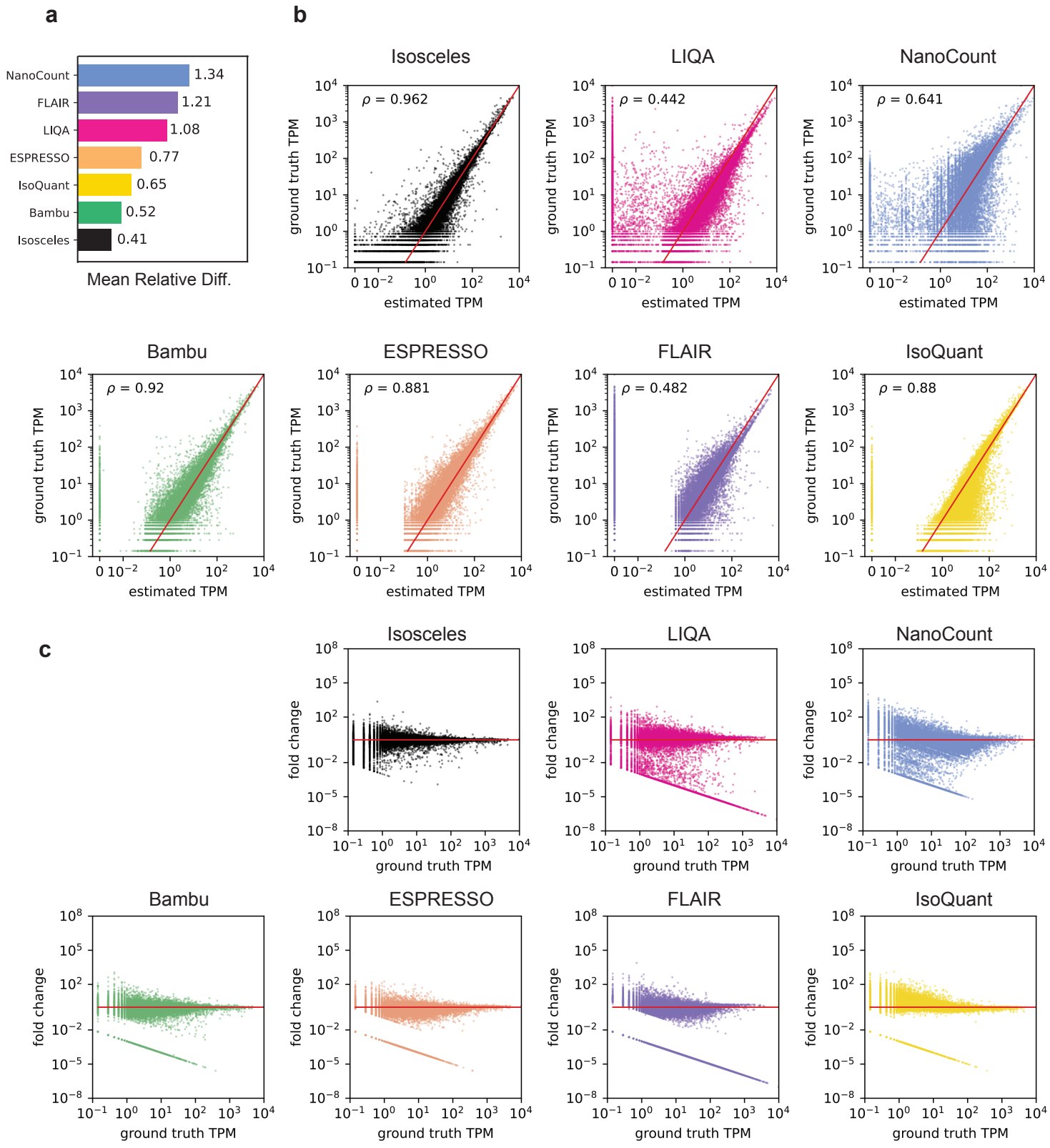
Supplementary Information

Accurate long-read transcript discovery and quantification at single-cell, pseudo-bulk and bulk resolution with Isosceles

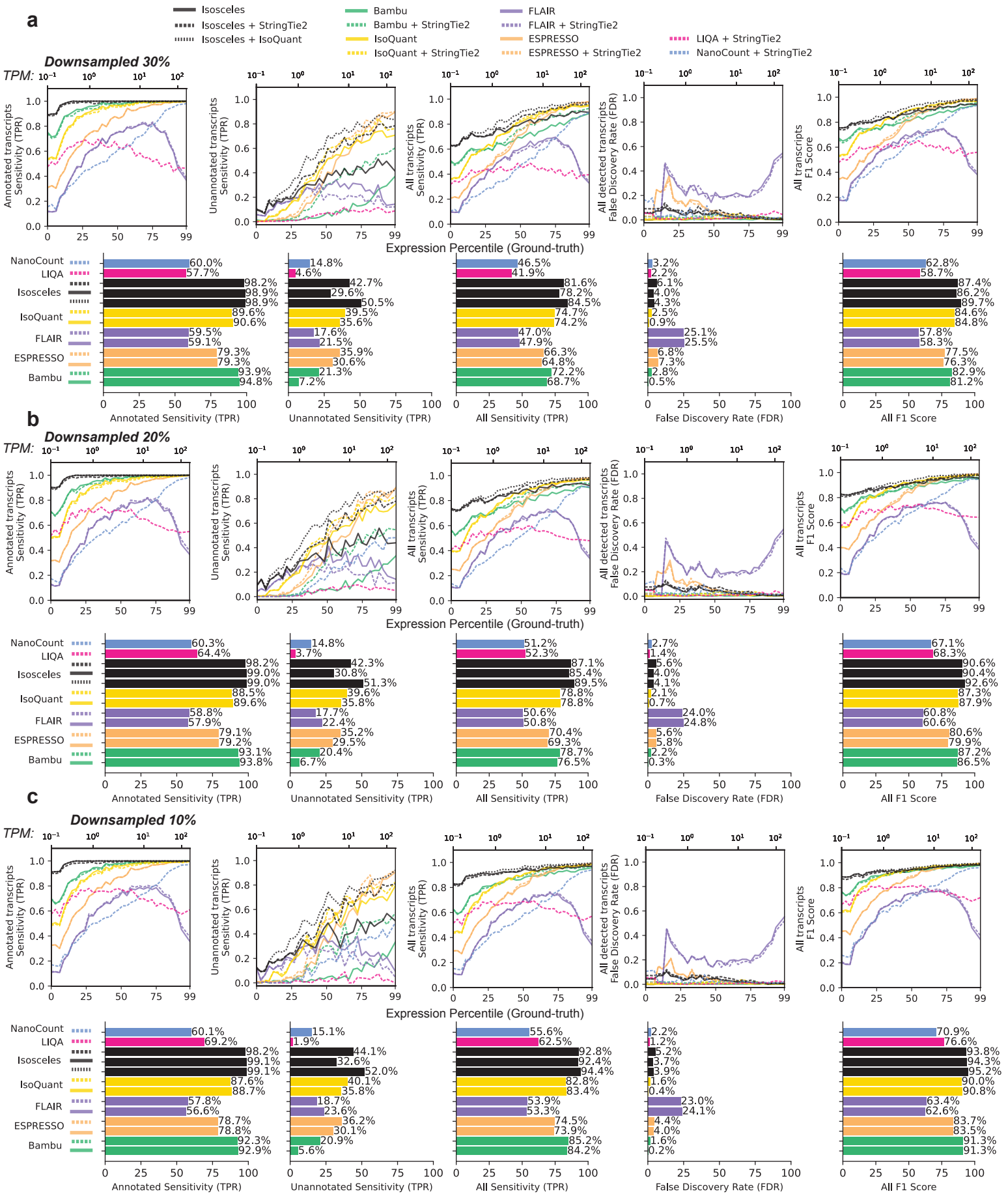
Michal Kabza[1], Alexander Ritter[2], Ashley Byrne[3], Kostianna Sereti[4], Daniel Le[3], William Stephenson[3], Timothy Sterne-Weiler[2,4,*]

[1] Roche Informatics, F. Hoffmann-La Roche Ltd., Poznań, Poland.
[2] Computational Biology & Translation, Genentech Inc., South San Francisco, USA.
[3] Department of Next Generation Sequencing and Microchemistry, Proteomics and Lipidomics, Genentech Inc., South San Francisco, USA.
[4] Department of Discovery Oncology, Genentech Inc., South San Francisco, USA.

