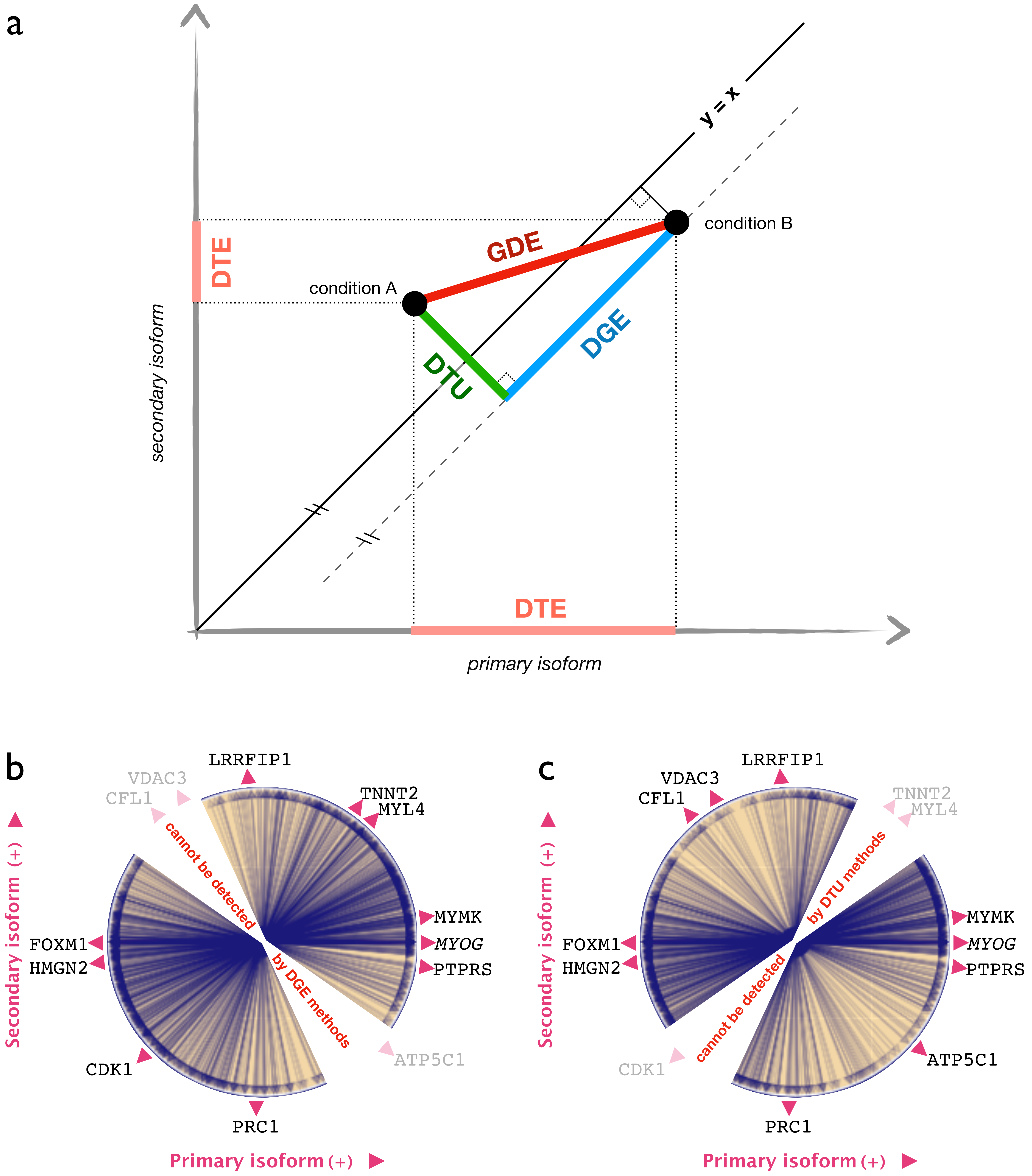
Supplementary Material

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

Vasilis Ntranos^, Lynn Yi^, Páll Melsted and Lior Pachter
^authors contributed equally

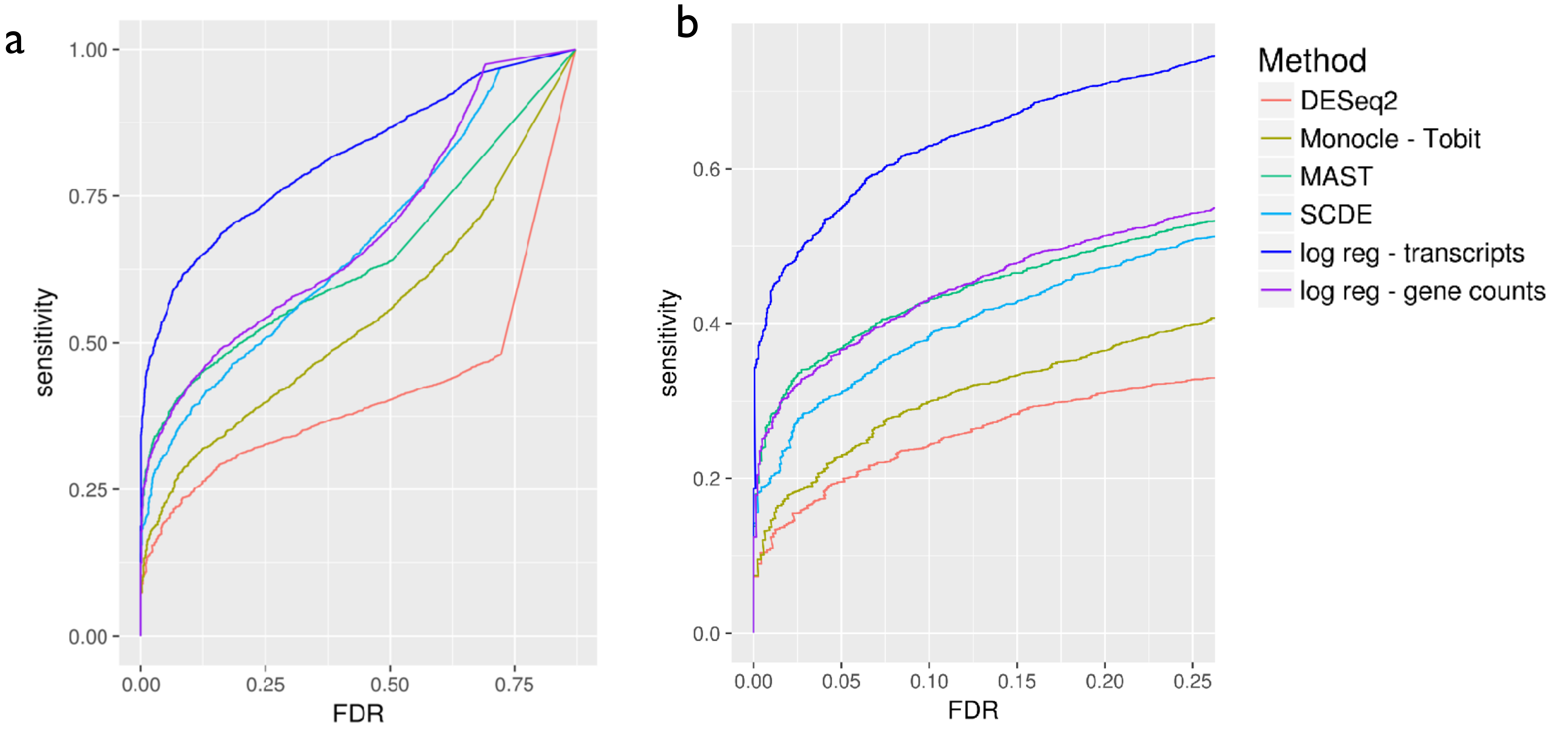
Supplementary Figure 1



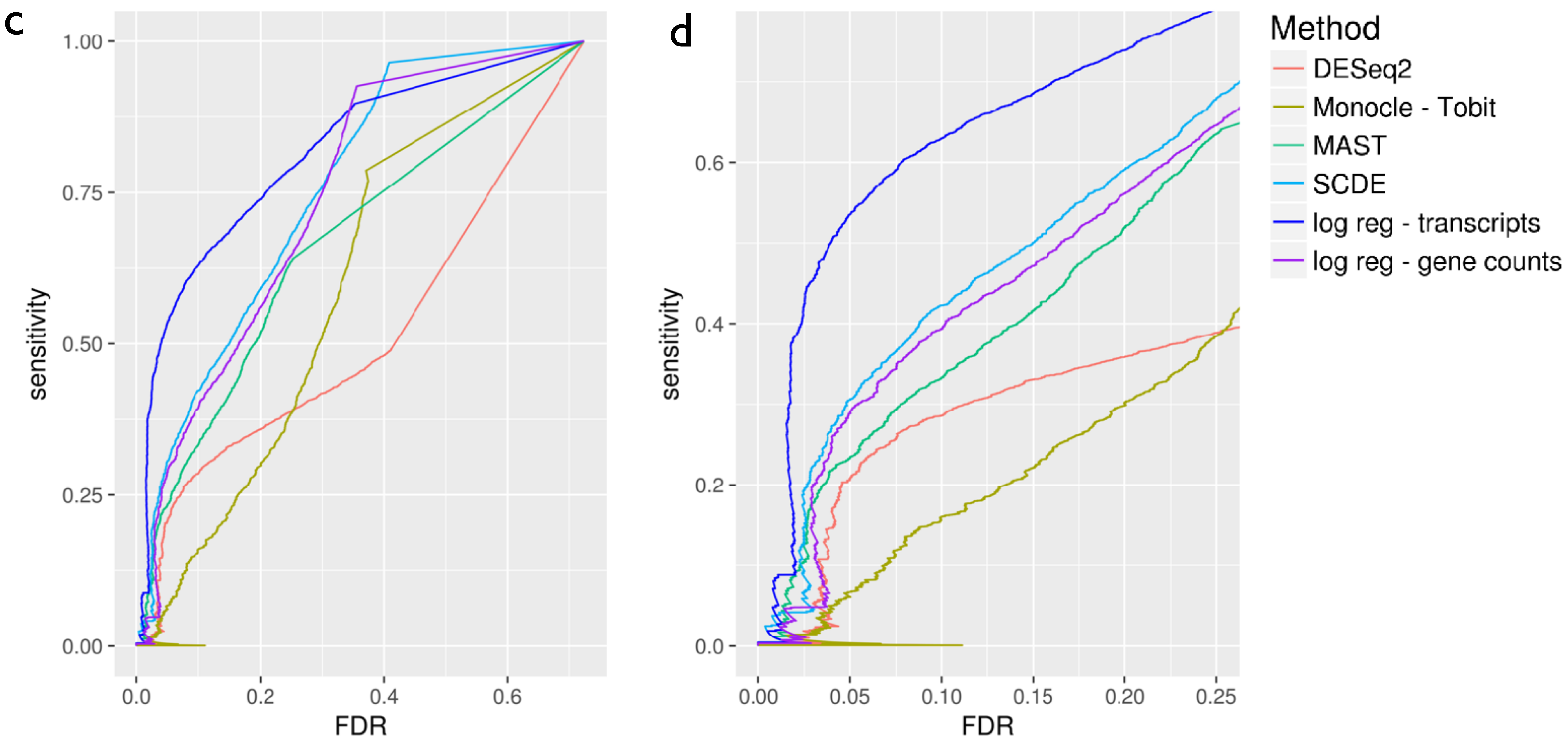
Relationship between differential expression methods. (a) Depiction of the difference in expression of a two-transcript gene in two cell types. The two black points correspond to gene expression in each of the two cell types: the x-coordinate depicts the expression of its first transcript and the y-coordinate, 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.

