Single Cell Data Analysis

Carissimo Annamaria

Single-Cell RNA Sequencing and Data Analysis – ELIXIR IIB TRAINING PLATFORM

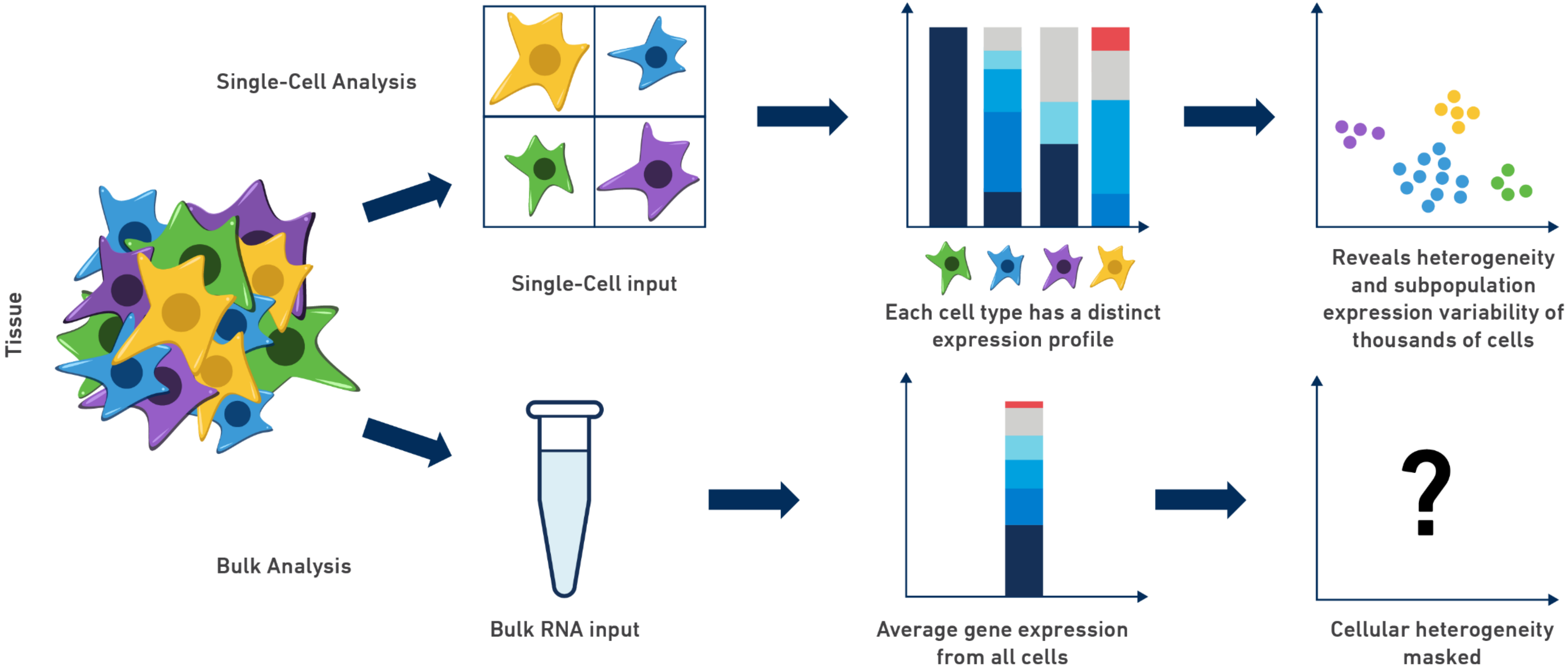
Tigem, Pozzuoli (Napoli), Italy | May 7-9, 2019 |

Outline

- Introduction
- Data Normalization
- Imputation
- Differential expression

Introduction

Bulk RNA-Seq vs ScRNA-seq


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

Introduction

Bulk RNA-seq

- Measures the **average expression level** for each gene across a large population of input cells
- Useful for comparative transcriptomics, e.g. samples of the same tissue from different species
- Useful for quantifying expression signatures from ensembles, e.g. in disease studies
- **Insufficient** for studying heterogeneous systems

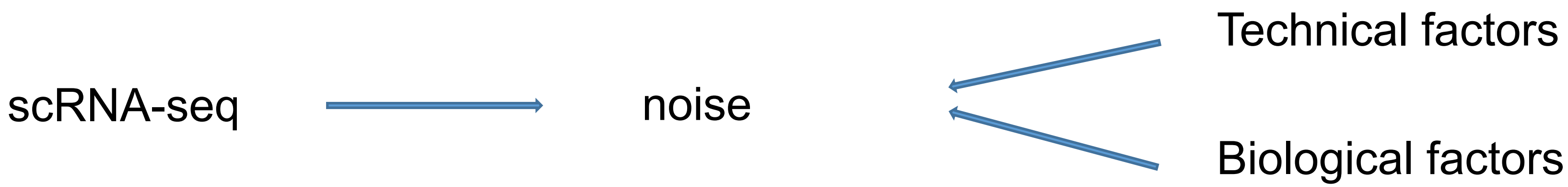
Introduction

scRNA-Seq

- Measures the **distribution of expression levels** for each gene across a population of cells
- Better understand the dynamics of gene expression pattern
- Allows to study new biological questions in which **cell-specific changes in transcriptome are important**, e.g. cell type identification, heterogeneity of cell responses
- Reveal heterogeneity within population of cells
- High dimensionality: thousand of cells

