Supplementary Information for

Impact of regulatory variation across human iPSCs and differentiated cells

Nicholas E. Banovich[1,‡*], Yang I. Li[2*], Anil Raj[2*], Michelle C. Ward[1,3], Peyton Greenside[4], Diego Calderon[4], Po Yuan Tung[1,3], Jonathan E. Burnett[1], Marsha Myrthil[1], Samantha M. Thomas[1], Courtney K. Burrows[1], Irene Gallego Romero[1,‡], Bryan J. Pavlovic[1], Anshul Kundaje[2], Jonathan K. Pritchard[2,5,6,†], Yoav Gilad[1,3,†]

[1]Department of Human Genetics, University of Chicago, Chicago, IL, USA.

[2]Department of Genetics, Stanford University, Stanford, CA, USA.

[3]Department of Medicine, University of Chicago, Chicago, IL, USA.

[4]Department of Biomedical Informatics, Stanford University, Stanford, CA, USA.

[5]Department of Biology, Stanford University, Stanford, CA, USA.

[6]Howard Hughes Medical Institute, Stanford, University, Stanford, CA, USA.

[‡]Present address. Translational Genomics Research Institute, Phoenix, AZ, USA (N.E.B.); School of Biological Sciences and Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (I.G.R.).

[*]These authors contributed equally to this work.

[†] **Corresponding author. Email: pritch@stanford.edu (J.K.P.); gilad@uchicago.edu (Y.G.).**

Contents

List of Tables

List of Figures

Supplementary Methods

1 iPSC generation

We reprogrammed LCLs into iPSCs using an episomal reprogramming approach described previously [1, 2]. Briefly, we transfected 1 million LCLs (Amaxa^{TM} Nucleofector^{TM} Technology; Lonza) with 1ug of oriP/EBNA1 PCXLE based episomal plasmids that contain the genes OCT3/4, SOX2, KLF4, L-MYC, LIN28, and an shRNA against p53 [1, 2]. Cells were cultured in suspension for seven days after transfection in hESC media (DMEM/F12 supplemented with 20% KOSR, 0.1mM NEAA, 2mM GlutaMAX, 1% Pen/Strep, 0.1% 2-Mercaptoethanol, 25ng/ul of bFGF and .5mM NaB). On the 8th day we plated a range of 8,000 - 32,000 transfected cells per well in a 6-well plate coated with gelatin and seeded with irradiated CF1 mouse embryonic fibroblasts (MEF). Four days after the initial plating NaB was removed from the hESC media. Within 21 days colonies were visible and manually passaged onto a freshly prepared gelatin plate seeded with MEF. Manual passaging continued weekly for ten weeks. After ten passages of growth cells were expanded and at least ten stocks of cells were cryopreserved. Colonies were then transitioned to feeder-free conditions and cultured for at least an additional three passages before collecting cell pellets for analysis. Feeder-free cultures were grown using 0.01mg/cm2 (1:100) hESC-grade Matrigel and Essential 8 (E8) media. Feeder-free passaging is enzymatic rather than manual and was performed using DPBS supplemented with 0.5mM EDTA.

2 iPSC characterization

All iPSC lines were characterized for pluripotency and stability using three methods. First, we confirmed the ability of lines to differentiate to all three germ layers using the embroyid body (EB) assay. Lines were manually dissociated from their culture dish in large pieces. This material was then cultured in a suspension plate using the hESC media described above without bFGF for one week, while dense spherical EBs form. EBs are then plated into 12 well plates with gelatin and cultured in EB medium (DMEM supplemented with 10% FBS, 0.1mM NEAA, 2mM GlutaMAX) for one week. EBs in each well were then immunostained for cell types from all three germ layers Supplementary Figure 10. Next, all lines were karyotyped to search for large genomic rearrangements Supplementary Figure 10. Lines were karyotypes by the WiCell Research Institute (Madison, WI). Only one line, 19128, showed large genomic rearrangements that were not known rearrangements segregating in the population. The rearrangement observed in this line is a hallmark rearrangement of follicular lymphoma and thus was likely present in LCLs rather than a result of the reprogramming process. Finally, a classifier, PluriTest [3] was applied to gene expression data (Illumina HumanHT-12 array) to assay pluripotency bioinformatically. The classifier compares gene expression levels from uncharacterized lines to a "gold standard" panel of embryonic stem cells and iPSCs. Two metrics are obtained from this method, a pluripotency score and a novelty score. The pluripotency score represents goodness of fit of canonical pluripotency genes in the sample. The novelty score represents the deviance of non-pluripotency genes in the sample. All of the lines here pass the suggested empirical threshold (Supplementary Figure 10).

3 iPSC-derived cardiomyocyte differentiation

Differentiation from iPSCs to cardiomyocytes was performed using slight modifications of existing protocols [4, 5]. Specifically, iPSCs cultured under feeder-free conditions were seeded onto a 10 cm dish three to five days prior to differen- tiation. When cells were 70-100% confluent E8 media was replaced with 'heart media' (RPMI (14-040-CM, Thermo Fisher Scientific) supplemented with B-27 Supplement, minus insulin (A1895601, Thermo Fisher Scientific), 2mM GlutaMAX (35050-061, Thermo Fisher Scientific), and 100mg/mL Pennicllin/Streptomycin (30002Cl, Corning)) with the addition of 1:100 Matrigel (35427, Corning) and 12uM of the GSK-3 inhibitor CHIR99021 trihydrochloride (4953, Tocris,) which activates WNT signaling (day 0) [4]. 'Heart media' was replaced 24 hours later (day 1). After an additional 48 hours media was replaced with new 'heart media' containing 2uM of the WNT inhibitor Wnt-C59 (5148, Tocris) [4] (day 3). Cells were cultured in the media with Wnt-C59 for 48 hours. Regular 'heart media' changes were performed from day 5 to day 14. Clusters of spontaneously beating cells were typically visible between 7 and 12 days. On day 14 heart media was replaced with 'lactate media' (RPMI without glucose (11879, Gibco), 75 mg/ml human recombinant albumin (A0237, Sigma), 213 ug/ml L-ascorbic acid 2-phosphate (sc228390, Santa Cruz Biotechnology), 5mM Sodium L-lactate (L7022, Sigma), and 100mg/mL Pennicillin/Streptomycin) 24930130. The 'lactate media' purifies cardiomyocytes using metabolic selection whereby the majority of non-cardiomyocyte cells cannot use lactate as their primary source of energy, leaving a culture significantly enriched for cardiomyocytes [6]. Every other day 'lactate media' changes were performed until day 20. By day 20 the cells had generally formed into large three-dimensional sheets of beating cells. To make a more uniform sheet of cells we dissociated the cultures using 0.05% trypsin (25-053-Cl, Thermo Fisher Scientific) and replated cells into six-well plates at a density of 1.5 million cells per well. Cells were then cultured in 'galactose media' (DMEM without glucose (A14430-01, Thermo Fisher Scientific), 1.7 mg/mL galactose (G5388, Sigma), 1mM Na pyruvate (11360-070, Gibco), 5mM HEPES (SH3023701, Thermo Fisher Scientific), 2mM GlutaMax, 10%FBS (4360-500, J R Scientific), and 100mg/mL Pennicillin/Streptomycin) to help to mature cardiomyocytes by forcing them to undergo aerobic metabolism [7, 8]. Regular media changes with 'galactose media' continued for the duration of the experiment. After an additional five days (day 25) cells were moved to an incubator at physiological oxygen levels (10%). Two days later (day 27) cells were subjected to electrical stimulation at 6.6 V/cm, 2 ms and 1 Hz (C-PaceEP Cell Culture EP Stimulator, IonOptix) for four days to help further mature the cells [9] and standardize beating rate across wells and lines. Cardiomyocytes were harvested on day 31/32. To establish the purity of iPSC-CMs flow cytometry was performed using a cardiac-specific marker, cardiac Troponin T (564767, BD Biosciences). All samples reported here were of a high purity (i.e. 40-97% of cells of each individual express cardiac Troponin T; see table S1)

Line	Purity
18499	80.6
18505	89.7
18511	76
18520	93.5
18852	75.4
18855	85
18870	87.1
18912	90
19098	96.9
19101	47.9
19108	81.4
19116	82.9
19128	40
19160	79.8

Table S1: Purity of iPSC-CMs as measured by flow cytometry using the cardiac troponin T marker..

4 LCL growth

LCLs were grown in cells in standard media (RPMI1640, 15%FBS, 0.1mM NEAA, 2mM GlutaMAX) and maintained at a density of between 200,000-500,000 viable cells/ml.

5 Sample Collection

After at least three passages in feeder-free conditions iPSCs were passaged into a 10cm culture dish. At near full confluence cells were enzymatically dissociated and counted. After dissociation all additional steps are performed on ice or in a temperature controlled centrifuge. One 10cm dish yields between 3 million and 15 million cells. From each line 400,000 cells were divided into two tubes to be used for ATAC-seq [10]. The tagmentation step of the ATAC-seq protocol was performed immediately on the two cell pellets containing 200,000 cells each. The library preparation of ATAC-seq samples was done in larger batches at a later time. The remaining material was split between three tubes for RNA and DNA extractions. We isolated RNA and DNA using the Zymo dual extraction kits (Zymo Research) with a DNase treatment during RNA extraction (Qiagen) on a single cell pellet from each line. 50 bp single-end RNA sequencing libraries were generated from extracted RNA using the Illumina TruSeq kit as directed by the manufacturer. Sequencing of samples was performed on an Illumina 2500. Extracted DNA was bisulphite-converted and hybridized to the Infinium MethylationEPIC array (Illumina) at the University of Chicago Functional Genomics facility.

iPSC-CMs were collected on ice using manual dissociation. We isolated RNA and DNA using the Zymo dual extraction kits (Zymo Research) with a DNase treatment during RNA extraction (Qiagen) on a single cell pellet from each line. RNA sequencing libraries were generated from extracted RNA using the Illumina TruSeq kit as directed by the manufacturer. 50bp single-end RNA-seq was performed on an Illumina 2500. A pellet containing 200,000 cells was used immediately for the tagmentation step of the ATAC-seq protocol was performed immediately on a pellet containing 200,000 cells. The library preparation of ATAC-seq samples was done in larger batches at a later time. LCLs were collected by pelleting 200,000 cells in a temperature controlled centrifuge. The tagmentation step of the ATAC-seq protocol was performed immediately on a pellet containing 200,000 cells. The library preparation of ATAC-seq samples was done in larger batches at a later time.

6 Molecular data processing

6.1 RNA-seq processing

RNA-seq from LCLs [11] and iPSCs were mapped using the STAR RNA-seq aligner [12] standard settings and processed using WASP to filter out reads that map with allelic bias [13]. RNA-seq reads from cardiomyocytes were mapped using Subread allowing for two mismatches and were also filtered using WASP for biases in allelic mapping. Reads overlapping SNPs were remapped to reduce reference bias as described previously [13]. Only reads with a MAPQ greater than 10 were retained. Read depths are provided in Table S2.

To determine the organs and cell-types with regulatory profiles most similar to our LCLs, iPSC, and iPSC-CMs **(Fig. 1B)**, we used kallisto [14] to quantify transcript (gencode v19) isoform abundances in our samples (LCLs, iPSC, iPSC-CM) along with samples from GTEx (Heart – Atrial Appendage, Heart – Left Ventricle, Bone Marrow, Pancreas, Testis, Brain, Kidney, Colon, Lung, Skin, and Blood) [15] and

ENCODE (H1-ESC, LCLs) [16]. We next identified a set of organ-regulated genes by using five samples for each organ. To do this, for each gene, we compared the distribution of transcript abundance in the organ to abundance of the transcript in all other organs using a Mann-Whitney U test. We then took 500 genes with the most differentially expressed transcripts (according to the Mann-Whitney U p-values) for each organ, which resulted in 5,000 organ-regulated genes. Lastly, we computed the Spearman correlation of the transcript abundances for these 5,000 genes for each sample pair.

6.2 ATAC-seq processing

Paired end ATAC-seq reads were mapped using bowtie2 allowing for two mismatches per read. The ATAC-seq protocol works by randomly inserting sequencing adapters into open chromatin via a tagmentation enzyme. One unfortunate side effect of this procedure is a high enrichment of reads originating from mitochondrial DNA (between 25%-75% of reads). Yet, we chose not to pursue an enrichment approach prior to sequencing, in order to avoid introducing additional variation to the nuclear gene expression data. After mitochondrial reads were removed, we re-mapped all nuclear reads using the WASP pipeline to filter out reads that did not uniquely map to the same genomic location after accounting for genetic variation [13]. We then removed all duplicate fragments (duplicates of both read pairs) and reads with a mapping quality (MAPQ) less than 10. Each mate represents an independent tagmentation event and therefore, after mapping and duplicate removal, reads were treated as single end in all future analyses. The quality of our ATAC-seq libraries in all three cell lines were very similar to that of ATAC-seq data in the LCL GM12878 cell line [10], ATAC-seq data in the H1hesc cell line (courtesy of the Greenleaf Lab) and DNase-seq data in the YRI LCL cell lines [17] (see figs. 11, 12, and 13). Final read depths are provided in Table S2 and the fractions of reads filtered out after each processing step are illustrated in Supplementary Figure 14.

6.3 DNase-seq processing

Previously collected DNase-seq from LCLs was used to assay chromatin accessibility. Reads were mapped using a custom mapper, which has been previously described in depth [18]. In this study counts per base were obtained directly from a previous study [17].

6.4 Methylation array processing

Methylation levels were assayed using the Infinium MethylationEPIC array (Illumina) in iPSCs and the Infinium HumanMethylation450 array (Illumina) in LCLs. Methylation data from LCLs were obtained from a previous study [19]. In iPSCs a number of steps were taken to ensure high quality data. First, to enable accurate quantification of methylation levels all probes that contained a SNP with a MAF greater than 5% in the population were removed. Next, we removed all CpGs that were not detected in at least 75% of individuals. CpGs on the X or Y chromosome were removed.

7 Reduced regulatory variation in iPSCs

To quantify the regulatory variation in gene expression, chromatin accessibility, and DNA methylation levels, we calculated the average square distance from the mean for each individual n as defined as:

$$V_n = \frac{N}{L(N-1)} \sum_{l=1}^{L} \frac{(x_{nl} - \bar{x})^2}{\bar{x}^2}$$

for loci l and locus mean $\bar{x}$. iPSCs have consistently lower variation compared with LCLs and iPSC-CMs across all three regulatory phenotypes (Supplementary Figure 1), and significantly lower variation in gene expression (LCLs: $p < 10^{-6}$; iPSC-CMs: $p < 10^{-5}$; Mann–Whitney U) and DNA methylation (LCLs; $p < 10^{-11}$) levels. Chromatin accessibility was significantly lower in iPSCs compared to iPSC-CMs ($p = 2 \times 10^{-4}$) but not LCLs ($p = 0.42$).

