nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12924-w OPEN

Biological process activity transformation of single cell gene expression for cross-species alignment

Hongxu Ding[1,3]*, Andrew Blair[1,3], Ying Yang[2] & Joshua M. Stuart[1]*

The maintenance and transition of cellular states are controlled by biological processes. Here we present a gene set-based transformation of single cell RNA-Seq data into biological process activities that provides a robust description of cellular states. Moreover, as these activities represent species-independent descriptors, they facilitate the alignment of single cell states across different organisms.

[1] Santa Cruz Genomics Institute and Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA. [2] Department of Genetics and Development, Columbia University Medical Center, New York, NY 10032, USA. [3] These authors contributed equally: Hongxu Ding, Andrew Blair. *email: hding16@ucsc.edu; jstuart@ucsc.edu

The advent of single-cell RNA-sequencing (scRNA-Seq) technologies has greatly advanced our understanding of cellular states[1]. However, the signal-to-noise ratio of scRNA-Seq data is usually poor, confounding cellular state interpretation. Considering that cellular states are controlled by genetic regulatory mechanisms[2], we propose using biological process activities (BPAs) in place of the expression of individual genes in the scRNA-Seq data, which leverages an ensemble of dozens of related genes. In this way, discrepancies in individual genes can be averaged out, yielding reproducible measurements unaffected by common technical noises such as batch effects[3] and drop-out events[4] (Fig. 1b–d).

Gene sets have been used extensively over the past few years to infer the activity of biological processes in many applications. The various catalogs of gene sets, e.g., Molecular Signature Database (MSigDB), group genes into categories of related function. Such gene sets allow particular pathways to be associated with the results of high-throughput assays. Gene set enrichment analysis (GSEA) is a particularly successful approach that summarizes the putative importance of a biological process using an ensemble of expression levels for a set of genes documented to play a role in a specific process[5].

We propose transforming gene expression levels into interpreted features using BPAs. The BPA transform is based on the published aREA method[6,7] that extends GSEA to interpret scRNA-Seq data as inferred process activities. Compared to previous studies[8,9], which used selected gene sets describing specific biological processes, e.g., cell cycle or TP53 pathway, for this study we use a comprehensive collection of gene sets, e.g., Gene Ontology (GO) Biological Process (BP), and the immunologic portion of the MSigDB collection called C7[10], contributing ~650 and ~1800 gene sets after filtering, respectively (see Methods). BPA's only option is to decide what gene sets to use for the transformation (e.g., gene sets can be selected based on their standard deviation, see Methods). For an individual cell and a specific pathway, the BPA transform runs an aREA-based single-cell pathway enrichment analysis to create an activity score from the expression levels of the gene set members of the pathway (Fig. 1a). In this way, the gene expression signature of an individual cell is converted to a BPA profile in which every pathway has a distinct activity score. Moreover, assuming the GO-BP terminology is consistent across species, GSEA can be used in each species separately to infer an activity for the same set of processes. Thus, inferences of activity for each category from single-cell RNA-Seq data can be compared across species even though the categories between species have different gene members. In this way, data sets of human and model organisms can be combined directly to reveal functionally analogous cell types across species without the need to predict orthologs. We demonstrate the utility of using BPAs to align human and mouse data sets to shed light on their comparative and species-specific biology in early embryo development and in the cell types comprising the immune system.

Results

BPA mitigates scRNA-Seq batch effects and drop-out. Distinct, batch-specific clusters can be observed among peripheral blood