Supplementary Figure 2

Simulations - Experimental Effect Sizes

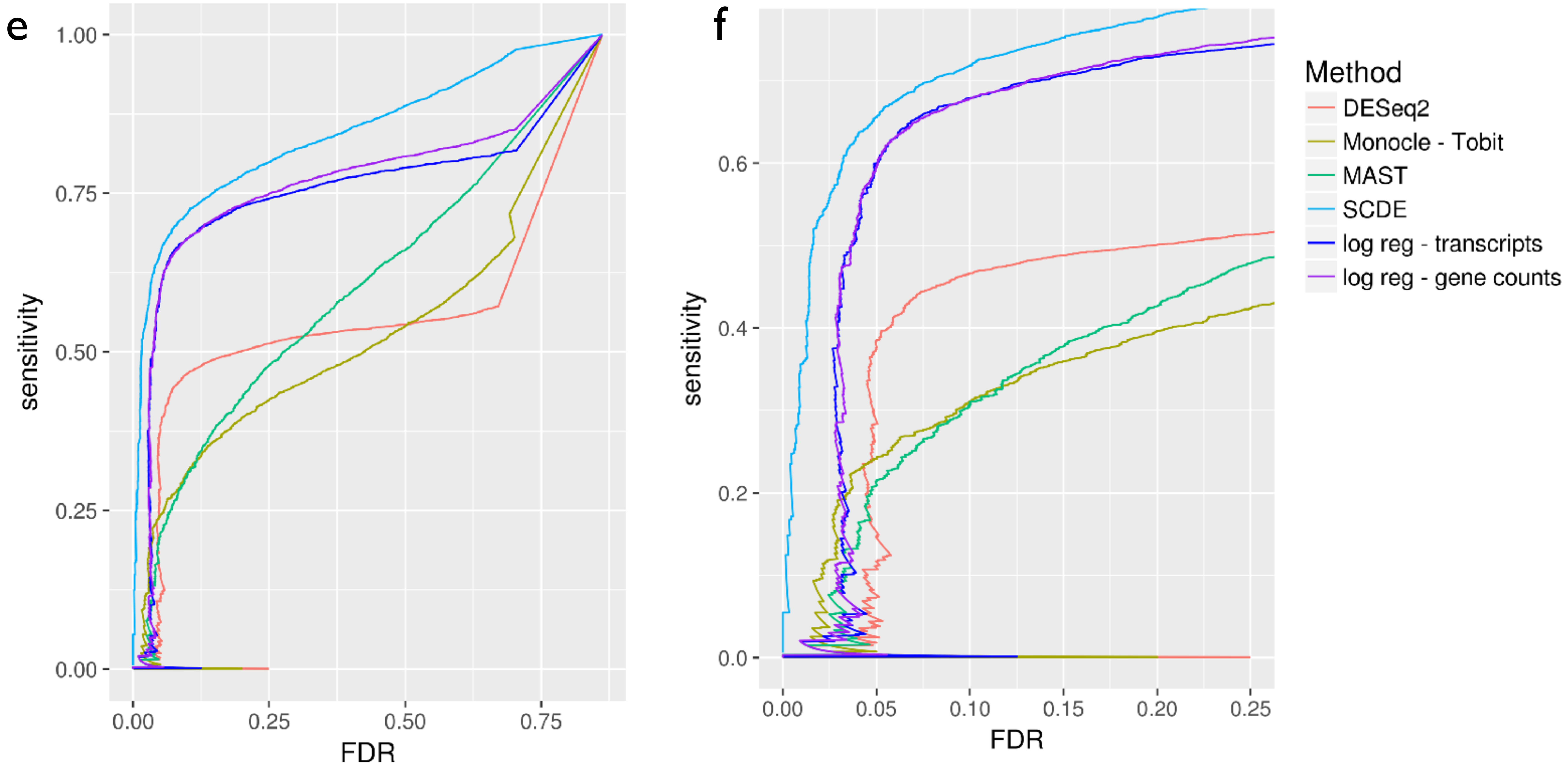


Simulations - Independent Effect Sizes



Supplementary Figure 2

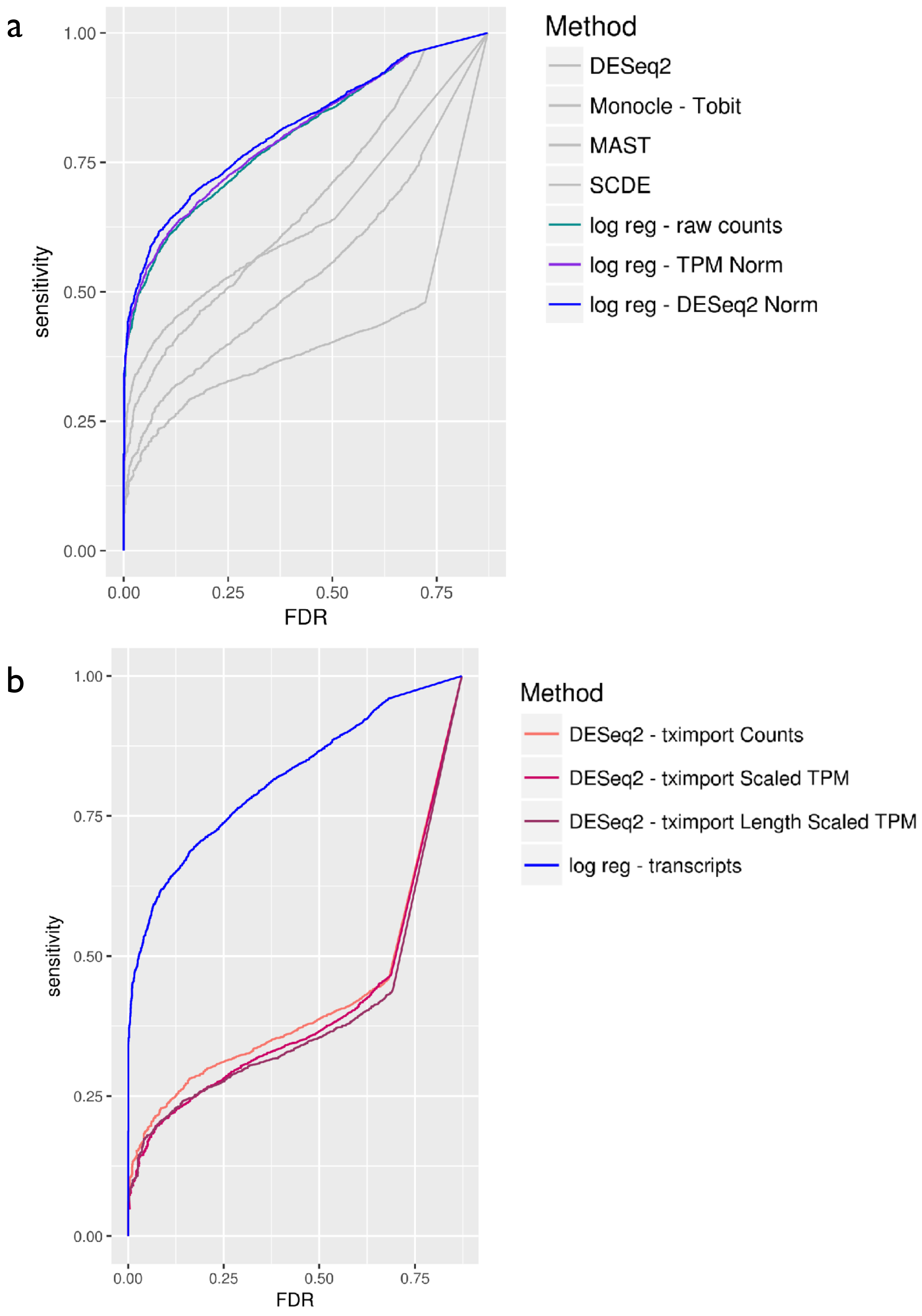
Simulations - Correlated Effect Sizes



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 univariate logistic regression on the summarized gene counts and is a DGE method

Supplementary Figure 3

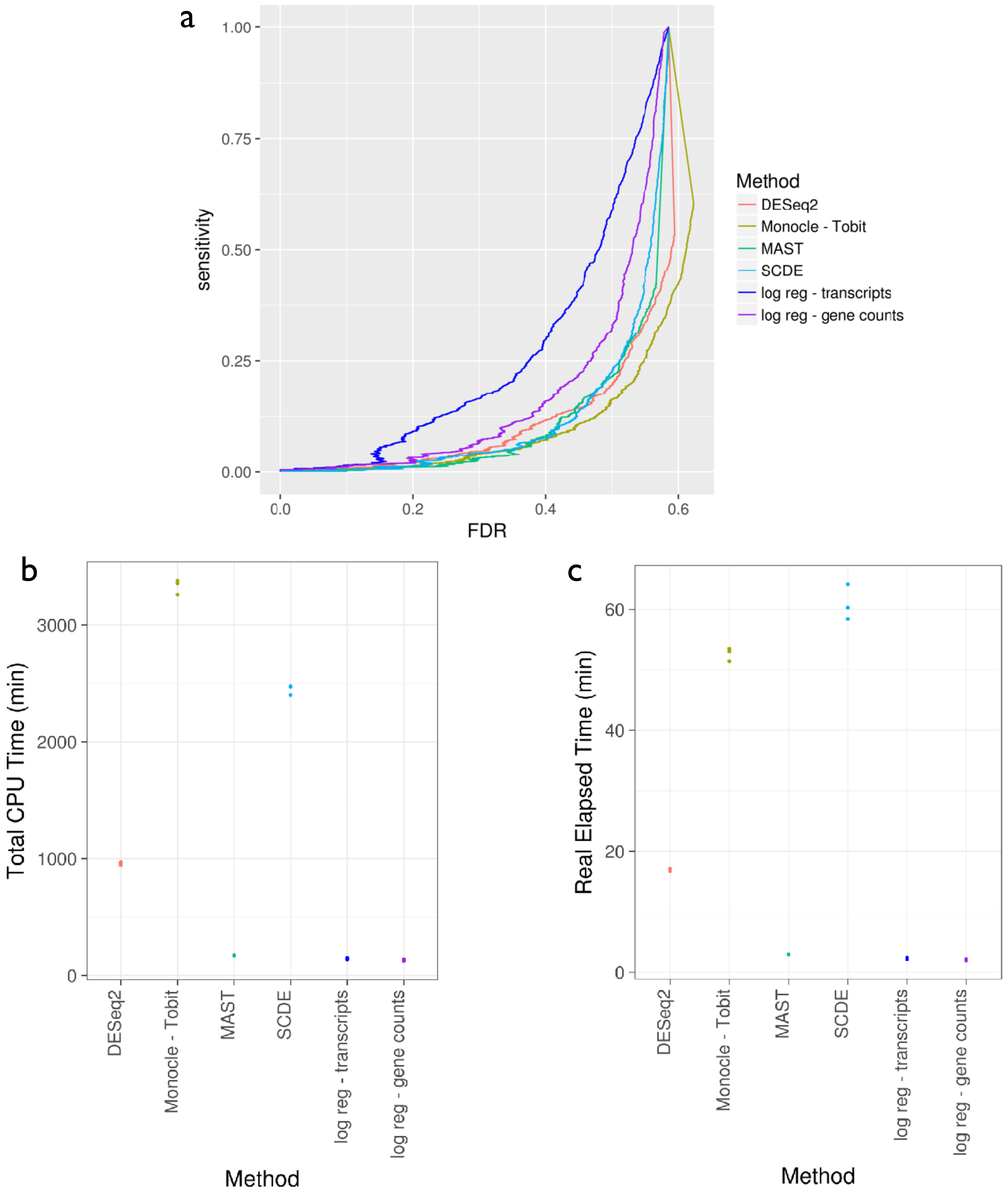
Simulations - Experimental Effect Sizes



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).