8 QTL mapping

8.1 Identifying eQTLs

To identify eQTLs in iPSCs and LCLs we transformed expression levels to a standard normal within each individual (iPSC: n= 59, LCL: n= 59). We next accounted for unknown confounders by removing principal components from the LCL (15 PCs) and iPSC (10 PCs) data. Genotypes were obtained using impute2 as described previously [20]. As in previous work, we were limited to examining putatively cis-acting genetic variants. Therefore, we only considered variants within 50kb of genes. To identify association between genotype and gene expression we used fastqtl [21]. After the initial regression, a variable number of permutations were performed to obtain a gene-wise adjusted p-value [21]. To identify significant eQTLs we used Storey's q-value [22] on the adjusted p-values. Genes with a q-value less than 0.1 are considered significant.

The sample size of iPSC-CMs in this study was too limited to identify eQTLs using a standard regression model. We therefore used the combined haplotype test (CHT) [13] to identify eQTLs. This method allowed us to identify eQTLs with small sample sizes by using both regression and allelic imbalance tests in combination. In this analysis, we focused on variants within 25kb of a gene. Following the procedure outlined by the

Sample	iPSC (ATAC)	LCL (ATAC)	LCL (DNase)	iPSC-CM (ATAC)
NA18486	32659004	-	42320431	-
NA18489	18879326	-	-	-
NA18498	30441944	-	42626294	-
NA18499	21191168	-	42508766	6198578
NA18501	35810874	15674600	33492742	-
NA18502	18752114	14083042	27546193	-
NA18504	-	36825244	27849233	-
NA18505	31937588	5371198	27446048	27664492
NA18507	-	-	74207029	-
NA18508	37029474	-	45294967	-
NA18510	15300428	-	46266665	-
NA18511	10675520	10560598	34325288	14151550
NA18516	-	-	28573172	-
NA18517	25455280	-	30218325	-
NA18519	17595712	32703926	44023326	-
NA18520	18571190	7302724	23360405	13295658
NA18522	56216208	7823794	23126812	-
NA18852	30521566	21607142	45901061	17157142
NA18853	11558304	-	29581654	-
NA18855	9065570	19680684	33802245	11861486
NA18856	13259732	-	30704705	-
NA18858	30453734	-	30070868	-
NA18859	13862852	-	29473448	-
NA18861	2635618	-	80235013	-
NA18862	19335654	-	32613319	-
NA18870	12410264	3287592	29196263	26946240
NA18907	26458794	-	44650582	-
NA18909	-	-	47794281	-
NA18912	23656490	87500262	27791308	12107644
NA18913	35096114	-	35757428	-
NA18916	-	-	31784568	-
NA19092	28648676	-	35404700	-
NA19093	3270958	-	13815173	-
NA19098	40165190	29180170	38472816	5325082
NA19099	40707618	-	29772954	-
NA19101	-	13478794	35503543	111719728
NA19102	45989426	6993378	-	-
NA19108	34009578	15842716	-	90819360
NA19114	4488602	-	48895531	-
NA19116	34003684	29583136	27680918	9386758
NA19119	15494768	-	29855656	-
NA19127	28119164	-	36647791	-
NA19128	28142480	34435242	-	26961610
NA19130	39817074	-	28993557	-
NA19131	-	-	35976533	-
NA19138	33720264	-	38512510	-
NA19140	15555202	-	34856598	-
NA19141	-	-	26834843	-
NA19143	22009676	-	35355157	-
NA19144	8856202	-	38393641	-
NA19147	-	-	48947479	-
NA19152	32112704	-	35249757	-
NA19153	14407548	-	30221567	-
NA19159	46076592	-	29652941	-
NA19160	26644496	4791876	36152805	46752622
NA19171	-	-	38884245	-
NA19190	25803656	-	36741524	-
NA19192	33147506	-	70691903	-
NA19193	10564818	-	39671681	-
NA19200	-	-	39434479	-
NA19201	-	-	36013821	-
NA19203	-	-	35304566	-
NA19204	29246864	-	40850302	-
NA19206	29077392	-	35333265	-
NA19207	21663330	-	35517782	-
NA19209	19455060	-	29632699	-
NA19210	10811076	-	-	-
NA19222	-	-	30295677	-
NA19223	-	-	34894059	-
NA19225	25977510	-	49138423	-
NA19239	10892712	-	108561800	-
NA19257	16167520	14729492	47206531	-

Table S2: Final sequencing read depths for chromatin accessibility

authors [22], we performed the CHT and one permutation of the CHT. We noted that our tests were not well calibrated, owing to the small number of samples. We therefore identified significant SNPs by performing Storey's q-value correction [22] on the null data. We then identified the largest p-value in the null data with a q-value less than 0.1. We used this p-value as a threshold in the non-permuted data to identify significant eQTLs.

8.2 Identifying meQTLs

To identify meQTLs in iPSCs and LCLs, we transformed methylation levels to a standard normal within each individual (iPSC: n= 58; LCL: n= 64) and principal components were removed to account for unknown confounders (iPSC: 6 PCs removed; LCLs: 5 PCs removed). In accordance with previous work, genetic variants within 3kb of a CpG were tested for associations with methylation levels. Methylation QTLs were identified using the fastqtl software [21] following the procedure described above. We inherently identified a larger number of meQTLs in iPSCs compared to LCLs due to the larger number of CpGs tested. However, we also compared only the CpGs shared across both arrays and found more meQTLs in iPSCs (n= 7,958; n= 5,738).

8.3 Identifying caQTLs

We first identified a comprehensive set of genomic loci that were open in iPSC, LCLs, or iPSC-CMs using the following approach. Focusing on the 12 individuals from whom we have data in all three cell types, we pooled the ATAC-seq data for all 12 individuals to create a chromatin accessibility track for each cell type; the data for each track corresponds to the count of transpositions (T_l) observed at each base pair. To account for differences in total read depth and differences in library quality (Supplementary Figure 11), we first converted the raw data track to a percentile track for each cell type as follows. Each position in the genome was assigned the total transposition count within the 100 bp window centered at that position, $S_l = \sum_{i=l-50}^{l+50} T_i$. Using an empirical distribution of S_l, each genomic position was assigned a percentile computed as $P_l = \frac{\sum_i \mathbb{I}[S_l > S_i]}{L}$, where $\mathbb{I}[\cdot]$ is the indicator function and L is the size of the genome. Given the percentile tracks for each cell type, we identified all genomic positions with $P_l >= 90$ in at least one cell type, added a 50 bp flank to either end of each such position to create open chromatin intervals, and merged overlapping intervals to obtain $2,533,844$ loci.

Starting with these 2.5 million genomic loci, we focused our analysis on common SNPs (MAF > 0.05) that were located within each locus to identify variants that are associated with and, likely, directly disrupt chromatin accessibility. For each cell type, we tested for association between these SNPs and the ATAC-seq signal across all individuals using the combined haplotype test in WASP [13]. Filtering out loci that did not contain a common SNP or did not have sufficient ATAC-seq reads overlapping the segregating SNPs, we ultimately tested 599,178 loci in iPSCs, 187,504 loci in LCLs, and 203,619 loci in iPSC-CMs. Using Storey's q-value [22], we identified significant caQTLs at 10% FDR (see table S3).

To identify distal caQTLs, we used ATAC-seq data from iPSCs (n=58) and DNase-seq data from LCLs (n=68). Chromatin accessibility levels were fit to a standard normal across individuals and qqnormed within individual [17]. Principal components were removed to account for unknown confounders (iPSCs: 1 PC removed; LCLs: 2 PCs removed). Associations between genetic variants within 500kb of a peak and chromatin accessibility levels were identified using fastqtl [21].

cell-type	eQTL	caQTL	distal caQTL	meQTL
LCL	2556 (n=85)	3902 (n=20)	2260 (n=68)	5738 (n=64)
iPSC	1441 (n=57)	20370 (n=57)	2130 (n=57)	29782 (n=57)
iPSC-CM	517 (n=13)	4045 (n=14)	-	-

Table S3: Counts of QTLs for each molecular phenotype in each cell type. Sample sizes are listed in parentheses.

8.3.1 Peak calling using MACS2 (Fig. 2B)

To identify a more stringent set of accessible regions in our cell types, we used MACS2 to call peaks in all individual ATAC-seq samples separately:

```
macs2 callpeak --treatment bamfile --gsize hs --format BAMPE -q 0.01
```

We next merged all peaks for each individual samples by cell-type, requiring that a peak has a 15X fold change enrichment over background signal.

9 Estimating QTL sharing

Storey and Tibshirani developed a method to estimate the true proportion of null statistics from a given p-value distribution [22]. This metric (π_0) can be used to calculate the proportion of significant tests from a p-value distribution by taking $1 - \pi_0$ (π_1). Here we calculate π_1 for eQTLs, caQTL, and meQTLs between cell types. To obtain a better estimate of the true sharing we generated π_1 statistics for a range of stringencies. Specifically, for eQTLs and caQTLs we calculated π_1 cumulatively from the top 150 most significant genes/loci to the top 2000 most significant genes/loci in intervals of 25 genes/lcoi. For meQTLs we calculated π_1 from the top 500 CpGs to the top 10,000 CpGs in intervals of 100 CpGs. As is clear from the density plots (Supplementary Figure 3), small deviations in threshold choice can create local valleys and peaks in sharing estimates. This method allows us to see sharing across a wide space of stringencies.

10 Identifying cell type specific and shared eQTLs/caQTLs

To identify cell type specific associations, we first removed loci that were tested in only one cell-type. Next, any locus with a significant association (even with a different lead variant) in both cell types was removed. An eQTL was considered cell-type specific if significant at an FDR of 10% in one cell type and a nominal p-value of greater than 0.05 in the second cell-type (as determined by fastqtl [21]). QTLs were considered shared if they were significant with a p-value of less than 10^{-5} in both cell types.

To define cell type specific caQTLs, we required a WASP p-value smaller or equal to 5×10^{-4} in one cell-type, and a p-value greater than 0.05 in the second cell-type. **(Fig. 2A,B,E)**

10.1 Linking cell-type specific caQTL to eQTL signal (Fig. 2A)

We used a one-sided Fisher's exact test to determine the level of significance at which the number of iPSC-specific caQTL that are also iPSC eQTLs is greater than the number of LCL-specific caQTLs are also iPSC eQTLs (and vice-versa). This yielded a p-value of 4.7×10^{-5} and 0.01 for the two comparisons, respectively. This result is robust with respect to various thresholds at which we defined LCL and iPSC eQTLs (e.g. 10^{-2}, 10^{-3},10^{-4},10^{-5}).

10.2 Accessibility at cell-type specific caQTLs (Fig. 2B)

To determine accessibility at cell-type specific caQTLs, we centered all windows (1500bp total) and computed the mean number of ATAC-seq reads falling at each position (s), normalized by total read counts for each individuals. We then computed the $\log_2$ signal plus 1 ($\log_2(s+1)$) signal at each base, ranked all windows by average signal across the window, and plotted in decreasing order of signal.

10.3 Accessibility at cell-type specific caQTLs (Fig. 2E)

We computed the mean accessibility at each open chromatin peak by summing over the number of reads falling into each peak and normalizing by total number of reads and peak length.

To obtain a set of iPSC-specific caQTLs that also affect expression of distal genes, we identified cell-type-specific caQTLs SNPs that were also associated with expression level of a nearby gene (100kb) in iPSC with

Cell-type	LCL-eQTL	not LCL-eQTL
LCL-specific caQTL	35	449
iPSC-specific caQTL	38	1309

Table S4: Contingency table for cell type specific caQTLs that are LCL eQTLs at p-value $< 10^{-3}$.

Cell-type	iPSC-eQTL	not iPSC-eQTL
LCL-specific caQTL	6	449
iPSC-specific caQTL	44	1309

Table S5: Contingency table for cell-type-specific caQTLs that are iPSC eQTLs at p-value $< 10^{-3}$.

a nominal p-value of at most 0.001. We noted that a few iPSC-specific caQTL SNPs were also associated with gene expression level in LCLs despite showing no association for local chromatin accessibility. This observation is consistent with the possibility that some caQTLs SNPs within chromatin accessibility peaks are not the causal SNP. We note however that for the vast majority of SNPs, the observations are consistent with a distal effect of caQTLs on expression levels of nearby genes in the corresponding cell type (Fig. 2D).

List of iPSC-specific caQTLs with effects at distal genes
`http://eqtl.uchicago.edu/yri_ipsc/Table_caQTL_eQTL_iPSC.txt`

Table S6: iPSC-specific caQTLs that also affect expression of distal genes

11 Hierarchical model to fine-map causal eQTLs and caQTLs

Starting with summary statistics computed when identifying eQTLs and caQTLs, we used a Bayesian hierarchical model (`https://github.com/rajanil/qtlBHM`) to classify causal SNPs by incorporating annotations such as chromatin states and TF binding sites. The summary statistics used by the method are the effect size and standard error; the method is explained in detail elsewhere [20].

For caQTLs, we computed the effect size (γ) from the allele-specific accessibility parameters reported by WASP (α and β) as $\gamma = \log \alpha - \log \beta$. To compute the standard error, we first computed the z-score from the reported p-value using a t-distribution with degrees of freedom as a function of the sample size. Since causal SNPs in caQTLs typically disrupt transcription factor binding sites, we first inferred binding for expressed TFs in the three cell types using msCentipede (`https://github.com/rajanil/msCentipede`) and used these inferred binding sites as genomic annotations in the hierarchical model. Inference under the model gives us the posterior probability that a locus is a caQTL, and the posterior probability that each SNP within the locus is the causal SNP conditional on the locus being a caQTL. Given these probabilities, we then computed the proportion of caQTLs that are explained by different TF binding annotations in each cell type.

