@uw.edu

Published 7 May 2015 on *Science* Express
DOI: 10.1126/science.aab1601

This PDF file includes:

Materials and Methods
Figs. S1 to S22
Table S1
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.aab1601/DC1)

Table S2

Materials and Methods

Cell Culture

All cells were cultured in incubators at 37°C with 5% CO2. GM12878 and HL-60 were maintained in RPMI 1640 medium (Gibco cat. no. 11875) containing 15% FBS, 100U/ml penicillin and 100μg/ml streptomycin. GM12878 flasks were counted and split to 300,000 cells/ml three times a week. HL-60 flasks were counted and split to 100,000 cells/ml three times a week. Patski embryonic kidney fibroblast and HEK293T cells were maintained in DMEM (Gibco cat. no. 11965) with 10% FBS, 100U/ml pencillin and 100μg/ml streptomycin. Patski cells were trypsinized three times a week and split 1:5 unless they were less than 50% confluent, in which case media was replaced and cells were allowed to continue growth. HEK293T cells were trypsinized and split 1:10 three times a week.

Sample Processing

Adherent lines were trypsinized, and then both adherent and suspension cells were washed once in PBS. Cells were combined and then lysed (2,500 cells per tagmentation reaction) using cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630; (4)) supplemented with protease inhibitors (Sigma). The isolated nuclei were then pelleted and resuspended in Nextera TD buffer (Illumina) for tagmentation. Nuclei were distributed onto 96-well plates and 2.5μM Tn5 (each well of the plate contained a unique barcode combination) was added to each well as previously described (17, 18). The tagmentation reaction was carried out for 30 minutes at 37°C and then the reaction was stopped by adding 40mM EDTA to each well. The nuclei were incubated for another 15 minutes at 37°C and then all nuclei were pooled. The pooled nuclei were then sorted on a FACSAria II cell sorter (BD). 15 nuclei (on the basis of forward- and side-scatter) were sorted into each well of a 96-well plate containing 20μl of EB buffer for the human/mouse mixture experiment. For the two experiments involving mixtures of human cells, nuclei were first stained with DAPI (Invitrogen) at a final concentration of 3μM and then 25 DAPI-positive nuclei were sorted into each well of a plate with 20μl EB buffer. Tn5 was released from the DNA and libraries were amplified 20 cycles while incorporating standard Nextera library barcodes following the protocol described in (17, 18). After the initial three experiments described in the paper, we switched to a set of 10bp library barcodes that were otherwise identical to the standard Nextera library adapters to allow us to sequence multiple experiments in parallel. After PCR, samples were pooled and cleaned with a Clean & Concentrator kit (Zymo). Sample concentrations were determined by Qubit (Invitrogen) and the libraries were run on a diagnostic 6% PAGE gel.

Sequencing

Libraries were sequenced on a MiSeq (Illumina) using a v2 300 cycle kit and a custom recipe (paired end 51bp reads with index reads that covered both the Tn5 barcode and the library amplification barcode) and custom primers as described in (17, 18). Libraries were loaded at 15pM. Base calls were converted to the qseq format with Illumina's offline basecaller (OLB v1.9.3) and then converted to fastq format with a custom pipeline. Barcodes that did not match any of the expected barcodes exactly were converted to the closest barcode if the edit distance was no greater than 3 and there were no secondary matches within an additional 1 base pair edit distance. Subsequently, reads were trimmed with Trimmomatic (24) and then mapped to an appropriate reference genome with BWA (25). For the human/mouse mixture, we created a chimeric reference genome of hg19 and mm9 after removing sequences that were not a part of the reference chromosomes. The human cell line mixtures were mapped to hg19 after removing sequences that were not a part of the reference chromosomes. Reads mapping to the mitochondrial genome and reads mapping with a quality less than 10 were removed. Finally, all fragments in the same library with duplicate start and end coordinates were removed using Picard (http://broadinstitute.github.io/picard). For RNA-seq, RNA was extracted with the RNeasy kit (Qiagen) and libraries were constructed with the standard TruSeq kit (Illumina). RNA quality and concentration were confirmed by Agilent Bioanalyzer, using the RNA 6000 Nano kit. Samples were multiplexed and sequenced (75bp single-end) on an Illumina NextSeq. Libraries were then mapped to the human genome (hg19) with TopHat (26) and gene expression levels were quantified with Cufflinks and Cuffdiff 2 (27).

Data Analysis

Calculating the barcode collision rate

An important parameter of our experimental design is the barcode collision rate (the fraction of barcodes that represent more than one nucleus out of all barcodes observed in a given experiment). The expected rate can be calculated in a straightforward manner based on the classic birthday problem (28) as a function of the total number of barcodes available to draw from (96 tagmentation barcodes in our experiments) and the number of nuclei sorted into each well of the PCR plate. If we assume that we are equally likely to sort nuclei from any of the tagmentation reactions into any given well of the PCR plate, we can calculate the expected collision rate as 96 – the expected number of barcodes representing one or zero nuclei. The expected number of barcodes not representing any nuclei in an experiment can be calculated as:

$$(\#\ of\ barcodes) * (1 - \frac{1}{\#\ of\ barcodes})^{(\#\ of\ nuclei\ sorted)} \qquad (1)$$

And the number of barcodes expected to represent exactly one cell is:

$$(\#\ of\ nuclei\ sorted) * (1 - \frac{1}{\#\ of\ barcodes})^{(\#\ of\ nuclei\ sorted - 1)} \qquad (2)$$

Because some collisions will involve nuclei of the same type, we can only observe approximately half of the barcode collisions that actually occur in our experiments. To estimate the actual barcode collision rate in each experiment, we: (1) Calculated the number of nuclei that appeared to be mixtures of two cell types; (2) Based on the ratio of cell types to one another within any given experiment, adjusted this count to account for the fact that collisions involving nuclei of the same cell type will be unobserved; (3) Made the simplifying assumption that collisions never involved more than two nuclei (as this should be true for the vast majority of collisions). We observe relatively high variance in the representation of the 96 transposase barcodes in our experiments (Fig.

S1A shows distribution of transposase barcodes for the all experiments experiments; Fig. S1B shows the distribution of PCR barcodes for comparison in the first three experiments described in the main text), however, our observed collision rates for the experiments seem well aligned with the theoretical expectation. We tried to estimate the effect the non-uniform distribution of barcodes might have on our observed collision rate by simulating barcodes drawn from a random Poisson process with the relative probabilities of each barcode set by the observed frequencies of transposase barcodes. However, the resulting total collision rates estimated for each experiment seemed to match reasonably well with the theoretical expectation (**Fig. S1C)**.

Determining accessible hypersensitive sites in single cells

To identify sites of accessible chromatin in individual cells, we first created reference maps of hypersensitive sites independently determined by the ENCODE Consortium using DNase I sequencing. To catalog hypersensitive sites in the GM12878/Patski comparison and the GM12878/HL-60 comparison, we downloaded hypersensitivity maps produced by the Stamatoyannopoulos lab from the ENCODE website (https://www.encodeproject.org). This group identified hypersensitive sites using an algorithm called 'Hotspot' (*29*) and provided hypersensitivity maps for two replicates for each sample. For each sample we downloaded the 'hotspot' files for each replicate. For the comparison of GM12878 and Patksi cells, hotspots for replicate samples were intersected and used as reference hypersensitive sites using BedTools (*30*). For the comparisons between GM12878 and HL-60, replicates were intersected separately for each cell line and then sites from the two cell lines were merged into a single reference using BedTools. For the GM12878 and HEK293 comparison, we downloaded hypersensitive sites produced by the Crawford lab (as maps for HEK293 were not produced by the Stamatoyannopoulos lab). This group produced hypersensitive maps (without replicates) using an algorithm called 'F-seq' (*31*). The DHS maps from each cell line were merged into a single reference using BedTools. In all analyses, hotspots or F-seq peaks overlapping regions of the genome blacklisted by ENCODE (https://sites.google.com/site/anshulkundaje/projects/blacklists) were filtered out. For the identification of sites differentially accessible between GM12878 and HL-60 (below), we also filtered out hypersensitive sites overlapping known CNVs from either cell line. HL-60 and GM12878 CNV coordinates determined by the ENCODE Consortium (*22*) were downloaded from http://genome.ucsc.edu. In addition, chromosomes 18, X and Y were excluded from analysis, as they are known to be aneuploid in HL-60. Reads overlapping ENCODE hypersensitive sites (*22*) were then identified with a custom python script using pysam (https://github.com/pysam-developers/pysam) to produce a matrix of the counts of reads overlapping each individual hypersensitive site (rows) for each individual cell (columns). This matrix was then converted to a binary matrix such that if a cell had at least one read overlapping a given hypersensitive site, it was considered accessible in that cell.

Comparing with bulk ATAC-seq

For comparison with previously published ATAC-seq data, we downloaded fastq files for the 500 cell replicates from GEO accession GSE47753 (*4*) and processed the samples through our mapping pipeline. Briefly, reads were first trimmed to for sequencing adapters and truncated to 51 bp with Trimmomatic, and then mapped to the human genome with BWA. Finally, duplicate reads were removed with Picard. After these steps, the number of reads overlapping each hypersensitive site were counted using pysam.

Evaluating the complexity of sequenced libraries

PCR duplication rates and library complexity estimates were determined for mappable (i.e. MAPQ >10), non-mitochondrial reads using the EstimateLibraryComplexity command in Picard. For comparison, the complexity of bulk ATAC-seq libraries generated from 500 cells (described above) was also estimated with Picard. To determine the number of sequence fragments per cell, the estimated total number of fragments in the library was divided by 500 (the number of cells included) for the bulk samples. Likewise, for the single cell library, the estimated total number of fragments from Picard was divided by 533 (the number of nuclei recovered in this experiment given our 500 read cutoff).

Identifying DHSs differentially accessible between cell types

Hypersensitive sites that are differentially accessible between GM12878 and HL-60 in our single cell ATAC-seq data were identified using logistic regression via the VGAM framework in R (*32*). For each site, accessibility was encoded as a binary response variable via the binomialff VGAM family function, which was supplied with default parameters. The best guess of each cell's type was used as a categorical predictor of accessibility. Cells were assigned to a particular cell type on the basis of the genome-wide relative proportion of reads mapping to cell line-specific hypersensitive sites (≥ 70% of cell type-specific reads mapping to DHSs from one of the two cell types). Cells that appeared to be a mixture of both cell types were excluded from this analysis. In addition, DHSs that were not observed in at least two cells and sites overlapping reported CNVs were excluded from this analysis. Statistical significance of observed differences in a site's accessibility between cell types was assessed by likelihood ratio test via the lrtest() function in VGAM. P-values were adjusted for multiple testing using the Benjamini and Hochberg method (*33*) and significant associations were identified at an FDR of 0.05.

Identifying enriched annotations for differentially accessible DHSs

Chromatin state information for GM12878 from the ENCODE Consortium was downloaded from the UCSC genome browser (http://genome.ucsc.edu/; "Combined" chromatin state track; (*22*, *34*)). To determine the chromatin states of differentially accessible sites we intersected bed files of the genomic coordinates for each using BedTools (*30*). To identify phenotypes associated with differentially accessible sites of the genome and modules of coordinately regulated chromatin accessibility, we first linked hypersensitive sites to the genes they regulate using links defined in (*2*) and then hypergeometric test function runGSAHyper() in the Piano package in R (*35*) to identify KEGG (*36*, *37*) and Reactome (*38, 39*) pathways enriched among the genes . As a background set for differentially accessible sites, we used all genes linked to accessible sites in GM12878 and HL-60 we had defined using ENCODE DNase I hypersensitivity data. The target set was the unique set of genes linked to the 1,666 sites that were differentially accessible in our logistic regressions. As a background set for modules of coordinately regulated chromatin accessibility, we used all genes linked to hypersensitive sites included in the latent semantic analysis.

