Interaction maps from incompletely dissociated primary tissue

Advances in single-cell biology has enabled us to investigate isolated single cells at an unprecedented resolution. Current methods can measure subtle changes in cell state, revealing specialized minor subpopulations within a cell type. However, functional characterization of such subpopulations is still challenging. A promising approach to untangling functional characterization is spatial transcriptomics, where genome-wide measurements of gene expression are enriched with spatial attributes. Such approaches allow us to determine the relative positions of cell subpopulations. Unfortunately, current spatial transcriptomic approaches suffer either from problems with sensitivity, are limited to a restricted panel of genes, or the spatial attributes are of low resolution.

In order to deconvolute the cell connectome, we sort multiplets of incompletely dissociated cells along with single cells from the same sample. The single cells provide a set of blueprints of possible sub-cell types, and subsequently we machine learning approaches to infer the composition of multiplets. Using this method we can measure the composition of a very large number of multiplets onto a lattice consisting of sub-cell types of arbitrary complexity.

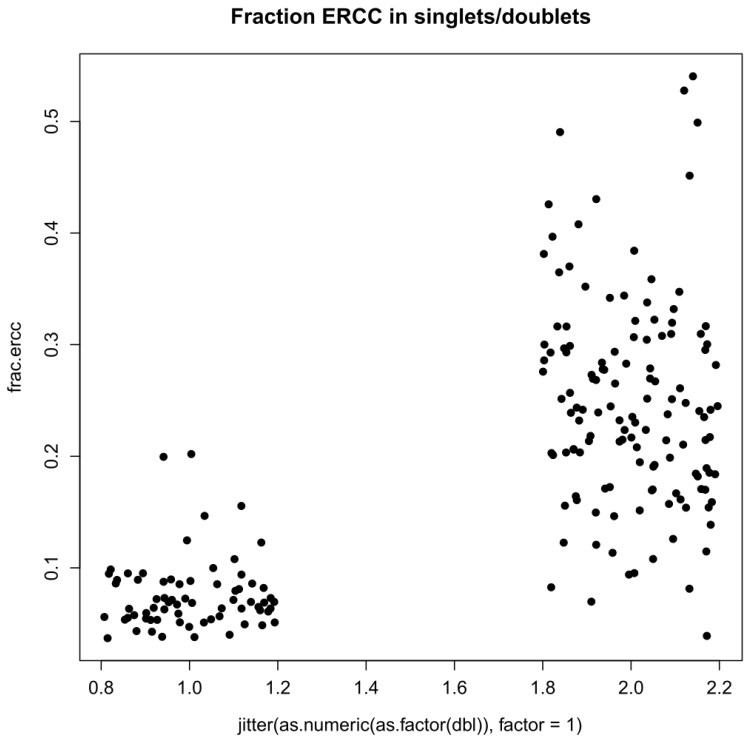

Figure 1. Fraction of reads from ERCC spike-in controls in multiplets (left) and singlets (right). A constant amount of non-human control RNA is spiked into each well when the plate is prepared. Multiplets have on average ~30% the number of ERCC reads, indicating that each contain on

average three cells. The fraction depends on various other things, such as the size of the cell and how much of the cell's RNA that was successfully recovered (can be a small fraction if the cell hit the wall of the plate, for example).

