Explore Single-cell sequencing dataset

Contents

1 Abstract 1

2 Background 1

2.1 Initial motivation of project 1

2.2 Similarity between single-cell dataset and typical computer science 1

2.3 Domain-specific knowledge 2

3 Data Source 2

4 Expected results and planned methods 2

4.1 Quality control 2

4.2 Using clustering to identify subpopulations of cells 2

4.3 Using LDA model to infer lineage of cell types 2

4.4 Predict developmental status of unseen new cell 2

1 Abstract

In cancer the heterogeneity of bulk cells widely exists and is one of the main obstacles to achieve either targeted cancer treatment or personalized medical care. To address the problem, single-cell sequencing technology emerges and provides the entire gene expression profile at individule cell level, which reaches highest resolution of investigating expression of cells. Here we utilize unsupervised machine learning to identify putative and established subpopulation and possible new marker genes for specific cell types. As the data itself is essentially a matrix with samples (i.e. cells) as columns and features (i.e. genes) as rows, the approaches in this project could be extended to other general classfication problem.

2 Background

2.1 Initial motivation of project

There exists emerging trend that big companies engaging in either internet or data are heading to healthcare areas, for example, Verily Life Sciences of Alphabet Inc. which is the umbrella company of Google, IBM Watson Health, etc., thus we are looking for some dataset which is not only dedicated to life sciences researches, but also has enough samples and features.

2.2 Similarity between single-cell dataset and typical computer science

Single-cell sequencing technique gives expression levels of entire genes at individual cell level, which means a big dataset composing of hundreds of cells (i.e. samples) with thousands of genes (i.e. features).

In typical computer science dataset, for example, developing recommendation systems at Amazon, customer segmentation at Bank of America, etc. They are normally starting from a data matrix with rows as customer ID and columns as descriptive features, e.g. name of watched movies watched, purchased products, annual salary, etc.

2.3 Domain-specific knowledge

Cell types: Different types of cells have different biological functions, which is the result of different gene expression profiles.

Marker genes: Genes that are exclusively expressed at specific type of cells.

3 Data Source

HSMMSingleCell: primary human skeletal muscle myoblasts (HSMM) were expanded under high mitogen conditions (GM) and then differentiated by switching to low-mitogen media (DM). The switch can be manually triggered by adding serum when cell culturing. 49-77 cells were captured at each of four time points (0, 24, 48, 72 hours) following serum switch.

It is a dataset with around 200-300 cells in total having around 50K genes. Each cell has a time point label: 0, 24, 48, 72.

4 Expected results and planned methods

4.1 Quality control

4.1.1 How many genes are detected per cell

4.1.2 Average expression level per gene

4.2 Using clustering to identify subpopulations of cells

4.2.1 Infer possible number of existed subpopulations

4.2.2 Hierarchical clustering

4.2.3 Dimension reduction via PCA, t-SNE for visualization

4.3 Using LDA model to infer lineage of cell types

4.4 Predict developmental status of unseen new cell