

Supplementary Information

Supplementary Figures

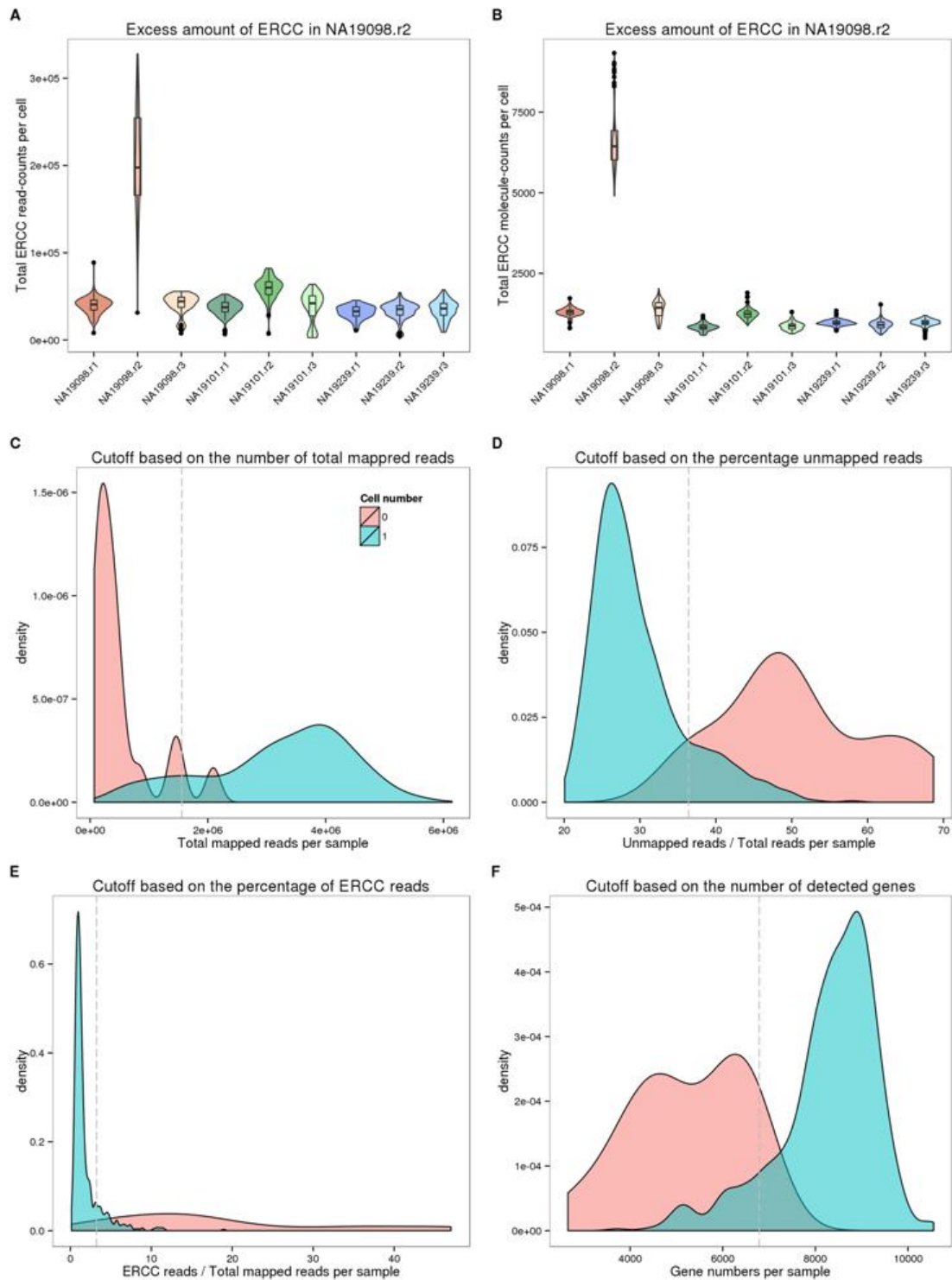


Figure S1. Removal of low quality samples.

Supplementary Figure S1. Removal of low quality samples. Violin plots of the total read-counts of ERCC spike-in controls in (A) and the total molecule-counts in (B) in single cell samples. The three colors represent the three individuals (NA19098 in red, NA19101 in green, and NA19239 in blue). (C-F) Density plots of the distributions of the total mapped reads in (C), the percentage of unmapped reads in (D), the percentage of ERCC reads in (E), and the number of detected genes in (F). The dash lines indicate the cutoffs based on the 95th percentile of the samples with no cells.

