

In the format provided by the authors and unedited.

Droplet scRNA-seq is not zero-inflated

Valentine Svensson 

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. e-mail: v@nxn.se

Methods

Estimating common dispersion

For the negative binomial distribution with mean μ and dispersion ϕ , the variance follows the relation

$$\text{Var} = \mu + \phi \cdot \mu^2.$$

For every gene the empirical variance and empirical mean was calculated. Using these the per-dataset ϕ parameter was fitted using the `curve_fit` function in the `scipy.optimize` package.

Estimating gene-wise dispersion

For each dataset and gene the (p, r) parameterization of the negative binomial distribution was used. The p and r parameters were optimized by maximizing the log likelihood. From the p and r parameters the mean and dispersion parameters can be calculated by $\mu = (1 - p) \cdot r/p$ and $\phi = 1/r$. The fast Python implementation from https://github.com/gokceneraslan/fit_nbinom was used.

Expected fraction zeros for Poisson distribution

For a Poisson distribution with mean μ the probability of observing a value of 0 is

$$P(0|\mu) = \exp(-\mu).$$

Data

Single cell RNA-seq data from 10x genomics was downloaded from <https://support.10xgenomics.com/single-cell-gene-expression/datasets/>. In particular, PBMC data was downloaded from http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_v3/pbmc_1k_v3_filtered_feature_bc_matrix.tar.gz. Data from NIH3T3 cells and HEK293T cells were extracted from http://cf.10xgenomics.com/samples/cell-exp/3.0.0/hgmm_5k_v3/hgmm_5k_v3_filtered_feature_bc_matrix.tar.gz. To separate NIH3T3 cells and HEK293T cells were first filtered to have at least 2,000 UMIs, and then assigned to HEK293T if they had 20 times more UMI's from human than from mouse, and vice versa for NIH3T3. Other cells were discarded. The count matrix for the GemCode data with ERCCs (Zheng et al 2017) was downloaded from http://cf.10xgenomics.com/samples/cell-exp/1.1.0/ercc/ercc_filtered_gene_bc_matrices.tar.gz.

Count matrices for Chromium droplets with human brain RNA and spike-ins (Svensson et al 2017) data were generated as previously described²⁷.

The count matrix for the InDrops data of droplets with K562 RNA and ERCC spike-ins was downloaded from GEO using accession GSM1599501.