Introduction

Bulk RNA-Seq vs ScRNA-seq



scRNA-seq → noise

Technical factors

Biological factors

Technical variability

low amount of mRNAs
amplification
dropouts events

Biological variability

stochastic nature of transcription

Introduction

scRNA-Seq data preprocessing

- Obtain RNA-Seq expression data
- Filter Cells - low quality cells
- Filter Fatures – lowly expressed genes
- Normalization
- Imputation

Normalization

Adjust for unwanted biological and technical effects that can mask the signal of interest

Several experimental sources of systematic biases due to:

- Sequencing depth
- Amplification
- Gene Length
- GC-content
- mRNA content

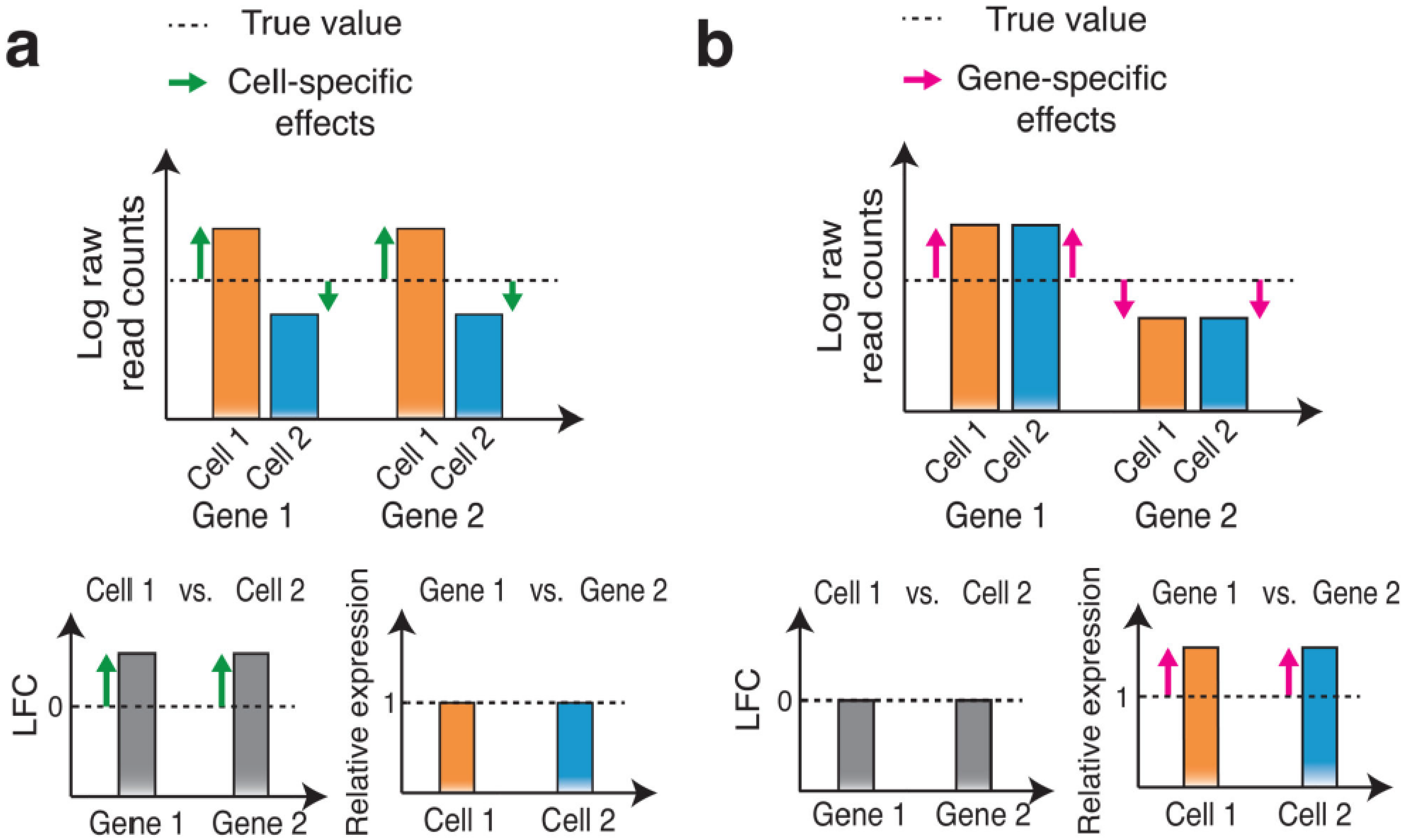
UMI – based protocol remove amplification biases