Identifying enriched annotations for coordinately accessible modules of DHSs

To identify modules of DHSs that are significantly enriched for specific pathways or annotations, we first linked DHSs within modules to the genes they regulate and then tested for significant enrichments of annotation terms using the hypergeometric test function as described above. To identify modules of DHSs that significantly overlap transcription factor binding in GM12878, we downloaded all transcription factor ChIP peak data sets for that cell type from ENCODE and intersected these maps with sites in the different modules and then used a hypergeometric test to determine significant overlaps.

Dimensionality reduction of chromatin accessibility data

For multidimensional scaling, pairwise Jaccard distances were calculated between cells based on the binary hypersensitive site usage matrix. These distance were then used to represent relationships between the cells in two dimensions using the cmdscale() function in R. All cells and all sites that were observed in at least one cell were included in these calculations. For latent semantic indexing of the cellular mixtures, we first filtered out cells that did not have at least 400 sites open in the binary hypersensitivity site usage matrix. We also filtered out hypersensitivity sites that were not used in at least 150 cells. The binary site usage matrix was then transformed using a term frequency–inverse document frequency algorithm. To do this, each site used in a cell was weighted by the total number of sites used in the cell. This value was then multiplied by the log of 1 + the inverse frequency of the site across all cells. Singular value decomposition was then performed on this transformed matrix. To visualize sensitivity, we produced a lower dimension representation of the data using the first 6 components of this singular value decomposition and capping the LSI scores at +/- 1.5. Cells and sites were clustered using Ward clustering on components 2-6, as component 1 appeared to be related to read depth. For latent semantic indexing of the GM12878 cells alone, sites used by fewer than 10% of cells were filtered out, as were cells with fewer than 400 sites open. A lower dimension representation of the data was produced with 6 components, excluding the first component and capping the LSI scores at +/- 1.5. Hierarchical clustering was performed with the “average” algorithm in this analysis.

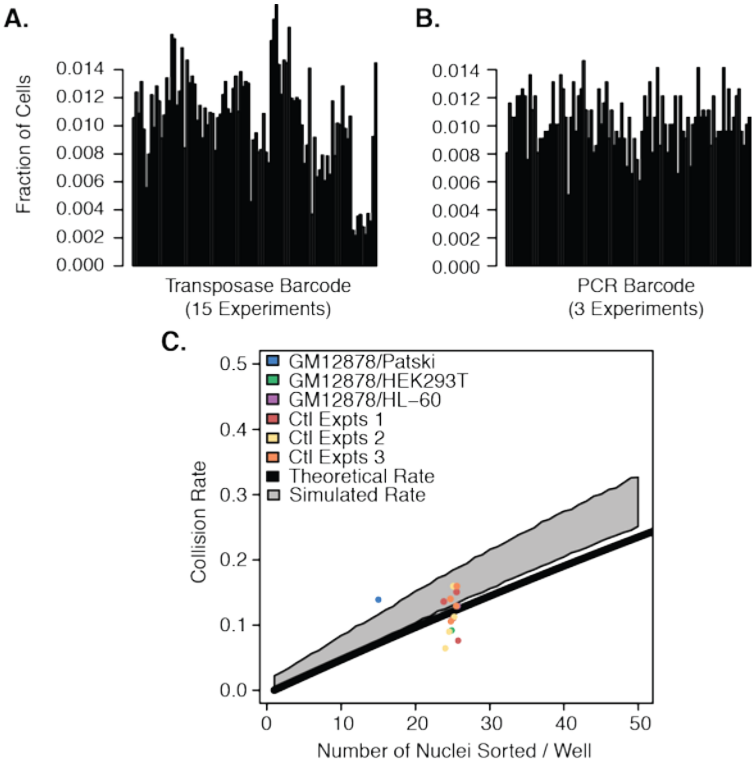

Fig. S1.

Expected and observed collision rate. (**A**) Barplot showing non-uniform distribution of transposase barcodes across all 15 experiments. (**B**) Barplot of PCR barcodes in three of the experiments (the other experiments used a different set of barcodes that cannot be directly compared here). (**C**) Observed/expected collision rate as a function of the number of nuclei sorted. Black line shows the expected collision rate based on the classic birthday problem. Gray ribbon shows 95% confidence interval of simulations that drew barcodes from a Poisson distribution while matching the observed frequencies of observed barcodes. Points show observed experimental collision rates. Points are jittered to avoid overplotting. Note: for the GM12878/Patski experiment, 15 unstained nuclei were sorted (instead of 25 DAPI-positive nuclei in the other experiments).

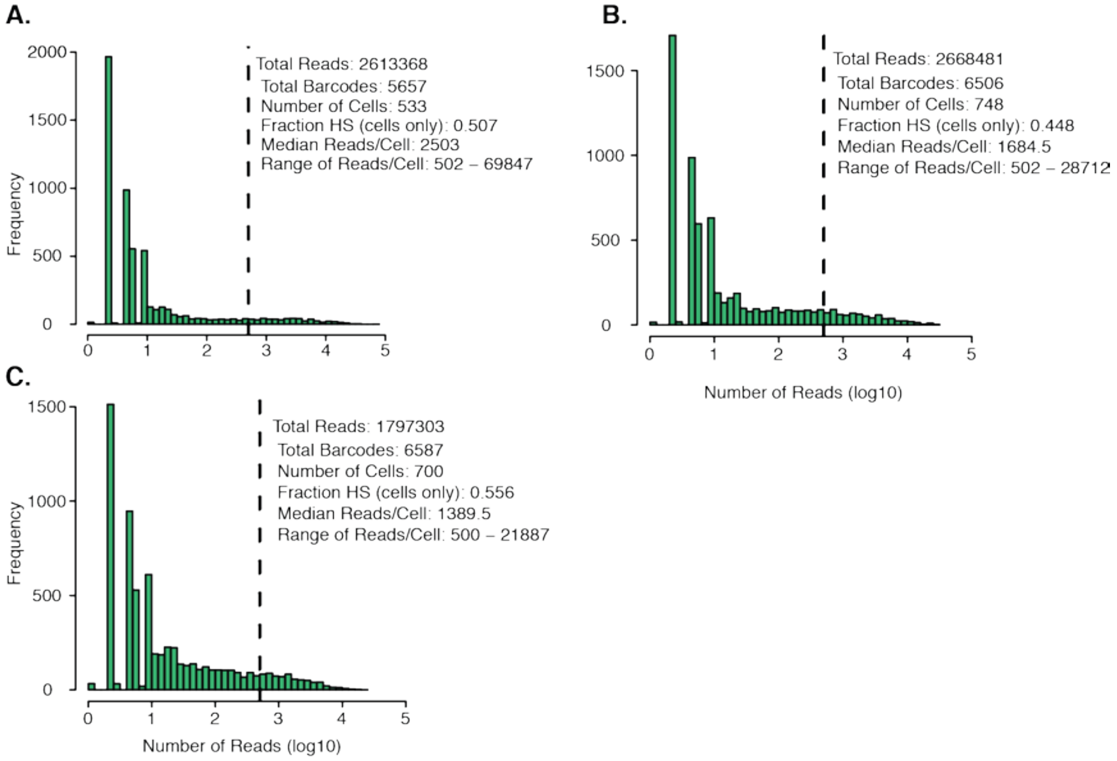

Fig. S2

Number of mappable reads assigned to each possible barcode combination. (**A**) Histogram of the number of reads (log-transformed) assigned to each barcode for the experiment mixing GM12878 (human) and Patski (mouse) cells. (**B**) Same as in (**A**) for the experiment mixing GM12878 and HEK293T cells. (**C**) Histogram for the experiment mixing GM12878 and HL-60. The dashed line marks 500 reads in each graph.

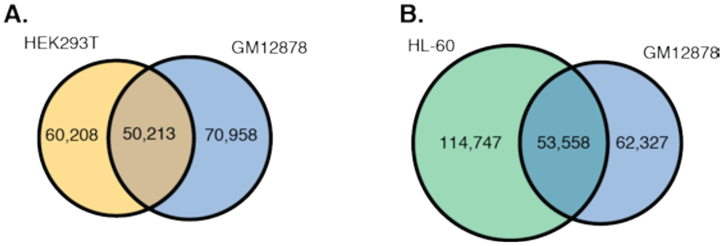

Fig. S3

Overlap of hypersensitivity sites from ENCODE. Venn diagrams depicting the overlaps of hypersensitive sites identified by the ENCODE Consortium for (**A**) GM12878 and HEK293T and (**B**) GM12878 and HL-60 (*24*). Note that there are different total numbers of sites for GM12878 in the two panels, because each pairwise comparison required the use of data from different labs with slightly different DNase-seq protocols.

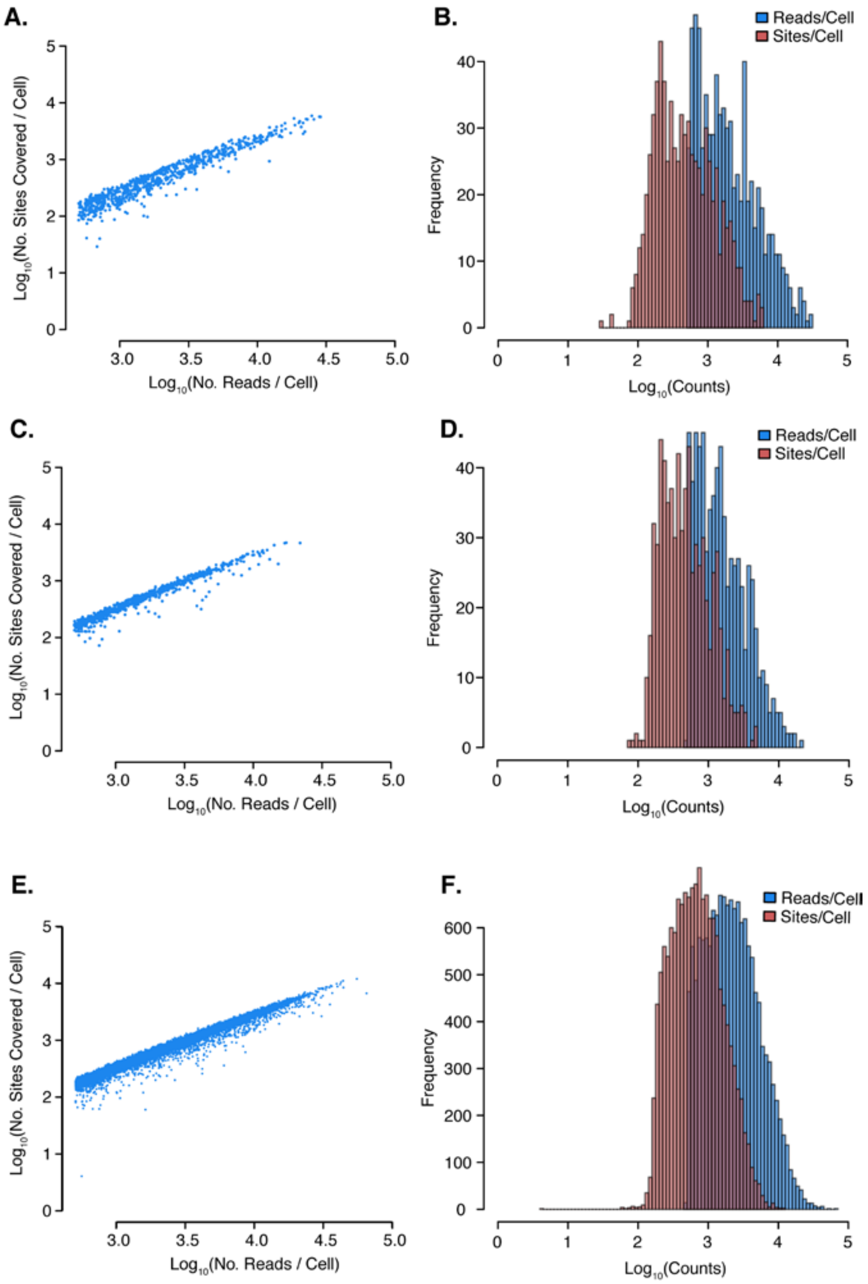

Fig. S4

Site coverage in individual cells compared to read depth in the human cell mixture experiments. (**A**,**C**,**E**) Scatter plot of the number of sites covered by each cell compared to the total number of reads recovered for that cell. (**B**,**D**,**F**) Histogram of the number of sites covered by each cell (in red) compared to the total number of reads recovered for that cell (in blue). (**A** and **B**) Site coverage for the GM12878/HEK293T mixture experiment. (**C** and **D**) Site coverage for the GM12878/HL-60 mixture experiment. (**E** and **F**) Site coverage for all GM12878/HL-60 experiments presented in the main text (includes the data from **C** and **D**).