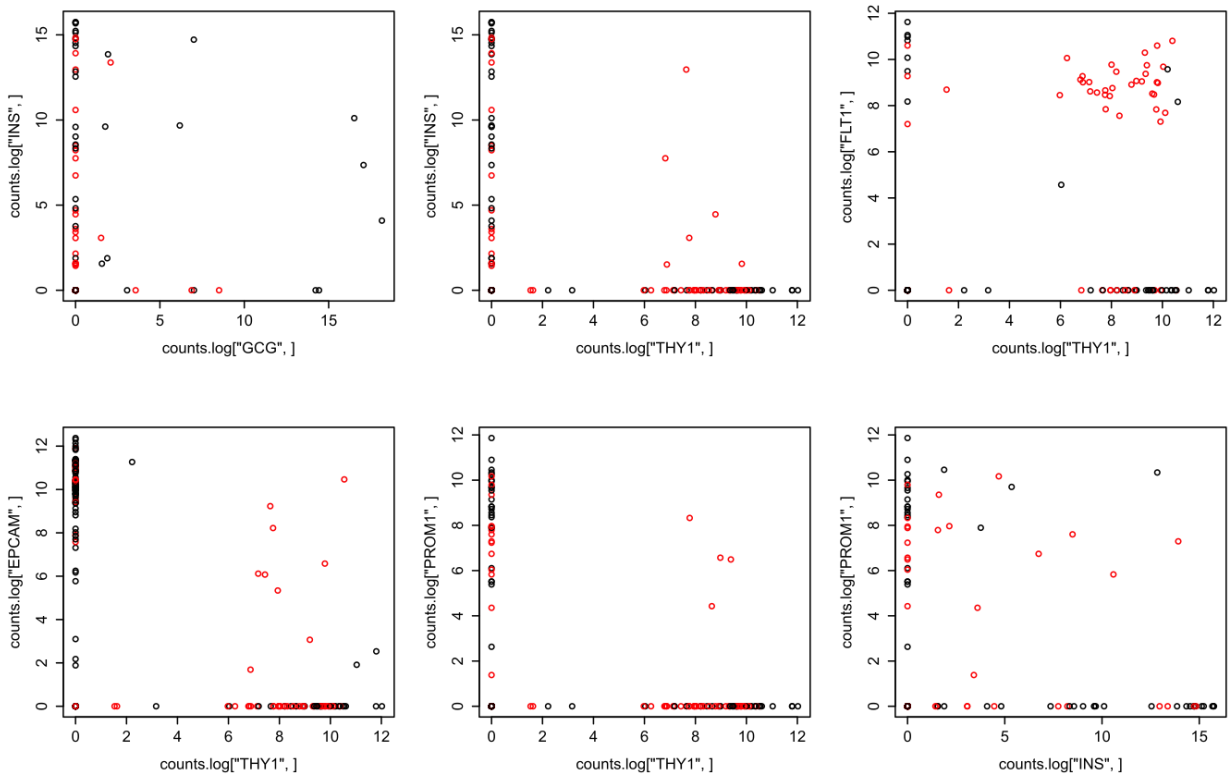

Figure 2. Co-expression of marker genes of various cell types in singlet (black dots) and multiplet (red dots) cells. INS: Beta cells (endocrine); GCG: Alpha cells (endocrine); THY1: Mesenchyme; FLT1 Endothelium; EPCAM: Epithelium, PROM1: Duct cells (endocrine progenitor). There seems to exist combinations of most major cell types. Alpha/Beta cell markers are some sort of negative control.