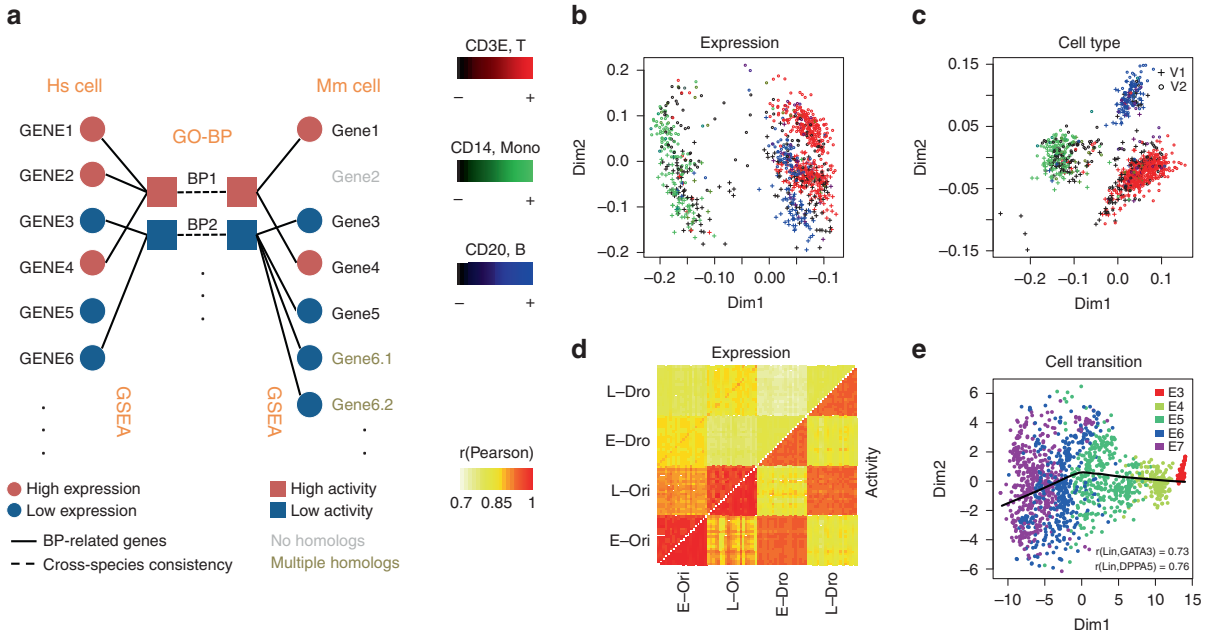

Fig. 1 Overview of BPA and performance on single-cell data sets. **a** Overview of biological process activity inference. Single=cell gene expression profiles for human (outer left column) can be compared with a mouse gene expression profile (outer right column) using transformed biological process activity profiles for human (inner left) and mouse (inner right) even though the gene members of each Gene Ontology Biological Process (GO-BP) are distinct in each species (outer links). **b** Single peripheral blood mononuclear cells (PBMCs) profiled using 10x Genomics V1 and V2 chemistry were visualized using transcript expression features. Cells were color-coded according to expression of B-cell, monocyte and T-cell-specific markers CD3E, CD14, and CD20, respectively. Two clusters for each cell type, corresponding to each of the Chromium chemistries are visible before the BPA transform. **c** Same as part B but PBMC data plotted after applying the BPA transform resulting in one cluster per cell type and no visible chemistry batch effect. **d** Drop-out events (Dro) were simulated into GTEx lung (L), and esophagus (E) bulk RNA-sequencing (Ori) data (Supplementary Fig. 2, see Methods); pairwise correlations between samples was computed and plotted, resulting from comparisons using the original gene expression features (upper matrix triangle) and using BPA features (lower matrix triangle); high correlations, red; low correlations, yellow. BPA preserves same tissue comparisons even for drop-outs and has lower on average cross-tissue correlation than using gene expression features. **e** BPA transform preserves the developmental order of embryo development of the original study[13] based on the pseudotime inferred from principal curve[14] construction in t-SNE space[15] (Supplementary Fig. 4)