Supplementary Figure 4

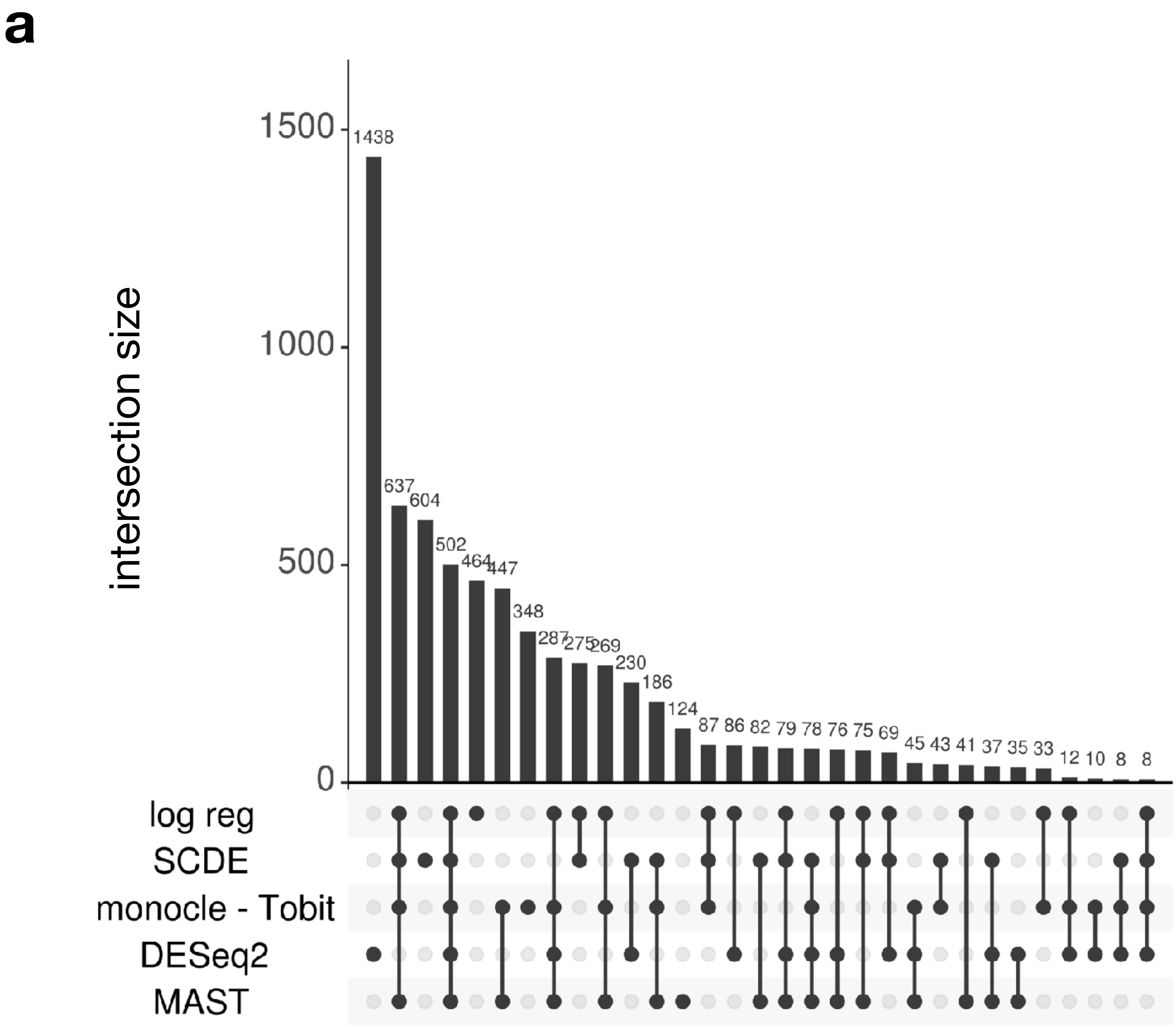
Splatter Simulations

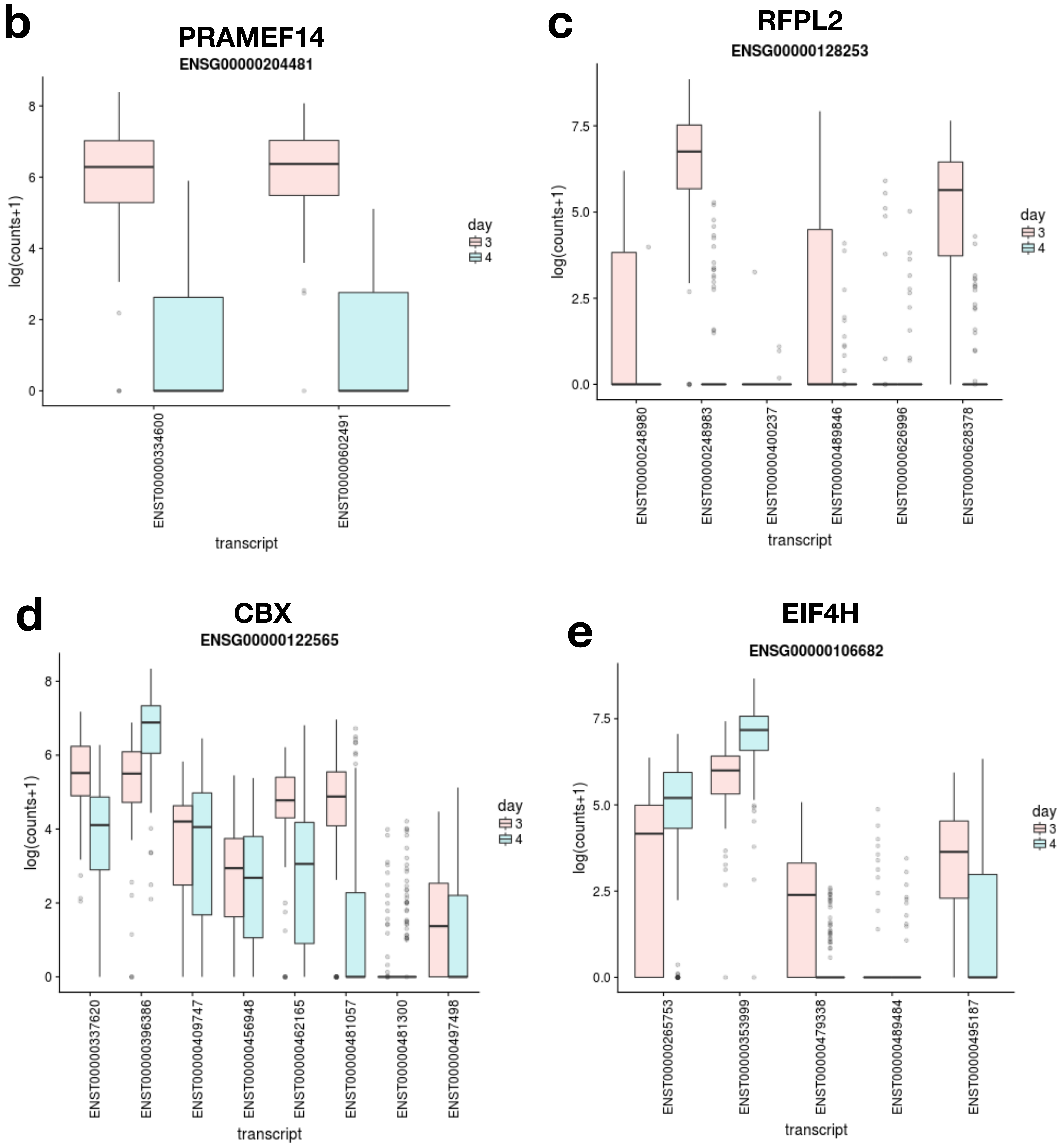


Splatter Simulations

Independently of the simulations described in Supplementary Figures 2-4, another simulation was generated using Splatter. Within Splatter, two groups of cells were simulated, each with 10% probability of producing differential transcripts, resulting in 19% differential transcripts between the two groups. The simulated counts are used as inputs into the differential expression methods for benchmarking. In (a) we plotted a sensitivity-FDR curve. In (b) and (c), we benchmarked the runtimes of these methods on the simulation, plotting the CPU time and the real elapsed time of three trials.

Supplementary Figure 5



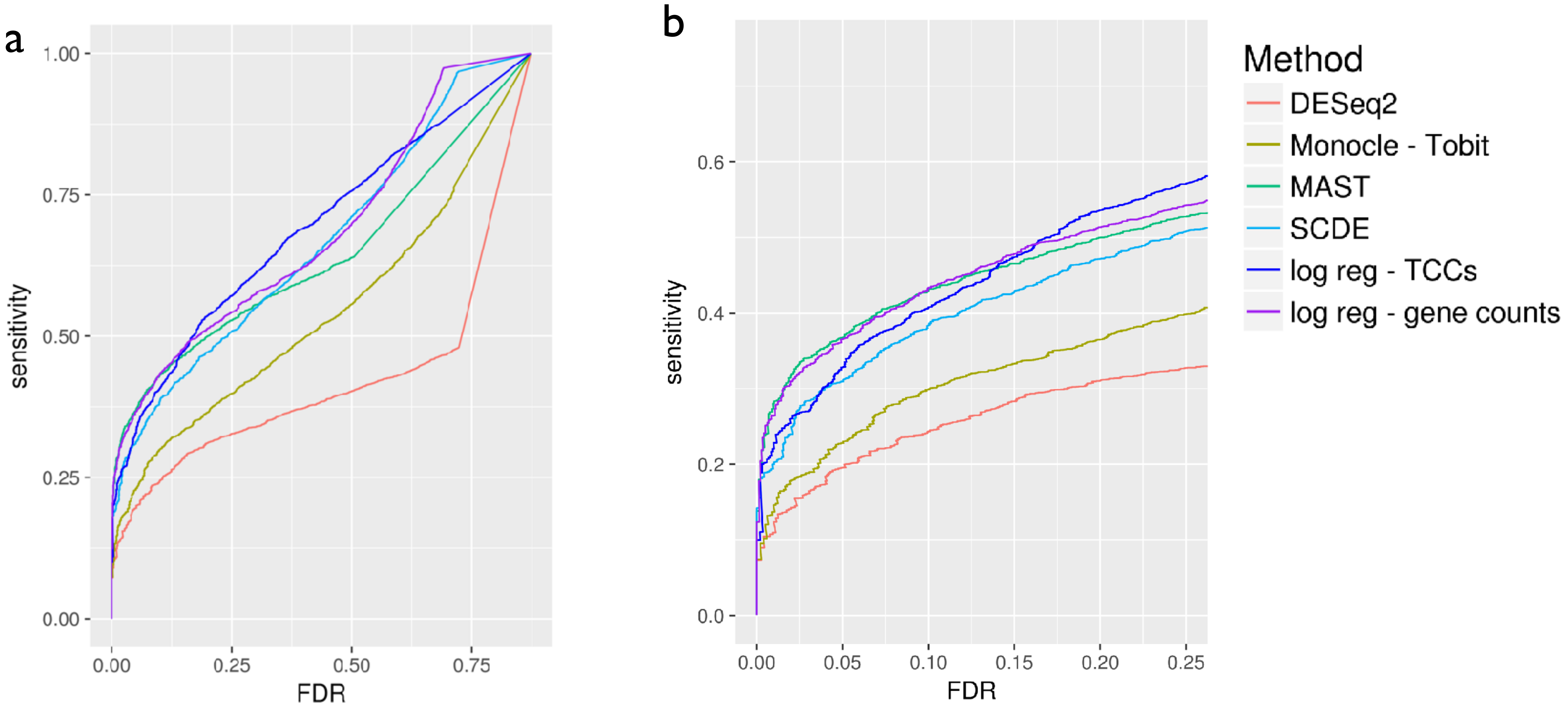


Analysis of embryonic dataset

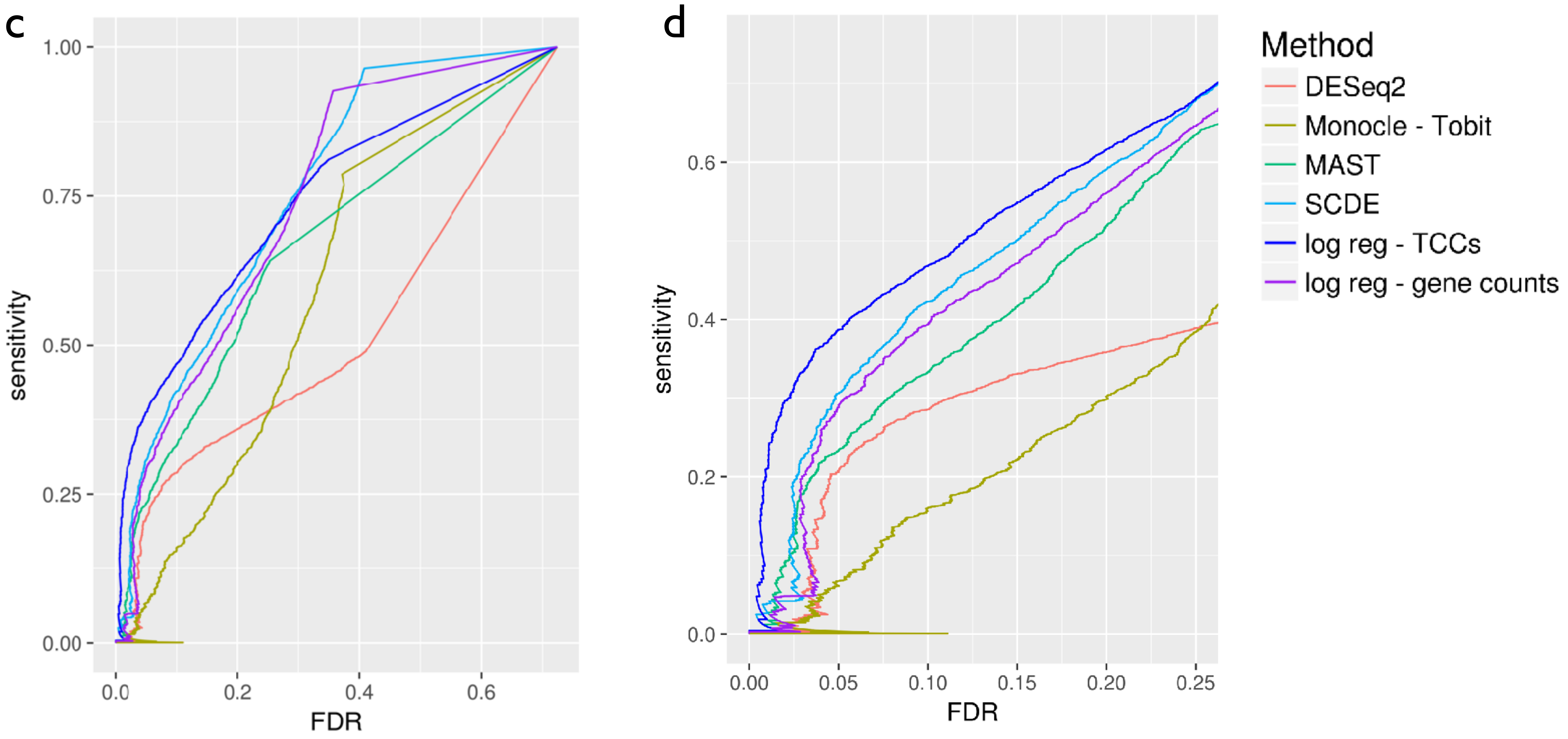
We used five different methods to find differential expressed genes between day 3 and day 4 post-fertilization preimplantation human embryonic cells. An UpSet plot shows sizes of the set intersections of the 3000 most significant genes from each method (a). We showcase transcript dynamics of two of the 502 genes that are in the intersection of all five methods (b, c) and two of the 464 genes that are in the set unique to logistic regression (d, e). In these figures, the quartiles are plotted (i.e. the box contains the 25-75th percentile), and outliers defined as farther than 1.5 * inter-quartile range are depicted as points.

