Cross-Species Alignment of Single Cell States with Biological Process Activity

Hongxu Ding, Andrew Blair and Joshua M. Stuart

UC Santa Cruz Genomics Institute and Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA

Correspondence should be addressed to H.D. (hding16@ucsc.edu) or J.M.S (jstuart@ucsc.edu)

H.D. and A.B. contribute equal to the study.

ABSTRACT: The maintenance and transition of cellular states are controlled by orchestrated biological processes. Here we show a transformation of gene expression of single cell RNA-Seq data using gene sets representing biological processes provides a robust description of cellular states. Moreover, as species-independent general descriptors of cellular states, the activity of these biological processes can be used to align single cell states across different organisms.

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 cellular states are controlled by orchestrated biological processes [2], we propose using biological process activities in place of the expression of individual genes. Biological process activity analysis is estimated from an ensemble of dozens of related genes (Figure 1A). In this way discrepancies in individual genes are averaged out, yielding reproducible measurements that are unaffected by common technical noises such as batch effects [3] and drop-out events [4] (Figure 1C,D).

Gene sets have been used for years to infer the activity of biological processes in many applications. The various catalogs of gene sets, e.g. Gene Ontology (GO) and 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. For example, gene set enrichment analysis (GSEA) summarizes the putative importance of a biological process by using the ensemble expression pattern of a set of genes documented to play a role in that specific process [5].

We extend this idea to transform and interpret scRNA-Seq data into inferred process association levels, henceforth called “activities”, using a large collection of gene sets. For this study, we used the Biological Process (BP) portion of the GO collection [6] as one source of gene sets. The gene expression signature of an individual cell can be transformed into a biological process activity profile using the gene members of the associated gene set. Moreover, because the GO-BP terminology is consistent across species, gene set enrichment analysis 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 the species (Figure 1A). In this way, datasets of human and model organisms can be composed directly to reveal functionally analogous cell types across species. We demonstrate the utility of using inferred biological activities to align human and mouse datasets to shed light on their comparative and species-specific biology in early embryo development and in the cell types comprising the immune system.

Distinct, batch-specific clusters can be observed among peripheral blood mononuclear cell (PBMC) scRNA-Seq datasets when gene expression profiles are used (Figure 1B), but no longer apparent when GO-BP activity features are used (Figure 1D). In this example, the clustering of the PBMCs recapitulates the B-cell, T-cells, and monocytes as denoted by the cell type-specific markers CD3E, CD14 and CD20, respectively. In addition, biological process activity 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 datasets of bulk tissue samples were taken from the GTEx lung (L) and esophagus (E) collection and labeled as “original” (Ori) while their counterparts containing simulated drop-out events were labelled “drop-out” (Dro) (Supplementary Figure 1, 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, since single cells will cluster according to drop-out extent rather than measured biological conditions. The inferred biological activity preserves within-tissue correlations, and reduces cross-tissue correlations (Figure 1C). Taken together, inferred biological process activity profiles produced clusters with distinctly enriched PBMC cell types according to marker gene expression (Figure 1D, Figure 2F-H and Supplementary Figure 3), as well as the known ordering of state transitions in a human preimplantation embryo dataset (Figure 1E, Figure 2A and Supplementary Figure 2; see Methods).

We next performed cross-species single cell state alignment using biological process activity 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 [7]. However, as orthologs are usually determined by computational analysis of protein sequence alone [8], the expression pattern of orthologous genes may not be the same across species [9]. The transformed dataset using GO-BP, on the other hand, provides a common set of terms from which detailed gene sets can be retrieved in a species-specific manner [6]. Each species can be analyzed separately using their species-specific gene sets and then merged across species at the biological process activity level assuming the ontology terms 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. While 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 stage to the 4-cell stage; the second step includes the 8-cell stage; and the third step includes only the morula stage. In mouse, the first step is relatively shorter, including the oocyte and pronuclear stages; the second step includes the 2-cell and 4-cell stages; and the last step includes the 8-cell and morula stages (Figure 2A). The data sets transformed with GO-BP produce the expected alignment of the three steps between the two species (Figure 2A), which was further confirmed objectively using

dynamic time warping (Supplementary Figure 5) [7][10], In addition, the stage-specific activation pattern of biological processes determined in the original study were recapitulated in both human and mouse cells (Figure 2B, C).

