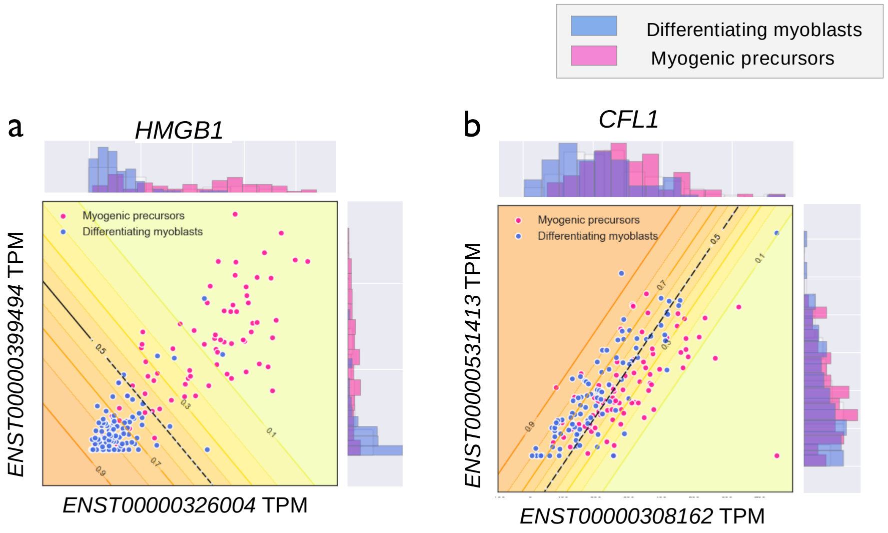
nature methods SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41592-018-0303-9

In the format provided by the authors and unedited.

A discriminative learning approach to differential expression analysis for single-cell RNA-seq

Vasilis Ntranos[1,2,7], Lynn Yi[3,4,7], Páll Melsted[5] and Lior Pachter[4,6]*

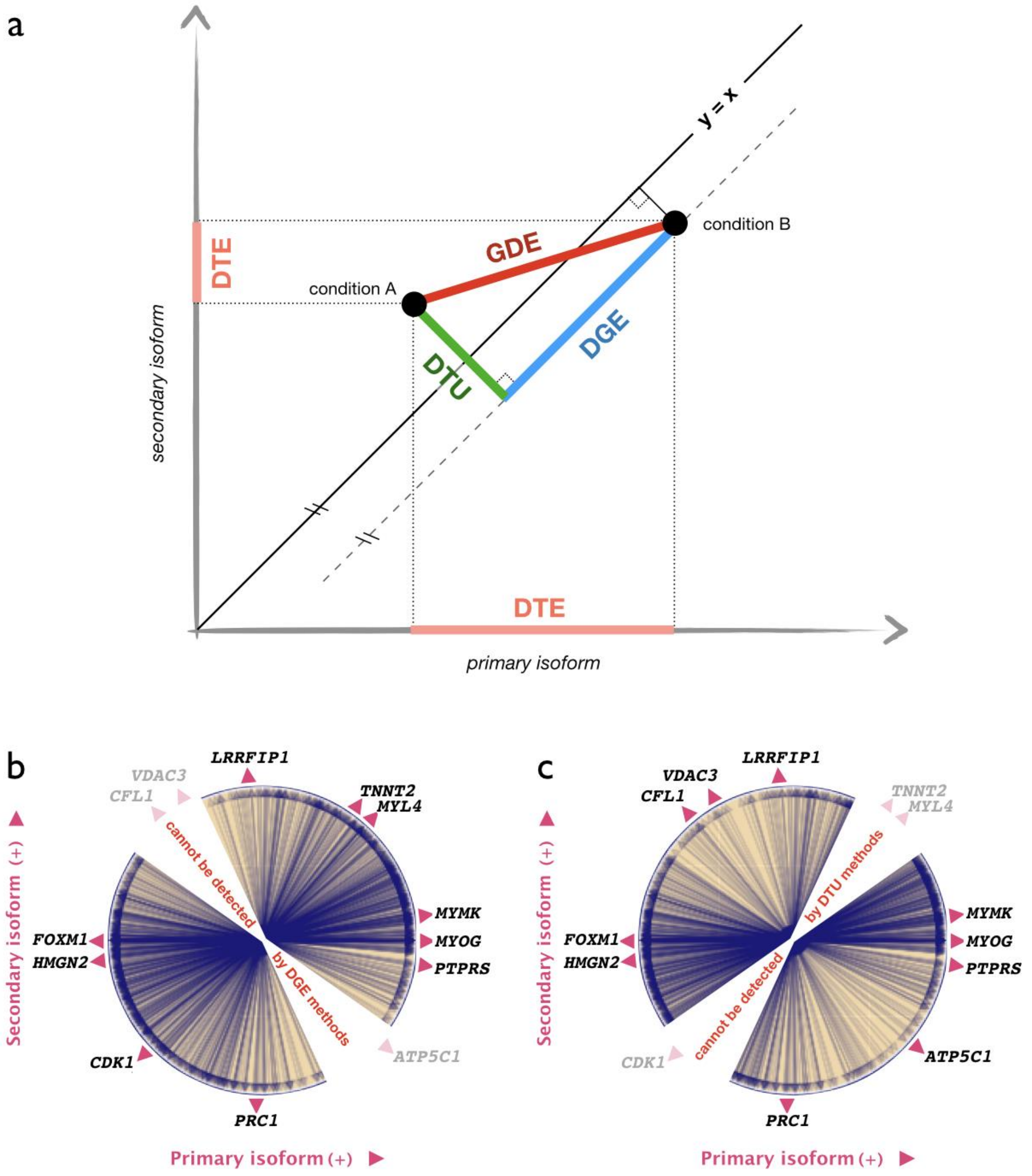
[1]Department of Electrical Engineering & Computer Science, UC Berkeley, Berkeley, CA, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [3]UCLA–Caltech Medical Science Training Program, UCLA, Los Angeles, CA, USA. [4]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [5]Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland. [6]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [7]These authors contributed equally: Vasilis Ntranos, Lynn Yi. *e-mail: lpachter@caltech.edu



Supplementary Figure 1

Examples of differential genes detected with logistic regression.

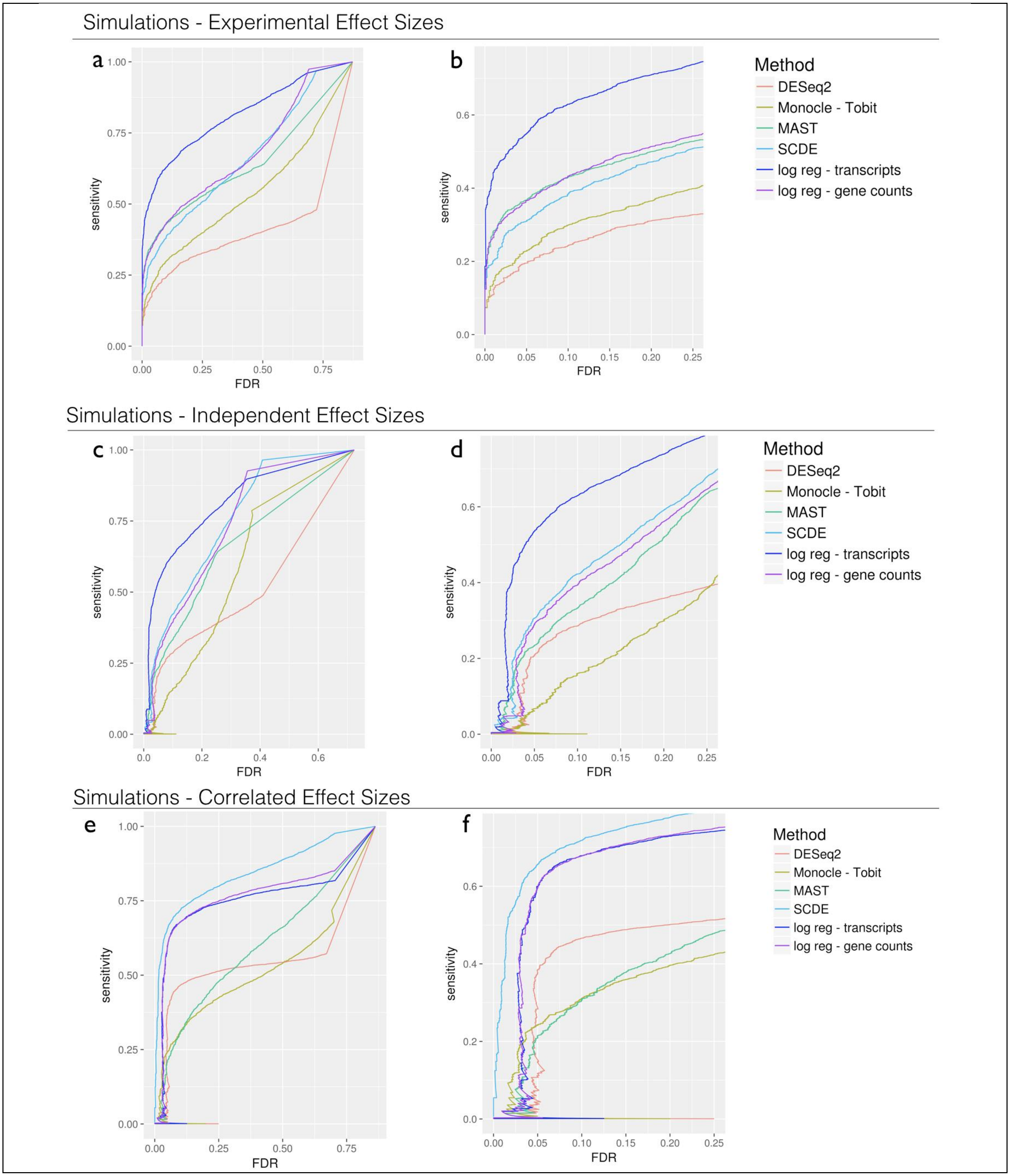
Two example genes undergo correlated changes in transcript expression **(a)** and isoform switching (**b**), respectively.



