About Single-cell RNA-Seq

Dr. Giorgio Giurato, PhD
ggiurato@unisa.it

Laboratory of Molecular Medicine and Genomics
Genomix4Life Srl
Department of Medicine, Surgery and Dentistry 'Schola Medica Salernitana'
University of Salerno

Training Course on Best practise for RNA-Seq data analysis

Sep 27-29, 2017 – Salerno

Bulk RNA-Seq *vs* Single-cell RNA-Seq

- Conventional “bulk” methods of RNA-Sequencing (RNA-Seq) process hundreds of thousands of cells at a time and average out the differences.

- In many context such bulk expression profiles are sufficient.

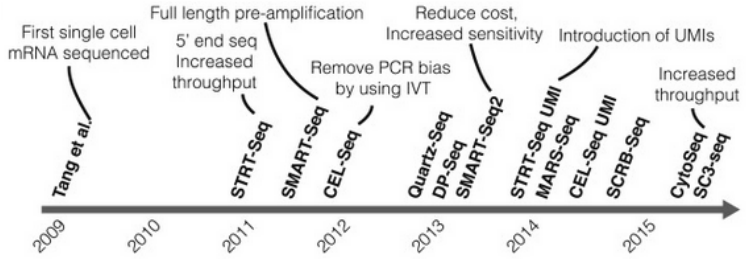
- For example in comparative transcriptomics, where the goal is to study the selection pressures that apply to gene expression levels between samples of the same tissue taken from different species.

- However, bulk RNA-Seq can mask or even misrepresent signals of interest.

Bulk RNA-Seq

Single-cell RNA-Seq

Single-cell RNA-Seq



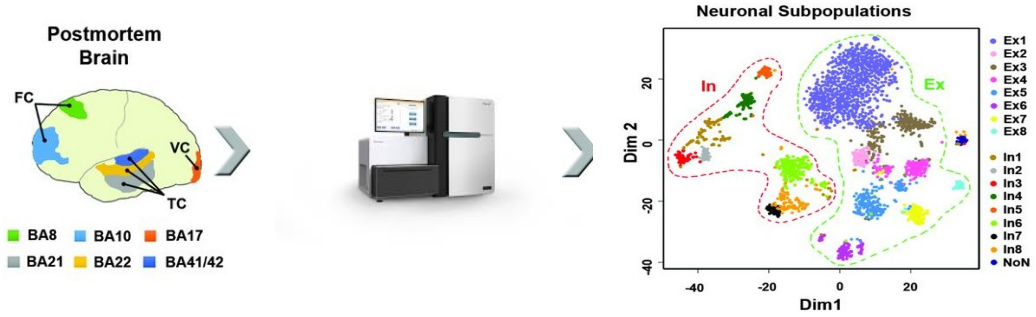
Kolodziejczyk et al., 2015, Molecular Cell

- Single-cell RNA-Seq (scRNA-Seq) has emerged as a revolutionary tool that allow to address scientific questions.

- There are important biological questions for which “bulk” measures of gene expression are insufficient.

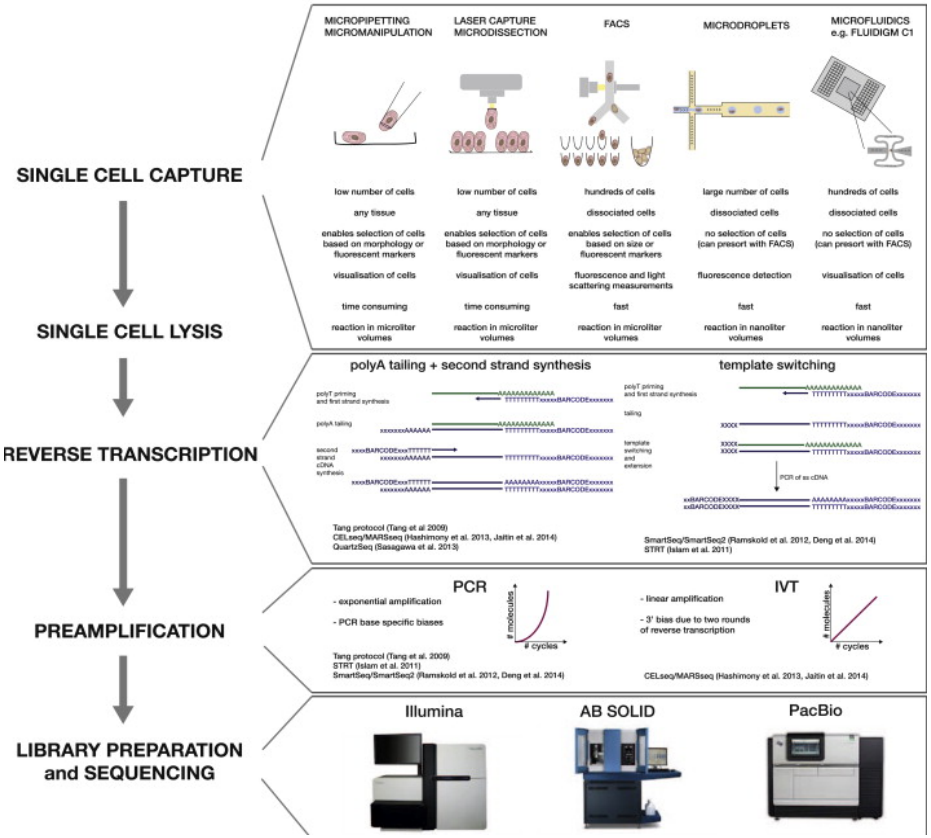
- During early development there are only a small number of cells, each of which can have a distint function and role.

