nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0303-9

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*]

Single-cell RNA-seq makes it possible to characterize the transcriptomes of cell types across different conditions and to identify their transcriptional signatures via differential analysis. Our method detects changes in transcript dynamics and in overall gene abundance in large numbers of cells to determine differential expression. When applied to transcript compatibility counts obtained via pseudoalignment, our approach provides a quantification-free analysis of 3′ single-cell RNA-seq that can identify previously undetectable marker genes.

Single-cell RNA-seq (scRNA-seq) technology provides transcriptomic measurements at single-cell resolution, making possible the identification and characterization of cell types in heterogeneous tissue. The problem of identifying transcripts or genes that are differentially expressed between cell groups in scRNA-seq is analogous to the problem of differential expression in bulk RNA-seq. Bulk RNA-seq differential expression methods can be applied directly to test for differences in transcript or gene expression between groups of cells[1], and methods that account for technical artifacts in scRNA-seq experiments, such as dropout modeling, seem to offer some advantages[2,3]. However, one aspect of scRNA-seq that current methods do not take advantage of is the large number of cells sampled in single-cell experiments. Furthermore, current scRNA-seq methods are mostly based on quantification of gene counts, thus precluding analysis of individual isoforms. In contrast, bulk RNA-seq is often used to study the dynamics of isoform expression, which have been shown to be important in both cell development[4] and pathology[5]. ScRNA-seq isoform analysis is more complicated than in bulk RNA-seq[6], but just as important[7]. Methods have been developed to test for differential transcript usage to allow investigation of these transcript dynamics from bulk RNA-seq, but such methods rely on the sampling of reads across isoforms. One of the challenges with scRNA-seq is that many methods produce data only from the 3′ ends of transcripts.

We show how prediction methods that take advantage of large numbers of cells and fully exploit all the transcript information that can be extracted from reads can greatly improve results both for differential gene expression and for differential transcript usage. We make use of logistic regression, which was considered when microarray gene expression assays were developed[8,9] but was not pursued owing to limited sample sizes. Now scRNA-seq provides the large number of samples required to accurately fit a logistic regression model. Instead of the traditional approach of using the cell labels as covariates for gene expression, we carry out logistic regression for each gene to predict cell labels from the quantification of constituent transcripts, when transcript quantifications can be accurately obtained. This is possible with certain technologies, such as SMART-Seq. Fitting the logistic regression model provides a linear combination of transcript quantifications that distinguishes cell groups, providing information about effect sizes of constituent transcripts, namely, the 'direction of change' (Fig. 1, Supplementary Fig. 1). Unlike traditional methods that test either for changes in overall gene abundance or for changes in transcript allocation, our method has the power to detect a change in any linear combination of transcript quantifications and provides a unified testing framework that eliminates the need for a dichotomy between differential gene expression and differential transcript usage methods (Supplementary Fig. 2).

In a simulation based on experimental effect sizes (Methods), logistic regression outperformed other existing scRNA-seq differential expression methods, even with different normalizations (Supplementary Figs. 3 and 4b). In cases where isoforms moved in concert, naive gene quantification by summing of isoform counts performed similarly to logistic regression (Fig. 1a–c, Supplementary Figs. 1a and 3e,f), but logistic regression also detected isoform switching (Fig. 1d–f, Supplementary Figs. 1b, 3c,d, and 5a). When applied to a dataset of differentiating myoblasts from Trapnell et al.[10], the method revealed diverse transcript dynamics across multiple genes known to be important for myogenesis (Fig. 1h). In addition to the dataset from ref. [10], we applied our method to a SMART-Seq2 dataset of embryonic cells to find genes with differential expression between days 3 and 4 post-fertilization, and compared the results with those of other methods. We showcase several genes undergoing isoform switching that were found only by our method (Supplementary Fig. 6). These results suggest that methods that test only for changes in overall gene expression are likely to miss a substantial proportion of differentially expressed genes (Supplementary Fig. 2b).

Transcript quantifications are biologically meaningful, but in some cases obtaining them might be infeasible, such as in cases where only the 3′ ends of transcripts are sequenced[11,12]. The reason is that transcripts of the same gene often share 3′ untranslated regions (UTRs) and therefore cannot be differentiated solely on the basis of 3′ end sequences. We therefore examined the possibility of conducting logistic regression directly on the transcript compatibility counts (TCCs) obtained via pseudoalignment. Pseudoalignment is a procedure that, for each read, finds a set of transcripts from which the read could have originated[13]. The sets of compatible transcripts are called equivalence classes, and the TCCs are the number of reads that map to each equivalence class[14]. TCCs were used by Ntranos et al.[15] as a more accurate, technology-independent