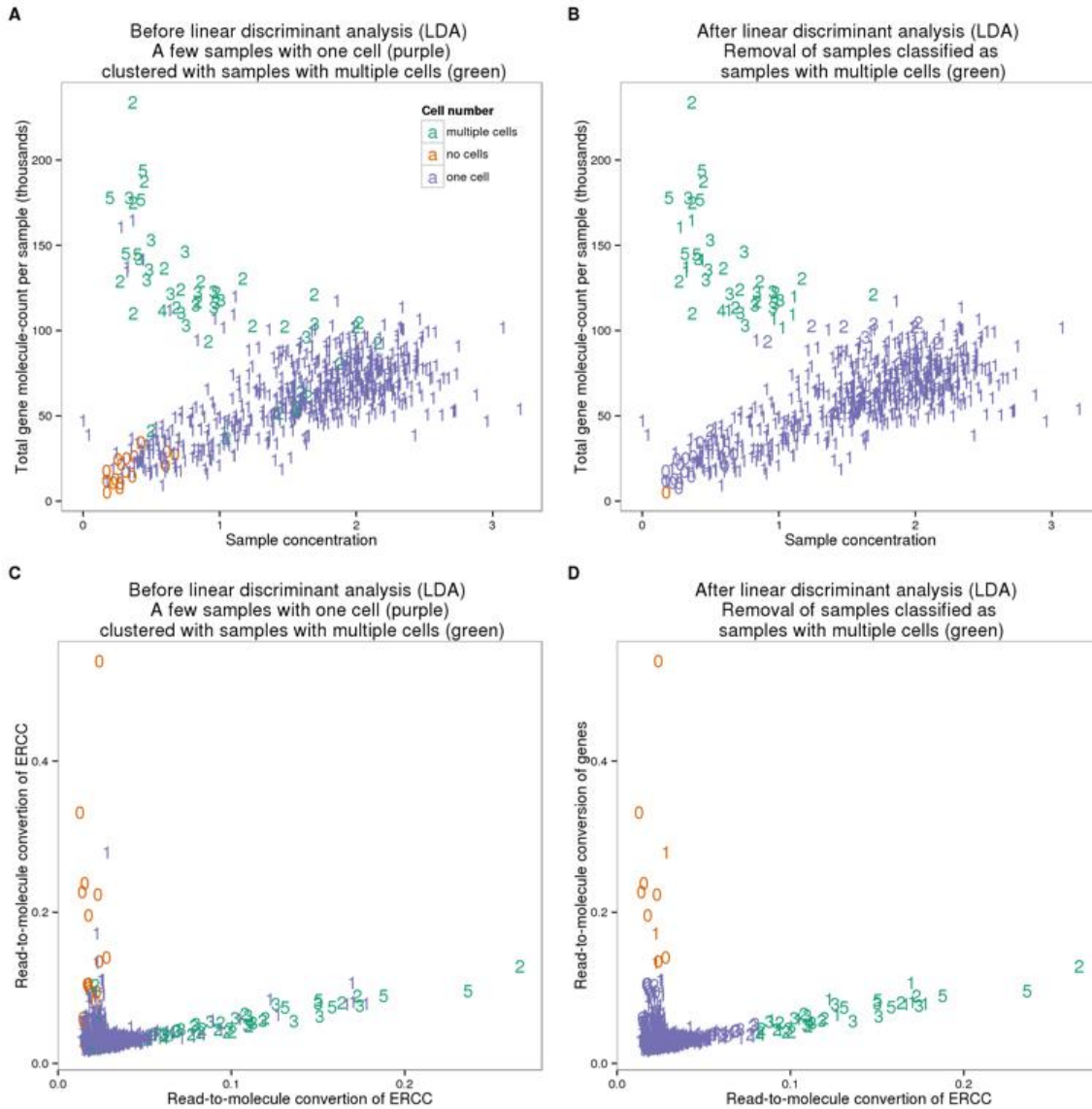


Figure S2. Removal of samples with multiple cells.

Supplementary Figure S2. Removal of samples with multiple cells. Scatterplots of the three groups of samples (no cell in green, single-cell in orange, and two or more cells in purple) before (A) and after (B) the linear discriminant analysis (LDA) using sample concentration of cDNA amplicons (ng/ l) and the number of detected genes. (C and D) Similarly, LDA was

performed to identify potential multi-cell samples using the read-to-molecule conversion efficiency (total molecule-counts divided by total read-counts per sample) of endogenous genes and ERCC spike-in controls. Scatterplots of before and after the LDA in (C) and (D), respectively. The numbers indicate the number of cells observed in each cell capture site.