We next performed biological process activity analysis to compare and align human and mouse immune cells. We included scRNA-Seq profiles of human PBMCs from a healthy donor (Chromium, Figure 1A, C and Figure 2F), mouse spleen and thymus (Tabula Muris, Figure 2G) [11] and human monocytes and dendritic cells (GSE94820, Figure 2H) [12]. 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 dataset, cell types (Figure 2F-H), as well as cell type-specific biological processes (Supplementary Table 1-3) were recapitulated using biological process activity, benchmarking the immunologic gene set in interpreting cellular states. Data sources, as well as cell types for the integrated analysis of the three datasets are shown in Figure 2D and E, respectively. Although some species-specificity was observed, cells are primarily clustered according to cell types, recapitulating T-cell and phagocyte (composed of monocytes and dendritic cells) populations. These results indicate that single cells from different experiments and across species can be aligned using biological process activity. Noticeably and against this trend, human and mouse B-cells failed to co-cluster with one another. We investigated this incongruity by measuring how each biological process was differentially activated in human vs. mouse B-cells. The top differential process was B-cell receptor signaling, with a higher activity in mouse compared to human. This suggests the mouse cells underwent B-cell mediated immune responses to a larger degree compared to the human cells when harvested (Supplementary Figure 4, Supplementary Table 4).

In summary, we have presented an enrichment analysis-based approach for inferring biological process activity among single cells. Transforming the transcript-level data into higher-level features representing cellular processes produces a dataset that is resistant to common technical noises in scRNA-Seq profiles. The transformed data preserves the integrity of cellular states and their transitions. Moreover, analysis in biological process activity space enables a straightforward comparison of cell states across platforms and species. Using this approach, model organisms can be directly

combined with human counterpart datasets to uncover inter-species commonalities and differences in evolution, normal development, and diseases at the resolution of individual cells.

REFERENCES

1. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. Mol Cell. 2015;58:610–20.
View Article PubMed Google Scholar

2. Xue Z, Huang K, Cai C, Cai L, Jiang C-Y, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500:593–7.
View Article PubMed Google Scholar

3. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36:421–7.
View Article PubMed Google Scholar

4. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.
View Article PubMed Google Scholar

5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
View Article PubMed Google Scholar

6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9. View Article PubMed Google Scholar

7. Alpert A, Moore LS, Dubovik T, Shen-Orr SS. Alignment of single-cell trajectories to compare cellular expression dynamics. Nat Methods. 2018;15:267–70.
View Article PubMed Google Scholar

8. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
View Article PubMed Google Scholar

9. Ginis I, Luo Y, Miura T, Thies S, Brandenberger R, Gerecht-Nir S, et al. Differences between human and mouse embryonic stem cells. Dev Biol. 2004;269:360–80.
View Article PubMed Google Scholar

10. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust. 1978;26:43–9.
View Article Google Scholar

11. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562:367–72.
View Article PubMed Google Scholar

12. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356. doi:10.1126/science.aah4573.
View Article PubMed Google Scholar

13. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016;48:838–47.
View Article PubMed Google Scholar

14. Ding H, Douglass EF Jr, Sonabend AM, Mela A, Bose S, Gonzalez C, et al. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. Nat Commun. 2018;9:1471.
View Article PubMed Google Scholar

15. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9 Nov:2579–605.
View Article Google Scholar

16. Ester M, Kriegel H-P, Sander J, Xu X, Others. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. aaai.org; 1996. p. 226–31.
View Article Google Scholar

17. Hastie T, Stuetzle W. Principal Curves. J Am Stat Assoc. 1989;84:502–16.
View Article Google Scholar

18. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016;165:1012–26.
View Article PubMed Google Scholar

METHODS

Biological process activity inference

Gene sets were downloaded from CRAN R package {msigdbr} that provides the MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp) dataset [5]. We included all 7 categories of MSigDB gene sets (C1-7), including positional gene sets (C1), curated gene sets (C2), motif gene sets (C3), computational gene sets (C4), GO gene sets (C5)

oncogenic gene sets (C6) and immunologic gene sets (C7) from 11 species including *Bos taurus*, *Caenorhabditis elegans*, *Canis lupus familiaris*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Sus scrofa*. In this study, we used GO gene sets (C5) biological processes (BP) subsets and immunologic gene sets (C7) from *Homo sapiens* and *Mus musculus*. To avoid bias, we exclude gene sets with too many (>100 for C5, >210 for C7) or too few (<50 for C5, < 190 for C7) genes. Specifically, C7 contains both up-regulated and down-regulated gene sets. Since single-cell RNA-sequencing profiles have high drop-out rate, the accuracy for quantifying under-expressed genes is low. Therefore we only took the up-regulated gene sets in C7 for all analyses. Enrichment analysis was performed using the aREA() function from the Bioconductor R package {viper}. aREA() function performs analytical rank-based enrichment analysis, which provides a computationally efficient analytical approximation of the widely-used GSEA [13][14].

