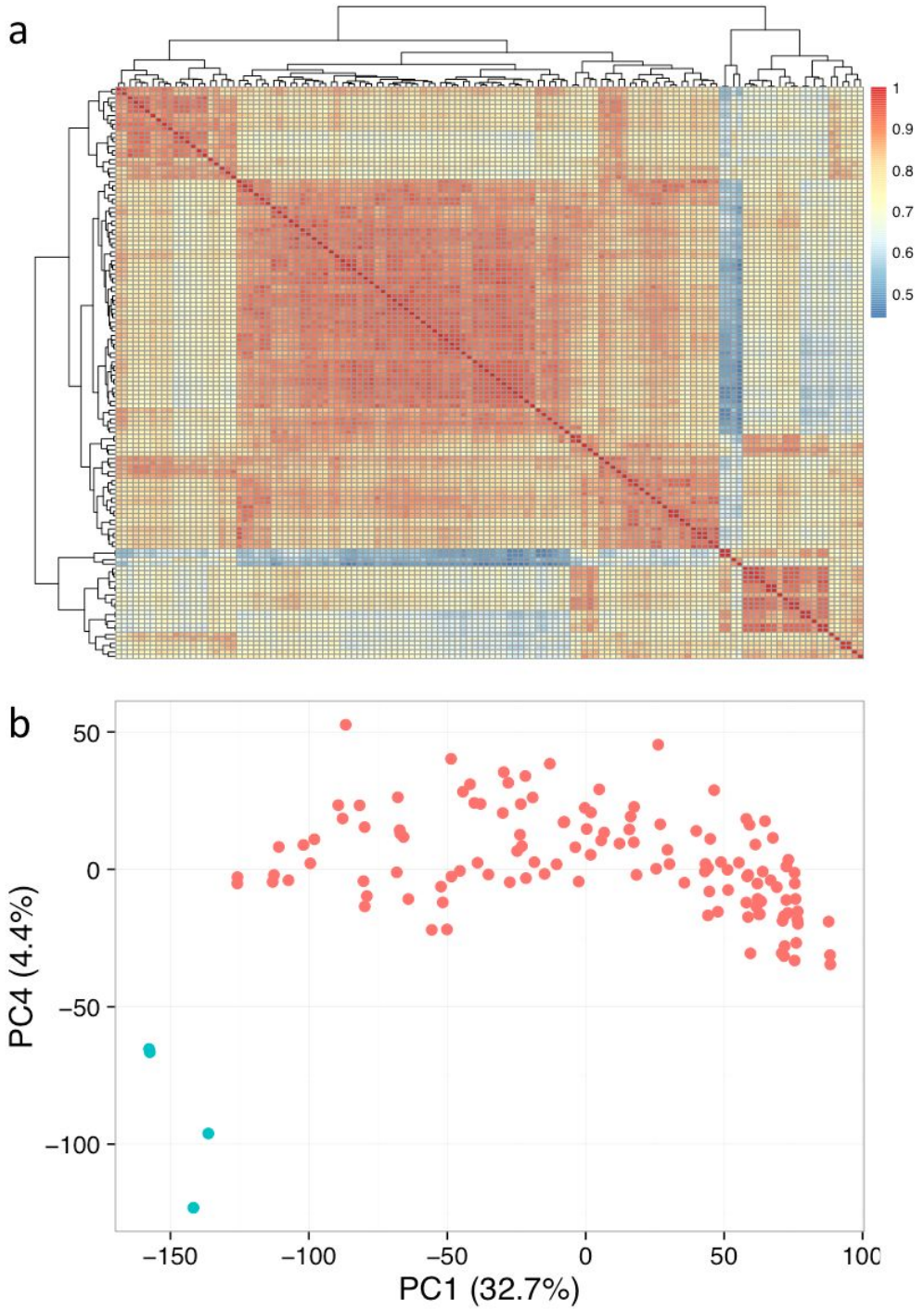
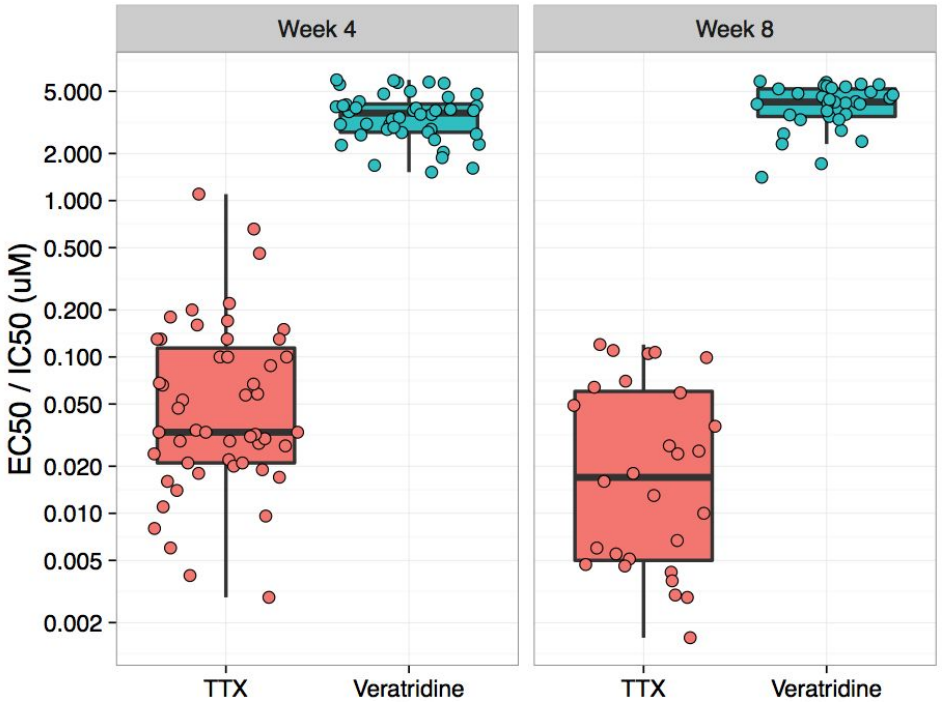
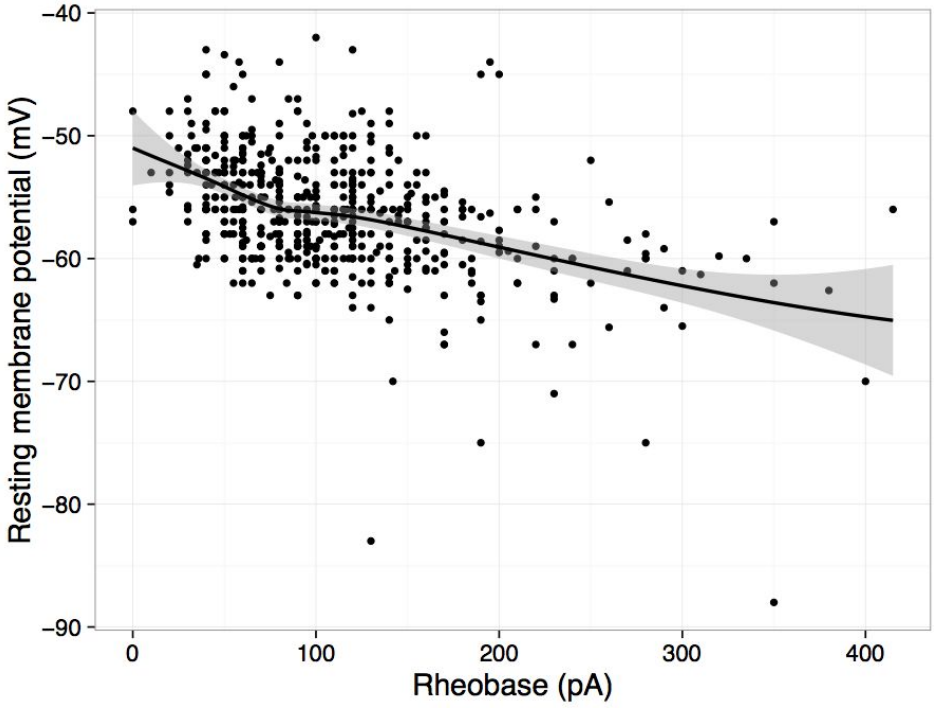
Supplementary Figures



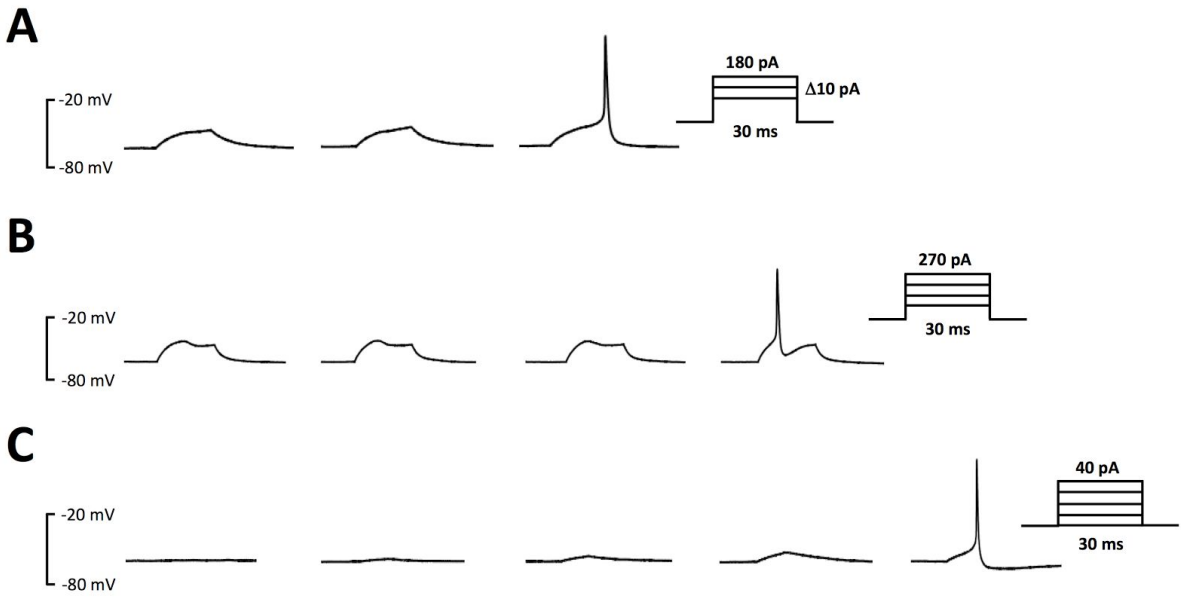
Supplementary Figure 1: Four RNA-seq sample outliers (cell lines koun_2, iakz_1, yuze_1_1, yuze_1_2) seen in (**a**) heatmap of sample-sample correlations and in (**b**) principal components plot (blue dots). The samples are outliers in the heatmap and in PC4, and moderate outliers in PC1.



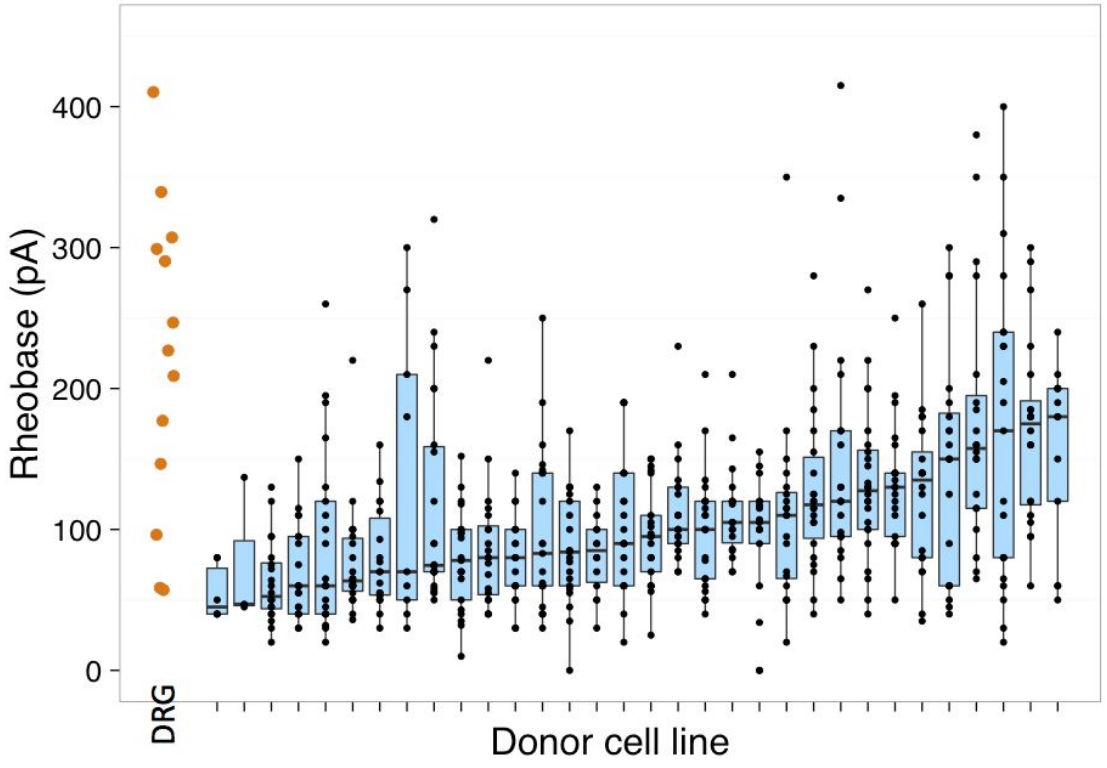
Supplementary Figure 2: Ca^{2+} response of IPSDSN cell cultures to veratridine and tetrodotoxin (TTX).



Supplementary Figure 3: Rheobase correlates with resting membrane potential in individual cells ($r = -0.44$, $p < 2x10^{-16}$). Points represent the rheobase and resting membrane potential measurements recorded by patch clamp from 615 individual cells across 31 donors.

