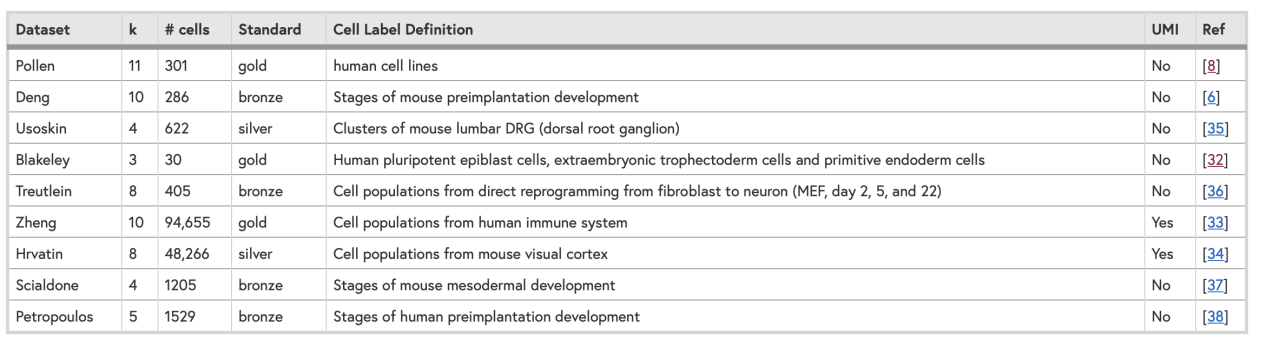
Candidate dataset

Deng dataset

GSE (raw file)



**- Usoskin dataset**

**Mouse,622 cells,3 batches, 4 cell types**

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.

- blakeley dataset

Human, 30 cells, 3 types, 7 batches(embryo)

GSE66507

-pancreas dataset

(human, 3000+cells, pre-filtering is needed, 6 individuals)

No UMIs, use spike-ins

<https://www.cell.com/cell-metabolism/fulltext/S1550-4131(16)30436-3?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1550413116304363%3Fshowall%3Dtrue#secsectitle0165>

<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5061/>

-tasic dataset

Tasic et al.3 characterized more than 1600 cells from

the primary visual cortex (V1) in adult male mice, using a set of established Cre

lines. Single cells were isolated by FACS into 96-well plates and RNA was reverse

transcribed and amplified using the SMARTer kit. Sequencing was performed

using the Illumina HiSeq platform, yielding 100 bp-long reads. We selected a subset

of three Cre lines, Ntsr1-Cre, Rbp4-Cre, and Scnn1a-Tg3-Cre, that label layer 4, layer 5, and layer 6 excitatory neurons, respectively. This subset consists of 379

cells, grouped by the authors into 17 clusters; we excluded the cells that did not

pass the authors’ quality control filters and that were classified by the authors as

“intermediate” cells between two clusters, retaining a total of 285 cells. Gene

expression was quantified by gene-level read counts. Raw gene-level read counts

and QC metrics (see below) are available as part of the scRNAseq Bioconductor R

package (https://bioconductor.org/packages/scRNAseq).

Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell

transcriptomics. Nat. Neurosci. 19, 335–346 (2016). <https://www.nature.com/articles/nn.4216>

Data pre-processed by Haghverdi et al., MNN

For the two hematopoietic datasets, we downloaded the read count matrix of the 1,920 cells proled by Paul et al. [27] and the 2,729 cells labeled as myeloid progenitor cells by Nestorowa et al. [34] from the NCBI Gene Expression Omnibus (GEO) with the accession numbers GSE72857 and GSE81682. Following Brennecke et al. [35], we sorted the genes according to their adjusted variance-mean ratio of expression levels in both datasets separately and focused on the 3,470 genes that are highly variable in both datasets.

Two of the pancreas datasets proled by the CEL-seq2 platform were downloaded from GEO with accession number GSE80176 [29] and GSE86473 [28]. The two datasets assayed by the SMART-seq2 platform were obtained from GSE85241 [36] and from ArrayExpress accession number E-MATB-5061 [30]. Following Haghverdi et al. [11], we excluded cells with low library sizes (< 100; 000 reads), low numbers of expressed genes (> 40% total counts 15 from ribosomal RNA genes), or high ERCC content (> 20% of total counts from spike-in transcripts) resulting in 7,095 cells. We selected the 2,480 highly variable genes shared by the four datasets according to Brennecke et al. [35] by sorting the ratio of variance and mean expression level after adjusting technical noise with the variances of spike-in transcripts. The cell types of the two datasets proled by the CEL-seq2 platform were labelled according to Lawlor et al. [28] and Grun et al. [29], with the GCG gene marking alpha islets, INS for beta islets, SST for delta islets, PPY for gamma islets, PRSS1 for acinar cells, and KRT19 for ductal cells. The cell types of the other two datasets assayed by the SMART-seq2 platform were provided in their metadata.

- pollen dataset

Human, 301 cells, 11 cell types, missing batch info

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053–8.

- 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 (UMI)

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 (umi)

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>

- Jaitin dataset(detct

GSE54006\_umitab.txt.gz A Processed data table. Each row contains the mRNA counts (post-filtering RMT counts) of a gene per each well (columns).

"Column names specify the batch ID and well ID, delimited by underscore."

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A. and others. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779.

- satija dataset

Satija R, Farrell JA, Gennert D, Schier AF et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015 May;33(5):495-502. PMID: 25867923