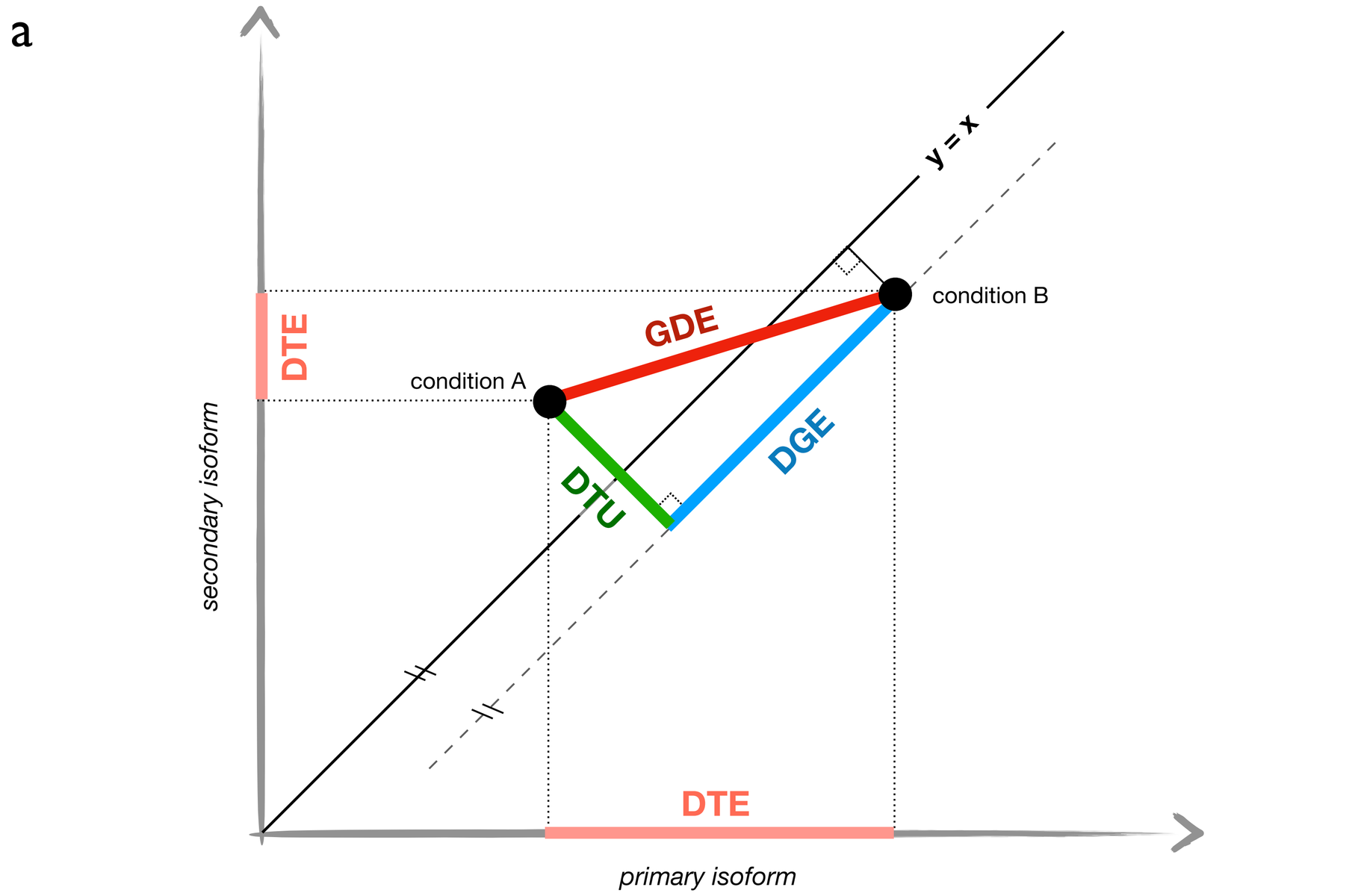
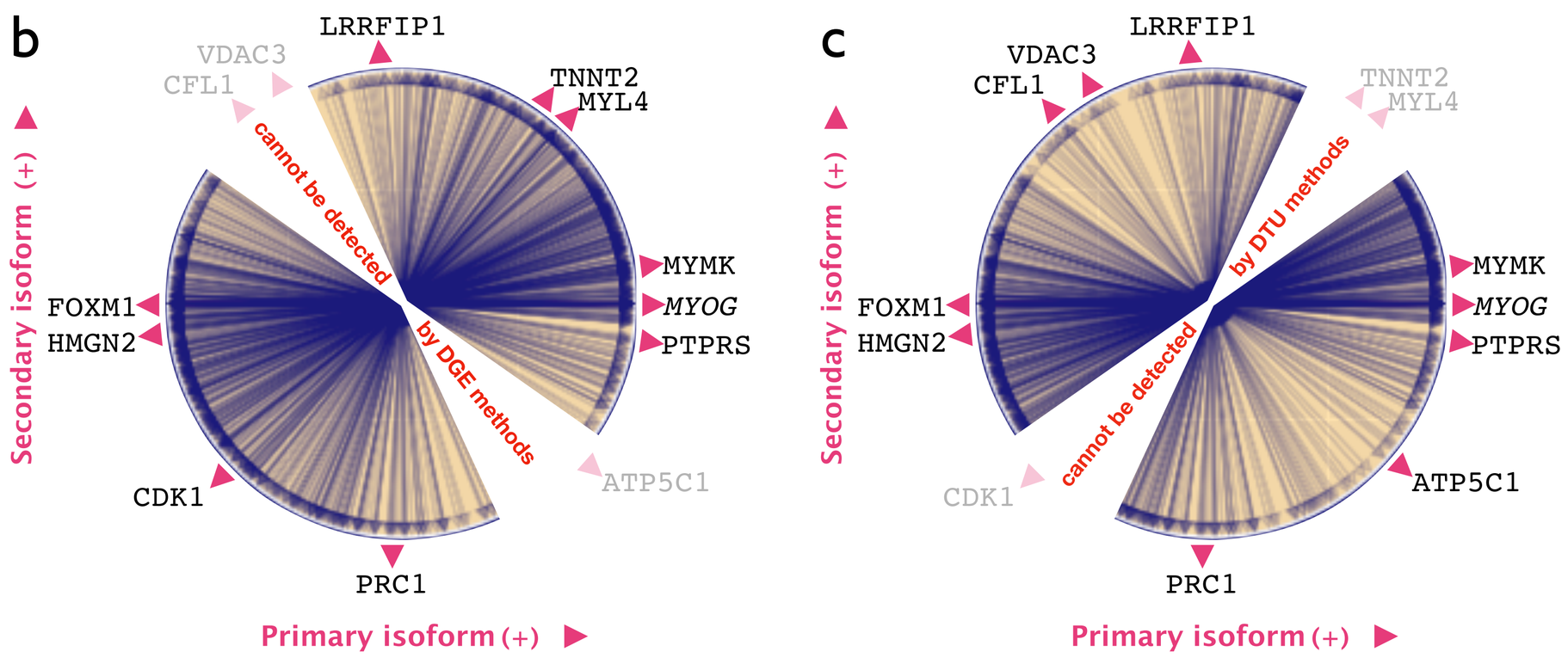
Supplementary Material

Identification of transcriptional signatures for cell types from single-cell RNA-Seq