- Complex tissues, such as brain tissues are composed of many distinct cell types.



Lake et al., 2016, Science

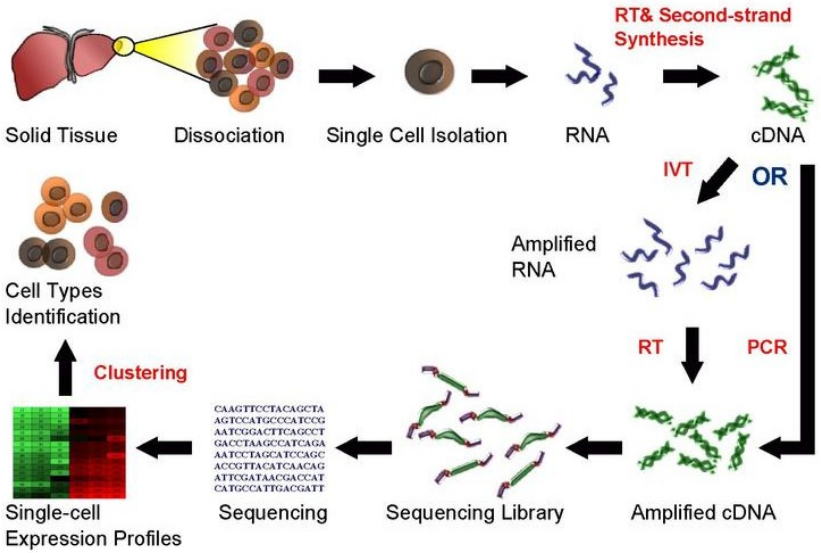
scRNA-Seq Workflow



Kolodziejczyk et al., 2016, Molecular Cell

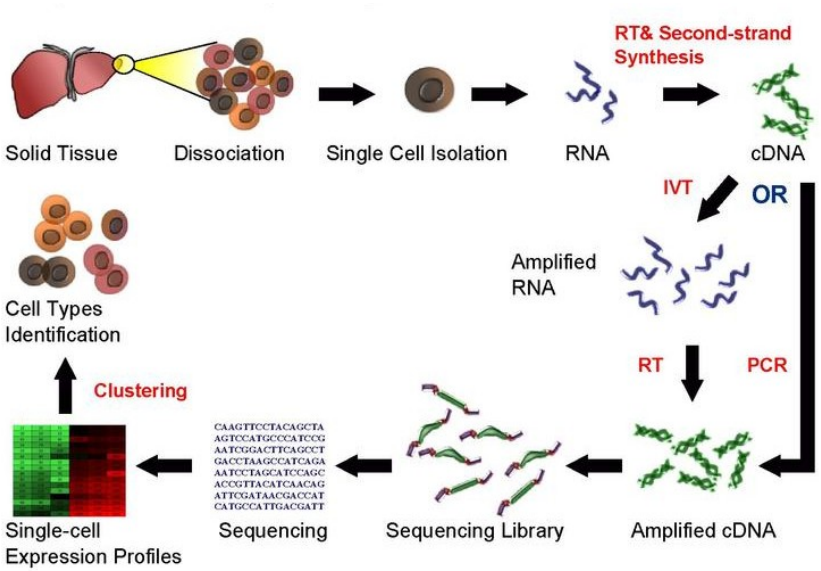
- In the current scRNA-Seq protocols RNA needs to be converted to cDNA for sequencing.
- The current scRNA-Seq method contains the following steps:
  - ❑ Isolation of single-cell and RNA
  - ❑ Reverse transcription
  - ❑ Amplification
  - ❑ Library generation and sequencing

scRNA-Seq Workflow



- A variety of noise and bias may be introduced in several steps of scRNA-Seq protocol.
- A critical step is represented by reverse transcription (RT) as the efficiency of the RT reaction determine the percentage of a cell's RNA population that is analyzed by the sequencer.
- In the amplification step, either PCR or In Vitro Transcription (IVT) is used to amplify cDNA.
- PCR efficiency on particular sequences will be exponentially amplified.
- On the other hand, libraries generated by IVT can produce sequences that may be transcribed inefficiently, generating imcomplete sequences.

Isolation of Single Cells



- There is currently no standardized technique for single-cell isolation.

- High-throughput methods for single-cell isolation include Fluorescence – activated cell sorting (FACS) and microfluidics.

- Both methods are accurate, automatic and capable isolating unbiased samples.

Experimental Design

- A well designed experiment is one that is:
  - ❑ Sufficiently powered.
  - ❑ Technical artifacts and biological features that may affect measurements are randomized, balanced.

- Incorporation of standards to facilitate quantitative comparisons of the expression level of each gene between cells.

- The use of spike-in molecules, taken from a different species from the cells of interest, is strongly recommended.

- The use of Unique Molecular Identifiers (UMIs) to barcode individual molecules.

- UMIs are tens of thousands of short DNA sequences (6-10 nucleotides in length), which are incorporated in molecules of interest before amplification, thus allowing biases to the accounted for.