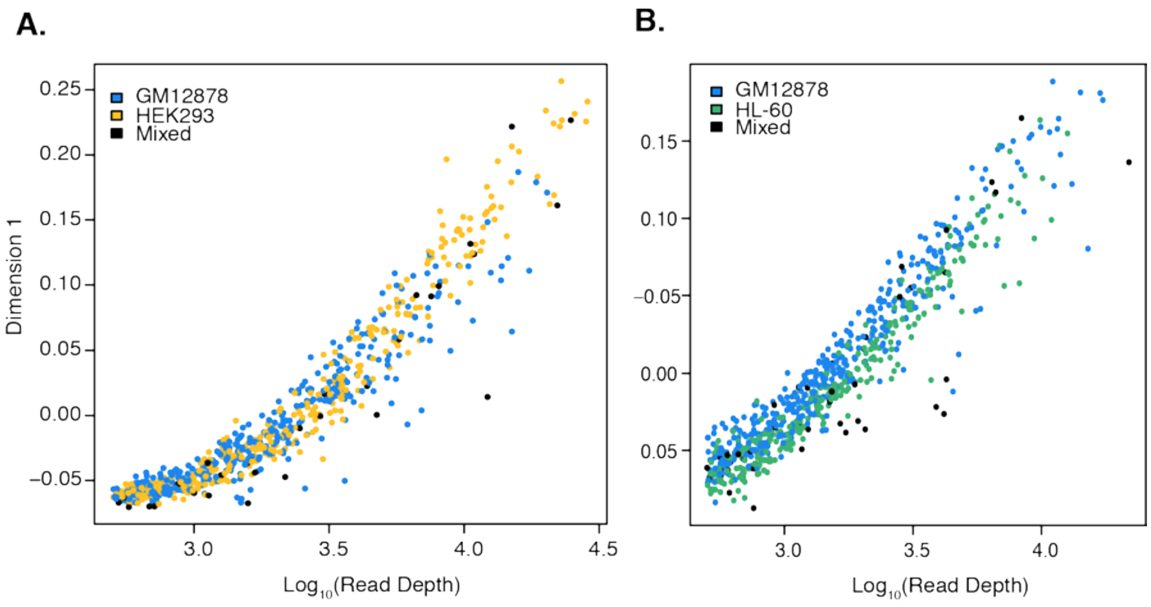

Fig. S5

Multidimensional scaling identifies read depth as a major factor in clustering single cells by cell type. (**A**) Plot of log-transformed read depth (x-axis) against dimension 1 of multidimensional scaling analysis (y-axis) based on Jaccard distances in experiment mixing HEK293T cells with GM12878 cells. (**B**) Same as in A for experiment mixing HL-60 and GM12878 cells.

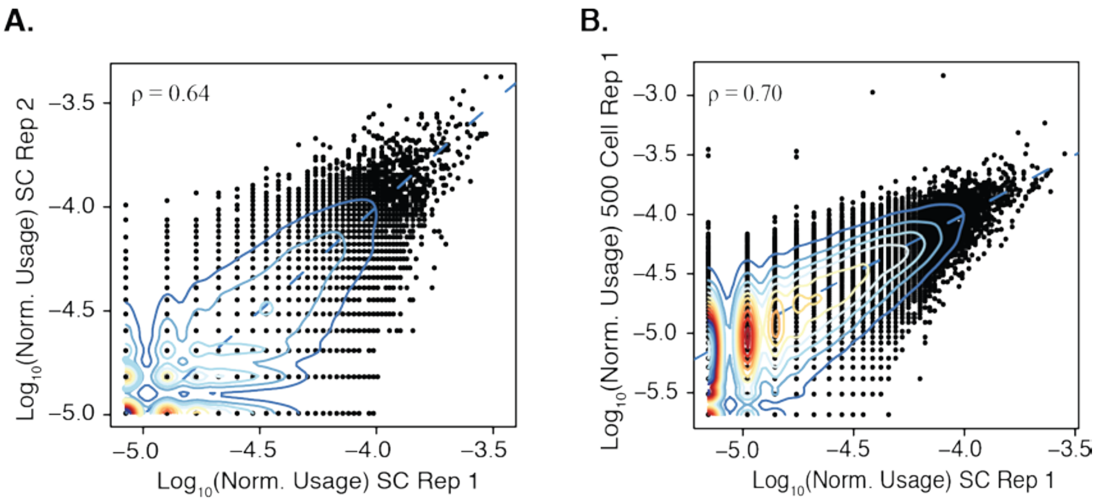

Fig. S6

Reproducibility of hypersensitive site usage for individual cells. (**A**) Scatterplot comparing site usage in GM12878 cells from two different experiments. X-axis: Log$_{10}$-transformed sum of GM12878 cells in which each site is used in the human/mouse mixture experiment normalized by the total site usage count for GM12878 cells in the GM12878/Patski experiment ('Normalized Usage'). Y-axis: Log$_{10}$-transformed normalized usage for GM12878 cells in the GM12878/HL-60 experiment. Dashed blue line is the identity line. (**B**) Log$_{10}$-transformed normalized usage for GM12878 cells in the GM12878/Patski experiment compared to the Log$_{10}$-transformed number of reads overlapping each hypersensitive site in a previously published bulk ATAC-seq sample generated from 500 cells (*4*) normalized by read depth across all DHSs. Dashed blue line is the identity line. In C and D, contours indicate density of points plotted in the region of the graph. Red indicates the highest density of points, while blue indicates lower densities.

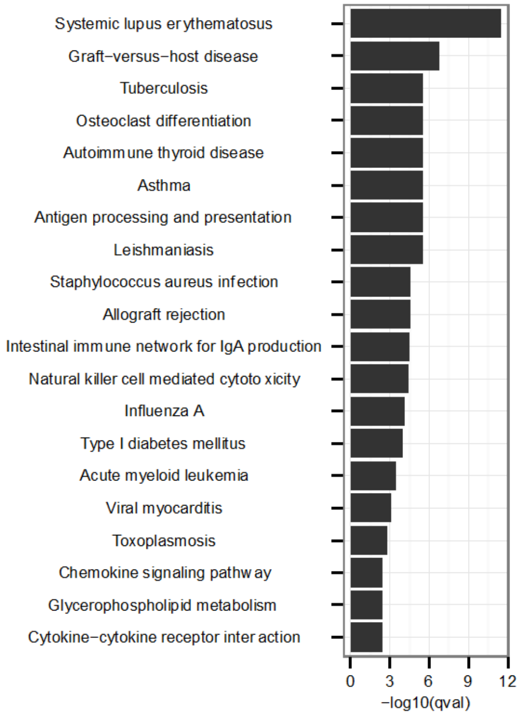

Fig. S7

KEGG pathways enriched in genes differentially accessible between GM12878 and HL-60 cells. Barplot of hypergeometric test q-value for enrichment of pathways (labeled on y-axis) for genes linked to significantly differentially accessible sites between GM12878 and HL-60.

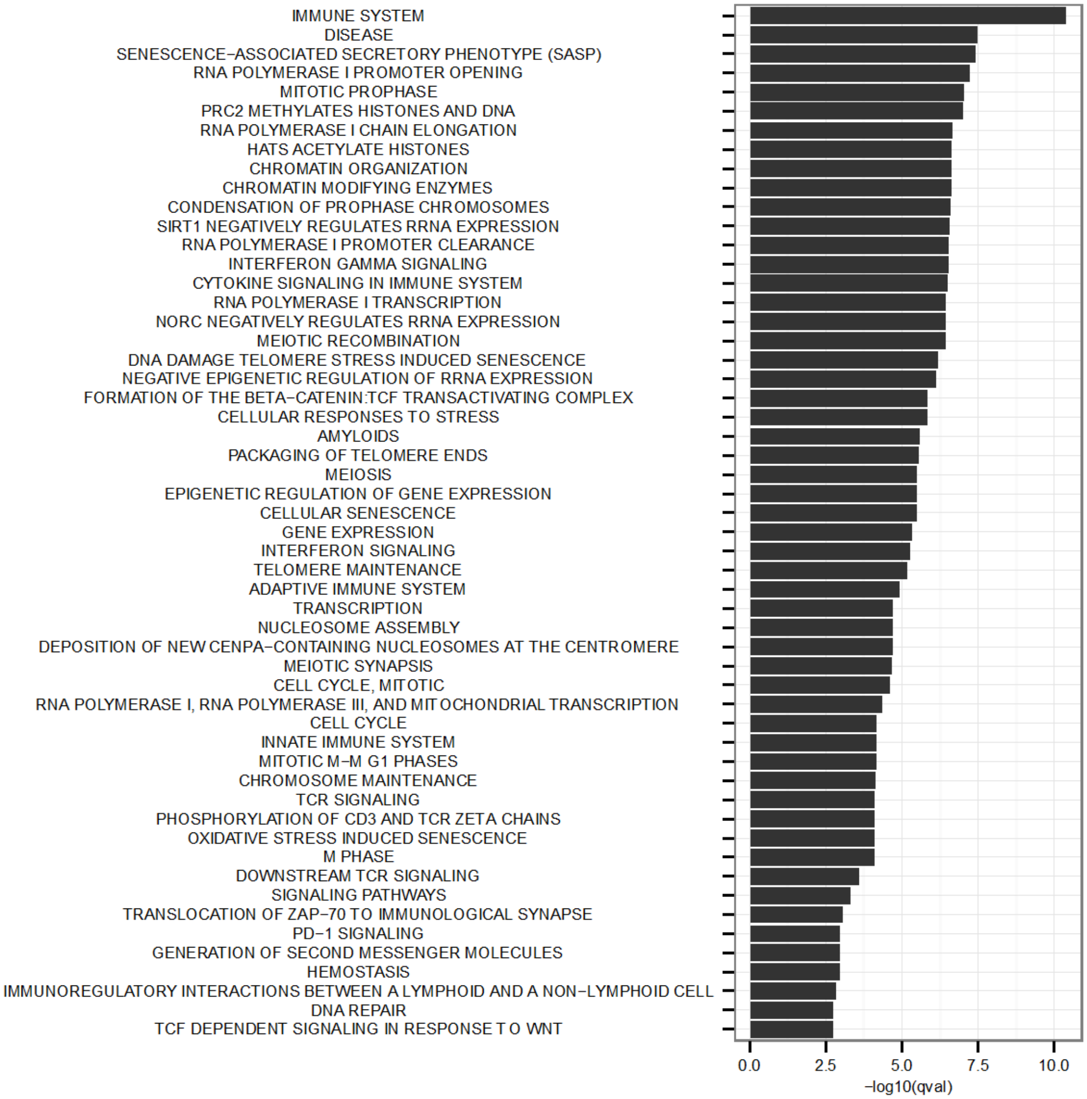

Fig. S8

Reactome pathways enriched in genes differentially accessible between GM12878 and HL-60 cells. Barplot of hypergeometric test q-value for enrichment of pathways (labeled on y-axis) for genes linked to significantly differentially accessible sites between GM12878 and HL-60.