Normalization

Cell/Gene specific effects

Systematic biases affect mesasurements of gene expression

Cell specific effects introduce bias in the LFC on raw read counts



Gene specific effects introduce bias in the LFC on raw read counts

Vallejos et al. *Nat Methods.* 2017 June

Normalization

Two types of normalization:

- *Within –sample normalization* which remove gene specific biases (GC content)

- *Between – sample normalization* which adjusts for effects related to distributional differences in read counts between cells (sequencing depth)

Normalization

Scaling Factor

Global-scaling factor normalization methods are inherited from bulk RNA-Seq data analysis

Motivation: bring cell –specific measure on to a common scale by standardizing a quantity of interest across cells by assuming that most genes are not differentially expressed

Methods:

- *Counts per Million* **CPM**
- *Upper quantile* **UQ**
- *Full quantile* ***FQ***
- Trimmed Mean of M-values **TMM**
- *DESEq normalization*

Normalization

Methods

CPM: Counts scaled by the number of reads ***N*** (total number of reads o library size) times one million

$$CPM(C) = \frac{c_{g.j}}{N} * 10^6 \quad g = 1, \dots, p \;\; j = 1, \dots . m \;\; N = \sum_{g=1}^{p} c_{g,.}$$

Standardizes the total number of reads between cells – library size normalization

UQ : scaling factor is proportional to the 75th percentile of the distribution of counts within each cell

FQ : all quantile of cell-specific cell are matched to a reference distribution

Quantile-based normalization methods are problematic in scRNA-Seq
due to the high frequency of zero counts

Normalization

Methods

TMM normalization: trims away extreme log-fold-changes to normalize the counts based on the remaining set on non differentially expressed genes

Procedure:

Step 1: double trimming based on log-fold changes ***M*** and absolute intensity ***A***
Step 2: the normalization factor is calculated using a reference sample

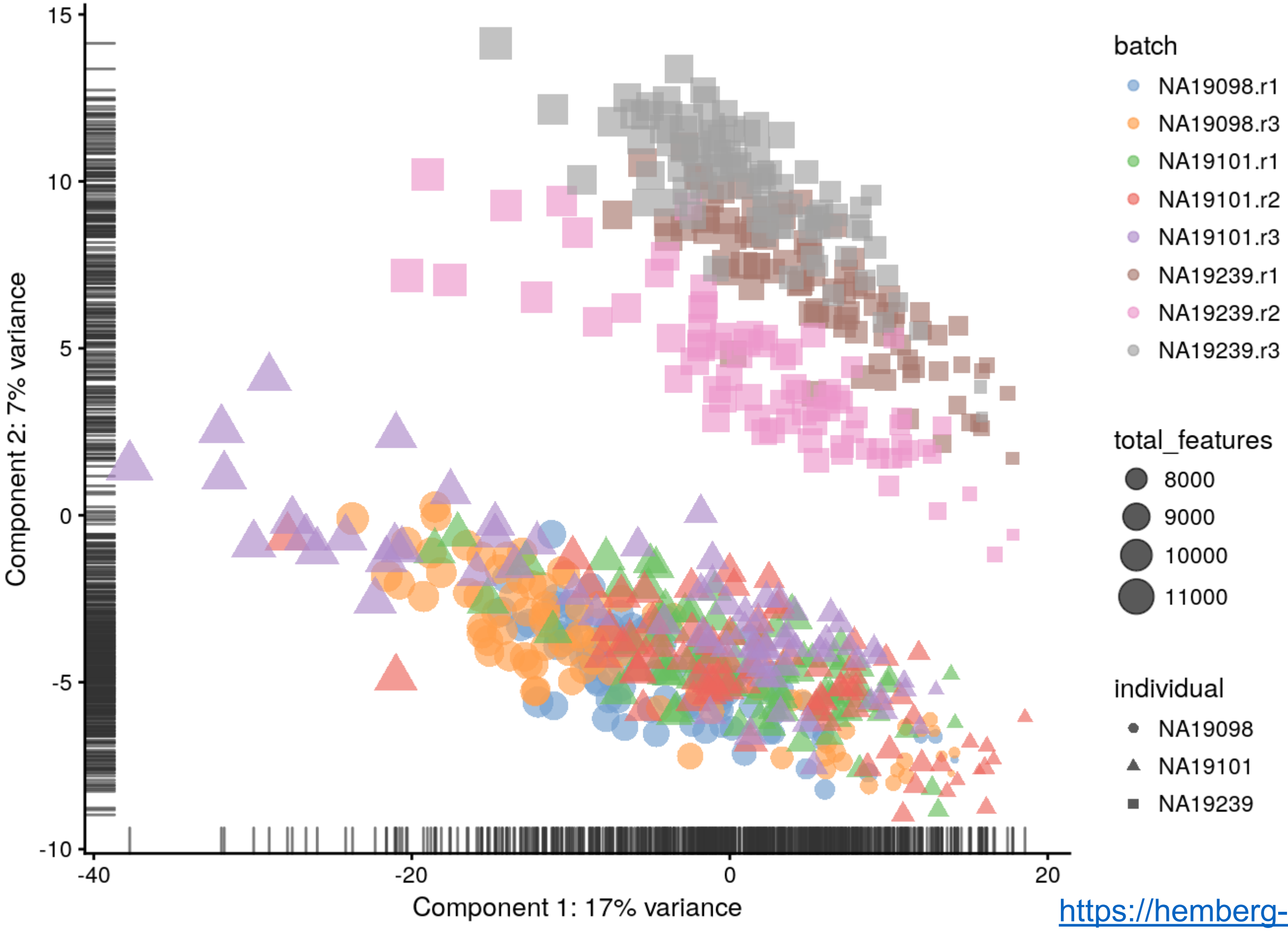
DESeq normalization: defines scaling factor estimates based on a pseudo-reference sample based on a geometric mean