mononuclear cell (PBMC) scRNA-Seq data sets when gene expression profiles are used directly (Fig. 1b), but are no longer apparent after BPA transformation (Fig. 1c), giving comparable, or even better performance compared to batch-correction algorithms[3,11,12] (Supplementary Fig. 1). In this example, the clustering of the PBMCs recapitulates the groups of B-cells, T-cells, and monocytes as denoted by the cell type-specific markers *CD3E*, *CD14*, and *CD20*, respectively. In addition, BPA is insensitive to drop-out events, illustrated through a controlled simulation in which drop-outs are introduced to mimic their distribution in real scRNA-Seq data. Complete RNA-sequencing data sets of bulk tissue samples were taken from the GTEx lung (L) and esophagus (E) collection and labeled as original (Ori), whereas their counterparts containing simulated drop-out events were labeled drop-out (Dro) (Supplementary Fig. 2, see Methods). We found that drop-out events appreciably decrease correlations within the same biological state. Moreover, correlations between different tissues with full data, e.g. r(L-Ori, E-Ori) was found to be higher than correlations between the same tissue type having drop-out data e.g. r(L-Dro, L-Dro) or r(E-Dro, E-Dro). Thus, artifacts in downstream analyses could be introduced when using transcript-level data containing drop-out events, as single cells will cluster according to the extent of the drop-out effect rather than the measured biological conditions. The inferred biological activity preserves within-tissue correlations, and reduces cross-tissue correlations (Fig. 1c). Taken together, inferred BPA profiles produced clusters with distinctly enriched PBMC cell types according to marker gene expression (Fig. 1d and Supplementary Fig. 3), as well as the known ordering of state transitions in a human preimplantation embryo data set (Fig. 1e, Fig. 2a and Supplementary Fig. 4; see Methods).

BPA facilitates cross-species comparisons. We next performed cross-species single-cell state alignment using BPA profiles in place of gene expression profiles. A previous effort used the expression pattern of one-to-one orthologous genes to align single cells across species[16]. However, as orthologs are usually determined by computational analysis of protein sequence alone[17], the expression pattern of orthologous genes may not be the same across species[18]. The BPA transform, on the other hand, provides a common set of terms from which detailed gene sets can be retrieved in a species-specific manner[10]. Each species can be analyzed separately using their species-specific gene sets and then merged across species at the BPA level assuming the pathway categories are equivalent, giving an overarching perspective of cellular states and transitions across various organisms.

We analyzed scRNA-Seq profiles reported in a human–mouse comparative study on embryo development. Although early embryo development is a continuous process, it can be roughly divided into three steps[2]. By analyzing human and mouse single cells separately, we find that the three steps were recapitulated. In human, the first step spans the oocyte to four-cell stage; the second step includes the eight-cell stage; and the third step includes only the morula stage. In mouse, the first step of the developmental timeline is relatively short compared with human, including the oocyte and pronuclear stages; the second step includes the two-cell and four-cell stages; and the last step includes the eight-cell and morula stages (Fig. 2a). The data sets transformed with GO-BP produce the expected alignment of the three steps between the two species (Fig. 2a), which was further confirmed objectively using dynamic time warping (Supplementary Fig. 5)[16,19]. In addition, the stage-specific activation patterns of biological processes determined