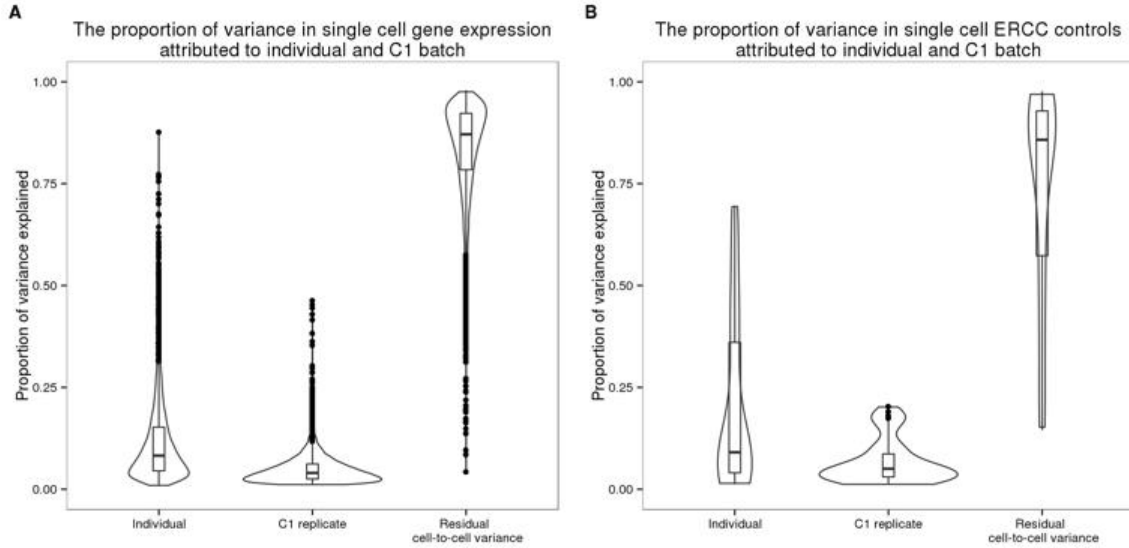


Figure S3. Sources of cell-to-cell variance in per-gene expression profile.

Supplementary Figure S3. Sources of cell-to-cell variance in per-gene expression profile. Violin plots of the proportion of per-gene cell-to-cell variance that was due to individual sample of origin, different C1 replicates, and other single cell sample differences. These results were calculated from the molecule counts before normalization and batch correction. Endogenous genes are shown in (A) and the ERCC spike-in controls in (B).

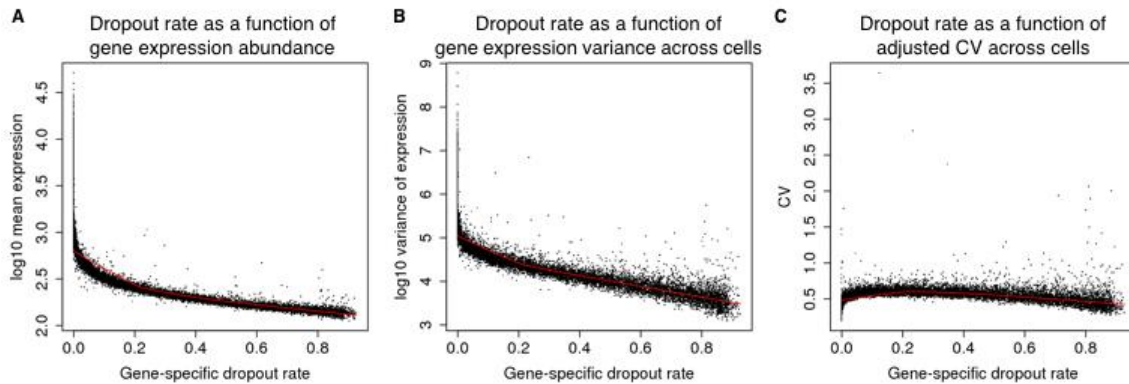


Figure S4. The gene-specific dropout rate

Supplementary Figure S4. The gene-specific dropout rate. The gene-specific dropout rate (the proportion of cells in which the gene is undetected) and its relationship with \log_{10} mean expression in (A), with \log_{10} variance of expression in (B), and with the CV in (C) of the cells in which the gene is expressed (cells in which at least one molecule of the given gene was detected). Each point represents a gene, and red lines indicate the predicted values using locally weighted scatterplot smoothing (LOESS).