[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

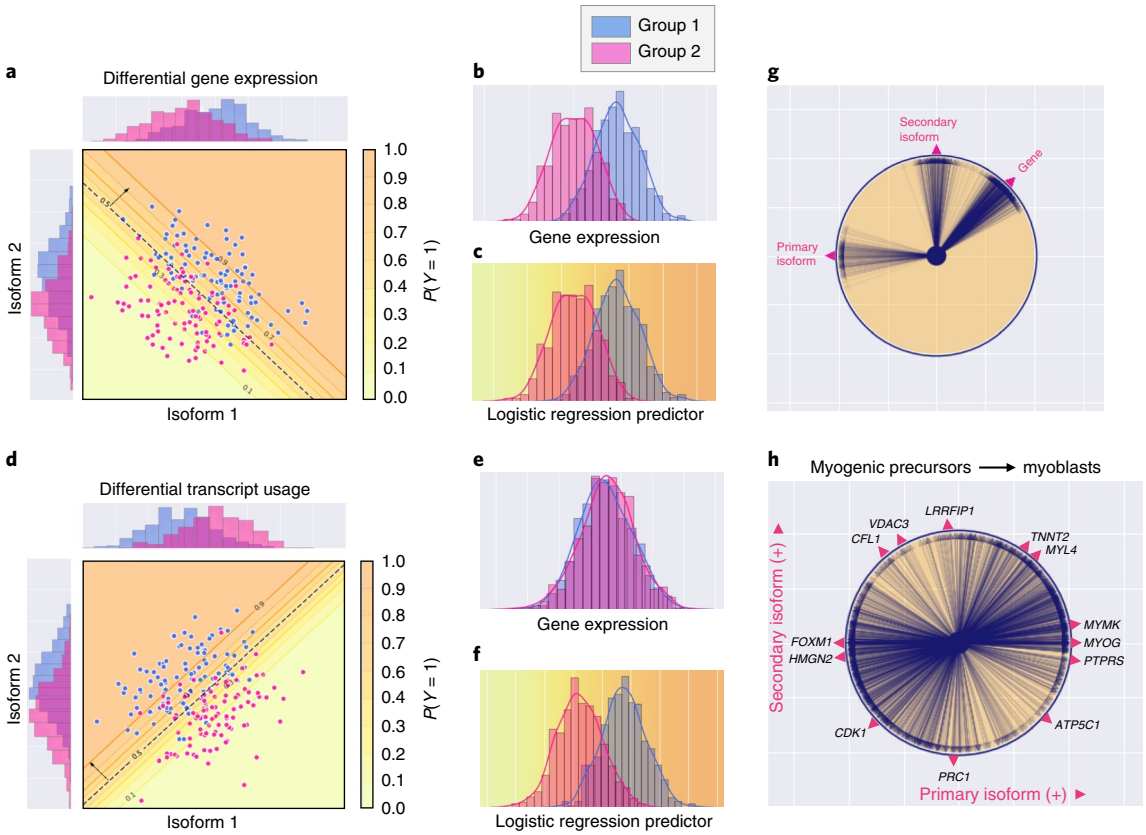

Fig. 1 | Logistic regression applied to scRNA-seq. Logistic regression can be used to detect differential gene expression at isoform-level resolution. **a**–**c**, Hypothetical scenario with two cell groups, group 1 and group 2, where both isoforms of a gene change with the same effect size. **a**, Each cell is plotted according to its two isoforms' abundances. The vector perpendicular to the classifier inferred by logistic regression (dashed line) indicates the direction of change. The orange shading corresponds to the probability of being in group 1 under the logistic regression model; cells farther above the dashed line are more likely to be in group 1 (darker orange), and cells farther below the dashed line are more likely to be in group 2 (lighter orange). **b**, Histogram of gene abundance. **c**, Histogram of the linear combination of transcript abundances learned by logistic regression along with the same probability gradient as in **a**. In this scenario, the linear combination found by logistic regression is the same as the summed gene abundances. **d**–**f**, Two isoforms have effect sizes in opposite directions (i.e., isoform switching). **e**, Histogram of gene abundance. **f**, Histogram of the linear combination of transcripts from logistic regression. **g**, Circle plot visualizing transcript dynamics for 1,000 genes, in which the directions of arrowheads correspond to the direction of the change for each gene. The *x*-axis corresponds to the primary isoform, and the *y*-axis corresponds to the secondary isoform. **h**, Directions of change for 1,308 genes in the dataset from Trapnell et al.[10] that were identified by logistic regression as differentially expressed between myogenic precursors and differentiating myoblasts. Pink arrowheads corresponding to known myogenic genes are marked along the edge of the circle.

transcriptomic signature for single-cell clustering because, unlike transcript quantifications, TCCs do not depend on a specific coverage model. In the case of 3′ sequencing where transcript quantification is infeasible, TCCs can be readily obtained via pseudoalignment, thereby maintaining the isoform-level information that is available in the data[15,16].