Supplementary Figure 6

Simulations - Experimental Effect Sizes

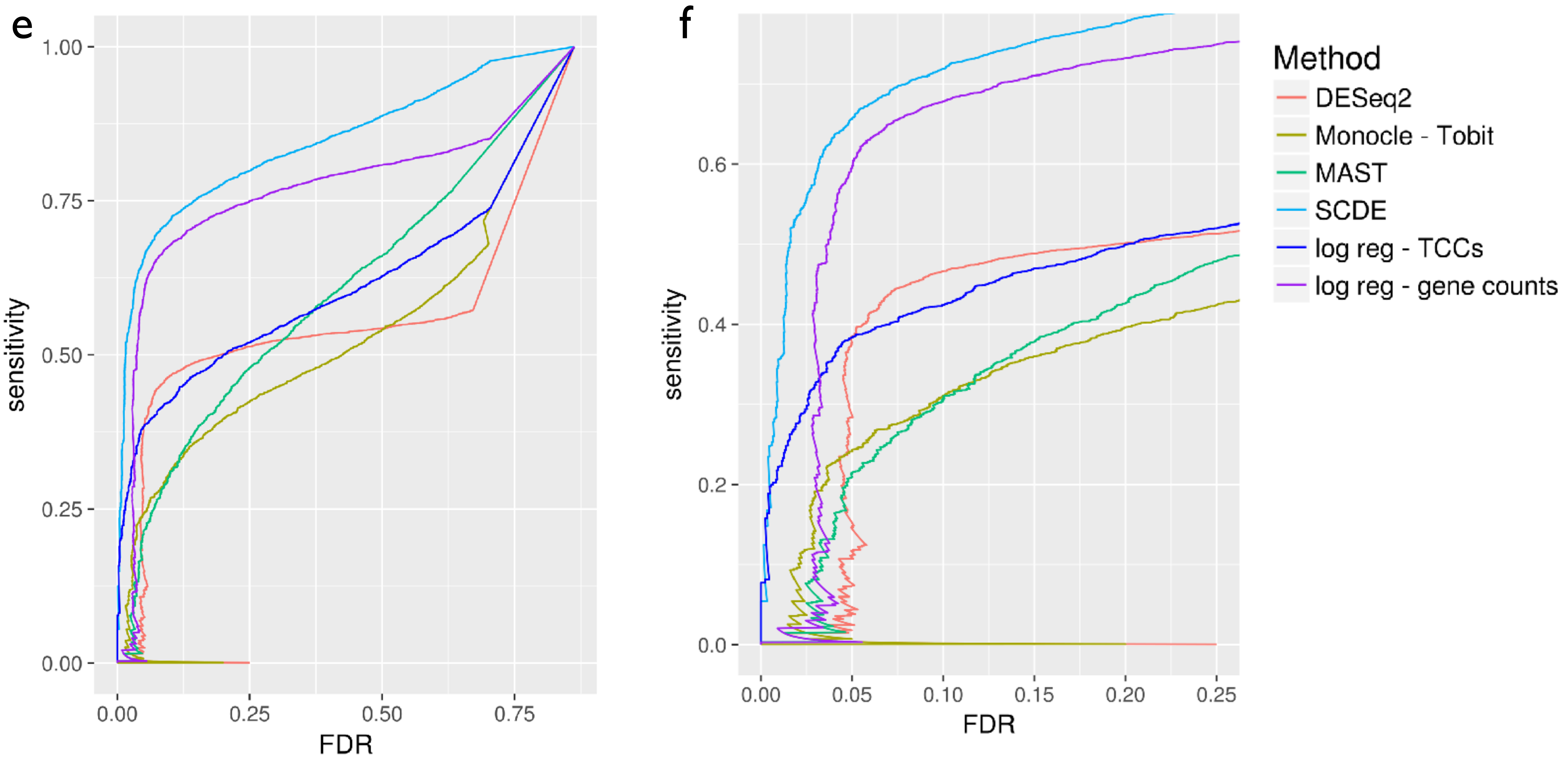


Simulations - Independent Effect Sizes



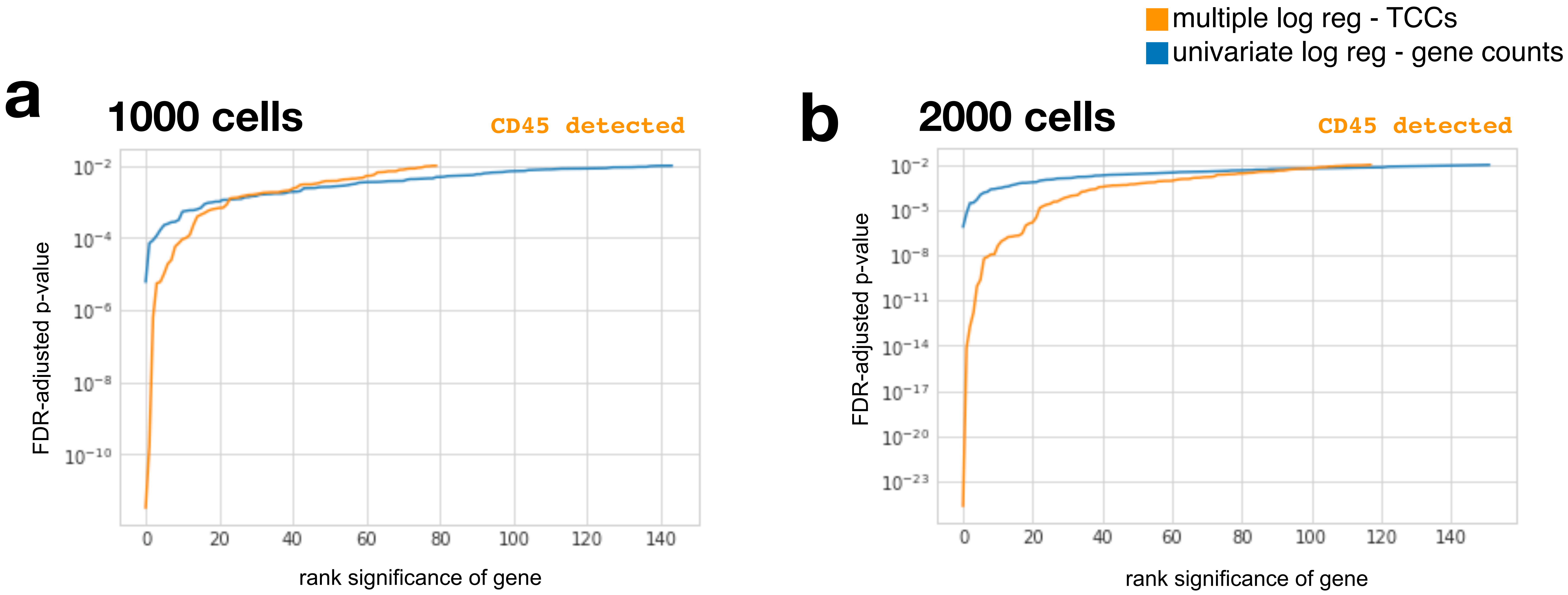
Supplementary Figure 6

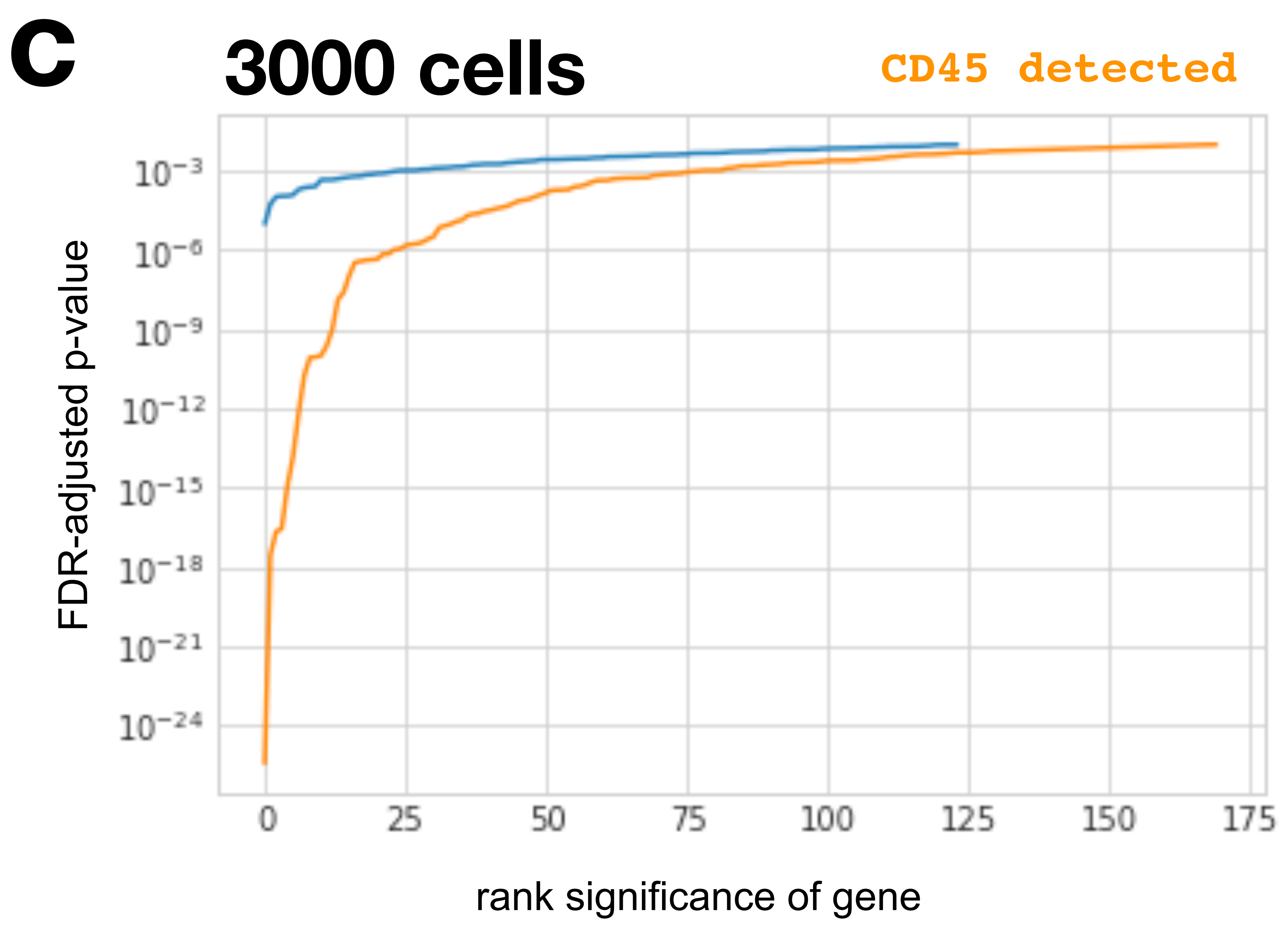
Simulations - Correlated Effect Sizes



Performance of differential expression methods on simulations. In the event that transcripts could not be quantified, performing logistic regression on TCCs is an alternative that also retains isoform-level information. On the same full-length simulation as in Supp Fig 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.