Vasilis Ntranos, Lynn Yi, Páll Melsted and Lior Pachter

Supplementary Figure 1

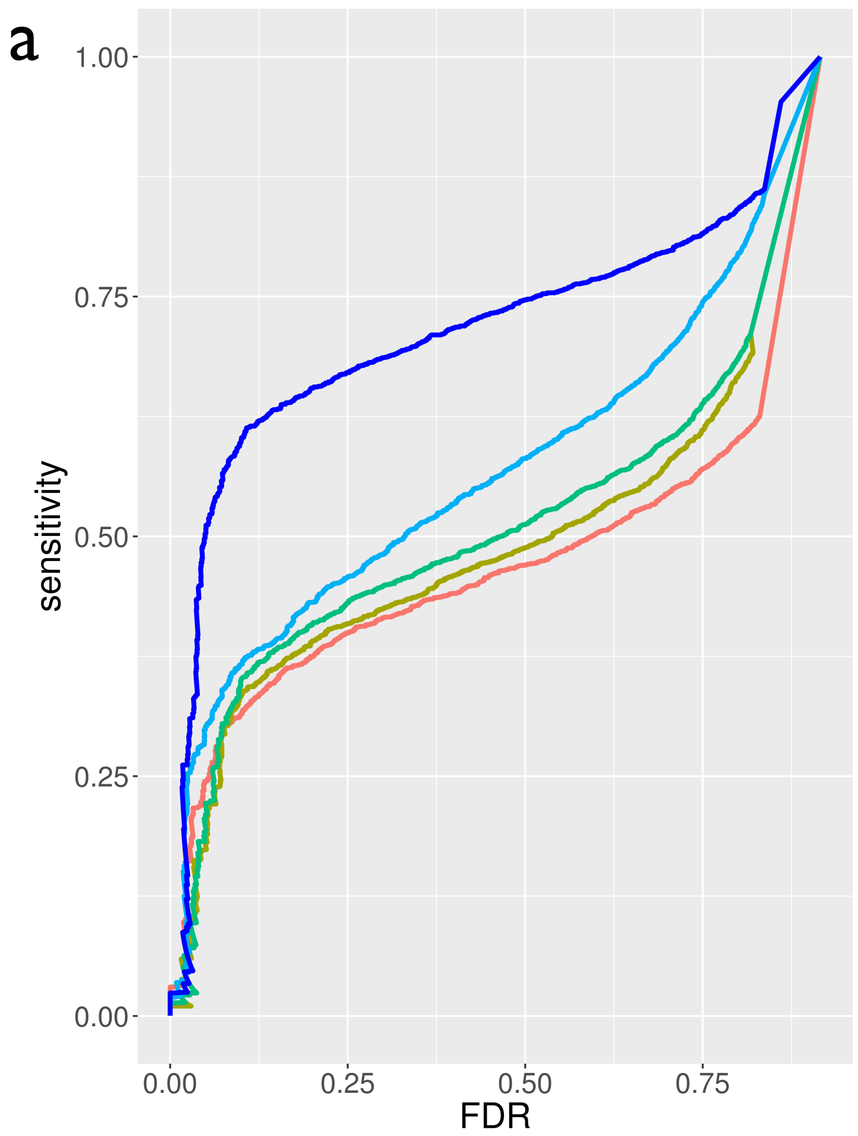


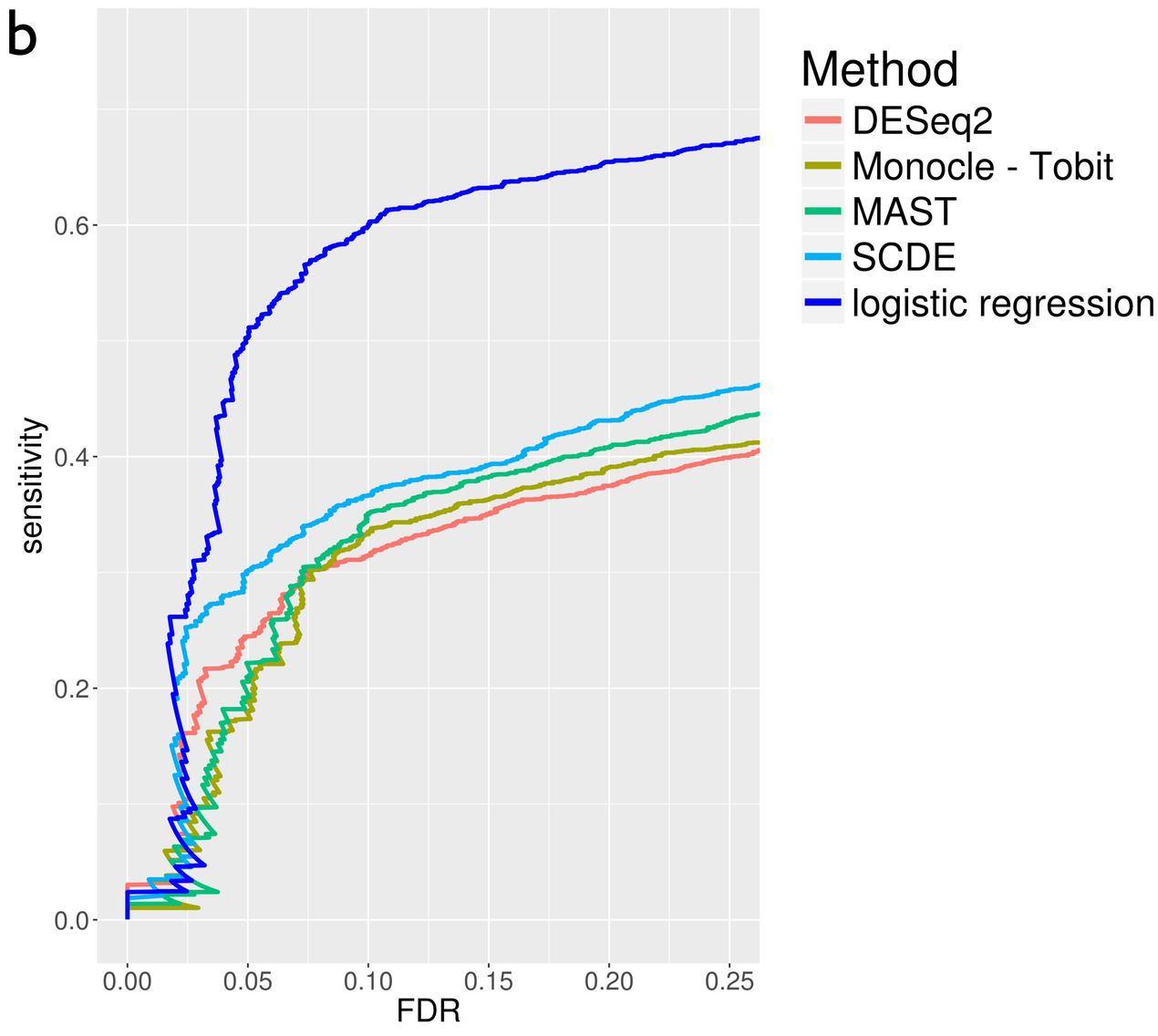


Relationship between differential expression methods. (a) Depiction of the difference in expression of a two-transcript gene in two cell types. The two black points correspond to gene expression in each of the two cell types: the x-coordinate of each point is the expression of its first transcript and 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; the changes are represented by differences between the projections of the points onto the line $y=x$ (blue segment). The projections correspond to summing transcript abundances. Traditional differential transcript usage (DTU) methods test for differential transcript allocation within a gene. This corresponds to 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.