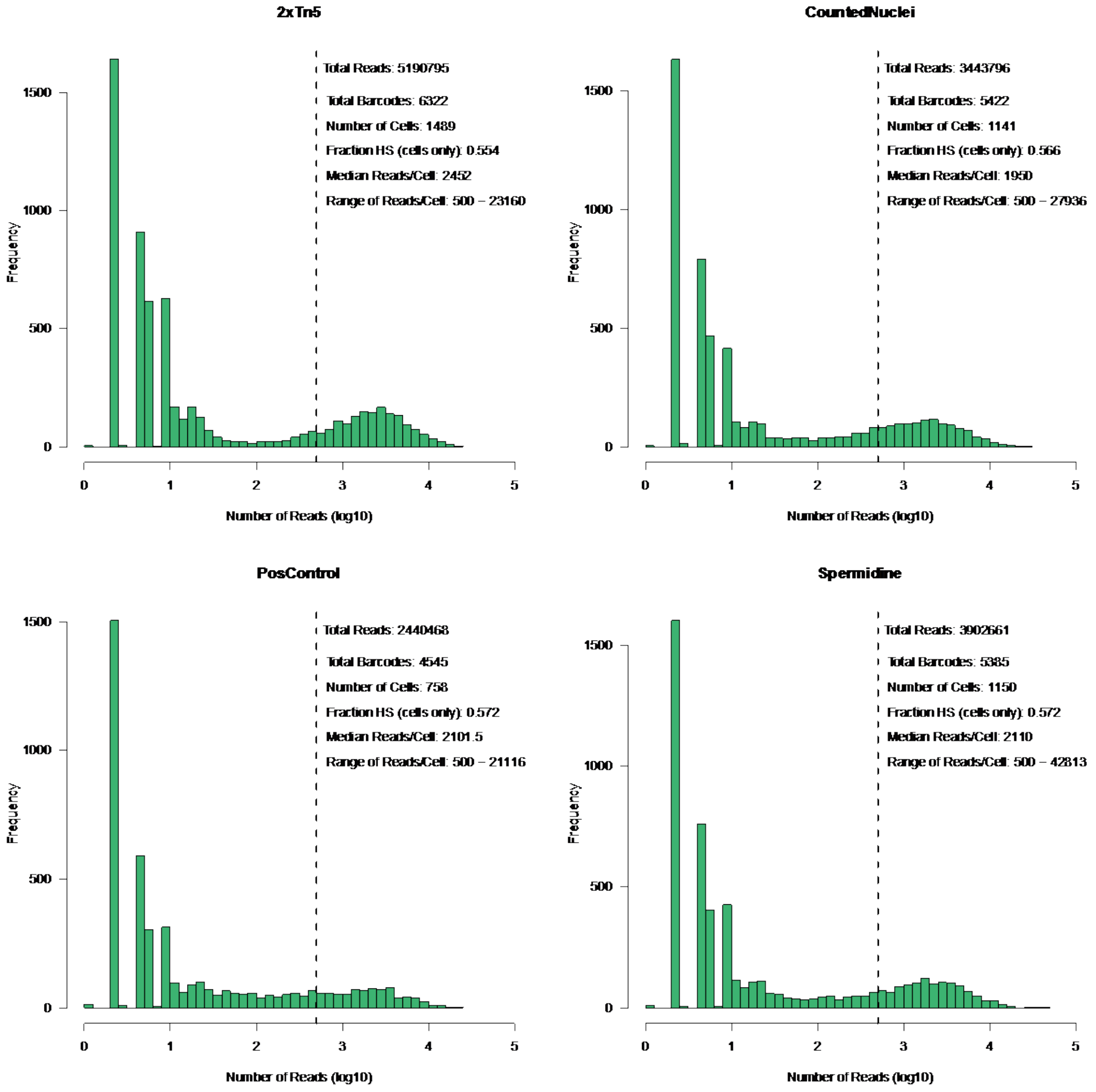

Fig. S9

Number of mappable reads assigned to each possible barcode combination on Experimental Condition Tests Day 1. As in Fig. S2, each panel represents the cells isolated in one experimental condition. The dashed line marks 500 reads in each graph. “2xTn5”=Doubled enzyme concentration; “CountedNuclei”=Counting nuclei instead of cells to determine starting material; ”PosControl”=Standard conditions described in the methods; “Spermidine”=Supplemented 40mM EDTA stop solution with 1mM Spermidine.

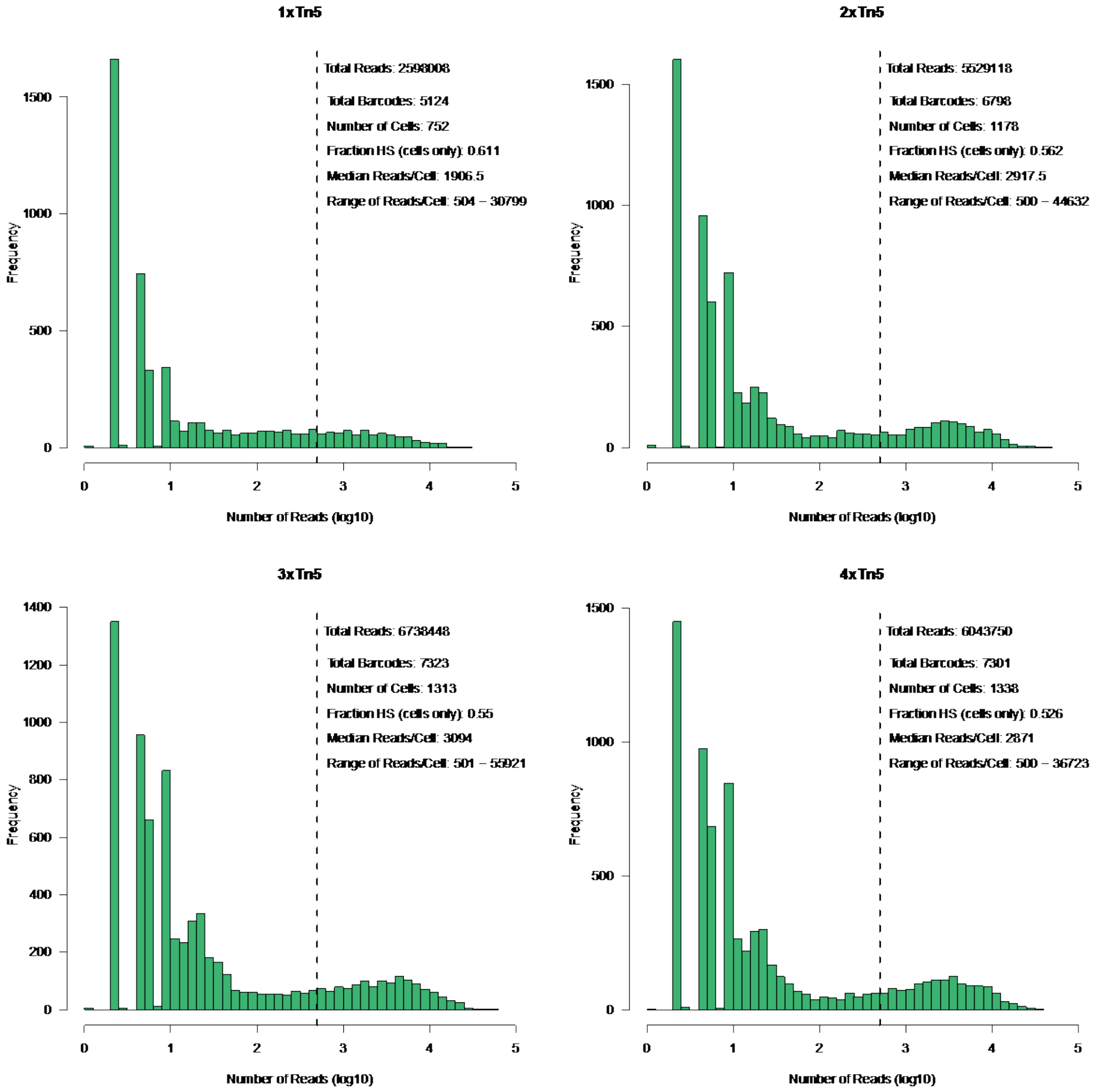

Fig. S10

Number of mappable reads assigned to each possible barcode combination on Experimental Condition Tests Day 2. As in Fig. S2, each panel represents the cells isolated in one experimental condition. The dashed line marks 500 reads in each graph. Names indicate varying enzyme concentrations.

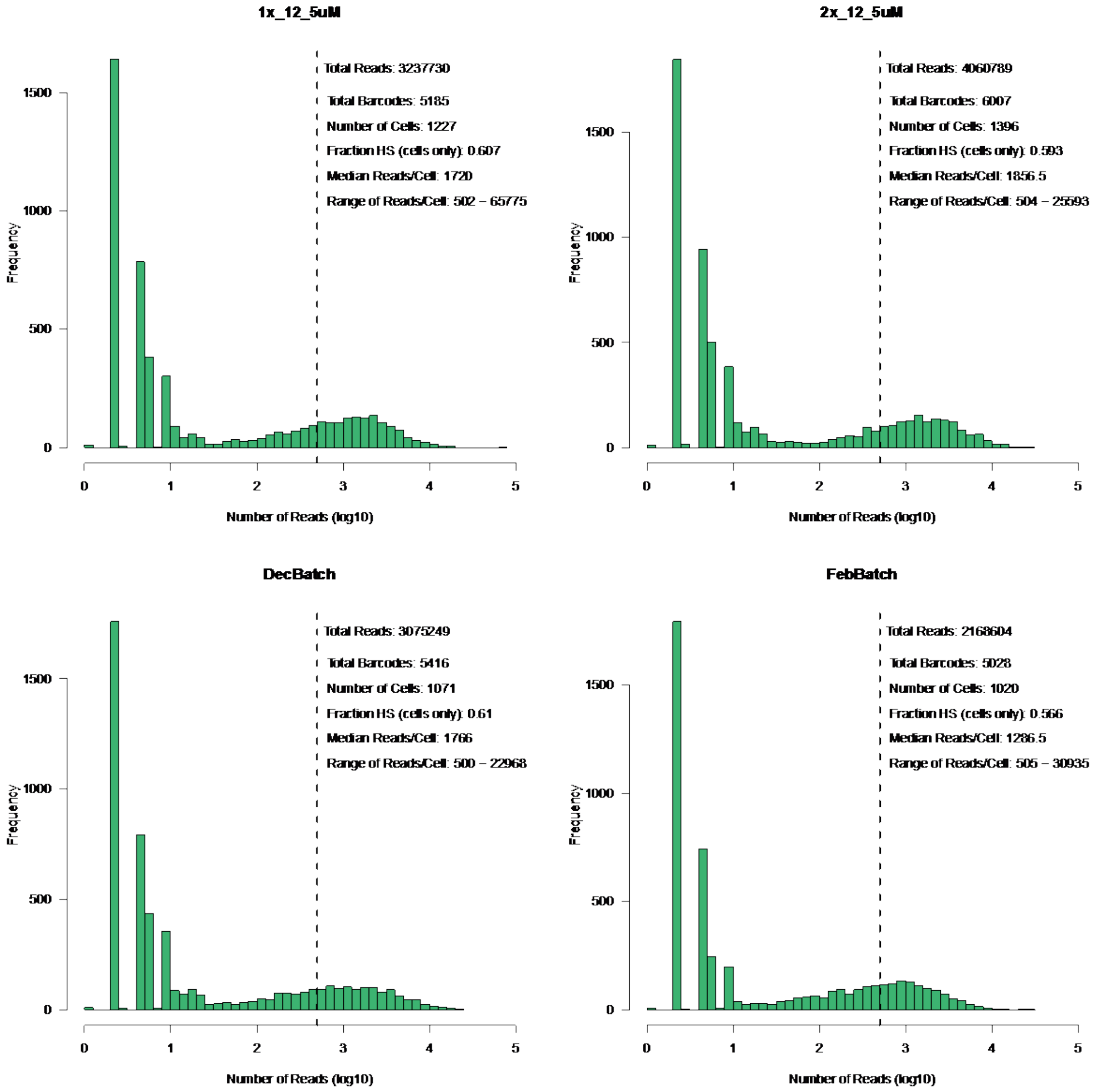

Fig. S11

Number of mappable reads assigned to each possible barcode combination on Experimental Condition Tests Day 3. As in **Fig. S2**, each panel represents the cells isolated in one experimental condition. The dashed line marks 500 reads in each graph. “1x_12_5uM” and “2x_12_5uM” used 12.5 or 25µM enzyme, respectively. “DecBatch” and “FebBatch” tested two batches of enzyme in parallel.