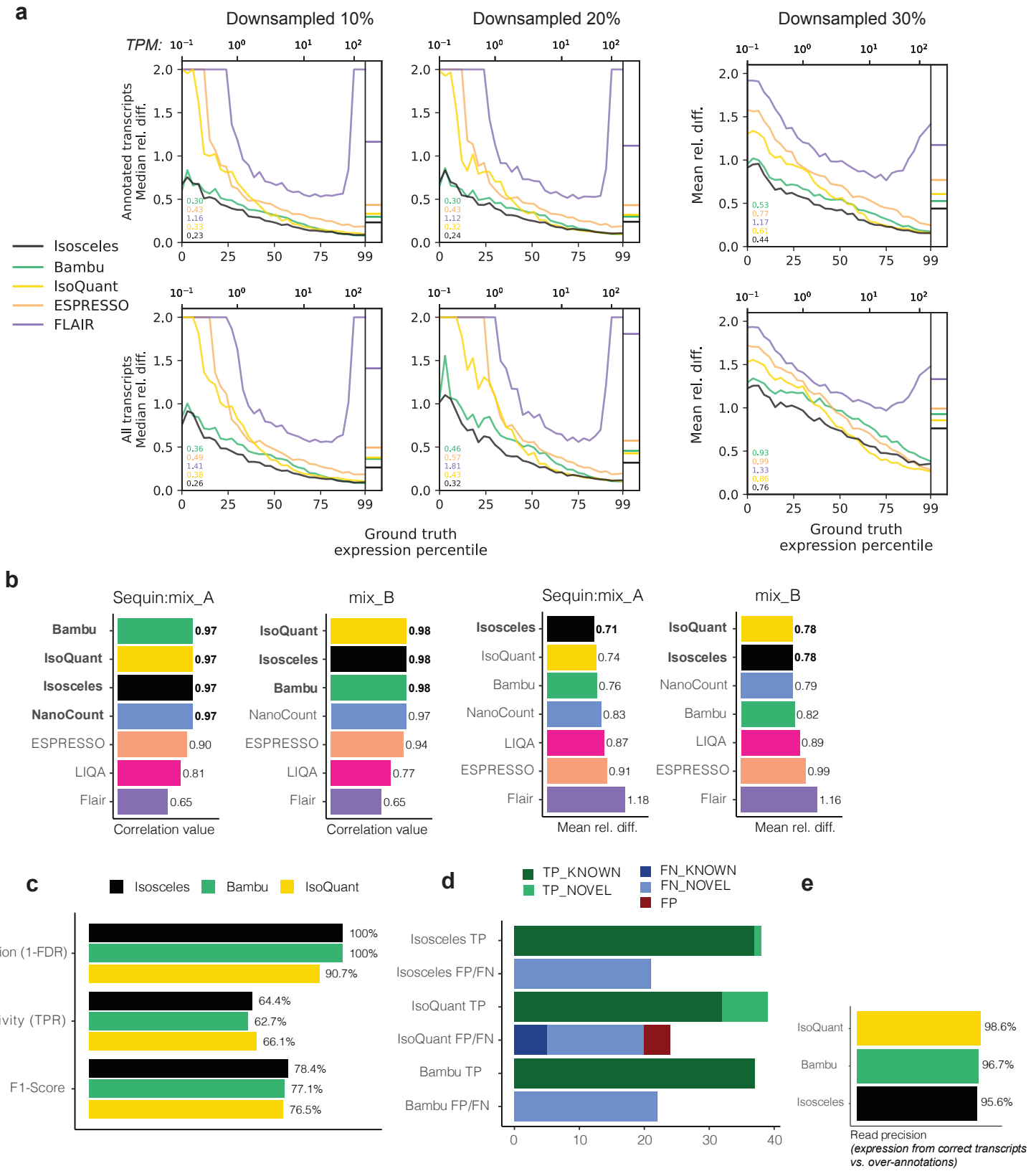
* Correspondence should be addressed to: sterneweiler.timothy@gene.com



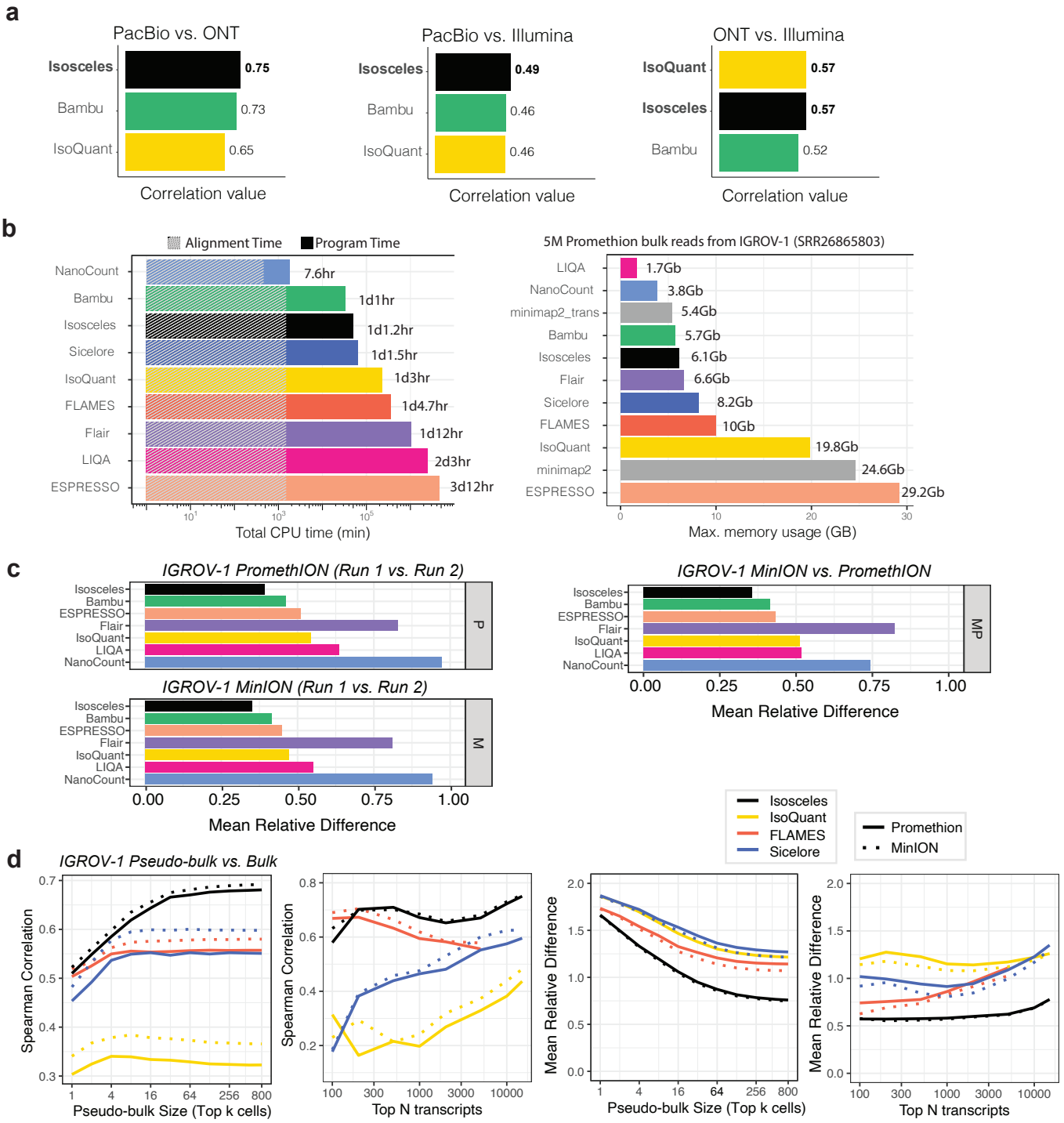
Supplementary Fig. 1– (a) Continued from Fig. 2a, mean relative difference of ground-truth vs. estimated TPM. (b) Scatter plots (with labeled Spearman coefficient) of estimated vs. ground-truth TPM values on log scale. Estimated TPM values below 0.001 are manually assigned a value of 0.001 on the plot. (c) MA plots of the fold change between estimated and ground-truth TPM vs. ground-truth TPM values on log scale. Estimated TPM values below 0.001 are manually assigned a value of 0.001 for the fold change calculation. Source data are provided as Source Data files.



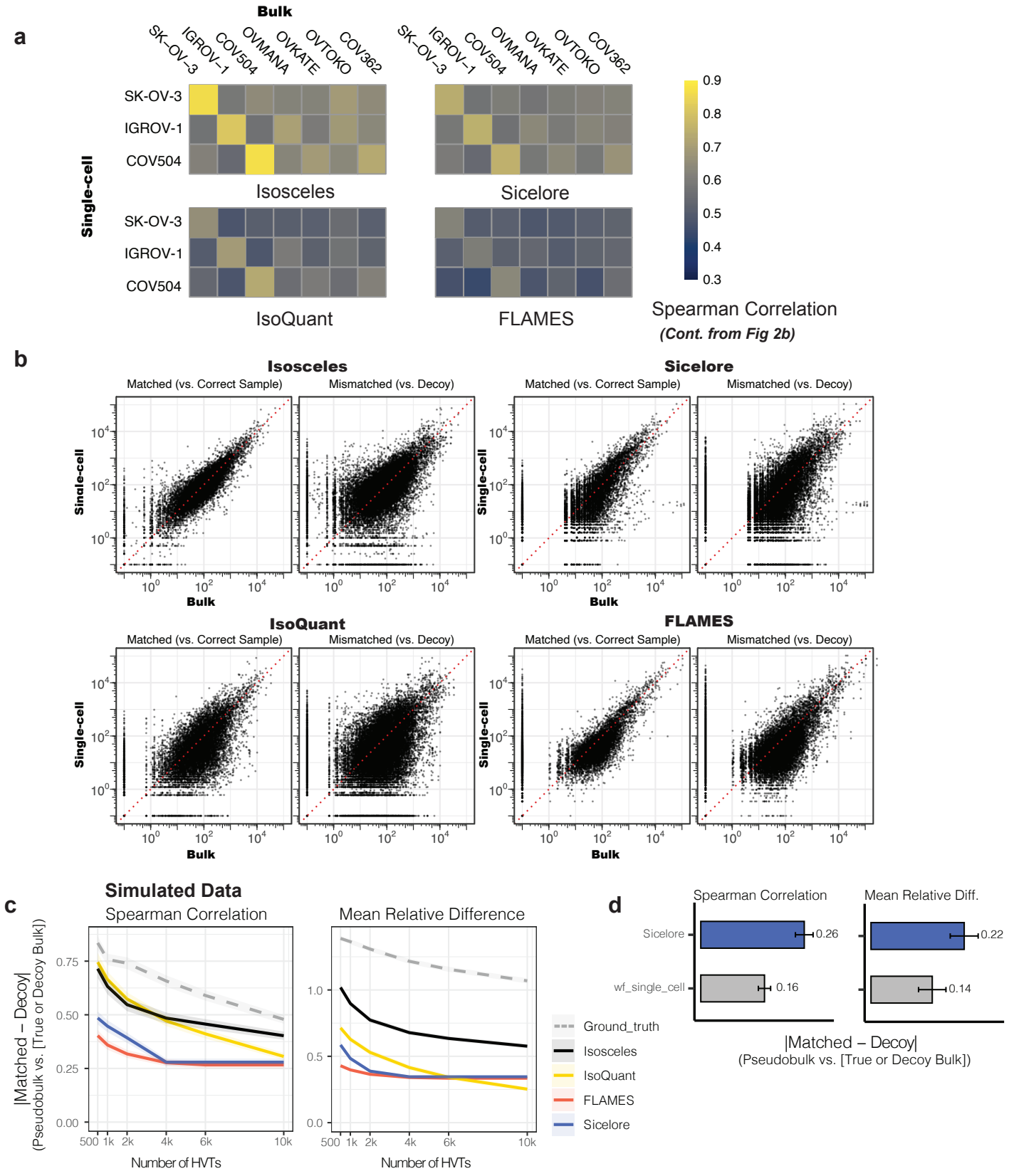
Supplementary Fig. 2– (a-c) Continuation of results from Fig. 2b for all combinations of programs (single-program shown as solid, and combinations given as dashed lines) tested at 