Overview of single cell RNA-Seq analysis

Vishal Koparde, Ph. D.

CCR Collaborative Bioinformatics Resource (CCBR)

Leidos Biomedical Research, Inc.

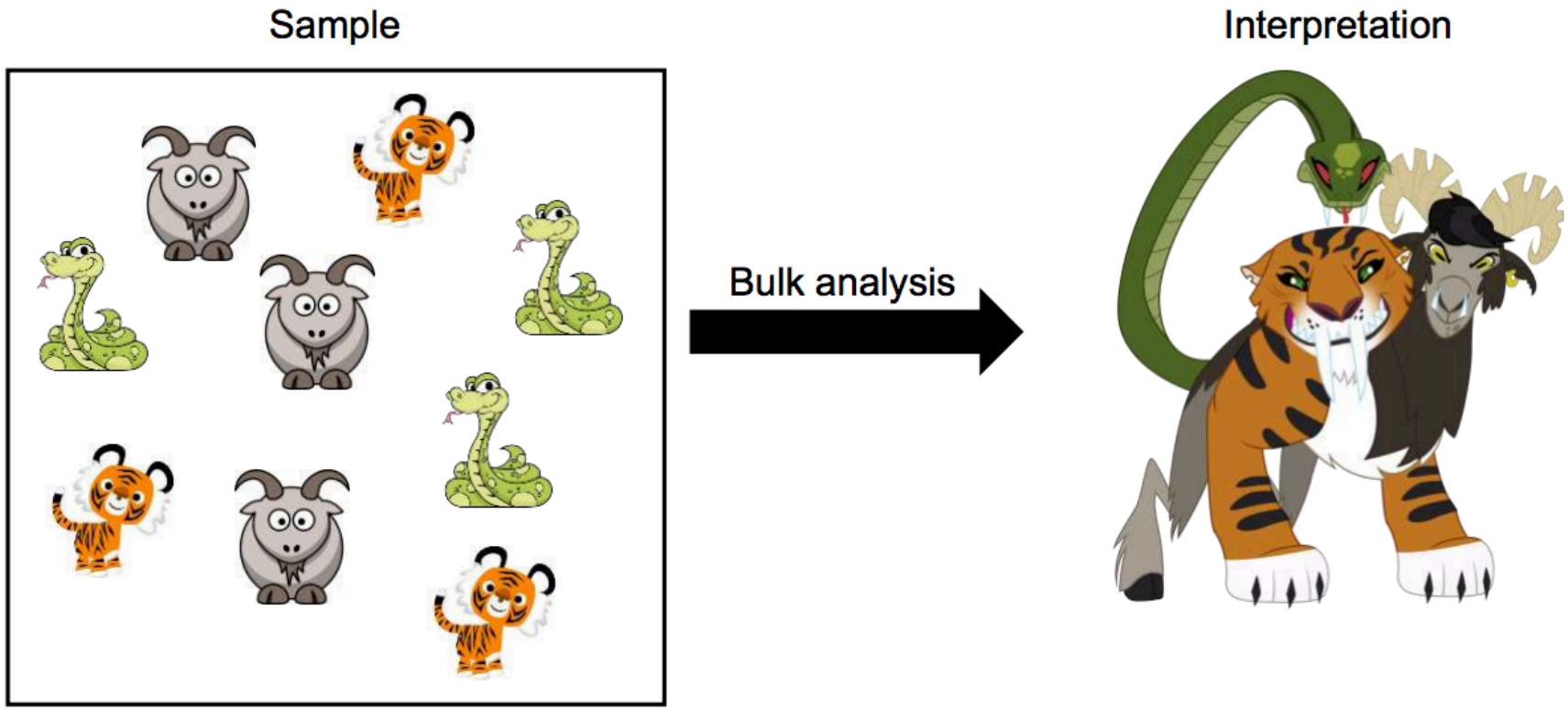
leidos
Leidos Biomedical Research, Inc.

NIH NATIONAL CANCER INSTITUTE
Center for Cancer Research

Outline

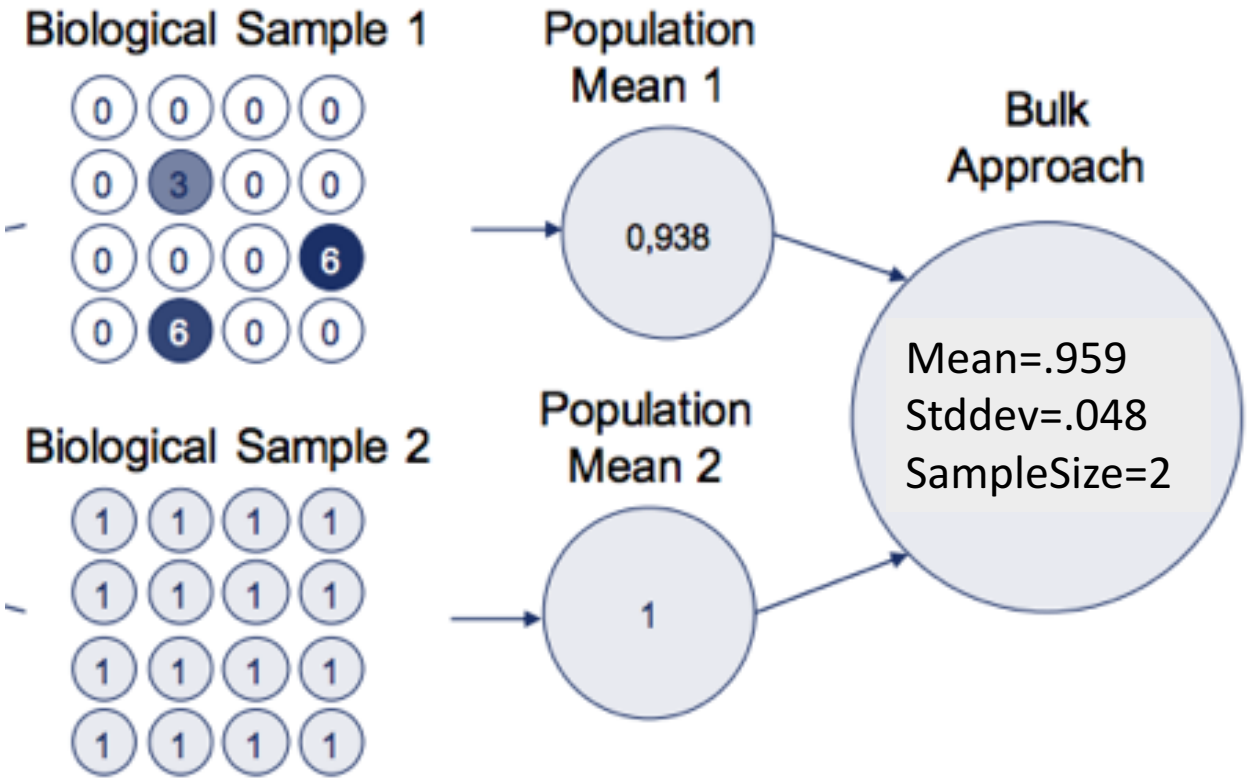
- Bulk RNA-Seq vs scRNA-Seq
- scRNA-Seq applications & challenges
- scRNA-Seq assays
- scRNA-Seq pipeline (UMI, QC, downstream analysis)
- t-SNE visualization
- CCBR scRNA-Seq pipeline

Bulk RNA-Seq: Averaging masks cellular heterogeneity

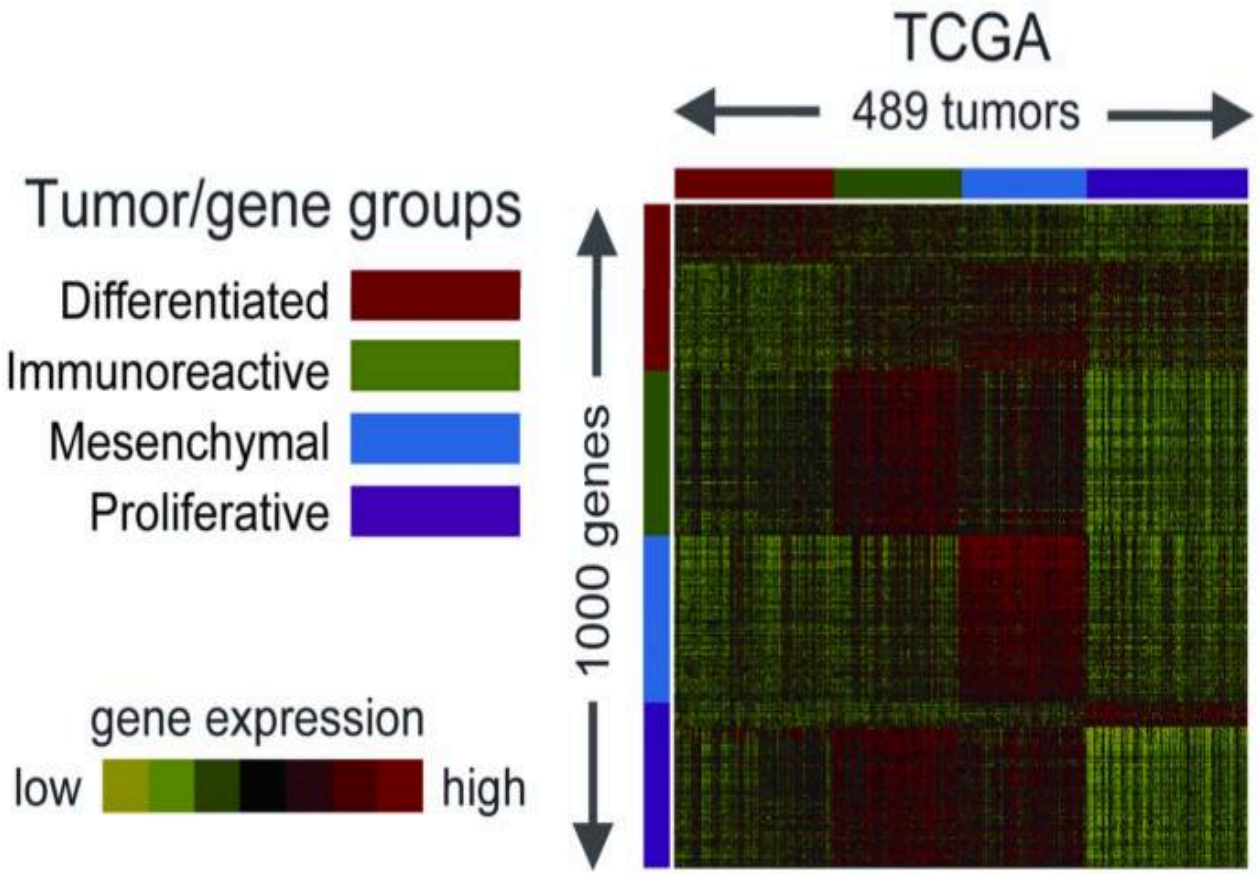


Bulk RNA-Seq

- The basic unit of research is the cell. But we are usually analyzing populations of cells and this can:
  - Lead to false positives from underestimating biological variability
  - Miss important biological divisions