Supplementary Figure 2

Relationship between differential expression methods.

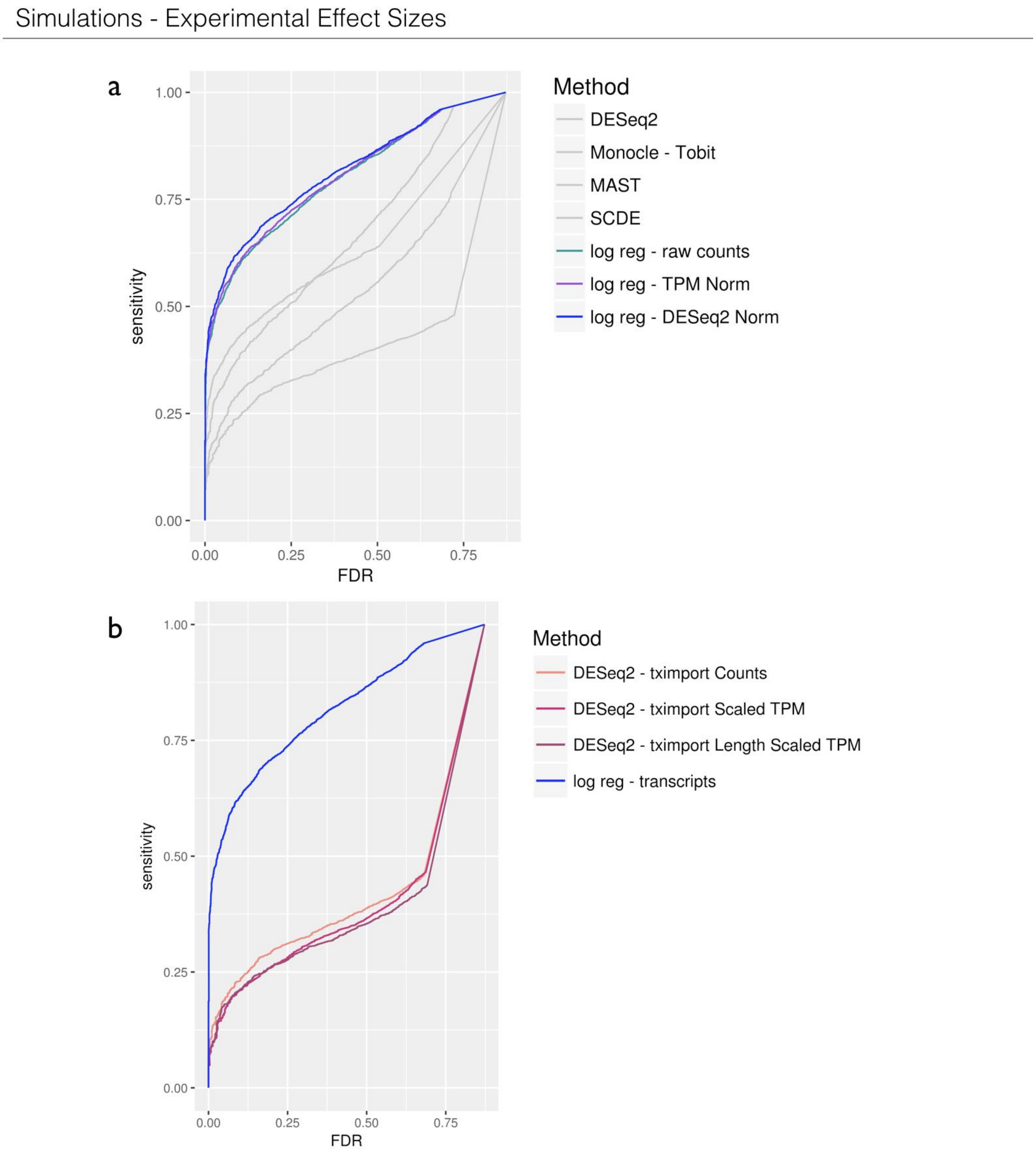
(a) Depiction of the difference in expression of a two-transcript gene in two cell types. The two black points correspond to gene expression in each of the two cell types: the x-coordinate depicts the expression of its first transcript and the y-coordinate, the expression of its second transcript. In differential transcript expression (DTE) tests, transcripts are independently assessed for differential expression, corresponding to independent testing with projections of the points onto the x-axis and y-axis (pink segments). Differential gene expression (DGE) tests are based on changes in overall gene expression; this change in overall gene abundance is proportional to the difference in the projections of the points onto the line *y=x* (blue segment). Traditional differential transcript usage (DTU) methods test for differential transcript allocation within a gene. Differences in transcript usage is proportional to differences of the projections onto the line *y=-x* (green segment), which is orthogonal to the DGE direction. Gene differential expression (GDE) is a moniker for changes between transcript abundances as reflected in the length of the line between them (red segment). Our proposed method uses logistic regression to find this line. (b) DGE methods have a “blind spot” for genes whose transcripts change only in relative abundance. Such transcripts can be detected by DTU. However, DTU has a blind spot for genes changing in overall abundance (c). Logistic regression for GDE has no blind spots, as differential analysis is performed in the detected direction of change.


Simulations - Experimental Effect Sizes
a
b
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - transcripts
log reg - gene counts
Simulations - Independent Effect Sizes
c
d
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - transcripts
log reg - gene counts
Simulations - Correlated Effect Sizes
e
f
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - transcripts
log reg - gene counts

Supplementary Figure 3

Performance of differential expression methods on simulations.

A scRNA-seq dataset containing two cell types, each with 105 cells, was simulated. In (a, b-zoomed in), effect sizes were derived from an experiment. In the independent effect size simulation (c, d), transcripts were independently chosen to be perturbed. In the correlated effect size simulation (e, f), genes were chosen independently to be perturbed, and all transcripts corresponding to the same gene were perturbed in the same direction with the same effect sizes. Four differential expression methods and three variants of logistic regression were tested on these simulations and their FDR-sensitivity plots are depicted. ‘log reg - transcripts’ is our GDE method, which performs logistic regression on the transcript quantifications. In contrast, ‘log reg - gene counts’ performs logistic regression on the summarized gene counts and is a DGE method.

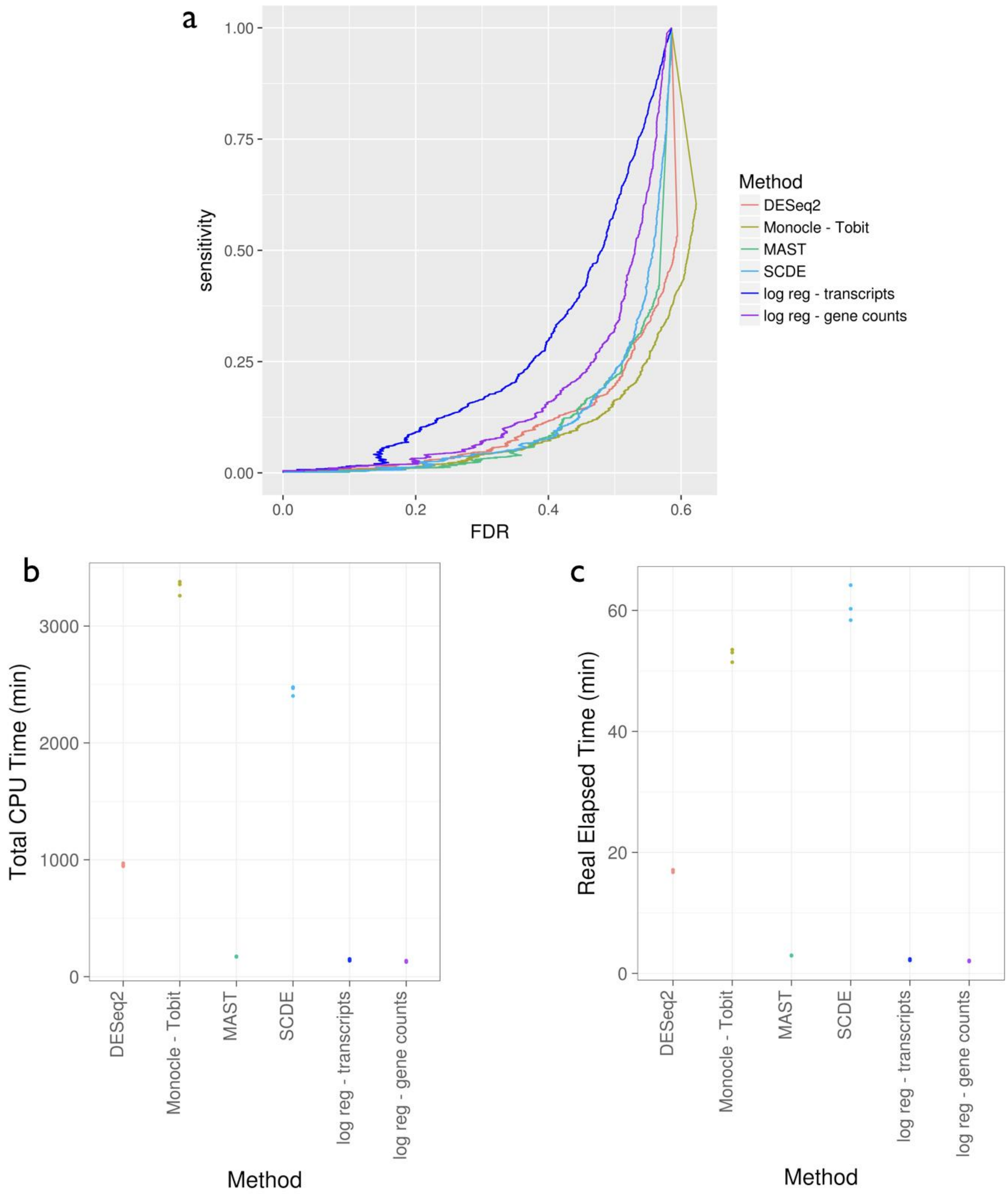


Supplementary Figure 4

Performance of logistic regression on the simulation based on experimental effect sizes.