On simulated data, the performance of logistic regression with TCCs was comparable to that of other methods (Supplementary Fig. 7a,b). Furthermore, it had more power to detect isoform switching (Supplementary Fig. 7c,d). To investigate whether logistic regression with TCCs confers an advantage over gene-count-based differential analysis of such data, we examined 10x Chromium scRNA-seq data from three human T cell populations that were purified with antibodies specific to different isoforms of CD45 (PTPRC)[11]. Using TCCs, we carried out pairwise differential analyses of purified CD45RO$^+$ memory helper T cells, CD45RA$^+$ naive helper T cells, and CD45RA$^+$ naive cytotoxic T cells, providing two positive controls (CD45RA$^+$ versus CD45RO$^+$) and a negative control (CD45RA$^+$ versus CD45RA$^+$) for the method. Logistic regression was able to detect differential expression of *CD45* (*PTPRC*) in the purified CD45RO$^+$ memory and CD45RA$^+$ naive T cell populations (Fig. 2a,b). This result was deemed impossible by Peterson et al.[17], who noted that 3′ mRNA sequencing alone could not resolve these markers. We confirmed that gene counts alone could not identify *CD45* as differential, and furthermore, we found that independent testing of TCCs reduced the statistical power. In contrast, in tests of CD45RA$^+$ naive helper T cells and CD45RA$^+$ naive cytotoxic T cells (Fig. 2c), *CD45* was not found to be significant with any method for this subsample, and there was little difference in overall *P* value distributions between independent testing of equivalence classes and multiple logistic regresion. A power analysis showed that logistic regression with TCCs found *CD45* to be differentially expressed after multiple testing correction (FDR < 0.01), whereas logistic regression on gene counts did not identify *CD45 at any cell number*. A distribution of the *P* values obtained by each method suggests that although both methods found the genes with the largest change in overall gene expression, only logistic regression with TCCs detected isoform switching (Supplementary Fig. 8).

Further examination of the transcripts corresponding to the equivalence classes identified by our method pointed to transcripts that were differentially expressed (Supplementary Fig. 9). Visual