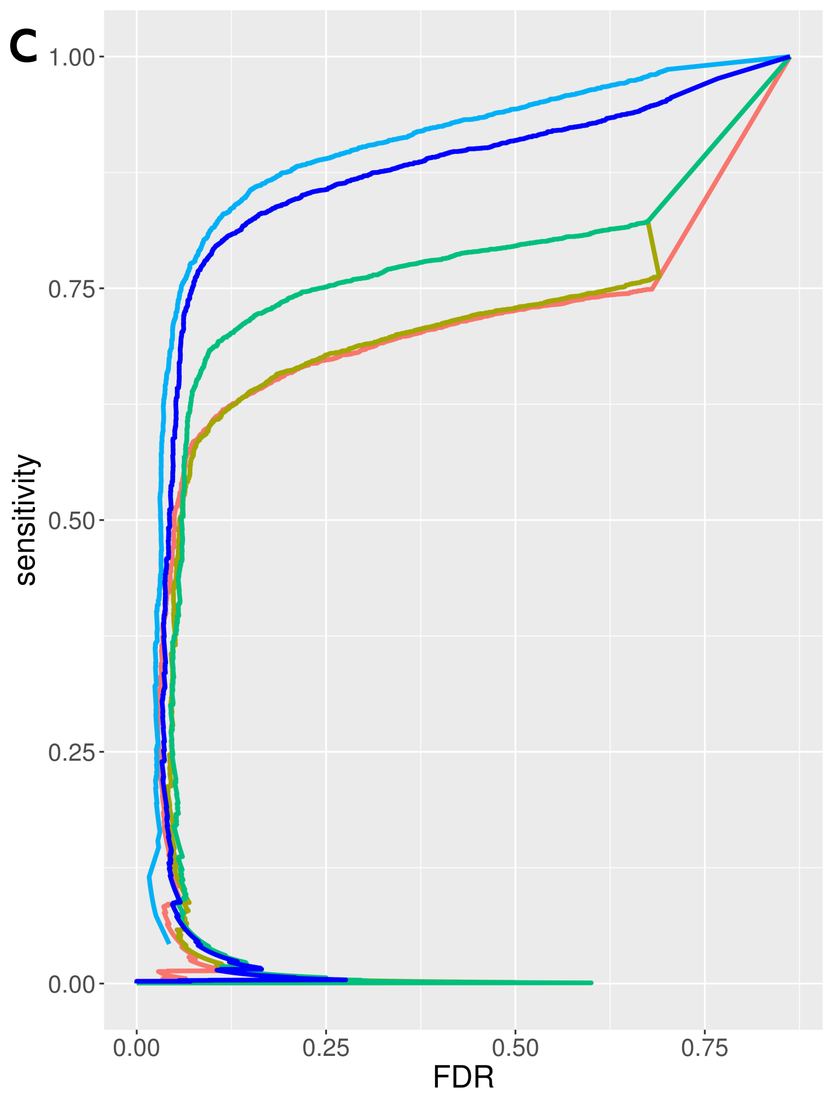
Supplementary Figure 2

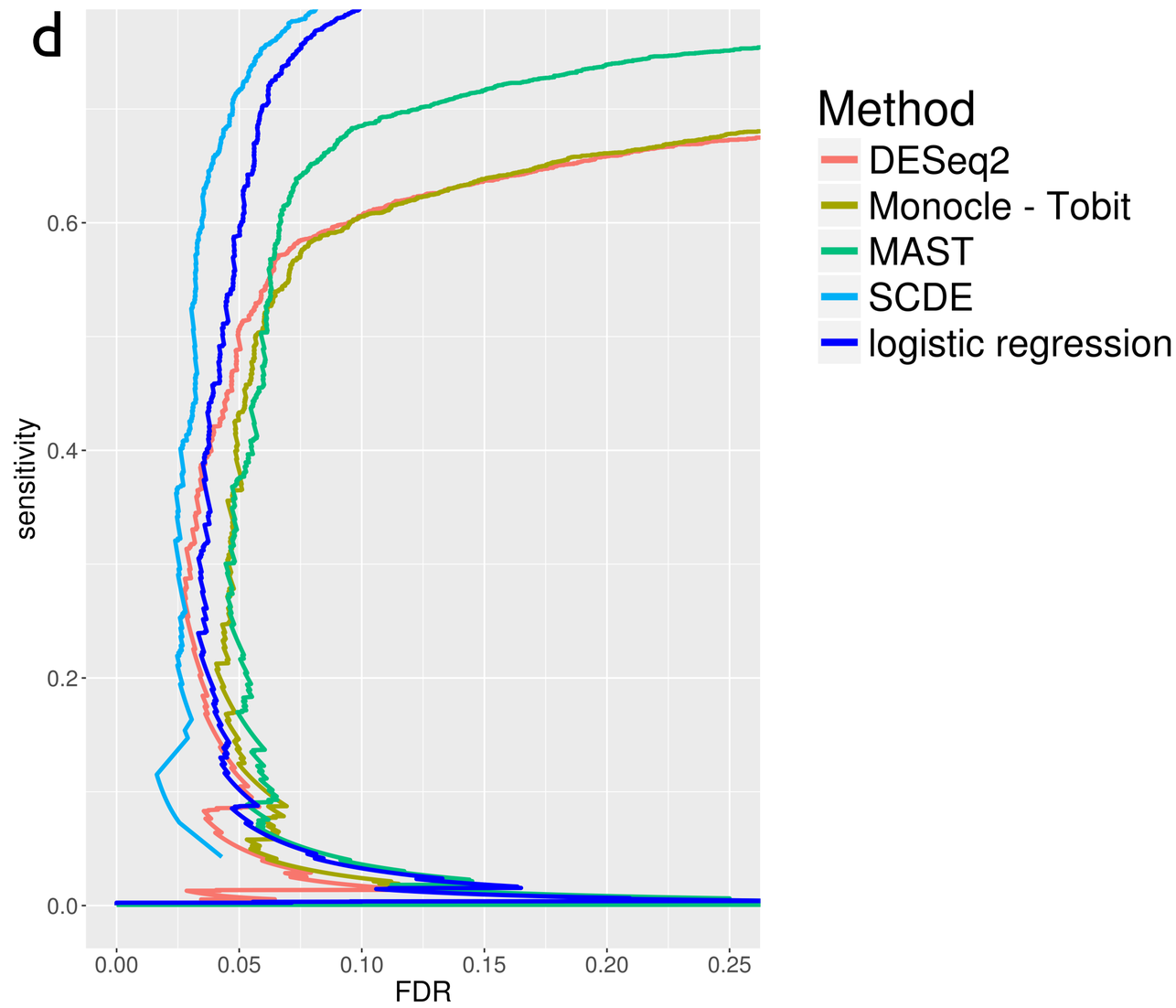
Simulations - Experimental Effect Sizes





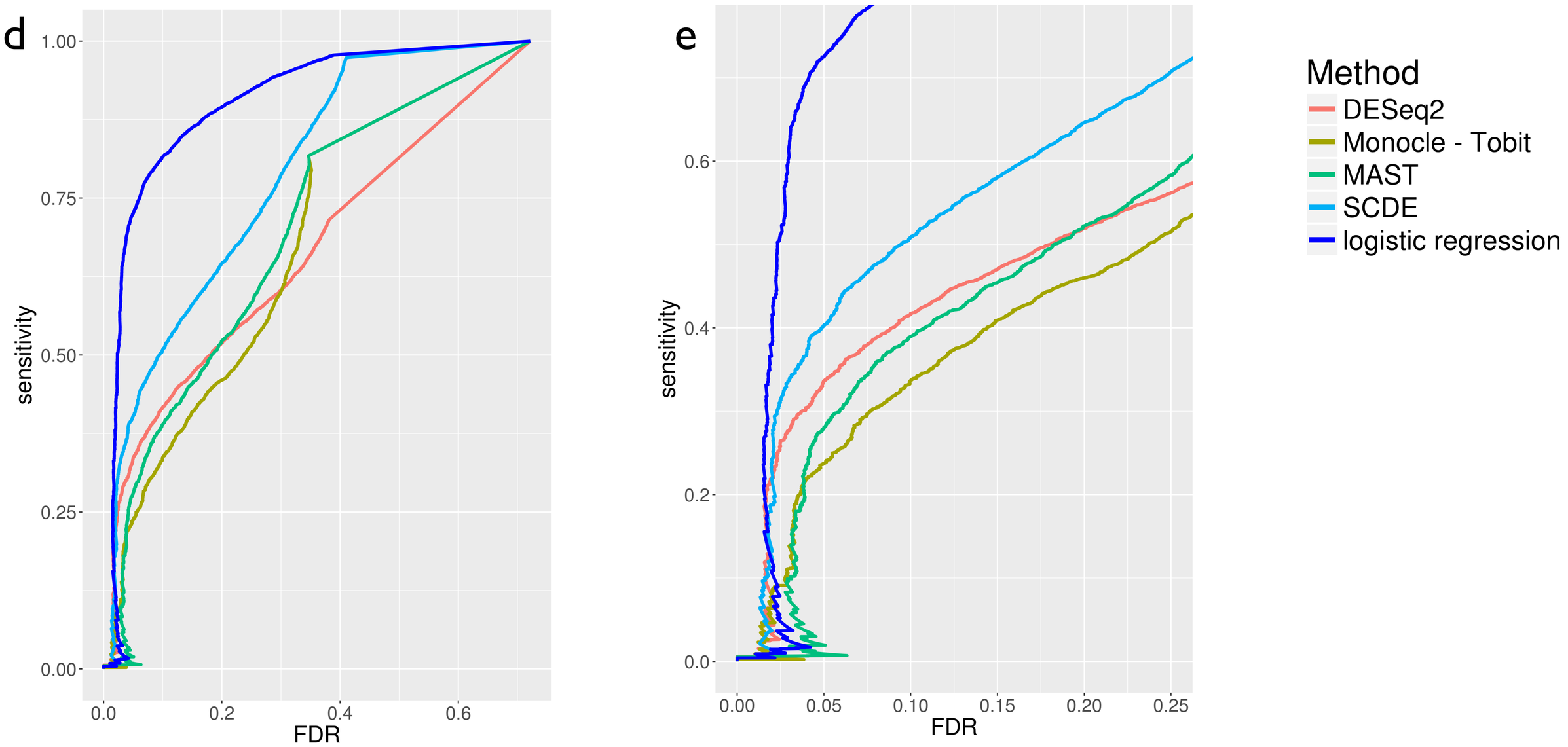
Simulations - Correlated Effect Sizes





Supplementary Figure 2

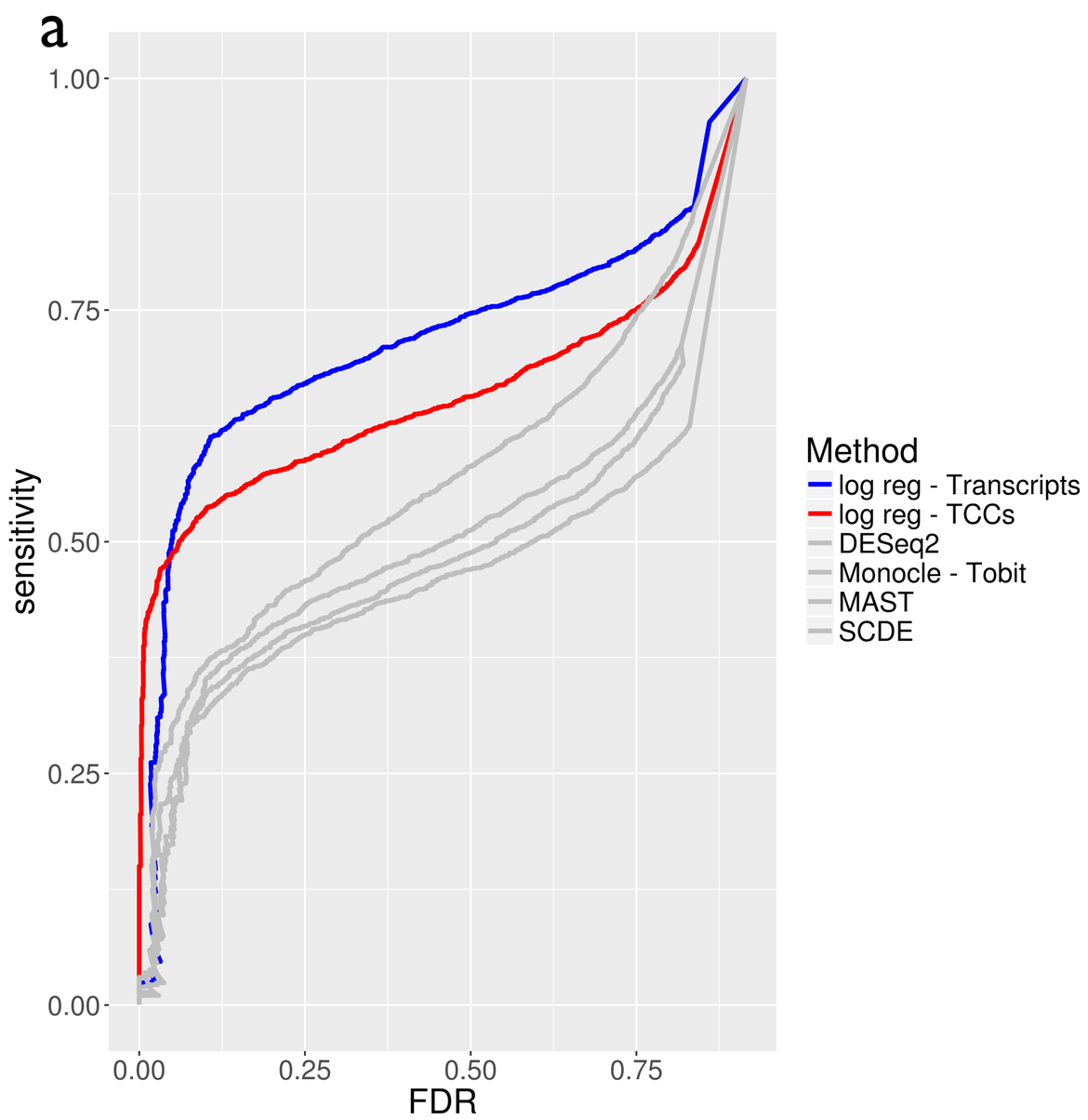
Simulations - Independent Effect Sizes