Supplementary Figure 7

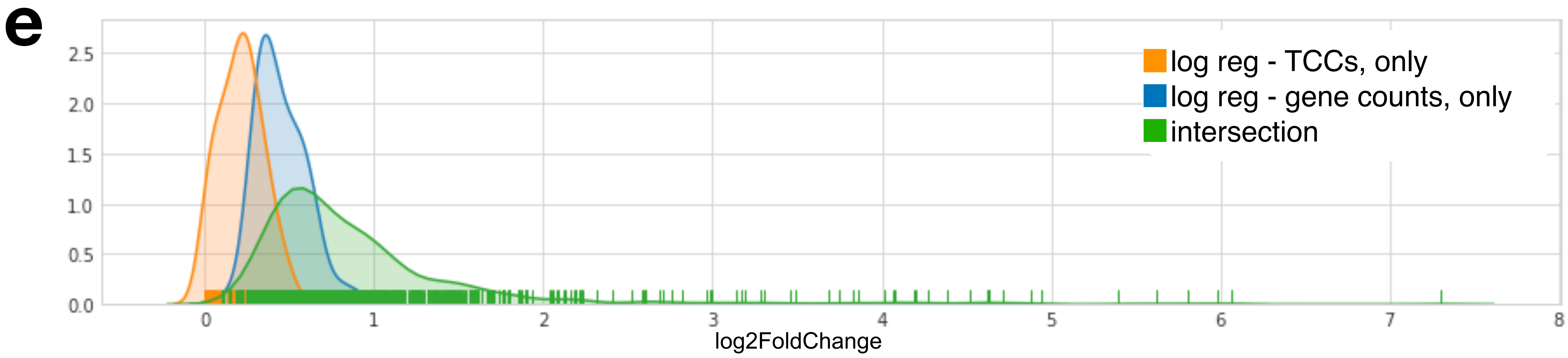




d

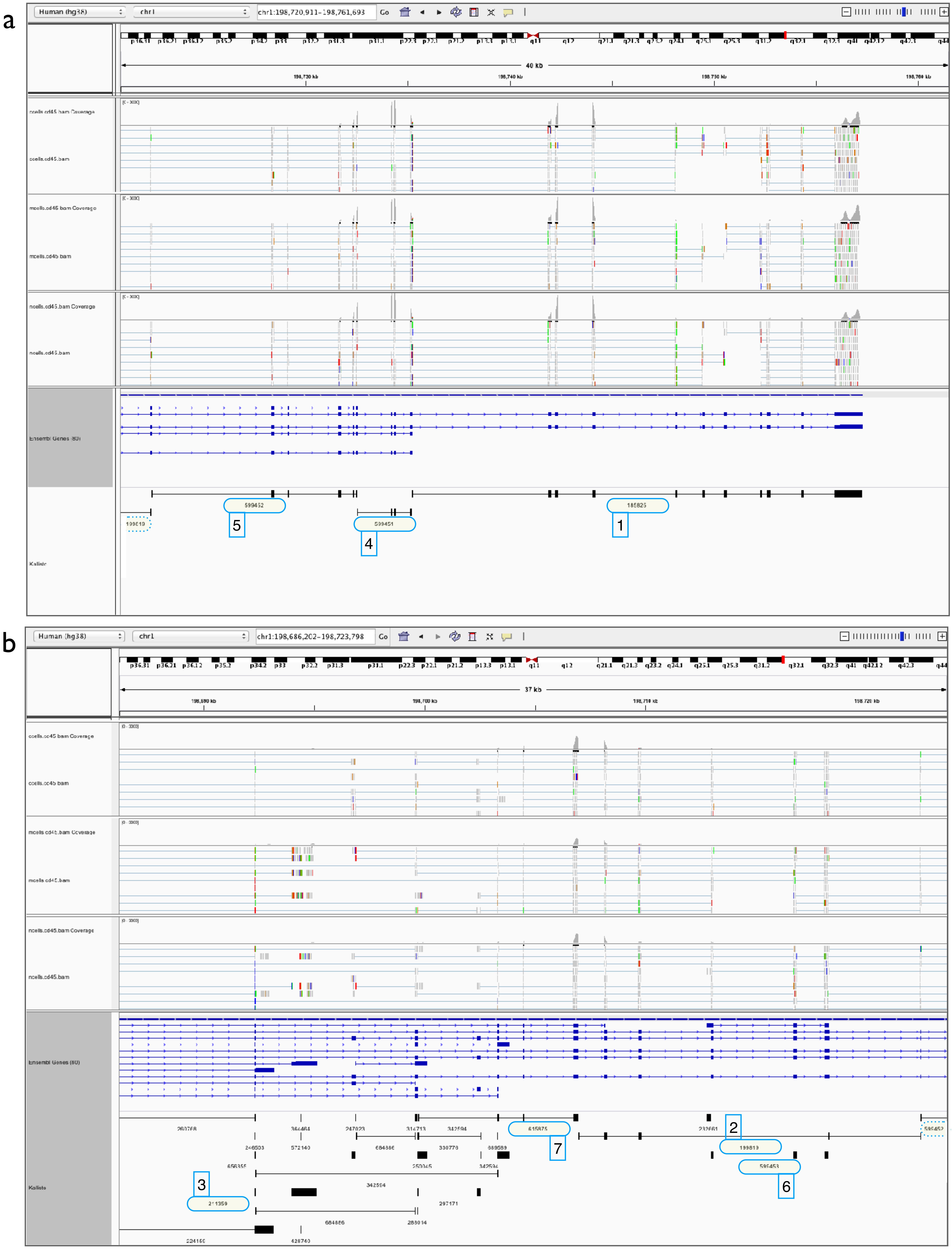
Number genes detected (FDR < 0.01)

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

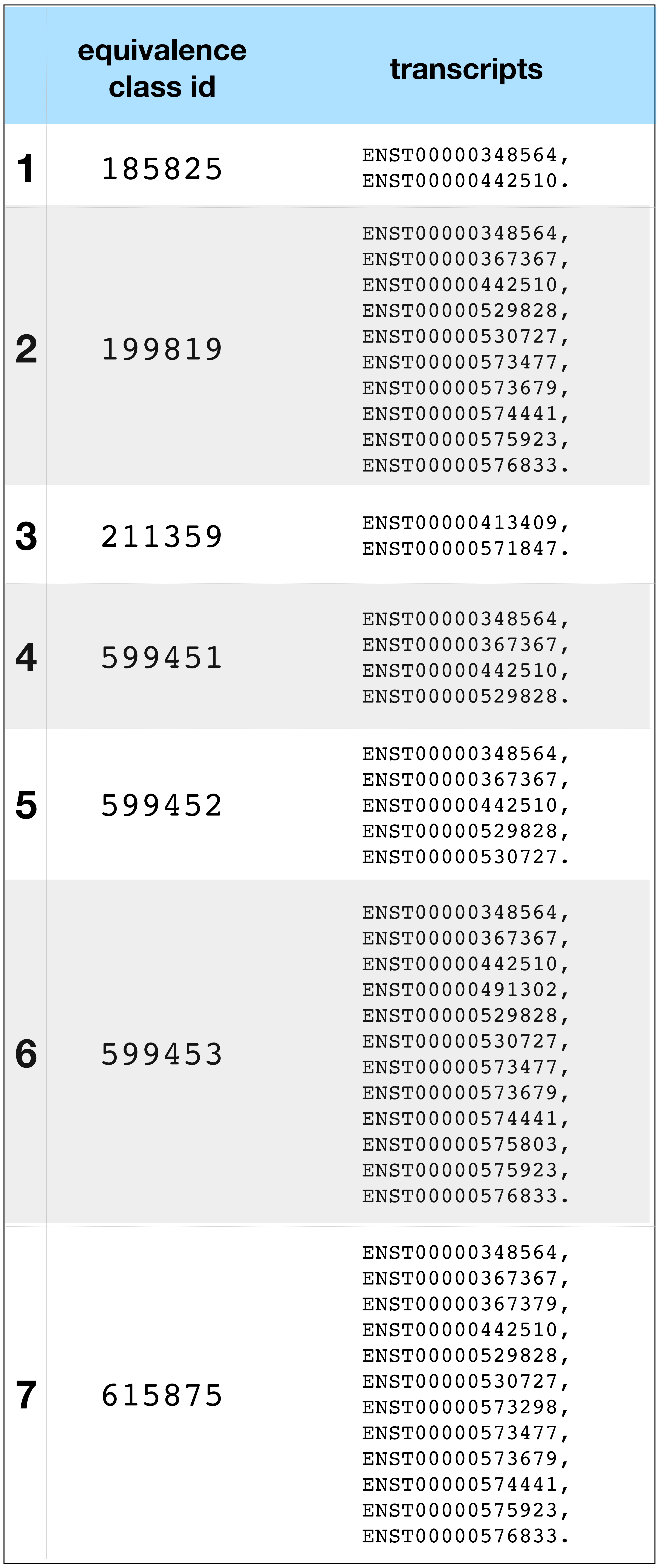


Power analysis of CD45. Using the PBMC dataset, we performed differential analysis between memory and naive T-cells at three levels of subsampling cells: 1000 cells (a), 2000 cells (b) and 3000 cells (c). We compared multiple logistic regression on TCCs with univariate logistic regression using gene 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.

Supplementary Figure 8



c

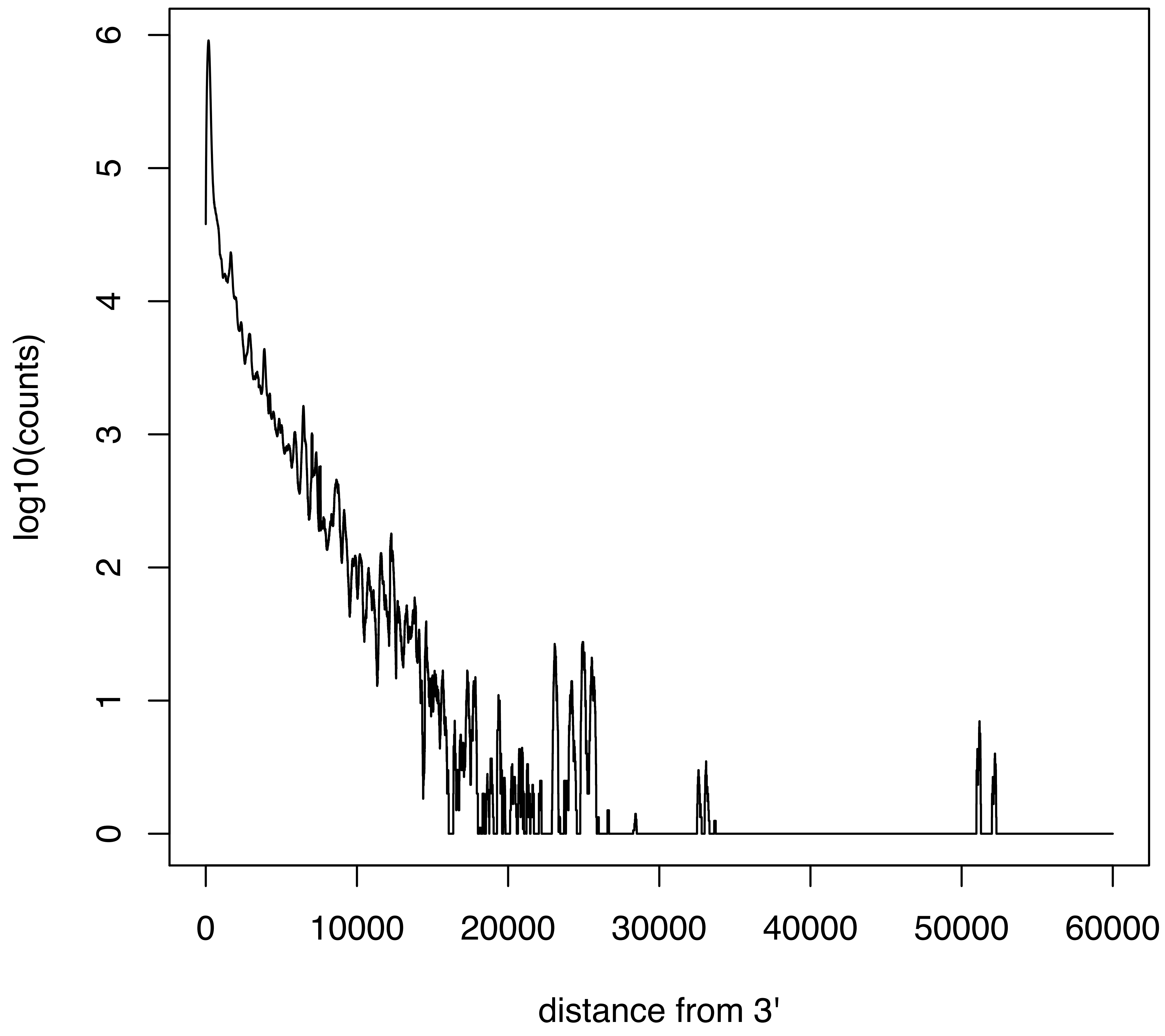


	equivalence class id	transcripts
1	185825	ENST00000348564, ENST00000442510.
2	199819	ENST00000348564, ENST00000367367, ENST00000442510, ENST00000529828, ENST00000530727, ENST00000573477, ENST00000573679, ENST00000574441, ENST00000575923, ENST00000576833.
3	211359	ENST00000413409, ENST00000571847.
4	599451	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.

