Computational Deconvolution of Complex Transcriptomics Data from Clinical Trials

Ting Gong[1*], Nicole Hartmann[2], Isaac S. Kohane[3], Volker Brinkmann[4], Bolan Linghu[1], Frank Staedtler [2], Martin Letzkus[2], Joseph D. Szustakowski[1]

[1] Biomarker Development, Novartis Institutes for BioMedical Research, Cambridge MA, USA, 02139

[2] Biomarker Development, Novartis Institutes for BioMedical Research, Basel, Switzerland

[3] Harvard Medical School, Children's Hospital Medical Center, Boston MA, USA 02115

[4]Department of Autoimmunity, Transplantation and Inflammation, Novartis Institutes for BioMedical Research, Basel, Switzerland

***Corresponding author**:

Ting Gong, Ph.D.

E-mail: ting.gong@novartis.com

Abstract

Samples collected from human subjects in clinical trials possess a level of cell type heterogeneity that can hinder or obfuscate the analysis of data derived from them. Particularly, expression profiles from blood represent a complex mixture of circulating cell types of varying origin, function.

We have developed approaches that explicitly built upon a linear latent variable model, in which expressions from a mixed cell population are modeled as the weighted average of expressions from different cell types. We estimate proportions of different pure cell/tissue types and gene expression profilings of distinct phenotypes, with a focus on complex samples collected in clinical trials, through constrained quadratic programming.

We have applied our methods to several well controlled benchmark data sets with known mixing fractions. Agreement between predicted and actual mixing fractions was excellent and robust to the experimental system. In addition, we have applied our method to more challenging mRNA expression profiling data from whole blood samples collected in clinical trials. Our method was able to predict mixing fractions for more than ten species of circulating cells, and was even able to provide accurate estimates for relatively rare cell types (<10% total population). The concordance of our predictions with measured Complete Blood Counts (CBC) in terms of Pearson correlation was beyond 0.8.

Our work demonstrates that precisely identifying and applying transcriptional patterns of purified cell types will serve as the first step towards a more comprehensive approach to address the changes in the blood transcriptome to disease and drug.