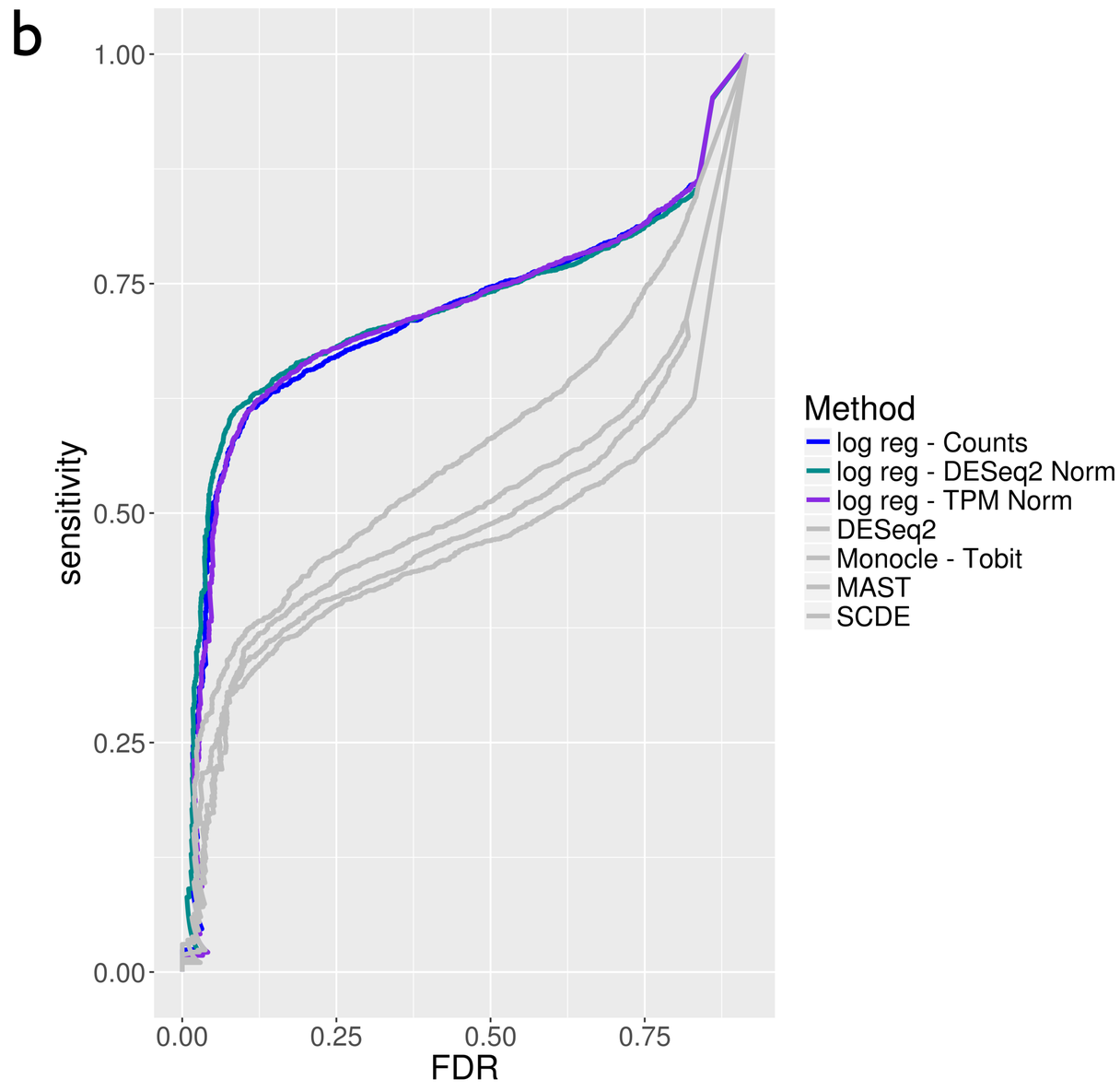


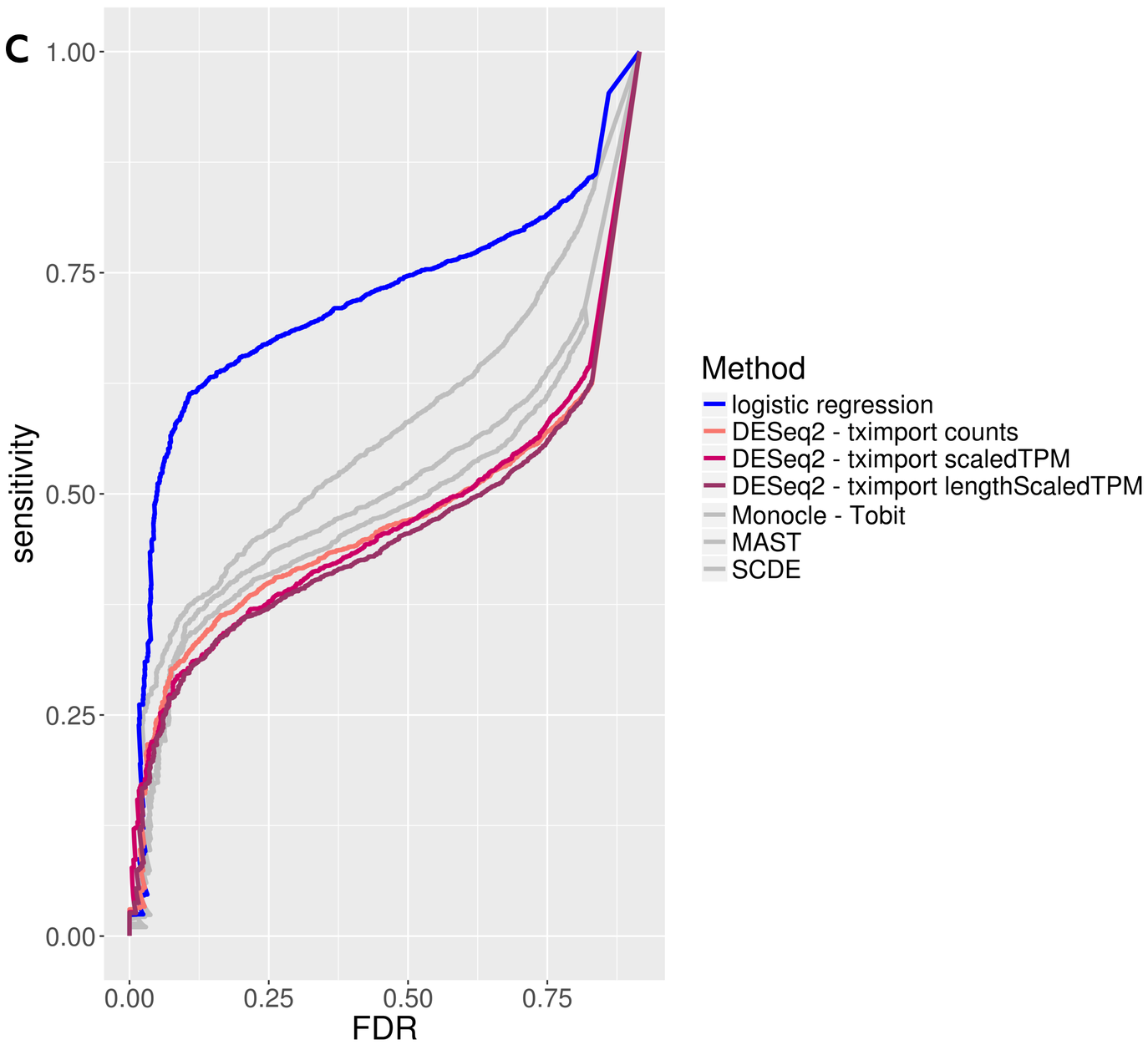
Performance of differential expression methods on simulations. A uniformly sequenced scRNA-seq dataset containing two cell types, each with 105 cells, was simulated. In (a, b-zoomed in), effect sizes were derived from an experiment. In the correlated effect size simulation (c, d), genes were chosen independently to be perturbed, and all transcripts corresponding to the same gene were perturbed in the same direction. In the independent effect size simulation (e, f), transcripts were independently chosen to be perturbed. Five differential expression methods were tested on these simulations and their FDR-sensitivity plots are depicted.