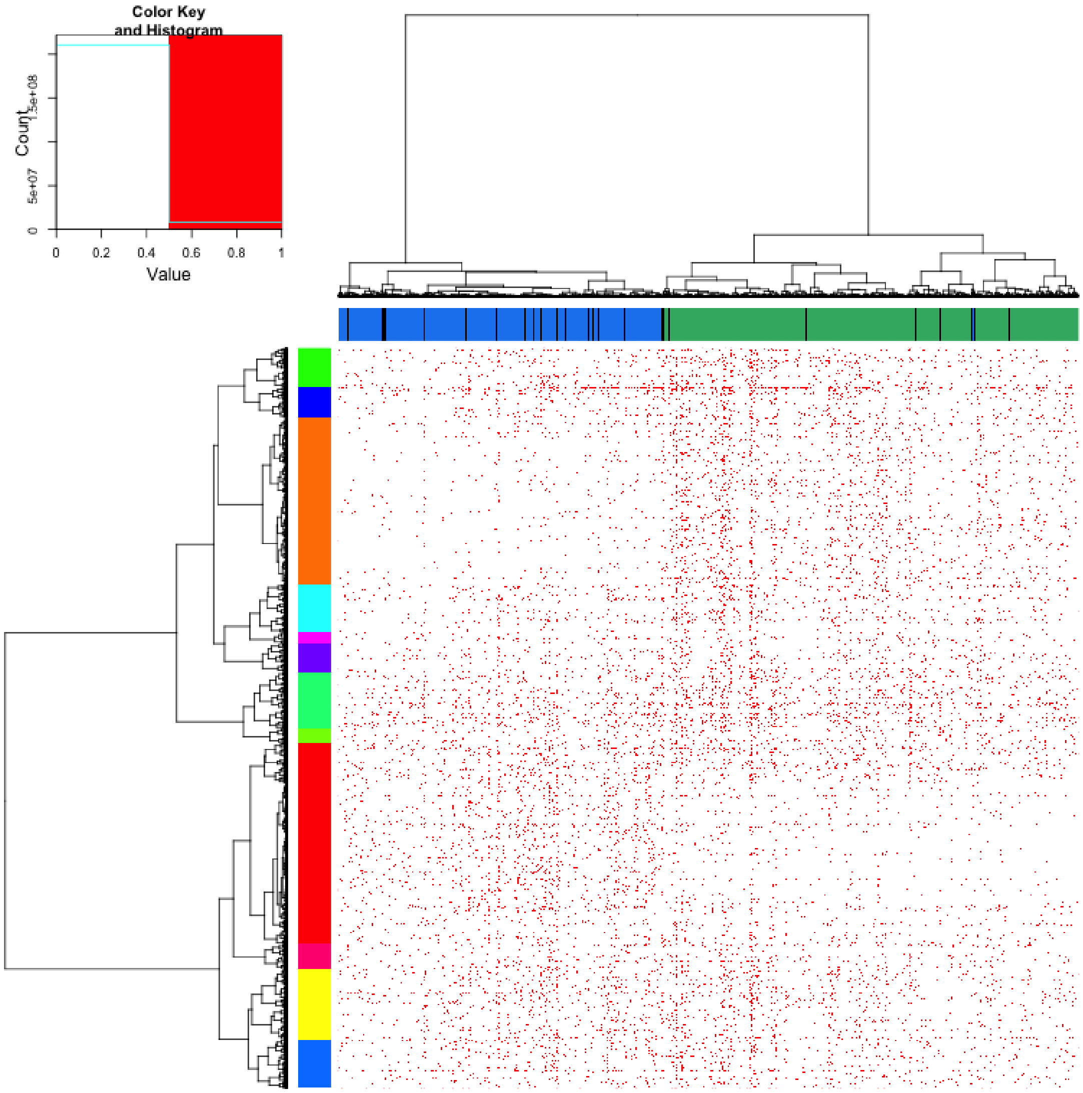

Fig. S12

Heatmap of binary site usage for GM12878/HL-60 mixture experiments. This heatmap is identical to **Fig. 3C**, except that binary site usage is plotted rather than lower dimensional representation generated from latent semantic indexing. Rows are hypersensitive sites and columns are cells.

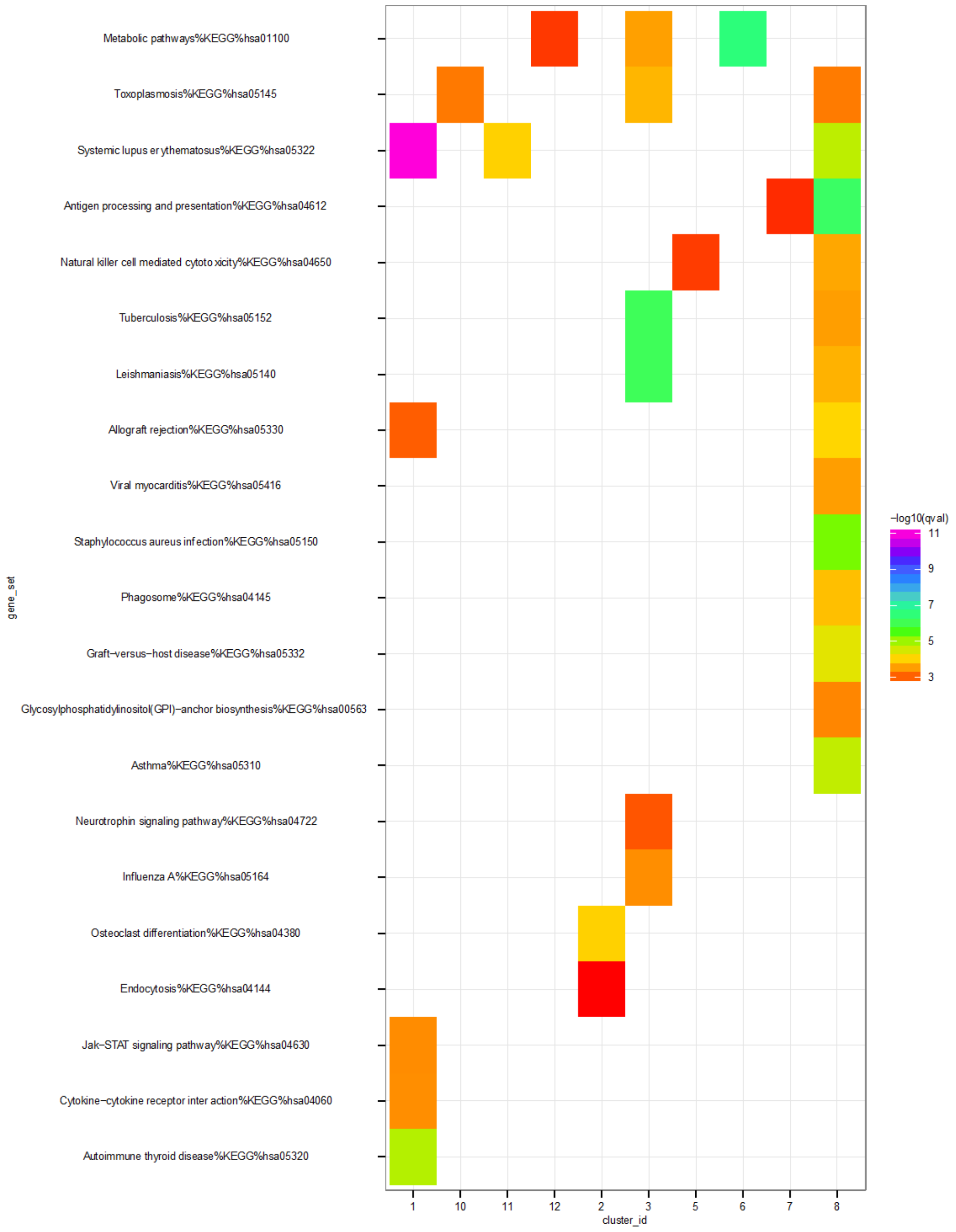

Fig. S13

KEGG pathways enriched in modules of coordinately regulated chromatin accessibility between GM12878 and HL-60 cells. Each row represents a pathway (labeled on y-axis) and each column represents a module from the heatmap presented in **Fig. 3C**. The color of each pathway/module combination indicates the q-value for enrichment of genes with that annotation term in that module.