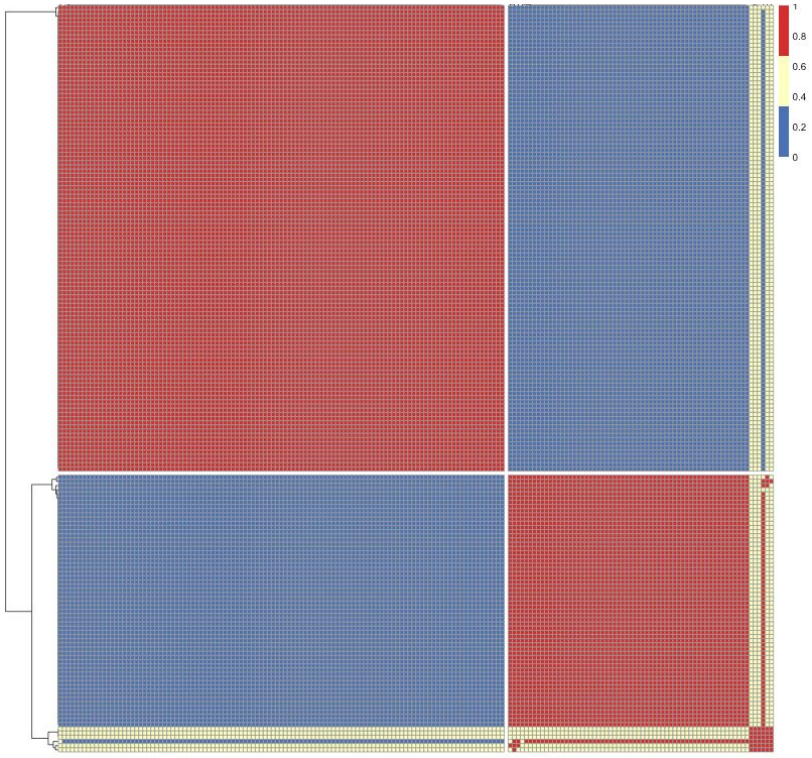


Supplementary Figure 4: Recordings from three different cells showing different rheobase thresholds: **A**) medium-high (180 pA), **B**) very high (270 pA), and **C**) low (40 pA). Rows show 2-4 current injection steps previous to the one that produced the action potential.



Supplementary Figure 5: The distribution of rheobase values for the 31 samples with electrophysiology recordings, as well as literature values for DRG (leftmost bar).

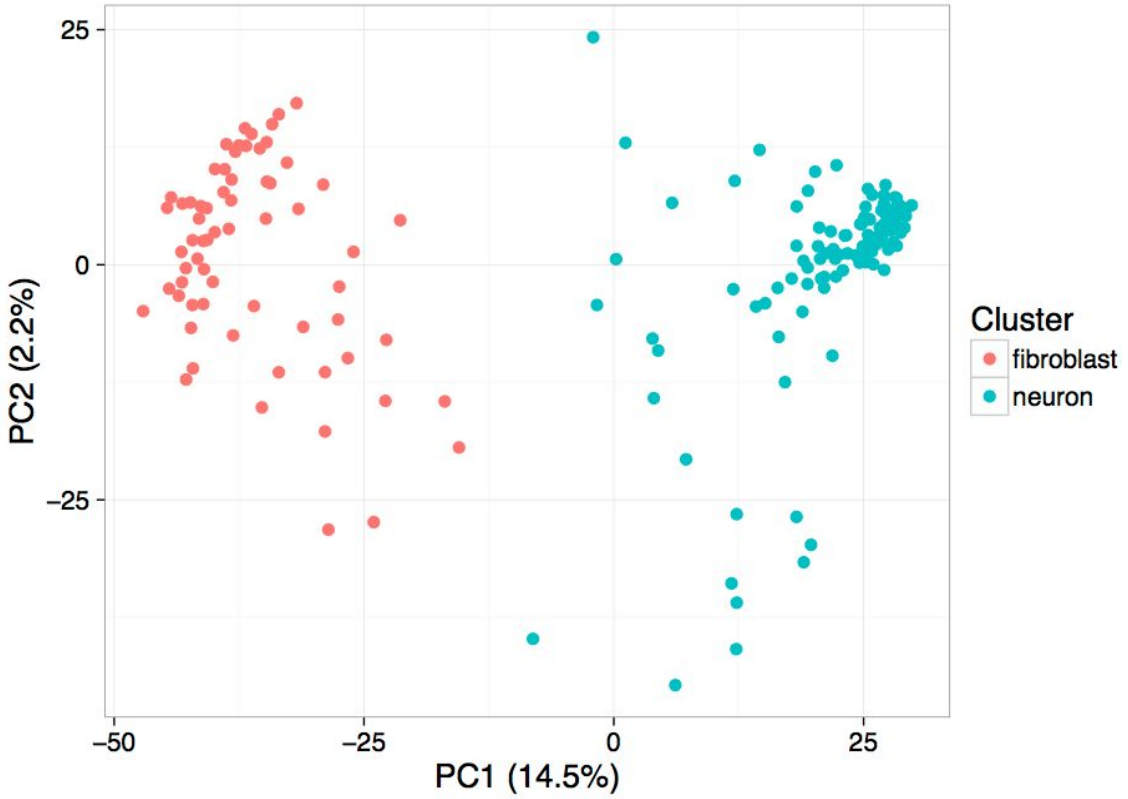
A Consensus matrix k = 2



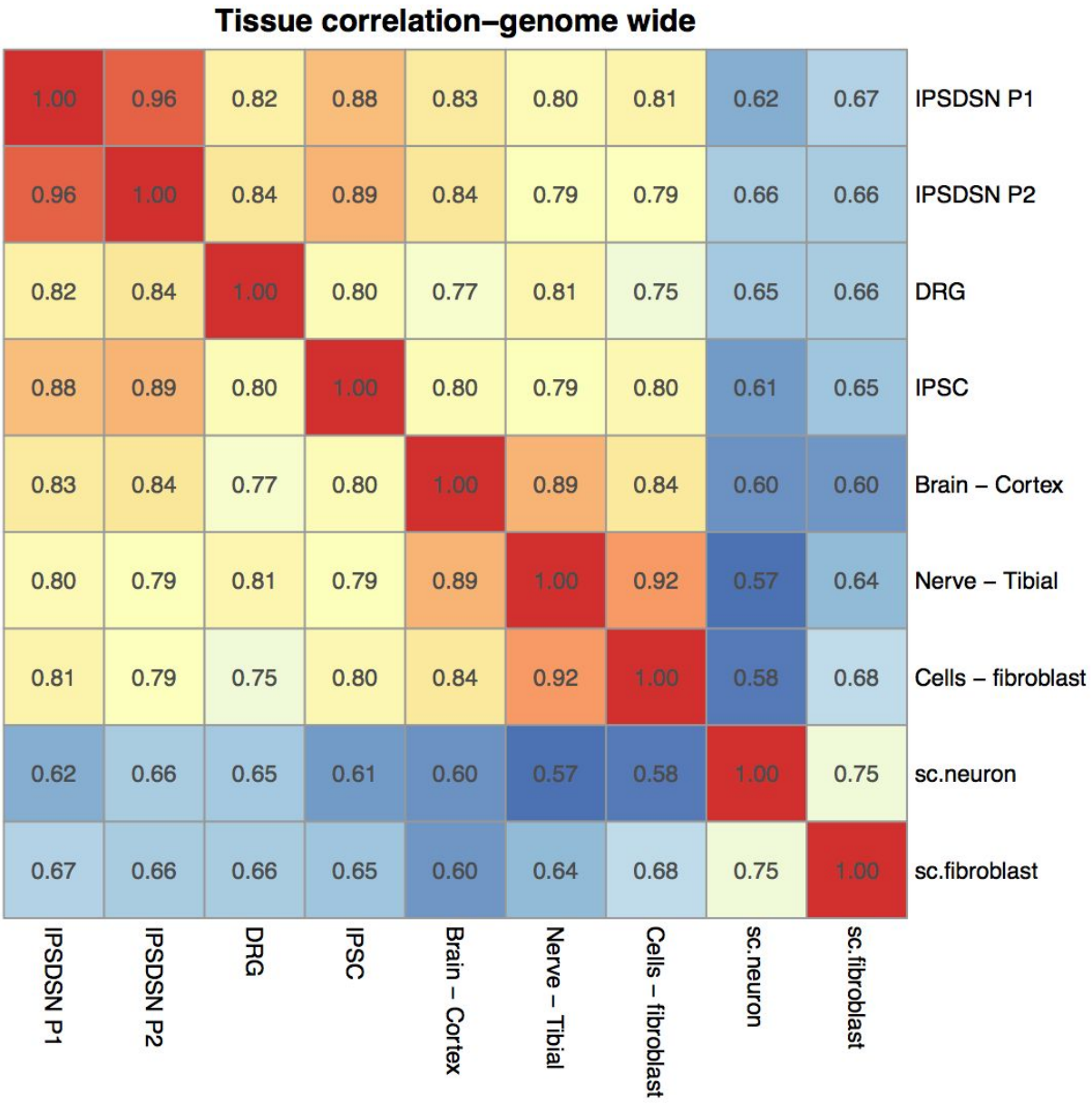
B Clustering scores

	k=2	k=3	k=4	k=5
Cluster 1	1.00*	0.98*	0.95*	0.99*
Cluster 2	0.89	0.90	0.93	0.86
Cluster 3		0.56	0.59	0.57
Cluster 4			0.54*	0.82
Cluster 5				0.53*

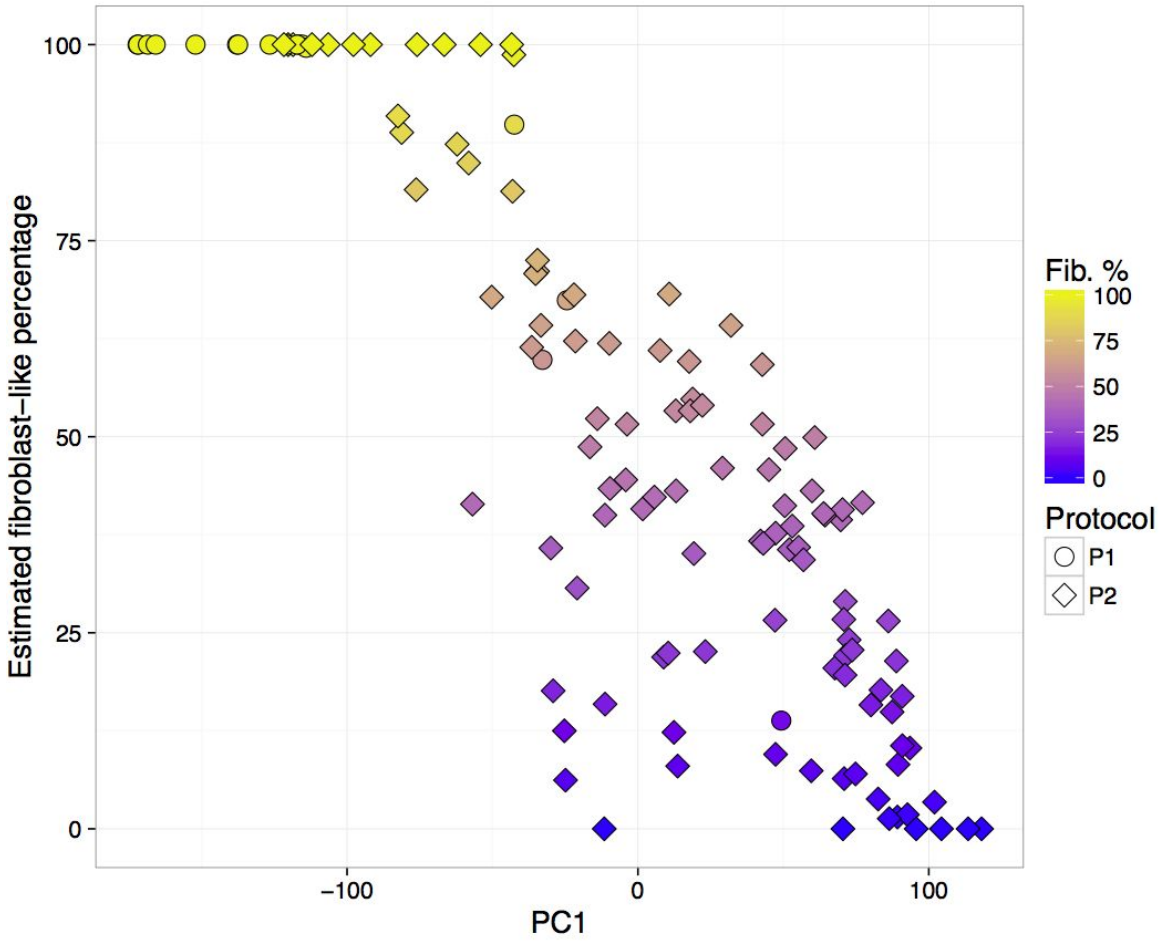
Supplementary Figure 6: SC3 clustering of single IPSDSN cells. We ran SC3 with gene expression counts from 177 single cells, specifying either k=2, 3, 4, or 5 clusters. Examining the consensus matrices (A) and clustering scores (B) for each of these cases suggests that the data are well described by 2 clusters of cells, as k=2 gives the highest average clustering score. For each value of k we looked at the differentially expressed genes for the clusters, as reported by SC3. These genes appeared to fall into two groups - neuronal genes (ion channels, neural developmental regulators) or fibroblast-like genes (collagens, extracellular matrix). Neuron-like clusters are marked by *.