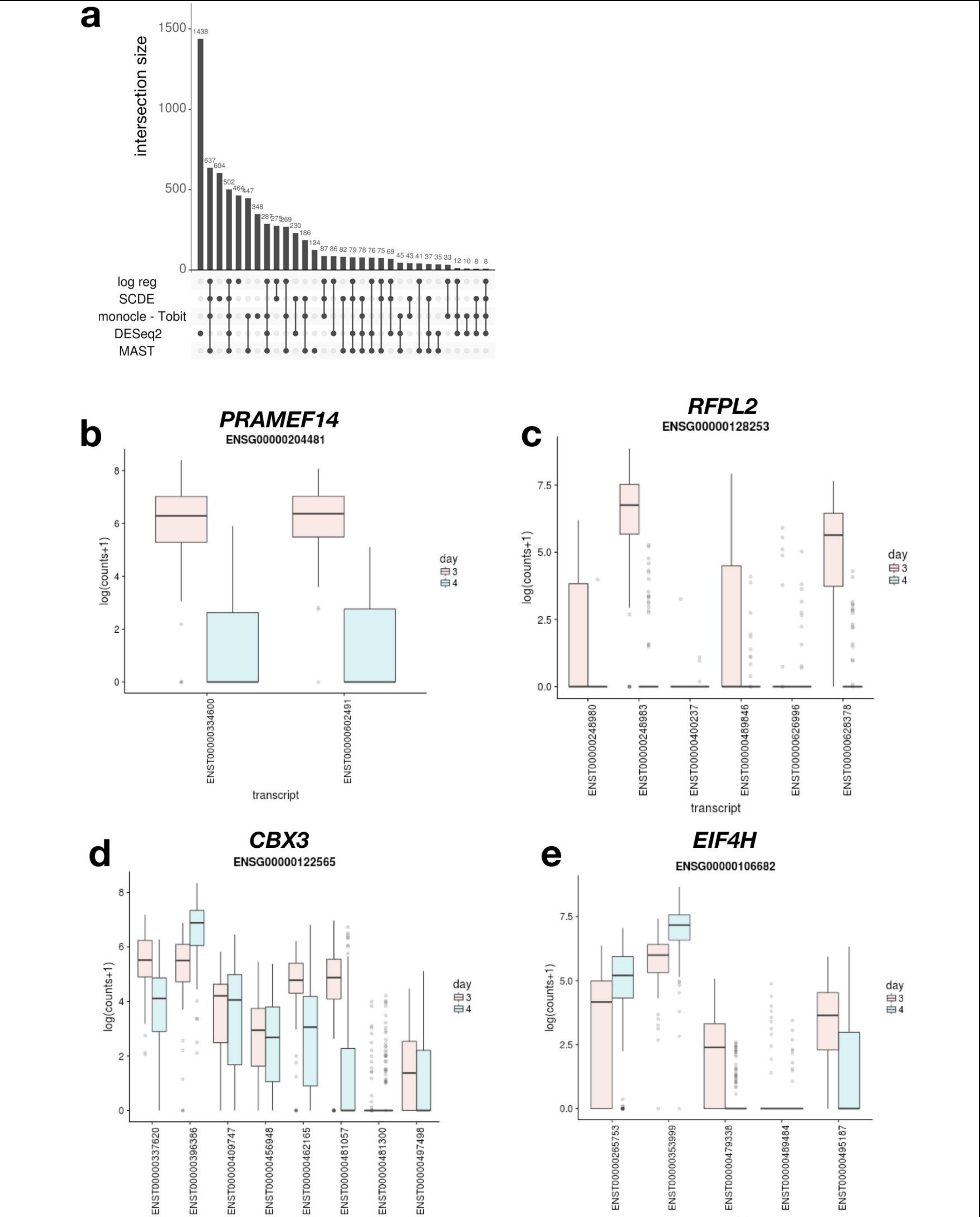
The simulation depicted in Supplementary Figure 2a,b was used to benchmark different parameters. In (a), three different normalization methods: transcript counts, size factor normalization from DESEq2, and transcript-per-million (TPM) normalization, were compared on this simulation. In (b), we compared tximport's three methods of summing transcript quantifications to gene quantifications prior to differential gene expression analysis with DESeq2 (b).

Splatter Simulations



Supplementary Figure 5

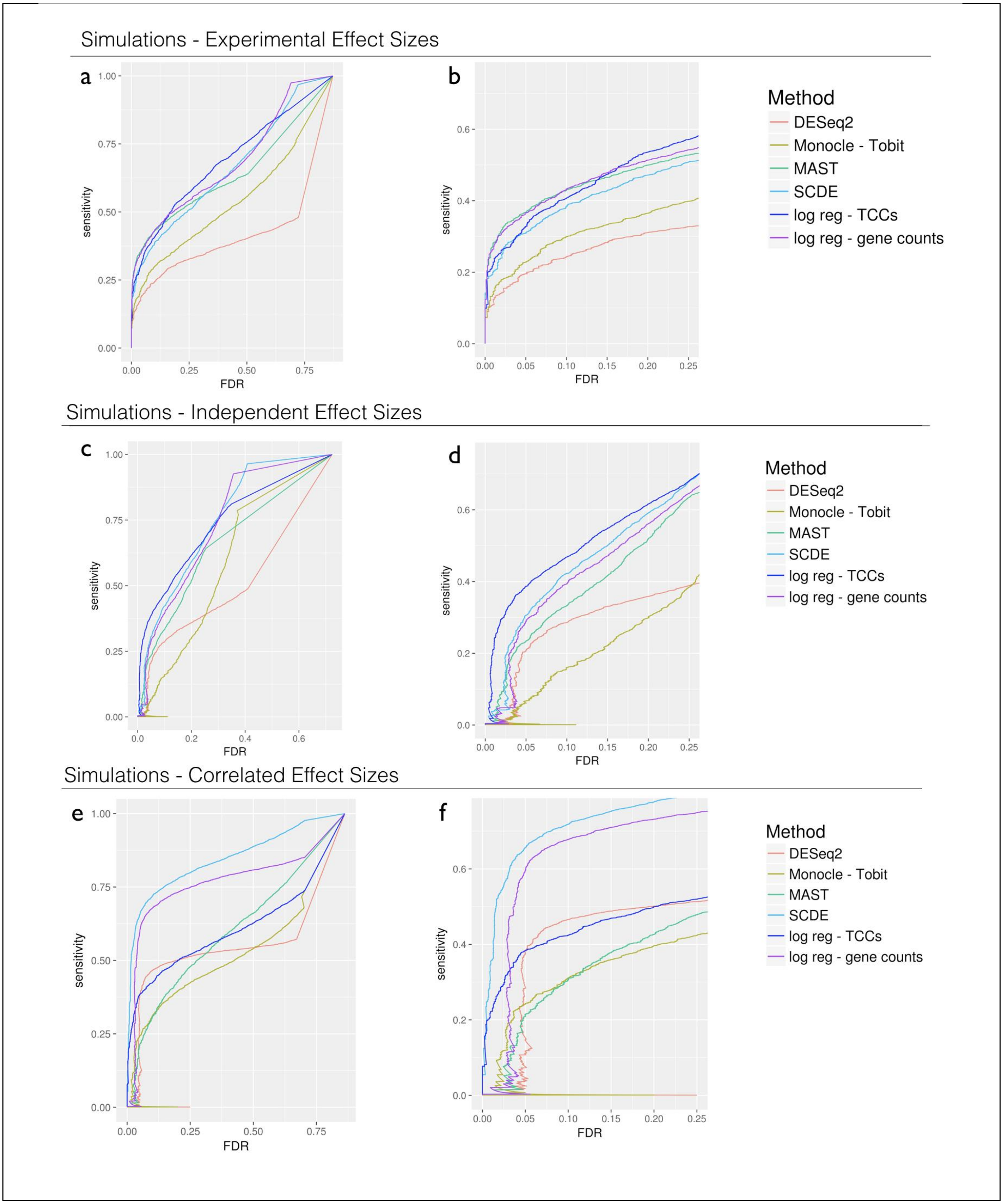
Splatter Simulations
Splatter simulations of two groups of cells, each with 10% probability of producing differential transcripts, resulting in 19% differential transcripts between the two groups. (a) sensitivity-FDR curve. (b,c) runtime benchmarks of the methods on the simulation, plotting the CPU time and the real elapsed time of three trials.