For eQTLs, the effect sizes and standard errors were directly obtained from the output of fastqtl [21] for iPSCs and LCLs, respectively, and computed from the output of WASP for iPSC-CMs. The output of qtlBHM included the inferred probabilities that a gene is an eQTL and the probability that a tested SNP is the causal SNP for a given eQTL gene. Among genes with posterior greater than 0.9 in one cell type, a gene was labeled as a shared eQTL if it had posterior greater than 0.9 in any other cell type, and it was labeled as a specific eQTL if it had posterior less than 0.1 in all other cell types. We were interested in characterizing the properties of shared and cell type-specific eQTLs, and to this end, we partitioned each genomic annotation into shared and cell type-specific annotations. For example, a particular genomic position was considered to be a shared promoter for iPSCs if the chromHMM state for that position is 'Promoter' in iPSCs and in at least one of the other two cell types, else the position is labeled an iPSC-specific promoter. For genomic annotations, we used chromHMM states inferred by the Roadmap Epigenomics Project (LCL: Gm12878 cell line, iPSC: H1hesc cell line, iPSC-CM: Right Ventricle) and caQTLs inferred previously. As before, we computed the proportion of eQTLs that are explained by each of the cell type-specific and shared genomic annotations (Supplementary Figure 15).

12 GTEx data

Two types of data were collected from the GTEx consortium [15]. First we obtained summary statistics from the *cis* eQTL analysis. Specifically, for every gene tested in a tissue, the *p*-value of the lead variant was obtained. To overlap with eQTLs identified in iPSC-CMs the variant identified in the GTEx data was tested in iPSC-CMs. The QQ-plot was generated from a limited number of tissues for clarity (figs. 2). Raw RNA-seq fastq files were obtained (15 samples each) in Heart − Atrial Appendage, Heart − Left Ventricle, Bone Marrow, Pancreas, Testis, Brain, Kidney, Colon, Lung, Skin, and Blood.

13 GWAS signal enrichments in gene expression data (Fig. 4A)

We used RolyPoly, a polygenic method that identifies trait-involved cell types by analyzing the enrichment of GWAS signal in cell type specific gene expression genome-wide. First, for each gene we calculate trait association scores by aggregating GWAS summary statistics from a window (10kb) centered on the TSS. Then, we estimate the individual contribution of each cell type to the observed gene score variance using a generalized linear regression model with normalized gene expression features. For each cell type we estimate an effect size coefficient and standard error, which we use for hypothesis testing. RolyPoly is available as an R package here `https://github.com/dcalderon/rolypoly`.

14 Stratified LDscore regression (Fig. 4B)

We obtained the LDscore regression software [23] from `https://github.com/bulik/ldsc` and ran it using

```
python ldsc.py --h2 sumstats_file --ref-ld-chr annotation --w-ld-chr
eur_w_ld_chr --overlap-annot --frqfile-chr 1000G.mac5eur. --out outfile
--print-coefficient
```

To obtain heritability enrichment estimates, we ran stratified LDscore regression [24] with each annotation separately plus annotations from the published baseline annotations, including:

Coding_UCSC, Coding_UCSC.extend.500, Conserved_LindbladToh, Conserved_LindbladToh.extend.500, Enhancer_Andersson, Enhancer_Andersson.extend.500, Enhancer_Hoffman, Enhancer_Hoffman.extend.500, Intron_UCSC, Intron_UCSC.extend.500, Promoter_UCSC, Promoter_UCSC.extend.500, SuperEnhancer_Hnisz, SuperEnhancer_Hnisz.extend.500, UTR_3_UCSC, UTR_3_UCSC.extend.500, UTR_5_UCSC, UTR_5_UCSC.extend.500, WeakEnhancer_Hoffman, and WeakEnhancer_Hoffman.extend.500

Summary statistics file were obtained from [25] (Rheumatoid Arthritis), [26] (Crohn's disease), the GIANT consortium (BMI), and CARDIoGRAMplusC4D (MI & CAD). To obtain a 95% confidence interval, we used $fc \pm 1.96s$, where fc and s are the LDscore fold enrichment estimates and standard error, respectively.

15 Neural network models for chromatin accessibility

Neural networks are highly expressive, nonlinear models that typically map a matrix representation of an input (here, DNA sequence of a genomic locus) to some output of interest (here, the chromatin accessibility state of that locus across cell types). Below, we describe in detail the different components of neural network models relevant to our problem of predicting chromatin accessibility, the model architectures of varying complexities used for our problem, our algorithm for learning the optimal model parameters, and key inferences made from our final model.

15.1 Model components

In deep neural networks, highly complex nonlinear mappings are typically constructed as a composition of several simpler nonlinear (or piece-wise linear) activation functions. Neural networks are typically represented by layers of variables (also called neurons) that connect an input layer of variables to an output layer of variables. A variable in each layer is a function of a linear combination of some subset of variables in the previous layer; the larger the number of layers, the deeper the neural network. The input layer consists of all the input variables in the model and the output layer consists of all the variables to be predicted.

Some commonly used intermediate layers are the dense layer, the convolutional layer, and the maxpool layer, and one commonly used activation function is the Rectified Linear Unit (ReLU).

In a dense layer with $\mathcal{J}$ variables, each variable depends on all the variables in the previous layer. Given a set of input variables $X_i, i \in \{1, \ldots, \mathcal{I}\}$, each output variable Y_j of the dense layer can be defined as

$$Y_j = f\left(\sum_{i=1}^{\mathcal{I}} d_{ij} X_i\right) \tag{1}$$

A convolutional layer with $\mathcal{L}$ units is specified by $\mathcal{L}$ convolutional filters, each filter is represented by a matrix $\mathbf{C}^l_{mn}, m \in \{1, \ldots, \mathcal{M}\}, n \in \{1, \ldots, \mathcal{N}\}$. Given an input matrix $X_{ij}, i \in \{1, \ldots, \mathcal{I}\}, j \in \{1, \ldots, \mathcal{J}\}$, each filter maps the input matrix X to an output matrix Y^l of dimension $\mathcal{P} \times \mathcal{Q}$ where $\mathcal{P} = \mathcal{I} - \mathcal{M} + 1$ and $\mathcal{Q} = \mathcal{J} - \mathcal{N} + 1$.

$$Y^l_{pq} = f\left(\sum_{m,n} \mathbf{C}^l_{mn} X_{p+m,q+n}\right) \tag{2}$$

A maxpool layer with width w contains a single maxpool operation that is applied to contiguous, spatially ordered variables. Given a set of spatially ordered variables $X_i, i \in \{1, \ldots, \mathcal{I}\}$, the maxpool operation gives as output a new set of spatially ordered variables $Y_j, j \in \{1, \ldots, \frac{\mathcal{I}}{w}\}$ defined as

$$Y_j = \max\{X_{w \times j}, X_{w \times j+1}, \ldots, X_{w \times (j+1)}\} \tag{3}$$

The maxpool operation effectively smooths over local perturbations in a signal of interest. For example, if co-binding of two transcription factors occur as long as their high-affinity sequence elements lie within 10-15bp from each other, this co-binding signal can be detected more robustly by including a maxpool operation of width 5.

Finally, a common choice for $f(x)$ is the ReLU, defined as $f(x) = \max\{0, x\}$.

15.2 Model architectures

We are interested in predicting the chromatin activity of a genomic locus across three cell types (iPSC, LCL and iPSC-CM) from the DNA sequence of the locus alone. Thus, the input to the neural network model is a one-hot encoding of the reference DNA sequence of length 500bp centered at the locus, and the Input Layer consists of 4×500 binary-valued variables. The output of a neural network model is a categorical variable $O \in \{1, \ldots, 7\}$ where the values of the variables denote the following

$$O = \begin{cases} 1, & \text{if open in iPSC-CM alone} \\ 2, & \text{if open in LCL alone} \\ 3, & \text{if open in iPSC-CM and LCL} \\ 4, & \text{if open in iPSC alone} \\ 5, & \text{if open in iPSC and iPSC-CM} \\ 6, & \text{if open in iPSC and LCL} \\ 7, & \text{if open in all three cell types} \end{cases} \tag{4}$$

Since we are specifically interested in detecting sequence features that capture cell-specific and shared chromatin activity, we only include loci that are open in at least one cell type; this helps us ignore obvious sequence features (e.g., GC content) that distinguish open and closed chromatin. We used the sigmoid activation function to model the probability of the categorical variable in the output layer.

Fig. 7 illustrates several neural network models of varying depth and complexity (including OrbWeaver) that all use the same input and output layers. While the parameters of all the intermediate layers need to be estimated for each of the models, the filters of the first convolutional layer in OrbWeaver were kept fixed to log-transformed position weight matrices (PWMs) of $1,320$ human transcription factors. For each TF, we used PWMs curated from two sources – TRANSFAC [27] and HT-SELEX [28] – and, if a TF had more than one PWM, we selected the PWM with the highest information content. This approach is distinct from that taken in DanQ [29], where the filters in the first convolutional layer were initialized at known PWMs, but were still treated as free parameters to be estimated.

15.3 Model learning

We used a training set of $282,088$ loci to learn the parameters of each model using Adadelta [30], a stochastic gradient descent algorithm that adaptively computes learning rates. At each gradient descent step, the gradient is computed using a random batch of 1000 loci and optimization is terminated after one pass through the entire dataset. Fig. 8 compares the performance of each of the models on a test dataset of $7,151$ loci on chromosome 18. We computed the prediction for each cell type by collapsing the probability $p(O)$ for each of the 7 categories to marginal probabilities of chromatin activity.

$$\begin{aligned} p(\text{open in iPSC}) &= p(O=4) + p(O=5) + p(O=6) + p(O=7) \\ p(\text{open in LCL}) &= p(O=2) + p(O=3) + p(O=6) + p(O=7) \\ p(\text{open in iPSC-CM}) &= p(O=1) + p(O=3) + p(O=5) + p(O=7) \end{aligned} \tag{5}$$

We observed a diminishing increase in accuracy with increasing depth, and found that including known PWM information in the neural network helped increase the accuracy substantially. We haven't systematically explored a large space of deep neural networks, and acknowledge the possibility that a better designed, more complex architecture could outperform OrbWeaver. However, our analyses suggest that the gain in accuracy that could be achieved by deeper, more complex neural networks could more easily be achieved by incorporating known information about mechanisms underlying the output of interest (in this case, TF binding) into simpler neural networks.

15.4 Identifying key transcription factors for each category of chromatin activity

Fixing the filters in the first convolutional layer to known TF PWMs allowed us to directly query and interpret the importance of each of the factors in predicting active chromatin belonging to one of the 7 categories. We computed importance scores using DeepLIFT [31], and for each of the 7 categories, we used loci belonging to that category if the model correctly predicted their category. For each locus, we calculated DeepLIFT scores on the input with respect to each filter in the first convolutional layer; this gives us a score for each TF at each position in the locus (i.e., a score matrix of size 1320×500). Note that these scores are different from those reported in the DeepLIFT paper, where DeepLIFT scores were typically computed with respect to the input layer rather than any intermediate layer. Each entry of the matrix quantifies the importance of the sequence affinity of a TF at a given position towards the category-specific prediction for that locus. For each locus, we designated the TF with the highest value in the score matrix as the TF that explains the chromatin activity state of that locus. For each category, we then computed a histogram across TFs of the proportion of loci explained by each TF. Table S7 lists the top 15 TFs for each category, and the proportion of sites explained by them.

iPS-CM only (frac)	LCL only (frac)	iPS-CM+LCL (frac)	iPSC only (frac)	iPSC+iPS-CM (frac)	iPSC+LCL (frac)	iPSC+LCL+iPS-CM (frac)
MEF2A (0.2359) [32, 33, 34]	**IRF1 (0.1917) [35, 36]**	**Cap1 (0.1264)** [37]	**Oct3/4 (0.3850) [38, 39]**	**TEAD4 (0.4068)**	**Rad21 (0.2699)** [40]	**Rad21 (0.4480)** [41]
TEAD4 (0.1835) [42]	IRF-6 (0.1598)	**c-Fos (0.1368)** [43]	**Sox2/6 (0.1358)** [44]	TEAD3 (0.1398)	**ZEB1 (0.1143)** [45]	**CTCF (0.1177) [40, 46]**
GATA2 (0.1840) [47]	**PU.1 (0.1470) [48, 49]**	**FoxH1 (0.0768)** [50]	**ZEB1 (0.0534)** [51]	**Oct3/4 (0.0592)**	SPIB (0.0424)	LRF (0.0224)
TEAD3 (0.0392)	PEBP (0.0937)	**MEF2A (0.0784)**	ZIC4 (0.0519)	**Sox2/6 (0.0416)**	EBF1 (0.0305)	Mad (0.0212)
GATA (0.0335)	RelB (0.0619)	**TEAD4 (0.0624)**	Rad21 (0.0509)	Rad21 (0.0365)	PRDM6 (0.0261)	ESET (0.0178)
NFIX (0.0387)	RPC155 (0.0426)	STAT5A (0.0440)	TEAD4 (0.0295)	Zbtb5 (0.0189)	NFKB2 (0.0261)	BRCA1 (0.0173)
NFE2 (0.0624)	IRF (0.0381)	PEBP (0.0368)	RPC155 (0.0163)	ZNF823 (0.0164)	RelB (0.0239)	CtBP1 (0.0157)
MRG2 (0.0395)	Oct-2 (0.0378)	ZNF154 (0.0328)	Sp2 (0.0137)	ZNF579 (0.0151)	Oct-2 (0.0218)	SP4 (0.0150)
ESRRB (0.0265)	NFKB2 (0.0263)	**IRF1 (0.0296)**	MAZR (0.0137)	**MEF2A (0.0151)**	ZNF480 (0.0196)	SMC-3 (0.0135)
JunB (0.0234)	PAX5 (0.0258)	STAT4 (0.0288)	KLF14 (0.0137)	c-MAF (0.0139)	**IRF1 (0.0185)**	sin3A (0.0132)
TFF-1 (0.0174)	PRDM1 (0.0175)	SPIB (0.0224)	RFX5 (0.0193)	NFIX (0.0202)	SNAI2 (0.0163)	ZNF823 (0.0130)
NF-1 (0.0120)	Six-5 (0.0108)	MEF-2C (0.0216)	SMAD1 (0.0127)	Gsc (0.0113)	PAX5 (0.0163)	NF-E4 (0.0124)
ZNF514 (0.0115)	IRF5 (0.0103)	Gata1 (0.0296)	INSM1 (0.0122)	NF-E4 (0.0101)	P50 (0.0163)	Rb (0.0122)
meis1 (0.0082)	EBF1 (0.0102)	TGIF (0.0168)	GRHL1 (0.0117)	NF-AT3 (0.0101)	**Oct3/4 (0.0163)**	ZNF579 (0.0116)
Hand1 (0.0061)	ZNF860 (0.0098)	MRG2 (0.0152)	SP4 (0.0112)	Sp2 (0.0088)	PEBP (0.0152)	NRF1 (0.0104)

Table S7: Key transcription factors that are predictive of chromatin activity in shared and cell-specific open chromatin loci. For each TF, the number in parentheses denote the fraction of all loci in that category that have the TF PWM as the most important variable in the neural network, as quantified by DeepLIFT scores. Key factors that are known to be important for cell-specific and shared regulatory activity are shown in bold, with citations to previous work indicated alongside them.