10%, 20%, and 30% of downsampled simulated transcripts (see Fig. 2b for additional legend and description). Source data are provided as Source Data files.



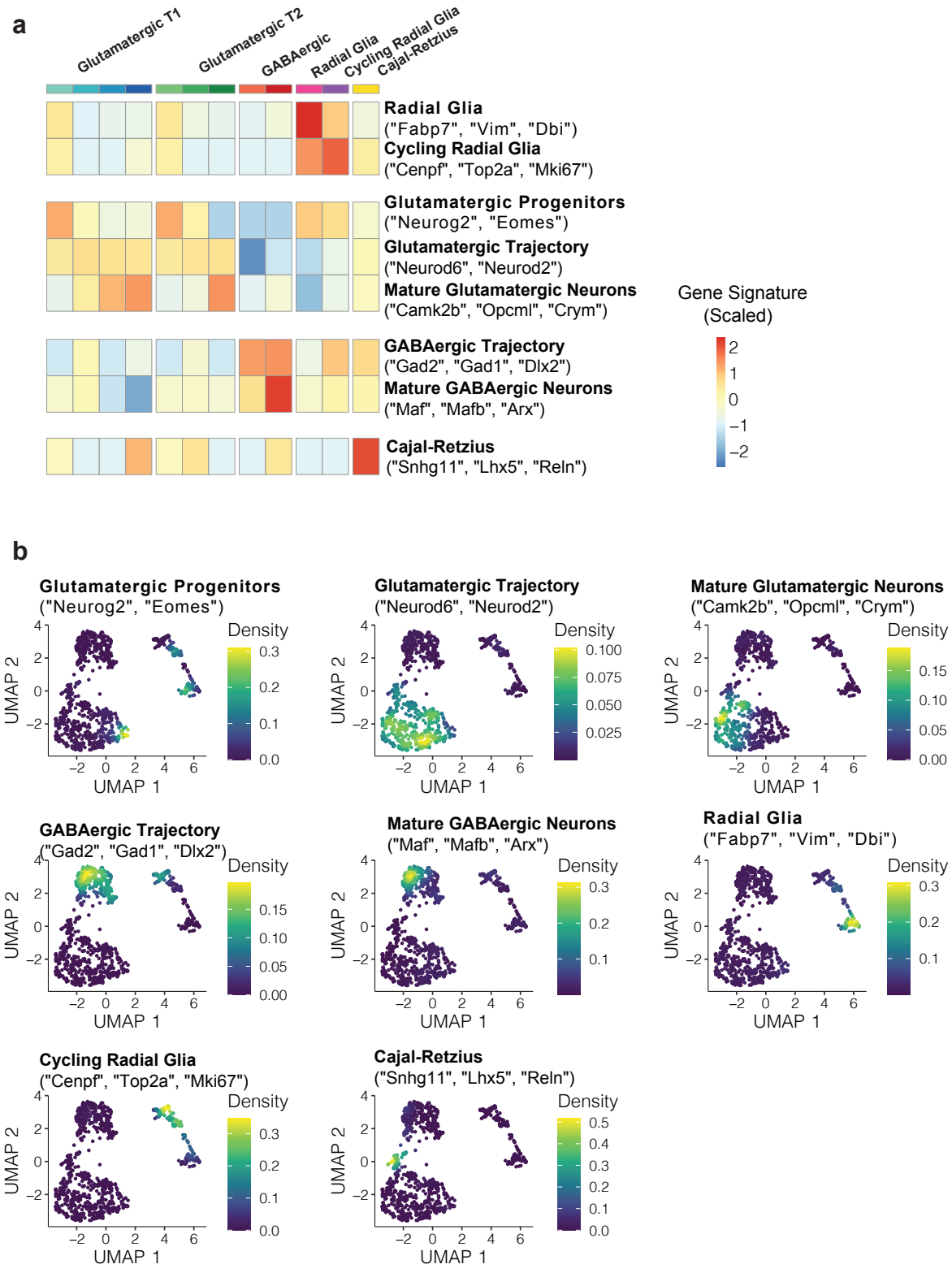
Supplementary Fig. 3– (a) Median relative difference for the additional 10% and 20% downsampling (left; see Fig. 2c for additional legend and description). Mean relative difference for the 30% downsample shown (10% and 20% is concordant, data not shown). (b) Spearman correlation and mean relative difference of ground-truth vs. method estimated TPM for Sequin mixes A and B. (c) Precision, recall, and F1 score metrics of spliced transcript detection using Lexogen provided annotated and withheld transcript sets for SIRV data (ie. insufficient annotation benchmark). (d) Number of true positives (TP), false positives (FP) and false negatives (FN) identified for SIRV data for panel c. (e) Precision of read assignment to correct annotations vs. decoys in SIRV data (using over-annotation benchmark, see Methods). Source data are provided as Source Data files.