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 (Fig 2) are boxed in blue on the IGV track (a, b) and their corresponding transcripts are tabulated (c).

Supplementary Figure 9

Read Distribution in 10x

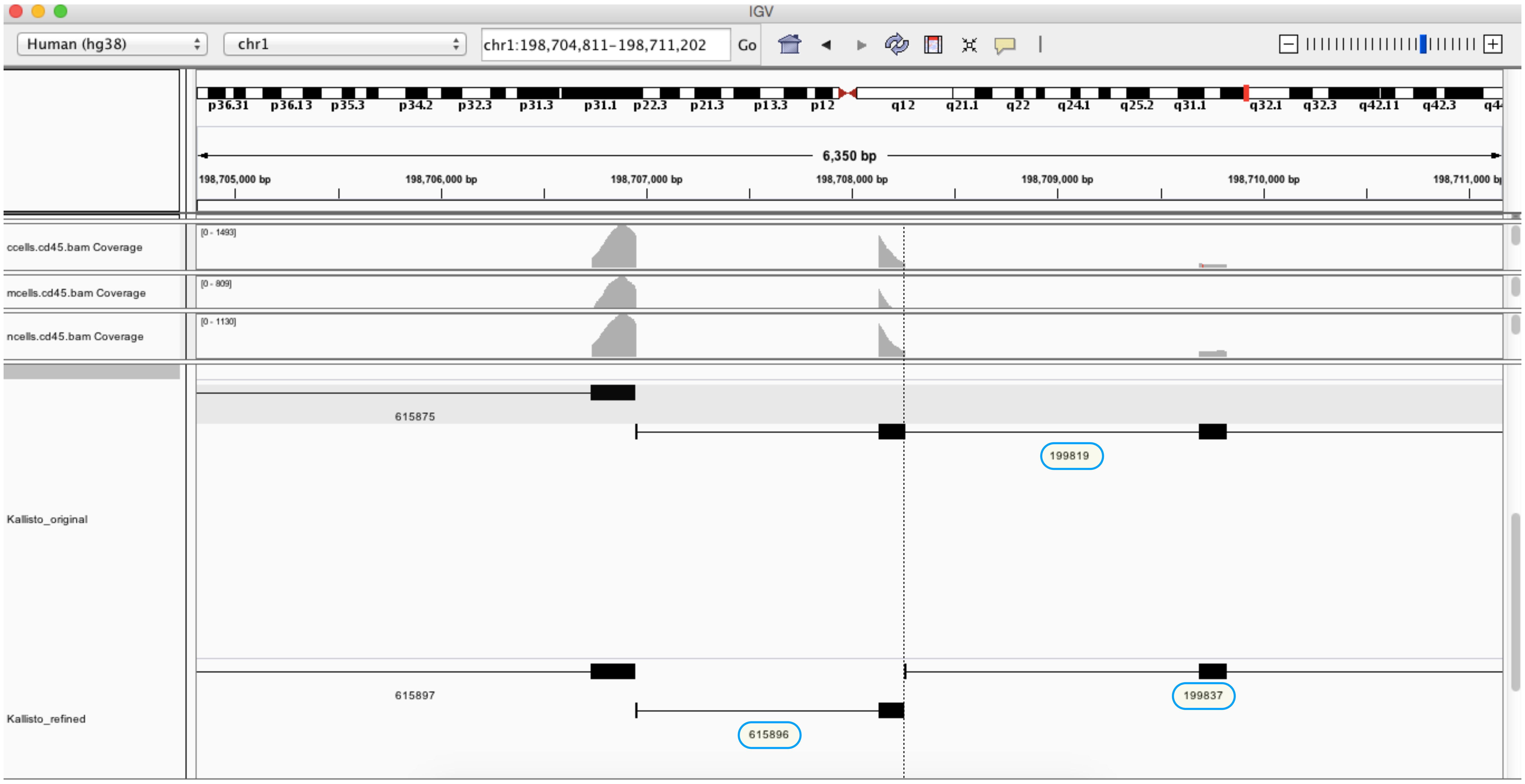


The distribution of read distance from the 3’ end from Zheng *et al.*, 2017. The substantial number of reads far from annotated 3’-ends suggests a large number of unannotated 3’ UTRs whose reads are informative when transcript compatibility counts are utilized.

Supplementary Figure 10

Transcriptome Updated with Newly Discovered 3'UTRs

a

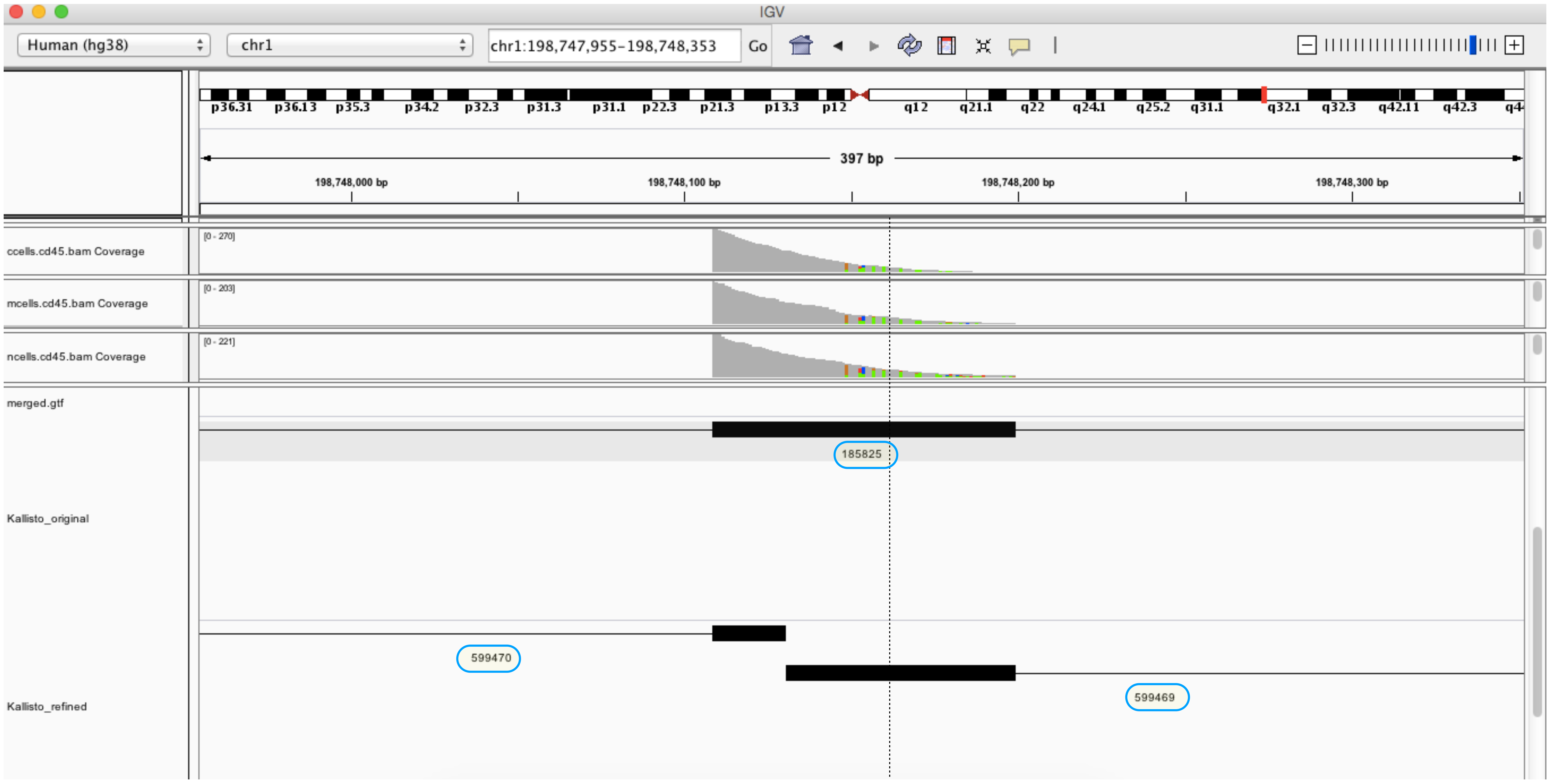


Human (hg38)

chr1

chr1:198,704,811-198,711,202

b

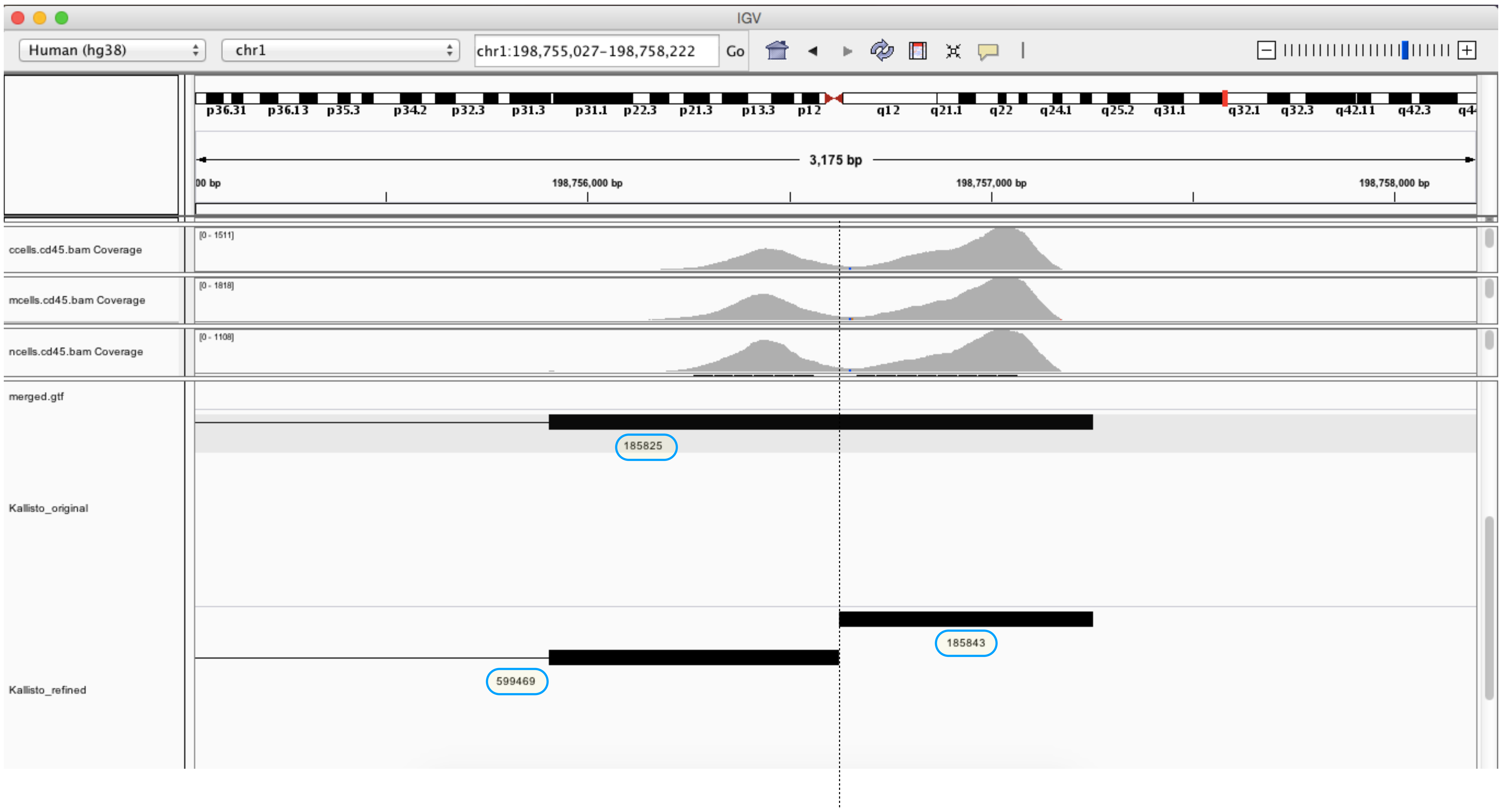


Human (hg38)