15.5 Predicting effects of genetic variation on chromatin accessibility

We now set out to compare the observed (estimated) effects of genetic variants on chromatin accessibility at loci tested for caQTLs, and the effects of genetic variants predicted by OrbWeaver. First, for each cell type, we used qtlBHM, a Bayesian hierarchical model, without any annotation to compute the probability that a locus is a caQTL (π_l) and the probability that a SNP is the causal variant for a locus conditional on the locus being a caQTL (π_s). Restricting to loci with $\pi_l > 0.99$ and $\pi_s > 0.99$, using a 500 bp window centered at the causal variant of each such locus, we computed the OrbWeaver prediction at each of the 240 haplotypes (corresponding to 120 YRI individuals). Partitioning the haplotypes based on the alleles of the causal SNP, we then computed the difference in the median prediction of chromatin activity between the reference and alternate alleles for each of the three cell types. Fig. 9 compares this predicted allelic difference in chromatin activity with the observed allelic imbalance in ATAC-seq reads $\left(\frac{1}{2}\frac{\beta-\alpha}{\beta+\alpha}\right)$, computed using the α and β parameters output by WASP for that SNP-locus pair.

16 Raw and processed data availability

All data will be made available at `http://eqtl.uchicago.edu/Home.html`.

Data	Accession
DNA methylation (LCL)	GSE57483 (GEO)
DNase-seq	GSE31388 (GEO)
RNA-seq (GEUVADIS)	E-GEUV-3 (ArrayExpress)
RNA-seq (iPSC, iPSC-CM)	GSE89895
ATAC-seq (iPSC, iPSC-CM)	GSE89895
DNA methylation (iPSC)	GSE89895

References

[1] K. Okita, Y. Matsumura, Y. Sato, A. Okada, A. Morizane, S. Okamoto, H. Hong, M. Nakagawa, K. Tanabe, K. Tezuka, T. Shibata, T. Kunisada, M. Takahashi, J. Takahashi, H. Saji, and S. Yamanaka. A more efficient method to generate integration-free human iPS cells. *Nat. Methods*, 8(5):409–412, May 2011.

[2] C. K. Burrows, N. E. Banovich, B. J. Pavlovic, K. Patterson, I. Gallego Romero, J. K. Pritchard, and Y. Gilad. Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet.*, 12(1):e1005793, Jan 2016.

[3] F. J. Muller, B. M. Schuldt, R. Williams, D. Mason, G. Altun, E. P. Papapetrou, S. Danner, J. E. Goldmann, A. Herbst, N. O. Schmidt, J. B. Aldenhoff, L. C. Laurent, and J. F. Loring. A bioinformatic assay for pluripotency in human cells. *Nat. Methods*, 8(4):315–317, Apr 2011.

[4] X. Lian, J. Zhang, S. M. Azarin, K. Zhu, L. B. Hazeltine, X. Bao, C. Hsiao, T. J. Kamp, and S. P. Palecek. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/β-catenin signaling under fully defined conditions. *Nat Protoc*, 8(1):162–175, Jan 2013.

[5] P. W. Burridge, E. Matsa, P. Shukla, Z. C. Lin, J. M. Churko, A. D. Ebert, F. Lan, S. Diecke, B. Huber, N. M. Mordwinkin, J. R. Plews, O. J. Abilez, B. Cui, J. D. Gold, and J. C. Wu. Chemically defined generation of human cardiomyocytes. *Nat. Methods*, 11(8):855–860, Aug 2014.

[6] S. Tohyama, F. Hattori, M. Sano, T. Hishiki, Y. Nagahata, T. Matsuura, H. Hashimoto, T. Suzuki, H. Yamashita, Y. Satoh, T. Egashira, T. Seki, N. Muraoka, H. Yamakawa, Y. Ohgino, T. Tanaka, M. Yoichi, S. Yuasa, M. Murata, M. Suematsu, and K. Fukuda. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell*, 12(1):127–137, Jan 2013.

[7] G. Wang, M. L. McCain, L. Yang, A. He, F. S. Pasqualini, A. Agarwal, H. Yuan, D. Jiang, D. Zhang, L. Zangi, J. Geva, A. E. Roberts, Q. Ma, J. Ding, J. Chen, D. Z. Wang, K. Li, J. Wang, R. J. Wanders, W. Kulik, F. M. Vaz, M. A. Laflamme, C. E. Murry, K. R. Chien, R. I. Kelley, G. M. Church, K. K. Parker, and W. T. Pu. Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat. Med.*, 20(6):616–623, Jun 2014.

[8] L. D. Marroquin, J. Hynes, J. A. Dykens, J. D. Jamieson, and Y. Will. Circumventing the Crabtree effect: replacing media glucose with galactose increases susceptibility of HepG2 cells to mitochondrial toxicants. *Toxicol. Sci.*, 97(2):539–547, Jun 2007.

[9] Y. C. Chan, S. Ting, Y. K. Lee, K. M. Ng, J. Zhang, Z. Chen, C. W. Siu, S. K. Oh, and H. F. Tse. Electrical stimulation promotes maturation of cardiomyocytes derived from human embryonic stem cells. *J Cardiovasc Transl Res*, 6(6):989–999, Dec 2013.

[10] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and 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(12):1213–1218, Dec 2013.

[11] T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill, E. T. Dermitzakis, X. Estivill, R. Guigo, E. Dermitzakis, S. Antonarakis, T. Meitinger, T. M. Strom, A. Palotie, J. F. Deleuze, R. Sudbrak, H. Lerach, I. Gut, A. C. Syvanen, U. Gyllensten, S. Schreiber, P. Rosenstiel, H. Brunner, J. Veltman, P. A. Hoen, G. J. van Ommen, A. Carracedo, A. Brazma, P. Flicek, A. Cambon-Thomsen, J. Mangion, D. Bentley, and A. Hamosh. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[12] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[13] B. van de Geijn, G. McVicker, Y. Gilad, and J. K. Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063, Nov 2015.

[14] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, May 2016.

[15] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y. H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[16] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward,

T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter,

J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W.

Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[17] J. F. Degner, A. A. Pai, R. Pique-Regi, J. B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, Feb 2012.

[18] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, 21(3):447–455, Mar 2011.

[19] N. E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, 10(9):e1004663, Sep 2014.

[20] Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, Apr 2016.

[21] H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, and O. Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, May 2016.

[22] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, 100(16):9440–9445, Aug 2003.

[23] B. K. Bulik-Sullivan, P. R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B.

Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, M. J. Cairns, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, B. Crespo-Facorro, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodriguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. Keong, S. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lonnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Muller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilainen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Soderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. D. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C.

Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, R. Adolfsson, O. A. Andreassen, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. B?rglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nothen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. O'Donovan. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, Mar 2015.

[24] H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P. R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, S. Purcell, E. Stahl, S. Lindstrom, J. R. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale, and A. L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235, Nov 2015.

[25] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[26] A. Franke, D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, C. Wijmenga, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Buning, A. Cohen, J. F. Colombel, M. Cottone, L. Stronati, T. Denson, M. De Vos, R. D'Inca, M. Dubinsky, C. Edwards, T. Florin, D. Franchimont, R. Gearry, J. Glas, A. Van Gossum, S. L. Guthery, J. Halfvarson, H. W. Verspaget, J. P. Hugot, A. Karban, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, C. Mowat, W. Newman, J. Panes, A. Phillips, D. D. Proctor, M. Regueiro, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, F. Seibold, A. H. Steinhart, P. C. Stokkers, L. Torkvist, G. Kullak-Ublick, D. Wilson, T. Walters, S. R. Targan, S. R. Brant, J. D. Rioux, M. D'Amato, R. K. Weersma, S. Kugathasan, A. M. Griffiths, J. C. Mansfield, S. Vermeire, R. H. Duerr, M. S. Silverberg, J. Satsangi, S. Schreiber, J. H. Cho, V. Annese, H. Hakonarson, M. J. Daly, and M. Parkes. Genome-wide

meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, 42(12):1118–1125, Dec 2010.

[27] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(Database issue):D108–110, Jan 2006.

[28] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, Jan 2013.

[29] D. Quang and X. Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, 44(11):e107, Jun 2016.

[30] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[31] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016.

[32] E. P. Ewen, C. M. Snyder, M. Wilson, D. Desjardins, and F. J. Naya. The Mef2A transcription factor coordinately regulates a costamere gene program in cardiac muscle. *J. Biol. Chem.*, 286(34):29644–29653, Aug 2011.

[33] R. Papait, P. Cattaneo, P. Kunderfranco, C. Greco, P. Carullo, A. Guffanti, V. Vigano, G. G. Stirparo, M. V. Latronico, G. Hasenfuss, J. Chen, and G. Condorelli. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. *Proc. Natl. Acad. Sci. U.S.A.*, 110(50):20164–20169, Dec 2013.

[34] A. Llucia-Valldeperas, B. Sanchez, C. Soler-Botija, C. Galvez-Monton, S. Roura, C. Prat-Vidal, I. Perea-Gil, J. Rosell-Ferrer, R. Bragos, and A. Bayes-Genis. Physiological conditioning by electric field stimulation promotes cardiomyogenic gene expression in human cardiomyocyte progenitor cells. *Stem Cell Res Ther*, 5(4):93, Aug 2014.

[35] M. Sjostrand, A. Johansson, L. Aqrawi, T. Olsson, M. Wahren-Herlenius, and A. Espinosa. The Expression of BAFF Is Controlled by IRF Transcription Factors. *J. Immunol.*, 196(1):91–96, Jan 2016.

[36] G. Yamada, M. Ogawa, K. Akagi, H. Miyamoto, N. Nakano, S. Itoh, J. Miyazaki, S. Nishikawa, K. Yamamura, and T. Taniguchi. Specific depletion of the B-cell population induced by aberrant expression of human interferon regulatory factor 1 gene in transgenic mice. *Proc. Natl. Acad. Sci. U.S.A.*, 88(2):532–536, Jan 1991.

[37] V. S. Peche, T. A. Holak, B. D. Burgute, K. Kosmas, S. P. Kale, F. T. Wunderlich, F. Elhamine, R. Stehle, G. Pfitzer, K. Nohroudi, K. Addicks, F. Stockigt, J. W. Schrickel, J. Gallinger, M. Schleicher, and A. A. Noegel. Ablation of cyclase-associated protein 2 (CAP2) leads to cardiomyopathy. *Cell. Mol. Life Sci.*, 70(3):527–543, Feb 2013.

[38] J. Nichols, B. Zevnik, K. Anastassiadis, H. Niwa, D. Klewe-Nebenius, I. Chambers, H. Scholer, and A. Smith. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, 95(3):379–391, Oct 1998.

[39] H. Niwa, J. Miyazaki, and A. G. Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, 24(4):372–376, Apr 2000.

[40] S. C. Degner, T. P. Wong, G. Jankevicius, and A. J. Feeney. Cutting edge: developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development. *J. Immunol.*, 182(1):44–48, Jan 2009.

[41] S. C. Degner, J. Verma-Gaur, T. P. Wong, C. Bossen, G. M. Iverson, A. Torkamani, C. Vettermann, Y. C. Lin, Z. Ju, D. Schulz, C. S. Murre, B. K. Birshtein, N. J. Schork, M. S. Schlissel, R. Riblet, C. Murre, and A. J. Feeney. CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proc. Natl. Acad. Sci. U.S.A.*, 108(23):9566–9571, Jun 2011.

[42] A. Benhaddou, C. Keime, T. Ye, A. Morlon, I. Michel, B. Jost, G. Mengus, and I. Davidson. Transcription factor TEAD4 regulates expression of myogenin and the unfolded protein response genes during C2C12 cell differentiation. *Cell Death Differ.*, 19(2):220–231, Feb 2012.

[43] Y. Ohkubo, M. Arima, E. Arguni, S. Okada, K. Yamashita, S. Asari, S. Obata, A. Sakamoto, M. Hatano, J. O-Wang, M. Ebara, H. Saisho, and T. Tokuhisa. A role for c-fos/activator protein 1 in B lymphocyte terminal differentiation. *J. Immunol.*, 174(12):7703–7710, Jun 2005.

[44] A. A. Avilion, S. K. Nicolis, L. H. Pevny, L. Perez, N. Vivian, and R. Lovell-Badge. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.*, 17(1):126–140, Jan 2003.

[45] C. N. Arnold, E. Pirie, P. Dosenovic, G. M. McInerney, Y. Xia, N. Wang, X. Li, O. M. Siggs, G. B. Karlsson Hedestam, and B. Beutler. A forward genetic screen reveals roles for Nfkbid, Zeb1, and Ruvbl2 in humoral immunity. *Proc. Natl. Acad. Sci. U.S.A.*, 109(31):12286–12293, Jul 2012.

[46] S. K. Balakrishnan, M. Witcher, T. W. Berggren, and B. M. Emerson. Functional and molecular characterization of the role of CTCF in human embryonic stem cell biology. *PLoS ONE*, 7(8):e42424, 2012.

[47] C. Mauritz, K. Schwanke, M. Reppel, S. Neef, K. Katsirntaki, L. S. Maier, F. Nguemo, S. Menke, M. Haustein, J. Hescheler, G. Hasenfuss, and U. Martin. Generation of functional murine cardiac myocytes from induced pluripotent stem cells. *Circulation*, 118(5):507–517, Jul 2008.

[48] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589, May 2010.

[49] Y. C. Lin, S. Jhunjhunwala, C. Benner, S. Heinz, E. Welinder, R. Mansson, M. Sigvardsson, J. Hagman, C. A. Espinoza, J. Dutkowski, T. Ideker, C. K. Glass, and C. Murre. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.*, 11(7):635–643, Jul 2010.

[50] A. Bondue and C. Blanpain. Mesp1: a key regulator of cardiovascular lineage commitment. *Circ. Res.*, 107(12):1414–1427, Dec 2010.