a
intersection size
1500
1000
500
0
1438
637
604
502
464
447
348
287
275
269
230
186
124
87
86
82
79
78
76
75
69
45
43
41
37
35
33
12
10
8
8
log reg
SCDE
monocle - Tobit
DESeq2
MAST
b
PRAMEF14
ENSG00000204481
log(counts+1)
8
6
4
2
0
ENST00000334600
ENST00000602491
transcript
day
3
4
c
RFPL2
ENSG00000128253
log(counts+1)
7.5
5.0
2.5
0.0
ENST00000248980
ENST00000248983
ENST00000400237
ENST00000489846
ENST00000626996
ENST00000628378
transcript
day
3
4
d
CBX3
ENSG00000122565
log(counts+1)
8
6
4
2
0
ENST00000337620
ENST00000396386
ENST00000409747
ENST00000456948
ENST00000462165
ENST00000481057
ENST00000481300
ENST00000497498
day
3
4
e
EIF4H
ENSG00000106682
log(counts+1)
7.5
5.0
2.5
0.0
ENST00000265753
ENST00000353999
ENST00000479338
ENST00000489484
ENST00000495187
day
3
4

Supplementary Figure 6

Analysis of embryonic dataset

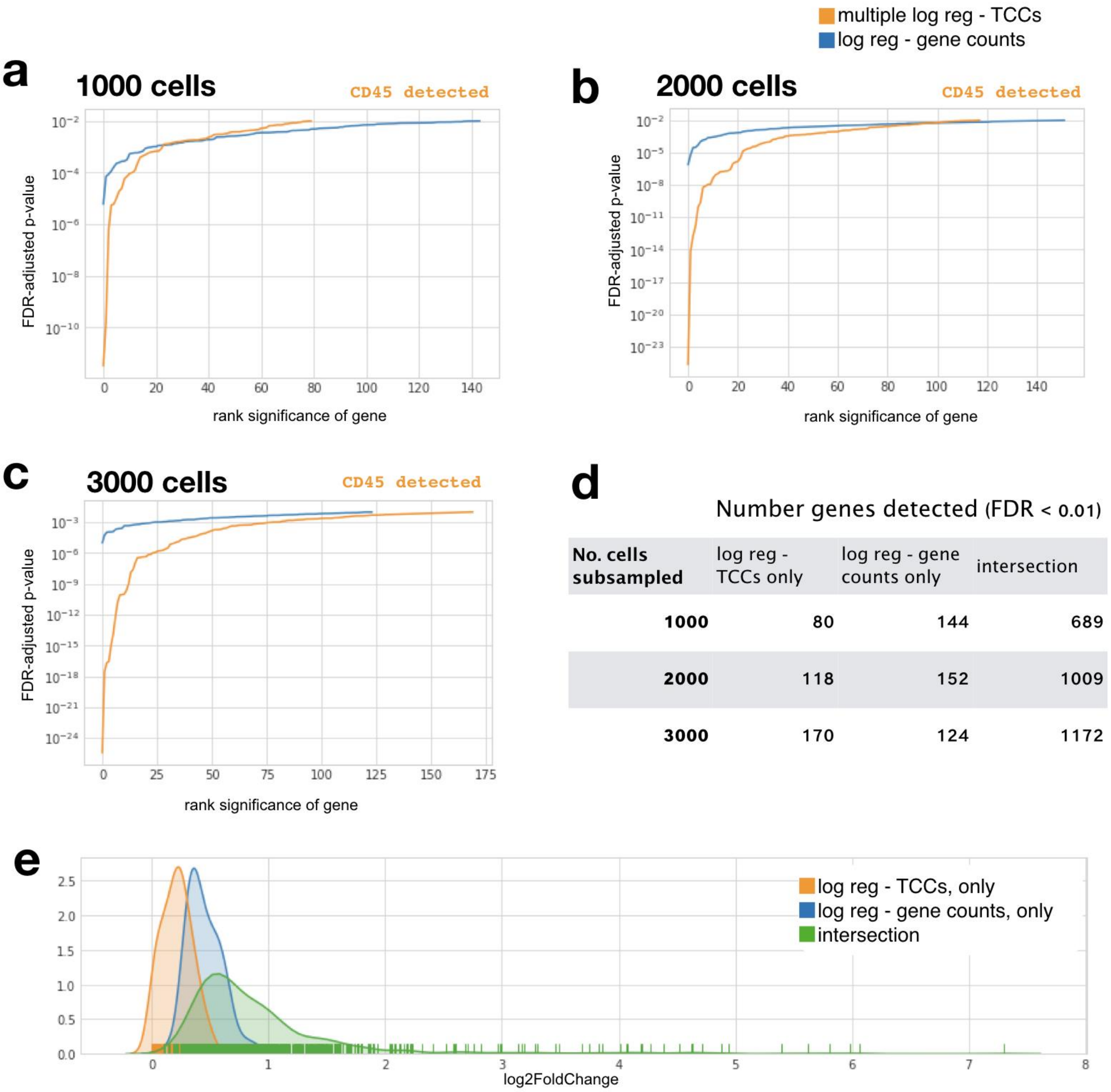
An UpSet plot showing sizes of the set intersections of the 3000 most significant genes found by each of five methods applied to day 3 (n=81 cells) and day 4 (n=190 cells) post-fertilization preimplantation human embryonic cells. (a) Transcript dynamics of two of the 502 genes in the common intersection of all five methods. (b, c) two of the 464 genes in the set unique to logistic regression. (d, e) Transcript expression in day 3 cells (n=81 cells) and day 4 cells (n=190) are visualized with boxplots, where the horizontal line denotes the median, the hinges denote the interquartile range (upper and lower hinges denote the 25 and 75 percentile, respectively), the whiskers extend to 1.5 * the interquartile range, and points correspond to cells outside the 1.5 * interquartile range.


Simulations - Experimental Effect Sizes
a
b
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - TCCs
log reg - gene counts
Simulations - Independent Effect Sizes
c
d
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - TCCs
log reg - gene counts
Simulations - Correlated Effect Sizes
e
f
sensitivity
FDR
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Method
DESeq2
Monocle - Tobit
MAST
SCDE
log reg - TCCs
log reg - gene counts

Supplementary Figure 7

Performance of logistic regression using TCCs on simulations.

On the simulation of two cell types each with n=105 cells (same simulation as Supplementary Figure 2), we benchmarked logistic regression using TCCs. In (a, b-zoomed in), effect sizes were derived from an experiment. In the independent effect size simulation (c, d), transcripts were independently chosen to be perturbed. In the correlated effect size simulation (e, f), genes were chosen independently to be perturbed, and all transcripts corresponding to the same gene were perturbed in the same direction with the same effect sizes.



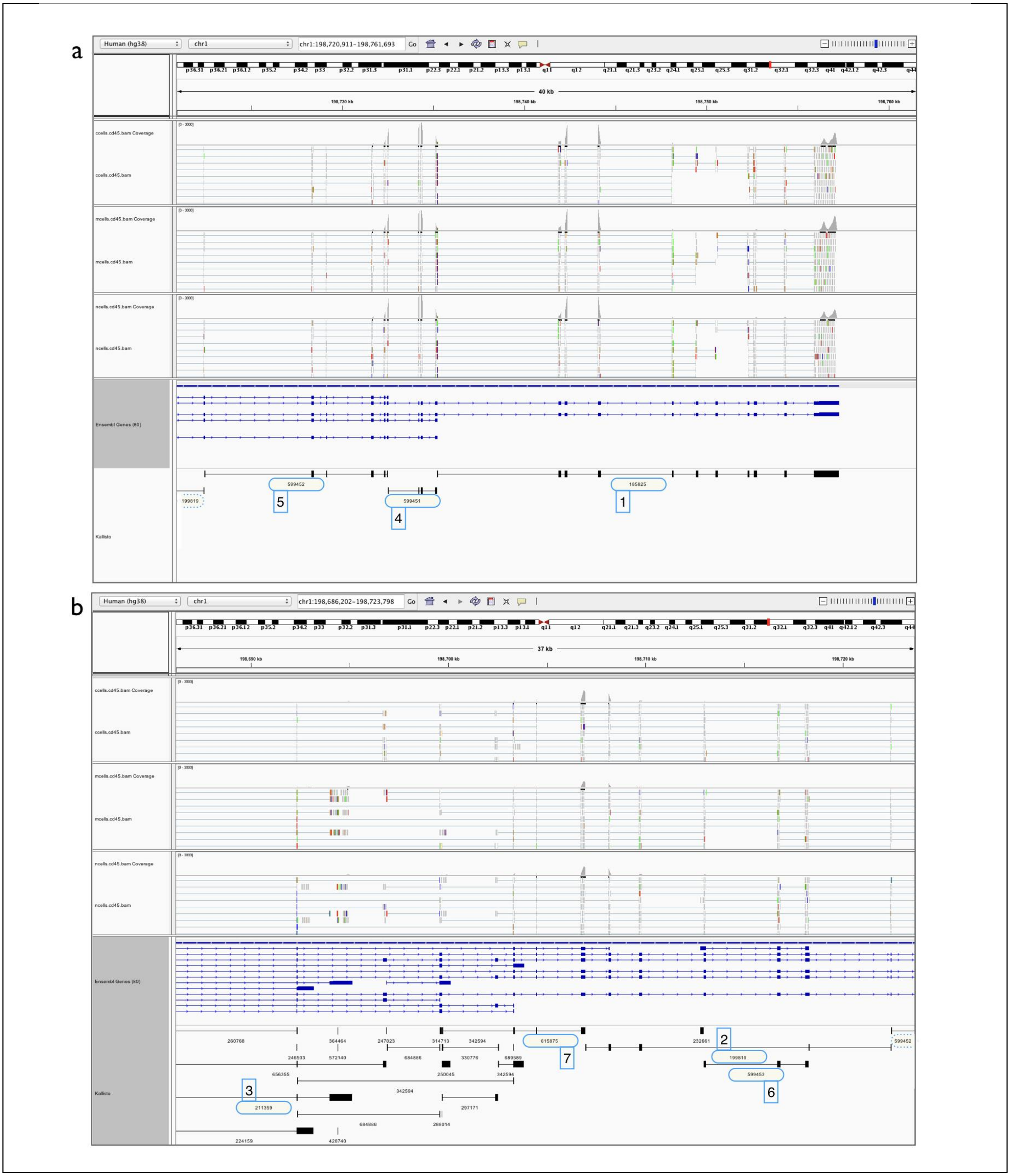
Number genes detected (FDR < 0.01)