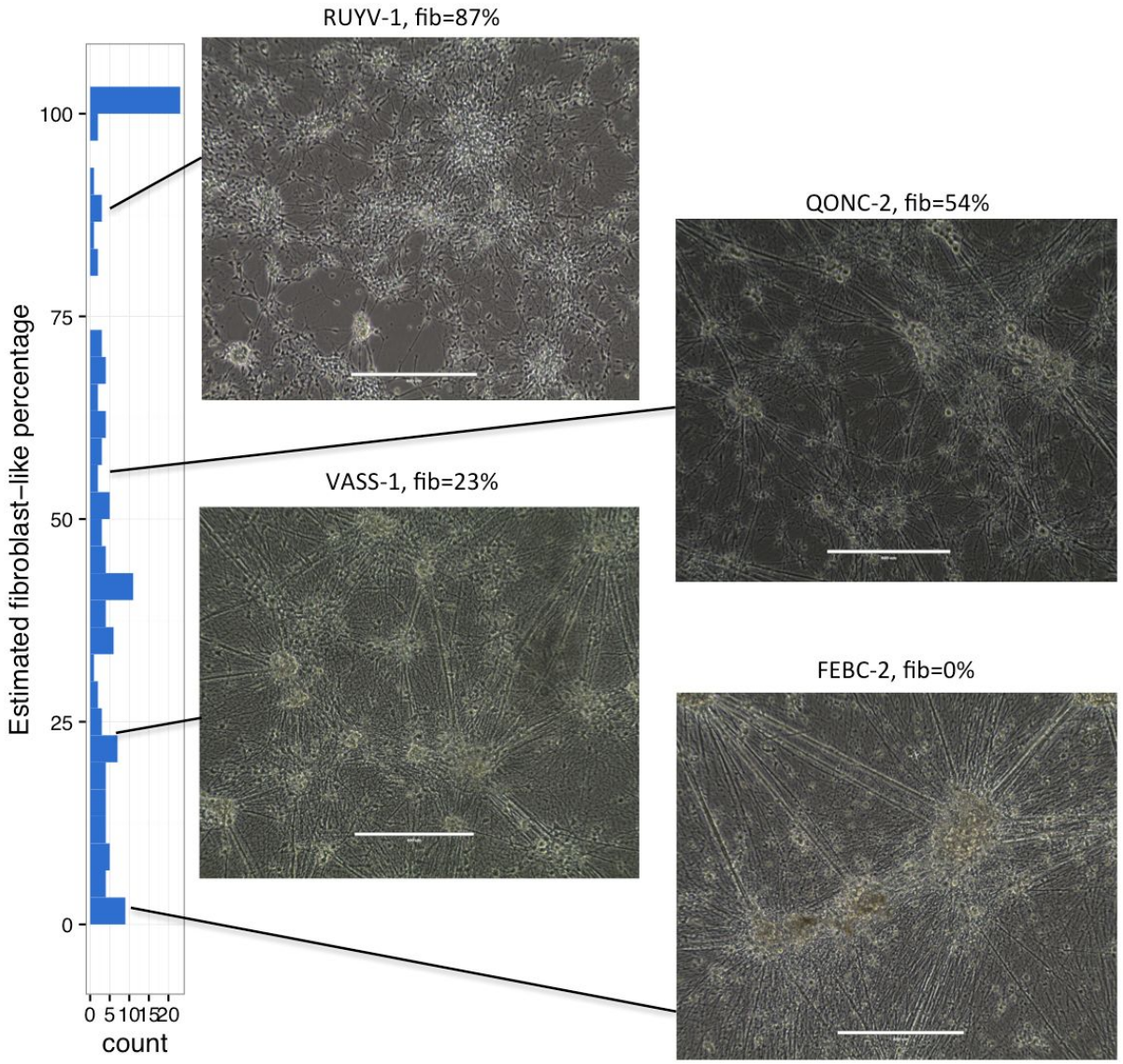
Supplementary Figure 7: Neuron and fibroblast-like IPSDSN cells form separate clusters in a PCA plot. Single cell counts were quantile-normal transformed per sample, and principal components were calculated and cells plotted showing principal components 1 and 2. Cells are colored based on their categorization by SC3, labeled by us as fibroblast or neuron.



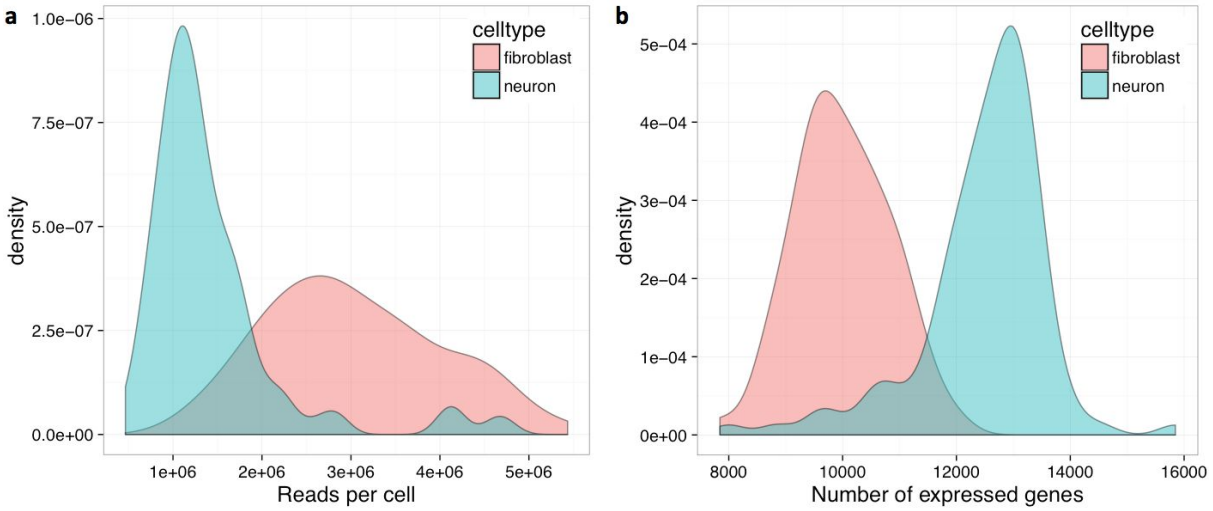
Supplementary Figure 8: Correlation of genome-wide gene expression in different tissues and cell cultures. Gene expression for IPSDSN single cells was averaged within a cluster (sc.neuron, sc.fibroblast). For each gene the mean expression (FPKM) in the group of samples was computed, and these values were correlated genome-wide across groups. Spearman correlation values are shown in each square. P1 and P2 protocol samples are highly similar to each other, and compare similarly to other tissues. Both single cell neurons and single cell fibroblast-like cells have similarity to DRG, although single fibroblast-like cells have greater similarity with GTEx fibroblasts.



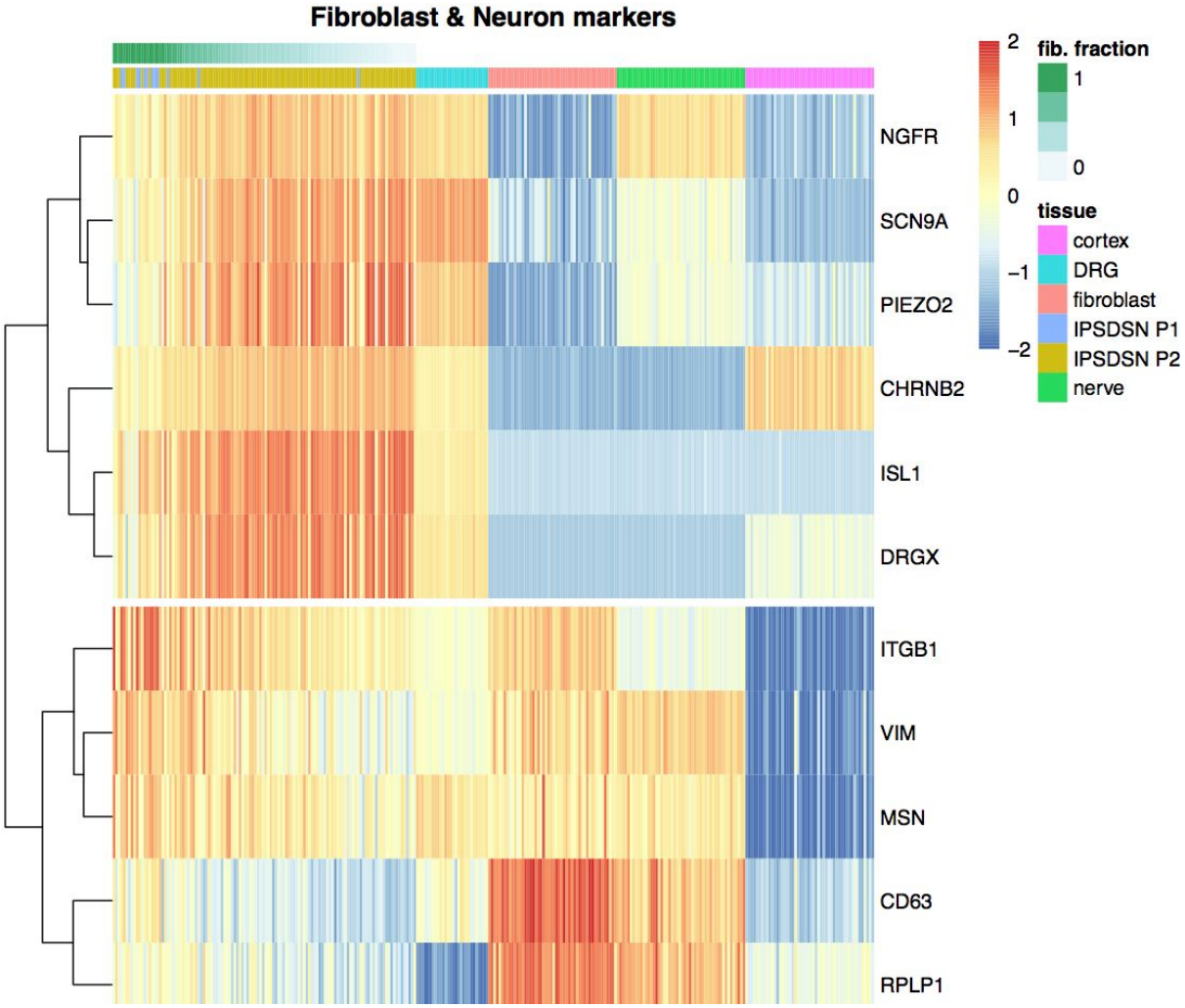
Supplementary Figure 9: Plot of estimated fibroblast percentage for bulk RNA-seq samples versus gene expression principal component 1, after excluding 5 outlier samples. Note that although many samples have estimated fibroblast percentage close to 100%, these samples visually contained a significant fraction of neuronal cells. The estimates represent the estimated fraction of RNA coming from fibroblast-like cells, not the number of cells.



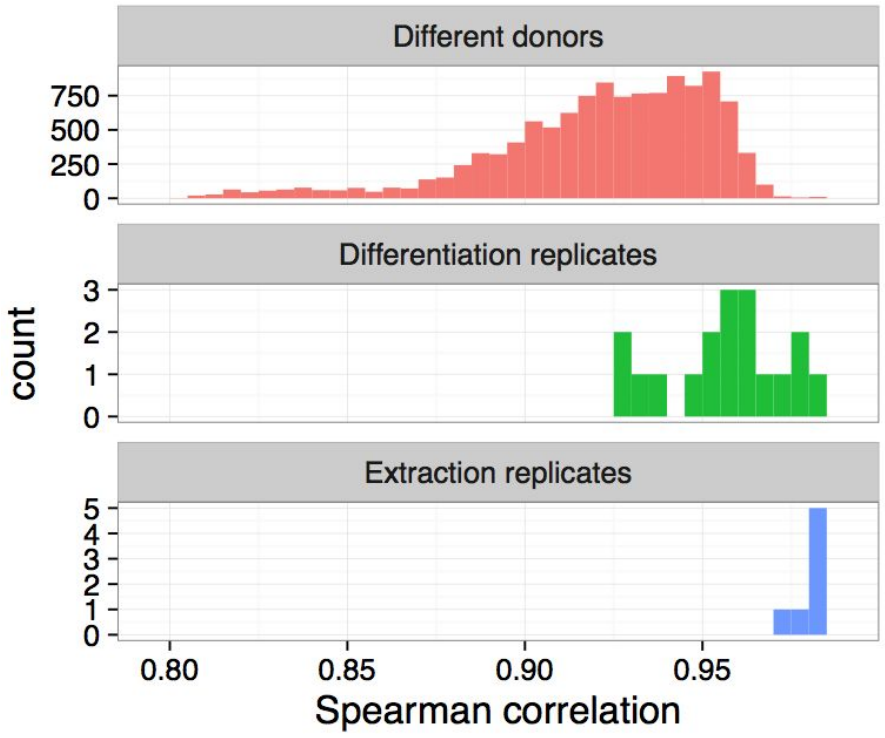
Supplementary Figure 10: Images of four cell lines, taken 4 weeks post induction of differentiation, with their estimated fibroblast-like content shown.



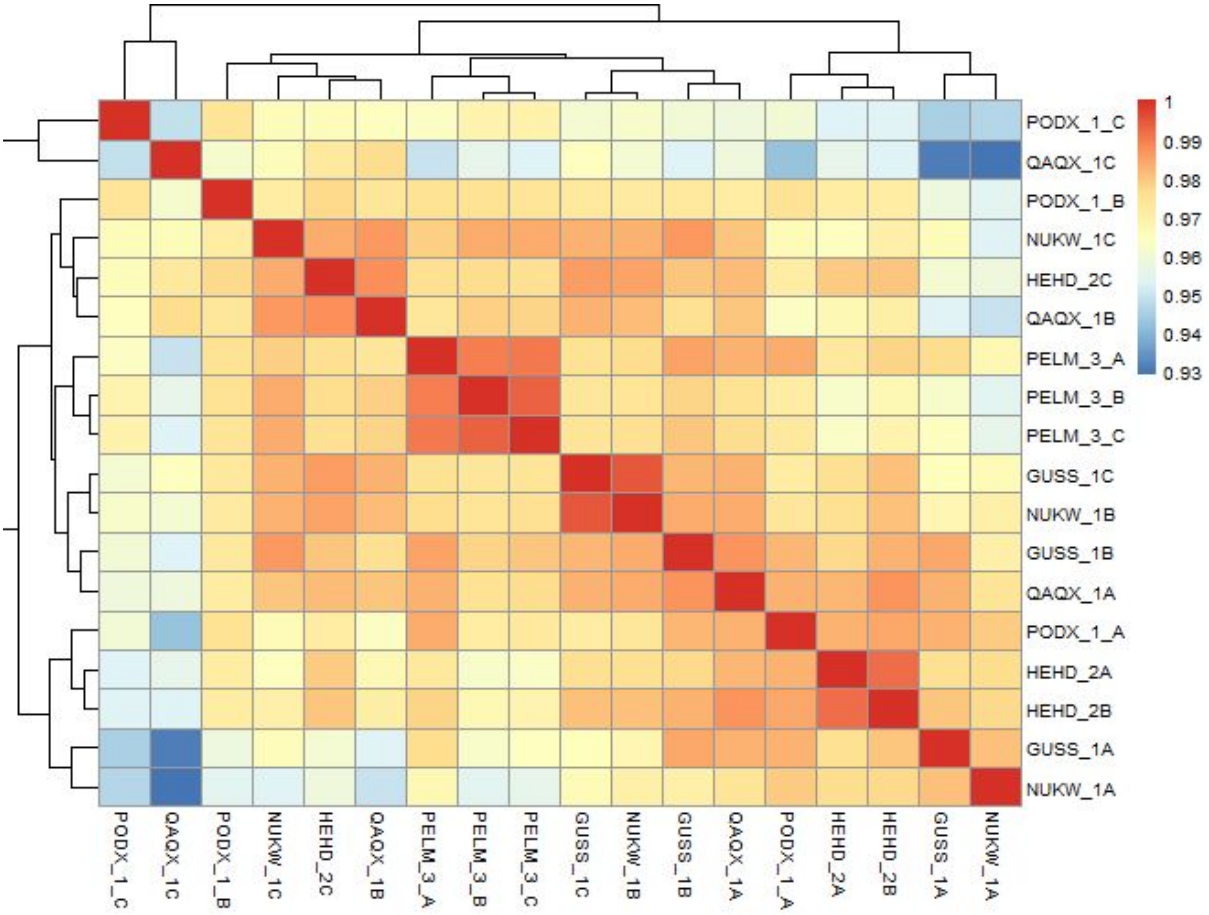
Supplementary Figure 11: a) Density plot of the number of reads per cell for neurons and fibroblast-like cells, based on 2-group clustering by SC2. **b)** Density plot of the number of genes with nonzero counts (“expressed”) for single cells.