Normalization

Examples

RAW data

PCA

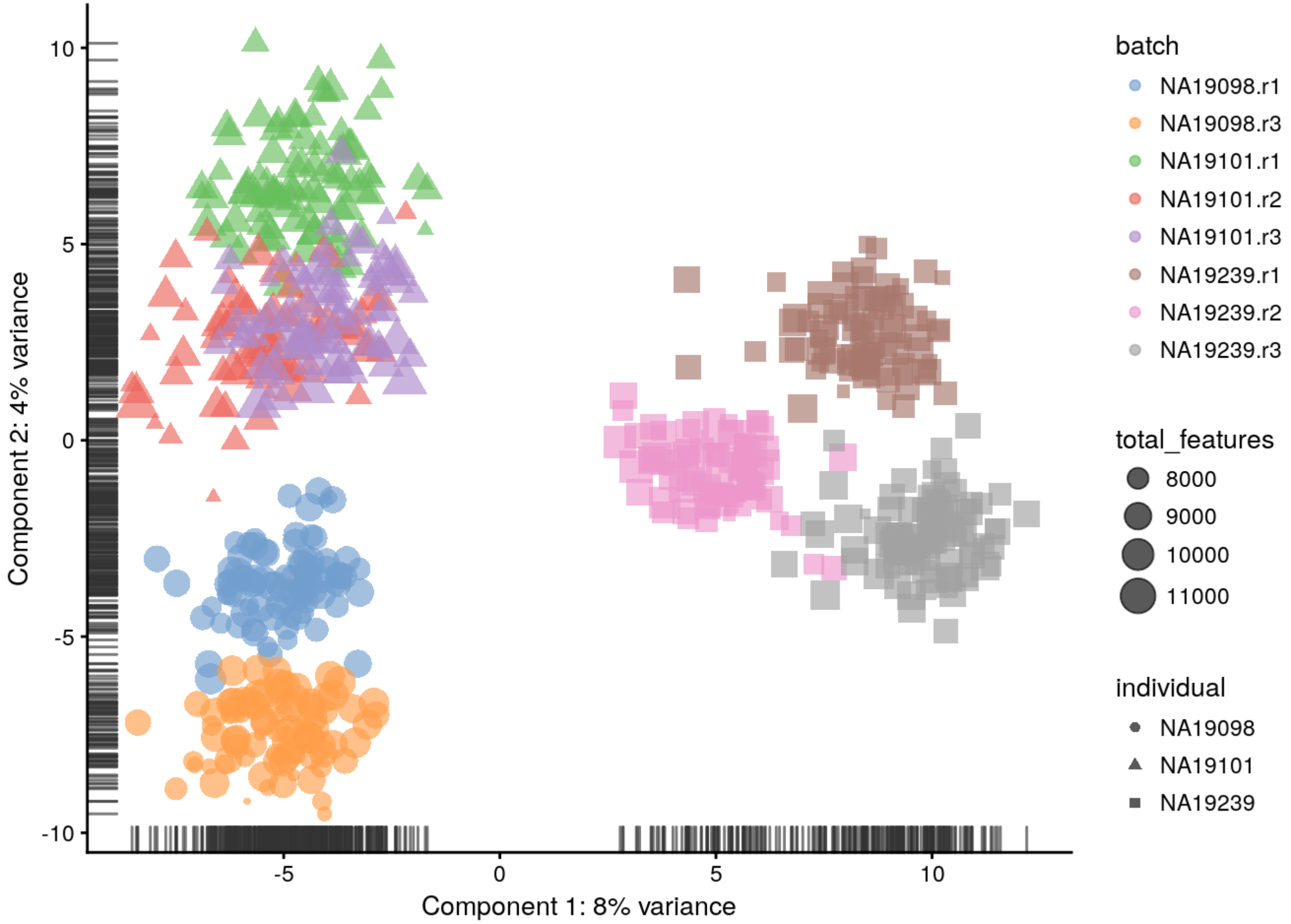


https://hemberg-lab.github.io/scRNA.seq.course/

Normalization

Examples

CPM data **PCA**



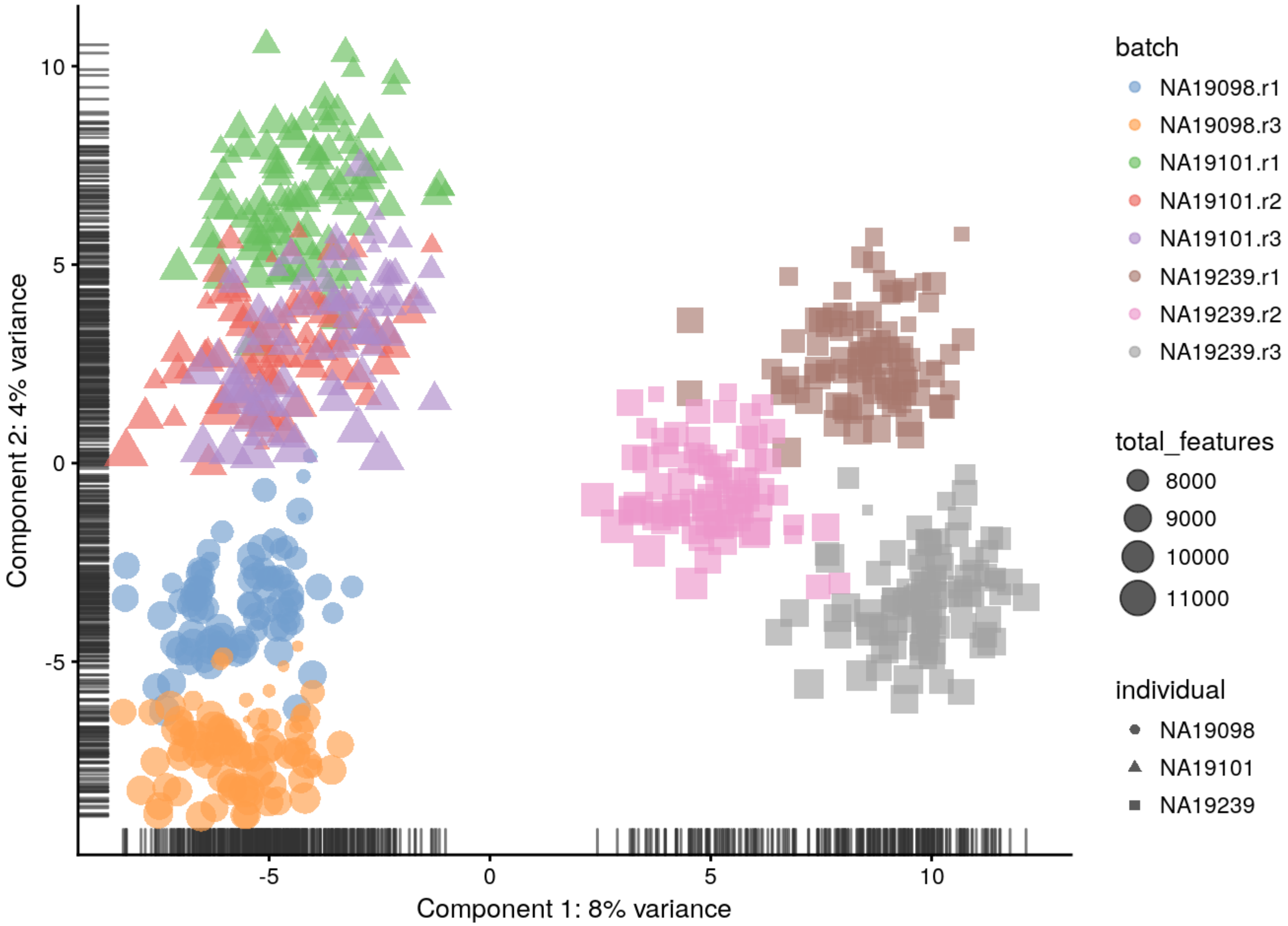
https://hemberg-lab.github.io/scRNA.seq.course/

Normalization

Examples

TMM normalized data

PCA



https://hemberg-lab.github.io/scRNA.seq.course/

Imputation

One of the main challenges when analyzing scRNA-seq data is the presence of **zeros**.

- The gene is not expressed in the cell and hence there are no transcripts to sequence
- The gene is expressed, but for some reason the transcripts is lost somewhere prior to sequencing
- The gene is expressed and transcripts is captured, but the sequencing depth is not sufficient to produce any reads.