Simulating the drop-out effect

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 absolute 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 relationship was determined empirically by analyzing Chromium and [18] datasets. Although absolute drop-out rate varies depending on cell types, such relationship stands. The simulation was done in lung and esophagus datasets separately, yielding 81.32% and 82.84% overall drop-out rate (Supplementary Figure 1).

Single cell heterogeneity analysis

For the cluster analysis used to detect cell types, we projected single cells from the original biological process activity space onto a two-dimensional space with t-SNE[15], using the Rtsne() function from the CRAN R package {Rtsne}. We then performed DBSCAN [16] clustering on the 2D space, using the dbscan() function from the CRAN R package {dbscan}. For pseudo-lineage analysis, we projected single cells from the

original biological process activity or expression space onto a 2D t-SNE space, followed by pseudo-lineage analysis using principal curves [17], which was adopted in the original study [18]. Principal curves were calculated using the principal.curve() function from the CRAN R package {princurve}.

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.

scRNA-Seq profiles for the human PBMC dataset 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 raw UMI counts and normalized by the sequencing depth as log2(TPM+1) for downstream analysis.

scRNA-Seq profiles for the human preimplantation embryo dataset, 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 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+1) for downstream analysis.

Tabula Muris datasets: https://www.nature.com/articles/s41586-018-0590-4. We downloaded the provided raw counts of spleen and thymus datasets 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.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

H.D. and J.M.S conceived and initiated the project. H.D. and A.B. performed the analysis. All authors prepared the manuscript.

COMPETING INTERESTS

All authors declare no competing interests.