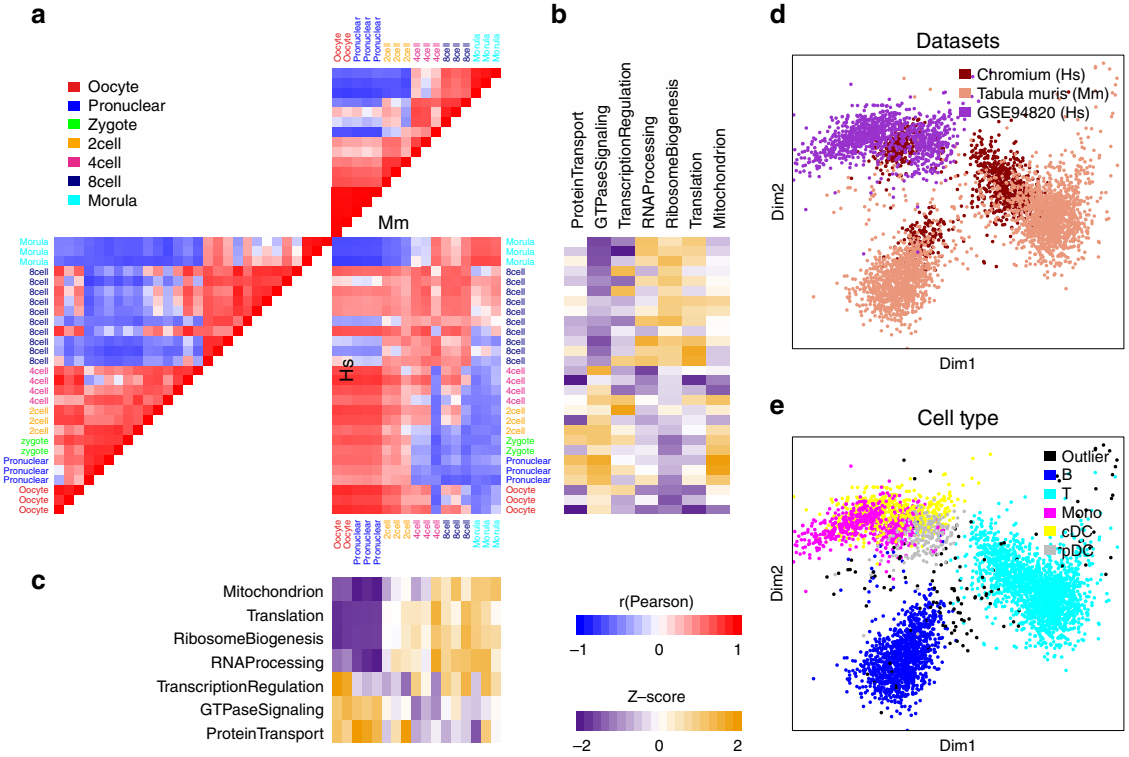

Fig. 2 BPA-based human–mouse integrative analysis. **a** Human and mouse early embryo single cells were taken from ref. [2]. Pairwise correlation of cells shown as a heatmap; high, red; low, blue. **b** Inferred activity of biological processes described in the original study for the mouse data set; high activities, yellow; low activities, purple. **c** Same as in **b** but for the human data set. **d** Multidimensional scaling (MDS) view of the cross-species BPA-integrated immune cell studies from human and mouse including two human studies—Chromium, brown; and GSE94820, purple—and one mouse study—Tabula Muris, peach. **e** Same data as in **d** showing the MDS view of the distinct immune cell types (colors) showing how cell types (e.g. T-cells, light blue) from both species cluster near one another

in the original study were recapitulated in both human and mouse cells (Fig. 2b, c).

We next performed BPA analysis to compare and align human and mouse immune cells. We included scRNA-Seq profiles of human PBMCs from a healthy donor (Chromium, Fig. 1a, c and Supplementary Fig. 3), mouse spleen and thymus (Tabula Muris, Supplementary Fig. 3)[20] and human monocytes and dendritic cells (GSE94820, Supplementary Fig. 3)[21]. To extend the scope of biological process selection, and to better describe the cellular states, we also included an immunologic gene set (see Methods). Within each individual data set, cell types (Supplementary Fig. 3), as well as cell type-specific biological processes (Supplementary Data 1–3) were recapitulated using BPA, benchmarking the immunologic gene set in interpreting cellular states. Data sources, as well as cell types, for the integrated analysis of the three data sets are shown in Fig. 2d, e. Although some species-specificity was observed, cells are primarily clustered according to cell types, forming clustered populations of T-cells and phagocytes (composed of monocytes and dendritic cells) contributed by both species.

To test the significance of this result, we used an analysis of variance to measure the separation of cells of the same cell type to those of the same species in the transformed data. Specifically, we calculated the variance within cell types (V_t) and the variance within species (V_s) and found that the ratio (V_t/V_s) was indeed significantly higher than unity ($P = 5.0 \times 10^{-12}$, F test, See Methods). In contrast, mapping genes to orthologs and using orthologous gene expression to combine the data sets resulted in a significantly lower level of species mixing within cell types (F test $P = 0.0035$). Visual inspection of the hierarchical clustering dendrogram of the BPA result (Supplementary Fig. 6) shows that the major division of the cells falls along the axis defining the B, T, and monocyte cell types and the human–mouse divisions occur as minor splits further down the tree. In contrast, clustering the ortholog gene expression data produces a dendrogram with a major split that falls along the species and data set distinctions. These results indicate that single cells from different experiments and across species can be combined effectively using BPA.