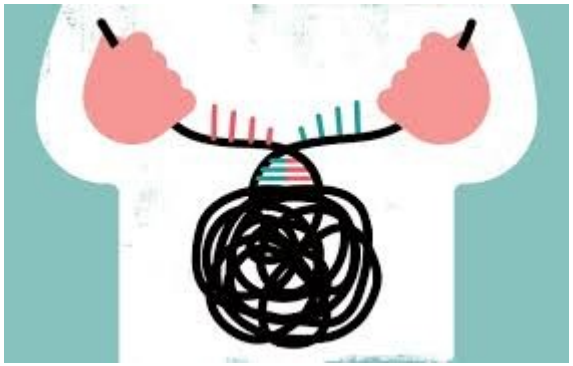
- In conjuction with spike-ins it is possible to estimate better the number of transcribed molecules, that is independent of amplification biases, which are a major source of techical variability.

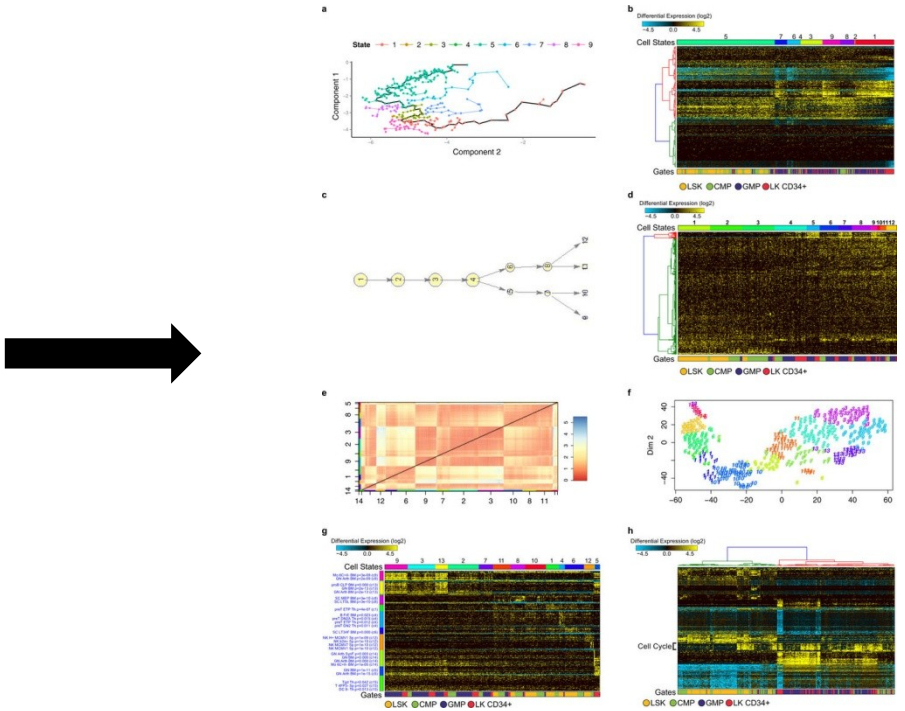
scRNA-Seq data analysis

- To ensure that scRNA-Seq data are fully exploited and interpreted correctly, it is crucial to apply appropriate computational and statistical methods.

- Given the widespread use of bulk RNA-Seq, many powerful tools for processing high-throughput transcriptome data already exist.

- scRNA-Seq poses several unique computational challenges that necessitate adoption of existing workflows, as well as the development and application of entirely new analytical strategies.





Transcription Quantification and Quality Control

- The analysis of scRNA-Seq data requires the careful execuction of different computational steps:
  - ❑ Quality Control
  - ❑ Read Alignment
  - ❑ Generation of gene expression counts
  - ❑ Normalization

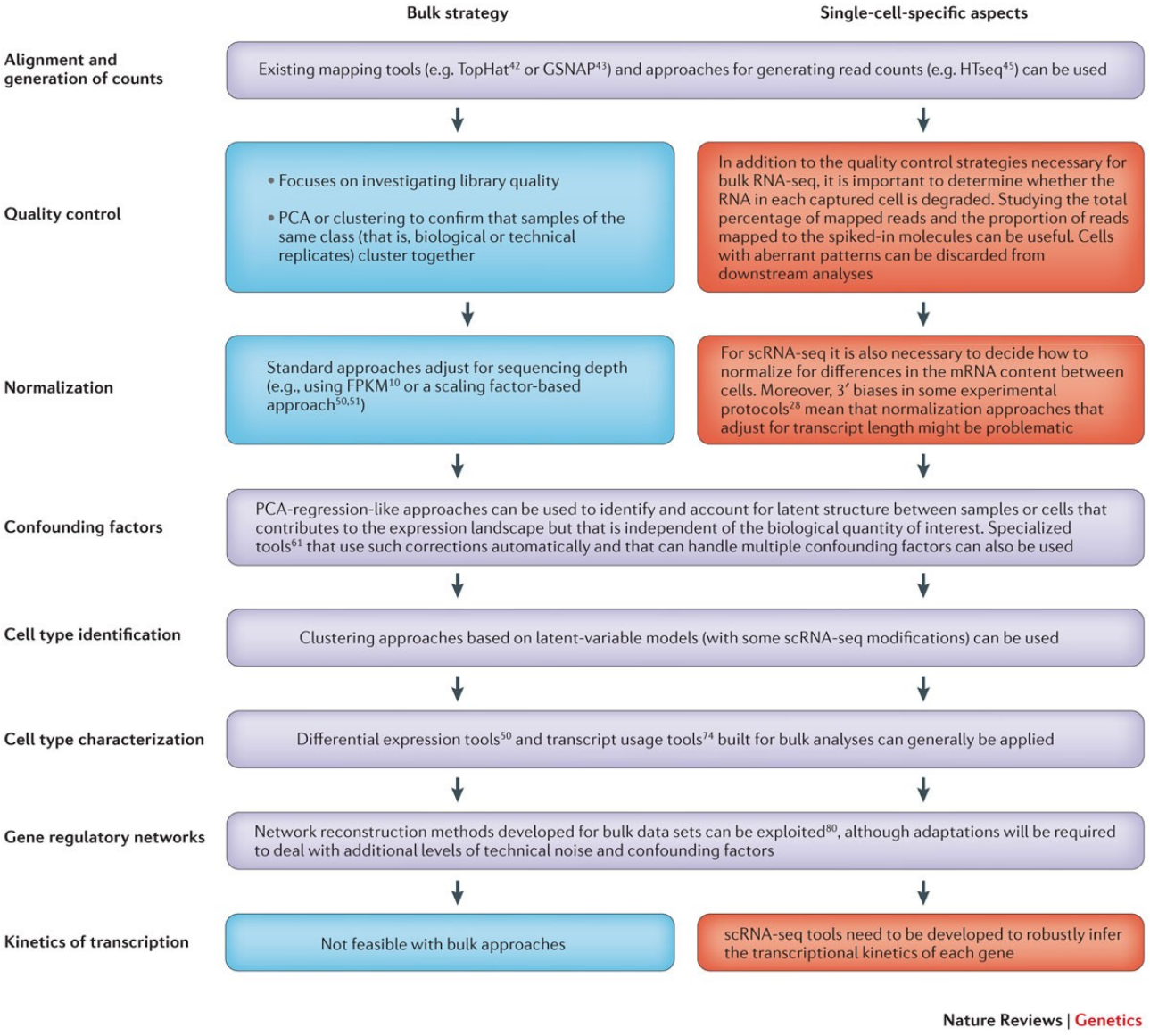
- For several of these tasks specific pipelines and tools have been developed.