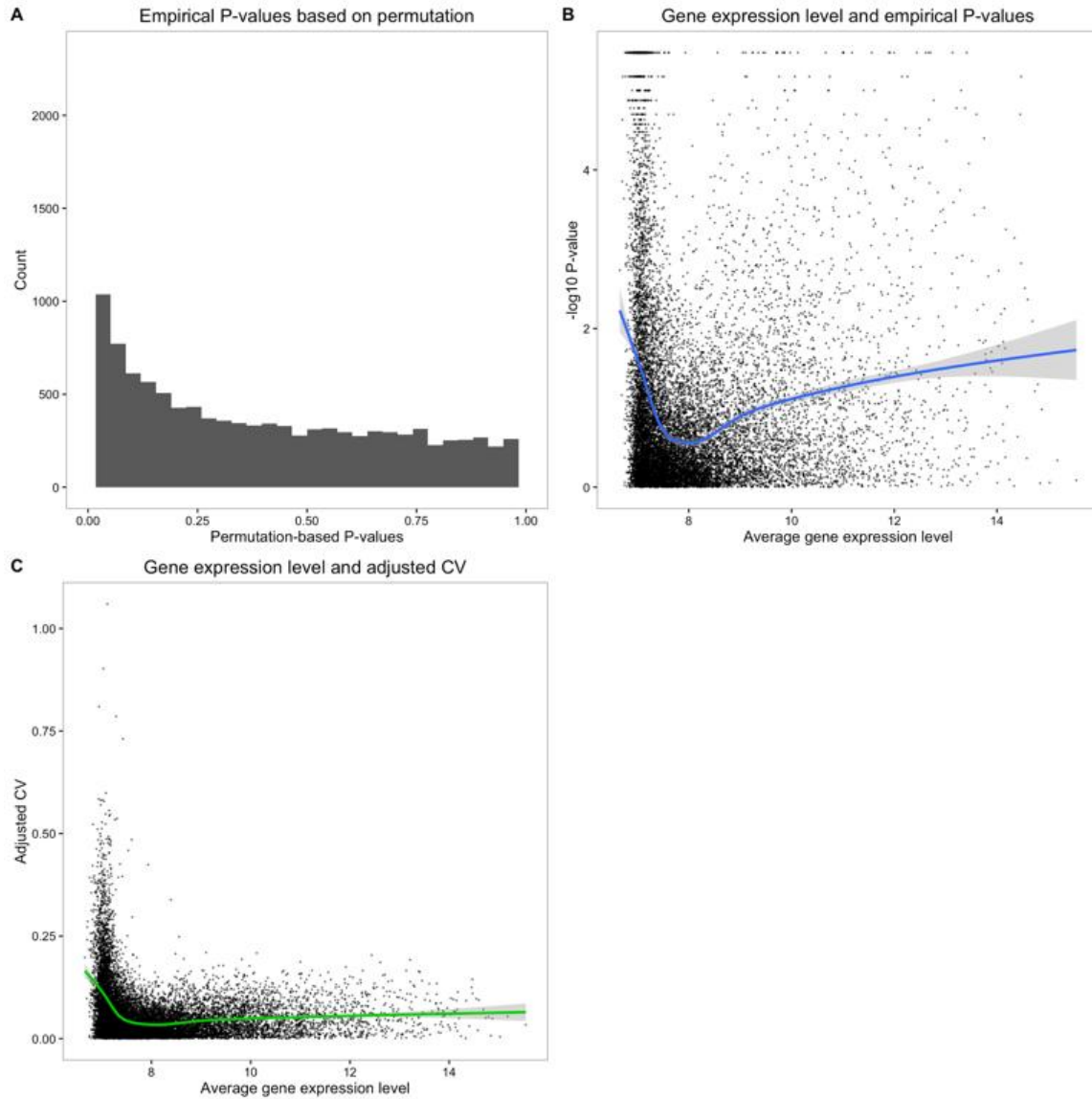
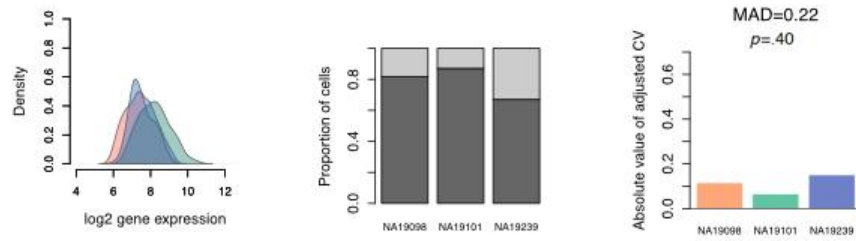


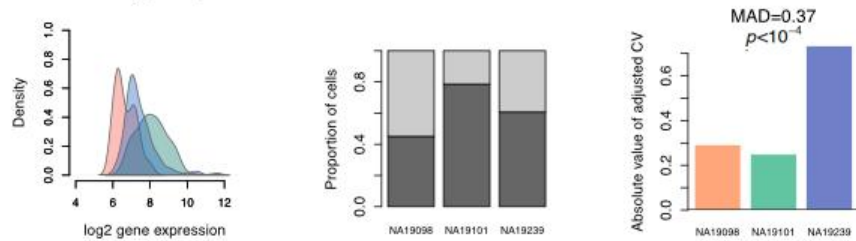
Figure S5. Permutation-based P -value.

Supplementary Figure S5. Permutation-based P -value. (A) Histogram of empirical P -values based on 300,000 permutations. (B) $-\log_{10}$ empirical P -values are plotted against average gene expression levels. Blue line indicates the fitted relationship between $-\log_{10}$ P -values and average \log_2 gene expression levels of cells that were detected as expressed, using locally weighted scatterplot smoothing (LOESS). (C) Median of Absolute Deviation (MAD) of genes versus average gene expression levels. Green line indicates the fitted relationship (LOESS) between the MAD values and average \log_2 gene expression levels of cells in which the gene was detected as expressed.