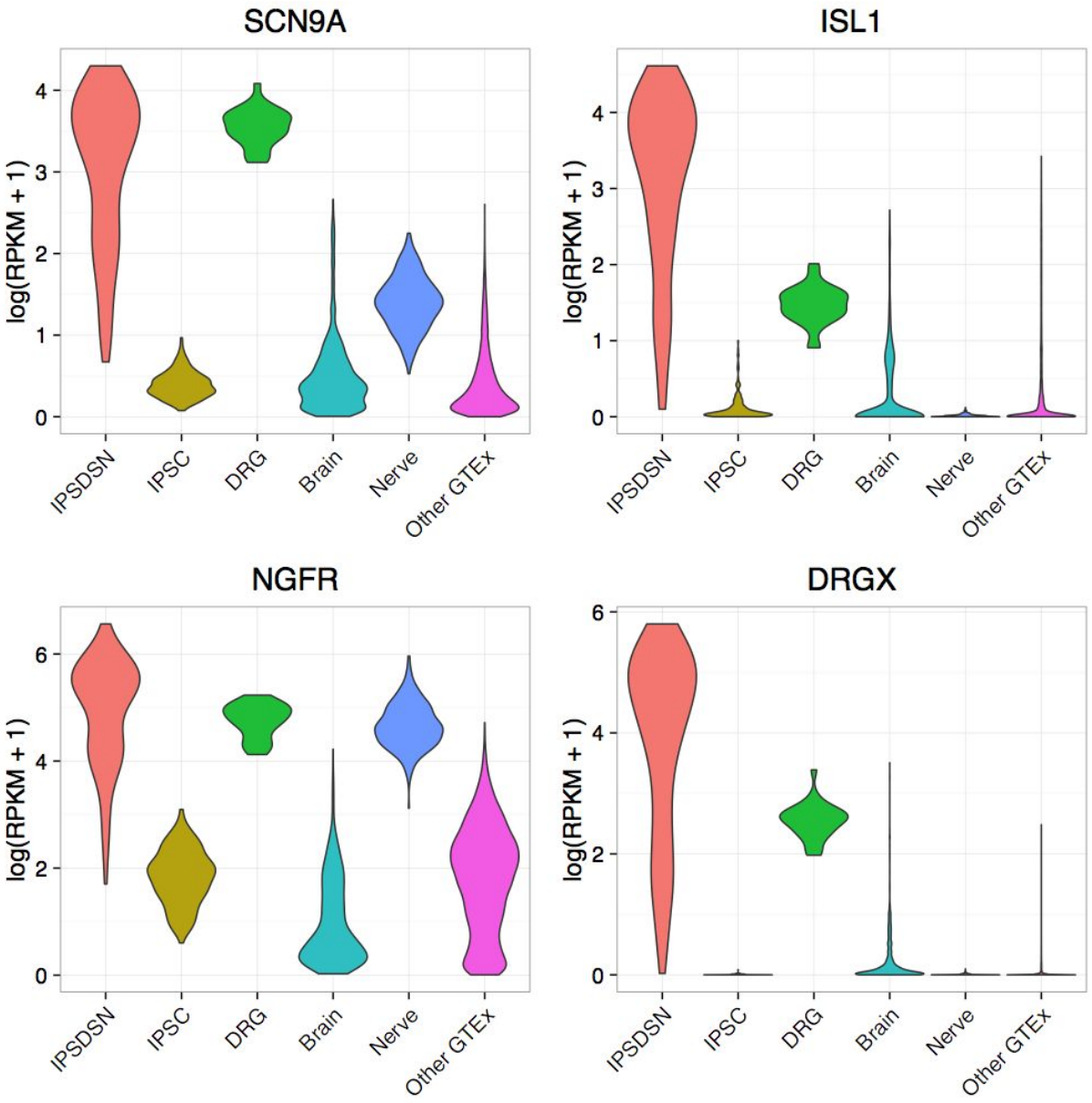
Supplementary Figure 12: Expression of sensory neuronal and fibroblast marker genes across IPSDSN samples, in comparison with DRG (N=28) and selected GTEx tissues (N=50 each). The overall similarity with DRG is high for neuronal markers, but less so for fibroblast markers. Gene expression determined as log10(FPKM +0.1), and then was mean-centered and Z-score normalized across samples for each gene, and plotted with pheatmap in R.



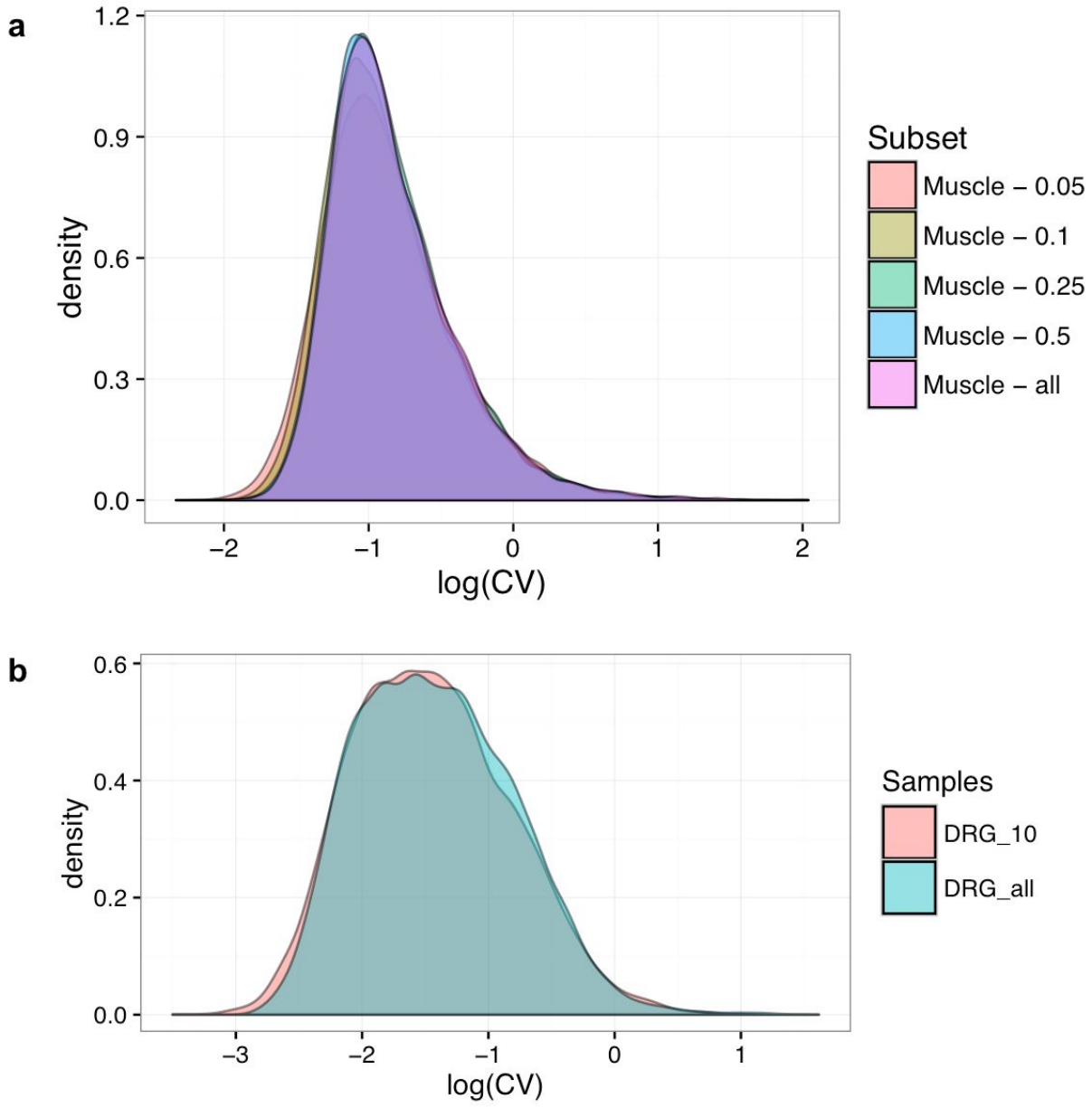
Supplementary Figure 13: Histogram of pairwise spearman correlation coefficients of gene expression among RNA extraction replicates (N=7), differentiation replicates from the same donor (n=6 donors, 3 replicates each), or across donors (n=94). RNA extraction replicates were highly repeatable (spearman ρ of 0.97 - 0.98). Differentiation replicates within a donor cell line were more variable (median ρ=0.96, range 0.93 - 0.98), but were more highly correlated than differentiations across donors (median ρ=0.93, range 0.80 - 0.98).



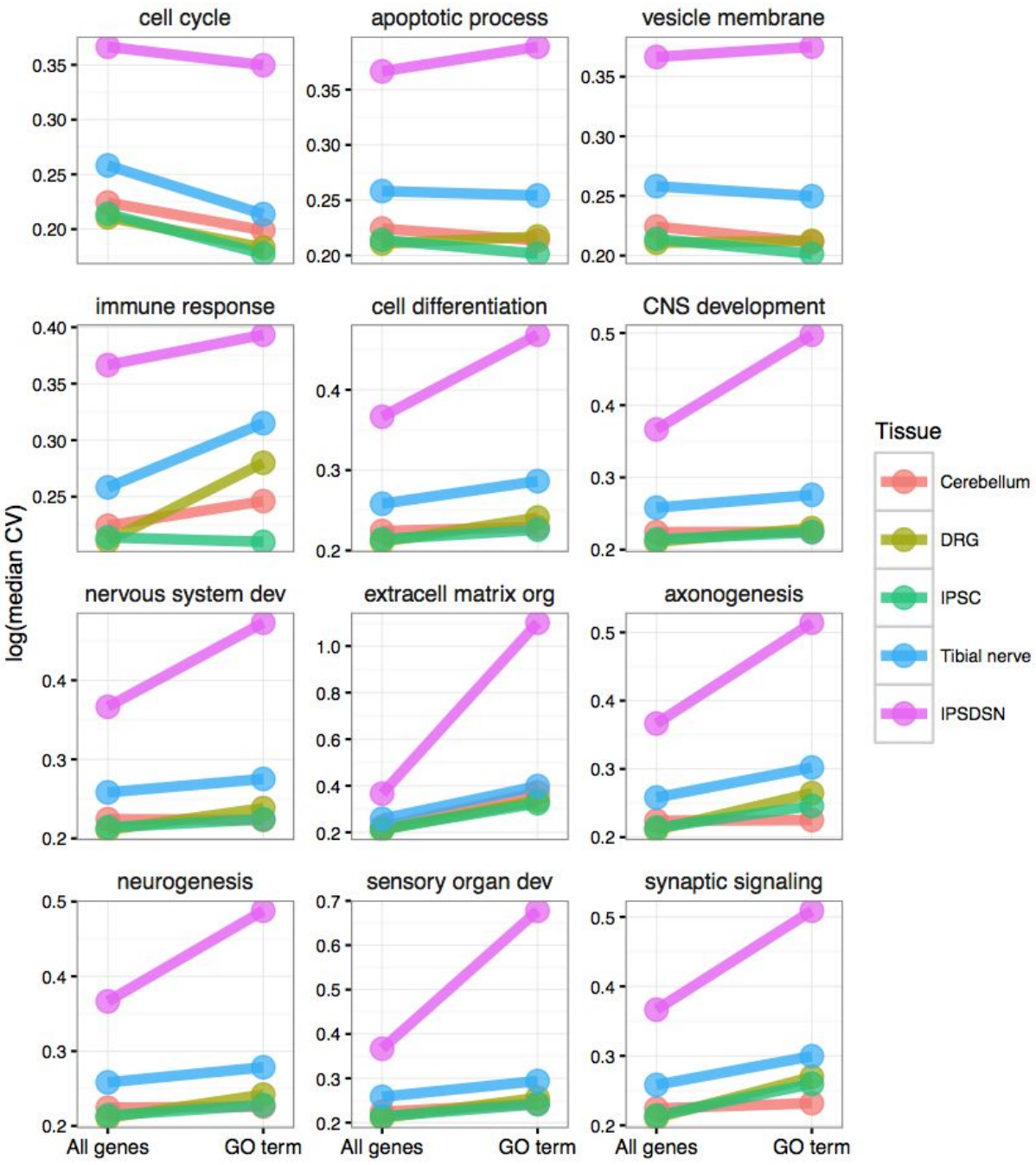
Supplementary Figure 14: Clustered heatmap of gene expression correlations between samples having biological (differentiation) replicates. Differentiation replicates for a given donor do not consistently cluster together, suggesting that variability induced during differentiation is greater variability across lines due to donor genetic background, clonal selection or cell culture conditions during reprogramming.