Bulk RNA-Seq

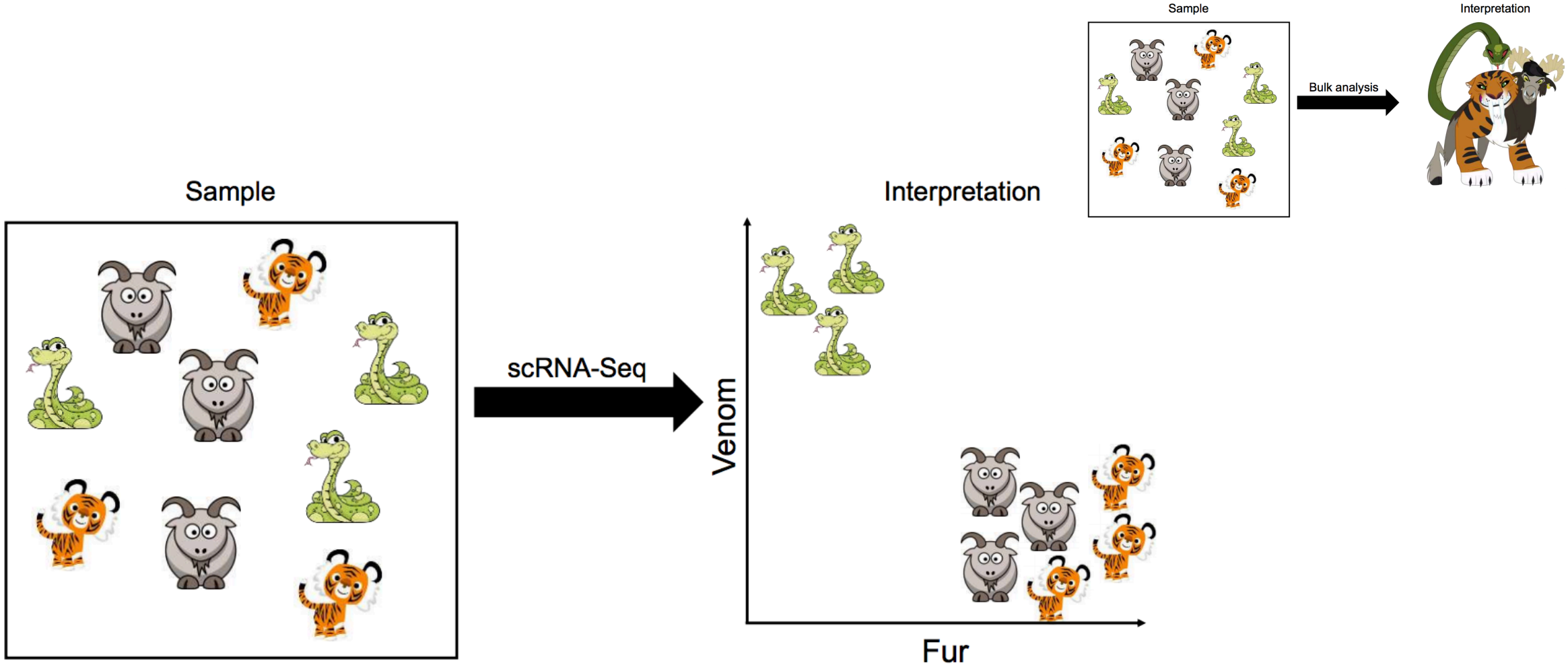


- Impossible to conclude if these differences are due to a uniform property of all cells within a tumor/gene group or due to additive effects of just a few sub-clones within a tumor/gene group
- Conclusions from bulk analysis can be representative of NOTHING

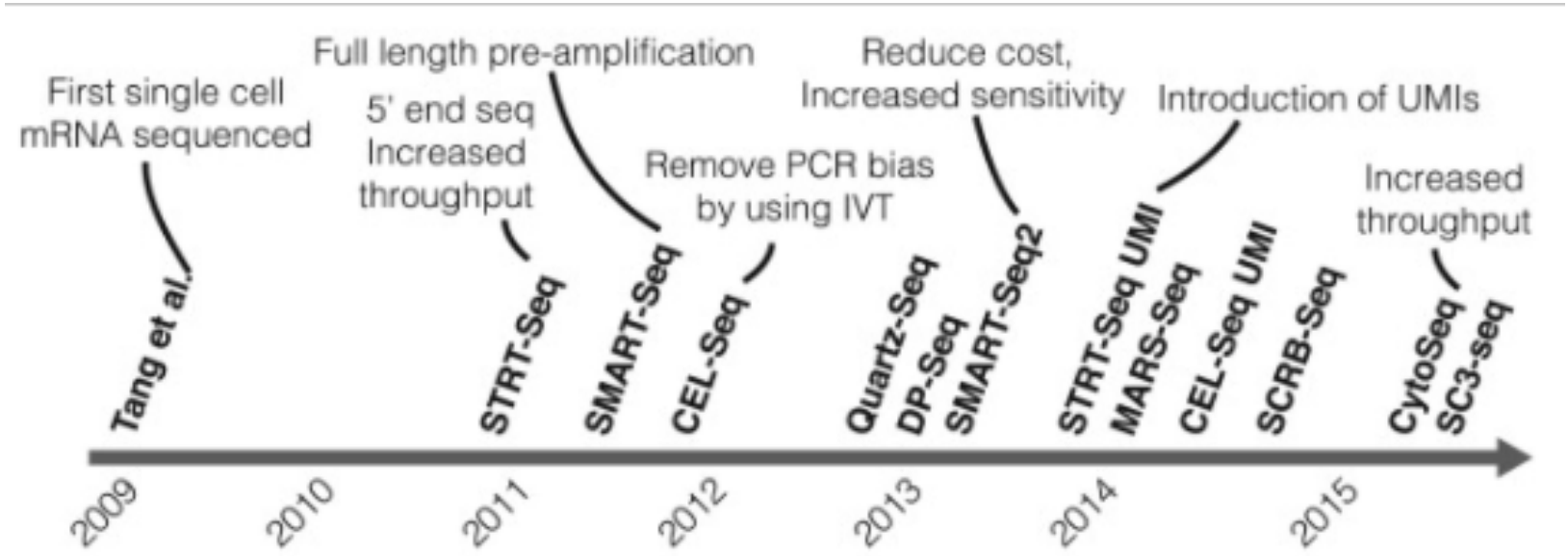
Nature. 2011 Jun 29;474(7353):609-15. doi: 10.1038/nature10166.

Integrated genomic analyses of ovarian carcinoma.

scRNA-Seq


Sample
Bulk analysis
Interpretation
Sample
scRNA-Seq
Interpretation
Venom
Fur

scRNA-Seq evolution



scRNA-Seq has been around for a decade, but has now matured with practical applications

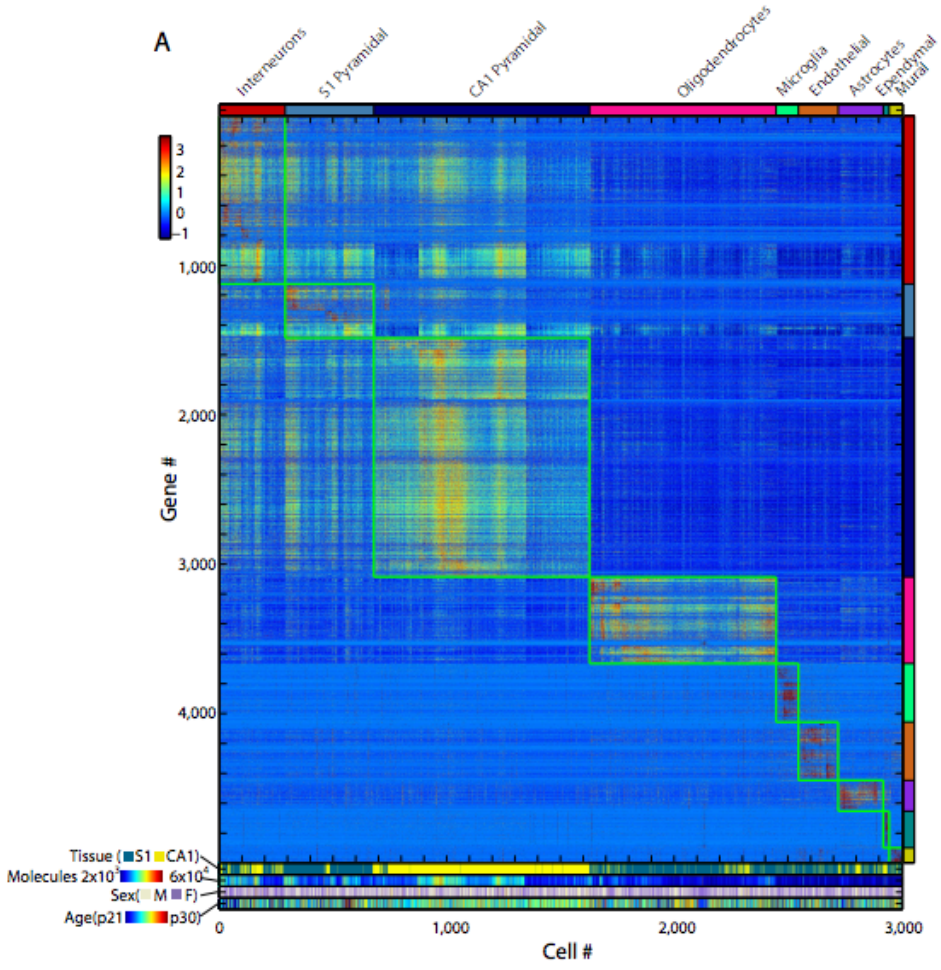
Mol Cell. 2015 May 21;58(4):610-20. doi: 10.1016/j.molcel.2015.04.005.

The technology and biology of single-cell RNA sequencing.

Kolodziejczyk AA[1], Kim JK[2], Svensson V[2], Marioni JC[1], Teichmann SA[3].

Potential of scRNA-Seq

- Investigate Transcript Heterogeneity
  - Differences in transcript abundance
  - Usage of alternate transcription initiation and polyadenylation sites
  - Alternative splicing and differential expression of transcript isoforms

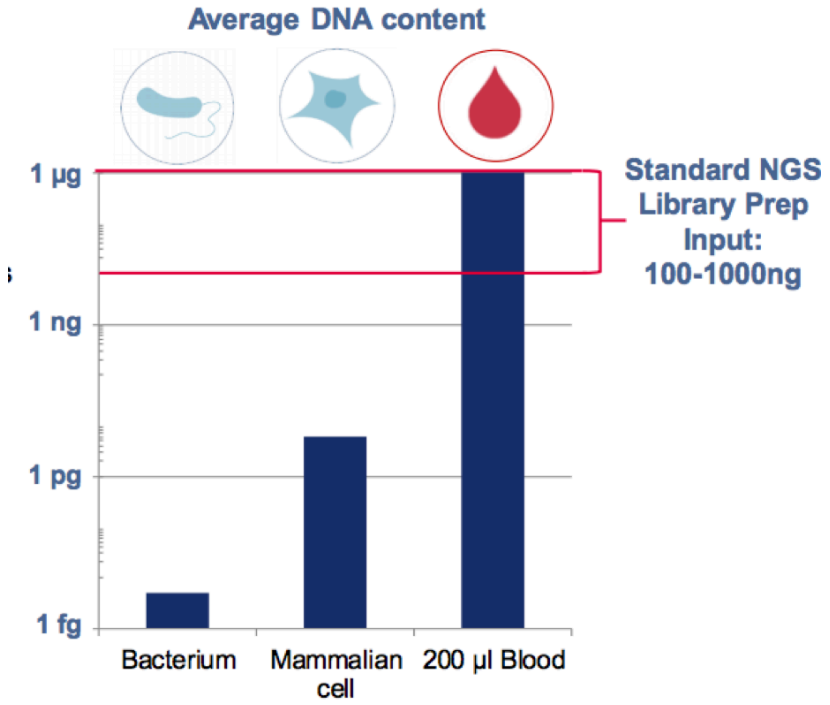


Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.