FIGURES

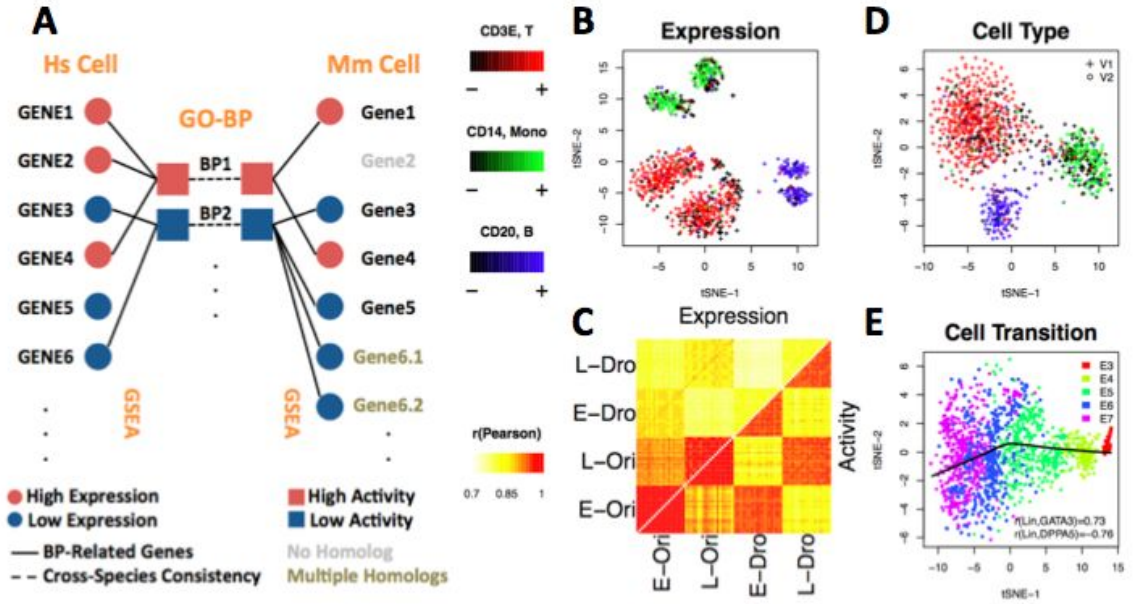


Figure 1. Biological process activity is resistant to technical noises in single-cell RNA sequencing profiles, providing accurate description of cellular heterogeneity. (A) Overview of biological process activity inference. Single cell gene expression profiles for human (outer left column) can be compared to 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 PBMCs (Peripheral Blood Mononuclear Cell) 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. (C) Drop-out events (Dro) were simulated into GTEx lung (L) and esophagus (E) bulk RNA sequencing (Ori) data (Supplementary Figure 1, see Methods); pairwise