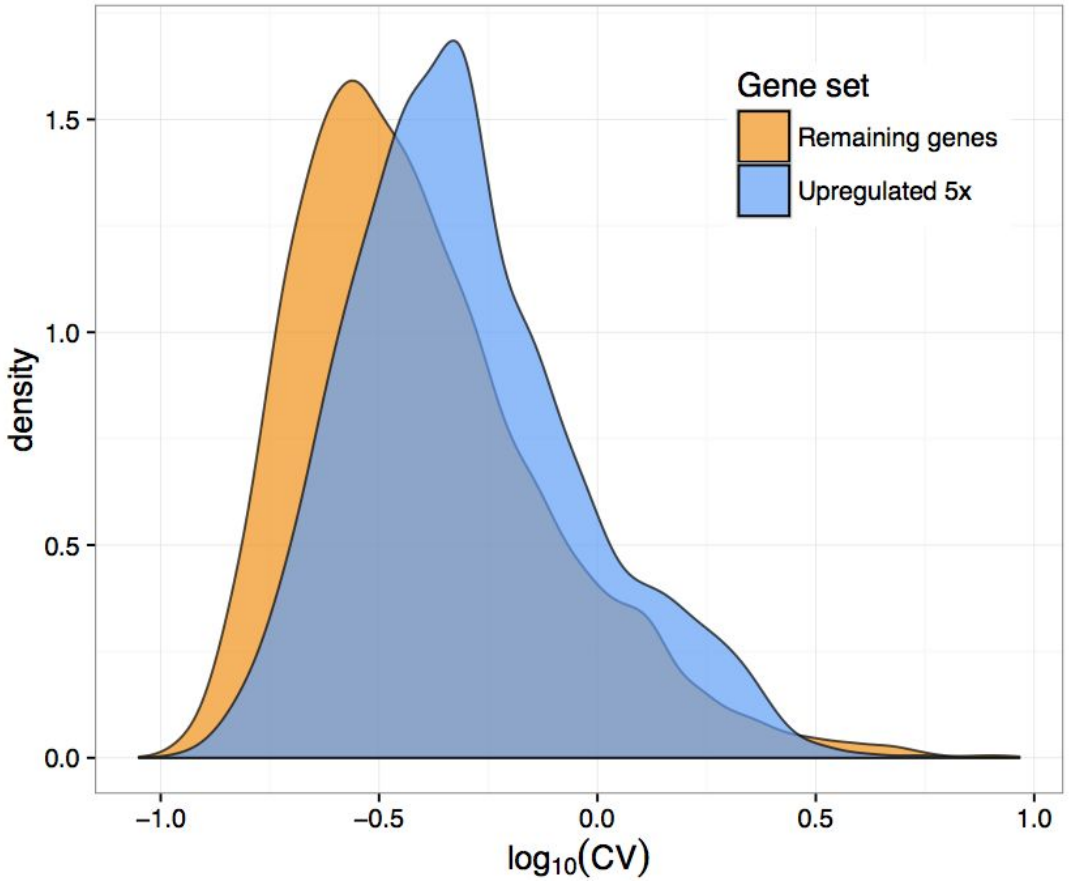
Supplementary Figure 15: Distribution of expression levels for selected sensory neuronal marker genes in IPSDSN, DRG, GTEx tibial nerve, GTEx brain, and all other GTEx tissues.



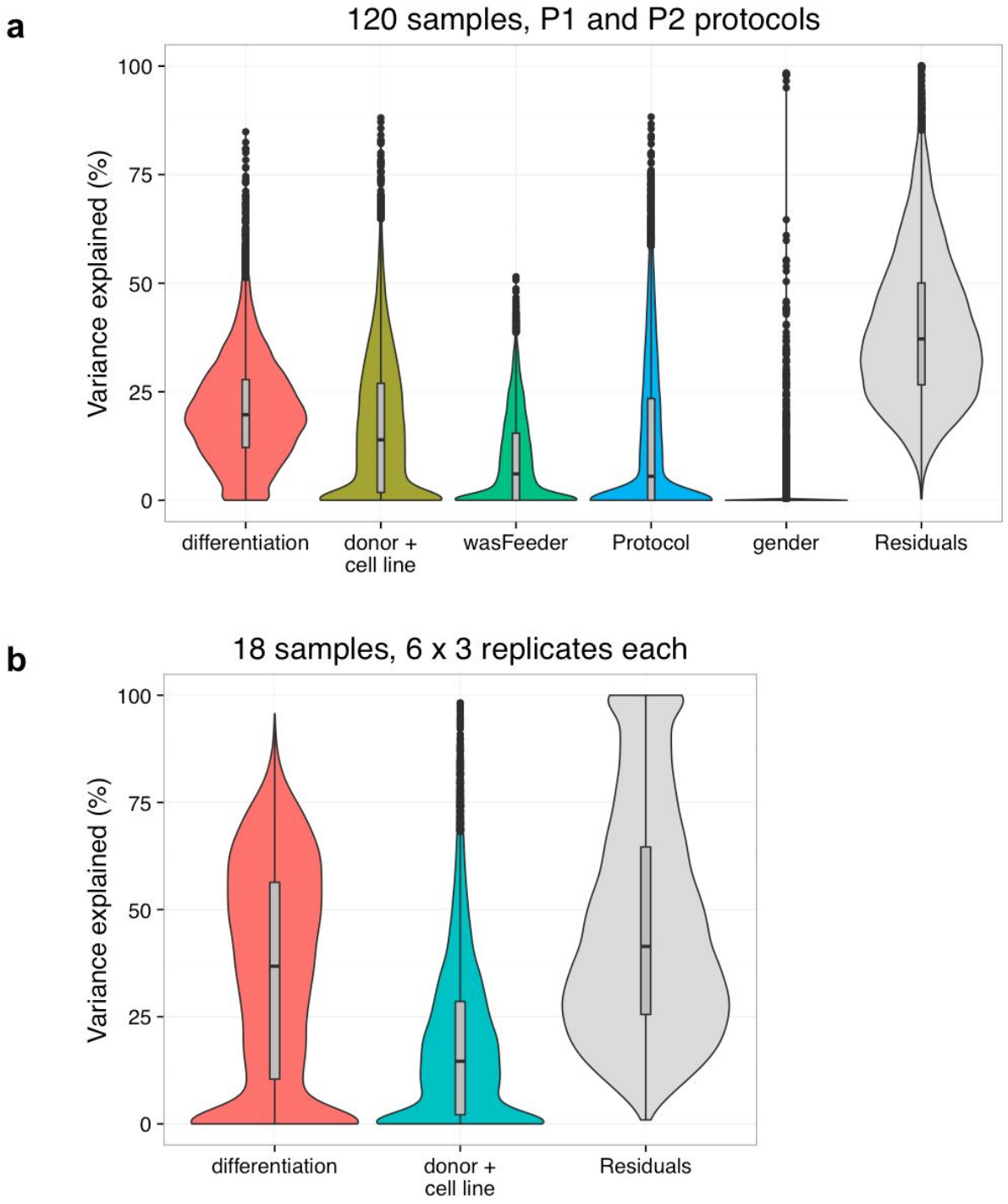
Supplementary Figure 16: (a) Sample size does not appear to affect the global distribution of gene coefficient of variation (CV). We used the GTEx tissue Skeletal Muscle (N=430), and selected random subsets corresponding to N/2, N/4, N/10, N/20 samples. We computed CV for all genes separately for each subsample. The legend indicates the fraction of samples subsampled. **(b)** The distribution of genes' coefficient of variation (CV) is equivalent when considering only 10 DRG samples from unique donors (DRG_10) or when including all 28 DRG samples (DRG_all), some of which come from different vertebrae of the same individuals.



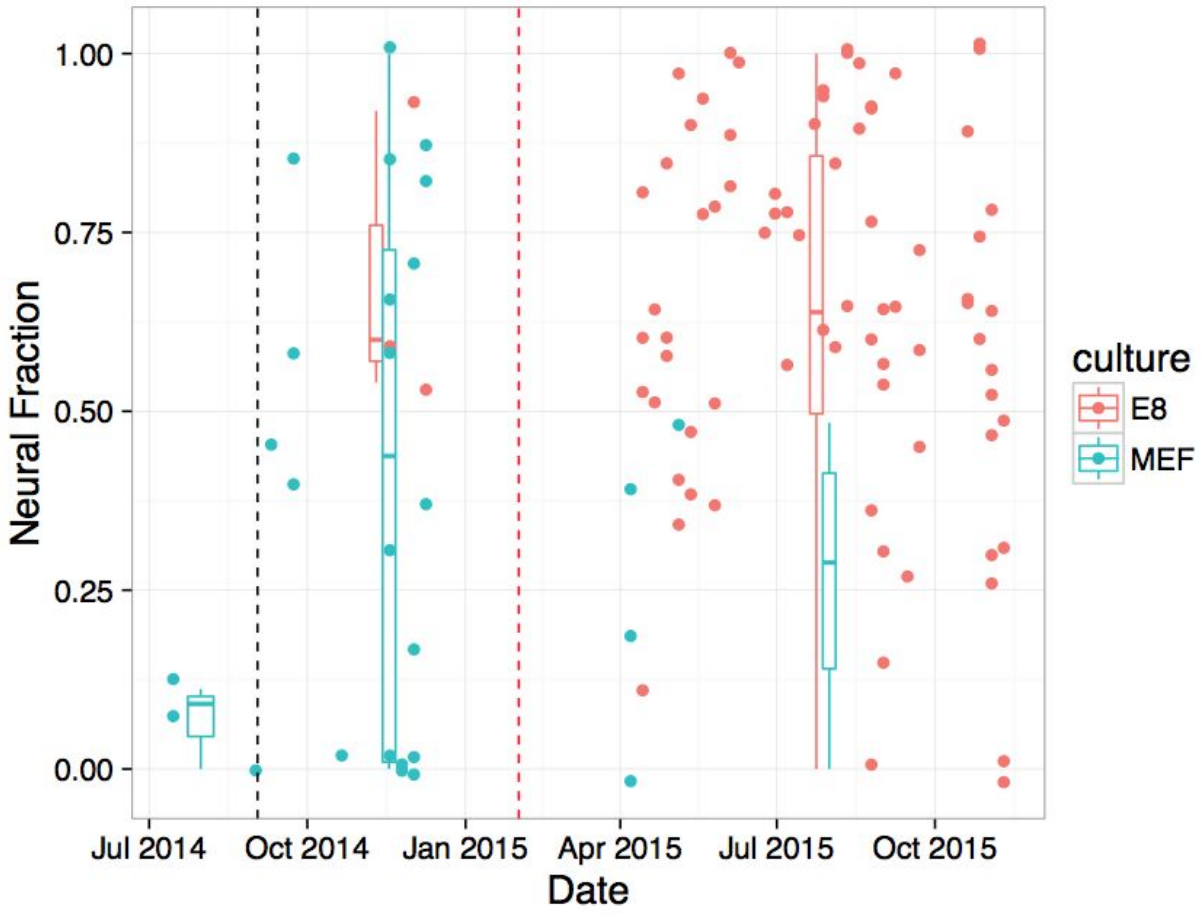
Supplementary Figure 17: Median variability of genes in specific GO terms is compared to median variability for all genes, separately for IPSDSNs, iPSCs, DRG, and GTEx cerebellum and tibial nerve. Genes categories related to neuronal function and differentiation have increased median variability in IPSDSNs relative to all genes, but this is much less the case for iPSCs or nervous tissue samples, and for other gene categories such as immune response or cell cycle genes.



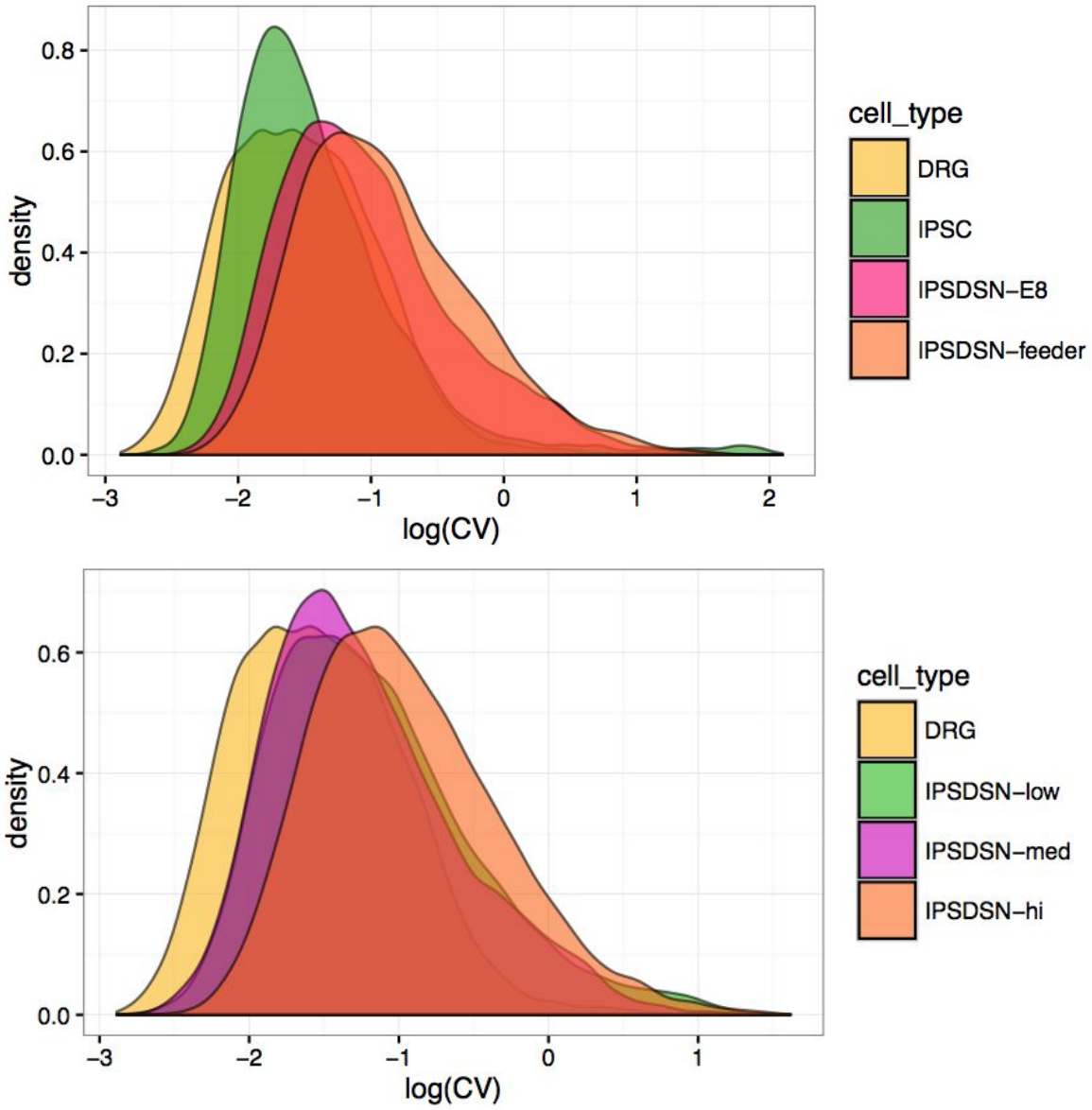
Supplementary Figure 18: Genes upregulated at least 5-fold upon differentiation from iPSC to IPSDSN (FDR < 1%, N=4246 genes) have increased variability relative to the remaining genes, despite similar levels of expression in the two groups (median/mean FPKM of upregulated genes: 4.1 / 15.6; remaining genes: 4.6 / 11.8).