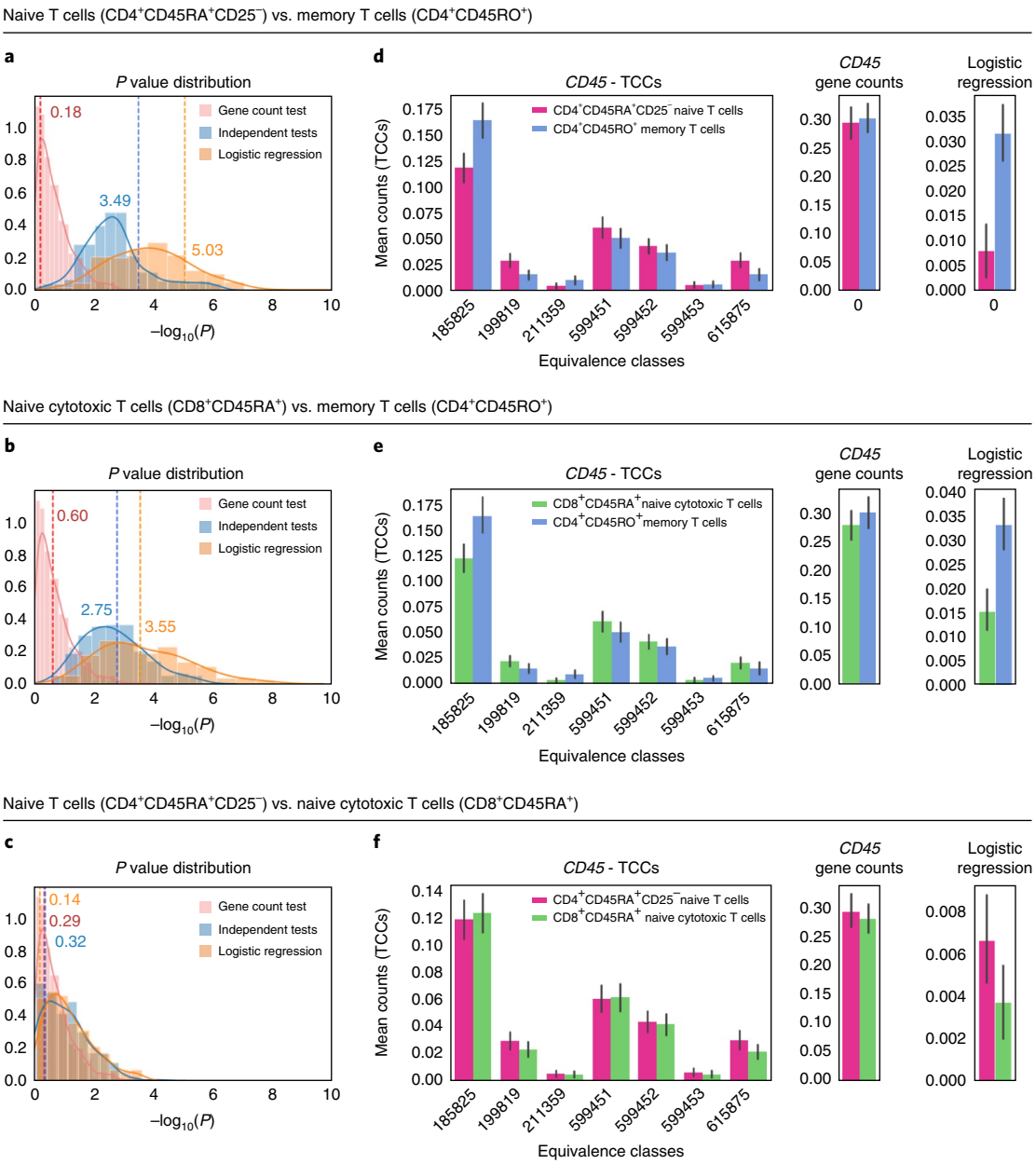

Fig. 2 | Logistic regression identifies *CD45* in purified T cell types. We carried out pairwise differential expression analysis of purified memory helper T cells, naive helper T cells, and naive cytotoxic T cells that were sequenced by 10x. **a**–**c**, *P* value distributions corresponding to three different differential expression methods, generated from 200 subsamples. Each subsample contained 3,000 cells of each cell type from the full dataset of 9,923 naive helper T cells, 9,994 memory helper T cells, and 11,914 naive cytotoxic T cells. The three methods used were our method of multiple logistic regression on TCCs ('logistic regression'), logistic regression on gene counts ('gene count test'), and logistic regression on each equivalence class followed by Bonferonni correction ('independent tests'). **d**–**f**, Bar plots corresponding to the expression profiles of a specific subsample; error bars correspond to the 95% confidence interval of the mean expression across 3,000 cells. The *P* value from each method for this particular subsample is marked as a dashed line on the corresponding *P* value distribution in **a**–**c**.

inspection of the differential equivalence classes identified for *CD45* revealed that the corresponding isoforms were being distinguished by virtue of alternative, unannotated 3′ UTRs (Supplementary Fig. 9a,b). To quantify the extent of isoform accessibility by 3′-end sequencing, we estimated the distribution of read pseudoalignments with respect to the annotated 3′ UTRs (Supplementary Fig. 9d). We found a substantial number of reads farther from 3′ ends than expected, which points to a large number of unannotated 3′ UTRs. To mitigate the effect of unannotated 3′ UTRs on our analysis, we updated the transcriptome with novel 3′ UTRs for *CD45* and redid the analysis. *CD45* expression remained differential (Supplementary Methods, Supplementary Figs. 10 and 11). Overall, our results are concordant with previous work on lymphocytic surface receptor isoform diversity[18]. Equivalence classes provide access to isoforms of genes other than *CD45*, and we found multiple other genes with evidence of isoform switching between memory and naive T cells (Supplementary Fig. 12).

To examine whether TCCs can be informative in a de novo scRNA-seq experiment, we analyzed a dataset consisting of 68,579 peripheral blood mononuclear cells sequenced at an average of

20,491 reads per cell[11]. After clustering and using known cell-type markers to annotate the clusters (Supplementary Fig. 13), we were able to recapitulate our previous *CD45* analysis: *CD45* was identified as differentially expressed between memory and naive T cells (Supplementary Fig. 14), which shows that TCC-based logistic regression can be applied to cell groups generated by unsupervised clustering as in standard, de novo scRNA-seq workflows.

Logistic regression is especially powerful for scRNA-seq because it leverages the large number of cells available in scRNA-seq experiments and incorporates isoform information for gene-level testing. It reveals the contribution of individual isoforms to the gene-level differential analysis, thus enhancing the interpretability of results. We have demonstrated the power of logistic regression for analysis of gene-level differential expression between two cell types, but the method extends to more general groupings. Furthermore, logistic regression can be carried out on all genes simultaneously to reveal gene markers that characterize cell types. Finally, our method scales effectively with both the number of reads and the number of cells, which is critical for processing increasingly large scRNA-seq datasets (Supplementary Fig. 5b,c).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0303-9.