correlations between samples was computed and plotted. Cellular cell states (D, Figure 2F and Supplementary Figure 3) and state transitions (E, Figure 2A and Supplementary Figure 2), are recapitulated using biological process activity (see Methods). Following the original study [18], pseudo-lineage was constructed using principal curves [17] on t-SNE space [15]. Correlation between pseudo-lineage distance and the expression of known lineage markers is shown.

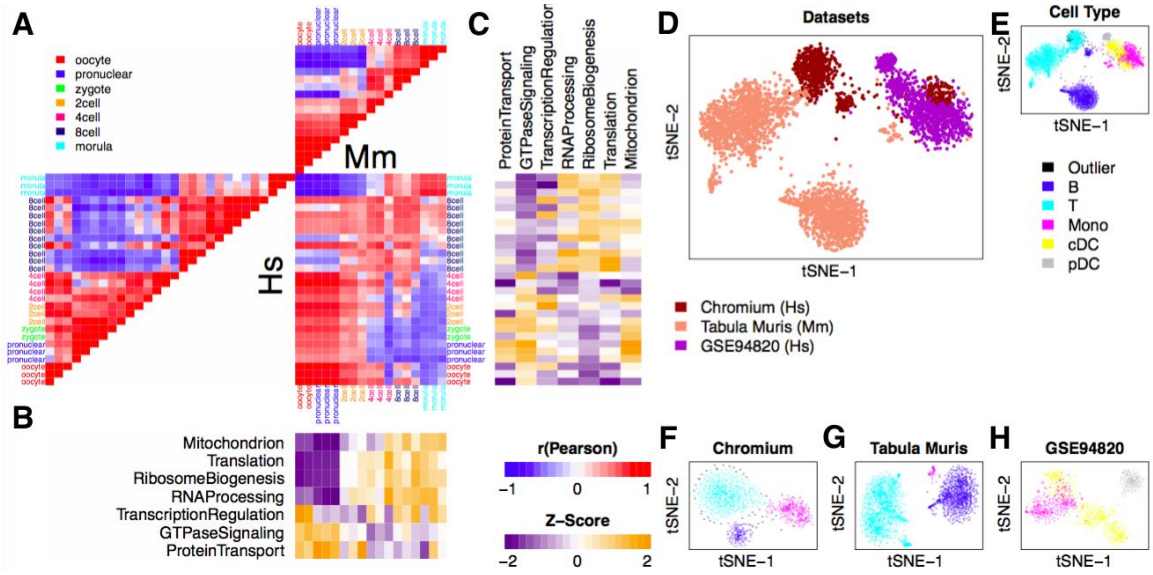
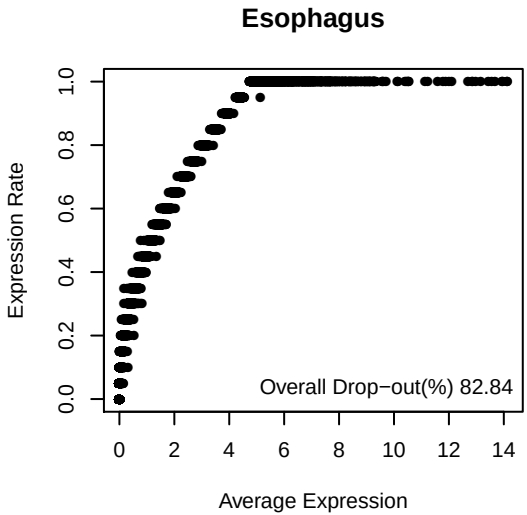
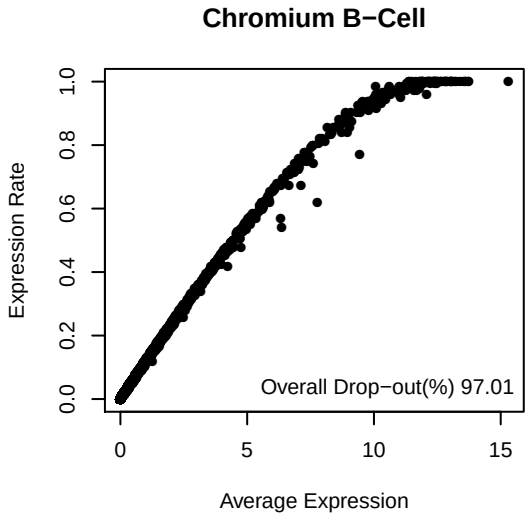
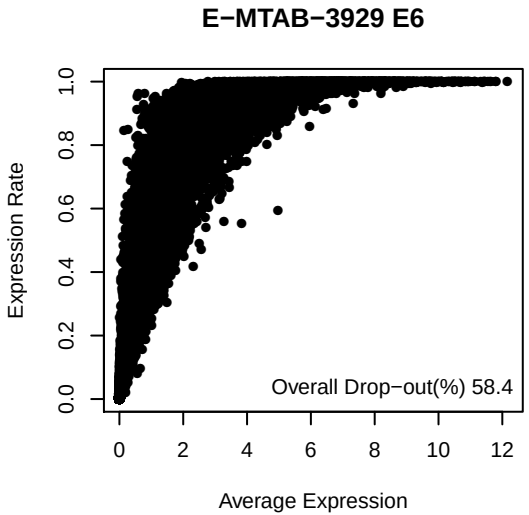
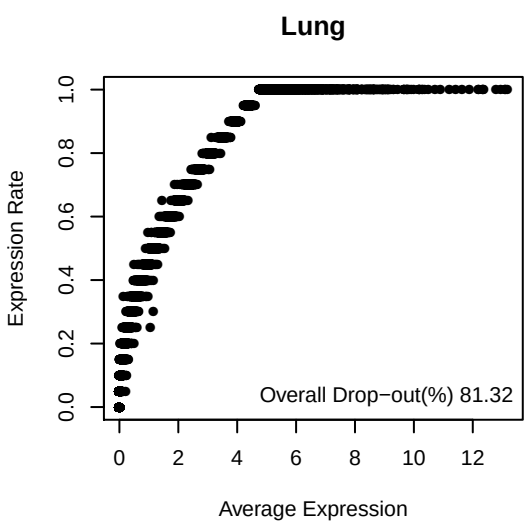