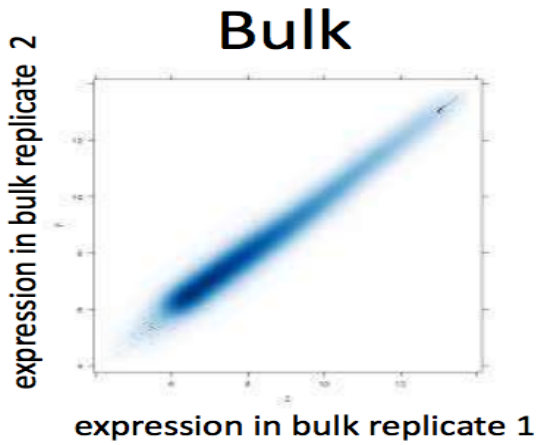
Zeisel A[1], Muñoz-Manchado AB[1], Codeluppi S[1], Lönnerberg P[1], La Manno G[1], Juréus A[1], Marques S[1], Munguba H[1], He L[2], Betsholtz C[3], Rolny C[4], Castelo-Branco G[1], Hjerling-Leffler J[5], Linnarsson S[5].

Challenges: scRNA-Seq

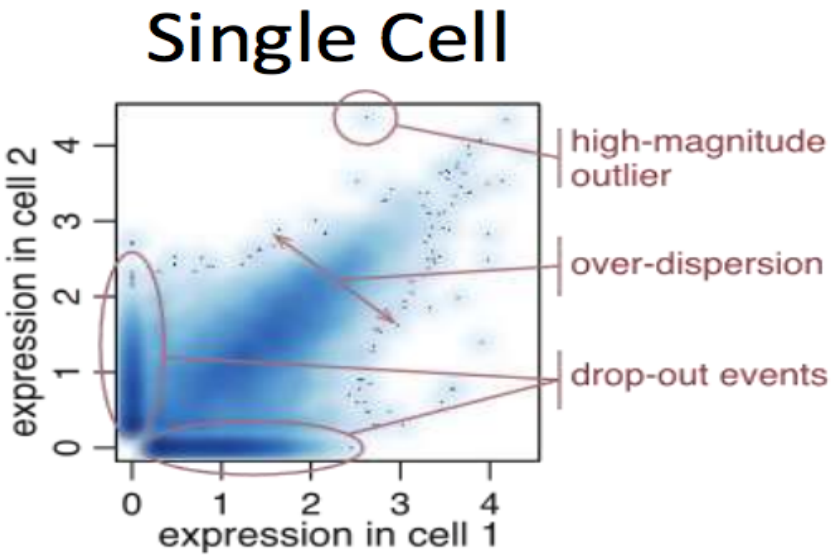
- Limited amounts of material
  - Illumina/PacBio require 1ng-500ng
  - 1 bacterium or 1 cancer cell contain 1fg-1pg
  - ~10-50pg of total RNA → 1-5% of which is mRNA
  - WTA/WGA become really important



Challenges: scRNA-Seq



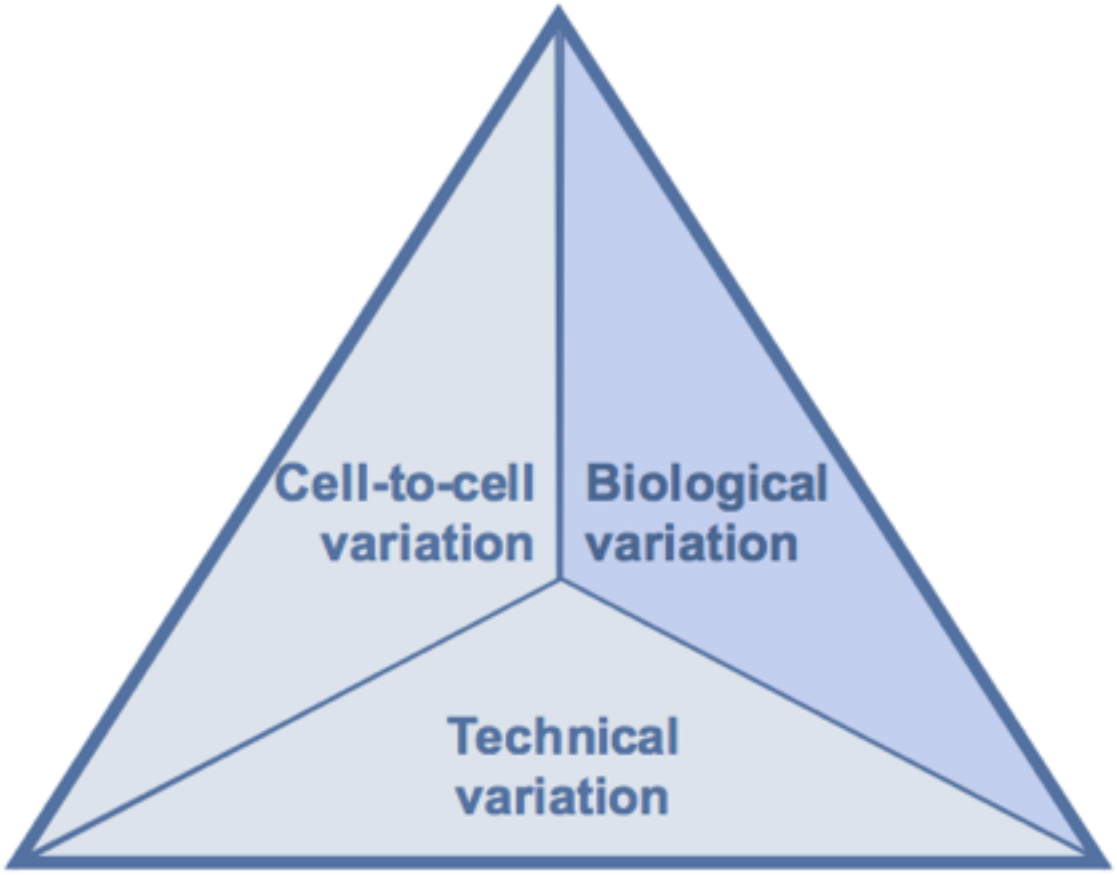
- Relatively high correlation is expected between samples



- Highly variable and noisy
- Many differences even between same type of cells
- Biological signal diminishes at cellular level and becomes comparable to technical variability

Challenges: scRNA-Seq

- Viability
- Cell stress
- Depends on application
- Budget considerations
- Technical variation
  - Cell capture
  - Library preparation
  - Inherent sequencing biases



scRNA-Seq : Different assays

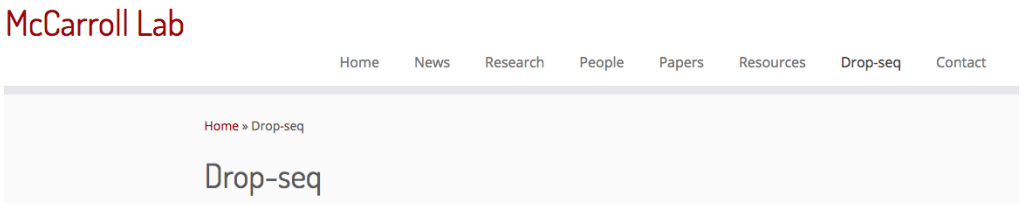
- There are many scRNA-Seq assays out there:
  - Some are commercialized … growing
  - Full transcript vs 3’
  - Isoform vs gene
  - Less or more automated
  - Different levels of throughput
  - Differences in costs
  - Differences in downstream bioinformatics analysis

SmartSeq2

- Full transcript scRNA-Seq
- Selects for poly-A tail
- 5’ switching for full transcript length capture
- Standard Illumina sequencing
  - Off-the-shelf products
- Hundreds of samples
- No UMIs used

DropSeq

- Droplet-based
- Throughput from hundreds to thousands
- 3’ End only
- Single cell transcriptomes attached to microparticles
- UMIs (unique molecular identifiers)
- Cell barcodes
- Drop-seq tools software



10X Genomics

- Droplet-based
  - GEM → Gel bead in EMulsion
- 3’ mRNA
- Standardized instrumentation and reagents
- More high-throughput scaling to tens of thousands of cells
- Less processing time
- Cell ranger software

10X Genomics Cell Ranger

- cellranger mkfastq … BCL to fastq
- cellranger count … fastq to count matrix
- cellranger aggr … combine fastqs from multiple runs into one count matrix
- cellranger reanalyze … count matrix to dimensionality reduction/clustering
- Python and R support included

Analysis pipelines differ a lot… why?

- scRNA-Seq datasets come in different flavors:
  - “Fishing expedition”: Dissociate tissue → get subpopulations!
  - Case vs Control
  - Molecular transformations over time (Pseudo time)
  - Perturbations eg. CRISPR Screening
- Very hard to have a single do-it-all pipeline for scRNA-Seq analysis