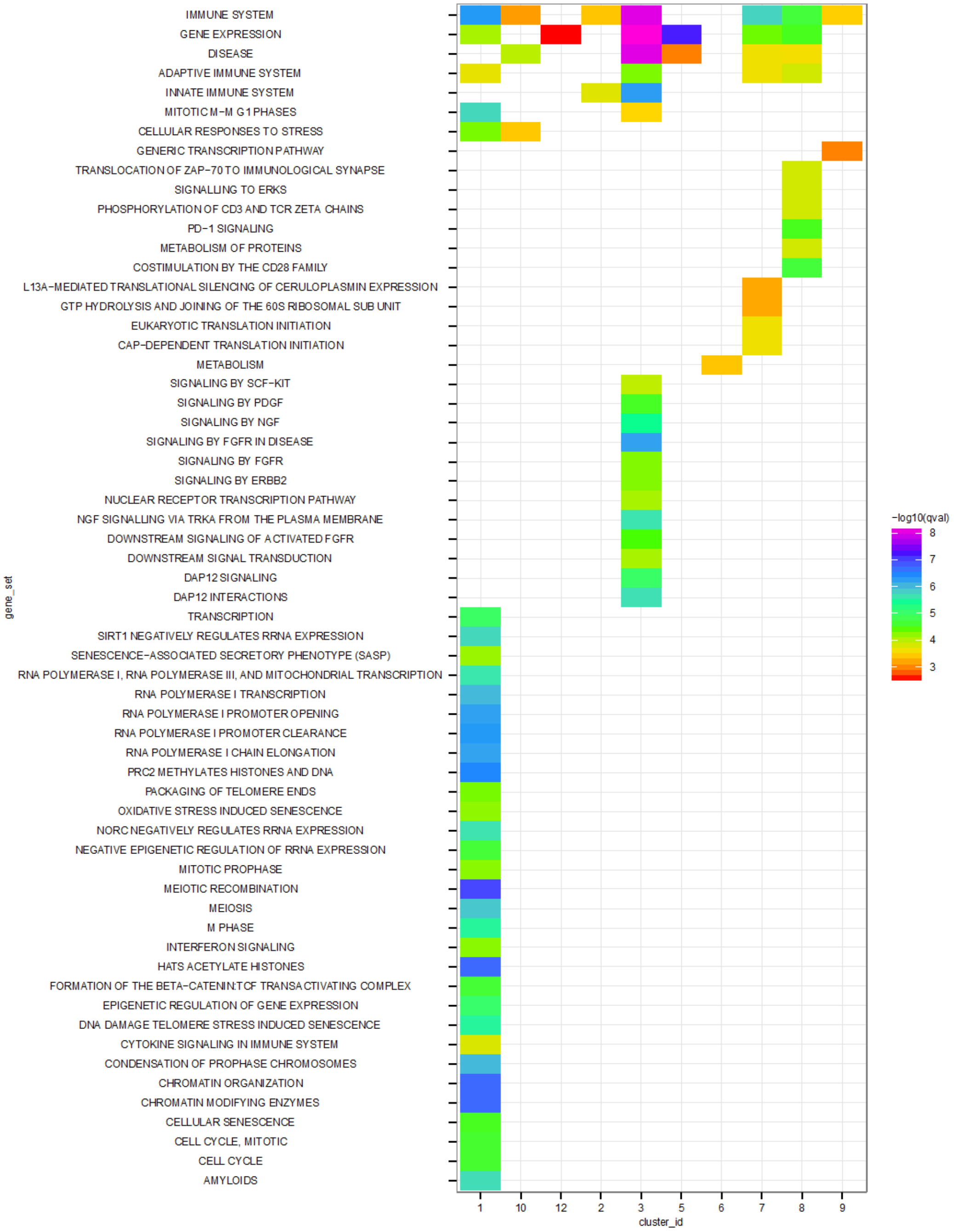

Fig. S14

Reactome pathways enriched in modules of coordinately regulated chromatin accessibility between GM12878 and HL-60 cells. Each row represents a pathway (labeled on y-axis) and each column represents a module from the heatmap presented in **Fig. 3C**. The color of each pathway/module combination indicates the q-value for enrichment of genes with that annotation term in that module.

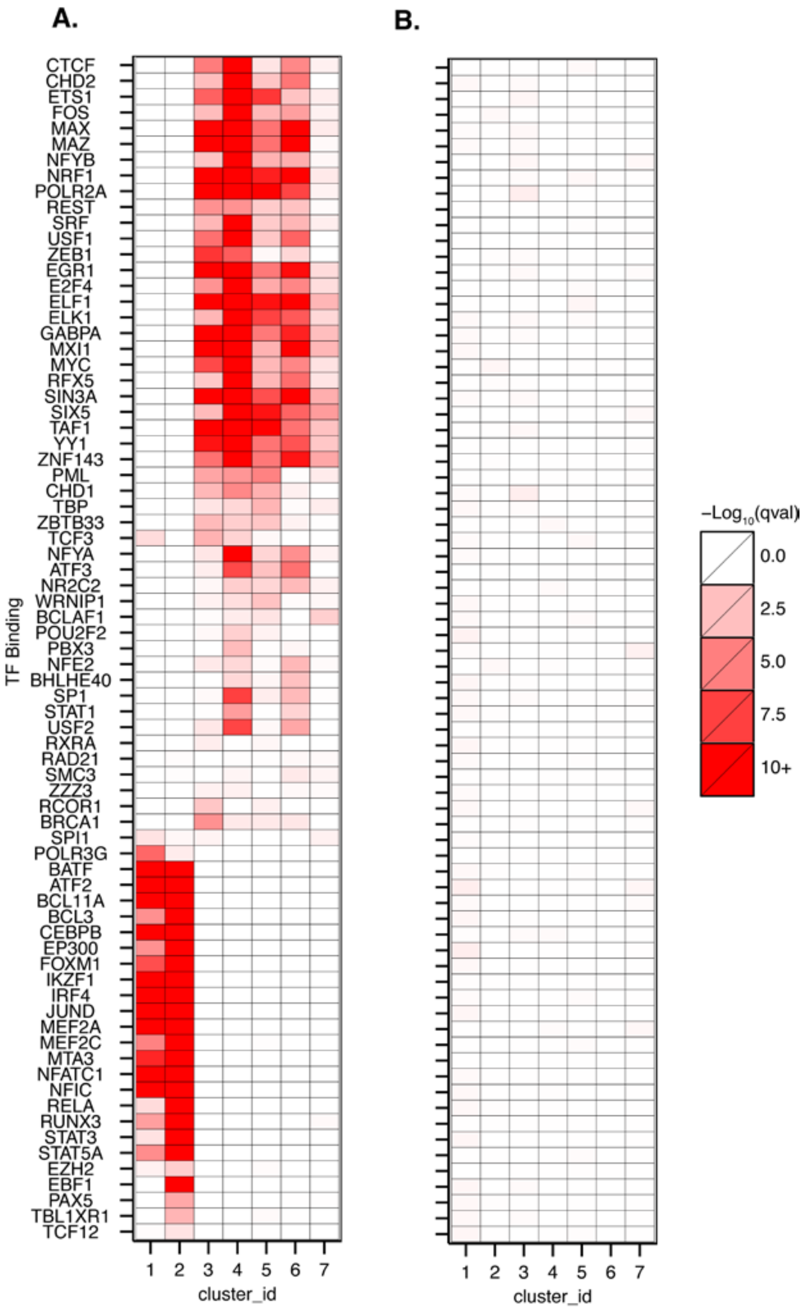

Fig. S15

Enrichments for transcription factor binding within specific modules of coordinately accessible regulatory elements in cells from all GM12878/HL-60 mixture experiments (4,118 GM128781 cells). (**A**) Enrichments for TF binding within specific modules. Overall, 71 of 75 TF had significant enrichments identified with at least one module at an FDR of 0.1. (**B**) The matrix of site usage was permuted while maintaining row and column sums. This permuted matrix was then used to cluster cells and sites and enrichments for binding were assessed. No significant enrichments were identified in the permuted data at an FDR of 0.1.

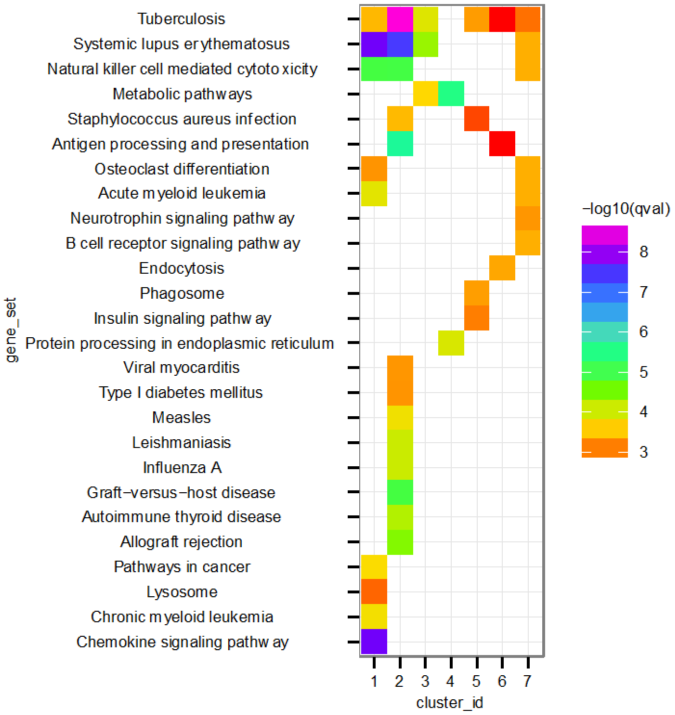

Fig. S16

KEGG pathways enriched in modules coordinately regulated between GM12878 subtypes. Each row represents a pathway (labeled on y-axis) and each column represents a module from the heatmap presented in **Fig. 4A**. The color of each pathway/module combination indicates the q-value for enrichment of genes with that annotation term in that module.

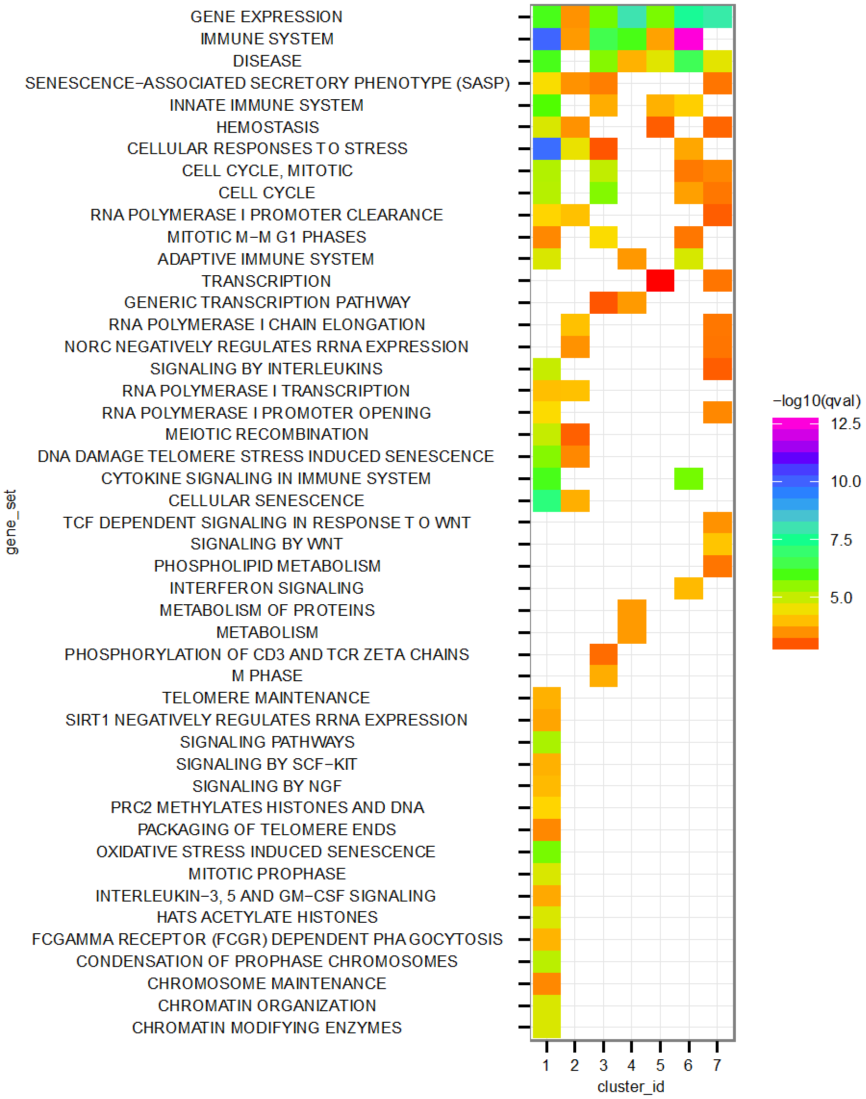

Fig. S17

Reactome pathways enriched in modules coordinately regulated between GM12878 subtypes. Each row represents a pathway (labeled on y-axis) and each column represents a module from the heatmap presented in **Fig. 4A**. The color of each pathway/module combination indicates the q-value for enrichment of genes with that annotation term in that module.