Received: 6 July 2018; Accepted: 13 December 2018;
Published online: 21 January 2019

References

1. Soneson, C. & Robinson, M. D. *Nat. Methods* **15**, 255–261 (2018).
2. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. *Nat. Methods* **11**, 740–742 (2014).
3. Finak, G. et al. *Genome Biol.* **16**, 278 (2015).
4. Yamazaki, T. et al. *Genes Dev.* **32**, 1161–1174 (2018).
5. Vitting-Seerup, K. & Sandelin, A. *Mol. Cancer Res.* **15**, 1206–1220 (2017).
6. Arzalluz-Luque, Á. & Conesa, A. *Genome Biol.* **19**, 110 (2018).
7. Gupta, I. et al. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2018/07/08/364950 (2018).
8. Xing, E. P., Jordan, M. I. & Karp, R. M. in *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning* (eds Brodley, C. E. & Pohoreckyj Danyluk, A.) 601–608 (Morgan Kaufmann, San Francisco, 2001).
9. Shevade, S. K. & Keerthi, S. S. *Bioinformatics* **19**, 2246–2253 (2003).
10. Trapnell, C. et al. *Nat. Biotechnol.* **32**, 381–386 (2014).
11. Zheng, G. X. et al. *Nat. Commun.* **8**, 14049 (2017).
12. Macosko, E. Z. et al. *Cell* **161**, 1202–1214 (2015).
13. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
14. Nicolae, M., Mangul, S., Măndoiu, I. I. & Zelikovsky, A. *Algorithms Mol. Biol.* **6**, 9 (2011).
15. Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L. & Tse, D. N. *Genome Biol.* **17**, 112 (2016).
16. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. *Genome Biol.* **19**, 53 (2018).
17. Peterson, V. M. et al. *Nat. Biotechnol.* **35**, 936–939 (2017).
18. Byrne, A. et al. *Nat. Commun.* **8**, 16027 (2017).

Acknowledgements

We thank N. Bray, J. Gehring and V. Svensson for discussion and comments on the manuscript, and H. Pimentel for assisting with the simulations. We thank A. Butler and R. Satija for implementing this method in Seurat. V.N., L.Y. and L.P. are partially funded by NIH R012017-0569.

Author contributions

V.N. developed the model during discussions with L.Y. and L.P, and analyzed the 10x PBMC dataset. L.Y. performed the simulations and analyzed the embryo SMART-Seq dataset. P.M. developed kallisto genomebam and assisted with analysis. All authors contributed extensively to the interpretation of the results and writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0303-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Model. We model cell membership as a function of gene expression. Let $X_{t,i}$ be the transcript abundance for transcript t in cell i. Let y_i be the indicator variable for the membership of cell i. Then for a gene g, the transcript abundances are linear predictors of cell membership.

$$y_i \mid X_{t,i} \sim \text{Bernoulli}\left(\sigma\left(c + \sum_{t \in g} B_t X_{t,i}\right)\right)$$

where $\sigma(t)$ is the logistic function, defined as $\sigma(t) = 1/(1+e^{-t})$, and c is a constant. This framework is multiple logistic regression, where for each gene the number of predictors is equal to the number of transcripts in the gene.

To obtain significance scores for each gene, we conducted a likelihood ratio test. The null model used for the likelihood ratio test was that cell membership does not depend on gene expression:

$$y_i \sim \text{Bernoulli}(\sigma(k))$$

where k is a constant. For each gene, the difference in the degrees of freedom between the alternative model and the null model is therefore equal to the number of transcripts contained in the gene.

If transcript quantifications are not available, TCCs may be used instead. Let $T_{e,i}$ be the TCC for equivalence class e in cell i. The TCC models are

$$y_i \mid T_{e,i} \sim \text{Bernoulli}\left(\sigma\left(c + \sum_{e \in g} B_e T_{e,i}\right)\right)$$

Our null models remain $y_i \sim \text{Bernoulli}(\sigma(k))$, and the differences in the degrees of the freedom are equal to the number of TCCs associated with each gene g.

Trapnell et al. analysis. We downloaded the preprocessed data from Trapnell et al.[10] from the conquer[1] database, which included the quantified transcripts-per-million (TPM) values and cell labels for 222 serum-induced primary myoblasts over time courses of 0, 24 and 48 h. We selected the 85 myogenic precursor cells and the 97 differentiating myoblast cells for differential expression analysis. We used Ensembl Homo_sapiens.GRCh38.rel84.cdna.all.fa to group 176,241 transcripts into 38,694 genes and tested each gene for differential expression between myogenic precursors and differentiating myoblasts with our method. Logistic regression was run using sklearn.linear_model.LogisticRegression(). After Benjamini–Hochberg correction, we obtained 1,308 significantly differentially expressed genes (FDR < 0.01). We visualized these genes in a circle plot by carrying out logistic regression on the primary and secondary isoforms, which are defined as the isoforms with the highest and second-highest average expression over all cells.