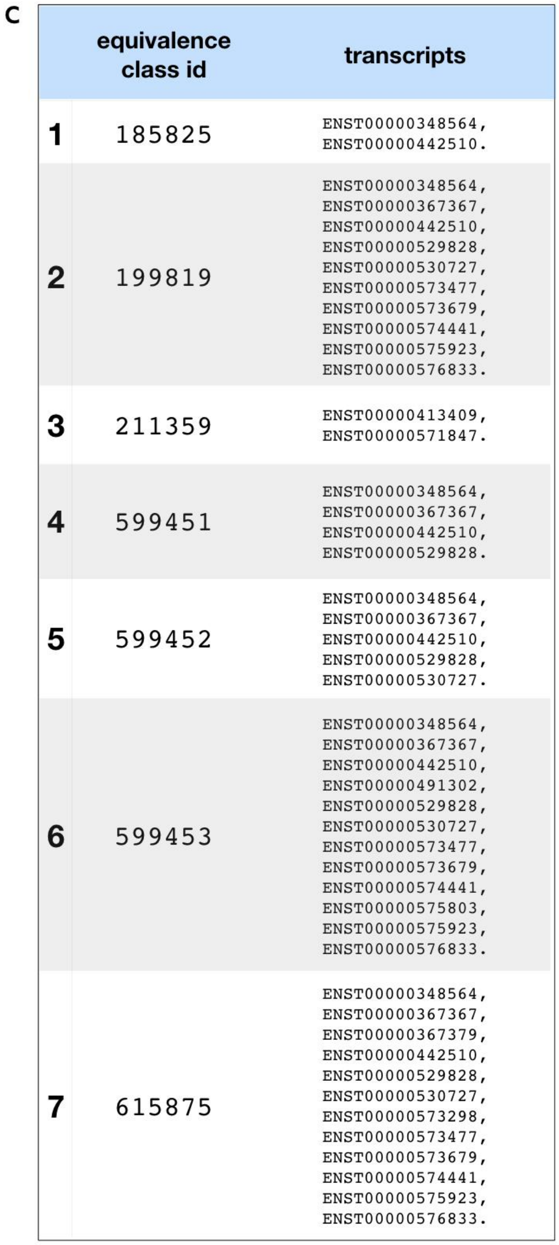
No. cells subsampled	log reg - TCCs only	log reg - gene counts only	intersection
1000	80	144	689
2000	118	152	1009
3000	170	124	1172

Supplementary Figure 8

Power analysis of CD45.

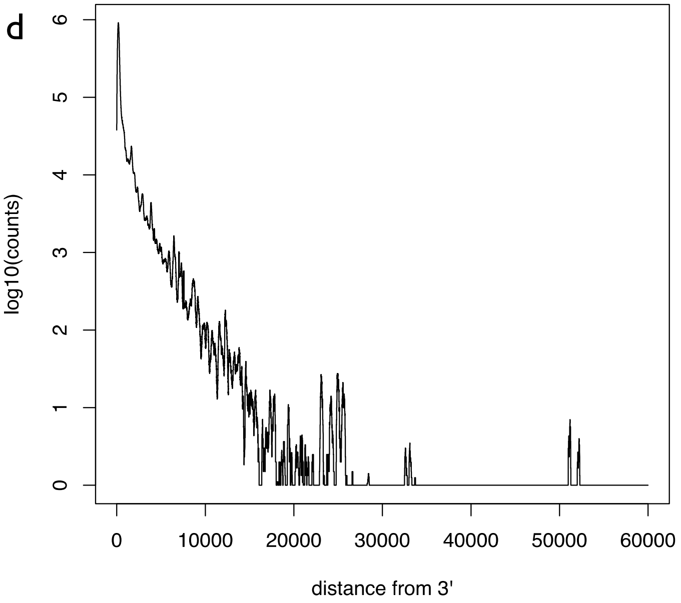
Using the PBMC dataset, we performed differential analysis between memory and naive T-cells at three levels of subsampling cells: 1000 cells (a), 2000 cells (b) and 3000 cells (c). We compared multiple logistic regression on TCCs with logistic regression using gene counts and performed Benjamini-Hochberg adjustment on p-values. At all three levels of subsampling, CD45 was found to be significant (FDR < 0.01) with logistic regression using TCCs, but not with gene counts. Furthermore, while there is a high overlap in the significant genes (FDR < 0.01) between both methods, there are genes that each method finds differential (FDR < 0.01) that the other does not (d). (e) shows the effect sizes on the overall gene counts discovered by each method uniquely compared to that in the intersection. Both methods identify genes with large effect sizes. Multiple logistic regression misses genes with small effect sizes but identifies genes with large changes in differential transcript usage.


a
Human (hg38)
chr1
chr1:198,720,911-198,761,693
Go
p36.31 p36.21 p36.12 p35.2 p34.2 p33 p32.2 p31.3 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
40 kb
198,730 kb
198,740 kb
198,750 kb
198,760 kb
ccells.cd45.bam Coverage
[0 - 3000]
ccells.cd45.bam
mcells.cd45.bam Coverage
[0 - 3000]
mcells.cd45.bam
ncells.cd45.bam Coverage
[0 - 3000]
ncells.cd45.bam
Ensembl Genes (80)
Kallisto
599452
5
185825
1
199819
599451
4
b
Human (hg38)
chr1
chr1:198,686,202-198,723,798
Go
p36.31 p36.21 p36.12 p35.2 p34.2 p33 p32.2 p31.3 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.3 q44
37 kb
198,690 kb
198,700 kb
198,710 kb
198,720 kb
ccells.cd45.bam Coverage
[0 - 3000]
ccells.cd45.bam
mcells.cd45.bam Coverage
[0 - 3000]
mcells.cd45.bam
ncells.cd45.bam Coverage
[0 - 3000]
ncells.cd45.bam
Ensembl Genes (80)
Kallisto
260768
364464
247023
314713
342594
615875
232661
2
599452
246503
572140
684886
330776
689589
7
199819
656355
250045
342594
599453
6
3
342594
211359
297171
684886
288014
224159
428740



C

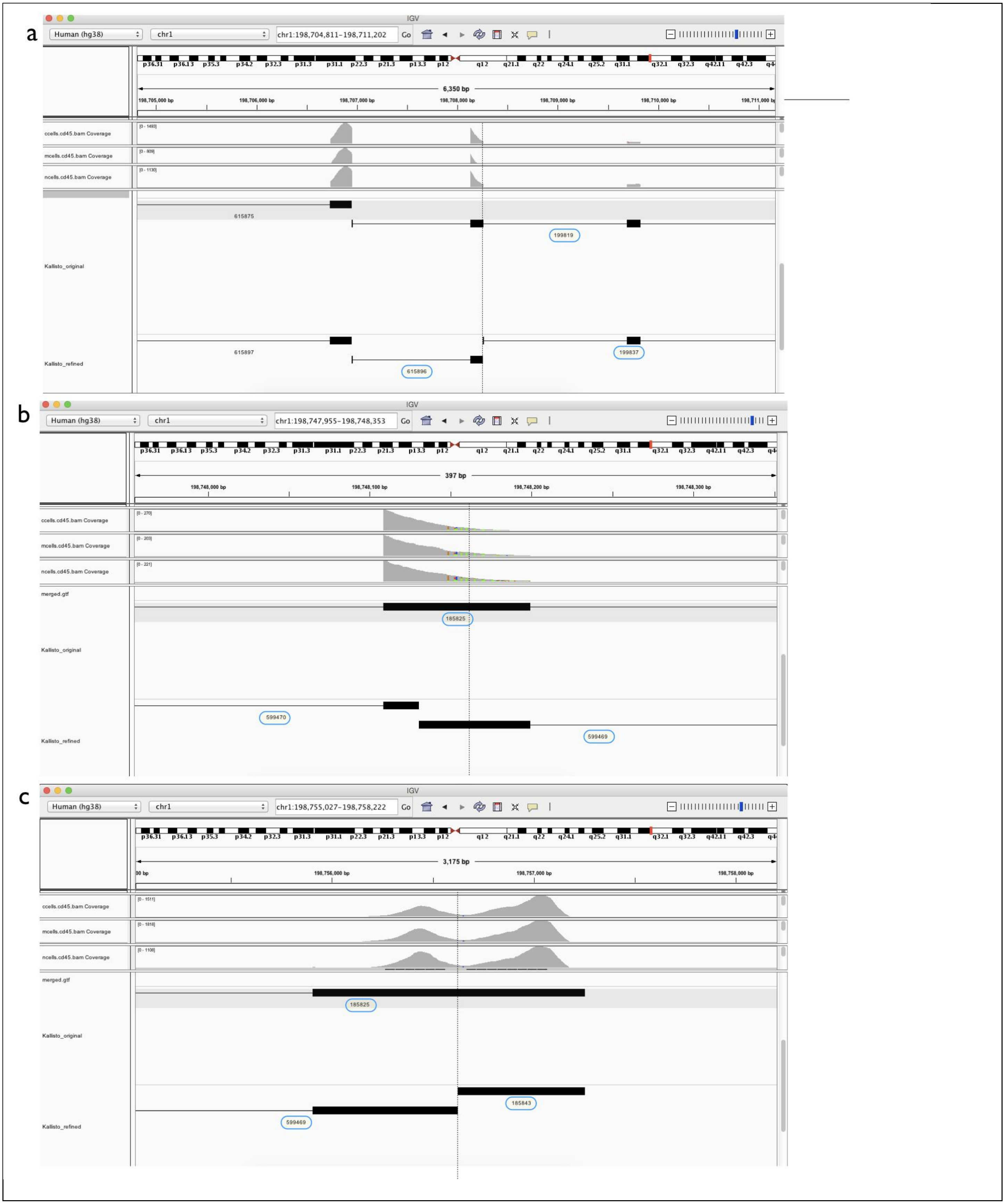
	equivalence class id	transcripts
1	185825	ENST00000348564, ENST00000442510.
2	199819	ENST00000348564, ENST00000367367, ENST00000442510, ENST00000529828, ENST00000530727, ENST00000573477, ENST00000573679, ENST00000574441, ENST00000575923, ENST00000576833.
3	211359	ENST00000413409, ENST00000571847.
4	599451	ENST00000348564, ENST00000367367, ENST00000442510, ENST00000529828.
5	599452	ENST00000348564, ENST00000367367, ENST00000442510, ENST00000529828, ENST00000530727.
6	599453	ENST00000348564, ENST00000367367, ENST00000442510, ENST00000491302, ENST00000529828, ENST00000530727, ENST00000573477, ENST00000573679, ENST00000574441, ENST00000575803, ENST00000575923, ENST00000576833.
7	615875	ENST00000348564, ENST00000367367, ENST00000367379, ENST00000442510, ENST00000529828, ENST00000530727, ENST00000573298, ENST00000573477, ENST00000573679, ENST00000574441, ENST00000575923, ENST00000576833.