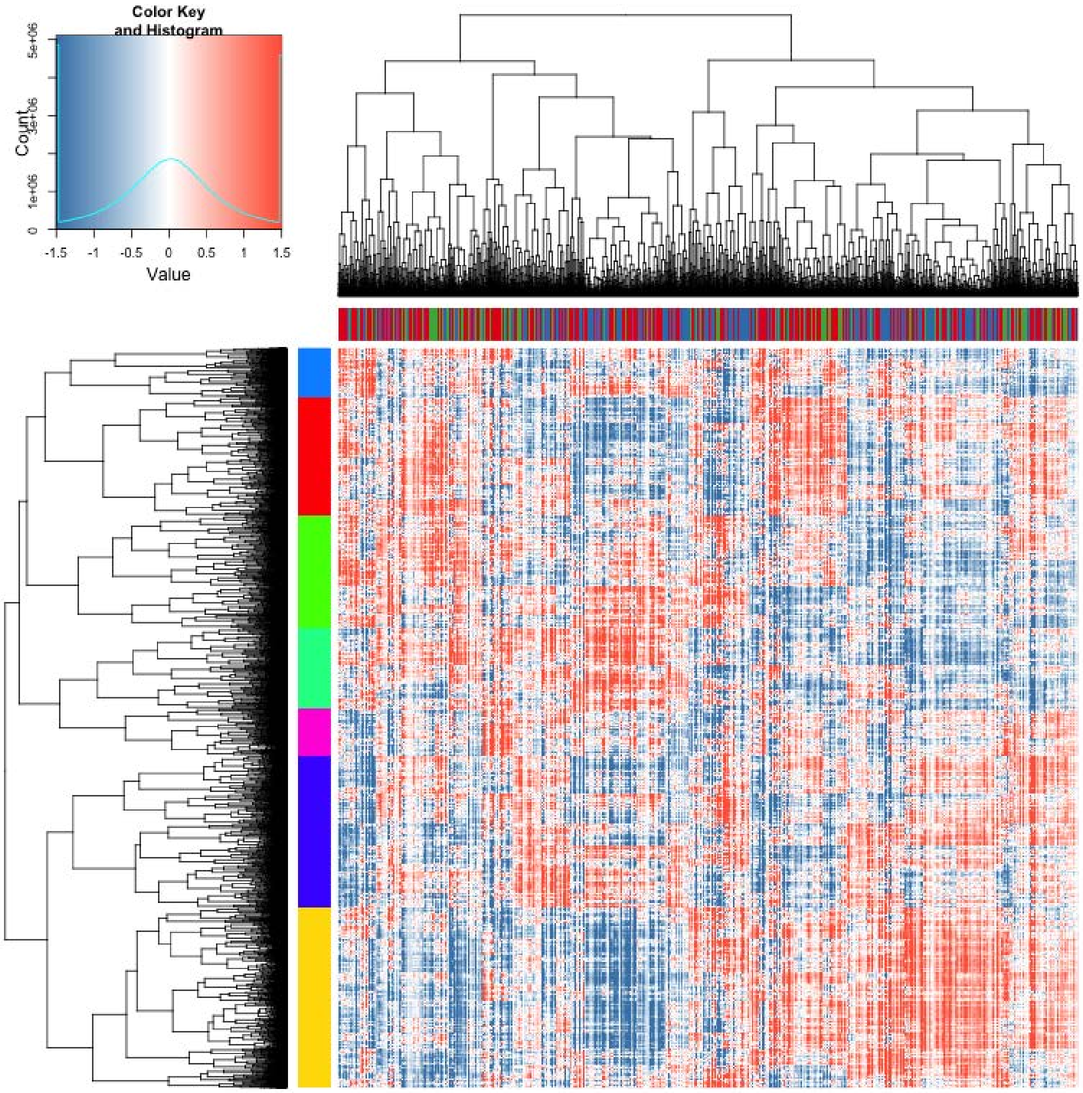

Fig. S18

Clustering of GM12878 subtypes not correlated with experiment date. Same heatmap as presented in **Fig. 4A**, except that color bar across columns is color-coded by the date on which the experiment was conducted (cells from 4 dates were included). Samples do not cluster by date.

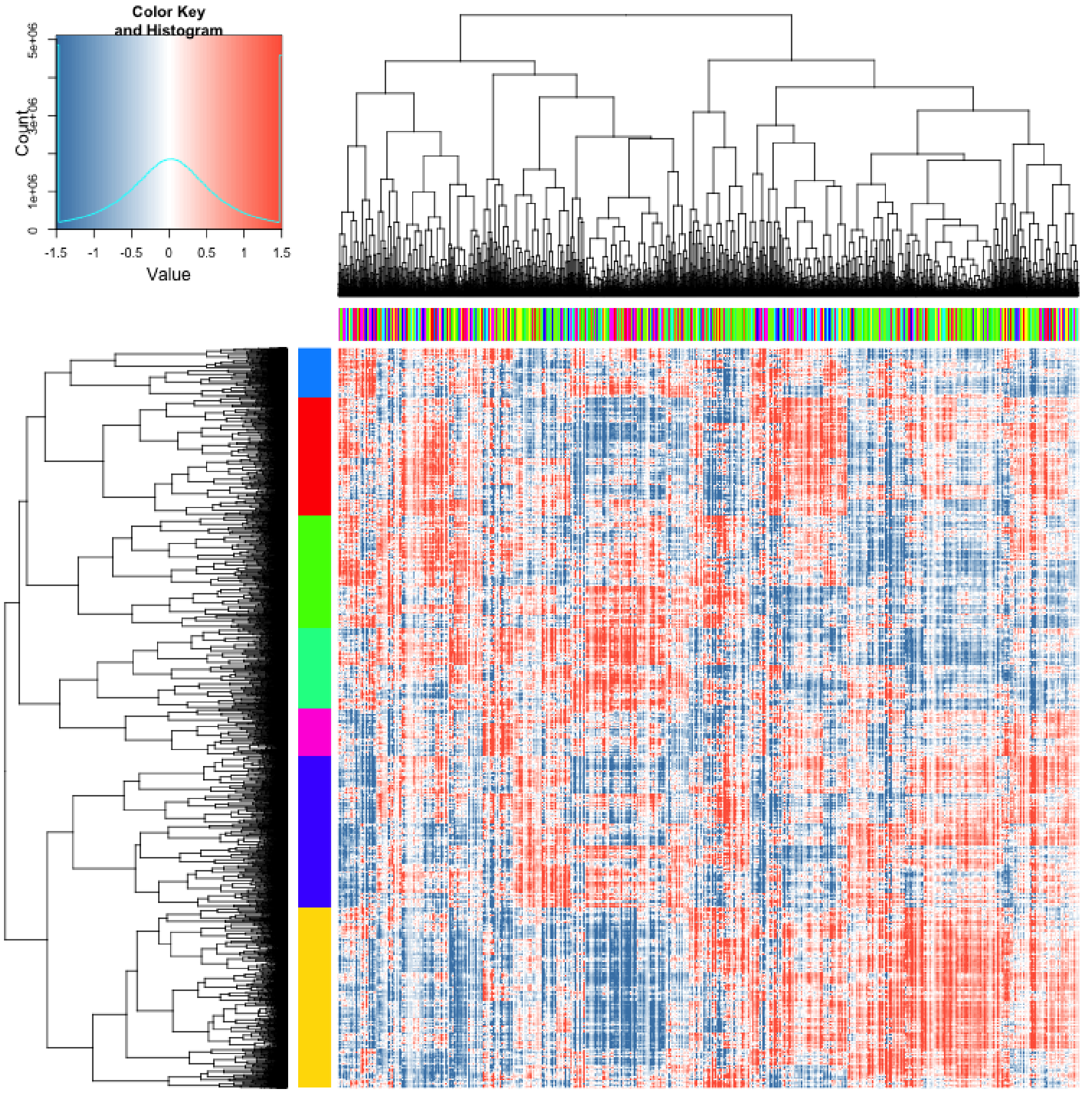

Fig. S19

Clustering of GM12878 subtypes not correlated with experimental condition. Same heatmap as presented in **Fig. 4A**, except that color bar across columns is color-coded by experimental condition. We tested several batches of enzyme, several enzyme concentrations and several enzymatic reaction stopping conditions during the four different experiments. Samples do not cluster by enzymatic condition.

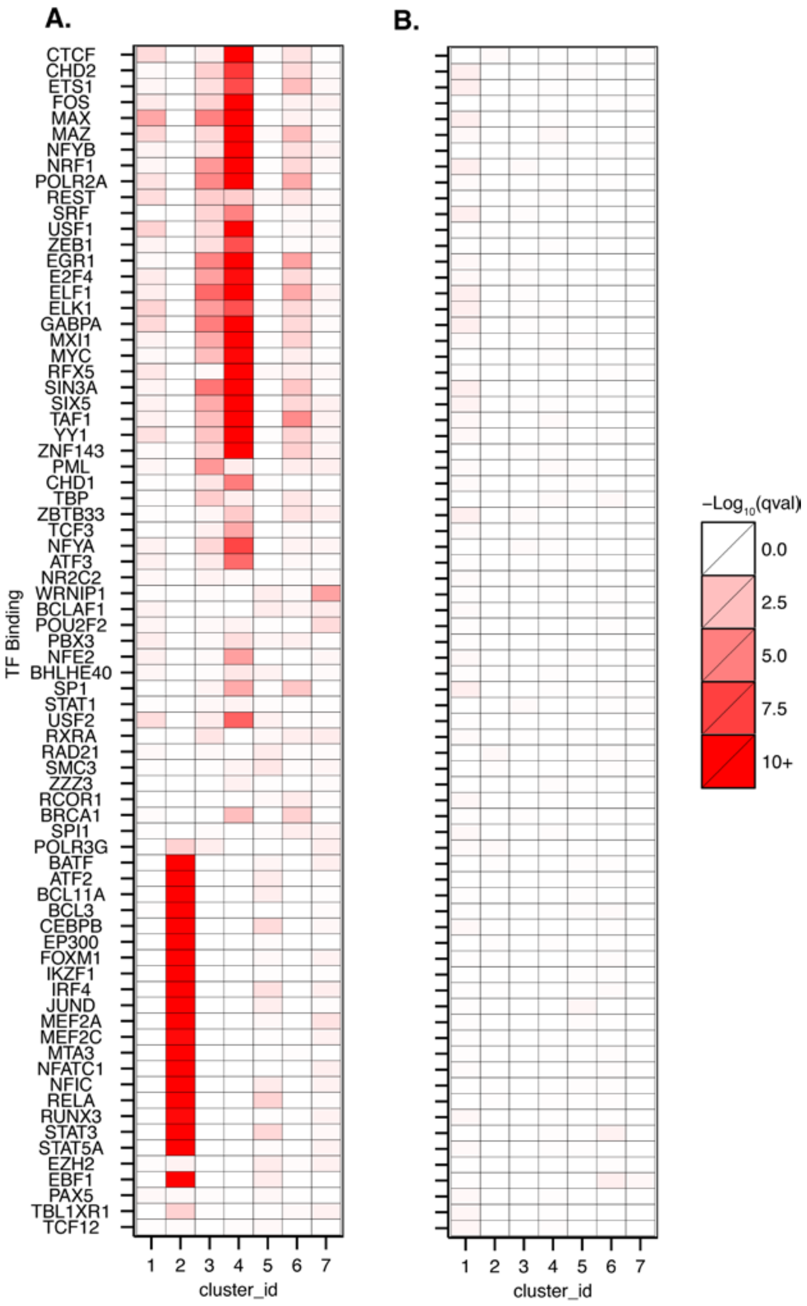

Fig. S20

Enrichments for transcription factor binding within specific modules of coordinately accessible regulatory elements in cells from the first experimental condition test day (1,879 GM12878 cells). (**A**) Enrichments for TF binding within specific modules. Overall, 63 of 75 TF had significant enrichments identified with at least one module at an FDR of 0.1. (**B**) The matrix of site usage was permuted while maintaining row and column sums. This permuted matrix was then used to cluster cells and sites and enrichments for binding were assessed. No significant enrichments were identified in the permuted data at an FDR of 0.1.