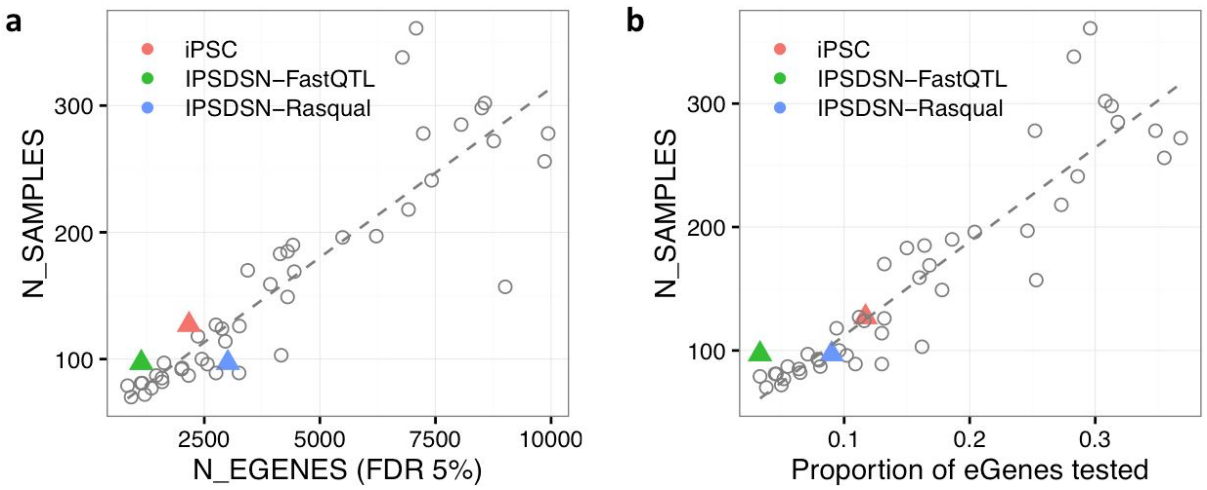
Supplementary Figure 19: Variance components analyses of IPSDSN gene expression. (**a**) Variance partitioning for 119 samples (13 P1 protocol, 106 P2 protocol). (**b**) Variance partitioning for 18 samples, from 6 iPSC lines differentiated 3 times each. All 18 samples were from E8-medium iPSC lines derived from females, and were differentiated with the P2 protocol.



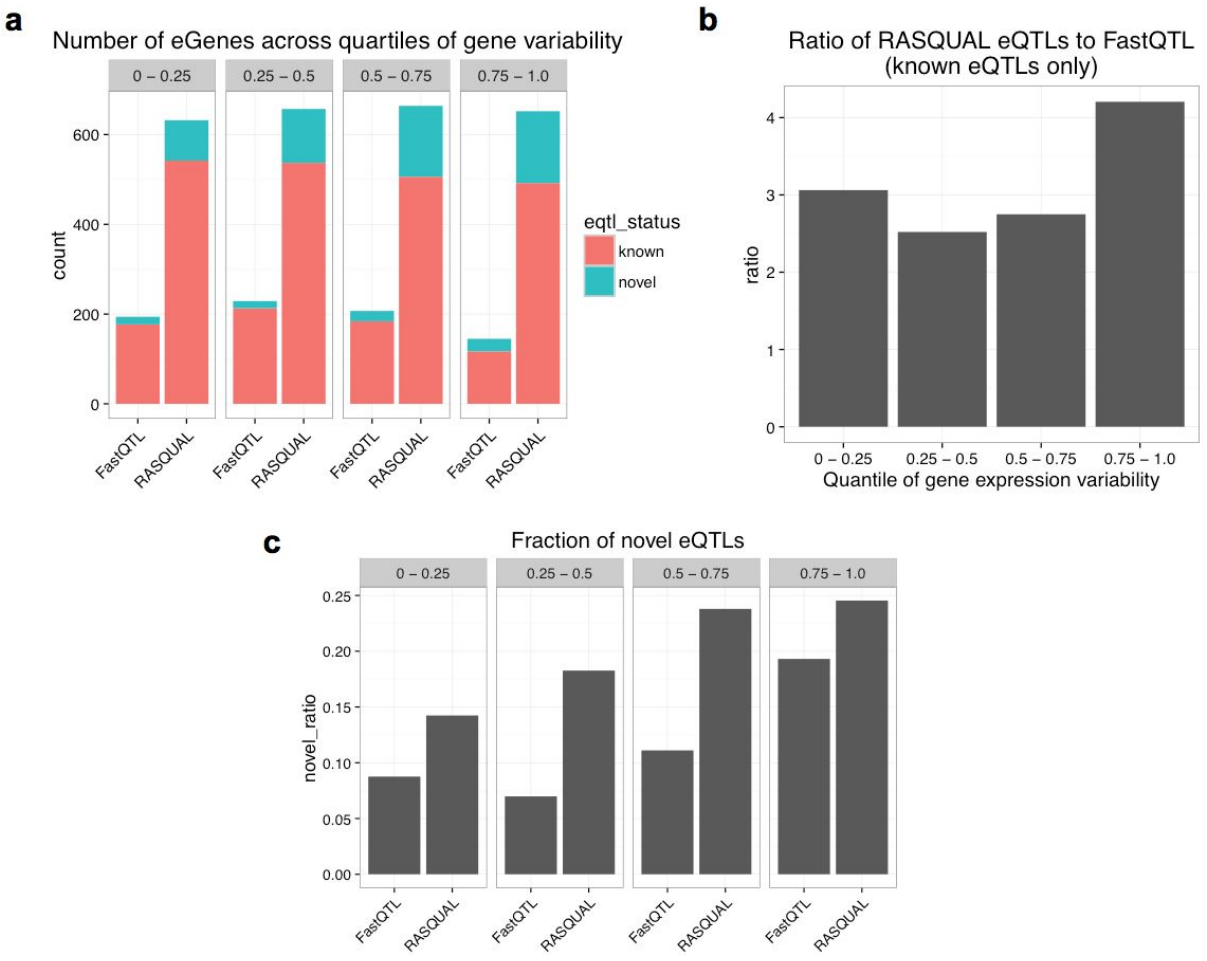
Supplementary Figure 20: Neural fraction in IPSDSN samples varies with date and iPSC culture condition. Only P2 protocol samples are shown, and each dot represents the estimated neural fraction for one sample. We used linear models in R to estimate the effects of culture conditions or differentiation date on neural fraction. We either used date as a continuous variable, or split differentiation date into 2 bins (red line) or 3 bins (black and red lines). In all models including both factors, culture conditions were more strongly associated with neural fraction than was date, and the association of culture conditions remained significant. In contrast, the association of date with was not significant with date as a continuous variable or split into 2 bins ($p > 0.3$), and was marginally significant with date split into 3 bins ($p = 0.038$).



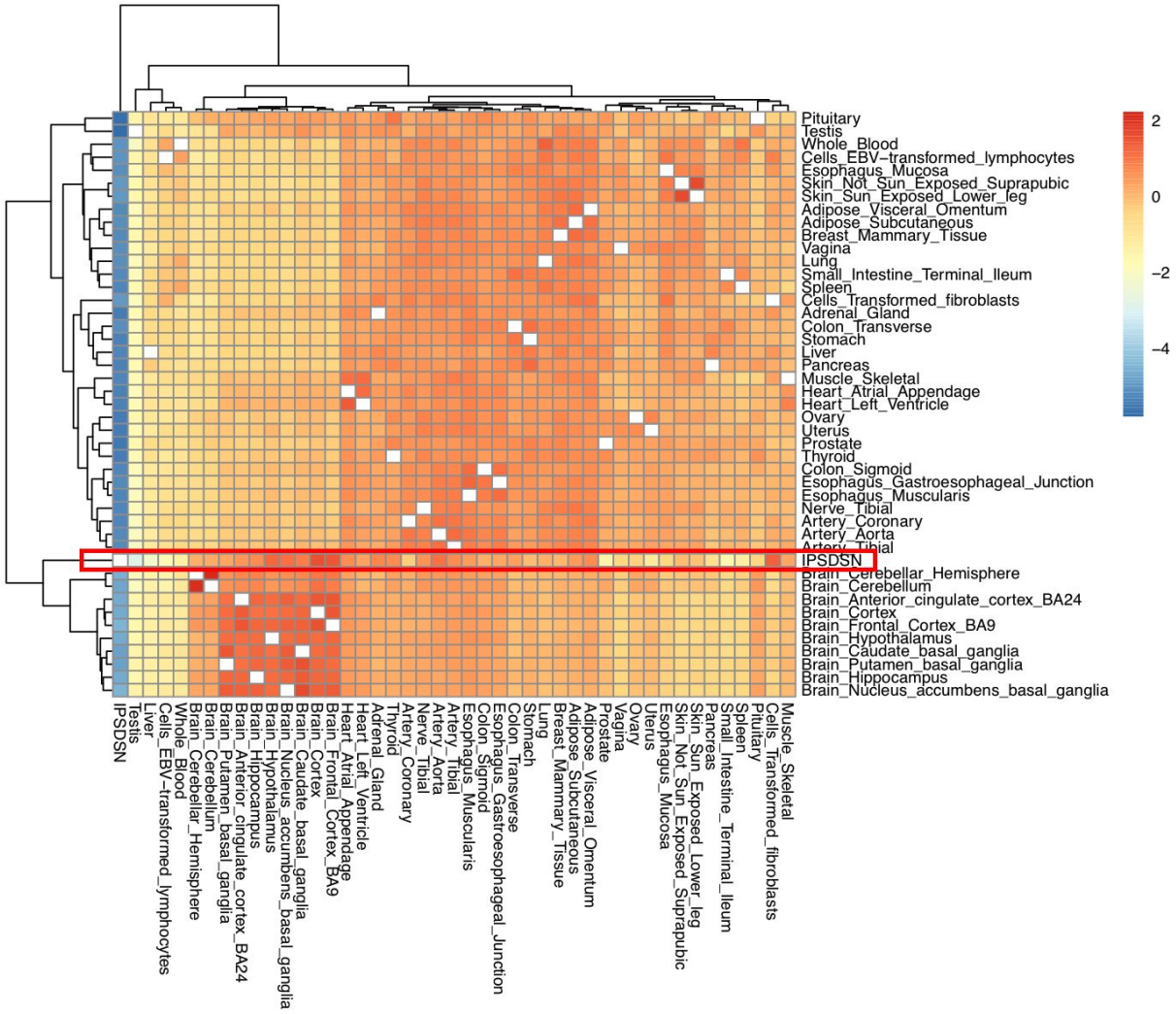
Supplementary Figure 21: The natural log of the coefficient of variation across samples is plotted for different subgroups of samples. **(a)** Samples from feeder-iPSCs (N=27) have slightly higher variability than those from E8-iPSCs (N=79), but samples from E8-iPSCs still have much higher variability than iPSC or DRG. **(b)** Comparing DRG and P2 protocol IPSDSN samples from E8-iPSCs only, separated based on fibroblast-like content: low (estimated < 20%, N=24), medium (20-50%, N=31), and high (> 50%, N=24). High fibroblast-like samples have slightly higher CV across all genes, but this accounts for only a fraction of the increased variability seen relative to primary DRG.



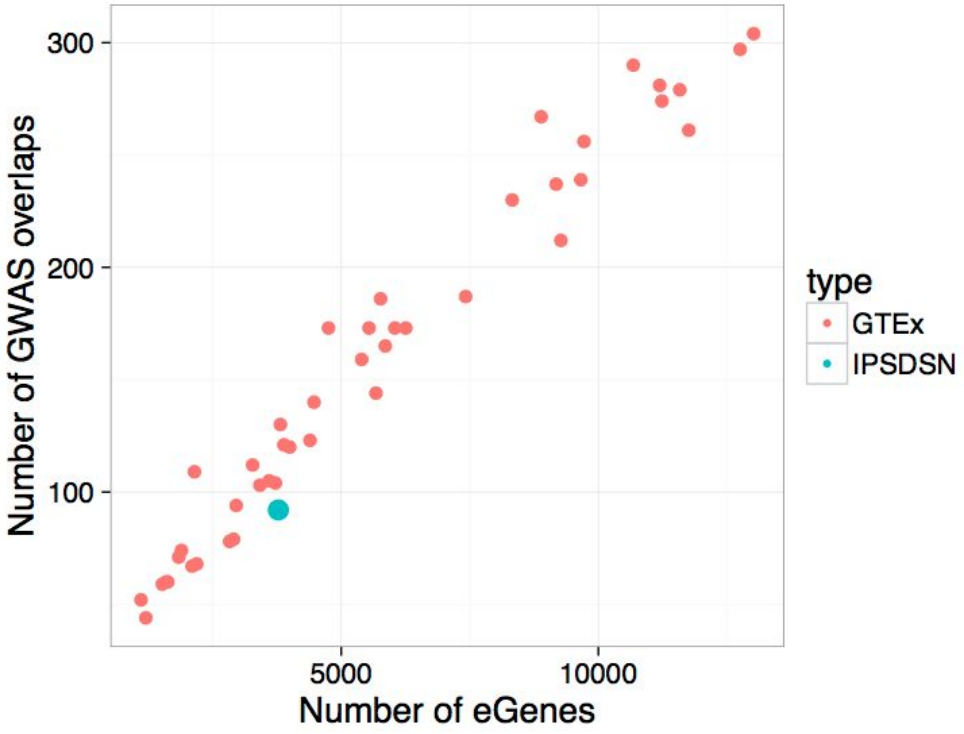
Supplementary Figure 22: (a) Number of eGenes discovered for IPSDSNs using either RASQUAL (Neuron_Rasqual) or FastQTL (Neuron_fastQTL), in relation to sample size, with other GTEx tissues also plotted. **(b)** The proportion of all tested genes called as eGenes.