Zheng et al. analysis. We obtained the raw reads for the three human peripheral blood mononuclear cell (PBMC) purified cell-subtype datasets described by Zheng et al.[11] (CD4$^+$CD45RA$^+$CD25$^-$ naive T cells, CD4$^+$CD45RO$^+$ memory T cells, and CD8$^+$CD45RA$^+$ naive cytotoxic T cells) from the 10x Genomics Support website[19]. We preprocessed the reads (barcode detection, error correction, and pseudoalignment) with the scRNA-Seq-TCC-prep kallisto wrapper (SC3Pv1 chemistry; https://github.com/pachterlab/scRNA-Seq-TCC-prep) to obtain the single-cell TCC matrix. After filtering out cells with total unique molecular identifier (UMI) counts outside the interval [1,000–30,000], we obtained 31,831 cells (9,923 CD4$^+$CD45RA$^+$CD25$^-$ naive T cells, 9,994 CD4$^+$CD45RO$^+$ memory T cells, and 11,914 CD8$^+$CD45RA$^+$ naive cytotoxic T cells, respectively). We selected all the equivalence classes that contained at least one isoform associated with the *CD45* gene and filtered out the ones with total UMI counts less than 0.25% of the total number of cells, that is, equivalence classes with fewer than ~79 UMI counts across all cells. This resulted in seven equivalence classes uniquely associated with subsets of the annotated isoforms of *CD45*. We obtained the gene counts for each cell by summing the TCCs. We carried out all three pairwise tests for differential expression between the purified cell subtypes using a multiple logistic regression model on the seven TCCs, a logistic regression model on the aggregated gene counts, and a logistic regression model independently on each equivalence class. We ran logistic regression with sklearn.linear_model.LogisticRegression(), and we used the likelihood ratio test to obtain P values for all three tests, as described in the "Model" section. For each pairwise test, we randomly subsampled 3,000 cells per group across 200 independent subsamples to generate P value distributions for each method.

We preprocessed the raw reads for the 68,000-PBMC dataset with the scRNA-Seq-TCC-prep kallisto wrapper to obtain the TCC matrix. Equivalence classes that mapped to multiple Ensembl gene names and cells with total UMI counts outside the interval [2,000–20,000] were filtered out. The resulting 65,444 cell by 95,426 equivalence class matrix was subsequently used for postprocessing and clustering with scanpy 0.2.6[20]. We used the same steps outlined in the "Zheng et al. recipe" provided by scanpy, except we selected the 5,000 most variable equivalence classes in lieu of the 1,000 most variable genes. To verify the clustering structure, we plotted the cells with t-distributed stochastic neighbor embedding using specific marker gene abundances obtained by summing all the constituent TCCs. Supplementary Fig. 14 focuses on the clusters that are most likely to correspond to populations of naive cytotoxic T cells (cluster A; CD8A$^+$CD4$^-$CCR7$^+$, 5,226 cells), naive helper T cells (cluster B; CD4$^+$CCR7$^+$, 12,424 cells), and memory helper T cells (cluster C; CD4$^+$S100A4$^+$CCR10$^+$, 4,173 cells). Clusters A and C corresponded to clusters 3 and 6 in Supplementary Fig. 13, whereas we obtained cluster B by manually merging clusters 1 and 2. We carried out pairwise differential expression tests on these three clusters using multiple logistic regression on TCCs and the likelihood ratio test (see "Model" section for construction of the likelihood ratio test). For comparison to our method, we also carried out logistic regression on gene counts, and independent logistic regressions on each TCC followed by Bonferroni correction. We obtained P value distributions by applying these three differential expression tests across 200 subsamples, each time subsampling 2,000 cells per cluster.

To estimate the distribution of read distances to the 3′ end, we pseudoaligned the reads from the three purified T cell populations to the transcriptome, using the pseudobam option of kallisto 0.44.0. In the case of read multiple alignment, the weight of the read was split evenly across all reported transcripts. The distance to the 3′ end was inferred from the transcriptome coordinates reported in the BAM file.

To detect novel 3′ UTR ends for *CD45*, we identified reads whose alignment extended past the UTR into the poly(A) tail and that kallisto pseudoaligned to *CD45*. These reads were clustered according to their pseudoaligned genomic coordinates. After discarding clusters corresponding to known 3′ UTR ends, we had three clusters remaining, corresponding to unannotated 3′ UTR ends and containing 69, 71, and 97 reads, respectively. For each of these clusters, we removed the poly(A) tail and generated a consensus sequence via multiple alignment with FSA[21]. We aligned the consensus sequence to the genome to determine the genomic coordinates of the novel 3′ UTR endpoint. We modified the reference transcriptome by creating a new version of each transcript belonging to *CD45* that overlapped the new 3′ UTR endpoint, which resulted in the addition of 13 novel transcripts. For visualization purposes, we also ran kallisto pseudobam with the updated GTF file (Supplementary Fig. 10).

Petropoulos et al. analysis. We downloaded the Petropoulos et al.[22] dataset, which contains quantifications for 1,529 human preimplantation embryonic cells, from the conquer[1] database. We used the provided Ensembl transcript and gene quantifications (counts and TPMs) to conduct differential analysis of day 3 and day 4 embryonic cells (271 total cells). The differential expression methods were run with the same normalization and filters as the simulations (see below). We used the method glm from R's native 'stats' library for logistic regression, by using the parameter "family = 'binomial'" with its default logit link function. We used UpSetR 1.3.3[23] to plot the size of the intersection sets of the 3,000 most significantly differentially expressed genes.