To evaluate the performance of BPA analysis in separating closely and distantly related species, we integrated human (E-MTAB-3929)[13], zebrafish (GSE66688)[22], and mouse (GSE65565)[23] embryo-related single cells. Single cells from human preimplantation embryos, especially in the early stages (E3 and E4), overlapped with mouse embryonic stem cells (Supplementary Fig. 7). This finding is consistent with the mouse ESCs being derived from blastocyst, an early-stage preimplantation embryo. Single cells from human preimplantation embryos and mouse ESCs are separated from single cells of the zebrafish embryo, indicating zebrafish are less related to the cells of human and mouse. On the other hand, owing to different scRNA-Seq platforms, gene expression profiles of the three data sets are not comparable. Therefore, expression-based Seurat[24] analysis failed in cross-species integration (Supplementary Fig. 8). This further confirms that the BPA transform is resilient to common technical batch effects among single-cell expression profiles.

Discussion

In summary, we have presented a GSEA-based approach for inferring BPAs within single cells. Transforming the gene-level data into interpreted features representing cellular processes produces a data set that is less influenced by common technical noises in scRNA-Seq profiles. The transformed data preserve the integrity of cellular states and their transitions. Moreover, analysis in BPA space enables a straightforward comparison of cell states across platforms and species with very few parameter settings and without the need for predicting orthologs. Using the BPA approach, model organism data sets can be directly combined with a human counterpart to uncover inter-species commonalities and differences in evolution, normal development, and diseases at the resolution of individual cells. Other transformations of the data are possible that could lead to similar benefits. For example, master regulator analysis approaches like VIPER[6,7] and SCENIC[27], which infer the activity of transcription factors from their targets, might also be used. Other transforms such as deep autoencoders[28] or network diffusion approaches[29] are also possible. Methods that preserve the explicit association with genes (e.g., VIPER, SCENIC, and network diffusion) can be run prior to BPA transformation and thus offer the potential for exploring a combination of approaches. Finally, in addition to single-cell data sets, BPA is applicable to bulk expression analysis (Supplementary Fig. 9), serving as a general approach for describing and combining biological states across data sets and species.

Methods

Gene set selection and filtering. Gene sets were downloaded from the CRAN R msigdbr package that includes the MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp) data set[5]. In this study, the biological process subset of MSigDB's GO gene sets (C5) and the MSigDB immunologic gene sets (C7) for *Homo sapiens*, *Mus musculus*, and *Danio rerio* were used in the presented analyses. C7 contains paired upregulated and downregulated gene sets. As single-cell RNA-sequencing profiles have high drop-out rates, the accuracy for quantifying underexpressed genes is low. To avoid using such genes, as well as redundancy caused by paired gene sets, we retrained only the up-regulated gene sets in C7 for analyses in the study. The gene sets were filtered by the number of genes in the gene set as aREA-based BPA tends to assign higher activities to larger gene sets. On the other hand, smaller gene sets yield less robust activities as indicated by an increase in the coefficient of variance as gene set sizes decrease (Supplementary Fig. 10). For the C5-based gene sets, we mitigated these biases by restricting the size of the gene sets to range between 50 and 100 genes. For the C7 gene sets, we controlled for size by selecting gene sets ranging between 190 and 210. After filtering, for C5, 685, 680, and 722 gene sets were kept in mouse, human, and zebrafish, respectively. For C7, 1741 and 2156 gene sets were kept in mouse and human, respectively (Supplementary Fig. 11).

Simulating drop-out effects in RNA-Seq data. We randomly selected 20 samples from each of the GTEx lung and esophagus bulk RNA-sequencing samples. To mimic the single-cell RNA-sequencing scenario, the simulated drop-out rate was determined by using a drop-out probability that is a function of the expression level of each transcript. For example, more lowly expressed transcripts have a higher likelihood of drop-out than those that are more highly expressed. Such a relationship was determined empirically by analyzing Chromium and Ginis et al.[18] data sets. Although the drop-out rate varies depending on the cell type, such a relationship still holds. The simulation was done in lung and esophagus data sets separately, yielding 81.32% and 82.84% overall drop-out rate (Supplementary Fig. 2).