Supplementary Figure 23: (a) Number of eGenes discovered by RASQUAL and FastQTL across quartiles of gene expression variability. The bottom 25% of eGenes ranked by sample-to-sample expression variability are at the left, while the top 25% are at the right. Bar colors show the number of eQTLs that overlap with a GTEx eQTL ("known") or where no GTEx tissue has $p < 0.01$ for our lead eQTL SNP ("novel"). As support for the claim that RASQUAL power gains do not come at the expense of an increased false positive rate, we see that the rate of novel eQTLs increases only slightly from the least to the most highly variable genes, and the trend is similar for FastQTL and for RASQUAL. **(b)** Restricting to known eQTLs discovered by either method, the ratio of the number of eGenes discovered by RASQUAL to FastQTL is highest in the highest quartile of gene expression variability, although power gains from RASQUAL are high across the board. **(c)** The fraction of novel eQTLs increases for both fastQTL and RASQUAL as gene variability increases, with RASQUAL overall finding a higher fraction of novel eQTLs. Although one explanation would be a higher false positive rate for RASQUAL, various properties of the novel eQTLs do not differ significantly from known eQTLs (expression levels, allele frequency, hardy-weinberg equilibrium, mapping bias, number of feature SNPs). It is possible that because GTEx used a linear model, RASQUAL finds true eQTLs that are more difficult to discover without examining allele-specific expression.



Supplementary Figure 24: Clustering of tissues based on eQTL effect sizes. IPSDSNs cluster most closely with GTEx brain tissues, although they also show elevated similarity in effect sizes with GTEx fibroblasts (second to last column). eQTL effect sizes were correlated pairwise between tissues. As expected, due to technical differences between GTEx and our project, IPSDSNs had globally lower similarity with GTEx than GTEx tissues did among themselves. We therefore Z-scaled the rows of the correlation matrix, which puts the similarities between tissues on the same scale. Intuitively, what we are thus asking is where IPSDSNs cluster based on their relative similarity with other GTEx tissues (clustering of rows). As can be seen in the highlighted row, IPSDSNs have higher relative similarity with GTEx brain (redder colour) than with other tissues. Clustering of columns, which contain unscaled pairwise correlations, shows that IPSDSNs are outliers in that they have globally lower similarity (bluer colour) with all GTEx tissues.



Supplementary Figure 25: The number of overlaps at LD R^2 > 0.8 between eQTLs in IPSDSNs and GWAS catalog SNPs is within the range seen for similarly powered tissues in GTEx. The number of eQTL-GWAS overlaps is heavily dependent on the number of eQTLs discovered, which is reflected in the tight linear relationship with the number of eGenes.

Supplementary Note

HIPSCI consortium members

Helena Kilpinen[1], Angela Goncalves[2], Vackar Afzal[3], Kaur Alasoo[2], Sofie Ashford[4], Sendu Bala[2], Dalila Bensaddek[3], Francesco Paolo Casale[1], Oliver J Culley[5], Petr Danecek[2], Adam Faulconbridge[1], Peter W Harrison[1], Annie Kathuria[5], Davis McCarthy[1,9], Andreas Leha[10], Shane A McCarthy[2], Ruta Meleckyte[5], Yasin Memari[2], Nathalie Moens[5], Filipa Soares[6], Alice Mann[2], Ian Streeter[1], Chukwuma A Agu[2], Alex Alderton[2], Rachel Nelson[2], Sarah Harper[2], Minal Patel[2], Alistair White[2], Sharad R Patel[2], Laura Clarke[1], Reena Halai[2], Christopher M Kirton[2], Anja Kolb-Kokocinski[2], Philip Beales[8], Ewan Birney[1], Davide Danovi[5], Angus I Lamond[3], Willem H Ouwehand[2,4,7], Ludovic Vallier[2,6], Fiona M Watt[5], Richard Durbin[2,11], Oliver Stegle[1], Daniel J Gaffney[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
[2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom.
[3]Centre for Gene Regulation & Expression, School of Life Sciences, University of Dundee, DD1 5EH, United Kingdom.
[4]Department of Haematology, University of Cambridge, Cambridge, United Kingdom.
[5]Centre for Stem Cells & Regenerative Medicine, King's College London, Tower Wing, Guy's Hospital, Great Maze Pond, London SE1 9RT, United Kingdom.
[6]Wellcome Trust and MRC Cambridge Stem Cell Institute and Biomedical Research Centre, Anne McLaren Laboratory, University of Cambridge, CB2 0SZ, United Kingdom.
[7]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom.
[8]UCL Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, United Kingdom.
[9]St Vincent's Institute of Medical Research, Fitzroy Victoria 3065, Australia.
[10]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
[11]Department of Genetics, University of Cambridge, Cambridge, United Kingdom.

RNA sequencing

Cells growing in T25 flasks were washed twice with PBS followed by addition of 600 mL of RLTPlus buffer. Cells were gently lifted from the flask, lysed, and transferred to 1.5 ml tubes. RNA and gDNA were isolated using AllPrep DNA/RNA Minikit (Qiagen). RNA was eluted in 33 uL of DNAse-free water and DNA eluted in 53 uL EB buffer. RNA libraries were prepared using the Illumina TruSeq strand-specific protocol, and were sequenced with paired-end reads (2x75) on Illumina Hiseq with V4 chemistry.

ATAC library preparation and sequencing

Nuclei isolation

Media was removed from T25 flasks and washed twice with 10 mL of room temperature D-PBS without calcium and magnesium. The adherent neuronal cultures were lifted by treating with 3 mL of Accutase (Millipore – SCR005) at room temperature for four minutes. The Accutase was quenched by adding 6 mL of 2% foetal bovine serum in D-PBS. The cells were transferred to a 15 mL conical tube and centrifuged at 300 g for 5 minutes at 4 °C. The cell pellet was resuspended in 1 mL of ice-cold sucrose buffer (10 mM tris-Cl pH 7.5, 3 mM $CaCl_2$, 2 mM $MgCl_2$ and 320 mM sucrose) and pipetted briefly to break up the large clumps before incubating on ice for 12 minutes. 50 µL of 10% Triton-X 100 was added to the sucrose-treated cells and mixed briefly before incubating on ice for a further 6 minutes. Nuclei were released by performing 30 strokes with a tight dounce homogeniser on ice. Approximately 1 x 10^5 nuclei were transferred to a 1.5 mL microfuge tube and centrifuged at 300 g for 5 minutes at 4 °C. All traces of the lysis buffer were removed from the nuclei pellet.

Tagmentation, PCR amplification and size selection

The tagmentation and PCR methods used here are in principle the same as that described in (Buenrostro et al. 2013), but with some modifications as described in (Kumasaka, Knights, and Gaffney 2016). The nuclei pellet was resuspended in 50 µL of Nextera tagmentation master mix (Illumina FC-121-1030) (25 µL 2x Tagment DNA buffer, 20 µL nuclease-free water and 5 µL Tagment DNA Enzyme 1) and incubated at 37 °C for 30 minutes. The tagmentation reaction was stopped by the addition of 500 µL Buffer PB (Qiagen) and purified using the MinElute PCR purification kit (Qiagen 28004), according to the manufacturer's instructions and eluting in 10 µL of Buffer EB (Qiagen). 10 µL of the tagmented chromatin was mixed with 2.5 µL Nextera PCR primer cocktail and 7.5 µL Nextera PCR mastermix (Illumina FC-121-1030) in a 0.2 mL low-bind PCR tube. The indexing primers used for amplification were from the Nextera Index kit (Illumina FC-121-1011), using 2.5 µL of an i5 primer and 2.5 µL of an i7 primer per PCR, totalling 25 µL. PCR amplification was performed as follows: 72 °C for 3 minutes and 98 °C for 30 seconds, followed by 12 cycles of 98 °C for 10 seconds, 63 °C for 30 seconds and 72 °C for 3 minutes. To remove the excess of unincorporated primers, dNTPS and primer dimers, Agencourt AMPure XP magnetic beads (Beckman Coulter A63880) were used at a ratio of 1.2 AMPure beads:1 PCR sample (v/v), according the manufacturer's instructions, eluting in 20 µL of Buffer EB (Qiagen). Finally, size selection was performed by 1 % agarose TAE gel electrophoresis, selecting library fragments from 120 bp to 1 kb. Gel slices were extracted with the MinElute Gel Extraction kit (Qiagen 28604), eluting in 20 µL of Buffer EB.

Illumina sequencing
A total of 31 ATAC-seq libraries each prepared with a unique Nextera i5 and i7 tag combination were pooled. Index tag ratios were assessed by a single MiSeq run and were balanced before being sequenced at two per lane with paired-end reads (2x75) on a HiSeq with V4 chemistry. However, rebalancing did not appear to work correctly, as the number of reads varied from a minimum of 17 million to a maximum of 987 million. However, 22 samples had over 100 million reads, and 30 samples had over 40 million reads. Across samples, a median of 56% of reads mapped to mitochondrial DNA.

Read alignment
We aligned reads to GRCh38 human reference genome using bwa mem v0.7.12. Reads mapping to the mitochondrial genome and alternative contigs were excluded. As for RNA-seq data, we used VerifyBamID v1.1.2 (Jun et al. 2012) to detect sample swaps. This revealed one mislabeled sample, which we then corrected. We used Picard v1.134 MarkDuplicates (https://broadinstitute.github.io/picard/) to mark duplicate fragments.

Peak calling
We used MACS2 v2.1.1 (Zhang et al. 2008) to call ATAC-seq peaks for individual samples with parameters '--nomodel --shift -25 --extsize 50 -q 0.01'. We defined a consensus set of peaks as regions in which peaks overlapped in at least 3 samples. At regions of overlap, the consensus peak was defined as the union of overlapping peaks. This resulted in 381,323 peaks, with 98% of peaks ranging from 82 - 1191 base pairs.

PCA plot clustering samples with GTEx tissues

We downloaded the GTEx v6 gene RPKM file (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz) as well asand sample metadata (GTEx_Data_V6_Annotations_SampleAttributesDS.txt) from the GTEx web portal (http://www.gtexportal.org/home/datasets). We computed gene RPKMs for all genes for the 28 DRG samples, the 119 sensory neuron samples (5 outliers removed), and 239 HIPSCI IPS samples. We used genes that were quantified in all of these sample sets, and where at least 50 GTEx samples had RPKM > 0.1. We passed log2(RPKM + 1) for 8553 GTEx samples to the bigpca R package to compute the first 5 PCs using the SVD method. We then determined sample loadings for each PC using the PC weights and log2(RPKM + 1) values for GTEx samples as well asand for our in-house samples, and plotted sample PC1 vs. PC2 values as Figure 1b.

Single-cell RNA sequencing

Blood-derived iPSCs from a single individual, who was not a HIPSCI donor, were differentiated to IPSDSNs in 3 batches using the P2 protocol, and were matured for 8 weeks. Dissociated cells were loaded onto a Fluidigm C1 for automatic cell separation, reverse transcription and amplification. Libraries were prepared from C1 chambers containing single cells using the Illumina Nextera XT kit. These were quantified with the Qubit dsDNA HS assay (Thermo Fisher) and KAPA Library Quantification Kit (KAPA Biosystems) and size-checked with Agilent Bioanalyser DNA 1000. Libraries were 96-way multiplexed and

sequenced by Illumina Nextseq500 (2x75bp). Reads were aligned to GRCh38 and Ensembl 80 transcript annotations using STAR v2.4.0d with default parameters. We excluded 9 cells expressing fewer than 20% of the ~56,000 quantified genes, and then used SC3[14] to cluster the remaining 177 cells based on expression counts. We examined alternative numbers of clusters from k=2 to 5 (Supplementary Figure 6). With two clusters, marker genes clearly identified one cluster (111 cells) as neuronal, whereas the other cluster (66 cells) had high expression of extracellular matrix genes reminiscent of fibroblasts. With 3 and 4 clusters, the sensory-neuronal cell cluster remained unchanged, and the fibroblast-like cluster became further subdivided. This suggests that a majority of the cells in this sample were terminally differentiated into sensory neurons, whereas the remaining cells were more heterogeneous in their gene expression.

To display marker gene expression (Figure 2a), we used DESeq2's variance stabilizing transformation, and then R's "scale" function to mean-center and normalize expression values across cells, and plotted the result using the pheatmap R package. To compare gene expression between single-cell clusters and bulk RNA-seq samples (Supplementary Figure 8), we computed the mean FPKM for each gene separately in single neurons and fibroblast-like cells. We subsetted to genes with nonzero expression in at least one GTEx tissue and in at least one of our tissues (iPSC, DRG, IPSDSN bulk, IPSDSN single cells), and computed the Spearman correlation between each pair of tissues.