Read simulation framework. We developed an scRNA-seq simulation framework that can simulate reads (https://github.com/pachterlab/NYMP_2018/tree/master/simulations). Parameters for the simulator were estimated using data from Trapnell et al.[10] In each simulation, cells were simulated from two different cell groups: a null group and a perturbed group, each with 105 cells. The null type was modeled after the cluster of proliferating myoblasts from the Trapnell et al.[10] dataset. Specifically, after quantification of the dataset using kallisto and clustering on TCCs, the cluster containing 105 cells with *MYOG* expression was identified and used as the basis of our simulations.

The nonzero TPMs from the myoblast cluster were used to estimate the parameters of a log-normal distribution for each transcript. To simulate the null cell type, we drew TPMs for each transcript from a log-normal distribution. Then, for each transcript, a subset of cells was chosen at random in which the transcript abundance was set to 0 ('dropout'). The percentage of dropout for each transcript was matched to the experimental dataset. Mathematically, given d_t as the dropout rate for transcript t, and with μ_t and σ_t parameterizing the mean and variance of the expressed component, the transcript expression is modeled as

$$x_t \mid \mu_t, \sigma_t, d_t = \begin{cases} 0 & \text{with probability } d_t \\ \text{lognormal}(\mu_t, \sigma_t) & \text{with probability } (1-d_t) \end{cases}$$

The use of the log-normal distribution on TPMs was motivated by the Tobit model in Monocle[2], the mixture of dropout and expression components was motivated by SCDE, and the masking of random cells with zeros to reach sufficient dropout was modeled after Splatter. The mean variance plots and the distributions of zeros were greatly concordant between the simulations and the experimental data after which they were modeled (Supplementary Fig. 15).

We prepared three different types of simulated data to reflect distinct perturbation scenarios and effect sizes. Transcript expression in fewer than 5 of the 105 cells was deemed too low, and these transcripts were filtered out of the perturbation. In the independent effects simulation, 30% of the transcripts that passed the filter (20,456 of 68,179 expressed transcripts) were chosen at random to be perturbed. For each transcript, a minimum effect size of twofold was drawn

from a truncated log-normal distribution. The direction of each perturbation was chosen uniformly at random (50% upregulated, 50% downregulated). In the correlated effect simulations, genes with all transcripts passing the filter also passed the filter. We chose 30% of the remaining genes (~5,220 of 17,390 genes) at random to be perturbed, and expressed transcripts (defined as expressed in ≥5 cells) of that gene were perturbed with the same effect size drawn from a truncated log-normal distribution at a minimum of 2. In the experiment-based simulations, the effect sizes were learned from Trapnell et al.[10] from the set of transcripts that DESeq2[24] found to be differentially expressed (P value < 0.05). The same transcripts were perturbed with their DESeq2-derived effect sizes in the simulation.

We applied the effect sizes to the mean expression, and we generated abundances per cell by sampling from the log-normal distribution truncated at zero. Using these cell-by-cell abundances, RSEM[25] generated paired-end reads uniformly distributed across transcripts, applying a model learned from a proliferating myoblast cell from the Trapnell et al.[10] dataset and a background-noise read percentage (parameter theta) of 20%. The number of reads per cell was learned from the myoblast cluster by fitting of a log-normal distribution of reads per cell ($\mu = 14.42$, $\sigma = 0.336$), corresponding to a mean of 193,000 paired-end reads per cell.

Splatter simulation framework. We also used Splatter[26], which simulates transcript counts directly instead of reads. The same 105 myoblasts from Trapnell et al.[10] used to model the simulations above were used to fit Splatter simulation parameters. Transcripts with more than 90% zeros were filtered from the simulation, leaving 47,606 transcripts to be simulated. We used Splatter's default parameters to simulate two groups, with a 0.1 chance of perturbation in each group, resulting in 9,095 perturbed transcripts and corresponding to a 19% perturbation rate across the two groups. The 47,606 transcripts were randomly assigned to 15,420 genes according the transcriptomic structure, and transcript counts were summed to provide gene counts. These transcripts corresponded to 6,393 perturbed genes across 15,420 total genes.

Simulation analyses and benchmarking. Logistic regression, Monocle's Tobit model[10], DESeq2 1.16.11[24], MAST 1.2.1[3], and SCDE 1.99.4[2] were used to benchmark the simulations in R. Monocle's Tobit model method, DESEq2, and MAST were invoked using Seurat's wrapper functionality through Seurat::FindMarkers[27]. We used the method glm from R's native 'stats' library to perform logistic regression, by using the parameter "family = 'binomial'" with its default logit link function.