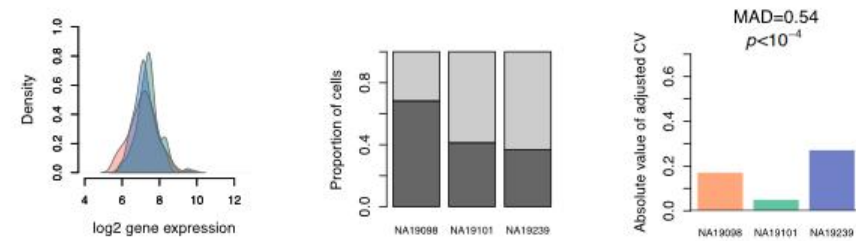
A. *NANOG* is not detected in all cells and exhibited no differential CV across individuals



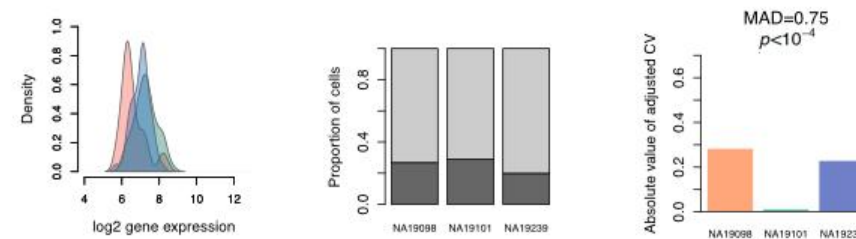
B. *TFPI* has higher expression level and variance but lower CV in individual NA19101



C. *SLC25A16* has similar expression level and significantly different CVs across individuals



D. *SLC9A5* has similarly high dropout rate and significantly different CVs across individuals



E. *SREK1IP1* has high expression level, low dropout rate, and significantly different CVs across individuals

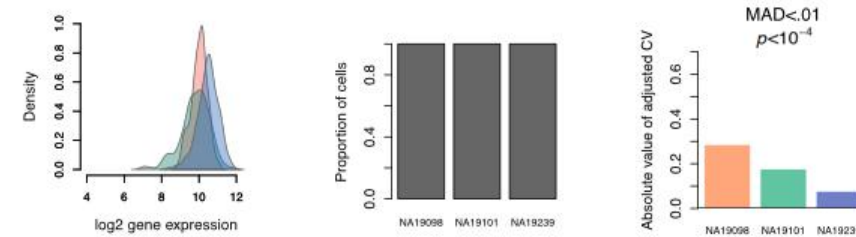


Figure S6. Inter-individual differences in regulatory noise.

Supplementary Figure S6. Inter-individual differences in regulatory noise. These 5 example genes illustrate various patterns of cell-to-cell gene expression variance. For each gene, the left panel shows the distribution of the log₂ gene expression levels (considering only cells in which the gene is detected as expressed), the middle panel shows the proportion of cells

in which the gene is detected as expressed (dark grey) and the dropout rate (light grey) for each individual, and the right panel shows the absolute value of adjusted CV for each individual, along with the corresponding gene-specific MAD (median of absolute deviation) value and *P*-value. The three colors in the upper and lower panel represent the individuals (NA19098 in red, NA19101 in green, and NA19239 in blue).

The gene expression level of pluripotency genes in single cell samples from the three individuals