New pipelines keep popping up…

Software packages

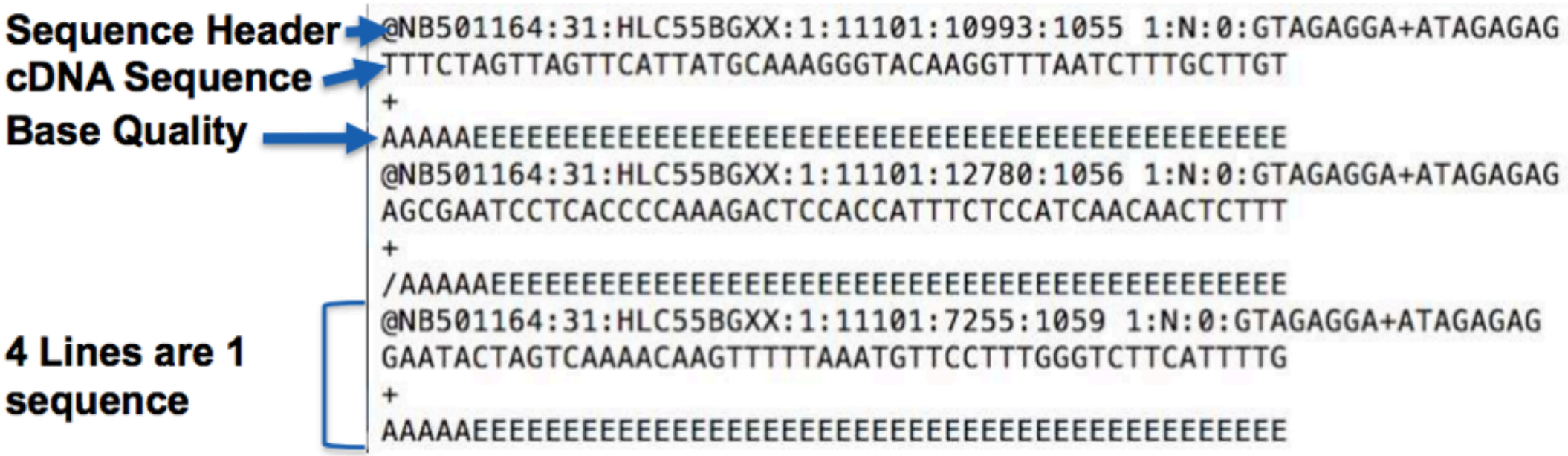
RNA-seq

- anchor - [Python] - ⚓Find bimodal, unimodal, and multimodal features in your data
- ascend - [R] - ascend is an R package comprised of fast, streamlined analysis functions optimized to address the statistical challenges of single cell RNA-seq. The package incorporates novel and established methods to provide a flexible framework to perform filtering, quality control, normalization, dimension reduction, clustering, differential expression and a wide-range of plotting.
- BackSPIN - [Python] - Biclustering algorithm developed taking into account intrinsic features of single-cell RNA-seq experiments.
- BASiCS - [R] - Bayesian Analysis of single-cell RNA-seq data. Estimates cell-specific normalization constants. Technical variability is quantified based on spike-in genes. The total variability of the expression counts is decomposed into technical and biological components. BASiCS can also identify genes with differential expression/over-dispersion between two or more groups of cells.
- BatchEffectRemoval - [Python] - Removal of Batch Effects using Distribution-Matching Residual Networks
- BEARscc - [R] - BEARscc makes use of ERCC spike-in measurements to model technical variance as a function of gene expression and technical dropout effects on lowly expressed genes.
- bonvoyage - [Python] - 📐 Transform percentage-based units into a 2d space to evaluate changes in distribution with both magnitude and direction.
- BPSC - [R] - Beta-Poisson model for single-cell RNA-seq data analyses
- CellCNN - [Python] - Representation Learning for detection of phenotype-associated cell subsets
- Cellity - [R] - Classification of low quality cells in scRNA-seq data using R
- cellTree - [R] - Cell population analysis and visualization from single cell RNA-seq data using a Latent Dirichlet Allocation model.
- clusterExperiment - [R] - Functions for running and comparing many different clusterings of single-cell sequencing data. Meant to work with SCONE and slingshot.
- CytoGuide - [C++,D3] - CyteGuide: Visual Guidance for Hierarchical Single-Cell Analysis
- destiny - [R] - Diffusion maps are spectral method for non-linear dimension reduction introduced by Coifman et al.(2005). Diffusion maps are based on a distance metric (diffusion distance) which is conceptually relevant to how differentiating cells follow noisy diffusion-like dynamics, moving from a pluripotent state towards more differentiated states.
- DeLorean - [R] - Bayesian pseudotime estimation algorithm that uses Gaussian processes to model gene expression profiles and provides a full posterior for the pseudotimes.
- dropClust - [R/Python] - Efficient clustering of ultra-large scRNA-seq data.
- ECLAIR - [python] - ECLAIR stands for Ensemble Clustering for Lineage Analysis, Inference and Robustness. Robust and scalable inference of cell lineages from gene expression data.
- embeddr - [R] - Embeddr creates a reduced dimensional representation of the gene space using a high-variance gene correlation graph and laplacian eigenmaps. It then fits a smooth pseudotime trajectory using principal curves.
- Falco - [AWS cloud] - Falco: A quick and flexible single-cell RNA-seq processing framework on the cloud.
- FastProject - [Python] - Signature analysis on low-dimensional projections of single-cell expression data.
- flotilla - [Python] - Reproducible machine learning analysis of gene expression and alternative splicing data
- GPfates - [Python] - Model transcriptional cell fates as mixtures of Gaussian Processes
- GiniClust - [Python/R] - GiniClust is a clustering method implemented in Python and R for detecting rare cell-types from large-scale single-cell gene expression data. GiniClust can be applied to datasets originating from different platforms, such as multiplex qPCR data, traditional single-cell RNAseq or newly emerging UMI-based single-cell RNAseq, e.g. inDrops and Drop-seq.
- HocusPocus - [R] - Basic PCA-based workflow for analysis and plotting of single cell RNA-seq data.
- ICGS - [Python] - Iterative Clustering and Guide-gene Selection (Olsson et al. Nature 2016). Identify discrete, transitional and mixed-lineage states from diverse single-cell transcriptomics platforms. Integrated FASTQ pseudoalignment /quantification (Kallisto), differential expression, cell-type prediction and optional cell cycle exclusion analyses. Specialized methods for processing BAM and 10X Genomics spares matrix files. Associated single-cell splicing PSI methods (MultIPath-PSI). Apart of the AltAnalyze toolkit along with accompanying visualization methods (e.g., heatmap, t-SNE, SashimiPlots, network graphs). Easy-to-use graphical user and commandline interfaces.
- MAGIC - [python or matlab] - Markov Affinity-based Graph Imputation of Cells (MAGIC).
- MAST - [R] - Model-based Analysis of Single-cell Transcriptomics (MAST) fits a two-part, generalized linear models that are specially adapted for bimodal and/or zero-inflated single cell gene expression data.
- mfa - [R] - Bayesian modelling of bifurcations using a mixture of factor analysers
- K-Branches - [R] - The main idea behind the K-Branches method is to identify regions of interest (branching regions and tips) in differentiation trajectories of single cells. So far, K-Branches is intended to be used on the diffusion map representation of the data, so the user should either provide the data in diffusion map space or use the destiny package perform diffusion map dimensionality reduction.
- M3Drop - [R] - Michaelis-Menten Modelling of Dropouts for scRNASeq.
- MAST - [R] - Model-based Analysis of Single-cell Transcriptomics (MAST) fits a two-part, generalized linear models that are specially adapted for bimodal and/or zero-inflated single cell gene expression data
- MIMOSCA - [python] - A repository for the design and analysis of pooled single cell RNA-seq perturbation experiments (Perturb-seq).
- Monocle - [R] - Differential expression and time-series analysis for single-cell RNA-Seq.

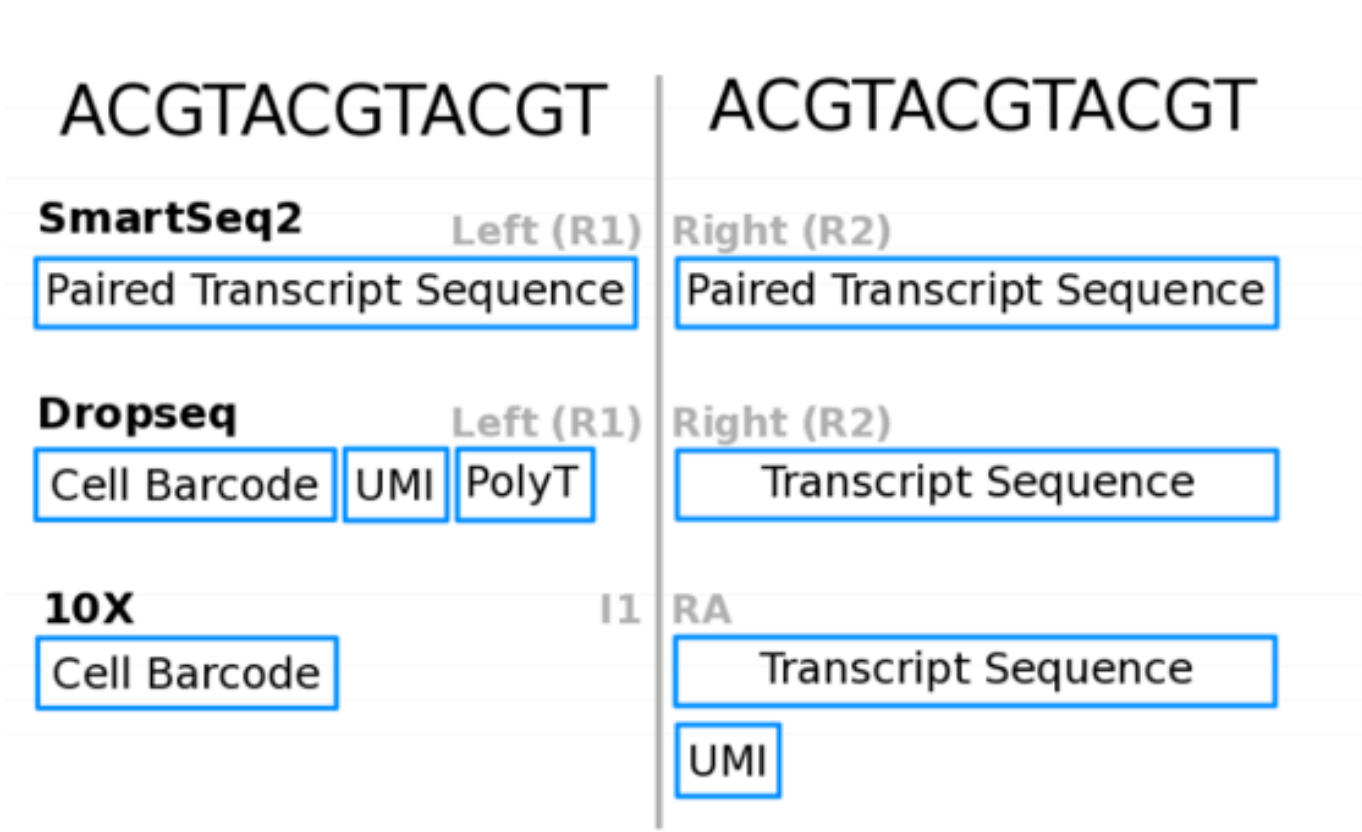
https://github.com/seandavi/awesome-single-cell