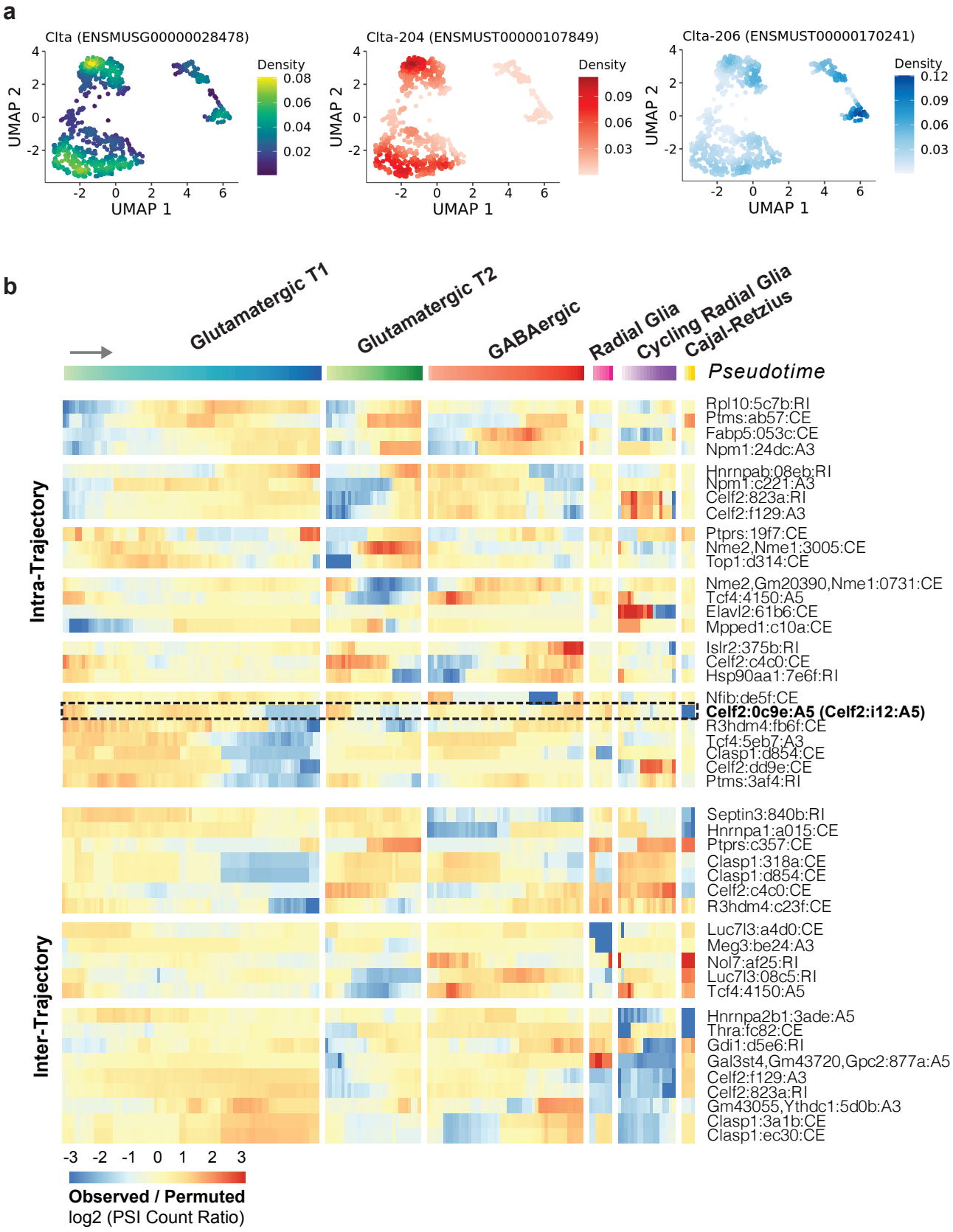
Supplementary Fig. 4– (a) Spearman correlation values of GM12878 cell line transcript quantifications obtained from sequencing with different platforms (Illumina, PacBio and ONT). (b) Total CPU time and peak RAM usage calculated according to analysis of a 5M read PromethION sample of sequencing IGROV-1 (SRR26865803). (c) Mean relative difference between two runs of the same IGROV-1 sample sequenced with a PromethION (P), MinION (M), or between the two (MP). (d) Correlations and relative difference of pseudo-bulk vs. bulk for top 5000 expressed transcripts in IGROV-1 cells as a function of the number of top ranked cells (by UMI count) included in the pseudo-bulk (left, middle right) and as a function of the top number of transcripts included for the top 64 cells (middle left, right). Source data are provided as Source Data files.