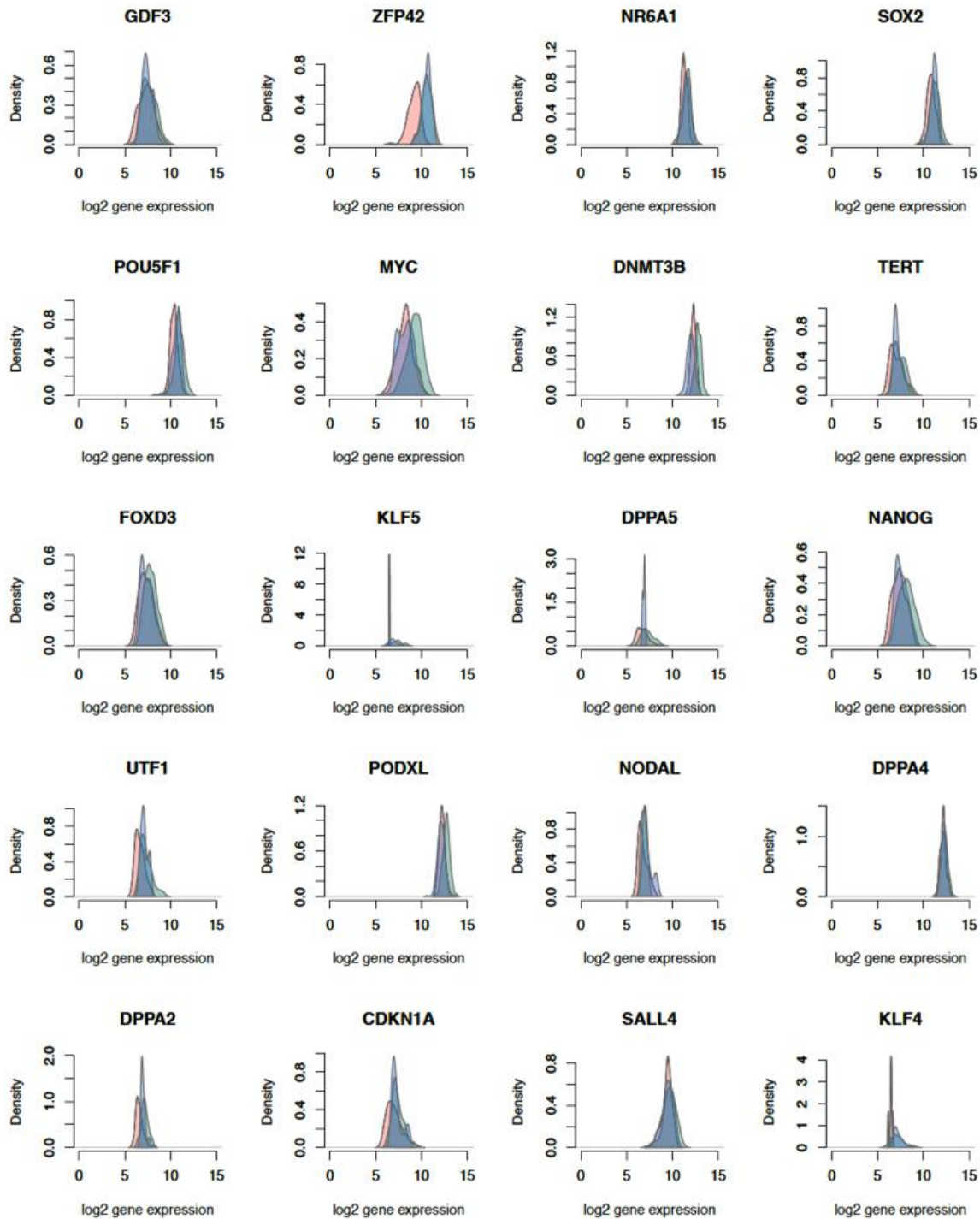


Figure S7. Cell-to-cell variation of pluripotency genes.

Supplementary Figure S7. Cell-to-cell variation of pluripotency genes. Density plots of the distribution of \log_2 gene expression of key pluripotency genes across all single cells by individual. The peaks with lower gene expression values (\log_2 around 4) represent the cells in

which the gene is undetected. The three colors represent the three individuals (NA19098 is in red, NA19101 in green, and NA19239 in blue).

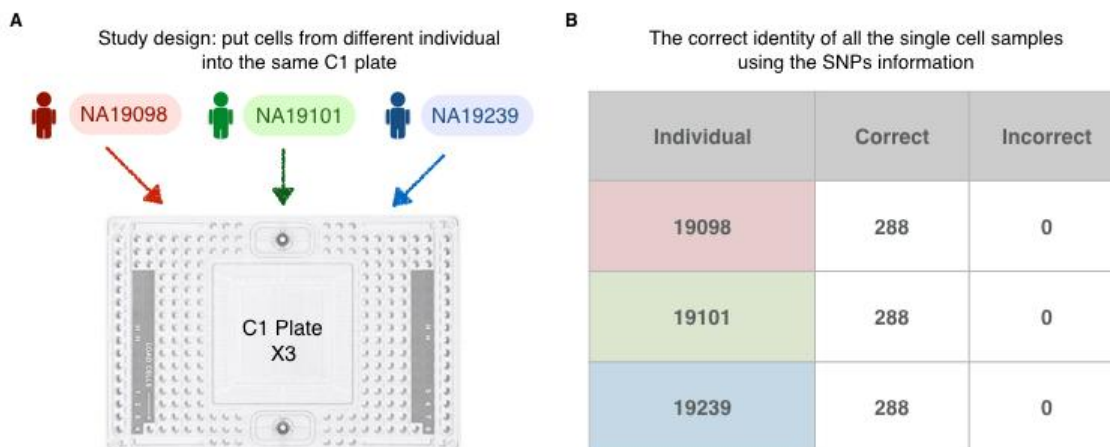


Figure S8. Proposed study design for scRNA-seq using C1 platform.

Supplementary Figure S8. Proposed study design for scRNA-seq using C1 platform. (A) A balanced study design consisting of multiple individuals within a C1 plate and multiple C1 replicates to fully capture the batch effect across C1 plates and further retrieve the maximum amount of biological information. (B) The correct identity of each single cell sample was determined by examining the SNPs present in their RNA sequencing reads.