Raw data

FASTQ File Format



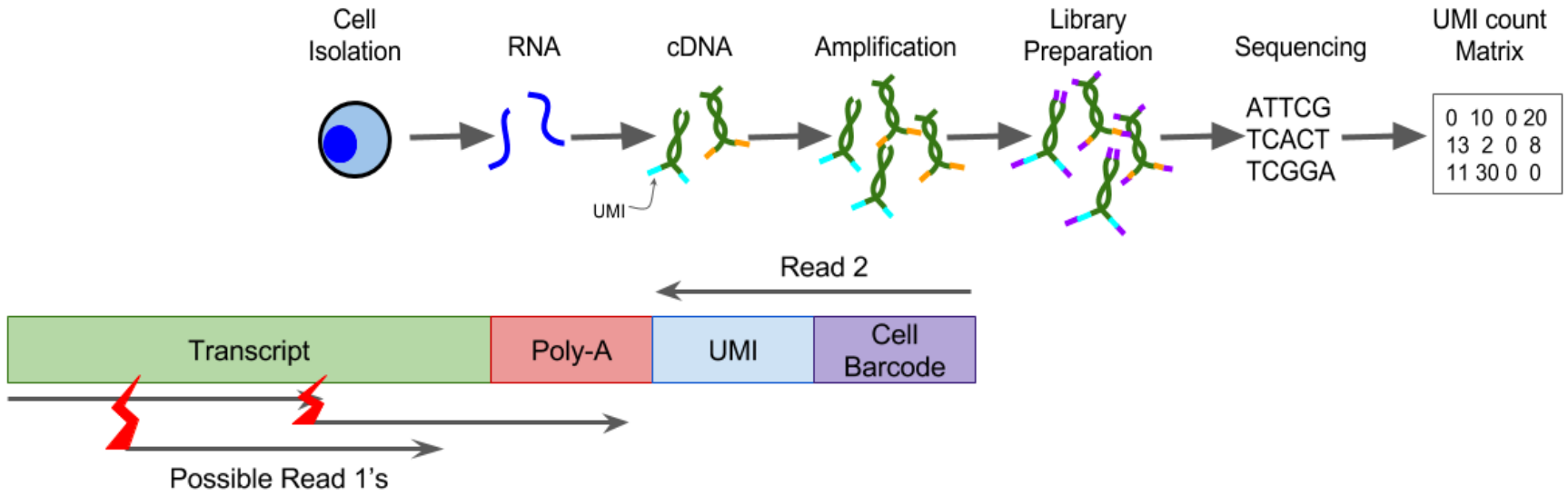
Assays differ by FASTQ content



- Dropseq and 10X is really just single end RNA Seq on a per cell level

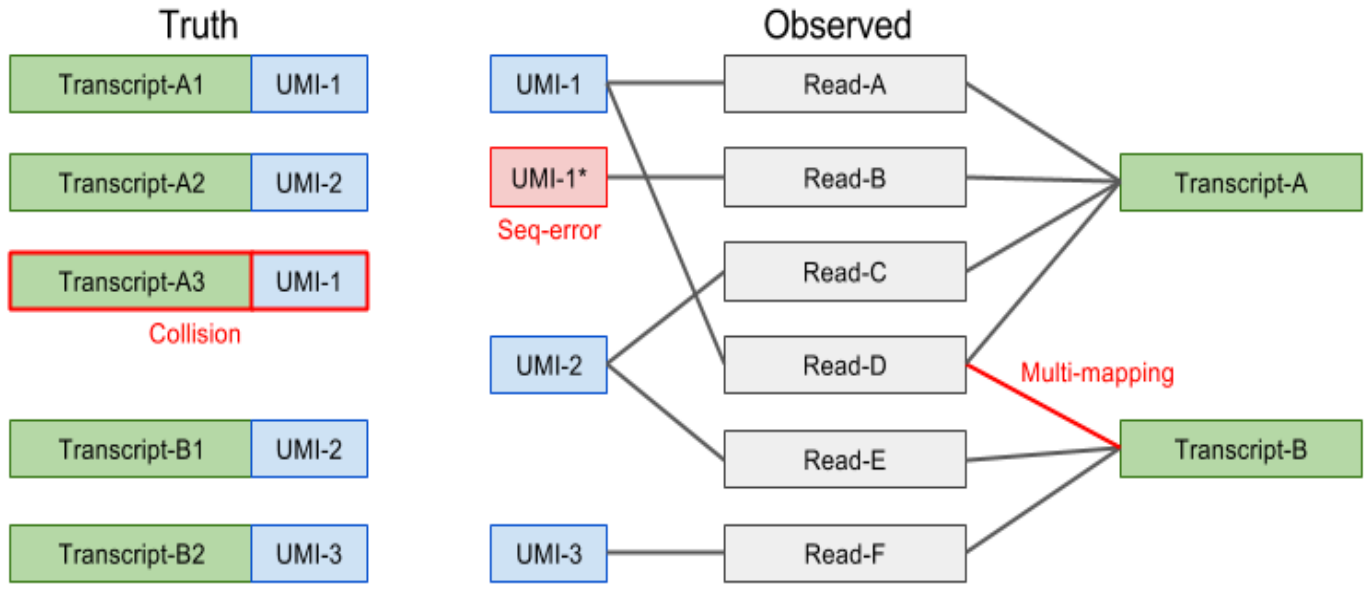
UMI

- Unique Molecular Identifiers are short (4-10bp) random barcodes added to transcripts during reverse-transcription.
- They enable sequencing reads to be assigned to individual transcript molecules and thus the removal of amplification noise and biases from scRNASeq data

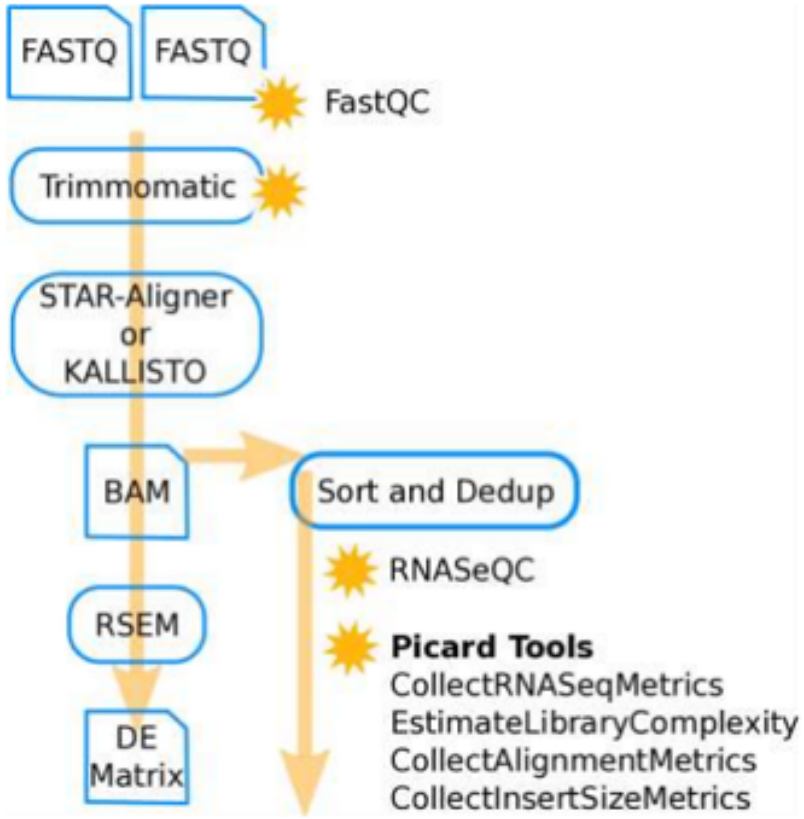


Issues with UMI

- Different UMI doesn't necessarily mean different molecule
- Different transcript doesn't necessarily mean different molecule
- Same UMI doesn't necessarily mean same molecule
- Cell-range performs built-in error correction for UMI-Read-Transcript mapping

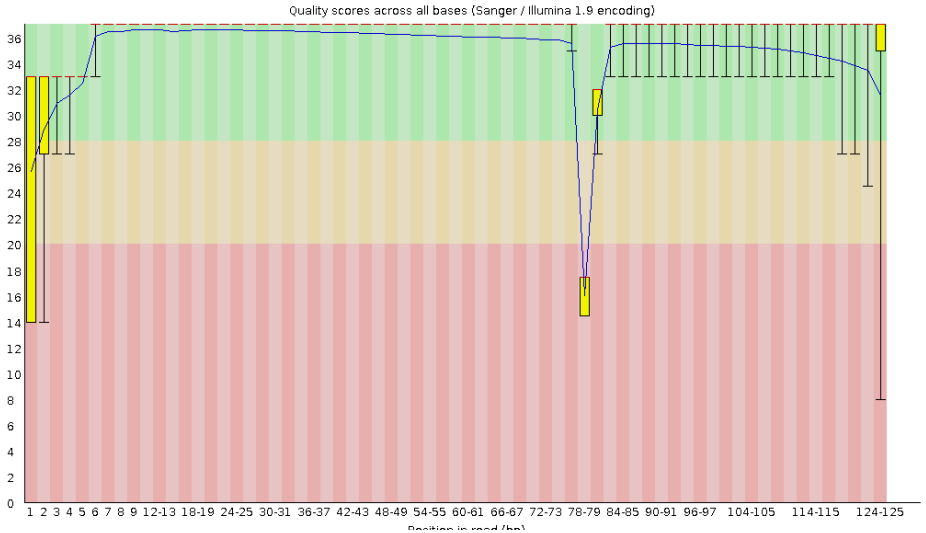


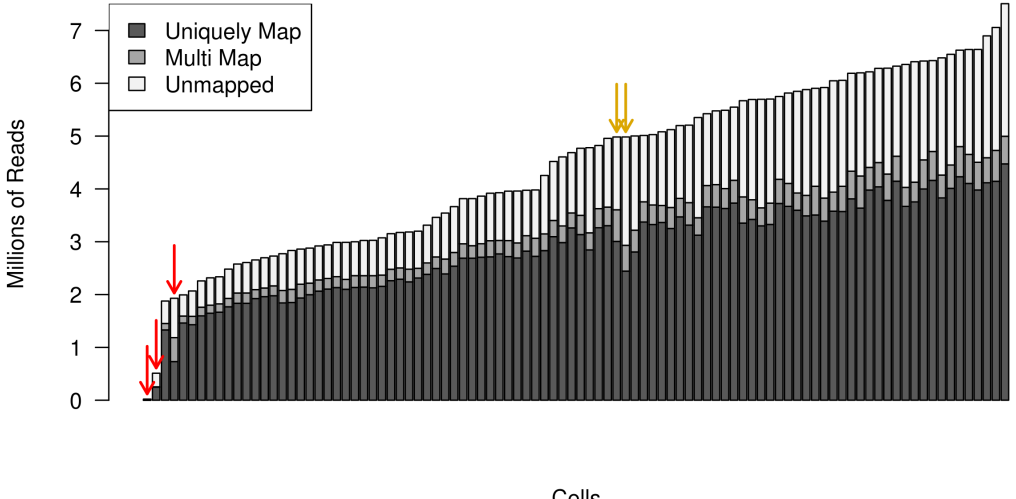
scRNA-Seq pipeline : QC

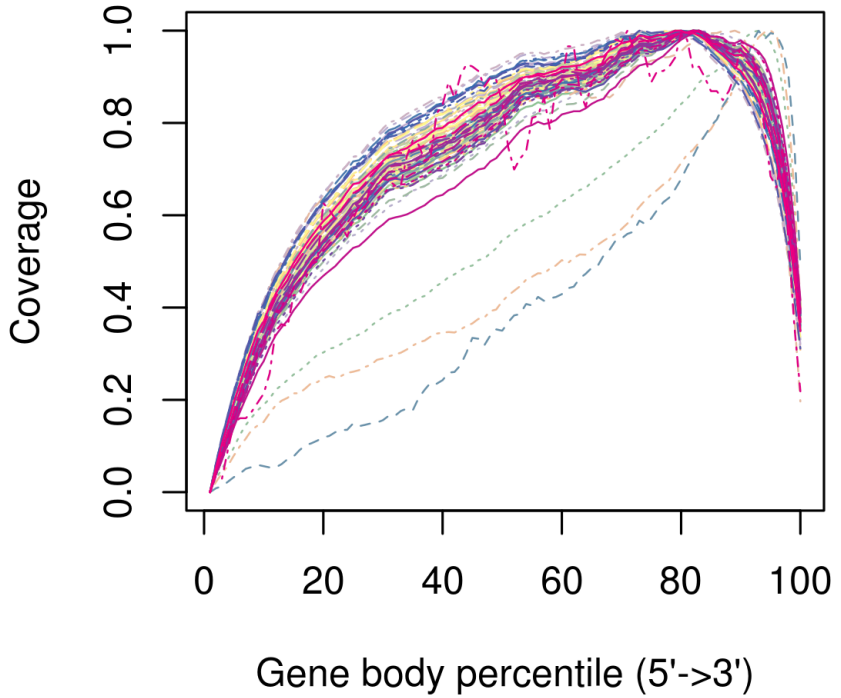


- Read alignments and much of the QC can be handled similar to bulk RNA-Seq
- Cell ranger uses masked human and mouse genome to exclude “problem” regions apriori

