

SOFTWARE

Open Access



# stochprofML: stochastic profiling using maximum likelihood estimation in R

Lisa Amrhein<sup>1,2</sup> and Christiane Fuchs<sup>1,2,3\*</sup>

\*Correspondence:  
christiane.fuchs@uni-bielefeld.de  
<sup>3</sup> Faculty of Business  
Administration and Economics,  
Bielefeld University,  
Universitätsstrasse 25,  
33615 Bielefeld, Germany  
Full list of author information is  
available at the end of the article

## Abstract

**Background:** Tissues are often heterogeneous in their single-cell molecular expression, and this can govern the regulation of cell fate. For the understanding of development and disease, it is important to quantify heterogeneity in a given tissue.

**Results:** We present the R package stochprofML which uses the maximum likelihood principle to parameterize heterogeneity from the cumulative expression of small random pools of cells. We evaluate the algorithm's performance in simulation studies and present further application opportunities.

**Conclusion:** Stochastic profiling outweighs the necessary demixing of mixed samples with a saving in experimental cost and effort and less measurement error. It offers possibilities for parameterizing heterogeneity, estimating underlying pool compositions and detecting differences between cell populations between samples.

**Keywords:** StochprofML, Stochastic profiling, Gene expression, Cell-to-cell heterogeneity, Mixture models, Deconvolution, Maximum likelihood estimation, R

## Background

Tissues are built of cells which contain their genetic information on DNA strings, so-called *genes*. These genes can lead to the generation of *messenger RNA (mRNA)* which transports the genetic information and induces the production of *proteins*. Such mRNA molecules and proteins are modes of expression by which a cell reflects the presence, kind and activity of its genes. In this paper, we consider such *gene expression* in terms of quantities of mRNA molecules.

Gene expression is stochastic. It can differ significantly between, e.g., types of cells or tissues, and between individuals. In that case, one refers to *differential gene expression*. In particular, cells can be differentially expressed between healthy and sick tissue samples from the same origin. Moreover, cells can differ even within a small tissue sample, e.g. within a tumour that consists of several mutated cell populations. Mathematically, we regard two populations to be different if their mRNA counts follow different probability distributions. If there is more than one population in a tissue, we call it heterogeneous. The expression of such tissues can be described by mixture



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.