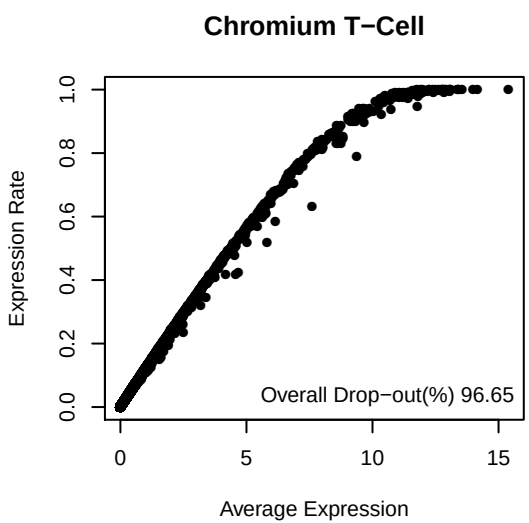
Figure 2. Aligning human and mouse single cell datasets using biological process activity. Human and mouse early embryo single cells were taken from [2] and (A) pairwise correlation of cells were calculated and displayed. (B, C) Inferred activity of biological processes involved in early embryo development described in the original study. For human and mouse immune cells profiling, (D) data sources, as well as (E) cell types were shown according to biological process activity-based single cell alignment. (F-H) Cell type analysis using biological process activity within individual datasets (Supplementary Figure 3, see Methods). Chromium, single-cell RNA sequencing profiles of human PBMC from healthy donor (Figure 1A and C); Tabula Muris, single-cell RNA sequencing profiles of mouse spleen and thymus [11]; GSE94820, single-cell RNA sequencing profiles of human monocytes and dendritic cells [12].