Dropouts events

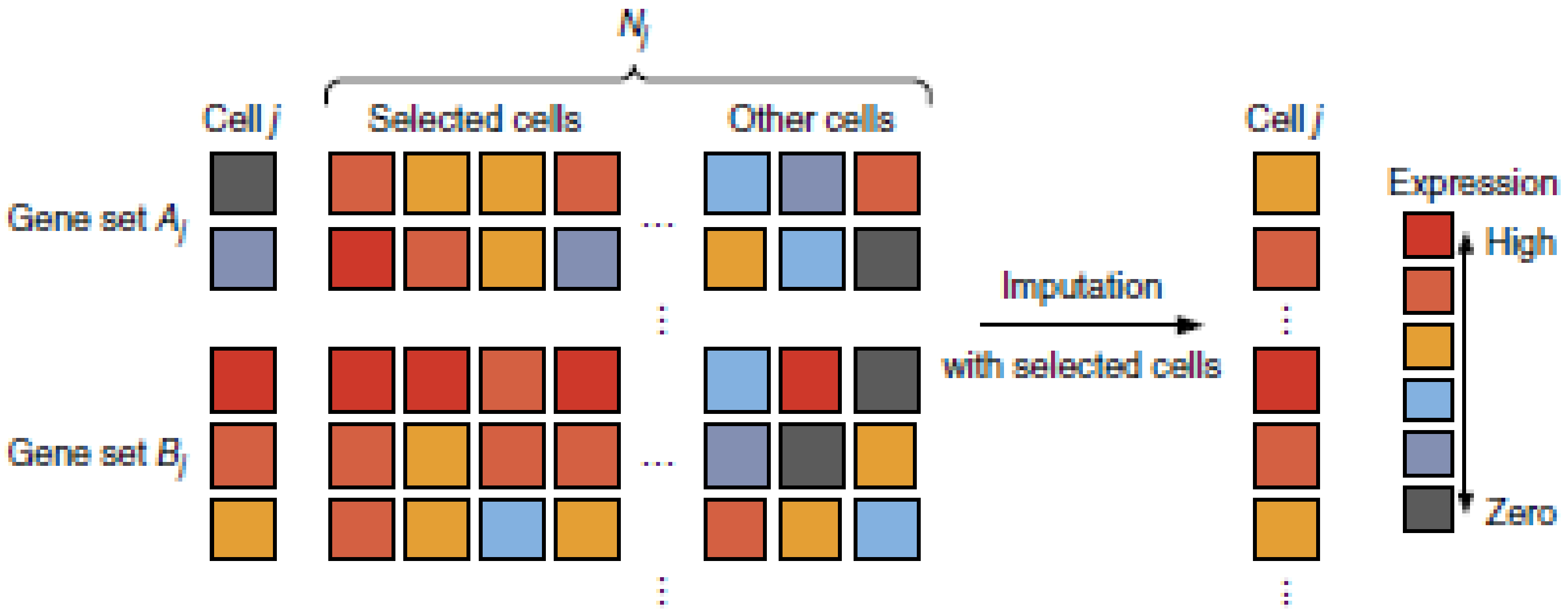
a gene is observed at a moderate expression level in one cell but undetected in another cell

Imputation

Methods

scImpute: a statistical method to accurately and robustly impute the dropouts in scRNA-Seq data.

scImpute first learns each gene's dropout probability in each cell based on a mixture model. Next, scImpute imputes the (highly probable) dropout values in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes unlikely affected by dropout events



Li et al. Nature Communication 2018

Imputation

Methods

scImpute can be applied before:

- dimension reduction of scRNA-seq data
- normalization of scRNA-seq data
- clustering of cell populations
- differential gene expression analysis
- time-series analysis of gene expression dynamics

https://github.com/Vivianstats/scImpute

Imputation

Methods

Markov Affinity-based Graph Imputation of Cells (MAGIC)

- imputes missing expression values by sharing information across similar cells, based on the idea of heat diffusion.

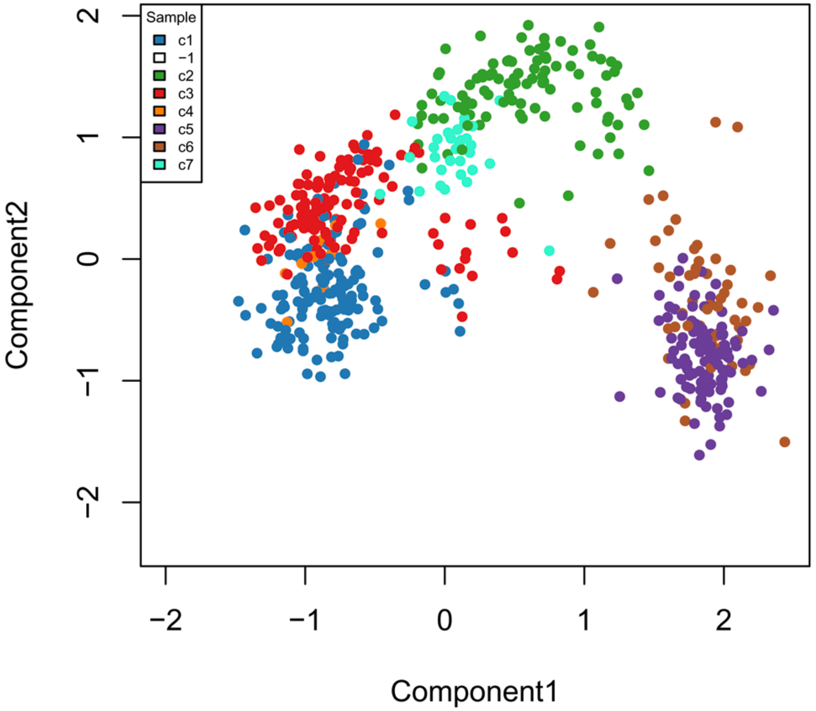
- create a Markov transition matrix, constructed by normalizing the similarity matrix of single cells.