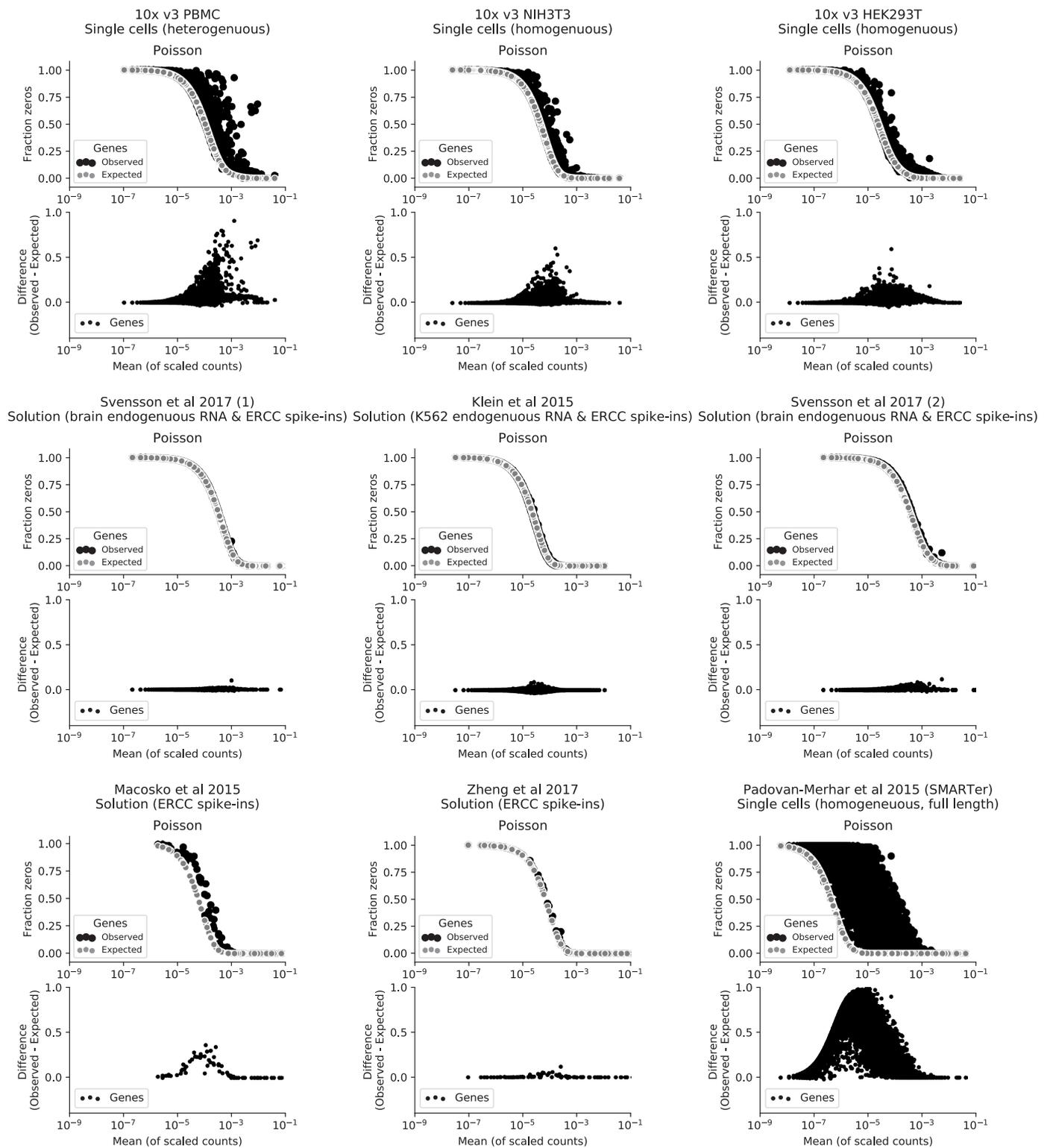
Data for Drop-seq data with ERCC spike-ins in droplets was downloaded from GEO with accession GSM1629193.

Full length SMARTer counts of human fibroblasts was downloaded from GSE66053 on GEO.

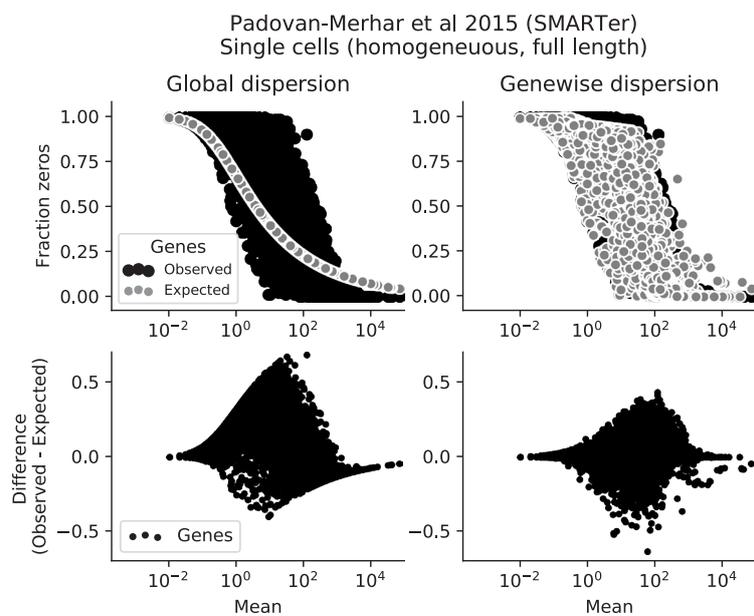
For ease of access, a folder containing all datasets with results in H5AD format³¹ have been deposited to CaltechDATA with DOI [10.22002/D1.1264](https://doi.org/10.22002/D1.1264)

All analysis and figure generation functions are available in Jupyter notebooks at the same CaltechDATA repository.

Supplementary Figure 1



Supplementary Figure 2



Zeros in full-length low-throughput scRNA-seq data. The fraction of wells with zero count for each gene is shown compared to the mean read count of the gene. Expected fraction with common dispersion displayed on left, and with gene-wise dispersion on right. Below plots the difference between observed and expected fractions is displayed.