Supplementary Figure 9

IGV visualization of pseudoalignments.

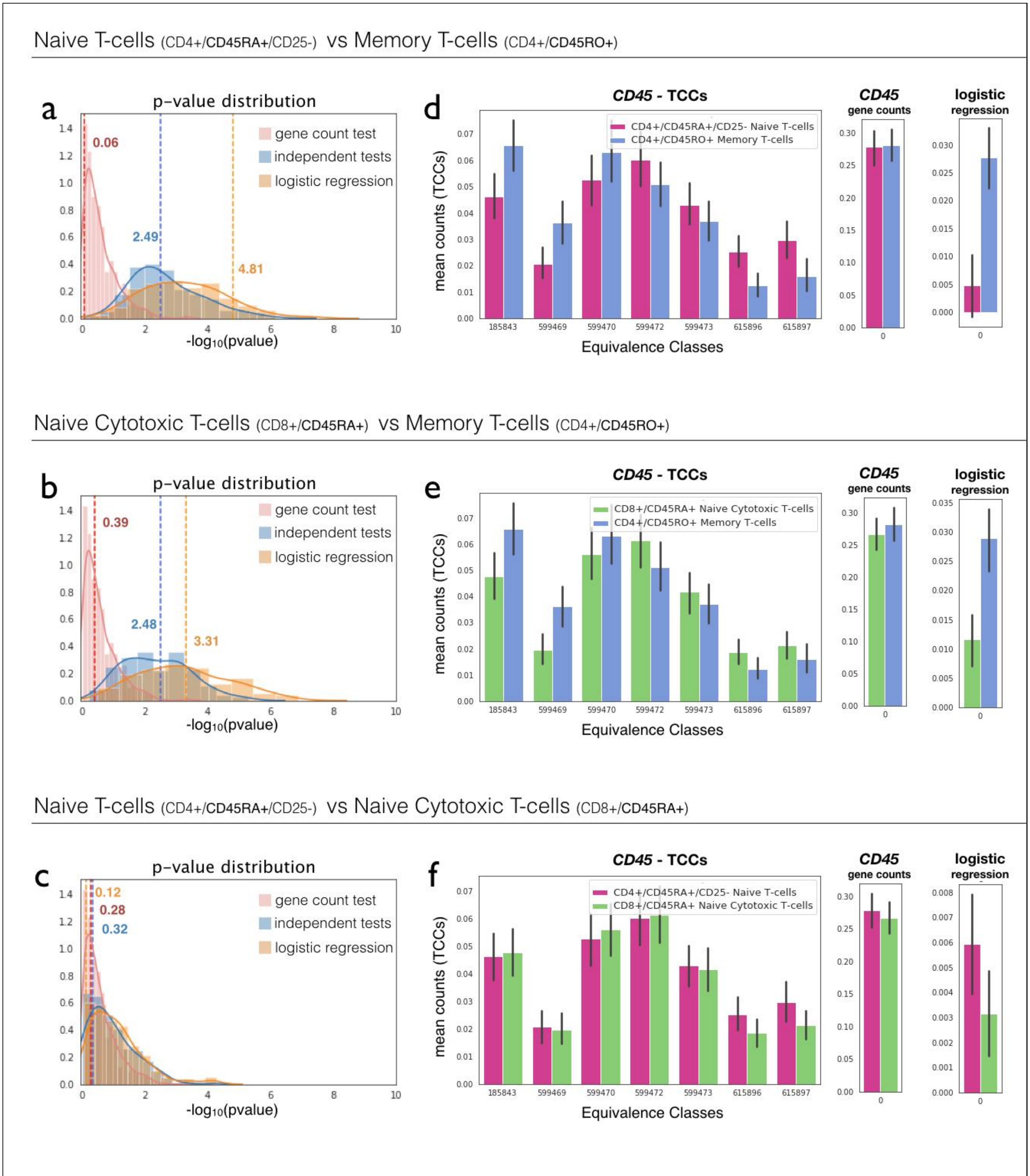
The kallisto v0.44.0 pseudobam option outputs a BAM file for each sample that can be visualized directly with IGV. Shown here are the pseudoalignments of the three purified T-cell types from Zheng *et al*., 2017 (a, b). The TCCs (track ‘kallisto’) are shown alongside their transcripts of origin (shown in track ‘Ensembl Genes’). TCCs used in the differential expression analysis from Figure 2 are boxed in blue on the IGV track (a, b) and their corresponding transcripts are tabulated (c). (d) shows the distribution of read distance from the 3’ end from Zheng *et al.*, 2017 as found with pseudobam. The substantial number of reads far from annotated 3’-ends suggests a large number of unannotated 3’ UTRs whose reads are informative when transcript compatibility counts are utilized.


a
IGV
Human (hg38)
chr1
chr1:198,704,811-198,711,202
Go
p36.31 p36.13 p35.3 p34.2 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q4
6,350 bp
198,705,000 bp
198,706,000 bp
198,707,000 bp
198,708,000 bp
198,709,000 bp
198,710,000 bp
198,711,000 bp
ccells.cd45.bam Coverage
[0 - 1493]
mcells.cd45.bam Coverage
[0 - 809]
ncells.cd45.bam Coverage
[0 - 1130]
615875
199819
Kallisto_original
615897
199837
615896
Kallisto_refined
b
IGV
Human (hg38)
chr1
chr1:198,747,955-198,748,353
Go
p36.31 p36.13 p35.3 p34.2 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q4
397 bp
198,748,000 bp
198,748,100 bp
198,748,200 bp
198,748,300 bp
ccells.cd45.bam Coverage
[0 - 270]
mcells.cd45.bam Coverage
[0 - 203]
ncells.cd45.bam Coverage
[0 - 221]
merged.gtf
185825
Kallisto_original
599470
599469
Kallisto_refined
c
IGV
Human (hg38)
chr1
chr1:198,755,027-198,758,222
Go
p36.31 p36.13 p35.3 p34.2 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q4
3,175 bp
00 bp
198,756,000 bp
198,757,000 bp
198,758,000 bp
ccells.cd45.bam Coverage
[0 - 1511]
mcells.cd45.bam Coverage
[0 - 1818]
ncells.cd45.bam Coverage
[0 - 1108]
merged.gtf
185825
Kallisto_original
185843
599469
Kallisto_refined