Name	For bulk cell populations or single cells?	Function	Ref
Fastqc	Bulk population	Mapping quality control	–
Kraken	Bulk population	Mapping quality control	46
GSNAP	Bulk population	Alignment	43
TopHat	Bulk population	Alignment	42
HTSeq	Bulk population	Obtaining expression counts	45
Single-cell normalization	Single cells	Normalization	33
Monocle	Single cells	Mapping transcripts on differentiation cascade	66
DESeq	Bulk population	Testing for differential expression	50
scLVM	Single cells	Accounting for confounding variation in scRNA-seq	61
Single-cell differential expression	Single cells	Testing for differential expression	55
Kinetics of transcription	Single cells	Identifying kinetic parameters	81

scRNA-seq, single-cell RNA sequencing. In this table, some common tools for the analysis of scRNA-seq data are described. We note that this list is not exhaustive, especially in relation to the suggested tools for analyses of bulk RNA-seq data sets, but instead is meant to give some examples of tools that can be used at all stages of scRNA-seq analysis. See Further Information.

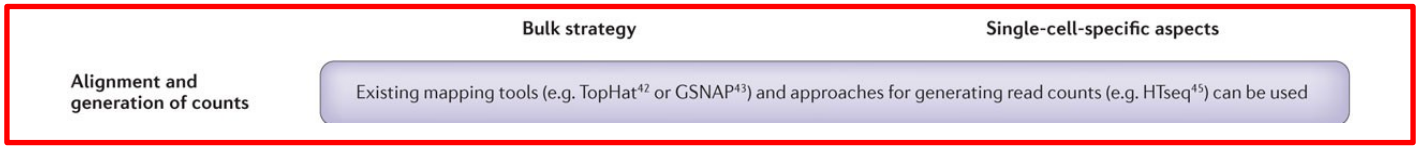
Stegle et al., 2015, Nature Genetics

Workflow data analysis



Stegle et al., 2015, Nature Genetics

Read Alignment



	Bulk strategy	Single-cell-specific aspects
Alignment and generation of counts	Existing mapping tools (e.g. TopHat[42] or GSNAP[43]) and approaches for generating read counts (e.g. HTseq[45]) can be used	

- Read alignment and quantification of expression values rapresent the first steps in the analysis of scRNA-Seq datasets.

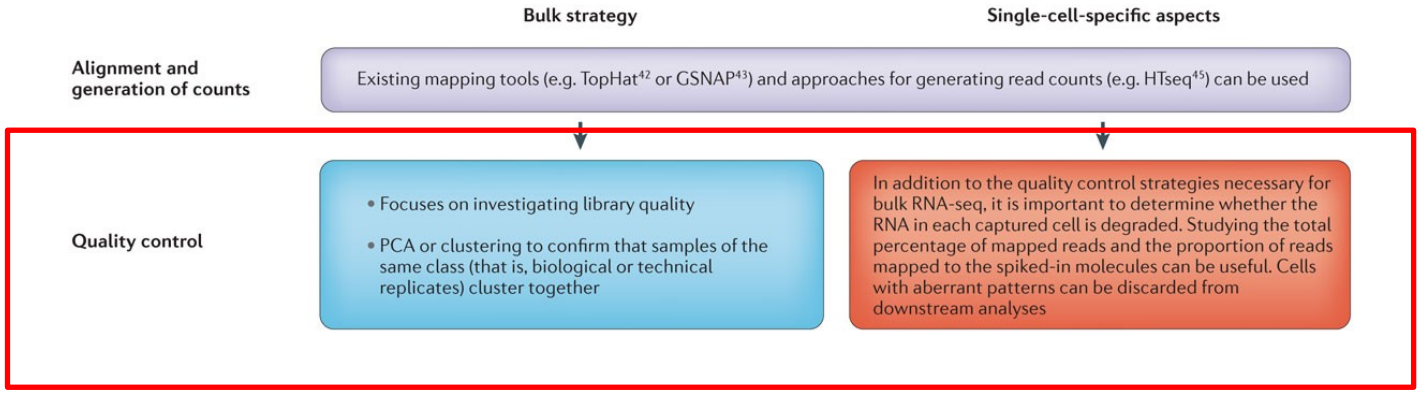
- In general, most of the methodology developed for bulk RNA-Seq can be reused for scRNA-Seq.