scRNA-Seq pipeline : QC



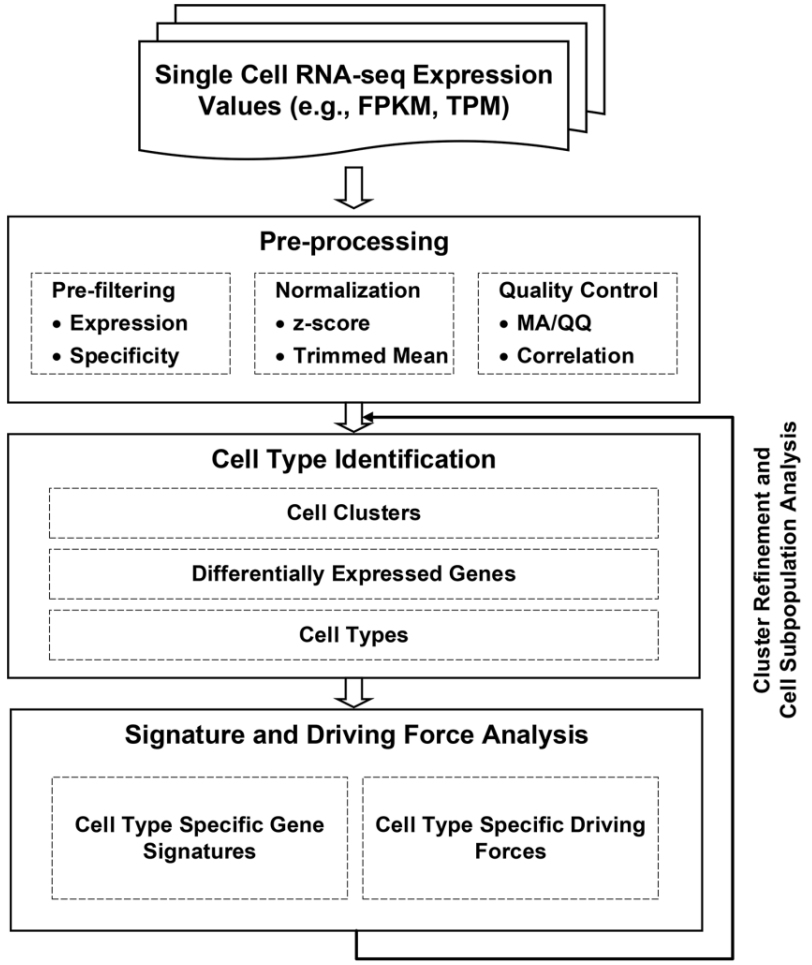




scRNA-Seq pipeline : Downstream analysis

- Complete end-to-end pipeline in R
  - SINCERA used by Cincinnati Children's Hospital

- R packages
  - Seurat
  - scater
  - SC3



Common steps of a scRNA-Seq pipeline

- Aligning to genome
- QC
- Read counting
- Filtering
- Normalization
- Clustering

- Selection of tools vary between dataset to dataset … depend on the bigger biological question

Count Matrix

	Cell 1	Cell2	Cell3	Cell4	...
Gene 1	0	0	3	10	
Gene 2	24	0	41	12	
Gene 3	175	284	93	162	
Gene 4	0	0	0	0	
Gene 5	36	0	32	21	
...	...	...	...	...	

- 5000-12000 genes per cell on 10X
- Most entries are zero → Sparse matrix
- Different efficient representation of sparse matrix in memory are used

Count Matrix as coordinate list

Can be optimal for dense matrices

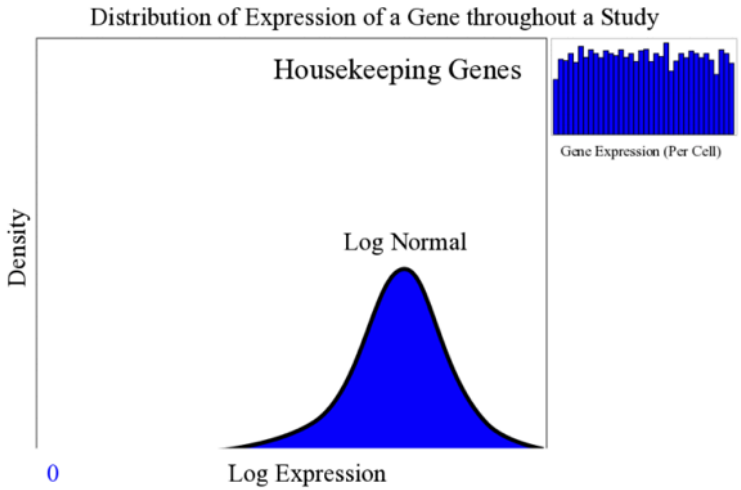
More optimal for sparse matrices

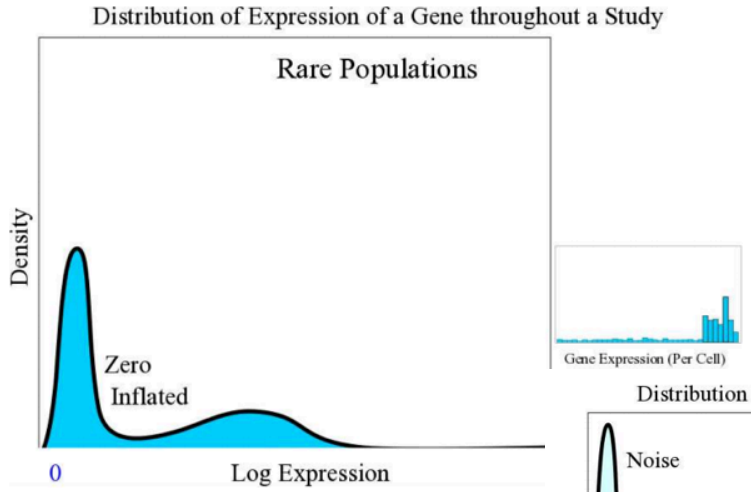
2D Arrays **vs** Coordinate List

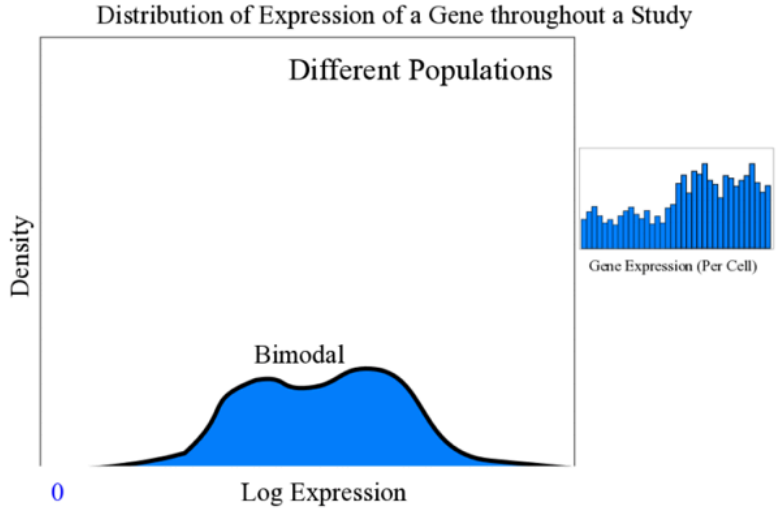
	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0
3	0	0	0	0	0	2	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	3