Supplementary Figure 10

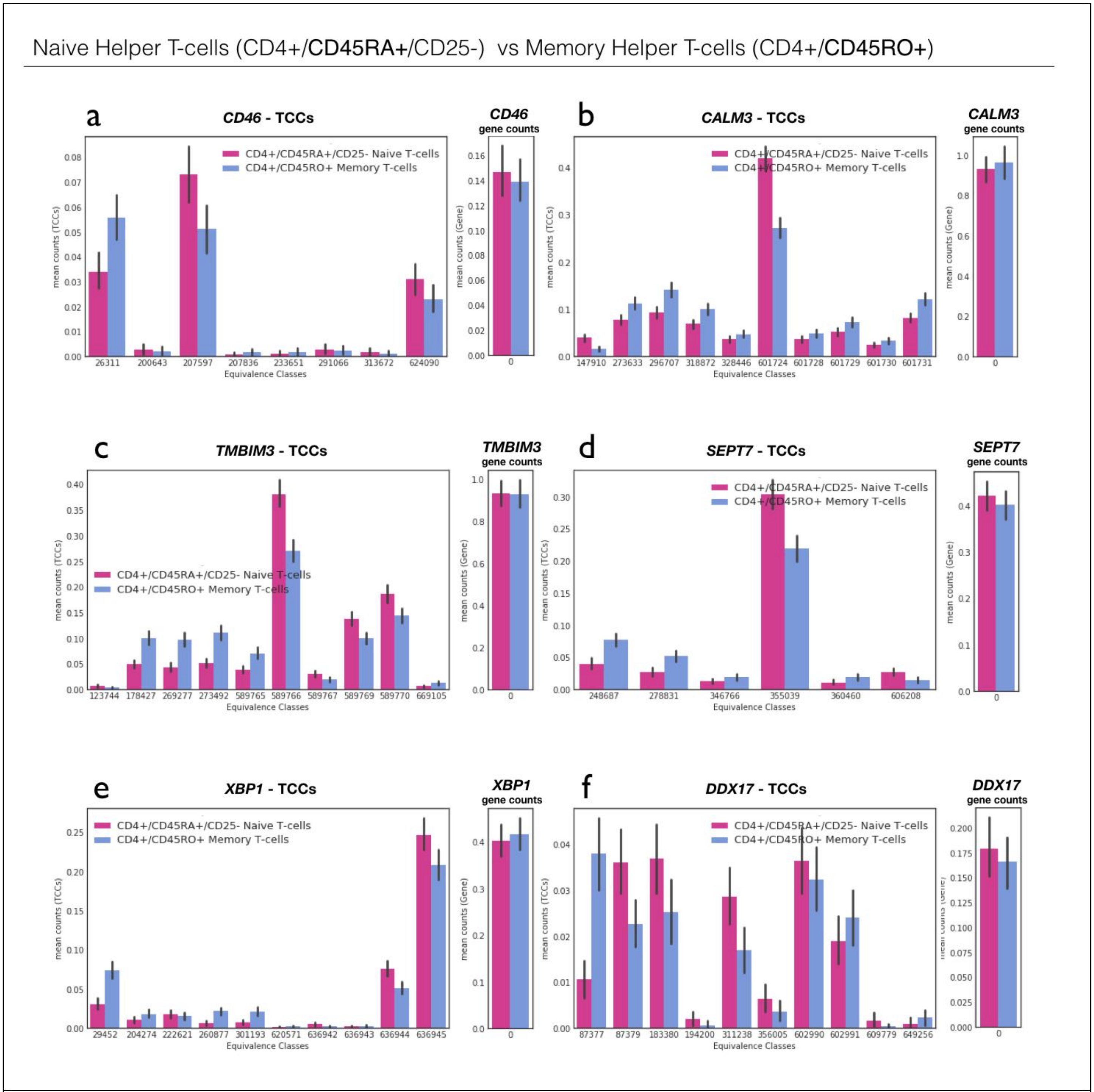
IGV visualization of TCC structure using a revised transcriptome with updated 3’UTRs.

After identifying three unannotated 3’UTRs from the Zheng *et al*., 2017, we modified the transcriptome to include these novel UTRs (see Supplementary Methods). This figure depicts ECs in original transcriptome (track ‘kallisto_original’) side-by-side with the ECs of the updated transcriptome (track ‘kallisto_refined’). Also included are coverage tracks for each of the three purified T-cell types from Zheng *et al*., 2017. An analysis shows the three newly inserted UTRs break up the previous ECs into more refined ECs. (a) shows that EC #199819 is refined into EC #615896 and #199837. In (b, c), EC #185825 is represented by 3 ECs in the refined version, EC #185243, #599469, and #599470.



Supplementary Figure 11

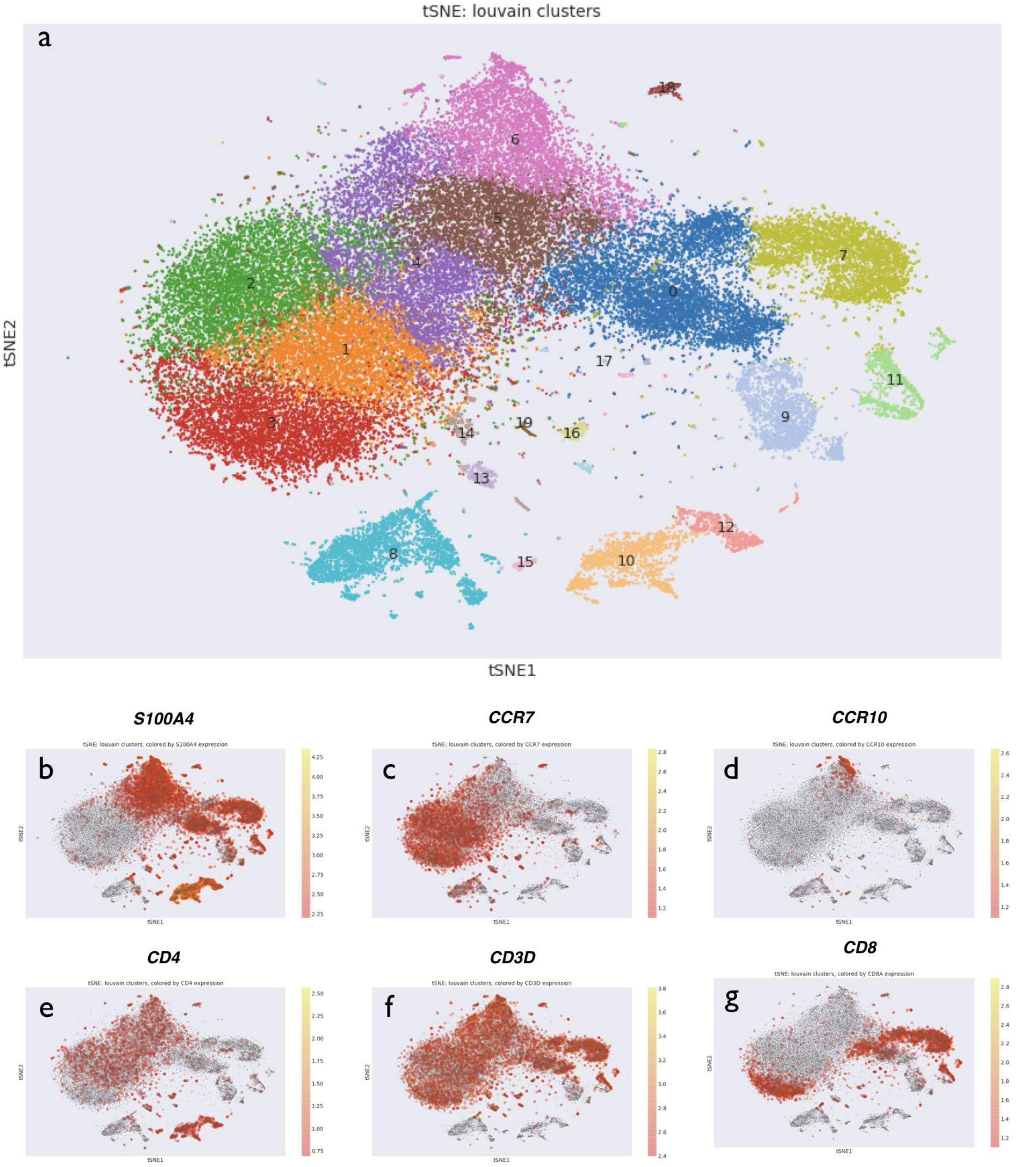
Reanalysis of CD45 using updated transcriptome.
The human transcriptome was updated with *de novo* 3'UTRs. Using the updated transcriptome, we performed pairwise differential expression tests on CD45 between the three purified T-cell types that were sequenced with the 10x technology. The p-value distributions in (a, b, c) were generated by 200 subsamples, where we subsampled n=3000 cells per cell type from the full dataset of n = 9923 naïve helper T-cells, n=9994 memory helper T cells, and n=11914 naïve cytotoxic T cells. On each subsample, we performed multiple logistic regression on the TCCs ('logistic regression'), logistic regression on gene counts ('gene count test') and independent logistic regression for each EC ('independent tests'). Corresponding p-values were obtained by using the likelihood ratio test (see Methods for the definitions of null and alternate models and degree of freedom for the likelihood ratio test). Bar plots in (d, e, f) depict the mean expression across 3000 cells corresponding to one of the 200 subsamples. The whiskers correspond to 95% confidence intervals of the mean expression across 3000 cells. CD45 remained differential with logistic regression on TCCs but not with gene counts. ECs#185843 and #599469, refined from EC #185825, remain differential between the memory T cell type and the naive T cell types (d, e). EC #615896, refined from EC #199819, remains differential between naive and memory helper T cells (d, e).



Supplementary Figure 12

Differential genes between naïve and memory helper T-cells.