chr1

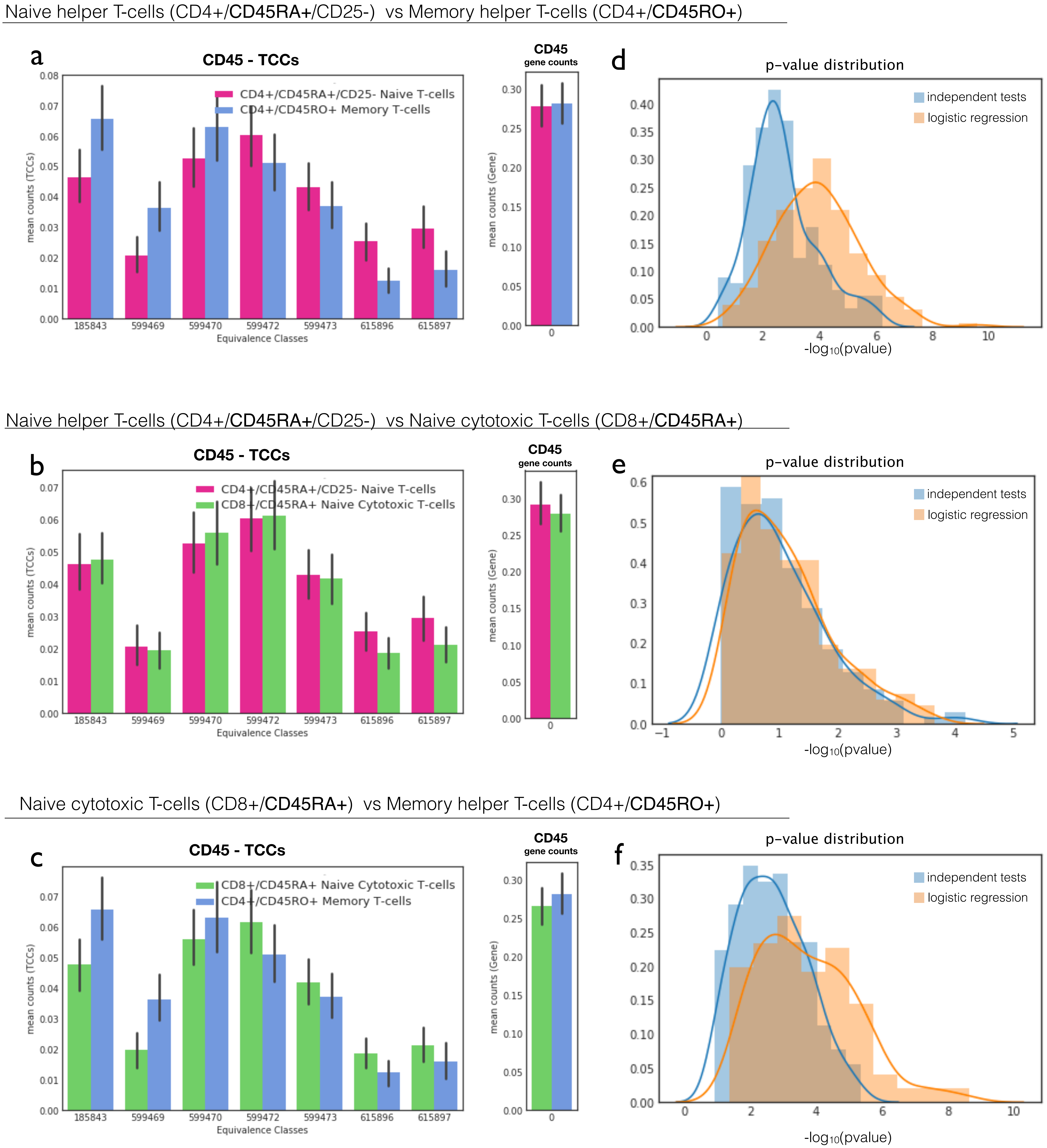
chr1:198,747,955-198,748,353

c



<u>IGV visualization of TCC structure from UTR modification.</u> 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.

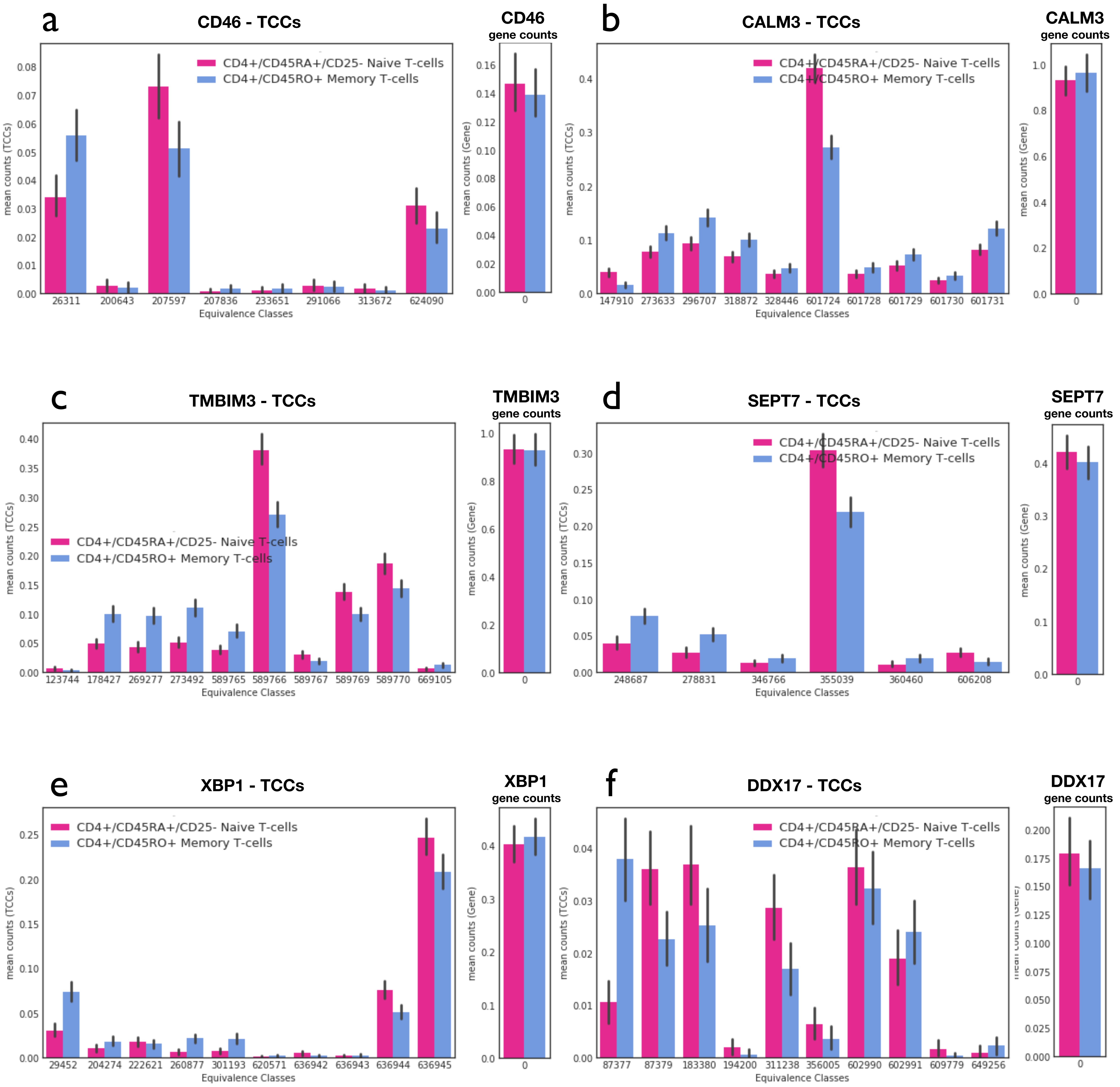
Supplementary Figure 11



Reanalysis of CD45 using updated transcriptome. The transcriptome was updated with new 3’UTRs. After obtaining TCCs using the new transcriptome, we performed logistic regression on CD45, which remained differential with 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 (a, b). EC #615896, refined from EC #199819, remains differential between naive and memory helper T cells. The p-value distributions in (d,e,f) were generated by subsampling n=3000 cells per group over 200 iterations and the error bars in (a,b,c) correspond to 95% confidence intervals.

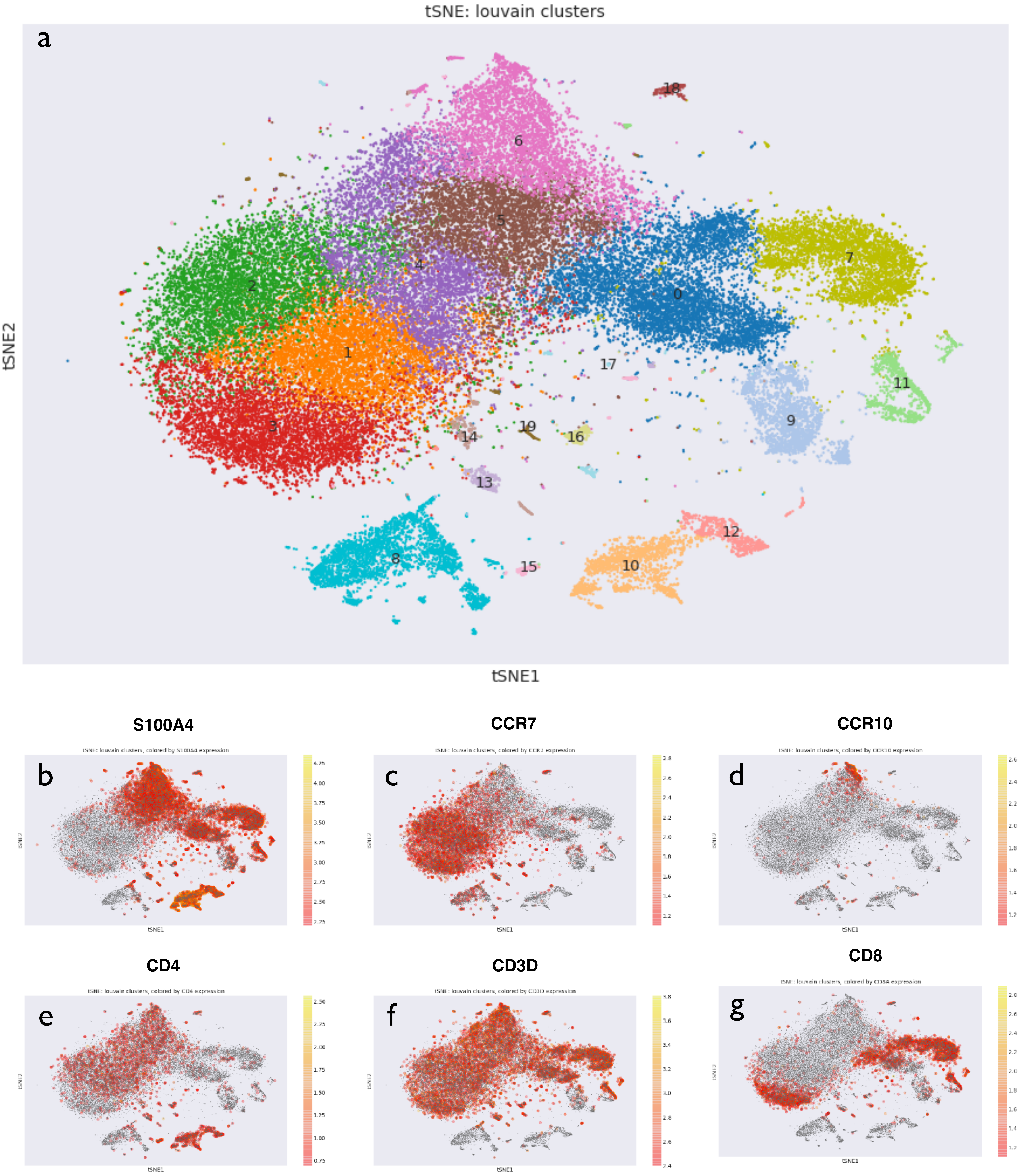
Supplementary Figure 12

Naive Helper T-cells (CD4+/**CD45RA+**/CD25-) vs Memory Helper T-cells (CD4+/**CD45RO+**)



Differential genes between naïve and memory helper T-cells. Naïve helper T cells and memory helper T-cells were purified in Zheng *et al*., 2017 and then independently sequenced with 10x technology. We performed differential expression between these cell types using logistic regression on TCCs and found several genes to be differential, including CD45. In contrast, these genes were not detected when examining only gene counts. The barplots were generated by randomly sampling n=3000 cells per group and the error bars correspond to 95% confidence intervals.