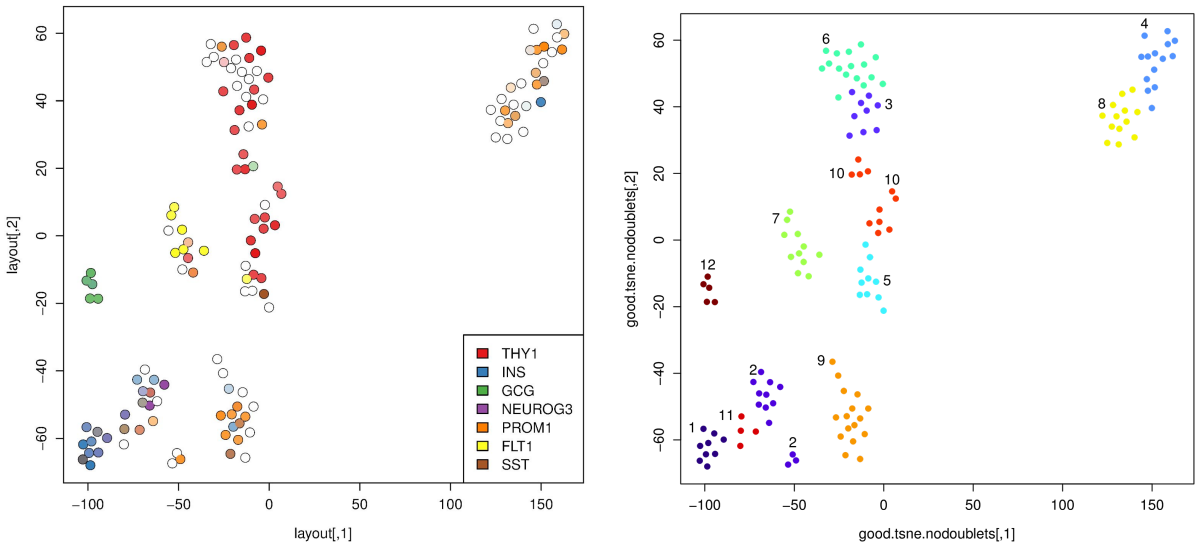

Figure 3. 2d tSNE projection of singlets based on pairwise pearson correlation between the transcriptional profile of the 2000 genes most highly expressed in a cell. Each cell is represented by a colored point. Left panel: color is by expression of marker genes. Right panel: points are colored by cluster identity. Clusters were determined using mclust, based on pariwise distances in tSNE (2d projected) space. Numbers indicate cluster identity (same indices as in Fig 4).