- If synthetic spike-ins are used, the reference sequence should be augmented with the DNA sequence of the spike-in molecules prior to mapping.

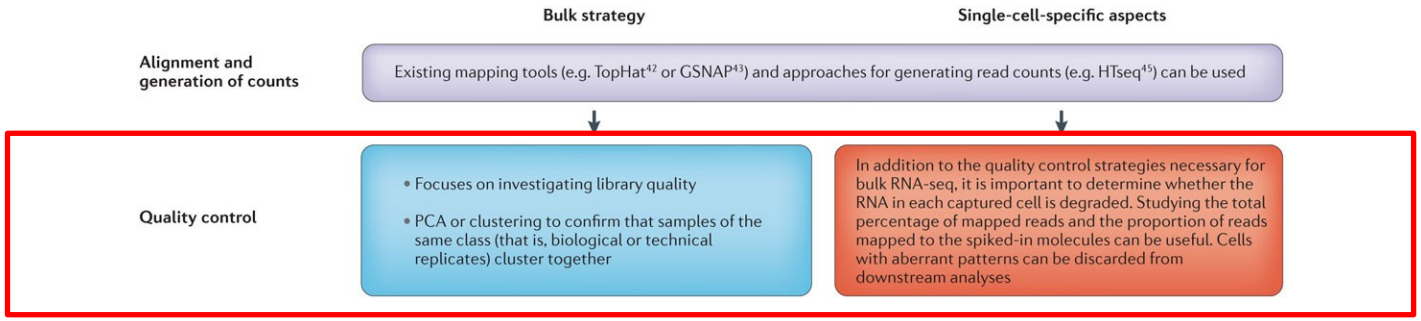
- When a UMIs protocol is used, the barcode attached to each read should be removed before the alignment.

- The mapped reads can be summarized to generate expression levels using the same approaches that are applied in conventional RNA-Seq experiments.

Quality Control



- Quality Control needs to be applied to the raw sequencinbg reads and also after alignment.
- This step is important to identify poor-quality libraries of individual cells.
- At level of raw sequence files, bulk RNA-Seq tools, such as FastQC or Kraken can be used.
- The data can be visualized using tools such Integrative Genomics Viewer.
- These steps will help to identify potential sample mix-up and external contamination or whether there was a problem with the sequencing itself as opposed to single cell caputere and amplification.



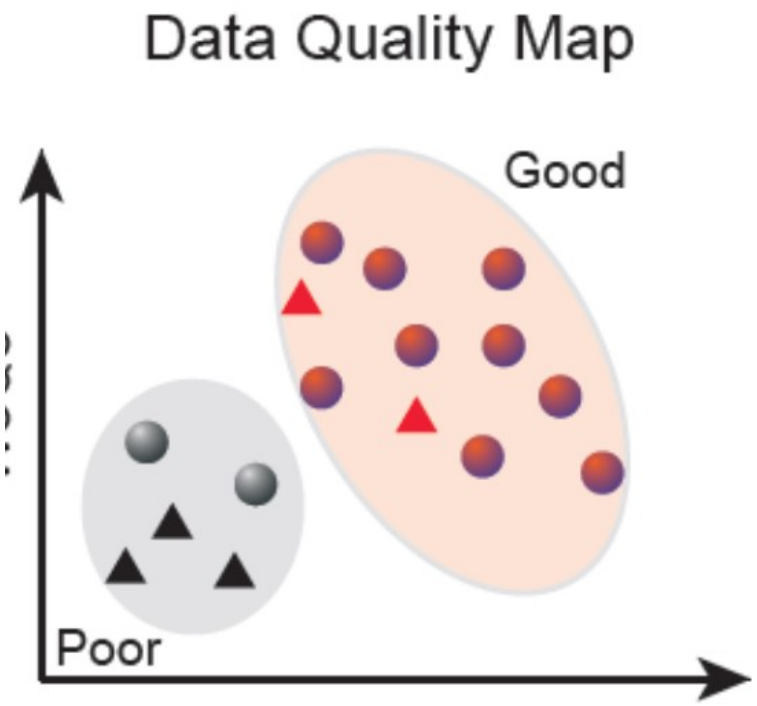
- Metrics can be taken in consideration to evalute quality of sequencing libraries:
  - ❑ Fraction of reads that map back to the genome of interest. If this value is low, it might indicate:
    - ✓ RNA is degradaded
    - ✓ external contamination
    - ✓ Cells were inefficilently lysed
  - ❑ Ratio of the number of reads mapped to the endogenous RNA to the number of reads mapped to extrinsic spike-ins:
    - ✓ a high proportion of reads mapped to the spike-ins would be indicative of a low quality of RNA in the cell of interest and might be a reason to discasrd cells.