[51] U. Wellner, J. Schubert, U. C. Burk, O. Schmalhofer, F. Zhu, A. Sonntag, B. Waldvogel, C. Vannier, D. Darling, A. zur Hausen, V. G. Brunton, J. Morton, O. Sansom, J. Schuler, M. P. Stemmler, C. Herzberger, U. Hopt, T. Keck, S. Brabletz, and T. Brabletz. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat. Cell Biol.*, 11(12):1487–1495, Dec 2009.

Supplementary Figures

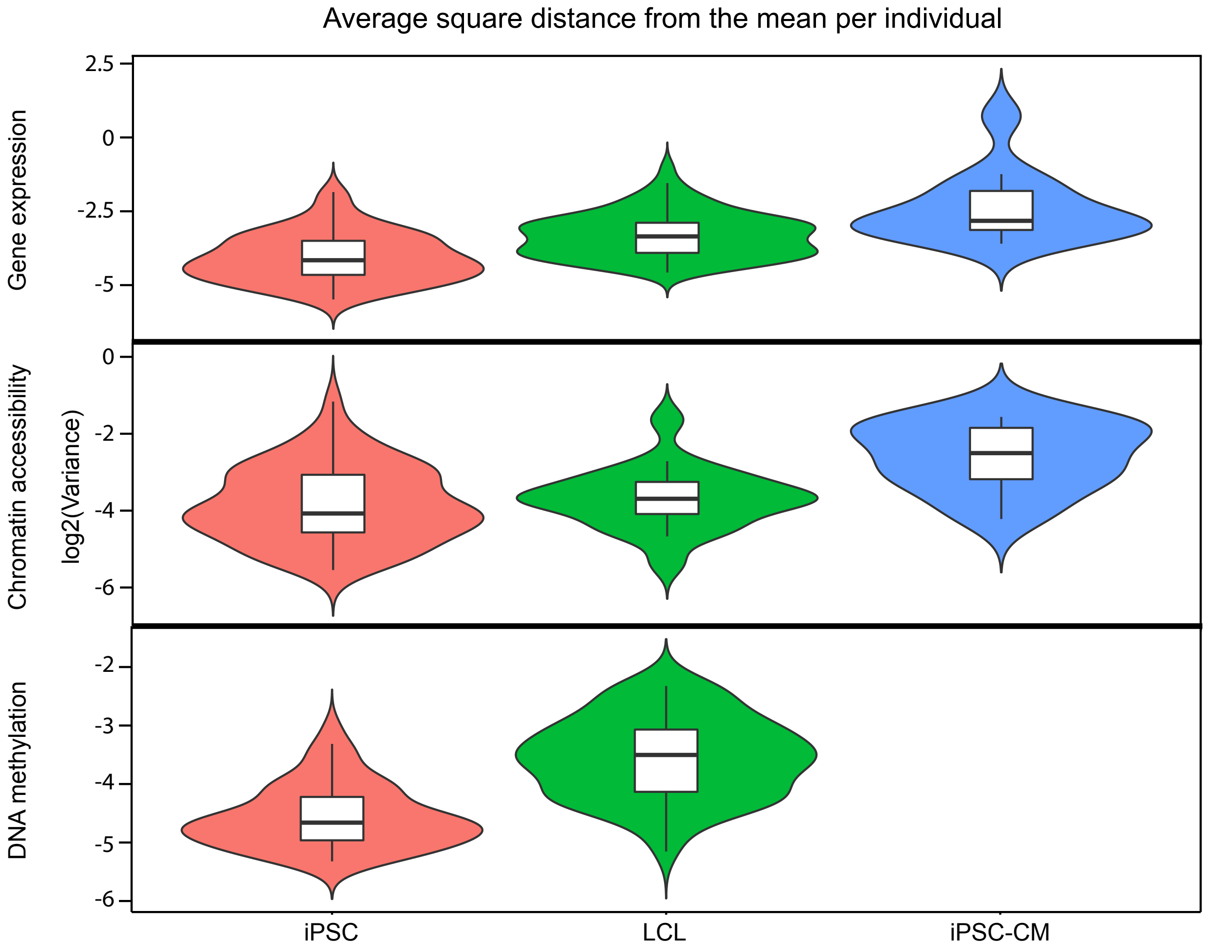

Figure 1: Violin plots representing per individual $\log_2$ of the average square distance from the mean for iPSC, LCL, and iPSC-CM gene expression, chromatin accessibility, and DNA methylation levels (methylation not assayed in iPSC-CMs).

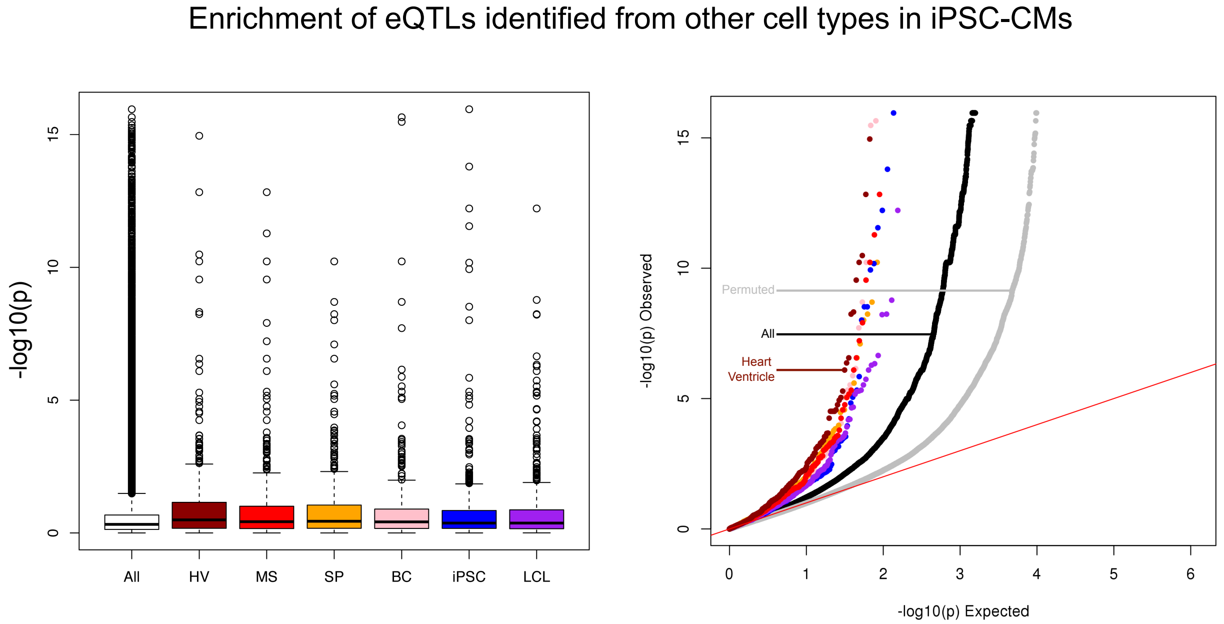

Figure 2: Boxplots of the $-\log_{10}$ p-value from eQTLs identified in iPSC-CMs conditioned on being previously identified as eQTLs in tissues from the GTEx consortium as well as iPSCs and LCLs. Specifically, the GTEx tissues heart left ventricle (HV), skeletal muscle (MS), spleen (SP), and the cortex of the brain (BC) were used. eQTLs identified in heart left ventricle show the greatest enrichment of associations with expression in iPSC-CMs.

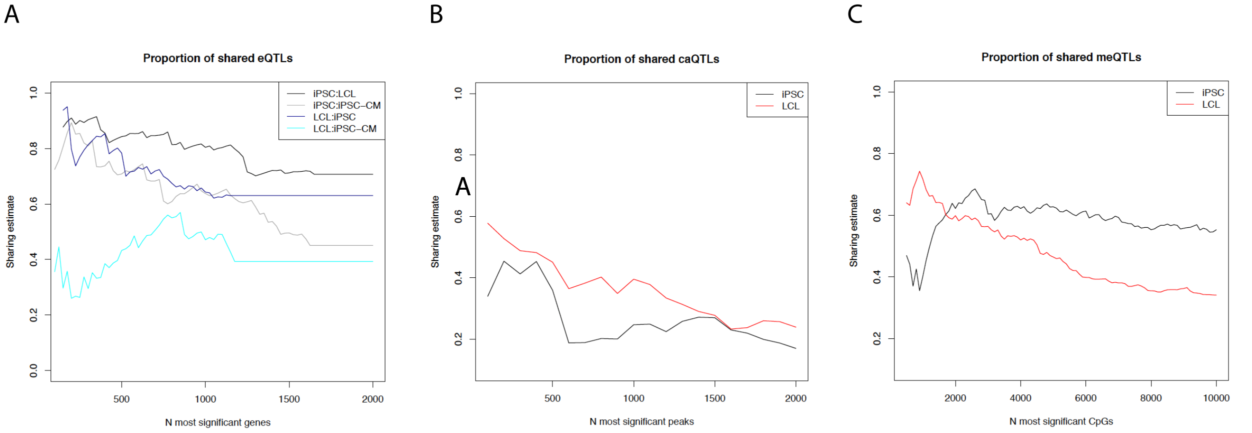

Figure 3: π_1 **estimates of sharing** of expression QTLs (A), chromatin accessibility QTLs (B), and DNA methylation QTLs (B) estimated across the top N most significant genes, peaks, and probes, respectively.

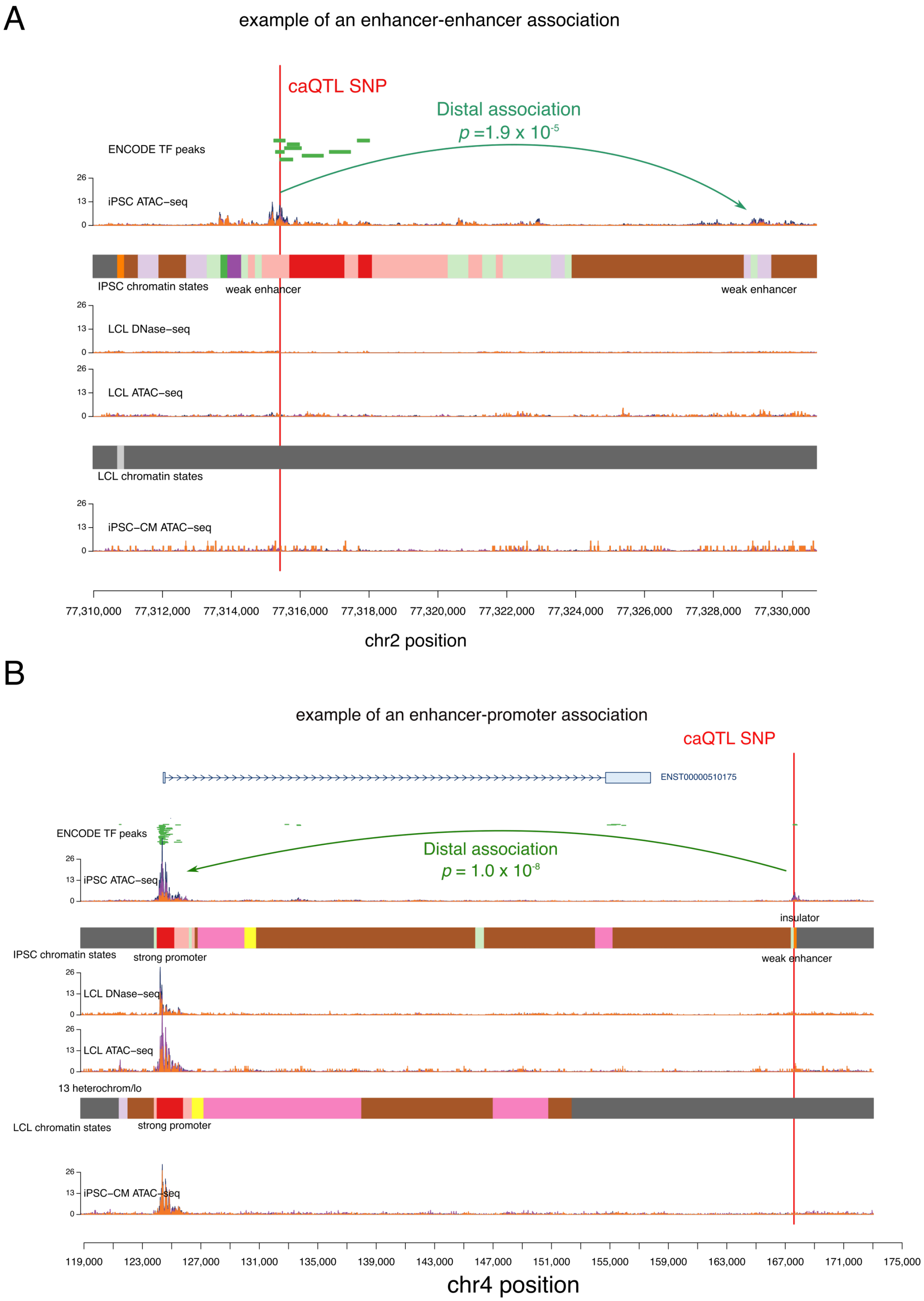

Figure 4: **Examples of distal caQTLs affecting a chromatin element interactions. (A)** Example of an iPSC-specific caQTL SNP located in a weak enhancer in iPSC and in LCL heterochromatin. This SNP affects accessibility at a distal enhancer element in iPSCs, but not in LCLs. **(B)** Example of another iPSC-specific caQTL SNP located in an iPSC enhancer and in LCL heterochromatin. This SNP has an iPSC-specific distal effect at a promoter that is active in all three cell types (iPSC, LCLs, and iPSC-CM)