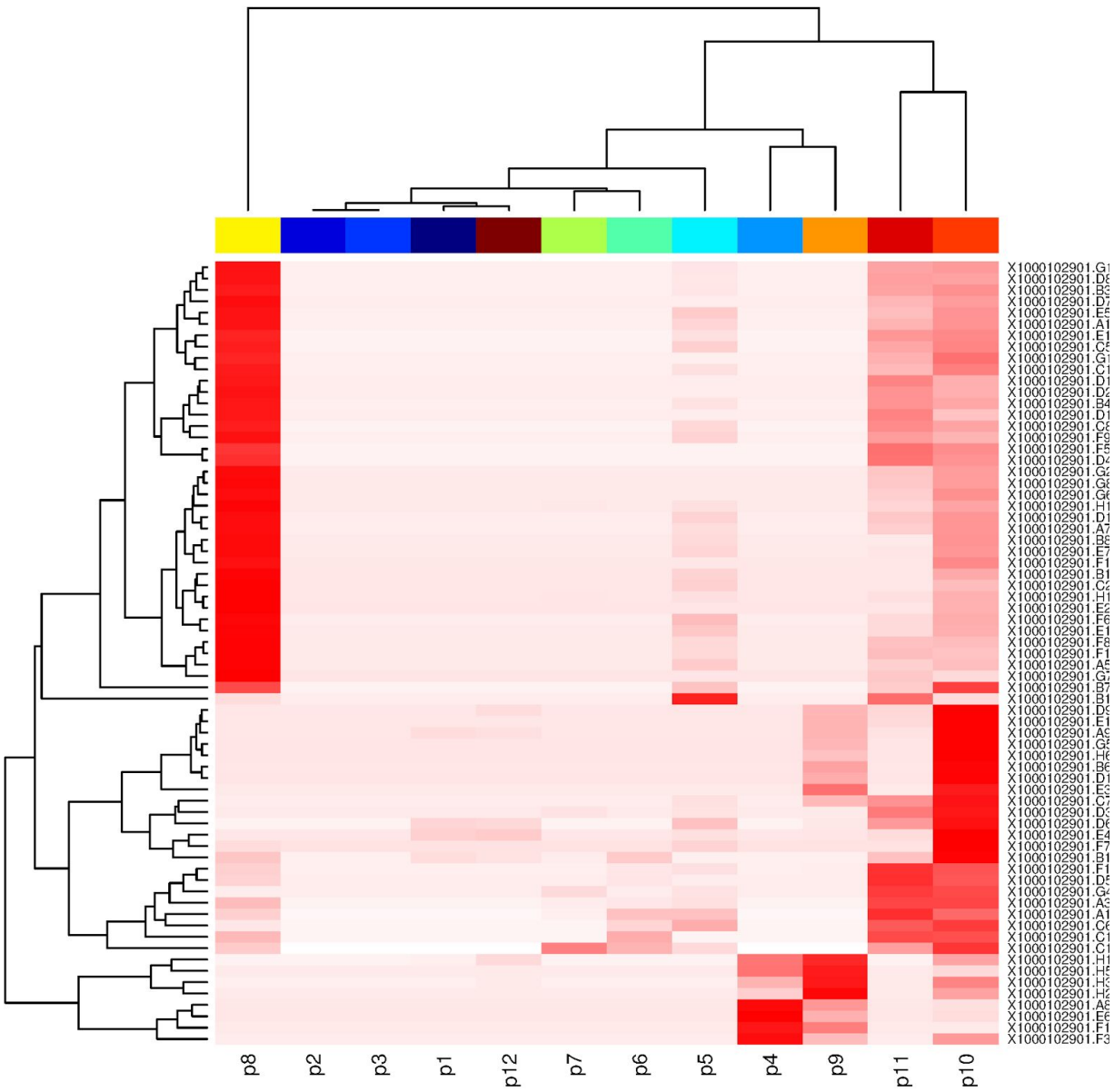

Figure 4. Test run! We used the bounds constraint BFGS method (“L-BFGS-B”) from the R package optimx, to find the linear combinations of cell types (clusters in Fig 3) that most closely approximates the multiplet gene expression profile. In most cases only a few major cell types are selected, which is reassuring. A large fraction of the multiplets contain ductal cells (p8, 4, 9). The other major groups are Mesenchymal cells (most notable p10, might be a sub-cell type in the cluster with p6 and/or p5). P11 represent cells that are differentiating into beta cells (INS-expressing cells). Might be that p5 is a mesenchymal subtype that is always found close to progenitor duct cells, not to differentiating endocrine.