- In the imputation of a single cell, the weights of the other cells are determined by the transition matrix.

https://github.com/KrishnaswamyLab/MAGIC

Differential expression

AIM Find genes that vary between cell type and state or in response to a perturbation



In single cell data, once the groups of cell have been identified one can find differentially expressed genes either by comparing gene expression between clusters in a pairwise manner.

Differential expression

What is a Hypothesis of a test?

- A hypothesis is an assumption about the population parameter.
  - A parameter is a characteristic of the population, like its mean or variance.
  - The parameter must be identified before analysis.

Differential expression

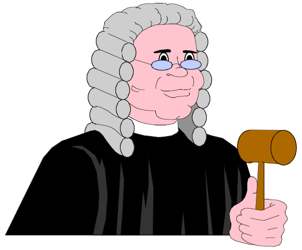
The Null Hypothesis, H_0

- States the Assumption to be tested

e.g. Our class mean age is 50 (H_0: $\mu=50$)

- Begin with the assumption that the null hypothesis is TRUE.

(Similar to the notion of innocent until proven guilty)



The Null Hypothesis may or may not be rejected, but our aim is to REJECT the null hypothesis!

Differential expression

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis

  e.g. The average age of our class is different from 50 (H_1: $\mu \neq 50$)

- Is generally the hypothesis that is believed to be true by the researcher!

Differential expression

We have to simultaneously test, for each gene, the null hypothesis: gene expression has not changed

For each gene j the test is expressed in term of a Statistic and a p-value

Null Hypothesis
Ho: $\mu j(WT)=\mu j(KO)$

Multiple testing corretion problem

Differential expression

Single cell data characteristics:

- Low library size
- High noise level
- Large fraction of dropouts events

It is not clear whether DE method tha have been developed for bulk RNA-Seq are suitable also for scRNA-seq

Differential expression

Methods for bulk RNA-Seq

- edgeR
- DESeq – DESeq2

Methods for scRNA-Seq

- MAST
- SCDE
- Monocle
- D3E
- SeuratBimod
- scDD

Non parametric test

- Wilcoxon test

Many others

Differential expression

Models

- Perhaps the simplest statistical model for count data is the Poisson, which has only one parameter.
- Under a Poisson model, the variance of the expression for a particular gene is equal to its mean expression.
- However, due to a variety of types of noise (both biological and technical), a better fit for read count data is usually obtained by using a *negative binomial* model, for which the variance can be written as:

  variance = mean + overdispersion x mean^2
- Since the overdispersion is a positive number, the variance under the negative binomial model is always higher than for the Poisson.

Differential expression

EdgeR

EdgeR

- Read counts are modeled by a negative binomial distribution
  For each gene, the variance is related to the mean by σ^2=μ + $\alpha\mu^2$ where α is the over-dispersion parameter
- The method estimates the gene –wise dispersions using a conditional maximum likelihood procedure
- An empirical Bayes procedure is used to shrink the dispersions towards a consensus value
- The *glmFit* function fit the data and the *glmLRT* compares the two conditions
- TMM normalization procedure is carried out to account for the different sequencing depths between the samples

Differential expression

DESeq

DESeq

- Read counts are modeled by a negative binomial distribution
- The variance of negative binomial distribution σ^2 is modeled as

$$\sigma^2 = \mu + s^2 v$$

where s is the size factor e v is the true concentration of reads

- a scaling factor normalization procedure is carried out to account for the varying sequencing depths of the different samples

Differential expression

MAST

MAST (Model-based Analysis of Single-cell Transcriptomics)

- models single-cell gene expression using a two part generalized linear model
- introduces the CDR (Cellular Detection Rate) as the fraction of genes expressed in a single cell
- CDR is modeled as a covariate
- Baysian framework to regularized model parameters

Take into account dropouts and bimodal expression distribution in which expression is either strongly non zeros or not detectable

Differential expression

SCDE

SCDE uses a bayesian approach to single cell differential expression analysis

- SCDE models the read counts computed for each gene using a mixture of Negative Binomial NB distribution and a Poisson distribution
- NB distribution models the transcripts that are amplified and detected
- The low-magnitude Poisson distribution model the unobserved or background-level signal transcripts that are not amplified (dropouts events)
- A Bayesian approach is used for differential expression