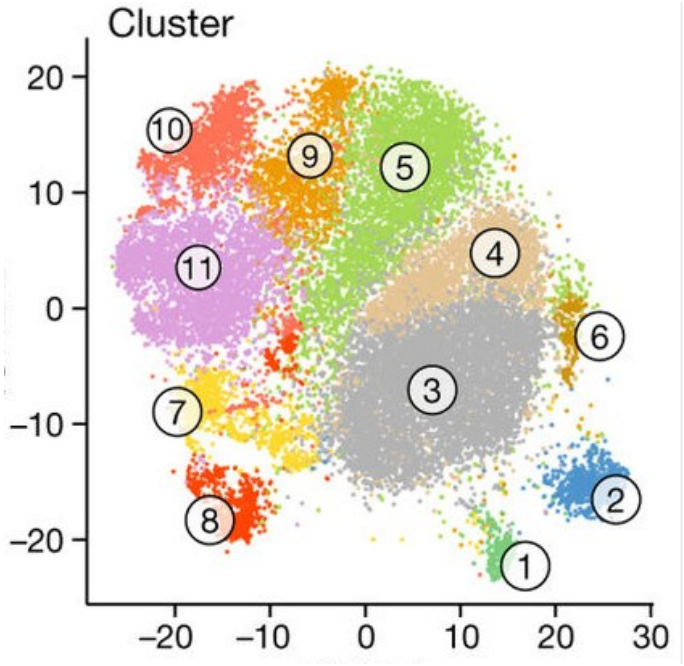
Principal Component Analysis

- The expectation when applying PCA is that good-quality cells cluster together and poor-quality cells are discterted.

- In some instance poor-quality cells may also form a distint cluster.

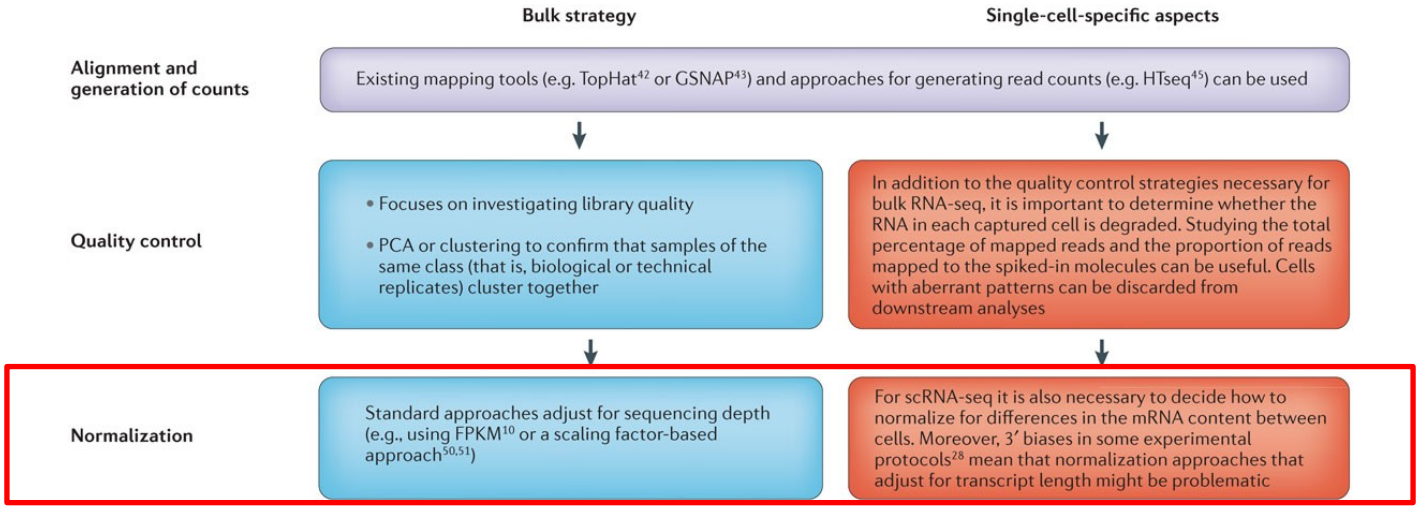
- It has been observed that poor-quality cells are often enriched in the expression of mitochondrial genes, which can cause them to cluster separately.