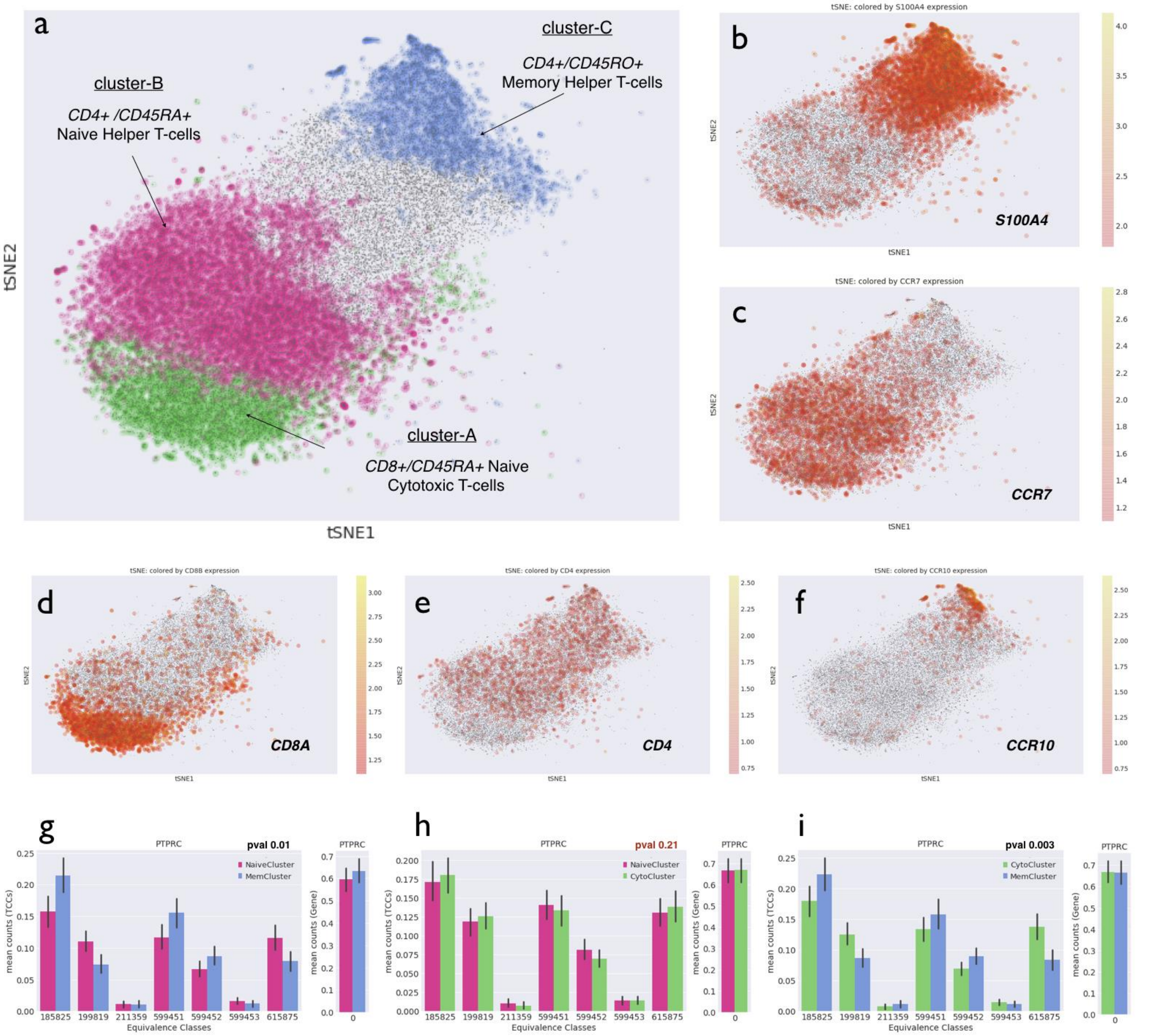
Naïve helper T cells (n = 9923 cells) and memory helper T-cells (n = 9994 cells) were purified in Zheng *et al.,* 2017 and then sequenced with 10x technology. The bar plots depict the mean expression of the several genes across a random subsample of 3000 naive helper T cells and 3000 memory helper T cells from the full dataset. The whiskers on the bar plot correspond to the 95% confidence interval of the mean expression. The genes depicted here were found to be significantly differentially expressed using logistic regression on TCCs on this subsample of cells, but were not detected when applying logistic regression on gene counts (see Methods).



Supplementary Figure 13

A *de novo* analysis of 68k PBMCs from Zheng *et al*. 2017.

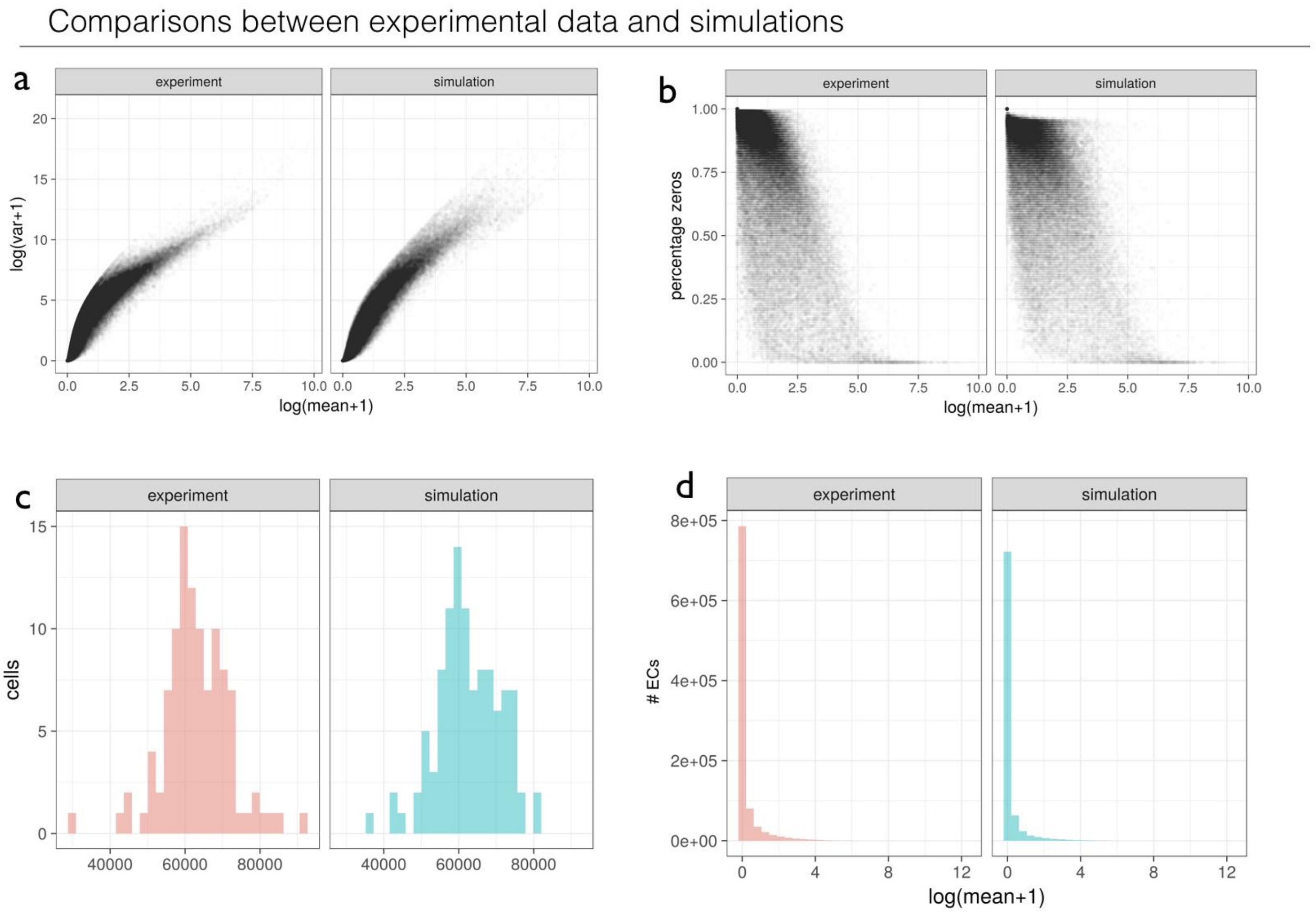
We obtained TCCs with kallisto pseudoalignment, clustered the cells (n= 65444 PBMCs) using the Louvain method in scanpy **(a)** and plotted the cells with known T-cell markers **(b-g)**. Naïve helper, memory helper and naïve cytotoxic T-cells formed distinct clusters, whereas Zheng *et al.* 2017 were unable to separate these cell types into distinct clusters.



Supplementary Figure 14

De novo analysis of T-cell clusters in 10x data.

Cells that most likely correspond to populations of naïve cytotoxic T-cells (Cluster A, CD8A+/CD4-/CCR7+, 5226 cells), naïve helper T-cells (Cluster B, CD4+/CCR7+, 12424 cells) and memory helper T-cells (Cluster C, CD4+/S100A4+/CCR10+, 4173 cells) were subsetted from the 10X dataset of 68K PBMCs (see Methods) and shown in TSNE space (a). Expression profiles of the known T-cell markers used to identify the three cell types are used to color the cells in TSNE space (b-f). We generated 200 subsamples of these cells, each with 2000 cells per cluster. For each subsample, we performed pairwise differential expression on CD45 across the three cell clusters using logistic regression on TCCs. The likelihood ratio test was used to obtain p-values from the logistic regression fit (see Methods for the definitions of of null and alternate models and degree of freedom used for the likelihood ratio test), and the average of the p-values across the 200 subsamples (g, h, i). The mean expression of PTPRC gene counts and TCCs corresponding to one subsample are depicted via bar plots (g, h, i), where the whiskers correspond to the 95% confidence interval of the mean expression.



Supplementary Figure 15

Comparisons between experiment and simulation.

The 105 simulated cells of the nonperturb group were compared to the 105 myoblast cells from Trapnell *et al*. on the basis of which they were simulated. (a) To compare the mean-variance relationship of transcripts between experimental and simulated data, each transcript's variance in TPM was plotted against its mean TPM expression. (b) To compare the extent of dropout, each transcript's proportion of zero expression across cells was plotted against its mean expression in TPM. We also compared the distribution in TCCs between the experimental and simulated data. Panel (c) depicts a histogram of the number of expressed ECs (i.e. nonzero TCCs) per cell. Panel (d) depicts a histogram of the mean expression across ECs.