Differential expression

Monocle

Monocle is a tool designed for single-cell RNA-Seq for ordering cells by process through differentiation stage

- Identifies genes that change significantly over the time
- Identifies genes that are differentially expressed across different cell type or conditions
- It uses a Generalized Linear Additive Model (GAM)

Differential expression

Discussion

- Prefiltering of genes is essential for obtaining a good and robust performance for several methods
- EdgeR tend to call lowly expressed genes with many zeros significant if they are present in the data, but otherwise performs well
- Methods developed for bulk RNA-Seq analysis doesn’t perform worse than those specifically developed for scRNA-Seq data, but sometimes show a stronger dependence on data prefiltering
- In particular, EdgeR and MAST have good performances

Soneson et al, Nature Methods 2018

Differential expression

Example

ARTICLE

Received 25 May 2016 | Accepted 1 Dec 2016 | Published 19 Jan 2017

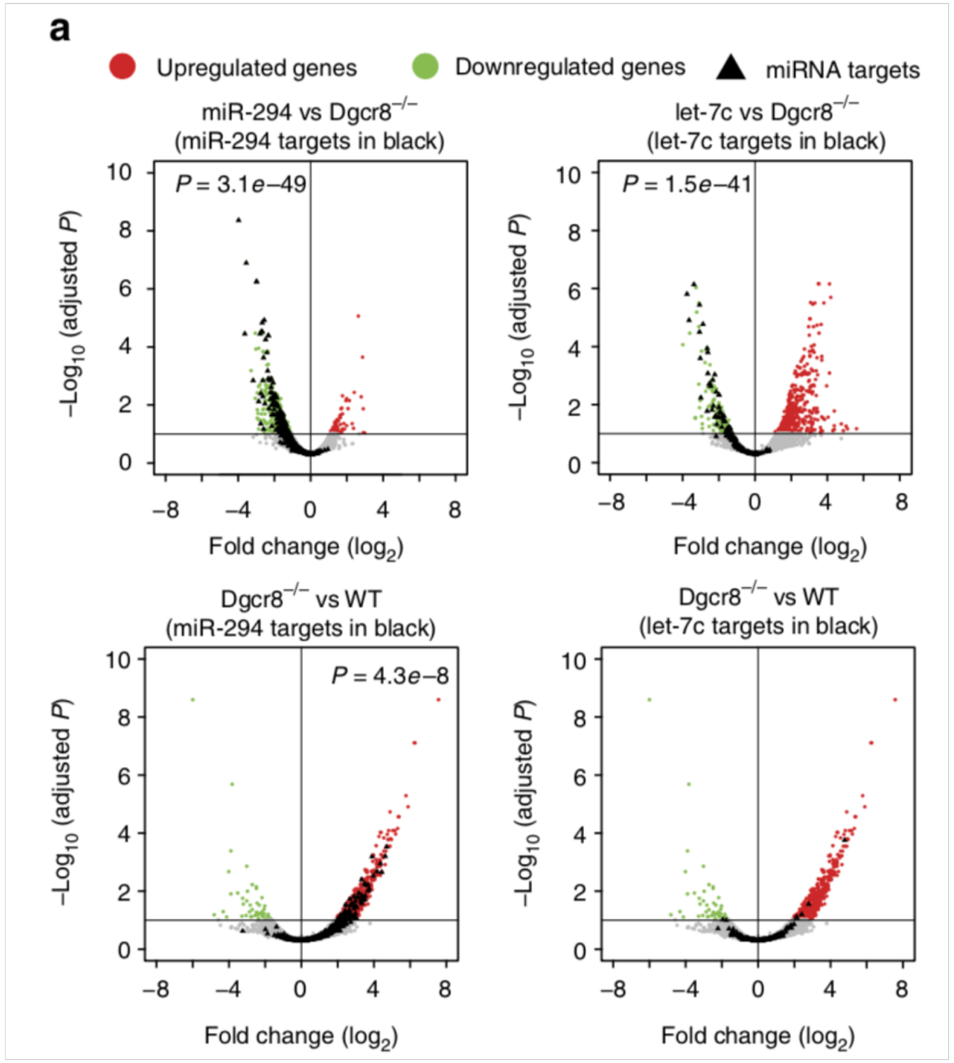
OPEN

The impact of microRNAs on transcriptional heterogeneity and gene co-expression across single embryonic stem cells

Gennaro Gambardella[1,*], Annamaria Carissimo[1,*], Amy Chen[2,3], Luisa Cutillo[1], Tomasz J. Nowakowski[2], Diego di Bernardo[1,4] & Robert Blelloch[2,3]

Differential expression

SCDE method



Conclusions

- Single-cell analysis is an exciting and rapidly expanding field
- Single cell improves our understanding of fundamental biological problems and helps us to better understand the nature and complexity of human disease in order to develop more effective therapies.
- Single-cell data present a number of intrinsic challenges, including systematic noise, the features of biological systems, and the sparsity and complexity of the data.
- Invest in development of new methods.