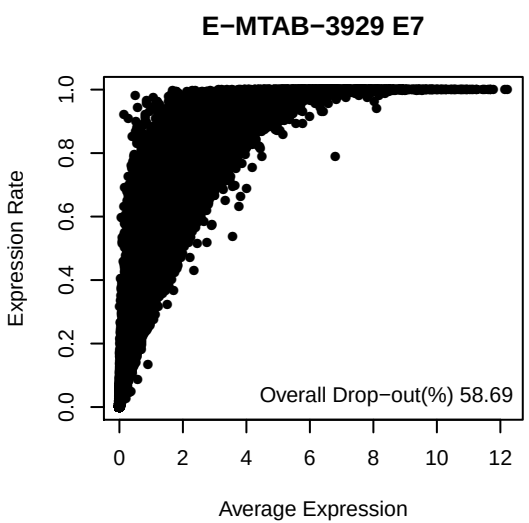

Esophagus
Expression Rate
Average Expression
Overall Drop−out(%) 82.84

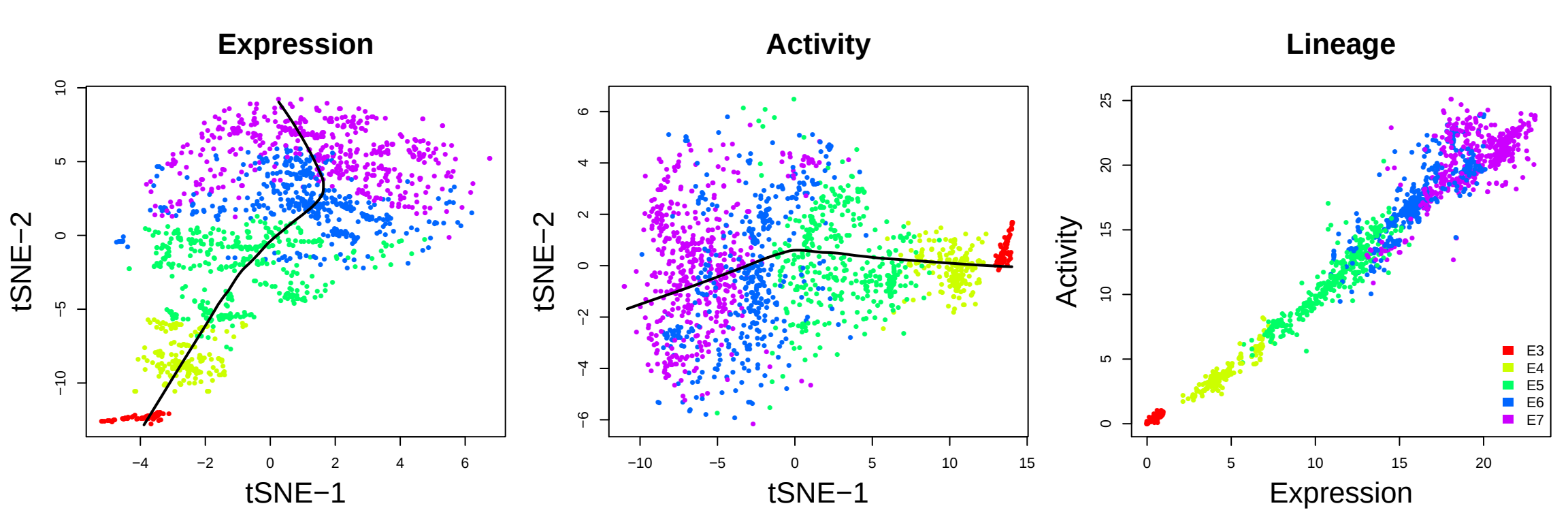

Chromium B−Cell
Expression Rate
Average Expression
Overall Drop−out(%) 97.01

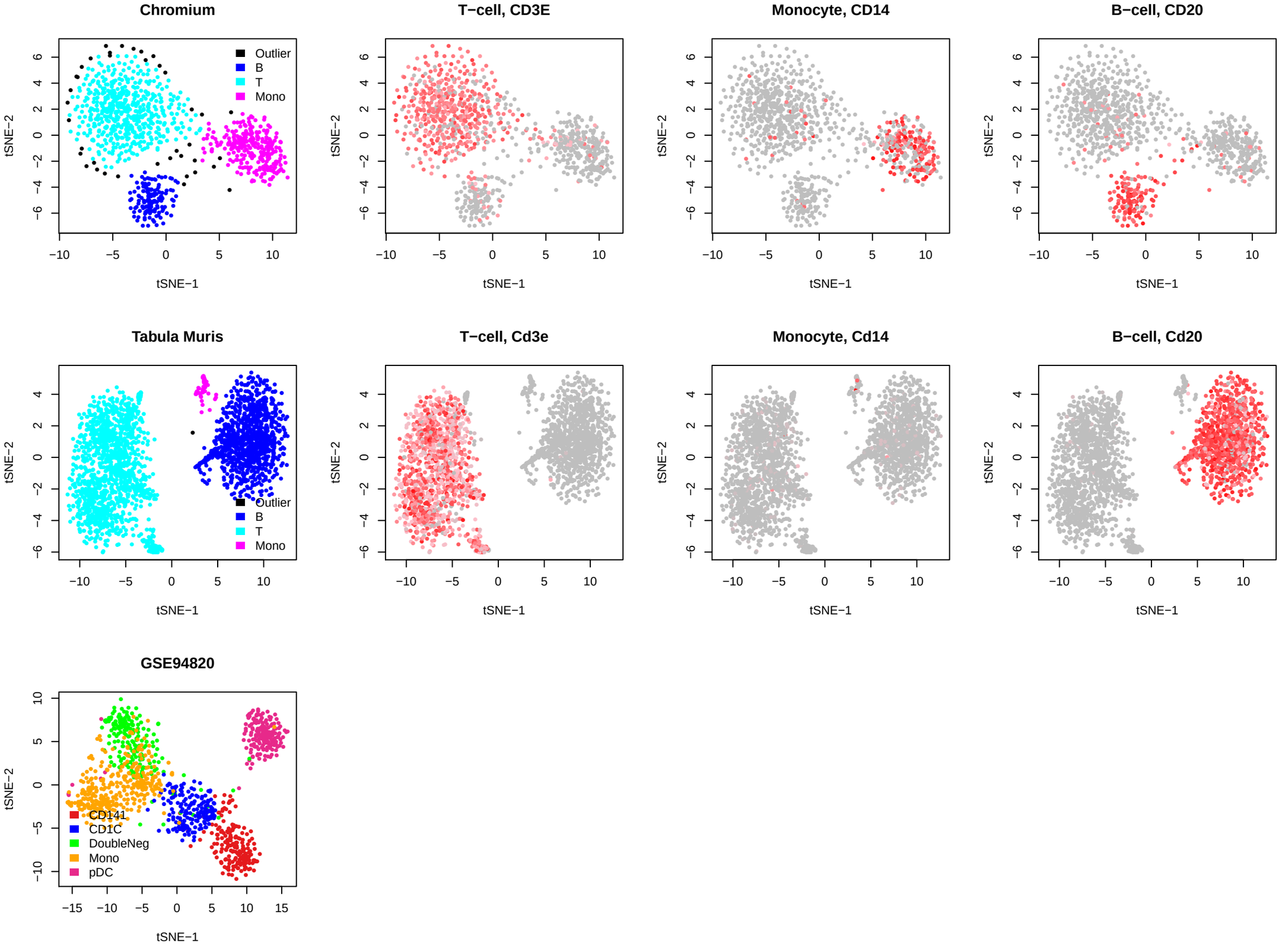

E−MTAB−3929 E6
Expression Rate
Average Expression
Overall Drop−out(%) 58.4

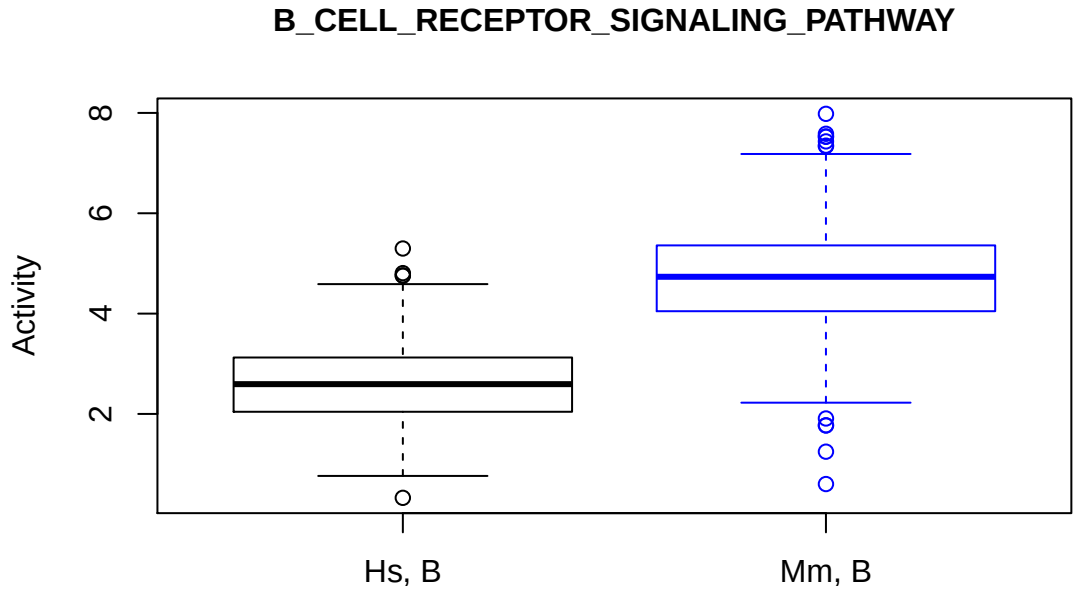

Lung
Expression Rate
Average Expression
Overall Drop−out(%) 81.32

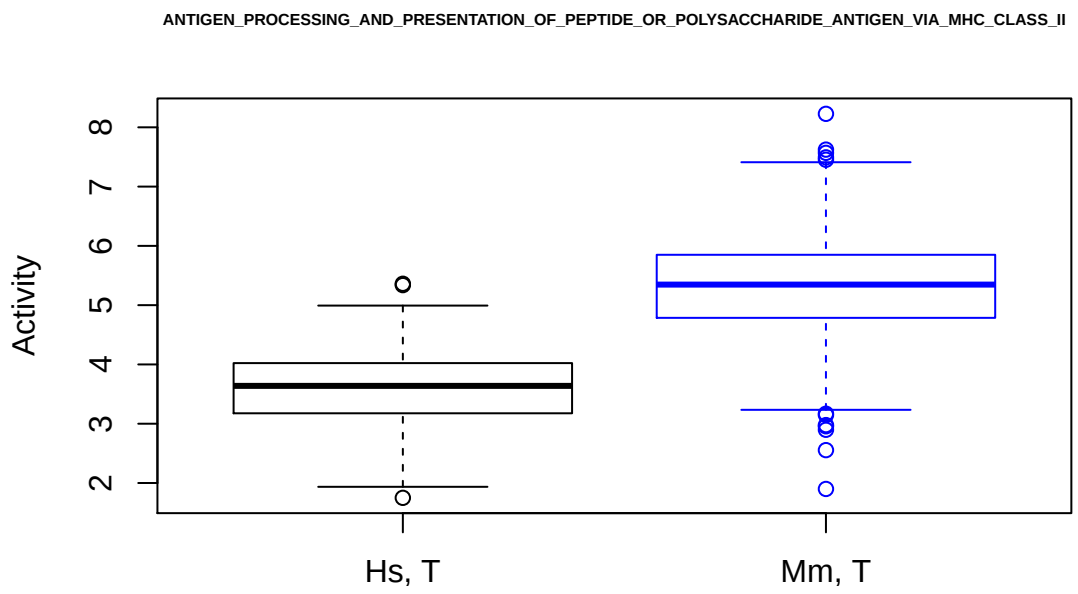

Chromium T−Cell
Expression Rate
Average Expression
Overall Drop−out(%) 96.65


E−MTAB−3929 E7
Expression Rate
Average Expression
Overall Drop−out(%) 58.69


Expression
tSNE−2
tSNE−1
Activity
tSNE−2
tSNE−1
Lineage
Activity
Expression
E3
E4
E5
E6
E7


Chromium
Outlier
B
T
Mono
T−cell, CD3E
Monocyte, CD14
B−cell, CD20
Tabula Muris
T−cell, Cd3e
Monocyte, Cd14
B−cell, Cd20
GSE94820
CD141
CD1C
DoubleNeg
Mono
pDC
tSNE−1
tSNE−2


B_CELL_RECEPTOR_SIGNALING_PATHWAY
Activity
2
4
6
8
Hs, B
Mm, B


ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II
Activity
2
3
4
5
6
7
8
Hs, T
Mm, T

Dynamic Time Warping