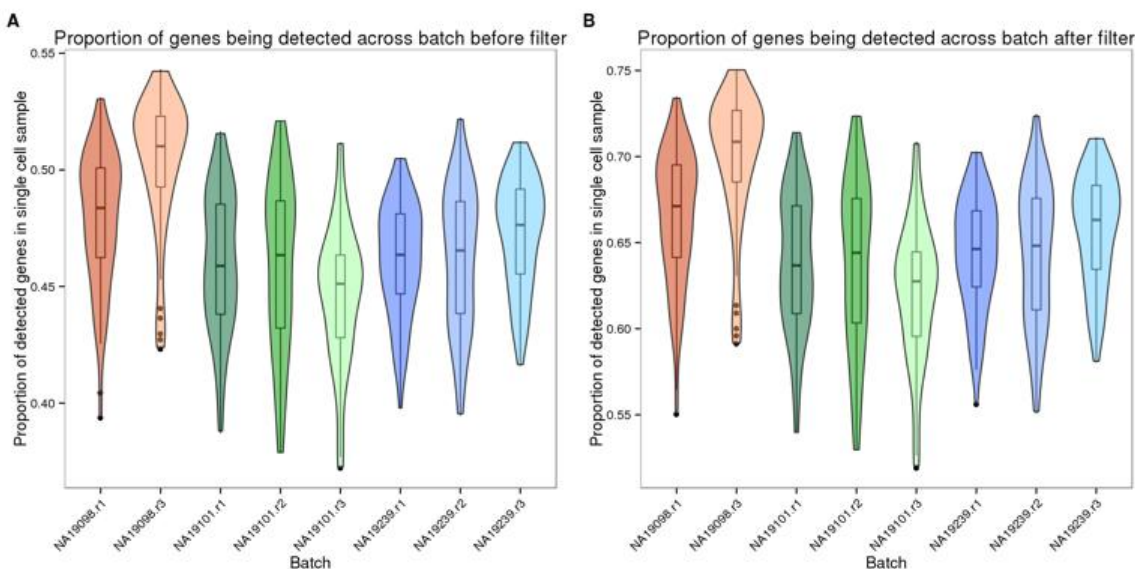


Figure S9. The proportion of genes detected in single cell samples.

Supplementary Figure S9. The proportion of genes detected in single cell samples. Violin plots of the proportion of genes detected, computed by the total number of detected genes in each single cell divided by the total number of genes detected across all single cells, before in

(A) and after in (B) the removal of genes with low expression. The three colors represent the three individuals (NA19098 is in red, NA19101 in green, and NA19239 in blue).