Supplementary Figure 3

Simulations - Experimental Effect Sizes



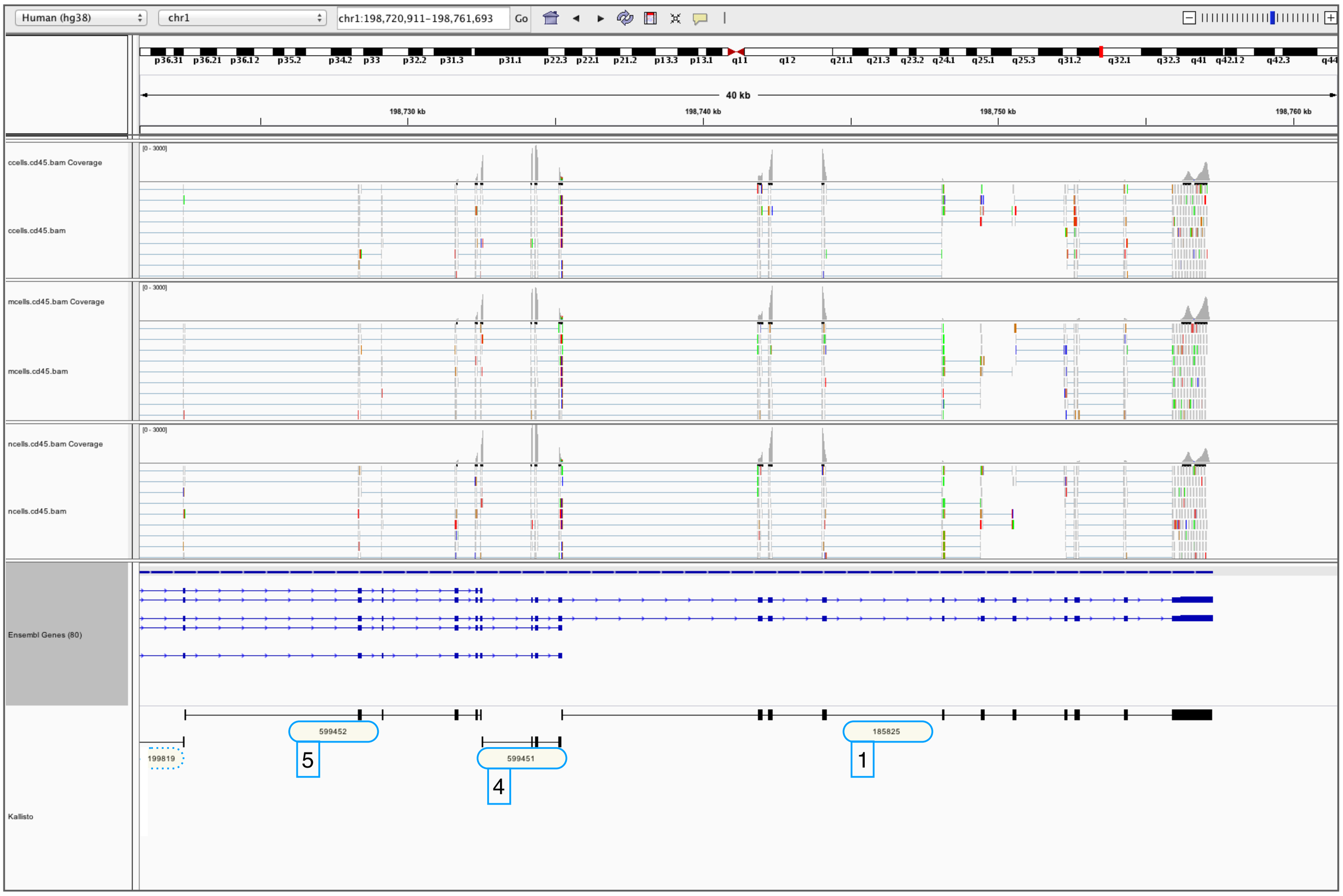




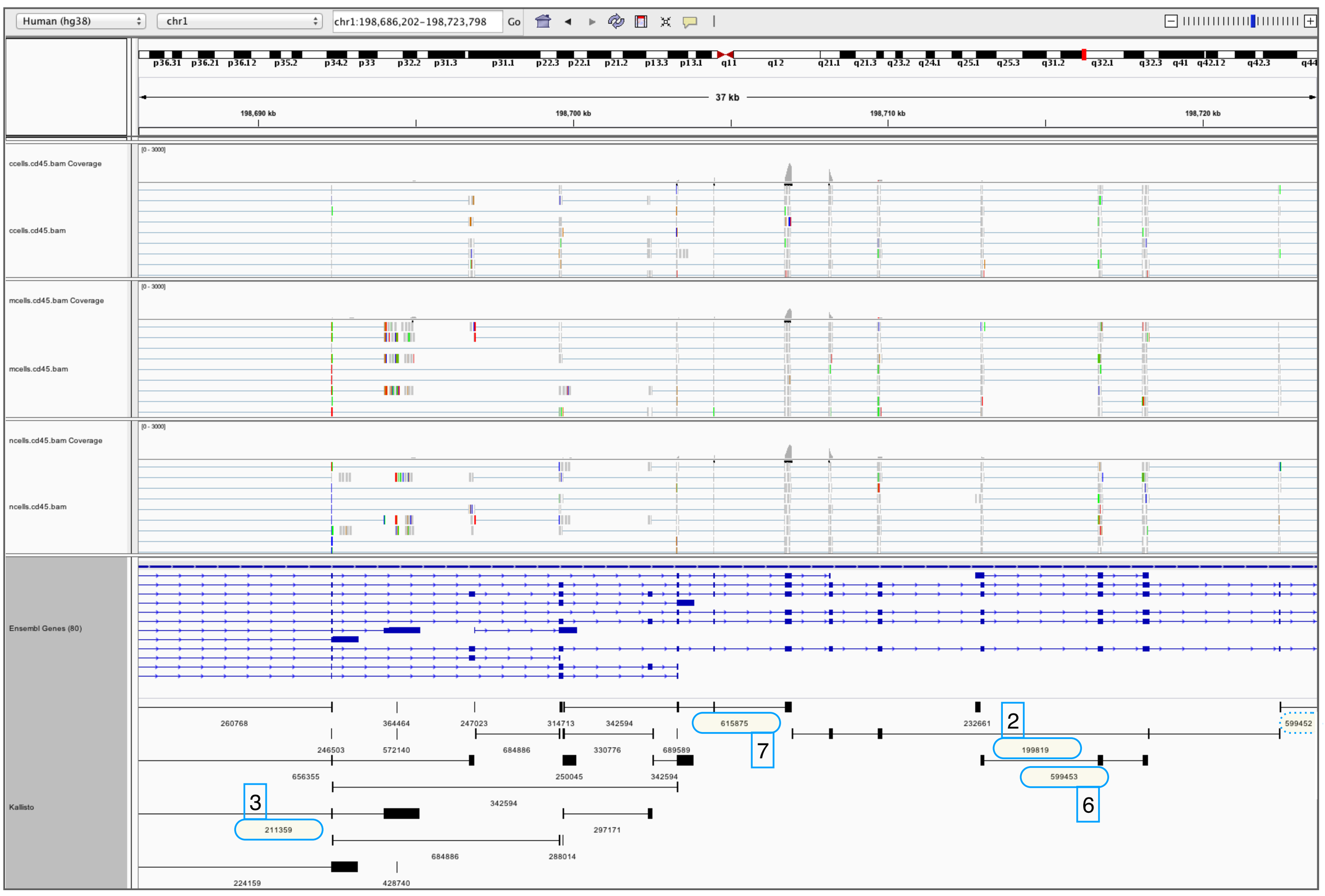
Performance of logistic regression on the simulation based on experimental effect sizes. The performance of logistic regression on transcript quantifications was compared with that of logistic regression on TCCs (a). Three different normalization methods: transcript counts, size factor normalization from DESEq2 and transcript-per-million (TPM) normalization, were also compared on the same simulation (b). Finally, we compared methods of summing transcript counts and abundances to gene counts and abundances using tximport, prior to differential gene expression analysis with DESeq2 (c).