Kolodziejczyk et al., 2015, Molecular Cell

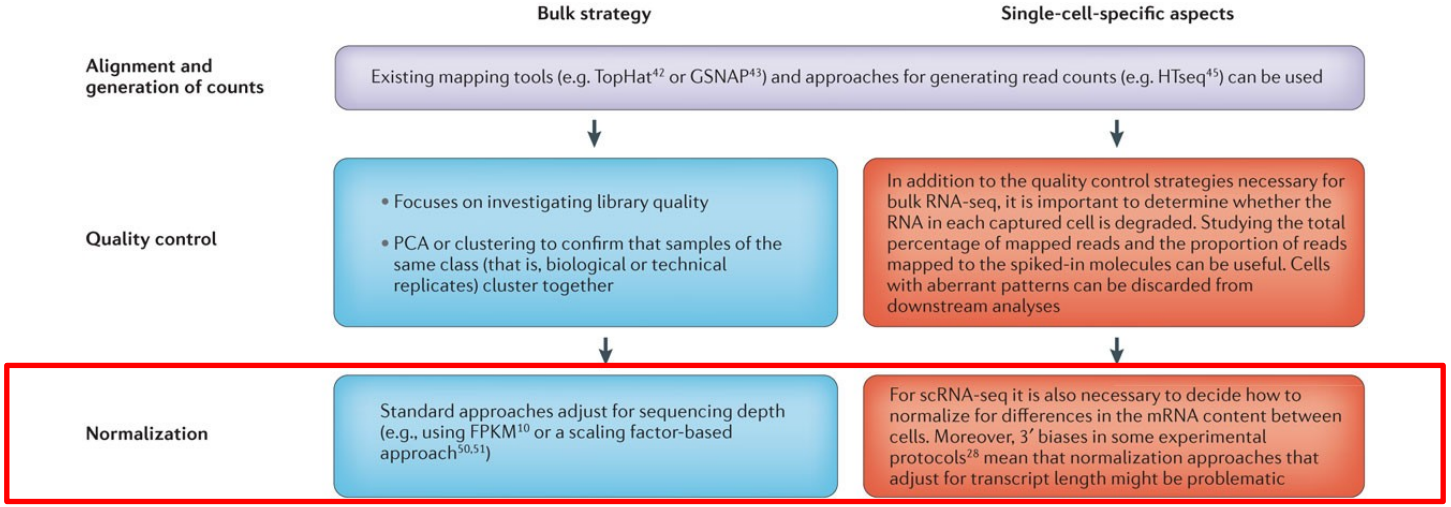
Normalization



- One import computational challenge in scRNA-Seq is to appropriately normalize the data.
- In bulk RNA-Seq data, the counts between different libraries are standardized by calculating quantities such as :
  - ❑"Fragments per Kilobase of Exon per Million Fragments Mapped" (FPKM),
  - ❑Transcripts per Million (TPM),
  - ❑size factor to make counts comparable betwwen libraries obtained from different samples.

- Different strategies can be considered for data generated **with** / **without** **UMIs**.

Normalization of scRNA-Seq data without UMIs

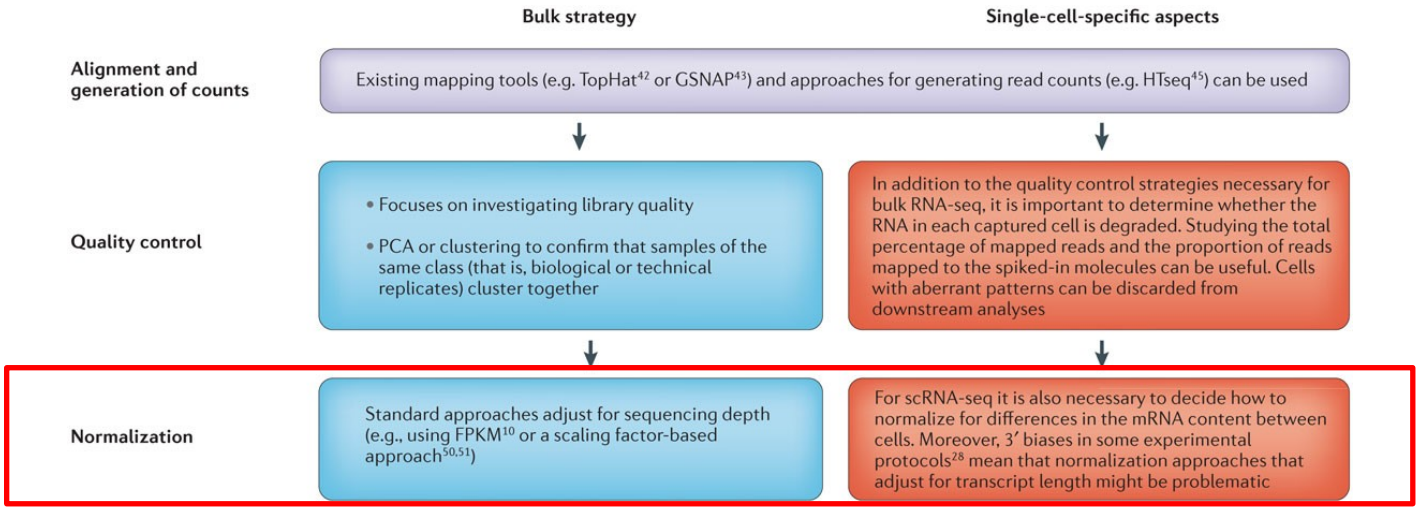


- When extrinsic spike-ins are used:
  - ❑ it is possible to estimate relative differences in the total RNA content between cells.
  - ❑ spike-in material is assumed to be costant across cells.

$$Ratio = \frac{\text{N. reads mapped on genome}}{N.\,reads\ mapped\ to\ spike-ins}$$