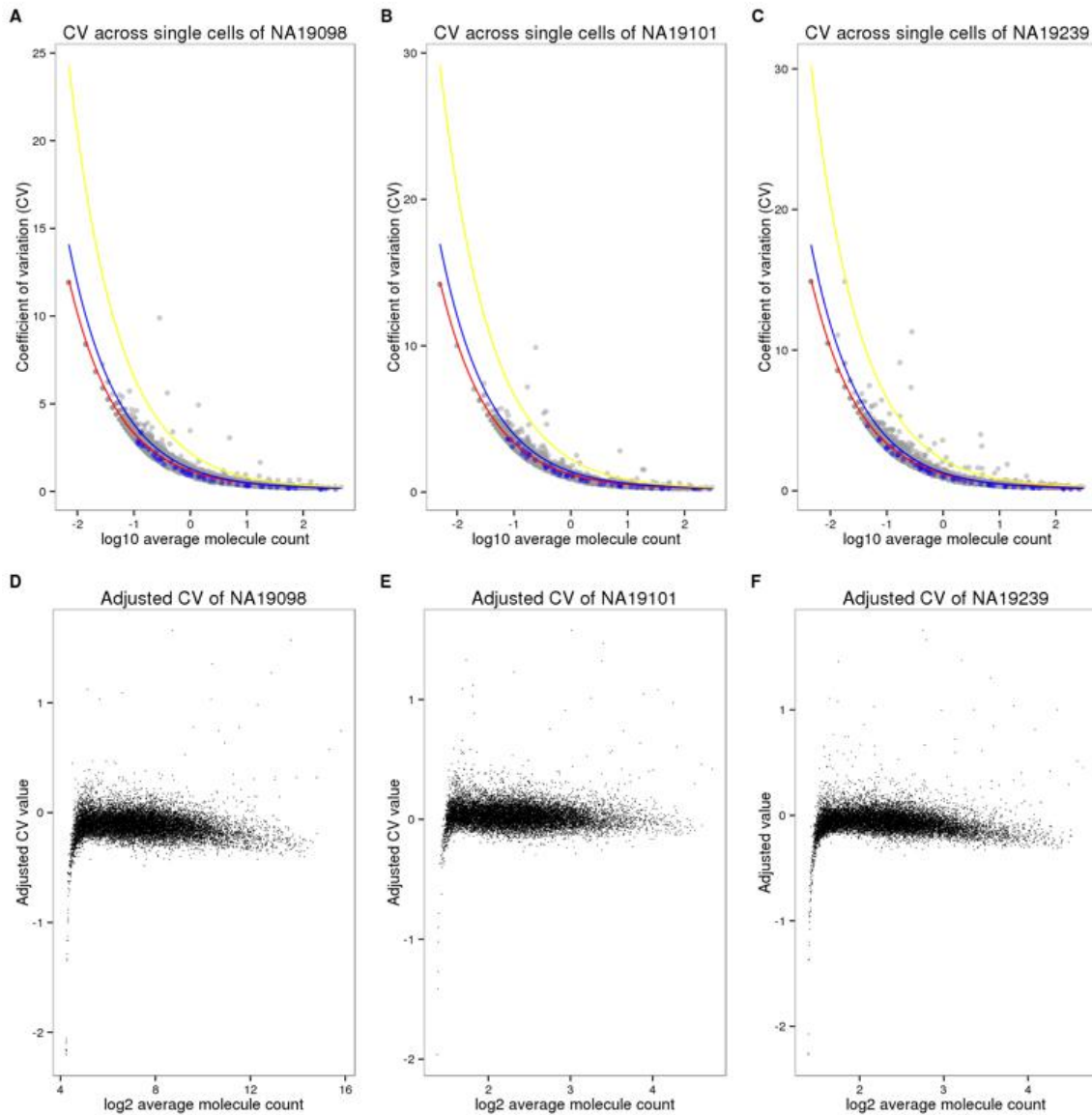


Figure S10. Coefficients of variation (CV) before and after adjusting for gene mean abundance.

Supplementary Figure S10. Coefficients of variation (CV) before and after adjusting for gene mean abundance. (A-C) CV plotted against average molecule counts across all cells for each individual²⁰. Grey points represent endogenous genes, and blue points represent ERCC spike-in controls. The curves indicate the expected CV under three different scenarios. Red curve depicts the expected CV of the endogenous genes while assuming a Poisson distribution with no over-dispersion. Likewise, blue curve depicts the expected CVs of the ERCC spike-in controls under the Poisson assumption. Yellow curve depicts the expected CVs of an over-dispersed Poisson distribution for which standard deviation is three times the ERCC spike-in

controls. (D-F) Adjusted CV values of each gene including all cells are plotted against \log_{10} of the average molecule counts for each individual.

Supplementary Tables

Supplementary Table S1. Data collection.

(A) iPSCs were sorted using the 10-17 μ m IFC plates with the staining of the pluripotency marker, TRA1-60. Single cell occupancy is the percentage of occupied capture sites containing one single cell. The average cDNA concentration was measured by the HT DNA high sensitivity LabChip (Caliper). (B) The 96 single cell libraries from one C1 plate were pooled and sequenced in three HiSeq lanes. The pooled samples were assigned across the four 8-lane flowcells.

A

Information and results of each C1 collections

Cell line	Passage	Input viability	No cell	Multiple cells	% single cell occupancy	TRA1-60 negative	% TRA1-60	Date	Ave cDNA con (ng/ul)	Replicate for seq
19098	12+15	89%	2	2	95.83	2	97.92	11/07/2014	N/A	(1)
19098	12+18	90%	2	4	93.75	0	100.00	11/14/2014	1.88	(2)
19098	12+20	83%	3	10	86.46	0	100.00	11/22/2014	1.40	(3)
19101	12+16	70%	2	3	94.79	9	90.63	11/13/2014	1.81	(1)
19101	12+19	94%	5	3	91.67	0	100.00	11/23/2014	1.38	(2)
19101	12+19	69%	1	19	79.17	0	100.00	11/24/2014	1.26	(3)
19239	12+16	85%	1	2	96.88	4	95.83	11/11/2014	1.60	(1)
19239	12+18	75%	1	6	92.71	5	94.79	11/17/2014	1.55	(2)
19239	12+19	93%	5	7	87.50	2	97.92	11/21/2014	1.70	(3)

B

The arrangement of samples for sequencing on four flowcells

Flowcell 1	Flowcell 2	Flowcell 3	Flowcell 4
Bulk	19098 (2)	19098 (3)	19239 (1)
19098 (1)	19239 (3)	19101 (1)	19101 (2)
19239 (2)	19098 (1)	19098 (2)	19098 (3)
19101 (3)	19239 (2)	19239 (3)	19101 (1)
19239 (1)	19101 (3)	Bulk	19098 (2)
19101 (2)	19239 (1)	19098 (1)	19239 (3)
19098 (3)	19101 (2)	19239 (2)	Bulk (all 9)
19101 (1)	Bulk	19101 (3)	

Table S1. Data collection.

Supplementary Table S2. High quality single cell samples.

List of the 564 high quality single cell samples.

Supplementary Table S3. Genes associated with inter-individual differences in regulatory noise.

List of genes that we classified the estimates of regulatory noise as significantly different across individuals (empirical permutation $P < 10^{-4}$). There are a total of 560 genes.

Supplementary Table S4. Gene ontology analysis of the genes associated with inter-individual differences in regulatory noise.