QTL calling

Expression QTLs

To call cis-eQTLs we used RASQUAL (Kumasaka, Knights, and Gaffney 2016), which leverages allele-specific reads in heterozygous individuals to improve power for QTL discovery, while accounting for reference mapping bias and a number of other potential artifacts. With RASQUAL a feature is defined by a set of start and end coordinates; for calling a gene eQTL these are the start and end coordinates for exons, whereas for an ATAC-seq peak these are the peak coordinates. RASQUAL requires as input the allele-specific read counts at each SNP within a feature. We used the Genome Analysis Toolkit (GATK) program ASEReadCounter (Castel et al. 2015) with options '-U ALLOW_N_CIGAR_READS -dt NONE --minMappingQuality 10 -rf MateSameStrand' to count allele-specific reads at SNPs (and not indels). We then annotated the AS read counts in the INFO field of the VCF used as input for RASQUAL. We used custom scripts to determine the number of feature SNPs in gene exons.

We used RASQUAL's makeCovariates.R script to determine principal components (PCs) to use as covariates, which determined 12 PCs as appropriate from the expression count data. We ran RASQUAL separately for each of 35,033 genes (19,796 protein-coding genes and 15,237 noncoding RNAs), passing in VCF lines for all SNPs and indels (MAF > 0.05, INFO > 0.8) within 500 kb of the gene transcription start site. We used the --no-posterior-update option in RASQUAL, as we found that not doing so led to some genes having miniscule p values, even with permuted data. To correct for multiple testing we used permutations; however, because RASQUAL is computationally intensive, it would not be possible to run a thousand or more permutations for every gene. Therefore we used an approach to balance power and computational time. To correct for the number of SNPs tested per gene, we used

EigenMT(Davis et al. 2016) to estimate the number of independent tests per gene, and then performed Bonferroni correction on a gene-by-gene basis. To estimate the false discovery rate (FDR) across genes, we used the --random-permutation option of RASQUAL and re-ran it once for every gene, saving the minimum p value (after eigenMT correction) of the SNPs tested for each gene. This gave a distribution of minimum p values across genes for the permuted data. To determine the FDR for eQTL discovery at a given gene, we use R to compute (#permuted data min pvalues < p) / (#real data min p values < p), where p is the minimum p value among SNPs for the gene in question. With this procedure we obtained 3,586 genes with a cis-eQTL at FDR 10% (2,628 at FDR 5%).

For QTL calling with FastQTL, we first computed principal components from the CQN-transformed gene expression matrix (cqn v5.0.2 (Hansen, Irizarry, and Wu 2012)). We ran FastQTL with permutations 31 separate times, in each run including the first N principal components (N=0...30) as covariates. For each run we used a cis-window of 500 kb, and included SNPs and indels with MAF > 0.05, INFO > 0.8, as we did for RASQUAL. We plotted the number of eGenes found in each of these runs, which plateaued and remained relatively stable at ~1,400 eGenes (FDR 10%) when anywhere from 16 to 30 PCs were used. We arbitrarily chose to use the FastQTL run with 20 PCs in downstream analyses.

ATAC QTLs

As we did for gene expression, we used featureCounts v1.5.0 to count fragments overlapping consensus ATAC-seq peaks and ASEReadCounter to count allele-specific reads at SNPs (and not indels) within peaks. We ran RASQUAL separately for each of 381,323 peaks, passing in VCF lines for SNPs and indels (MAF > 0.05, INFO > 0.8) within 1 kb of the center of the peak. Since >99.9% of peaks were less than 2 kb in size, this meant that we tested effectively all SNPs within peaks. As we did when calling eQTLs, we ran RASQUAL with the --random-permutation option for every gene, and determined FDR as described above. Note that in this case we used Bonferroni correction based on the number of SNPs tested, without using EigenMT, due to the small size of the windows tested. With this procedure we obtained 6,318 ATAC peaks with a cis-QTL at FDR 10%.

Splice QTLs

We downloaded LeafCutter from Github (https://github.com/davidaknowles/leafcutter) on April 17, 2016. We used the LeafCutter bam2junc.sh script to determine junction counts for each sample, followed by leafcutter_cluster.py. This resulted in 254,057 junctions in 59,736 clusters. To focus on splicing events likely to be significant, we applied a number of filters, including: (a) removing junctions accounting for less than 2% of the cluster reads, (b) removing introns used (i.e. having at least 1 supporting read) in fewer than 5 samples, (c) retaining only clusters where at least 10 samples had 20 or more reads in the cluster. This yielded a filtered set of 95,786 junctions in 30,591 clusters. We first determined the read proportions for all junctions within alternatively excised clusters. We then Z-score standardised each junction read proportion across samples, and then quantile-normalised across introns. We used this as our phenotype matrix for input to FastQTL to test for associations between intron usage and variants within 15 kb of the center of each intron. We chose a cis-window size of 30 kb (2 x 15 kb) because >91% of introns are < 30 kb in size,

and so this tests variants near exon/intron boundaries for the great majority of introns, while maximising power.

We ran FastQTL in nominal pass mode 31 times specifying the first 0 to 30 principal components as covariates, and examined the number of intron QTLs with minimum SNP p value < 10^{-5}. This showed that the number of QTLs plateaued when 5 PCs were used, and so we used 5 PCs in subsequent runs. We next ran FastQTL with 10,000 permutations to determine empirical p values for each alternatively excised intron. To correct for the number of introns tested per cluster, we used Bonferroni correction on the most significant intron p value per cluster. We then used the Benjamini-Hochberg method to estimate FDR across tested clusters. This yielded 2,079 significant SNP associations for intron usage (sQTLs) at FDR 10%.

For significant sQTLs we used bedtools closest with GRCh38 release 84 to annotate the gene(s) nearest the lead SNP for the association. To ensure we had relevant genes, we filtered the annotation to include only genes where one of the exon boundaries matched the intron boundary for the sQTL.

Electrophysiological recordings

Six coverslips per line were placed singularly into a 12-well plate and washed 1x with 1 ml DPBS (+/+). The coverslips were then coated with 1 ml of 0.33 mg/ml growth factor reduced matrigel for > 3 hr at room temperature. D14 cells were prepared at 1.6e6/ml in 15 ml media, then diluted in NB media to create a 0.3e6/ml suspension. The coverslips were transferred into a 12-well plate and 1 ml cell suspension was added. Plates were incubated at 37°C (5% CO2) for 24hr, after which the coverslips were transferred to a 12-well plate with 2 ml media. Cells were treated with Mitomycin C (0.001 mg/ml for 2hr hours at 37°C) post-plating on days 4 and 10. Media was changed twice weekly.

Patch-clamp experiments were performed in whole-cell configuration using a patch-clamp amplifier 200B for voltage clamp and Multiclamp 700A or 700B for current clamp controlled by Pclamp 10 software (Molecular Devices). Experiments were performed at 35°C or 40°C controlled by an in-line solution heating system (CL-100, Warner Instruments). Temperature was calibrated at the outlet of the in-line heater daily before the experiments. Patch pipettes had resistances between 1.5 and 2 MΩ. Basic extracellular solution contained (mM) 135 NaCl, 4.7 KCl, 1 $CaCl_2$, 1 $MgCl_2$, 10 HEPES and 10 glucose; pH was adjusted to 7.4 with NaOH. The intracellular (pipette) solution for voltage clamp contained (mM) 100 CsF, 45 CsCl, 10 NaCl, 1 $MgCl_2$, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with CsOH. For current clamp the intracellular (pipette) solution contained (mM) 130 KCl, 1 $MgCl_2$, 5 MgATP, 10 HEPES, and 5 EGTA; pH was adjusted to 7.3 with KOH. The osmolarity of solutions was maintained at 320 mOsm/L for extracellular solution and 300 mOsm/L for intracellular solutions. All chemicals were purchased from Sigma. Currents were sampled at 20 kHz and filtered at 5 kHz. Between 80% and 90% of the series resistance was compensated to reduce voltage errors. Rheobase was measured in current clamp mode by injecting increasing 30 milliseconds current steps until a single action potential was evoked. Intersweep intervals were 2 seconds. Current clamp data was analyzed using Spike2 software (Cambridge Electronic Device, UK) and Origin 9.1 software (Originlab).

Identifying tissue-specific eQTLs

We determined the set of tissue-specific eQTLs using the same procedure and code as in the HIPSCI project (Kilpinen et al. 2017). Briefly, we considered the full cis eQTL output of sensory neuron eQTLs and 44 tissues analyzed by the GTEx Project (Consortium et al. 2015). To enable comparison, lead SNP positions for sensory neuron eQTLs were first lifted back from GRCh38 to GRCh37 using Crossmap (Zhao et al. 2014). For each discovery tissue (including sensory neurons), we tested for the replication of all lead eQTL - target eGene pairs reported at FDR 5%. If the lead eQTL variant was not reported in the comparison tissue, then the best high-LD proxy of the lead variant ($r^2 > 0.8$ in the UK10k European reference panel) was used as the query variant. Replication was defined as the query variant having a nominal eQTL $p < 2.2x10^{-4}$ (corresponding to $p = 0.01 / 45$, where 45 refers to the total number of tissues tested) for the same eGene. We then extracted eGenes for which the lead eQTL did not show evidence of replication in any other tissue ($p > 2.2x10^{-4}$) or could not be tested (i.e. was not measured or reported as expressed in any other tissue).

This analysis gave 954 eGenes where the eQTL is specific to sensory neurons (Supplementary Table 15). We note that some of these “tissue-specific” eGenes could be due to the difference in QTL-calling methods used, notably that we used RASQUAL, a method incorporating both allele-specific and population-level expression variation. Therefore, some of the tissue-specific eGenes we report may actually be present more broadly in GTEx tissues but missed by the linear QTL model used in GTEx. Among the 1403 eGenes called by FastQTL, 208 were tissue-specific to IPSDSNs.

QTL overlap with GWAS catalog

The GWAS catalog was downloaded from https://www.ebi.ac.uk/gwas/ on 2016-5-08. To determine overlap between variants in the GWAS catalog and our lead QTLs, we first extracted all lead variants (both QTLs and GWAS catalog variants) from the full VCF file. We used vcftools v0.1.14 (Danecek and Seriani 2008) to compute the correlation R^2 between all lead variants within 500 kb of each other among our samples. We determined overlap separately for eQTLs, sQTLs, and ATAC QTLs, and retained only overlaps with $R^2 > 0.8$ between lead variants. Note that a given GWAS variant may be in LD with an eQTL for more than one gene, and vice versa, an eQTL for a single gene may be in LD with more than one GWAS catalog entry.

We used QTL-GWAS overlap for two purposes: first, to find individual cases where a QTL is a strong candidate as a causal association for the GWAS trait, and second, to determine whether any GWAS catalog traits are enriched overall for overlap with sensory neuron QTLs. For the first goal, we considered all overlaps with GWAS catalog associations having $p < 5x10^{-8}$, i.e. did not filter any redundant overlaps. These overlaps are reported in Supplementary Tables 20 (for eQTLs), 21 (for sQTLs), and 22 (for ATAC QTLs).

To determine whether our QTL overlaps were enriched in any specific GWAS catalog traits relative to other traits, we computed overlap with all GWAS catalog SNPs ($p < 5x10^{-8}$) but

sought to eliminate redundant overlaps. For traits that were reported with differing names (e.g. “Alzheimer's disease (cognitive decline)” and “Alzheimer's disease in APOE e4-carriers”), we grouped these into a single trait name (e.g. “Alzheimer's disease”). We then sorted overlaps by decreasing LD R^2, and kept the single overlapping QTL with the highest R^2 for each GWAS catalog entry. Similarly, we removed duplicates with the same reported GWAS catalog SNP and trait, such as when successive GWAS of the same trait report the same SNP association. We counted the number of such unique GWAS-QTL overlaps separately for eQTLs, sQTLs, and caQTLs, and we report these in Table 1. To avoid bias due to correlation between GWAS power and LD patterns, we restricted our analysis to the 41 traits with at least 40 GWAS catalog associations. We then considered the binomial probability of the observed overlap with each trait, with the expected overlap frequency being the proportion of QTL overlaps among all trait associations (6.2%). After correcting for multiple testing, no traits showed significantly greater overlap with our QTL catalog than other traits.

To test for overall enrichment of QTL overlapping with GWAS catalog SNPs, we downloaded the 1000 genomes VCF files (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and subsetted these to the EUR samples. We used vcftools to identify all SNPs in LD $R^2 > 0.8$ with a GWAS catalog SNP and removed duplicate SNPs. We used our IPSDSN eQTL lead SNPs as input to SNPsnap (https://data.broadinstitute.org/mpg/snpsnap/), and computed 1000 random sets of SNPs using default parameters to match for LD partners, MAF, gene density, and distance to nearest gene. We determined the number of occurrences of eQTL lead SNPs in the GWAS catalog SNP + LD partners, and did the same for the 1000 matched SNP sets. The IPSDSN eQTL lead SNPs had more overlaps (92) than any of the matched sets (median: 58, range 37-87). Note that this number of overlaps is fewer than the number we report in Supplementary Table 20; this is because we detect more overlaps when using LD from our own samples than when using 1000 genomes LD patterns, which is expected since 1000 genomes EUR LD does not perfectly reflect LD in our data. We performed the same overlapping process for lead eQTL SNPs from each GTEx tissue, and plotted the number of overlaps per tissue in Supplementary Figure 25.

References

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (September): 195.

Danecek, P., and G. Seriani. 2008. “An Efficient Parallel Chebyshev Pseudo-Spectral Method for Large Scale 3D Seismic Forward Modelling.” In *70th EAGE Conference and Exhibition Incorporating SPE EUROPEC 2008*. doi:10.3997/2214-4609.20147862.

Davis, Joe R., Laure Fresard, David A. Knowles, Mauro Pala, Carlos D. Bustamante, Alexis Battle, and Stephen B. Montgomery. 2016. “An Efficient Multiple-Testing Adjustment for

eQTL Studies That Accounts for Linkage Disequilibrium between Variants." *American Journal of Human Genetics* 98 (1): 216–24.

Hansen, Kasper D., Rafael A. Irizarry, and Zhijin Wu. 2012. "Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization." *Biostatistics* 13 (2): 204–16.

Jun, Goo, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. 2012. "Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data." *American Journal of Human Genetics* 91 (5): 839–48.

Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. "Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs." *Nature* 546 (7658): 370–75.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. "Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq." *Nature Genetics* 48 (2): 206–13.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014. "CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies." *Bioinformatics* 30 (7): 1006–7.