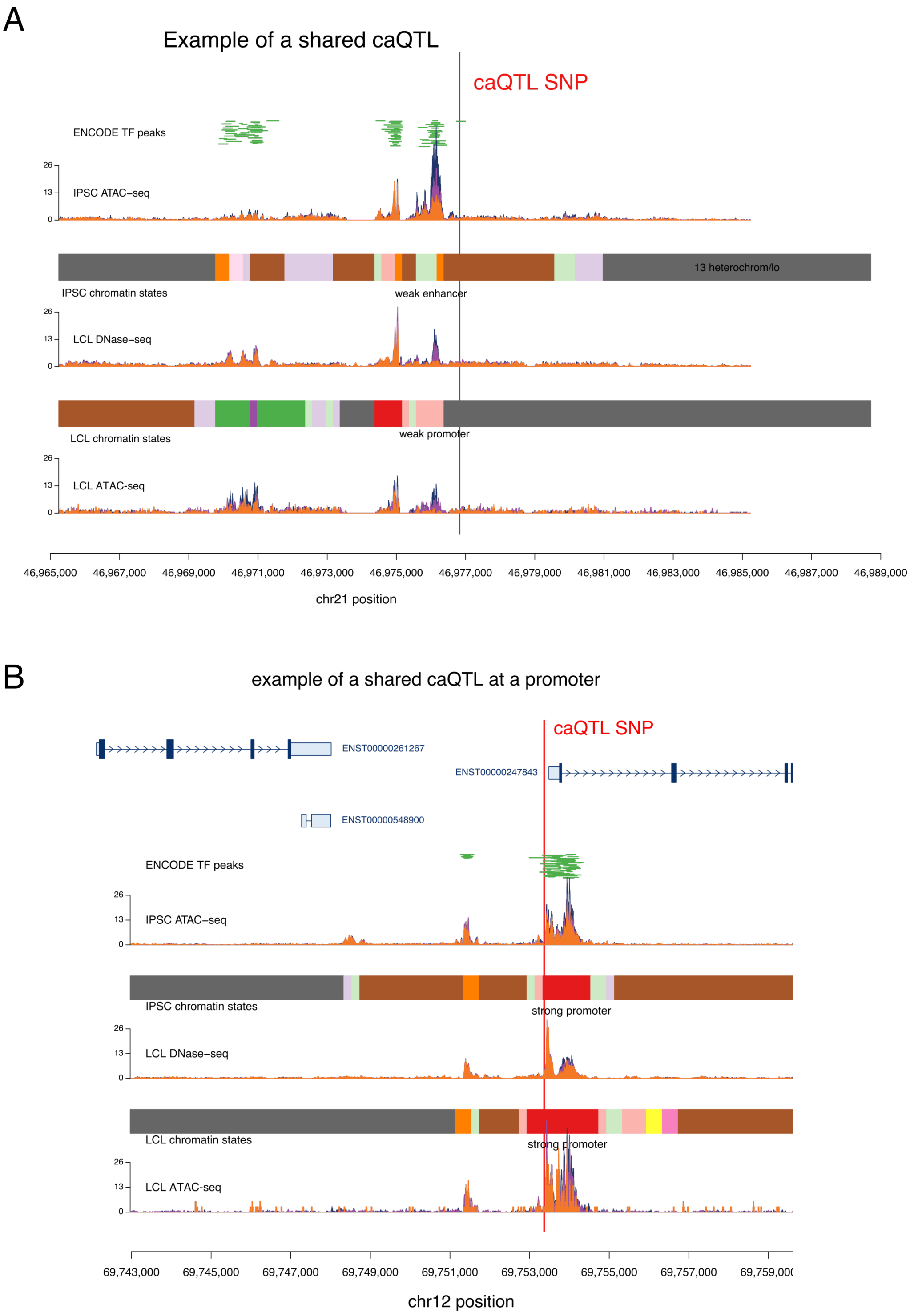

Figure 5: **Examples of shared caQTLs in LCLs and IPSCs.** **(A)** Example of a shared caQTL in an intergenic region. The associated caQTL SNP is located near a shared accessibility peak, and overlaps a predicted transcription factor binding site from ENCODE. The QTL effect is seen in both LCL ATAC-seq and DNase-seq data. **(B)** Example of a shared caQTL in a strong promoter region as determined by chromHMM predictions.

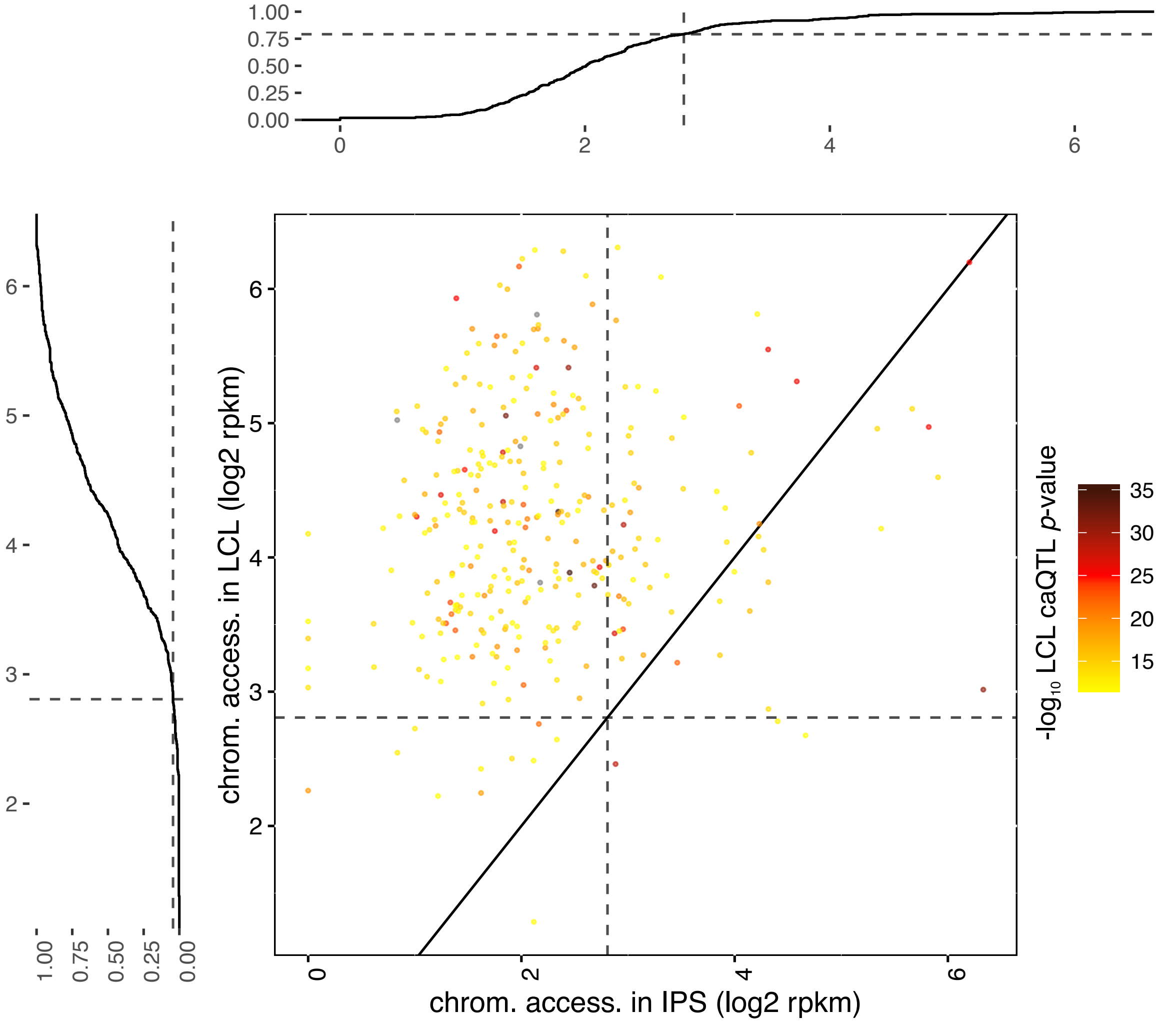

Figure 6: **LCL-specific caQTLs chromatin accessibility signal in LCLs and iPSCs.** Cumulative distributions of LCL-specific caQTL chromatin accessibility in LCL and iPSC show that most LCL-specific caQTLs (¿90%) are located in accessible regions in LCLs but only about 20% are located in iPSC-accessible regions.

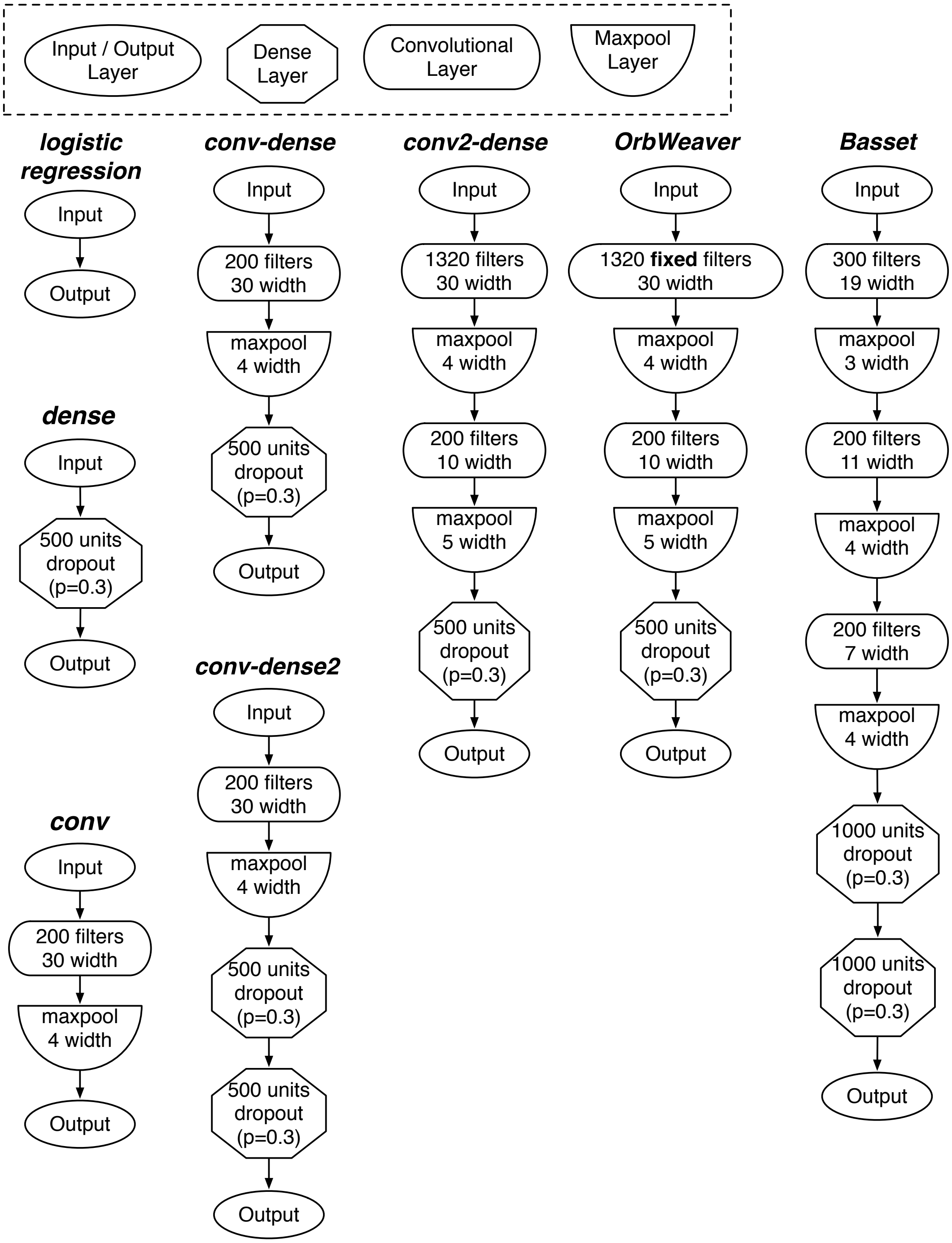

Figure 7: **Deep neural network architectures for predicting chromatin activity.** We illustrate deep neural network models of increasing complexity, including Fangtooth and Basset, each of which are built by layering simple functions. Each model lists the number of neurons (variables) in each layer, and the activation function used in all the dense and convolutional layers is the Rectified Linear Unit (ReLU). The first layer of filters in Fangtooth were fixed to log-transformed position weight matrices (see supplementary text). The Basset model was re-implemented using Theano and the Keras framework, and the batch-normalization steps were eliminated since they provided no increase in model accuracy for our problem at the cost of a large increase in model learning time.

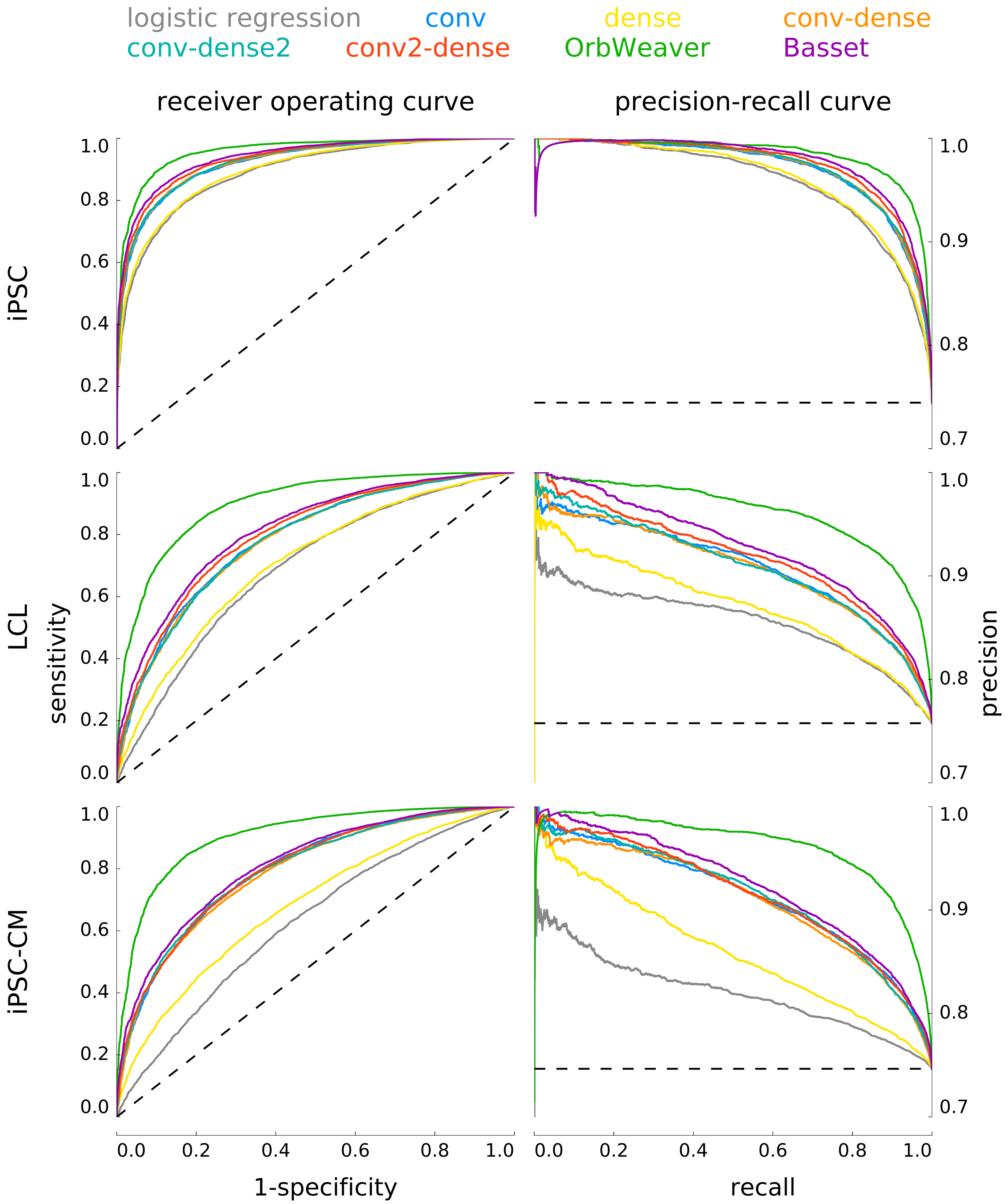

Figure 8: **Comparing the accuracy of neural network architectures.** The accuracies of each of the neural network models are compared in each cell type; the accuracies are represented using both the Receiver Operating Curve and the Precision-Recall curve. Notably, the gain in accuracy with each additional convolutional and dense layer was much smaller than the gain achieved by fixing the first convolutional filters to known position weight matrices of human transcription factors. In particular, Fangtooth, a simpler neural network with known PWMs achieved a substantially higher accuracy than Basset, a more complex neural network.