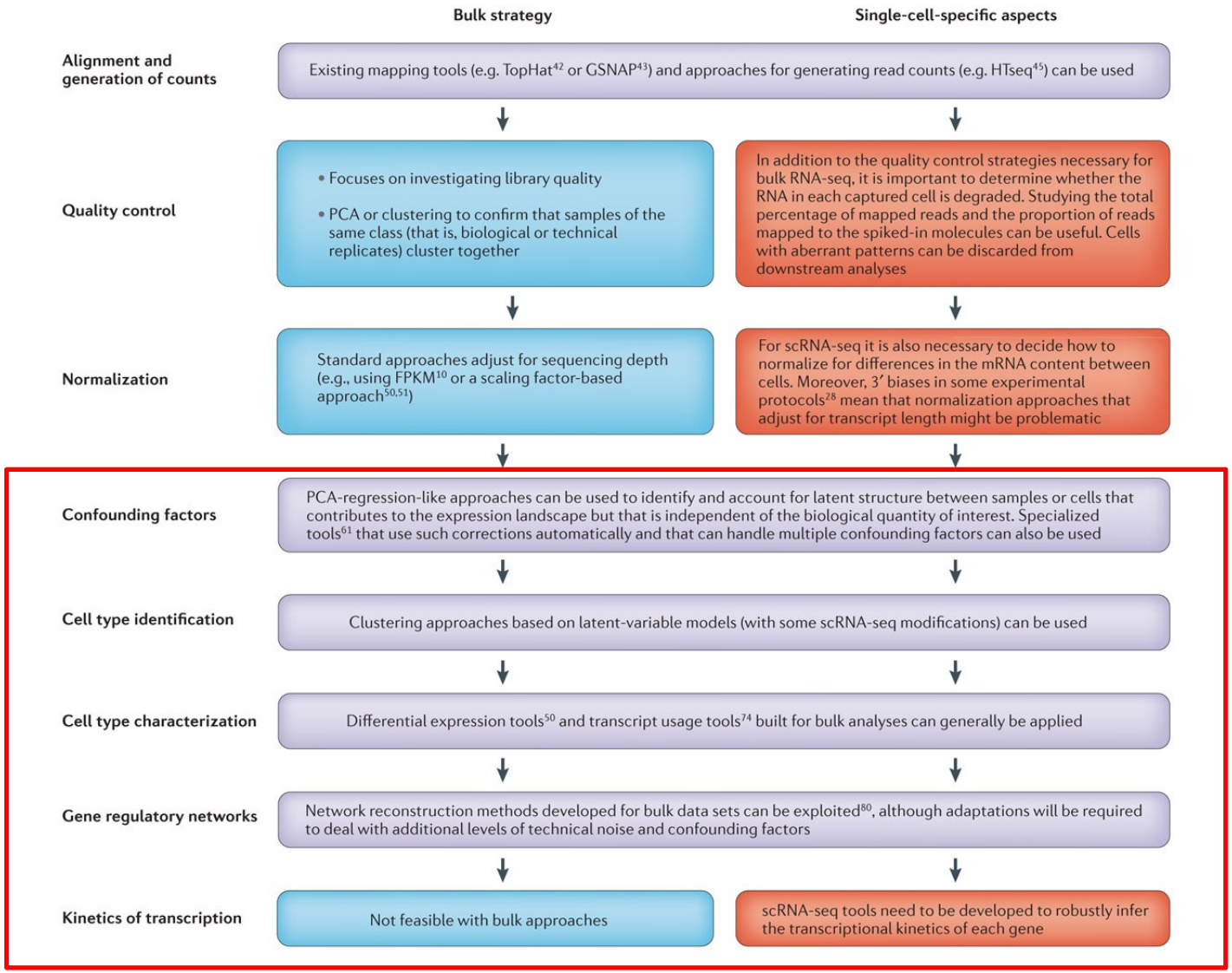
❑ratio allows differences in the amount of RNA within a cell to be inferred.

Normalization of scRNA-Seq data with UMIs



- When UMIs are used, the number of UMIs linked to each gene can be used as a direct measure of the number of cDNA molecules associated with that gene.

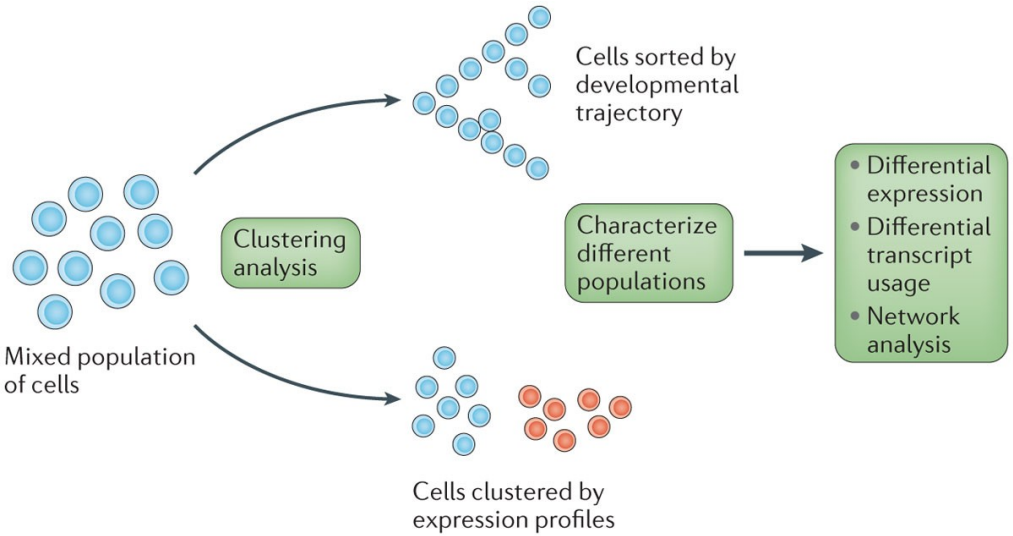
- However, despite the advantage of UMI-based protocols, technical sources of expression variability between cells cannot be fully excluded.



Nature Reviews | Genetics

Biological Insights

- The three biological questions for which scRNA-Seq can provide insights that are not obtainable via bulk RNA-Seq are:
  - ❑ Identification of cell type and cellular state:
    - ✓ scRNA-Seq can be used to address hidden tissue heterogeneity: by clustering cells on the basis of their expression profile.
  - ❑ Differential expression and transcript isoforms.
  - ❑ Identifying highly variable genes.



Stegle et al., 2015, Nature Genetics

Reviews

• The technology and biology of single-cell RNA sequencing, Kolodziejczyk et al., 2015, Molecular Cell

• Computational and analytical challenges in single-cell transcriptomics, Stegle et al., 2015, Nature Reviews Genetics

• Design and computational analysis of single-cell RNA-sequencing experiments, Bacher and Kendziorski, 2016, Genome Biology