2	2	1
6	3	2
8	6	3

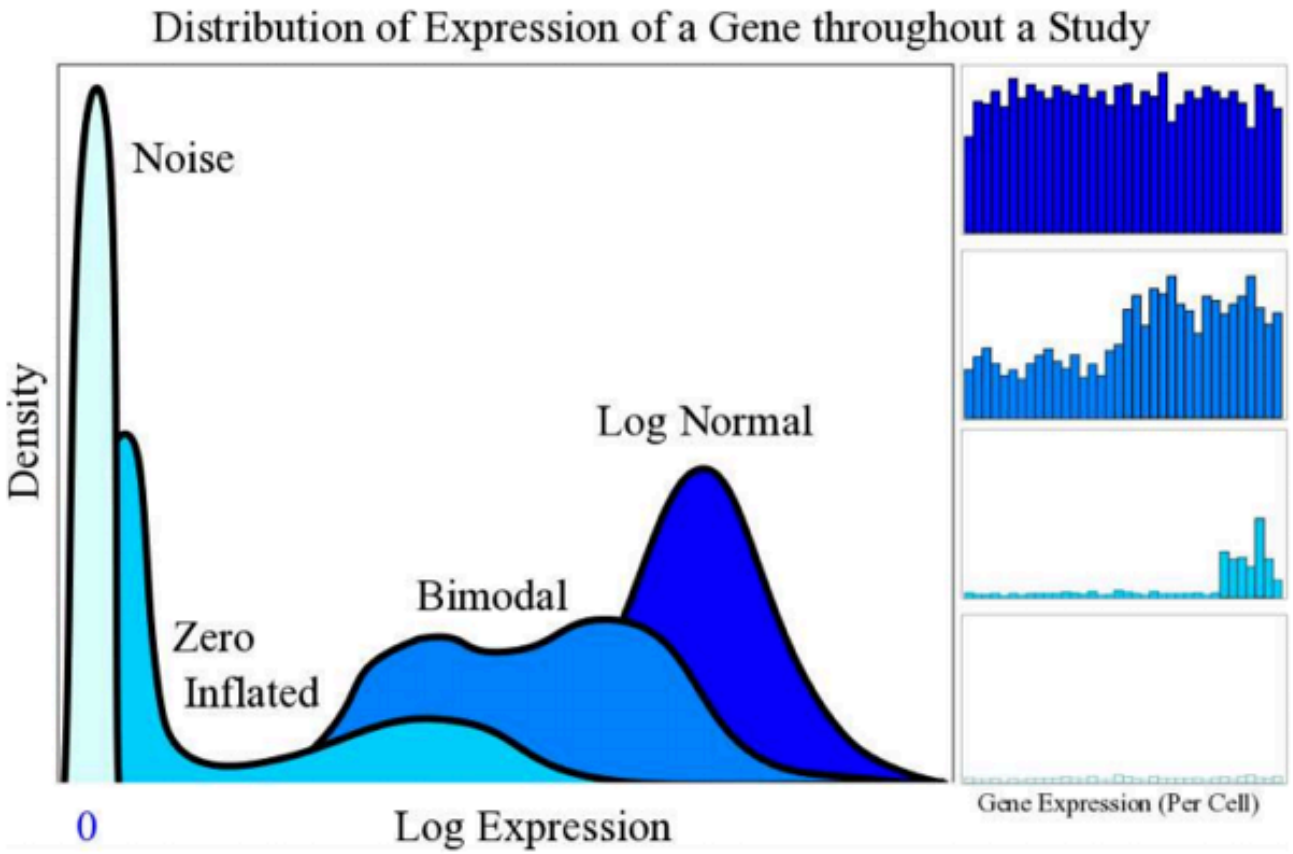
Genes have different distributions



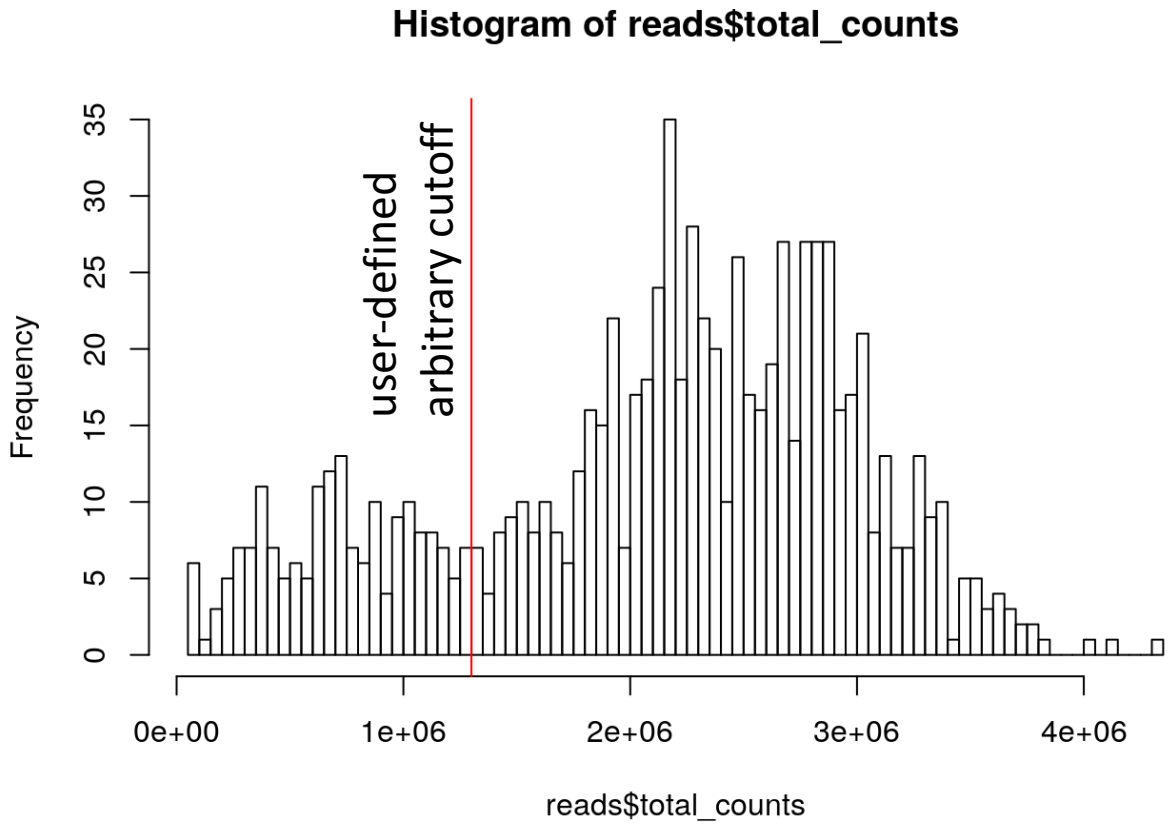




Genes have different distributions

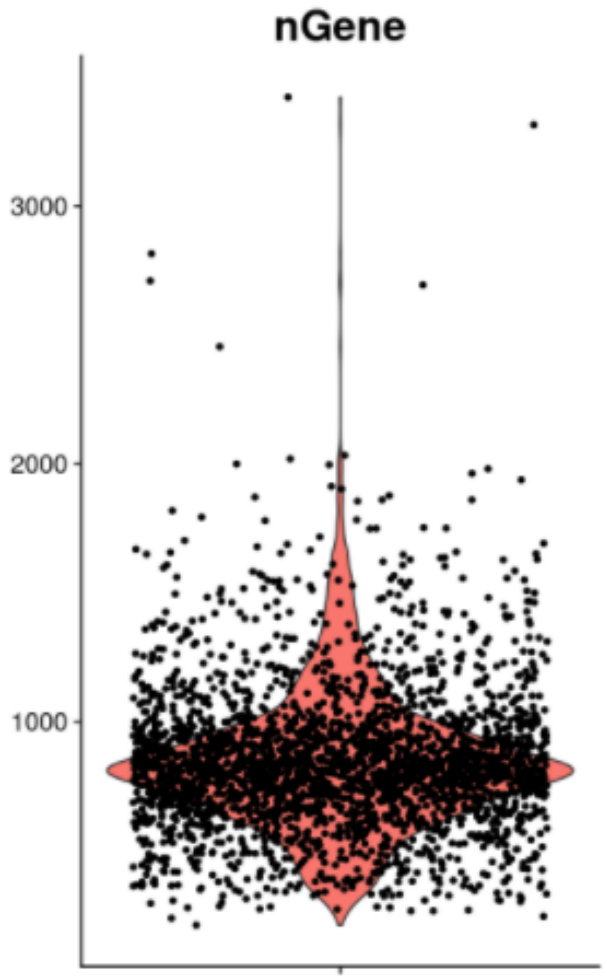


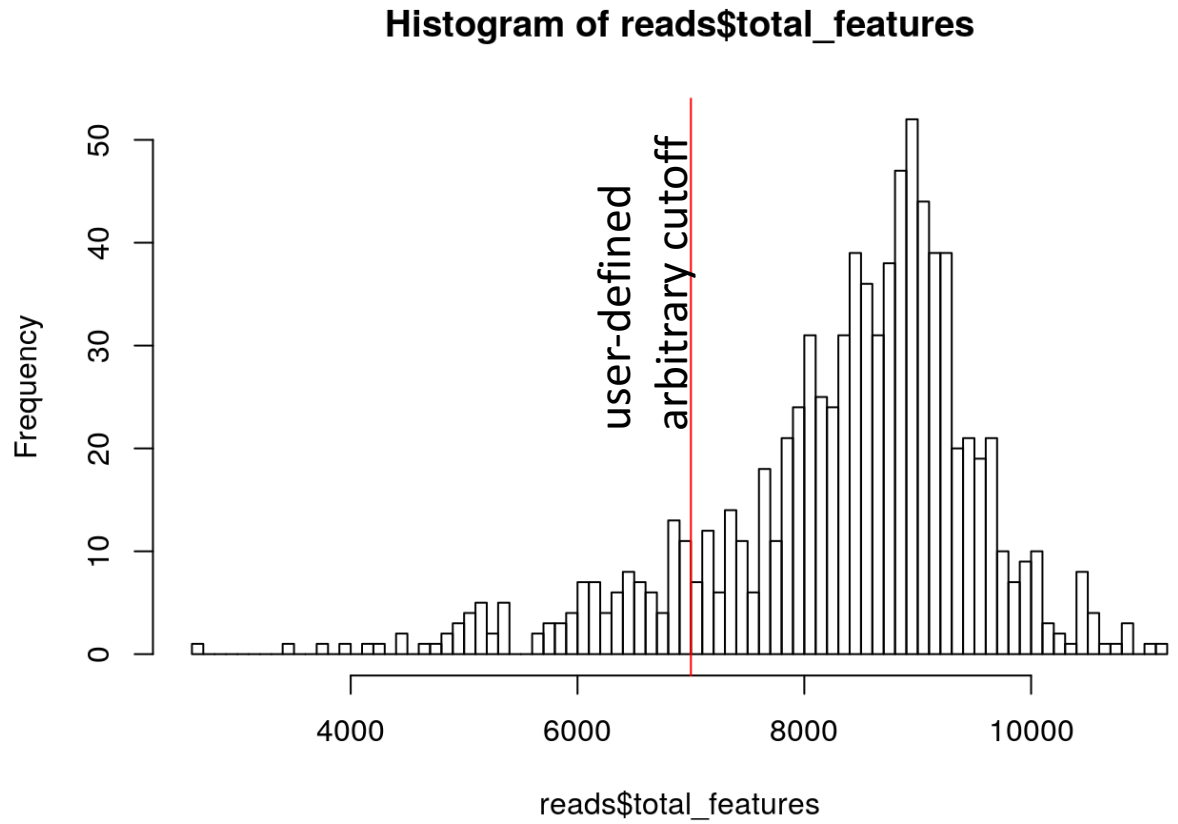
Filtering : Number of reads



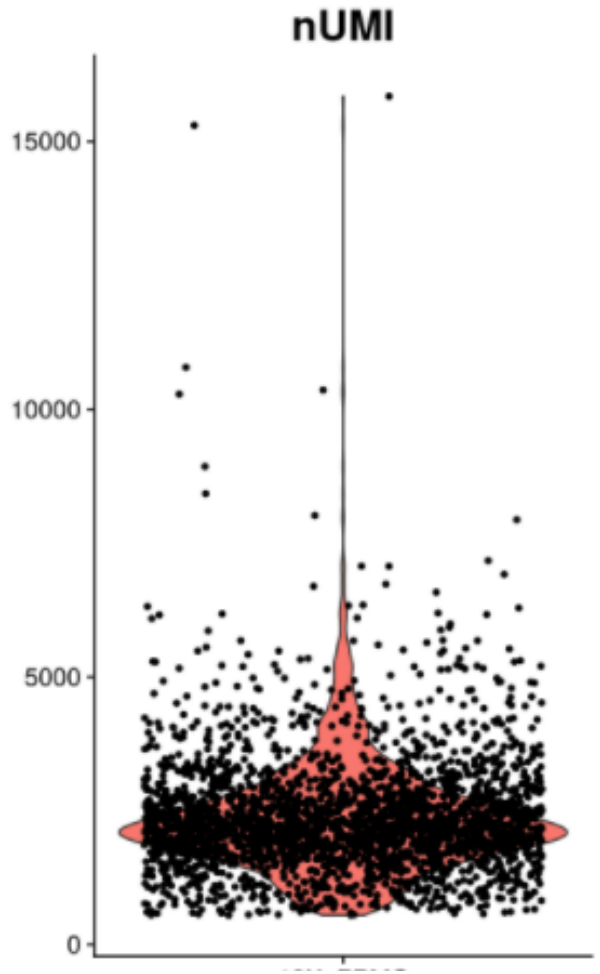
- Cell barcodes with low number of reads are excluded