The FASTQ files output from the RSEM simulations were quantified with kallisto v0.44.0. tximport[28] was used to aggregate transcript-level counts and abundances into gene-level counts and abundances before inputting into the various methods. In contrast, the Splatter[26] simulation did not require read quantification, as transcript counts were directly simulated. To afford each method its optimal input, we used normalizations native to each method. For SCDE, and DESeq2, the gene counts were used as input. For Monocle and MAST, the TPM abundances were used as input. For our method, we used DESeq2's library-size method of normalization on transcript counts before carrying out logistic regression. To apply DESeq2's method, we calculated size factors based on the transcript counts using DESeq2::calculateSizeFactors, and obtained the normalized counts by dividing by the cell's size factor. For all methods, we filtered out genes/transcripts with zero expression in >90% of cells from the analysis with logistic regression.

To perform logistic regression using TCCs, we filtered out ECs that contained transcripts from multiple genes and ECs with >90% zeros. Additionally, genes with fewer than four cells per TCC were filtered from analysis. TCCs were normalized with DESeq2's size factor method.

We benchmarked the accuracy of the methods by evaluating their trade-off between sensitivity and FDR. FDR is defined as the number of false positives divided by the number of total declared positives. We ranked the genes by significance (i.e., lowest to highest P value) and then calculated and plotted the FDR and sensitivity at each level of significance.

In addition to benchmarking the methods by accuracy, we evaluated the runtimes of the methods on the Splatter simulation. Every method was run in series three times on the same dataset on a machine with 40 cores and 350 GB. The runtimes were benchmarked with R's system.time(). All methods were run using a single core, except SCDE, which was run with its default 20 cores. The real elapsed time and the total processing time, calculated as the sum of the user time and the system time, are plotted in Supplementary Fig. 4.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

The code required to conduct the simulations and reproduce the analyses is available at https://github.com/pachterlab/NYMP_2018. We also have provided the Github repository that was zipped at the time of manuscript acceptance as Supplementary Software.

Data availability

The myogenesis dataset (Trapnell et al.[10]) is available on the conquer database and on GEO as series GSE52529. The dataset on embryogenesis is available on the conquer database (Petropoulos et al.[22]). The 10x PBMC dataset is available from the 10x Genomics Support website[19].

References

19. 10x Genomics. Single cell gene expression datasets. *10x Genomics Support* https://support.10xgenomics.com/single-cell-gene-expression/datasets (2018).
20. Wolf, F. A., Angerer, P. & Theis, F. J. *Genome Biol.* **19**, 15 (2018).
21. Bradley, R. K. et al. *PLoS Comput. Biol.* **5**, e1000392 (2009).
22. Petropoulos, S. et al. *Cell* **165**, 1012–1026 (2016).
23. Conway, J. R., Lex, A. & Gehlenborg, N. *Bioinformatics* **33**, 2938–2940 (2017).
24. Love, M. I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
25. Li, B. & Dewey, C. N. *BMC Bioinformatics* **12**, 323 (2011).
26. Zappia, L., Phipson, B. & Oshlack, A. *Genome Biol.* **18**, 174 (2017).
27. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. *Nat. Biotechnol.* **33**, 495–502 (2015).
28. Soneson, C., Love, M. I. & Robinson, M. D. *F1000Res.* **4**, 1521 (2015).

nature research

Corresponding author(s): Lior Pachter

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used publically available RNA-seq datasets from GEO and the conquer database.
Data analysis	We provide the code used to generate the analysis and figures in our Github repository here: https://github.com/pachterlab/NYMP_2018. RSEM1.3.0 was used to simulate reads and Splatter 1.2.2 to simulate counts; DESeq2 1.16.11, MAST 1.2.1, SCDE 1.99.4, Seurat 2.0.1 for benchmarking; scanpy 0.2.6 for 10x PBMC preprocessing and clustering; R3.4.1's glm() function and sklearn 0.19.1's linear_model.LogisticRegression() function within python 3.6.2 for performing logistic regression; UpSetR 1.3.3 for generating visualizing intersecting sets.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The myogenesis dataset (Trapnell et al., 2014) is available on conquer database and on GEO as Series GSE52529. The dataset on embryogenesis is available on the conquer database (Petropoulos et al., 2016). The 10x dataset on PBMCs is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	On the publically available datasets, we chose to include all cells in our analyses, except for the analysis of the 10X PBMC datasets, where we chose to perform 200 subsamples at 3000 cells per cell type per subsample. We decided to generate subsamples so we can examine p-value distributions for several methods instead of obtaining one p-value per method for the entire dataset. We believed 3000 cells to be a reasonable approximation for the number of cells per cluster in a large experiment and 200 subsamples to generate accurate distributions.
Data exclusions	No data were excluded.
Replication	We did not need to reproduce any wet-lab experiments since our work was entirely computational. We did perform 200 subsamples as part of computational replication.
Randomization	We did not need randomization for experiments since we did not perform wet-lab experiments for this manuscript. However, in our subsampling, we were able to pick cells randomly from each cluster computationally.
Blinding	We did not need blinding since we had no randomized control trials.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging