Candidate dataset

- Usoskin dataset

622 mouse neuronal cells from the dorsal root ganglion,classified into 11 categories.

Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA-sequencing. Nat Neurosci. 2014;18(1):145–53.http://www.nature.com/doifinder/10.1038/nn.3881.

- 10x Genomics PBMC dataset

Correspond to 2700 single cells sequenced on an Illumina NextSeq 500 using UMIs

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. <http://www.nature.com/doifinder/10.1038/>ncomms14049.

- Islam dataset

represents 44 embryonic fibroblasts and 48 embryonic stem cells in the mouse, sequenced on an Illumina Genome Analyzer II

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21(7):1160–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3129258&tool=pmcentrez&rendertype=abstract.>

- Trapnell dataset

from the preprocessed single-cell data repository CONQUER (http://imlspenticton.uzh.ch:3838/conquer).

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. http://www.ncbi.nlm.nih.gov/pubmed/24658644. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4122333. <http://www.nature.com/articles/nbt.2859.>

-zeisel

characterized 3,005 cells from the primary somatosensory cortex (S1) and

the hippocampal CA1 region, using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing. Gene expression was quantified by unique molecular identifier (UMI) counts. In addition to gene expression measures, we have access to metadata that can be used to assess the methods: batch, sex, number of mRNA molecules. Raw UMI counts and metadata were downloaded from

http://linnarssonlab.org/cortex/.

Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–42 (2015).

<https://science.sciencemag.org/content/347/6226/1138/tab-figures-data>

- mESC dataset.

Kolodziejczyk et al. [37] sequenced the transcriptome of 704 mouse embryonic stem cells

(mESCs), across three culture conditions (serum, 2i, and a2i), using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing. We selected only the cells from the second and third batch, after excluding the samples that did not pass the authors’ QC filtering. This allowed us to have cells from each culture condition in each batch and resulted in a total of 169 serum cells, 141 2i cells, and 159 a2i cells. In addition to gene expression measures, we have access to batch and plate information that can be included as covariates in our model. Raw gene-level read counts were downloaded from http://www.ebi.ac.uk/teichmann-srv/espresso/. Batch and plate information was extracted from the sample names, as done in Lun & Marioni

Kolodziejczyk, A. A. et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 17, 471–485 (2015).

- Glioblastoma dataset.

Patel et al. [6] collected 672 cells from five dissociated human glioblastomas.Transcriptional profiles were generated using the SMART-Seq protocol. We analyzed only the cells that passed the authors’ QC filtering. The raw data were downloaded from the NCBI GEO database (accession GSE57872). Reads were aligned using TopHat with the following parameters: –rg-library Illumina –rgplatform Illumina –keep-fasta-order -G -N 3 –read-edit-dist 3 –no-coverage-search -x 1 -M -p 12. Counts

were obtained using htseq-count with the following parameters

(http://www-huber.embl.de/HTSeq/doc/count.html): -a 10 -q -s no -m union. We applied the dimensionality reduction methods to the 1,000 most variable genes.

Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014).

- OE dataset.

Fletcher et al. [29] characterized 849 FACS-purified cells from the mouse olfactory epithelium(OE), using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing. Gene-level read counts were downloaded from GEO (GSE95601; file GSE95601\_oeHBCdiff\_Cufflinks\_eSet\_counts\_table.txt.gz).

As done in Perraudeau et al. [31], we filtered the cells that exhibited poor sample quality using SCONE [36] (v. 1.1.2). A total of 747 cells passed this filtering procedure. To compare with the original results, we also re-analyze the final repertoire of 13 stable clusters found in Fletcher et al. [29], consisting of 616 cells, downloaded from https://github.com/rufletch/p63-HBC-diff. See Fletcher et al. [29] for details on the original analysis and Perraudeau et al. [31] for details on the ZINB-WaVE based workflow.

Fletcher, R. B. et al. Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution. Cell Stem Cell 20, 817–830 (2017).

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95601>