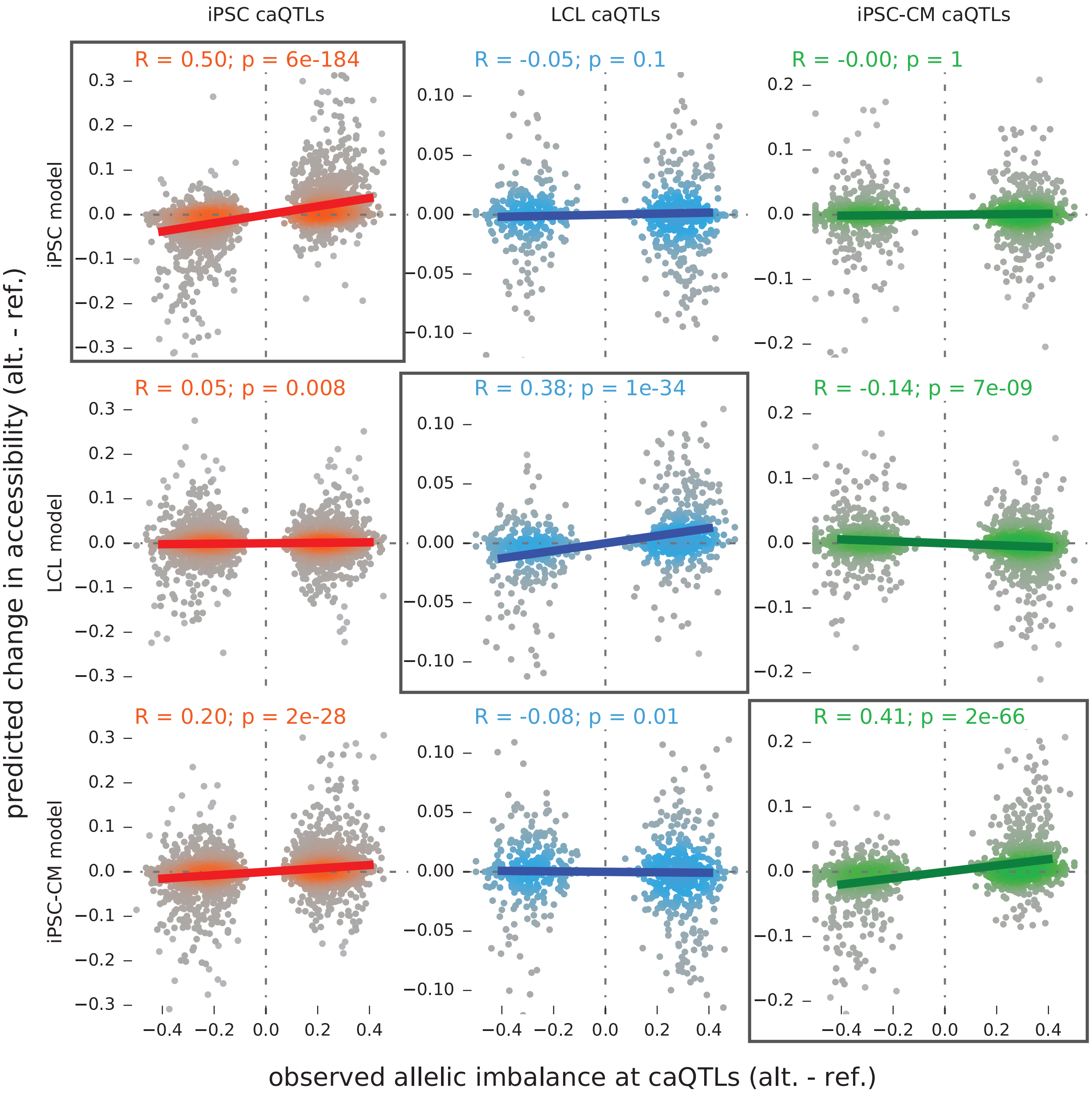

Figure 9: **Predicting genetic effects on chromatin activity using neural network models.** Observed allelic imbalance in ATAC-seq reads at caQTLs (estimated using WASP; x-axis) is compared with the difference in predicted chromatin activity between haplotypes tagged by the two alleles of the caQTL SNP (y-axis); both quantities are computed focusing on the causal caQTL SNP identified using qtlBHM, a Bayesian hierarchical model. Each column of panels corresponds to observed genetic effects in each of the three cell types (each point is a locus identified by qtlBHM as a caQTL with posterior greater than 0.99 in that cell type), and each row of panels corresponds to the genetic effects predicted by the Fangtooth model relevant to each of the three cell types. Notably, in each cell type, predictions of genetic effect using a model for a different cell type has very low or no agreement with the observed allelic imbalance in chromatin activity (off-diagonal panels).

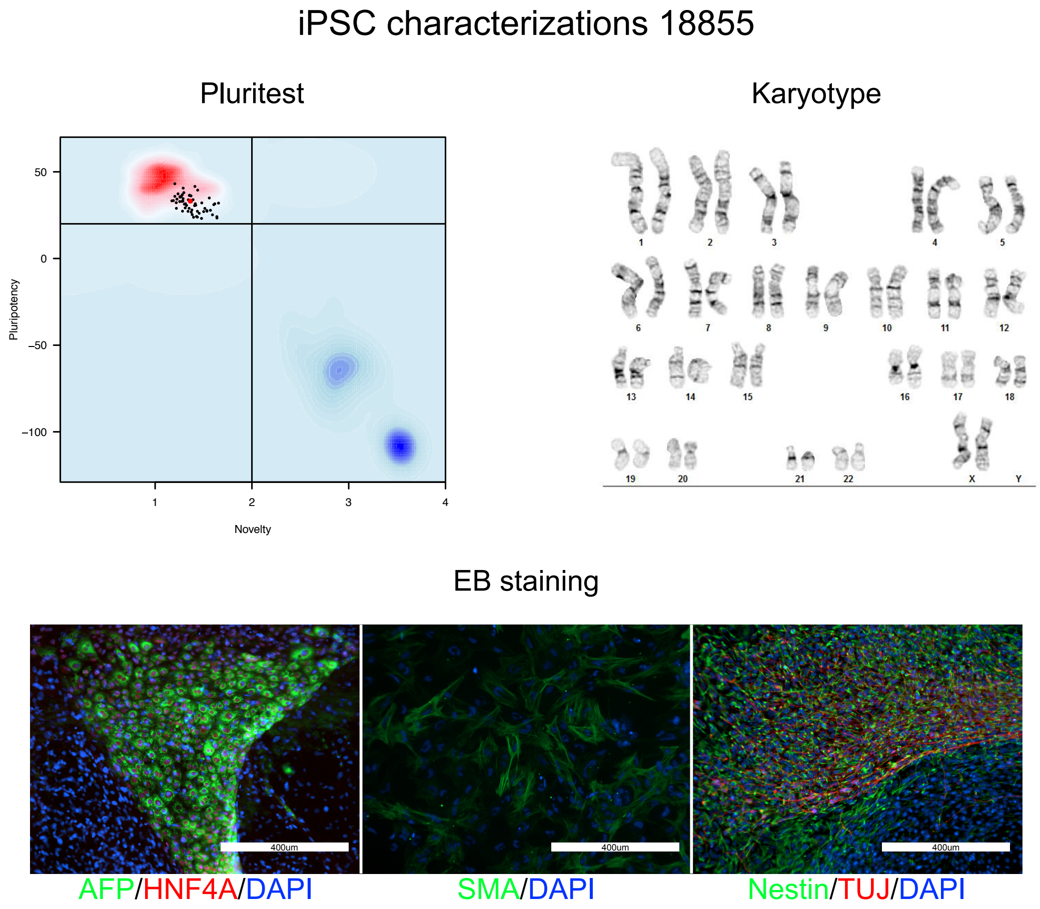

Figure 10: Pluritest results were plotted for each individual. The red dot corresponds to the NA18855 sample. Black lines represent the empirical quality thresholds. Karyotype results are shown for NA18855. Immunostaining performed on spontaneously differentiated tissue derived from embryoid bodies (EBs) confirms cells types from all three germ layers - endoderm (AFP/HNF4A), mesoderm (SMA), and ectoderm (Nestin/TUJ/Map2). QC for other lines can be found at `http://eqtl.uchicago.edu/yri_ipsc/iPSC_QC.pdf`.

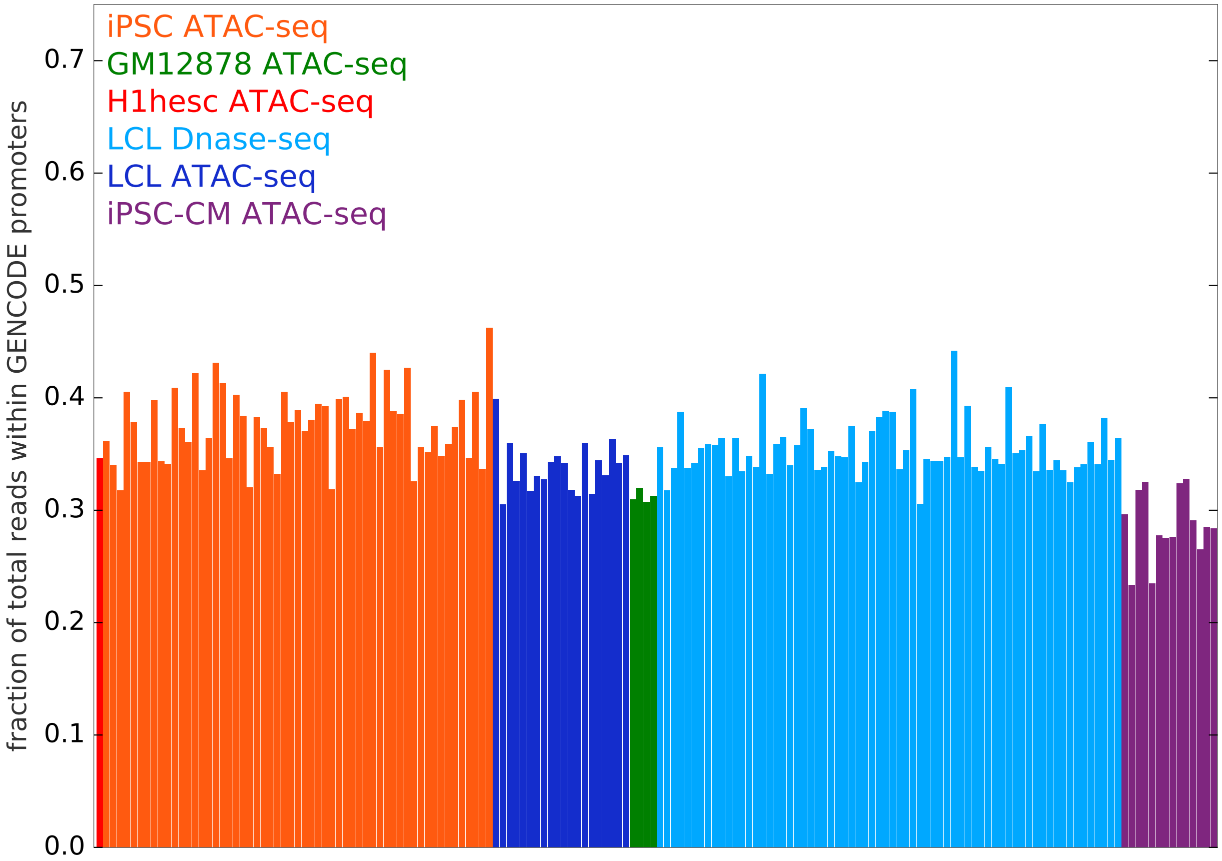

Figure 11: **Quantifying library quality using proportion of reads in promoters.** The proportion of ATAC-seq reads within promoters in YRI LCLs is comparable to the proportion of ATAC-seq reads in GM12878 [10] and the proportion of DNase-seq reads within promoters in the same group of YRI LCLs [17]. The proportion of ATAC-seq reads for iPSCs is also comparable to the proportion of ATAC-seq reads for H1hesc (courtesy of the Greenleaf Lab). The proportion of ATAC-seq reads for iPSC-CMs is slightly lower compared to other cell types indicating that a larger fraction of the transpositions in these libraries occur within distal enhancers.