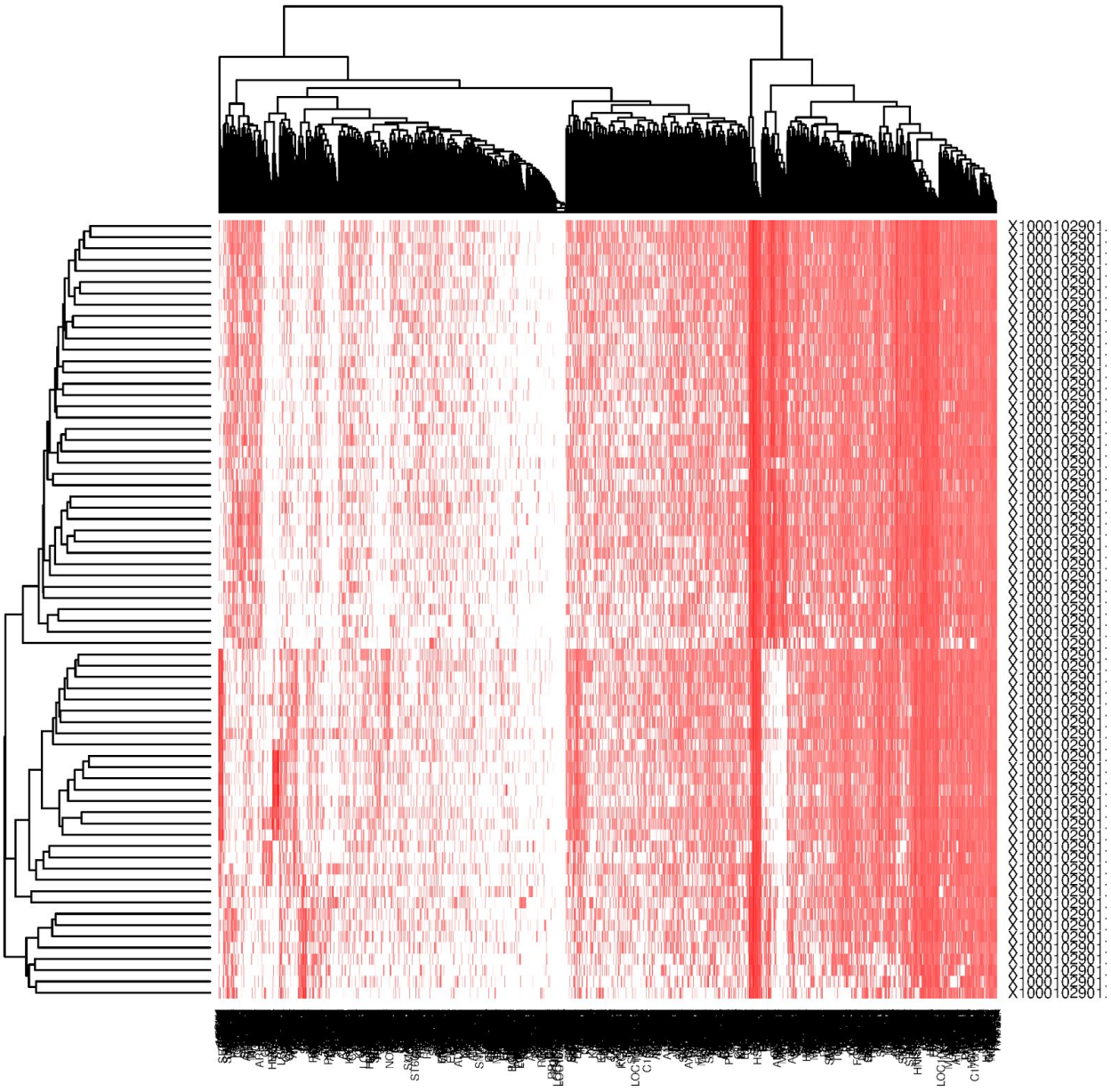

Figure 5. Comparison clustering using the top 1000 highly expressed genes. It is a mess, right?

Plans:

Methodologically, the project has two main parts - the experimental part and the computational part. The experimental part is very straightforward (although not obvious for most single-cell researchers, it seems), however the computational part can get pretty complicated.

There are basically two ways of doing the analysis.

1) Define cell types as averages across single cells that are similar to each other. Then determine the most likely combination of these cell types that would give rise to the

multiplet transcription profile. This is a robust approach since it sidesteps the rather high noise levels in single-cell data. I used an already existing parameter optimization method in Fig4, above, but one could use a more realistic version, by limiting the maximum number of cell types that one can fit for example (currently the model is always a positive weight for each cluster).

2) Directly fit a number of single-cell profiles to the multiplet profile. Eg. which 2-4 single cells out of the hundreds we profiled would best explain the multiplet profile? This is in a way a more interesting approach since it should be able to find more rare combinations, but it seems like it suffers from overfitting (only did some very preliminary tests).

There might be a good intermediate way between these two approaches. Either call many different clusters, classified into “superclusters” (which would be analogous to sub-cell types and cell types, respectively) and use 1), or use something similar to 2), but on averages of a small number of cells (in a way, call clusters in the same process as fitting multiplet profiles). Maybe best to go with the miniclusters idea first.

Either way, the end result would be sets of physically interacting sub-cell types. The hope would be to find different sub-cell types that tend to interact in different patterns.

In terms of writing a compelling paper, it will be very important to find some interesting result that can be verified in situ. For example, one could imagine a mesenchymal sub-cell type that is associated with endothelial cells and that can be distinguished by some mRNA transcripts.

Random extra text:
Cell state is dictated by internal state and external signalling. A

are dictated by direct interactions between cells

For many problems, the most important In other methods, the spatial attributes are larger than a single cell, precluding