CS 289 Final Project:
Modelization of single-cell gene expression along a timeline

Daniel Amar Ariane Lozac'hmeur Hector Roux de Bézieux Alexandre Vincent

March 23, 2018

Motivation and goal: Recent development in New Generation Sequencing (NGS) has allowed investigators to conduct high-throughput assays, simultaneously analyzing thousands of biological factors of interest. One such type of assay is single-cell RNA-seq, where the expression levels of every gene are recorded for a single cell. This technique leads, however, to large levels of dropouts: genes with low counts are not detected and appear as "technical zeros" in the final count matrix, which lead to bias in downstream analysis. Modeling this drop-out phenomenon has been the focus of many studies [1]. Recently, models focusing on zero-inflated negative binomial have proven efficient in many cases [2]. However, those models assume independent expression across samples. In a development process like embryogenesis, mean gene-expression levels at previous times help infer present gene-expression levels. We therefore propose to develop a new classification model that could incorporate those informations.

Dataset Description: We find five datasets containing gene expression data of mus musculus embryos, from Qiaolin Deng et al. in 2014[3], Zhigang Xue et al. in 2013 [4]Biase et al. in 2014 [5], Maud Borensztein [6] and Xiaoying Fan [7]. Each dataset will be considered individually, to avoid the problem of between-datasets normalization. A preliminary pass at the dataset filtered and normalized the read counts. Then, using slingshot [8], we obtain, for each dataset, a pseudo time. Those pre-processing steps where all conducted using R and Bioconductor. The final output that will be used for this project is a list of arrays, one per experiment. Each row of an array represents a gene, each column a time point and the array cell contain the gene count for that time point.

Method: All subsequent analysis will be conducted in python. Bases on techniques proposed previously, we will consider each gene individually and independently. For each gene, we model counts using a mixture model. Each read can come either from a dropout or another distribution to be defined (preliminary ideas include truncated Gaussian and negative binomial). For all non-dropouts counts, we can also fit a running mean along the time-line. Zero-counts are weighted by their posterior probability of being non-dropouts. This setting naturally suggest implementing an EM algorithm but other approaches will be considered.

Since this process will need to be implemented across 9000 genes and several datasets, the trade-off between accuracy of the model and computational speed will be a key factor in model selection.

References

[1] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740, 2014.

[2] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. Zinbwave: A general and flexible method for signal extraction from single-cell rna-seq data. *bioRxiv*, 2017.

[3] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[4] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell rna [thinsp] sequencing. *Nature*, 500(7464):593–597, 2013.

[5] Fernando H Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome research*, 24(11):1787–1796, 2014.

[6] Maud Borensztein, Laurène Syx, Katia Ancelin, Patricia Diabangouaya, Christel Picard, Tao Liu, Jun-Bin Liang, Ivaylo Vassilev, Rafael Galupa, Nicolas Servant, et al. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. *Nature Structural and Molecular Biology*, 24(3):226, 2017.

[7] Xiaoying Fan, Xiannian Zhang, Xinglong Wu, Hongshan Guo, Yuqiong Hu, Fuchou Tang, and Yanyi Huang. Single-cell rna-seq transcriptome analysis of linear and circular rnas in mouse preimplantation embryos. *Genome Biology*, 16(1):148, Jul 2015.

[8] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *bioRxiv*, page 128843, 2017.