Supplementary Figure 13



S100A4

CCR7

CCR10

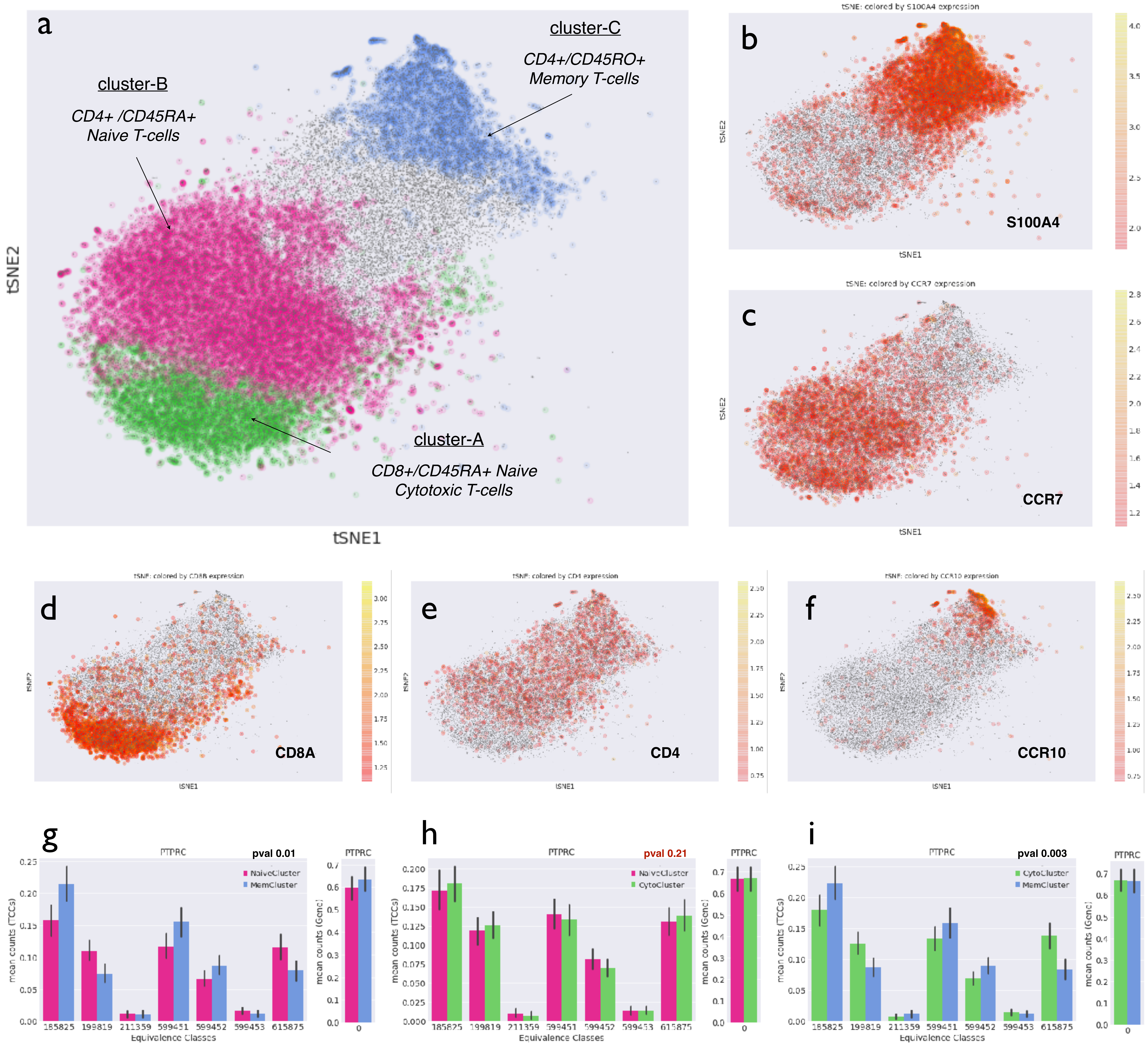
CD4

CD3D

CD8

A *de novo* analysis of 68k PBMCs from Zheng *et al.* 2017. We obtained TCCs with kallisto pseudoalignment, clustered the cells using the Louvain method (a) and plotted the cells with known T-cell markers (b-g). By using TCCs, we were able to differentiate naïve helper, memory helper and naïve cytotoxic T-cells into distinct clusters that are separable. In contrast, Zheng *et al*. 2017 were unable to separate these cell types into distinct clusters.

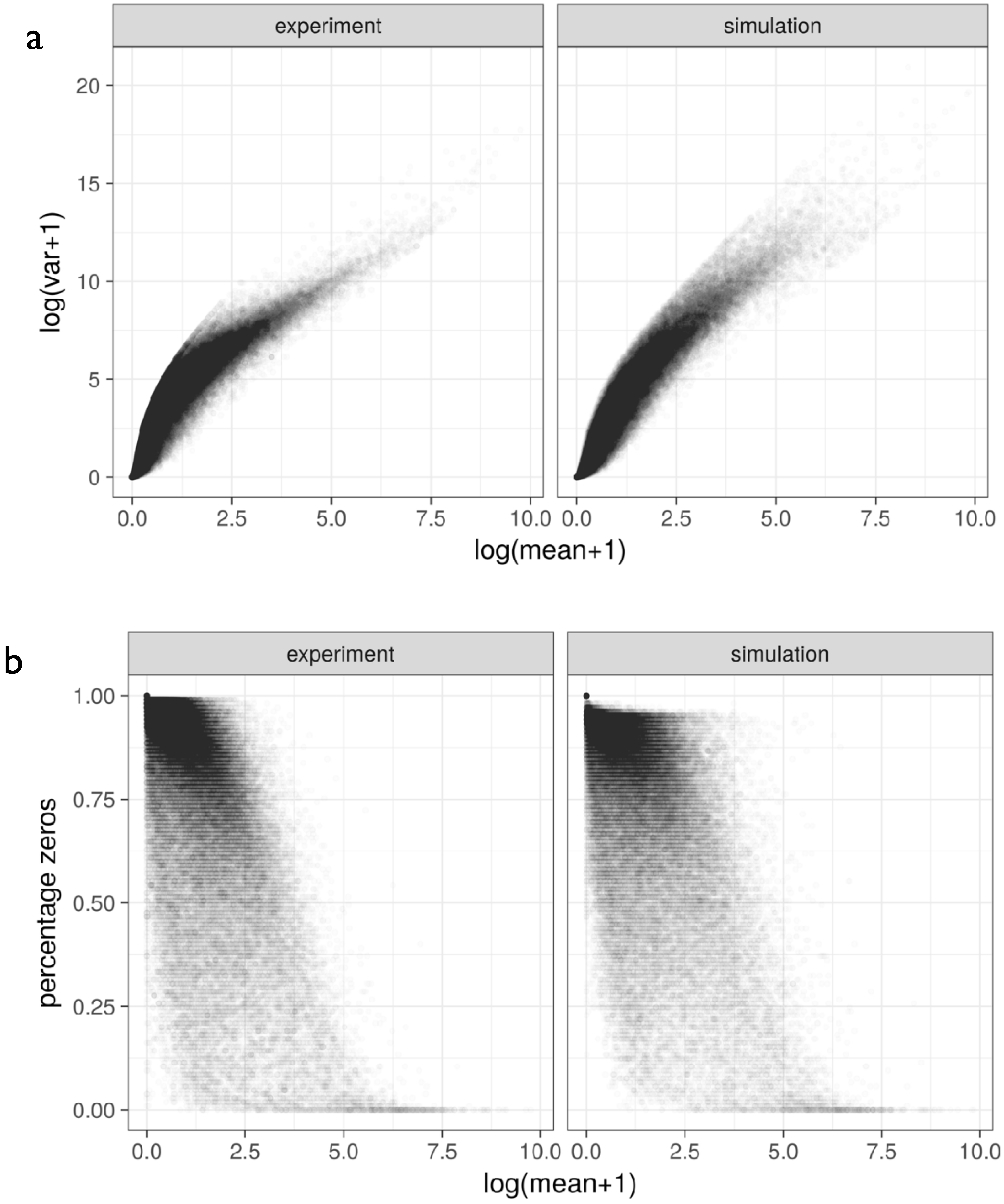
Supplementary Figure 14

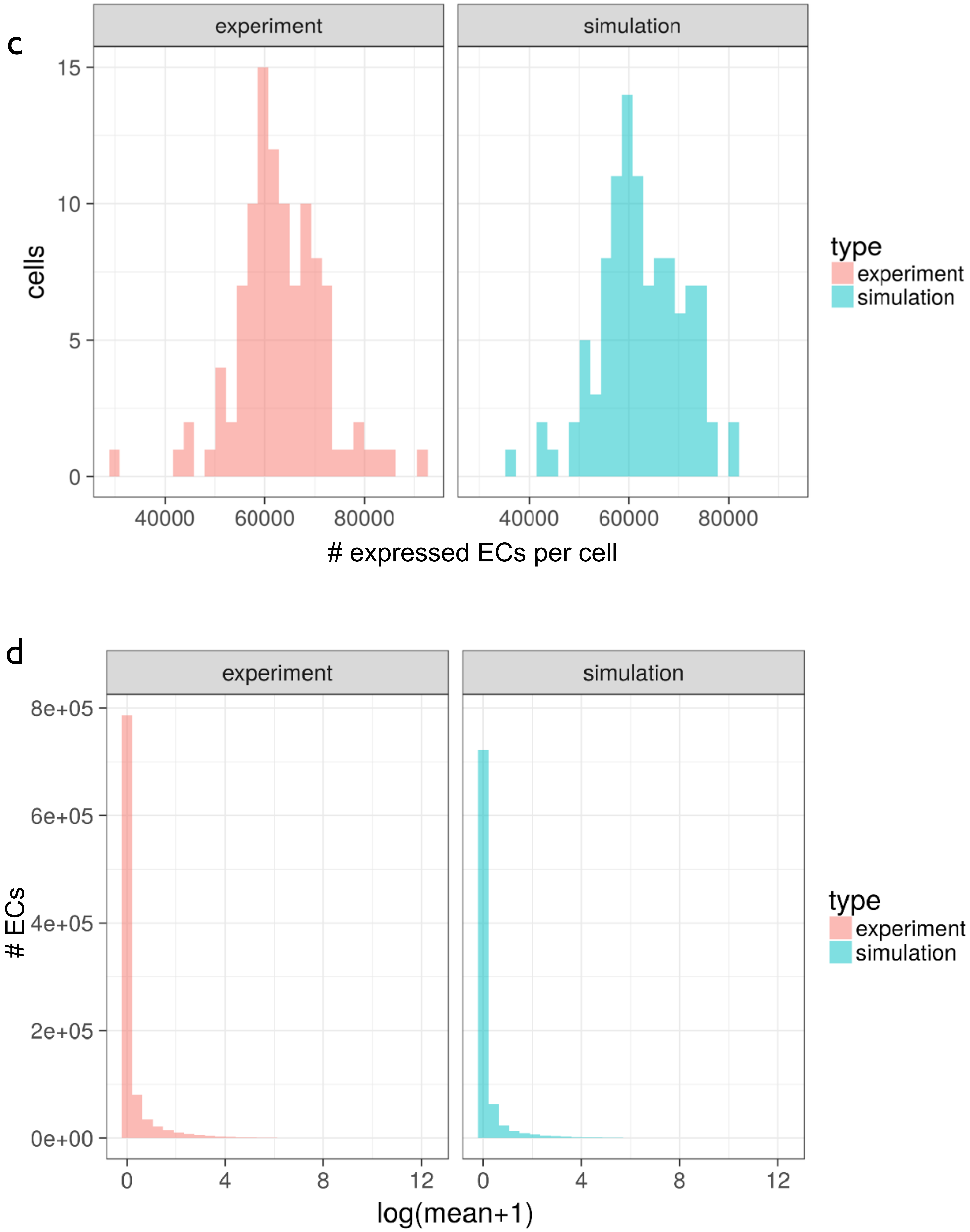


***De novo* analysis of T-cell clusters in 10x data.** A subset of the cells in the 10x data containing naïve, memory and cytotoxic T-cells was analyzed and clustered using TCCs (a). Known naïve, memory and cytotoxic T-cell markers were plotted (b-f) and used to identify the cell clusters. Logistic regression performed on the TCCs in three pairwise differential expression tests, which revealed that CD45 is differential between naïve and memory T-cells (g) and between cytotoxic and memory T-cells (i) with p-value = 0.01 and 0.003 respectively, but not between naïve and cytotoxic T-cells with p-value = 0.21. The p-values repoted in (g,h,i) were averaged over 200 iterations by randomly subsampling n=2000 cells per group and the error bars correspond to 95% confidence intervals.

Supplementary Figure 15

Comparisons between experimental data and simulations





Comparison between experiment and simulation

Our simulated group of non perturbed cells was compared to the myoblast cells from Trapnell *et al*. upon which they are simulated. To compare the mean-variance relationship, each transcript's variance in TPMs was plotted in log-log scale against its mean TPM expression (a). To compare the extent of dropout, each transcript's proportion of zero expression across cells was plotted against its mean TPM expression (b). We also compared the distribution in TCCs between the experimental and simulation data. (c) depicts a histogram of the number of expressed ECs (i.e. nonzero TCCs) per cell. (d) depicts a histogram of the expression in each EC.