# Modeling changes in cell populations between groups of samples of scRNA-seq data

Modrák Martin1, Niederlová Veronika2,3, Neuman Vít4, Šumník Zdeněk4, Štěpánek Ondřej2

1 Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic

2 Laboratory of Adaptive Immunity, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic.

3 Department of Cell Biology, Faculty of Science, Charles University in Prague, Czech Republic.

4 Department of Pediatrics, Second Faculty of Medicine, Charles University in Prague & Motol University Hospital, Prague, Czechia

Once cell populations have been identified in single cell sequencing data, it may be desirable to understand how the abundance of those populations changes between two or more groups of samples. Taking the assignment of cells to populations as given, two interesting issues arise: First, we need to address the compositional nature of the data, i.e. that the total number of cells sequenced is determined by the experimental design and we thus obtain only relative abundances. Second problem is that cell populations typically form a hierarchical structure. This needs to be reflected in modeling and forces us to ask whether it is more meaningful to look at abundance relative to the total or relative to parent population.

We argue that the compositional structure could be relatively easily bypassed as measuring the total number of cells per sample is often possible and even when it is not, the problem is not huge. We further show that it is natural to model the hierarchical structure in a hierarchical model (i.e. nested random effects) – this model can then answer questions about both absolute abundance and abundance relative to a parent population. Finally, we take note of an – as far as we are aware – unsolved problem of disentangling discrete population structure and continuous latent states of cells within population (cell cycle, proliferation).

We illustrate the issues and possible solutions with an example of comparing populations of peripheral blood T‑cells from children newly diagnosed with type 1 diabetes mellitus and healthy donors.