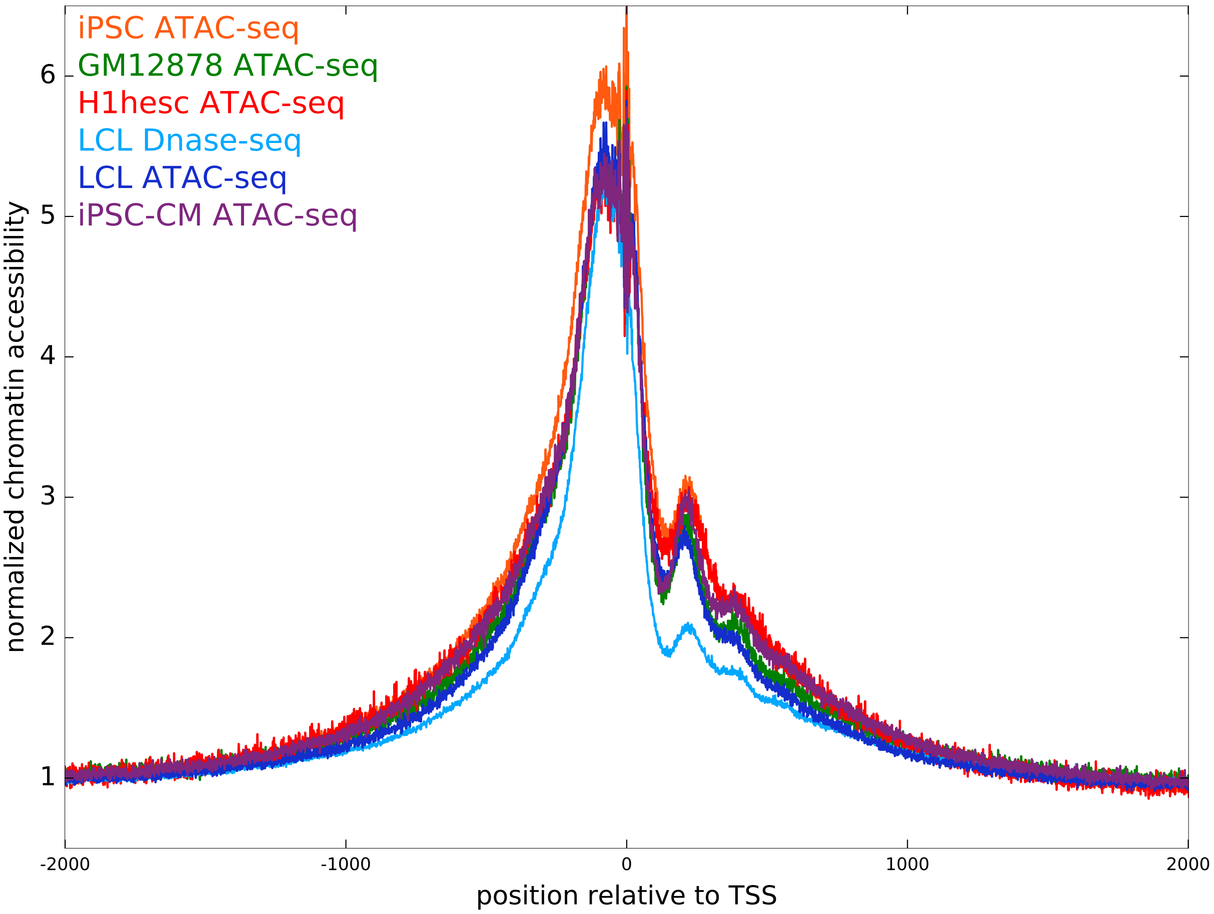

Figure 12: **Quantifying library quality using chromatin accessibility around TSS.** Chromatin accessibility in each cell type is computed as the count of Dnase nicks or ATAC transpositions centered at the transcription start site, normalized by the accessibility signal 2kb upstream of the TSS, mean-averaged across all GENCODE genes and median-averaged across multiple individuals or libraries. The accessibility profile across all ATAC-seq libraries are highly similar to each other and slightly broader than the accessibility profile of DNase-seq libraries.

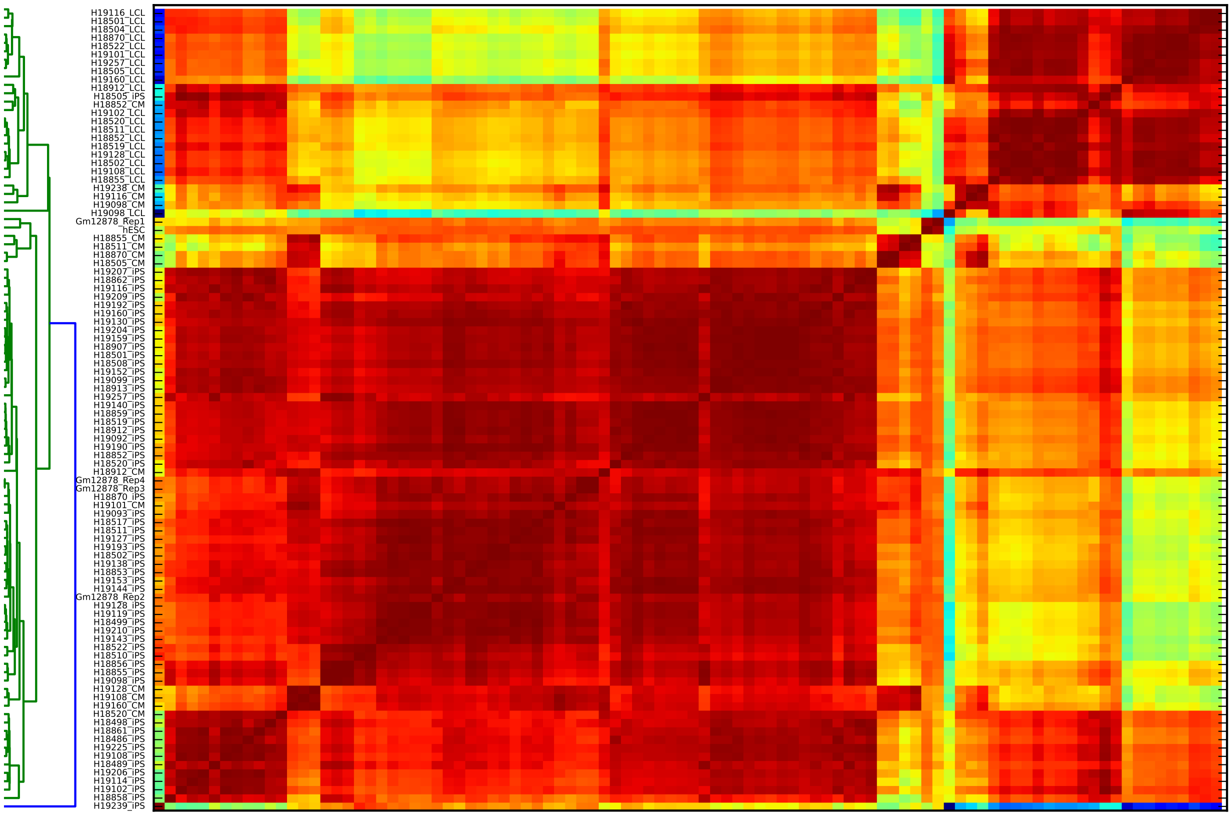

Figure 13: **Comparing the fragment-size distributions between ATAC-seq libraries.** A heatmap illustrating the similarity in the fragment-size distributions between several ATAC-seq libraries of LCLs, iPSCs, and iPSC-CMs, along with previously published data for LCL [10] and H1hesc (courtesy of the Greenleaf Lab). In general, libraries for each cell type cluster together, with the exception of a few iPSC-CM libraries. All iPSC and iPSC-CM libraries had very similar characteristics, with a substantial fraction of fragments spanning one or two nucleosomes. The LCL libraries were subtly different, with a much smaller fraction of fragments spanning nucleosomes.

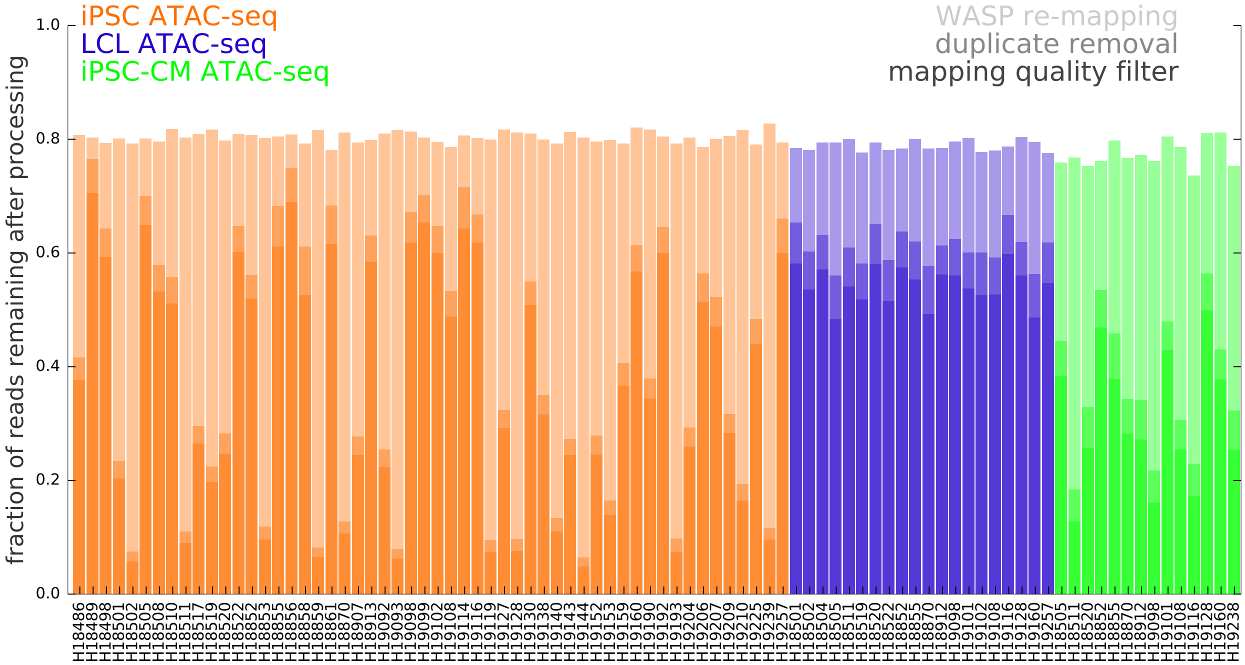

Figure 14: **Quantifying ATAC-seq read loss at various data processing steps.** At each step of processing the ATAC-seq data, the fraction of nuclear DNA reads filtered out varied across cell types and individuals. The fraction of reads filtered out by WASP remapping was relatively stable at ≈ 20%, as was the fraction removed using a mapping quality filter of 10 (≈ 5%). The fraction of reads removed due to filtering out duplicate fragments varied substantially across libraries.

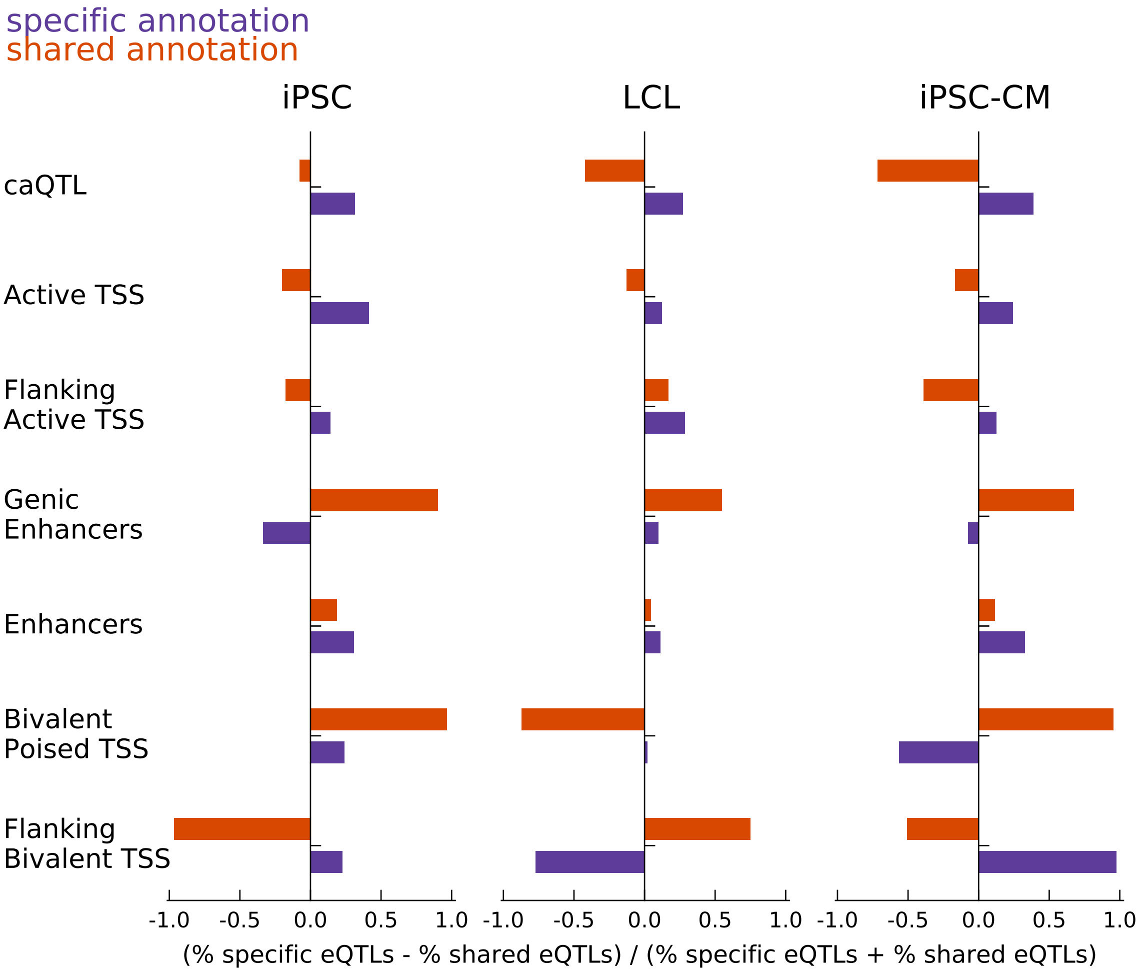

Figure 15: **Quantifying eQTLs explained by chromatin annotations.** For a set of genomic annotations that are specific to a cell type or shared between two or more cell types, we illustrate the difference in the proportion of shared eQTLs and the proportion of cell-specific eQTLs explained by each of the annotations. Importantly, we find that shared caQTLs and shared active TSS explain a larger fraction of shared eQTLs than cell-specific eQTLs. Enhancers, in general, explain a larger fraction of cell-specific eQTLs than shared eQTLs. In contrast, shared enhancers within genes explain a larger fraction of cell-specific eQTLs than shared eQTLs.