BPA transformation of single-cell RNA-Seq data. Single sample GSEA (ssGSEA) was used to quantify activity profiles from the original gene expression[25]. ssGSEA was performed using the aREA() function from the Bioconductor R viper package. The aREA() function performs a rank-based enrichment analysis, which provides a computationally efficient analytical approximation of the widely-used GSEA tool[6,7]. As the normalized enrichment scores given by the aREA() function are essentially z scores they can be directly merged across data sets without any further normalization.

The original gene expression is provided as the input signature to ssGSEA, to quantify a measure of activity for each biological process. This contrasts the typical use of single-cell GSEA analysis that quantifies a relative activity by subtracting the average expression signature of the entire data set from the expression of each individual cell[6,7]. In this study, we used the log-transformed original single-cell expression as the signatures for inference instead of using a differential measure that would strongly depend on the composition of the data set and possibly impact cross-data set integration (Supplementary Fig. 12).

Many genes occur in multiple pathways, creating redundancies among the gene sets owing to the overlap of their members. However, we found that down weighting gene sets according to their overlaps using shadow analysis in the Bioconductor R viper package[6,7] produced BPAs with very little difference from their unweighted versions (see Supplementary Fig. 13). For this reason, we chose to keep all of the genes within a biological process to insure a comprehensive representation and to use the unweighted gene sets to analyze the data in this study.

We tested the sensitivity of BPA to select relevant pathways using a data set with known cell types. Our assumption is that relevant pathways should contain some genes with expression levels that vary between cells of different cell types. Thus, the standard deviation (SD) across all cells may reflect that a gene set encodes cell type-related information. As shown in Supplementary Fig. 14, within the heterogeneous Chromium 10x PBMC data set, the gene sets in the MSigDB collection gave overall higher SD values compared with random gene sets (negative control), indicating that some pathway sets reflect immune cell-related information. This suggests that the SD values could be used to select-specific biological processes to analyze a given data set. In support of this, context-matching biological processes of the immune system (C7) gave higher SD values than general pathway gene sets (C5) on average for the PBMC data set (Supplementary Fig. 15).

Statistical test, dimensionality reduction, and clustering. We used a statistical test to assess the degree to which data sets of different species are concordant either in the original gene expression space or upon transformation using BPA. Intuitively, cells of the same cell type but from different species should overlap thus maintaining the separation of distinct cell types yet mixing the species. To quantify this notion, we used an analysis of variance in which we measured the variance within species (V_s) and within cell types (V_t) and calculated their ratio V_t/V_s. Values higher than unity indicate cells of the same cell types are closer together (lower variance) compared with a cell of a different cell type from the same species (higher variance). The ratio follows an F-distribution with $(N–M_t)$ and $(N–M_s)$ degrees of freedom where N is the number of cells, M_t is the number of cell types ($M_t = 3$ for the PBMC data set) and M_s is the number of data sets ($M_s = 3$, including PBMC, Tabula Muris and GSE94820 for the human–mouse analysis). To combine the original gene expression vectors across species without transforming with BPA, we mapped all genes from human and mouse to their orthologous counterparts using the human–mouse mapping table available in the R Bioconductor biomaRt package.

To visualize both the BPA transformed and original gene expression spaces, we produced a multidimensional scaling (MDS) dimensionality reduction using the cmdscale() function in the R stats package. To detect cell types, DBSCAN[26] clustering on the 2D MDS space was performed using the dbscan() function in the R CRAN dbscan package. To be consistent with the pseudotime analysis of the original study[13], we plotted the data in 2D with t-SNE[15] using the Rtsne() function from the R CRAN tsne package, followed by pseudo-lineage analysis using principal curves[14]. Principal curves were calculated using the principal.curve() function from the CRAN R princurve package. We also performed hierarchical cluster analysis on the data sets to produce the dendrograms illustrating the BPA data divides cell type more significantly than species compared with using the original gene expression data associated with predicted orthologs.