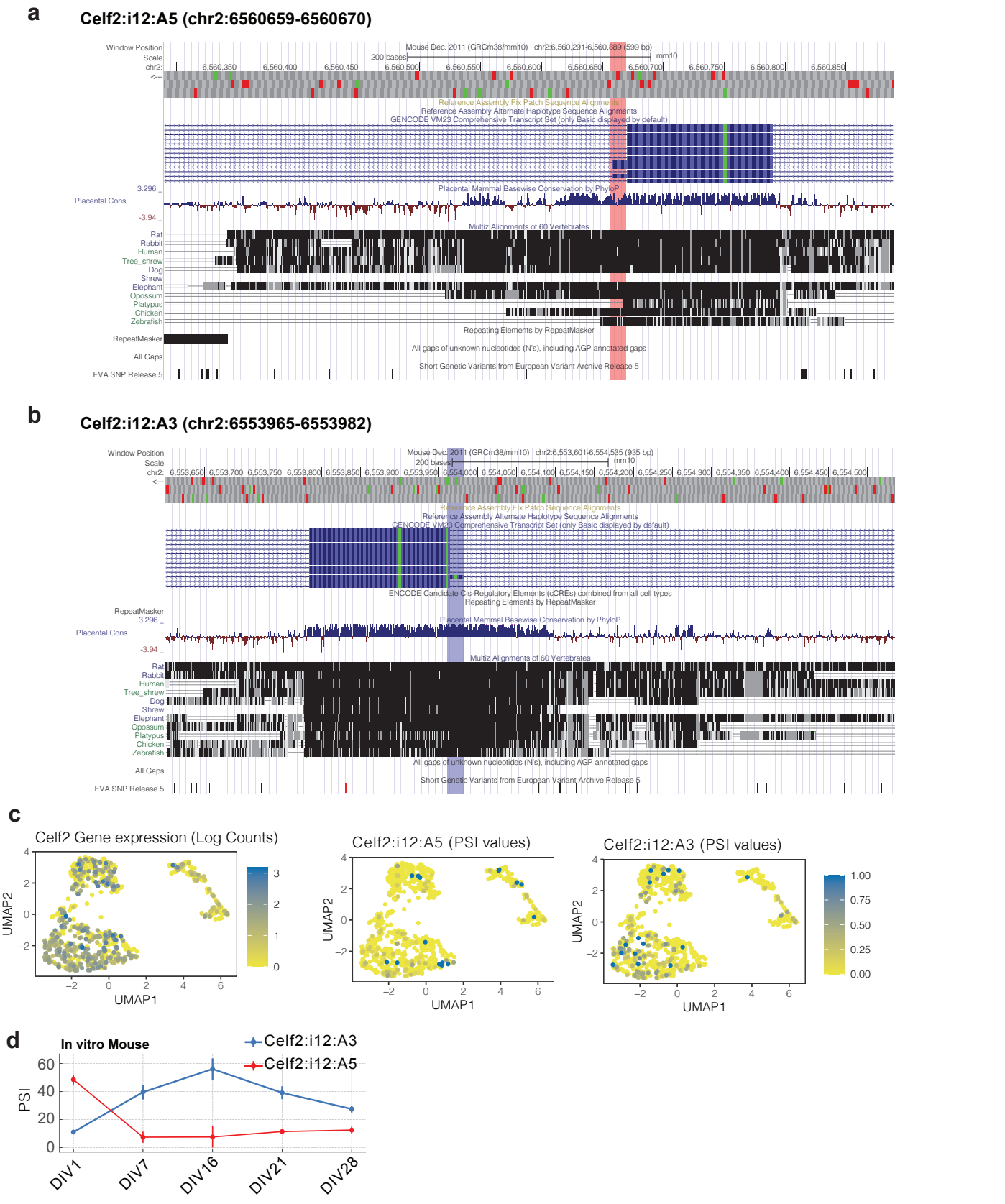
Supplementary Fig. 5– (a) Continued from Fig. 3b, Spearman correlations for each cell line in pseudo-bulk (by genetic identity) vs. the seven bulk nanopore sequenced ovarian cell lines for the top 4,000 HVTs. (b) Overlaid scatter plots of all matched (left) and decoy (right) comparisons, where each point is a transcript from one of the comparisons. (c) Continued from Fig. 3c, absolute difference between simulated matched and decoy cell lines across a range of 500-10,000 highly variable transcripts comparing mean relative difference and Spearman correlation metrics (shaded ribbons provide the upper and lower bounds of std. error, dashed gray line represents the ground-truth; see Methods) (d) Mean relative difference and Spearman correlation of Isosceles results preprocessed with Sicelore and wf-single-cell across matched and decoy comparisons for the top 4,000 highly variable transcripts (error bars show std. error). Source data are provided as Source Data files.



Supplementary Fig. 6– (a) Heatmap of marker gene signature expression (log-scaled mean gene TPM) along clusters colored according to Fig. 4a. (b) Density plots of each of the corresponding gene signatures (mean log-transformed gene counts) overlaid on the UMAP. Source data are provided as Source Data files.



Supplementary Fig. 7– (a) Exemplar result of 'isoform switching' analysis, the gene Clta is consistent with the findings highlighted in the original study. The two isoforms 204 and 206 show a distinct expression difference between the major clusters of radial glia and glutamatergic neurogenesis trajectory. (b) Expanded version of Fig. 4b plot (with the same scales), but labeled by gene : short hash id : event type. The short hash id matches the 'psi_event_label' column label Supplementary Data 1, which contains the genomic coordinates and Ensembl gene IDs. The event type abbreviations are explained in the Methods, eg. CE = Core Exonic interval, A5 = Alternative 5' splice site, A3 = Alternative 3' splice site, RI = Retained Intron. Source data are provided as Source Data files.



Supplementary Fig. 8– (a) UCSC Genome Browser snapshot of Celf2 intron 12 for the A5 event. Red bar marks the alternative region included by the A5 event, and matches the region marked in red from Fig 4d. (b) Same as panel a but for the A3 event, which matches the blue region in Fig 4d. (c) Extended set of plots matching Fig 4c, but raw gene expression counts, and raw PSI values. (d) PSI values for Celf2:i12:A5 and Celf2:i12:A3 across a mouse in vitro longitudinal glutamatergic neuron differentiation time series (all sample groups have n=2 samples; source accession identifiers listed in Supplementary Data 2) [30,31]. Source data are provided as Source Data files.