Supplementary Figure 4

a



b



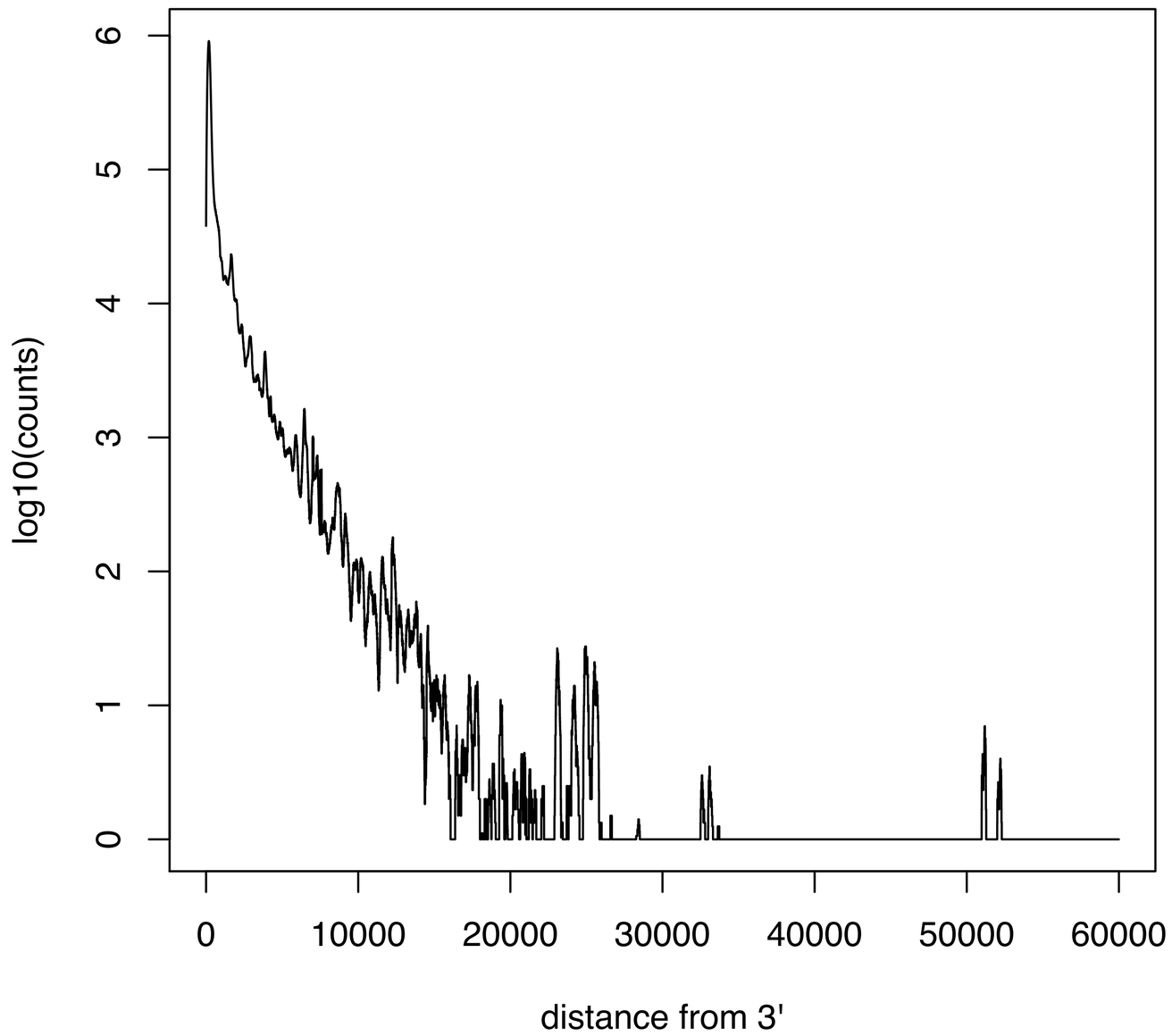
Supplementary Figure 4

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 pseudobam option outputs a BAM file for each sample that can be visualized 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).

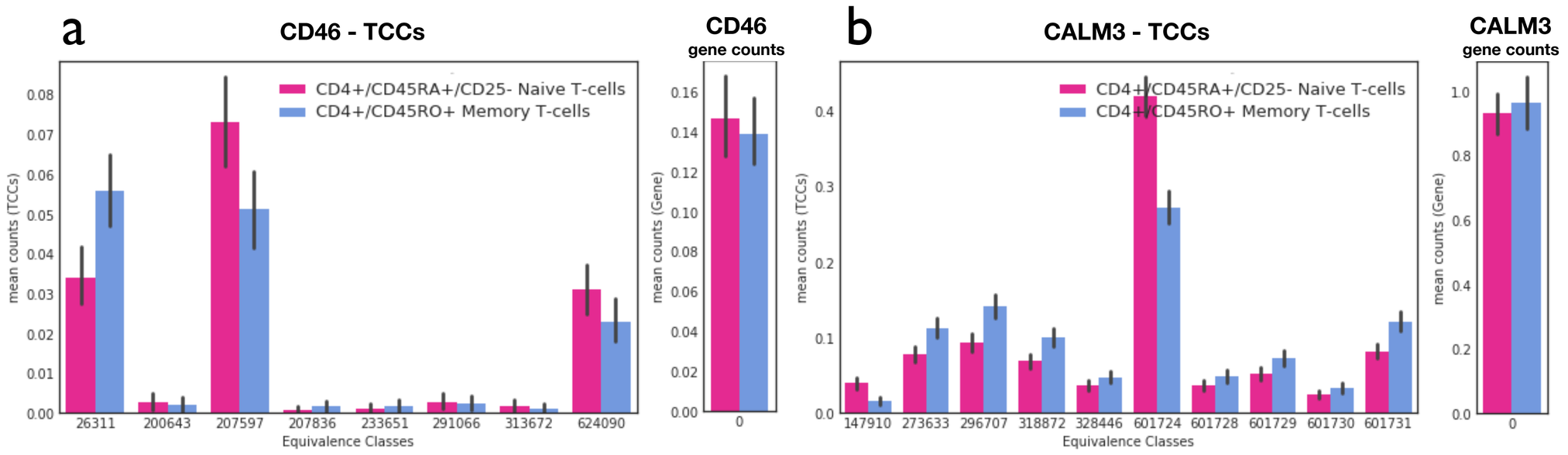
Supplementary Figure 5



<u>The distribution of read distance from the 3' end from Zheng *et al.*, 2017.</u> The substantial number of reads distal to annotated 3'-ends suggests a large number of unannotated 3' UTRs whose reads are informative when transcript compatibility counts are utilized.

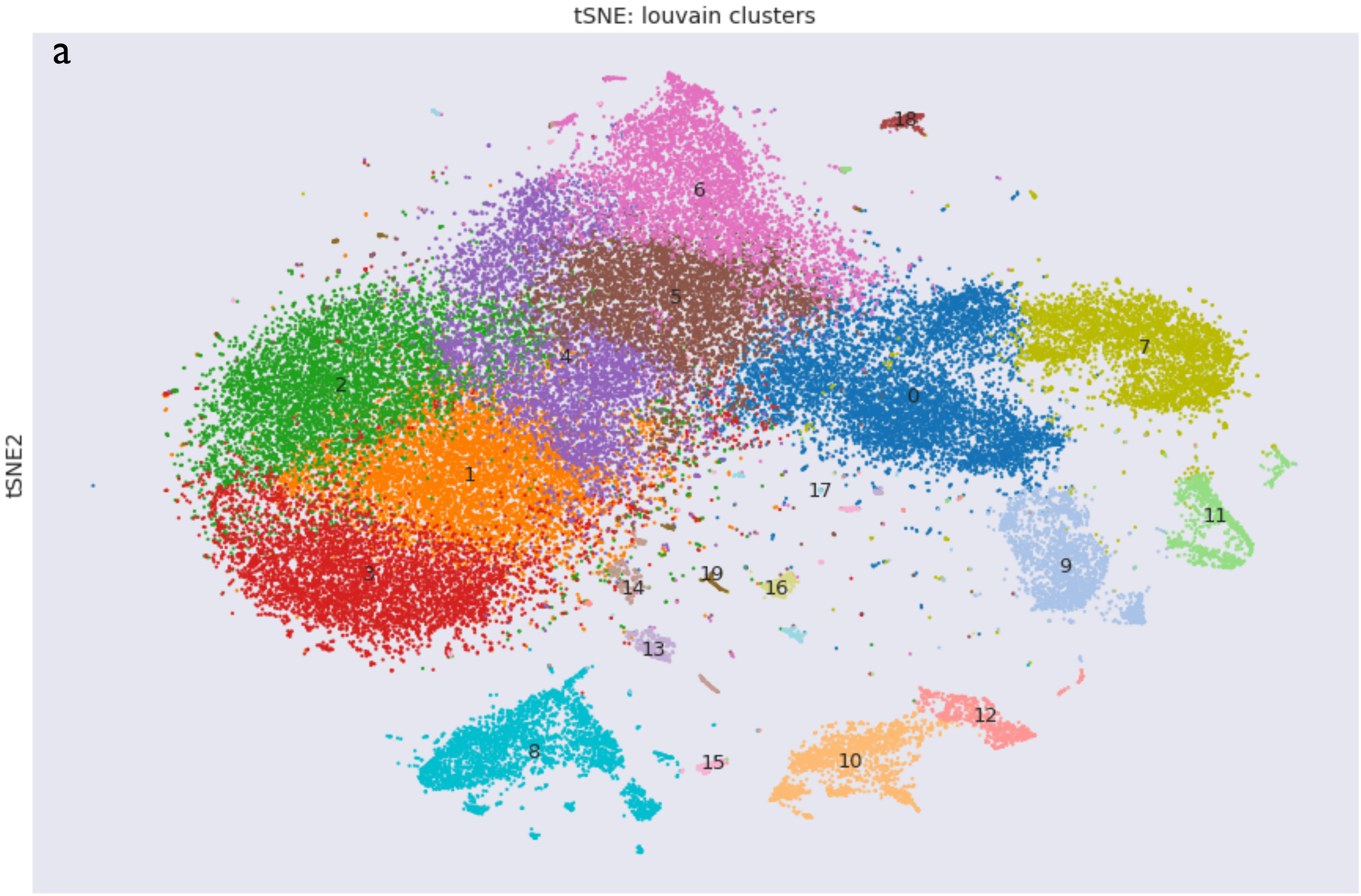
Supplementary Figure 6

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



<u>Differential genes between naïve and memory helper T-cells.</u> 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.

Supplementary Figure 7

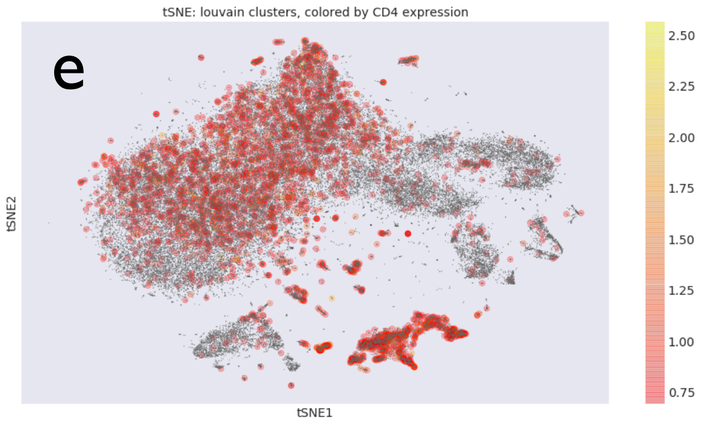


S100A4

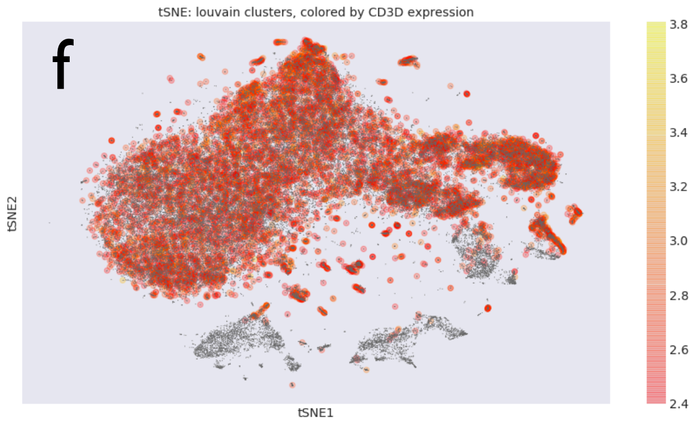
CCR7

CCR10

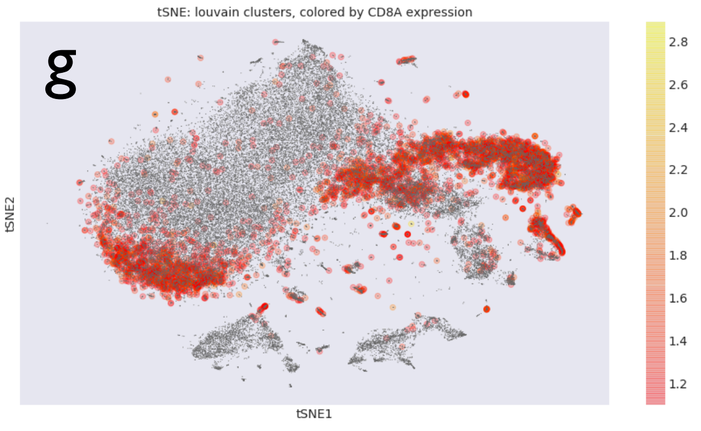
CD4



CD3D

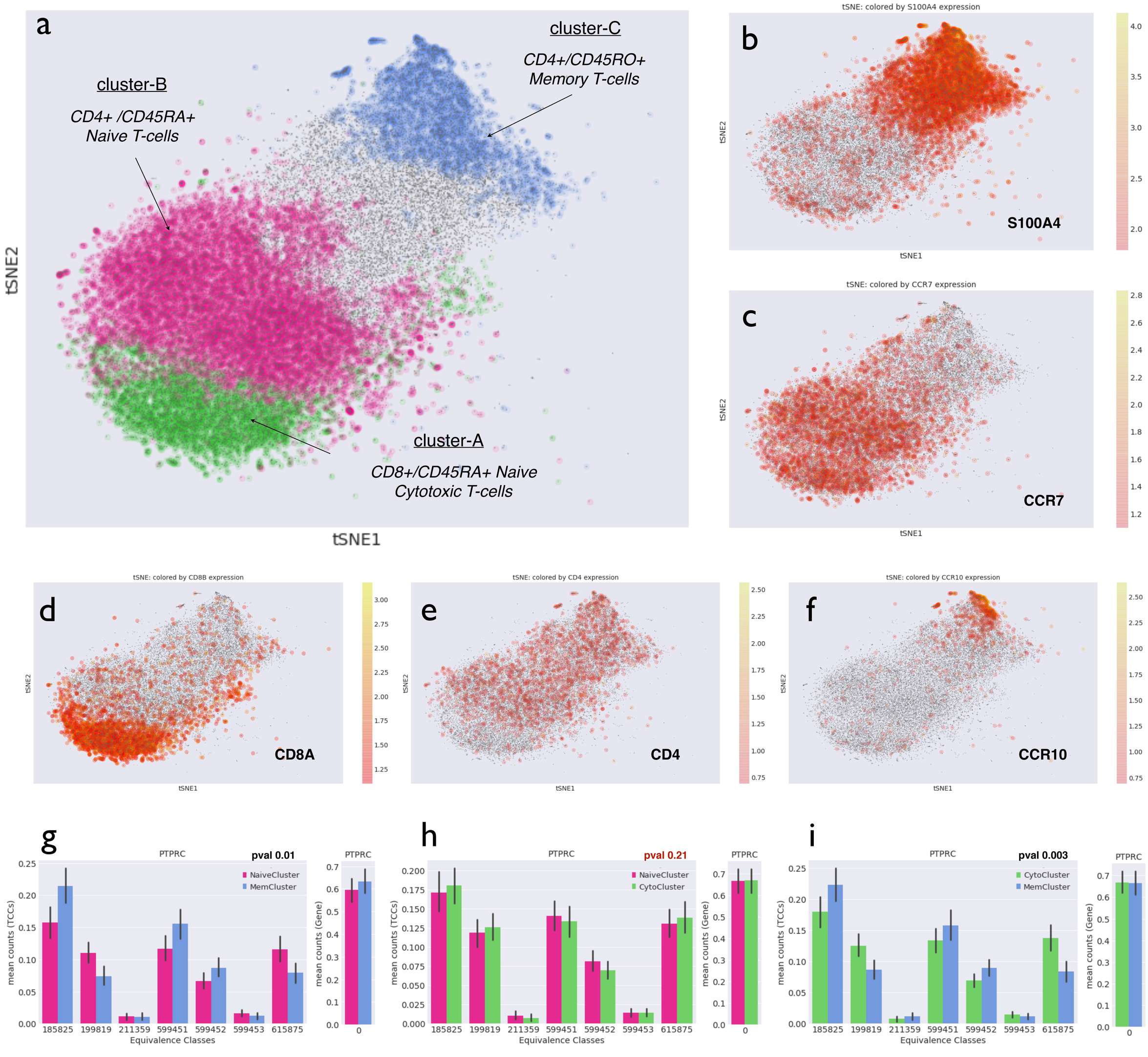


CD8



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 8



***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.