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

Data availability

GTEx bulk RNA-sequencing profiles can be found from the website: https://gtexportal.org/home/. We downloaded the provided normalized expression profiles and log-transformed them into log2(RPKM +1) for downstream analysis. Mouse and human esophageal epithelium normalized bulk RNA-Seq expression profiles: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116272. scRNA-Seq profiles for the human PBMC data set were taken from healthy donors generated using 10x Genomics V2 and V1 chemistry and available from: https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc4k, https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k. We downloaded the provided UMI counts and normalized by the sequencing depth as log2(TPM +1) for downstream analysis. scRNA-Seq profiles for the human preimplantation embryo data set, including time point: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3929/. We downloaded the provided normalized expression profiles and log-transformed them into log2(RPKM +1) for downstream analysis. scRNA-Seq profiles for the mouse embryonic stem cell data set: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525. We downloaded the provided normalized expression profiles and log-transformed them into log2(TPM +1) for downstream analysis. scRNA-Seq profiles for the zebrafish embryo data set: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66688. We downloaded the provided normalized expression profiles and log-transformed them into log2(TPM +1) for downstream analysis. scRNA-Seq profiles for the human and mouse early embryos, including time point annotations: https://www.nature.com/articles/nature12364. We downloaded the provided normalized expression profiles and log-transformed them into log2(RPKM +1) for downstream analysis. scRNA-Seq profiles for the human monocytes and dendritic cells, including cell type annotation: http://science.sciencemag.org/content/356/6335/eaah4573. We downloaded the provided normalized expression profiles and log-transformed them into log2(TPM +1da) for downstream analysis. Tabula Muris data sets: https://www.nature.com/articles/s41586-018-0590-4. We downloaded the provided counts of spleen and thymus data sets and normalized by the sequencing depth as log2 (CPM +1) for downstream analysis. All relevant data and analysis results are available from the authors.

Code availability

All scripts are available at https://github.com/hd2326/BiologicalProcessActivity.

Received: 5 February 2019; Accepted: 7 October 2019;
Published online: 25 October 2019

References

1. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
2. Xue, Z. et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593 (2013).
3. Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421 (2018).
4. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740 (2014).
5. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
6. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838 (2016).
7. Ding, H. et al. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* **9**, 1471 (2018).
8. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155 (2015).
9. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
10. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
11. Lin, Y. et al. H. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci.* **116**, 9775–9784 (2019).
12. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685 (2019).
13. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
14. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502–516 (1989).
15. Maaten, L. V. D. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
16. Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. methods* **15**, 267 (2018).
17. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
18. Ginis, I. et al. Differences between human and mouse embryonic stem cells. *Dev. Biol.* **269**, 360–380 (2004).
19. Sakoe, H., Chiba, S., Waibel, A. & Lee, K. F. Dynamic programming algorithm optimization for spoken word recognition. *Read. Speech Recognit.* **159**, 224 (1990).
20. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium. *Nature* **562**, 367 (2018).
21. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
22. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495 (2015).
23. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
24. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411 (2018).
25. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108 (2009).
26. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings* **96**, 226–231 (1996).
27. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083 (2017).

28. Hu, Q. & Greene, C. S. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pac. Symp. Biocomput.* **24**, 362–373 (2019).
29. Elyanow, R., Dumitrascu, B., Engelhardt, B. E., & Raphael, B. J. netNMF: A network regularization algorithm for dimensionality reduction and imputation of single-cell expression data. RECOMB Proceedings 2019 (2019).

Acknowledgements

J.M.S. was supported by a grant 5R01GM109031 from the NIGMS. J.M.S. and H.D. were supported by a grant from the Chan-Zuckerberg Initiative's Human Cell Atlas portals project. H.D. was supported by a gift from Seagate Technology. J.M.S. and A.B. were supported by grant GC1R-06673-C from the California Institute for Regenerative Medicine's Center of Excellence for Stem Cell Genomics.

Author contributions

H.D. and J.M.S. conceived and initiated the project. H.D., A.B. and Y.Y. collected public data sets and performed the analysis. H.D., A.B. and J.M.S. prepared 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/s41467-019-12924-w.

Correspondence and requests for materials should be addressed to H.D. or J.M.S.

Peer review information *Nature Communications* thanks Laleh Haghverdi, Xuegong Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.