Learning cell identities and (post-)transcriptional regulation using single-cell data

Michielsen, L.C.M.

SUMMARY

Tissues in the human body, and especially the brain, are heterogeneous and consist of many different cell types. Cell types can be defined by the genes expressed in a cell, and these expressions are controlled by unique cell-type-specific (post-)transcriptional mechanisms. Diseases can perturb these control mechanisms, and thus affect cell types differently. Consequently, understanding which cell type is affected by a disease is crucial information when developing new drugs or treatments. Single nucleotide polymorphisms (SNPs) in the DNA can be associated with diseases, but approximately 95% of such SNPs fall in the non-coding region of the DNA. Usually, it is unknown whether these variants are causal, and which gene and cell type they affect. Studying gene expression at the single-cell level could reveal such disrupted mechanisms.

Current advances in single-cell RNA sequencing have greatly improved our understanding of heterogeneous tissues and led to the discovery of many new cell types. However, this new technology also presents computational challenges. For example, when comparing datasets from different cohorts (e.g., across many different individuals) it is important to annotate cells consistently. To ensure such consistency, it is essential to annotate cells using classification methods instead of currently practiced clustering methods that are subjective and time-consuming. To facilitate this transition, in this thesis, we benchmarked cell-type classification methods and developed computational methods to automatically build reference atlases using multiple already labeled single-cell datasets. We show how such reference atlases can be deployed to automatically annotate new (unlabeled) single-cell datasets, as well as how they can be updated continuously using new labeled single-cell datasets.

Having established a more consistent cell-type annotation across single-cell datasets, we return to establishing a relationship between mutations and their effect on gene expression. Hereto, we study sequence-to-expression models that can predict an alteration in expression when a mutation is observed. Given that gene expression mechanisms are cell-type specific, we introduce sequence-to-expression models based on single-cell data to make cell-type-specific predictions. We use these models to show that certain mutations are indeed changing gene expression, increasing our understanding of transcriptional regulation.

Next to differences in gene expression between cell types, cell types might express different isoforms of a gene (i.e., different combinations of exons included in an mRNA molecule). Again, this can be altered by mutations in the DNA. Advances in single-cell long-read sequencing enabled measuring which cell types express which isoforms. We leveraged this data and propose a novel approach in which we adapted our sequence-to-expression models to predict cell-type-specific isoform usage. This opens a new avenue for looking at cell-type-specific alterations.

Taken together, we introduce a variety of computational methods to enhance single-cell RNA sequencing data to improve our understanding of cellular heterogeneity.