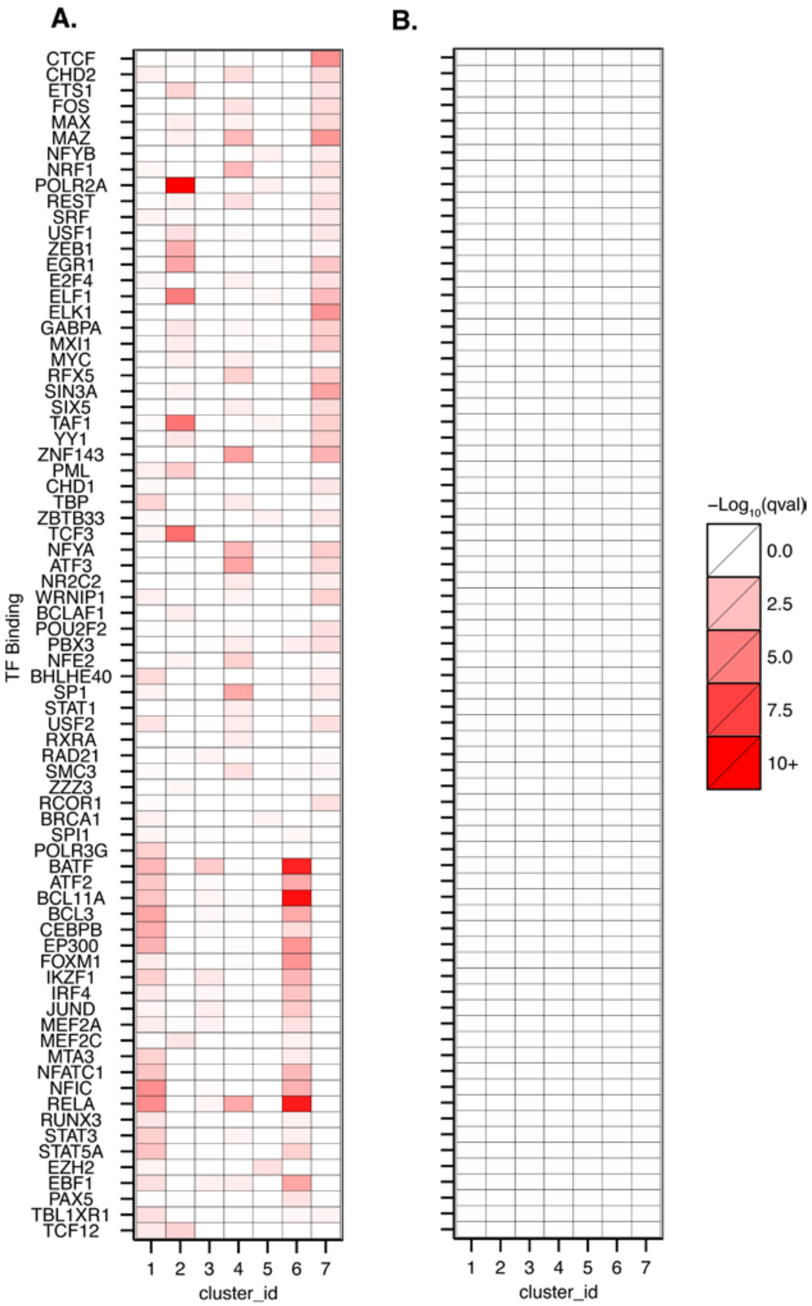

Fig. S21

Enrichments for transcription factor binding within specific modules of coordinately accessible regulatory elements in cells from the second and third experimental condition test days (2,022 GM12878 cells). (**A**) Enrichments for TF binding within specific modules. Overall, 62 of 75 TF had significant enrichments identified with at least one module at an FDR of 0.1. (**B**) The matrix of site usage was permuted while maintaining row and column sums. This permuted matrix was then used to cluster cells and sites and enrichments for binding were assessed. No significant enrichments were identified in the permuted data at an FDR of 0.1.

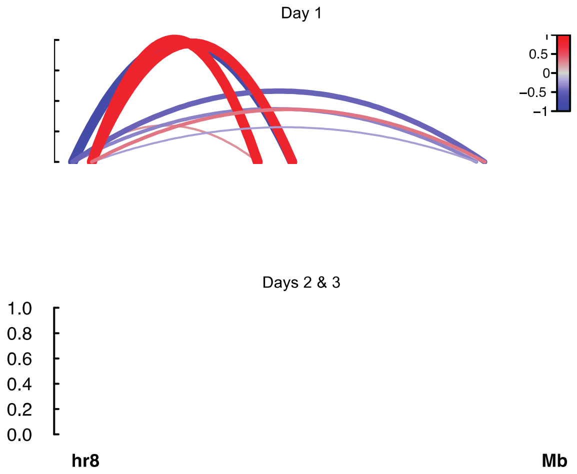

Fig. S22

Co-accessibility patterns between putative enhancers and TSS elements of LYN as measured with independent biological replicates. Experimental condition test days 2 and 3 were pooled to construct a replicate with similar numbers of cells as experimental condition test day 1. Latent semantic indexing was performed on each subset independently, and co-accessibility was calculated by Pearson correlation between the resulting regularized values.

Table S1.

Genomic coverage of DNase I hypersensitivity maps. This table reports the genomic coverage of hypersensitivity sites used in our analyses after combining replicates and merging between cell types. “Cell Type of DHS Map” describes which comparison the map was used for in the main text. “Patski” and “GM12878” were used in the original inter-species comparison, while the other two maps were used for the human cell type comparisons. “Reference genome” is listed as “mm9” for mouse cells or “hg19” for human cells. “Genome Size (bp)” was the non-NA size estimate from http://genomewiki.ucsc.edu/index.php/Hg19_Genome_size_statistics on 12/2/14). “No. of Elements” lists the number of individual number of elements included in each map. “Genome Coverage (bp)” lists the total bases covered by the reference maps. “Genome Coverage (fraction)” is the “Genome Coverage (bp)” divided by “Genome Size (bp)”.

Cell Type of DHS Map	Reference Genome	Genome Size (bp)*	No. of Elements	Genome Coverage (bp)	Genome Coverage (fraction)
Patski	mm9	2,620,345,972	159,424	66,787,300	0.025
GM12878	hg19	2,897,310,462	124,844	43,196,806	0.015
GM12878 vs. HEK293T	hg19	2,897,310,462	181,379	86,020,710	0.030
GM12878 vs. HL-60	hg19	2,897,310,462	230,632	97,103,964	0.034

*Genome sizes from http://genomewiki.ucsc.edu/index.php/Hg19_Genome_size_statistics on 12/2/14

Table S2.

Tests of differential accessibility from single cell ATAC-seq data comparing GM12878 and HL-60. This table contains all 52,479 sites tested for differentially accessibility. It is too large to include here and so is supplied as a separate file. For this analysis, sites overlapping copy number variants identified by the ENCODE Consortium for either cell type were filtered out. In addition, chromosomes 18, X and Y were removed, because HL-60 is known to be trisomic for all three. The columns of the file are as follows:

Column 1 – Chromosome of element
Column 2 – Genomic start coordinate of element
Column 3 – Genomic end coordinate of element
Column 4 – P-value from our binomial test of differential accessibility between the two cell types
Column 5 – Significance indicator. “0” = not significant after multiple test-correction. “1” = significant after multiple test-correction.

References

1. A. B. Stergachis, S. Neph, A. Reynolds, R. Humbert, B. Miller, S. L. Paige, B. Vernot, J. B. Cheng, R. E. Thurman, R. Sandstrom, E. Haugen, S. Heimfeld, C. E. Murry, J. M. Akey, J. A. Stamatoyannopoulos, Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013). Medline doi:10.1016/j.cell.2013.07.020

2. R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, J. A. Stamatoyannopoulos, The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012). Medline doi:10.1038/nature11232

3. A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, G. E. Crawford, High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008). Medline doi:10.1016/j.cell.2007.12.014

4. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013). Medline doi:10.1038/nmeth.2688

5. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011). Medline doi:10.1038/nature09807

6. A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke, S. R. Quake, Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014). Medline doi:10.1038/nmeth.2694

7. D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014). Medline doi:10.1126/science.1247651

8. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014). Medline doi:10.1126/science.1245316

9. A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, A. Regev, Single-cell transcriptomics reveals

bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013). Medline doi:10.1038/nature12172

10. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014). Medline doi:10.1038/nbt.2859

11. S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, G. Kelsey, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014). Medline doi:10.1038/nmeth.3035

12. T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, P. Fraser, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013). Medline doi:10.1038/nature12593

13. J. Gole, A. Gore, A. Richards, Y. J. Chiu, H. L. Fung, D. Bushman, H. I. Chiang, J. Chun, Y. H. Lo, K. Zhang, Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013). Medline doi:10.1038/nbt.2720

14. H. C. Fan, G. K. Fu, S. P. A. Fodor, Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015). Medline doi:10.1126/science.1258367

15. A.-E. Saliba, A. J. Westermann, S. A. Gorski, J. Vogel, Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014). Medline doi:10.1093/nar/gku555

16. X. Pan, Single cell analysis: From technology to biology and medicine. *Single Cell Biol.* **3**, 106 (2014). Medline doi:10.4172/2168-9431.1000106

17. A. Adey, J. O. Kitzman, J. N. Burton, R. Daza, A. Kumar, L. Christiansen, M. Ronaghi, S. Amini, K. L. Gunderson, F. J. Steemers, J. Shendure, In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014). Medline

18. S. Amini, D. Pushkarev, L. Christiansen, E. Kostem, T. Royce, C. Turk, N. Pignatelli, A. Adey, J. O. Kitzman, K. Vijayan, M. Ronaghi, J. Shendure, K. L. Gunderson, F. J. Steemers, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014). Medline doi:10.1038/ng.3119

19. Materials and methods are available as supplementary materials on *Science* Online.

20. A. Adey, H. G. Morrison, Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang, J. Shendure, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010). Medline doi:10.1186/gb-2010-11-12-r119

21. F. Yang, T. Babak, J. Shendure, C. M. Disteche, Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* **20**, 614–622 (2010). Medline doi:10.1101/gr.103200.109

22. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). Medline

23. A. Regev, The Human Cell Atlas; http://www.genome.gov/Multimedia/Slides/GSPFuture2014/10_Regev.pdf.

24. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). Medline doi:10.1093/bioinformatics/btu170

25. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

26. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013). Medline doi:10.1186/gb-2013-14-4-r36

27. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013). Medline doi:10.1038/nbt.2450

28. R. von Mises, Über Aufteilungs und Besetzungs-Wahrscheinlichkeiten. *Rev. la Fac. des Sci. l'Université d'Istanbul, N.S.* **4**, 145–163 (1939).

29. S. John, P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, J. A. Stamatoyannopoulos, Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011). Medline doi:10.1038/ng.759

30. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

31. A. P. Boyle, J. Guinney, G. E. Crawford, T. S. Furey, F-Seq: A feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008). Medline doi:10.1093/bioinformatics/btn480

32. T. W. Yee, C. J. Wild, Vector generalized additive models. *J. R. Stat. Soc., B* **58**, 481–493 (1996).

33. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995). http://www.jstor.org/stable/2346101

34. M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, W. S. Noble, Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013). Medline doi:10.1093/nar/gks1284

35. L. Väremo, J. Nielsen, I. Nookaew, Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41**, 4378–4391 (2013). Medline doi:10.1093/nar/gkt111

36. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000). Medline doi:10.1093/nar/28.1.27

37. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* **42** (D1), D199–D205 (2014). Medline doi:10.1093/nar/gkt1076

38. D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D'Eustachio, The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42** (D1), D472–D477 (2014). Medline doi:10.1093/nar/gkt1102

39. M. Milacic, R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, L. Stein, Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers* **4**, 1180–1211 (2012). Medline doi:10.3390/cancers4041180