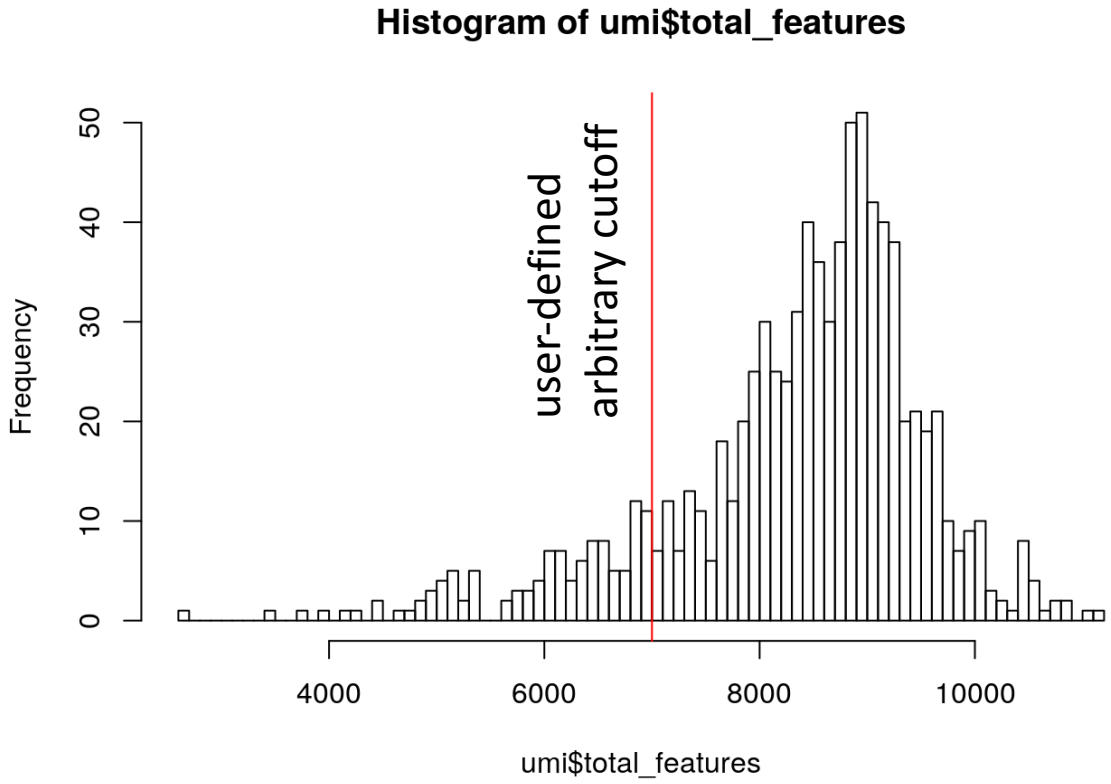
Filtering : Number of genes



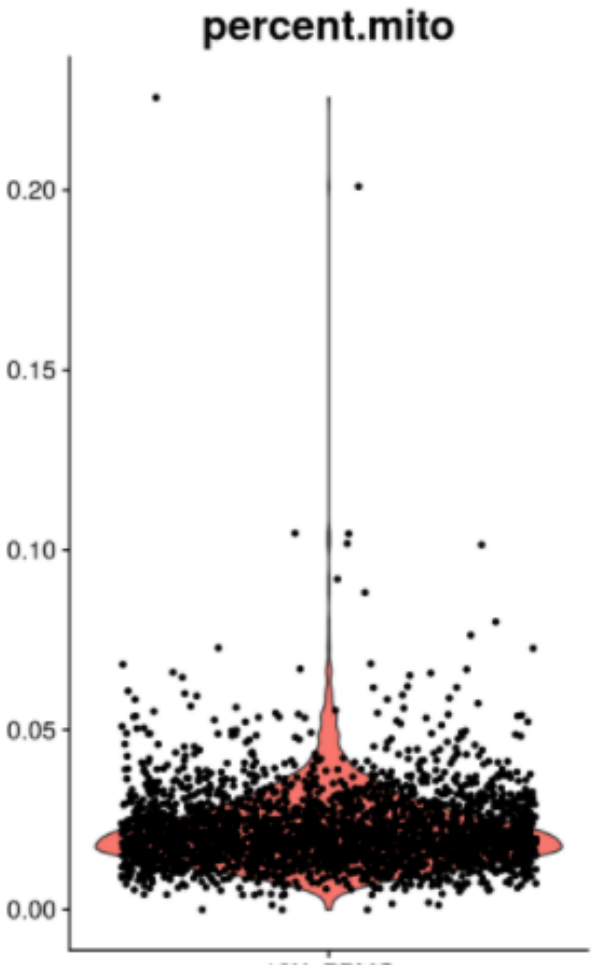


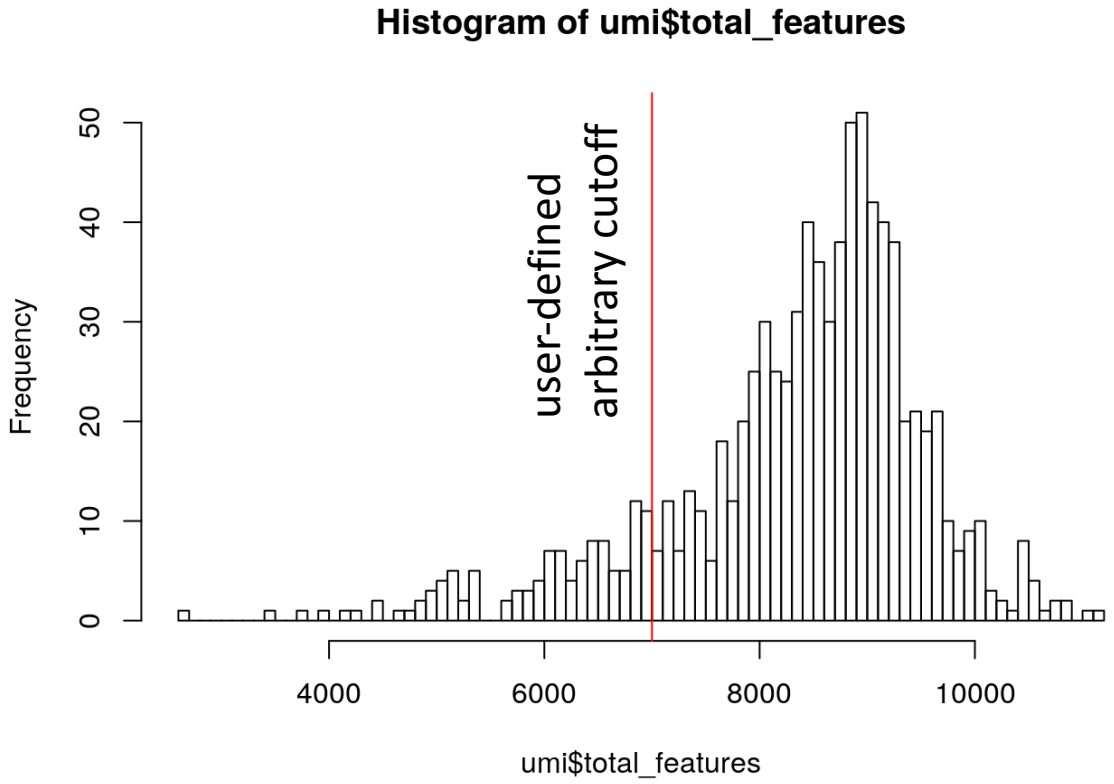
Filtering : Number of UMIs



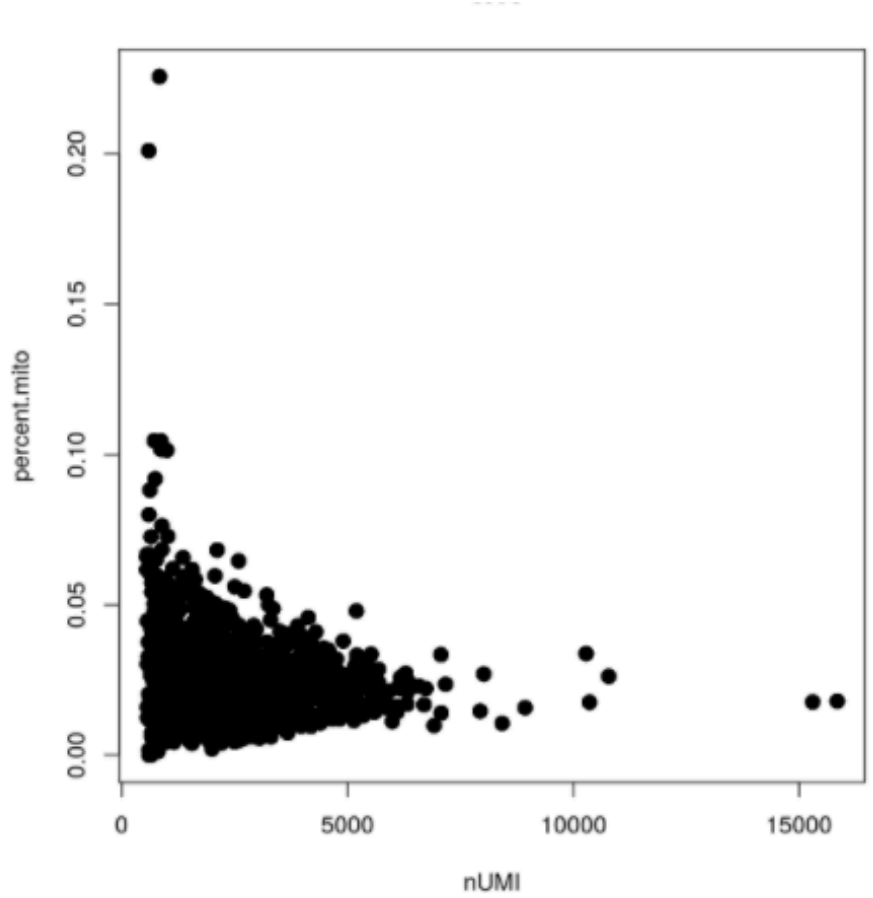


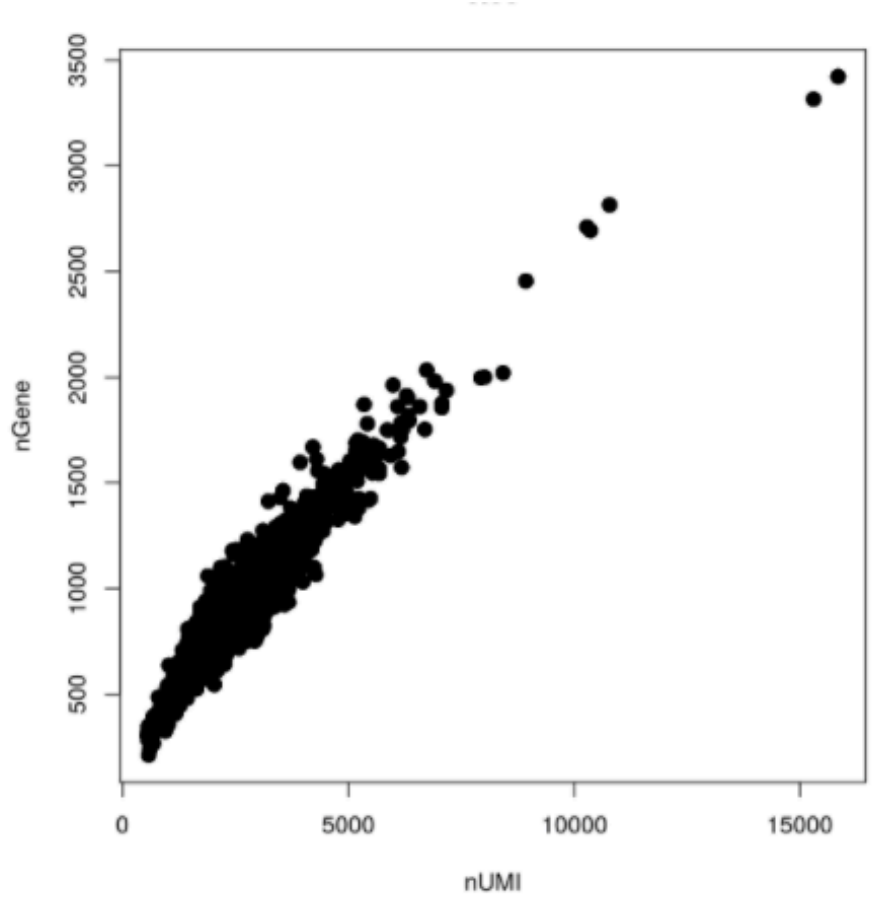
Filtering : Mitochondrial %





Filtering : Outliers are “multiple offenders”





Automatic Filtering

Celloline: A pipeline for mapping and quality assessment single cell RNA-seq data

cellity: Classification of low quality cells in scRNA-seq data using R

The `cellity` package contains functions to help to identify low quality cells in scRNA-seq data. It extracts biological and technical features from gene expression data that help to detect low quality cells.

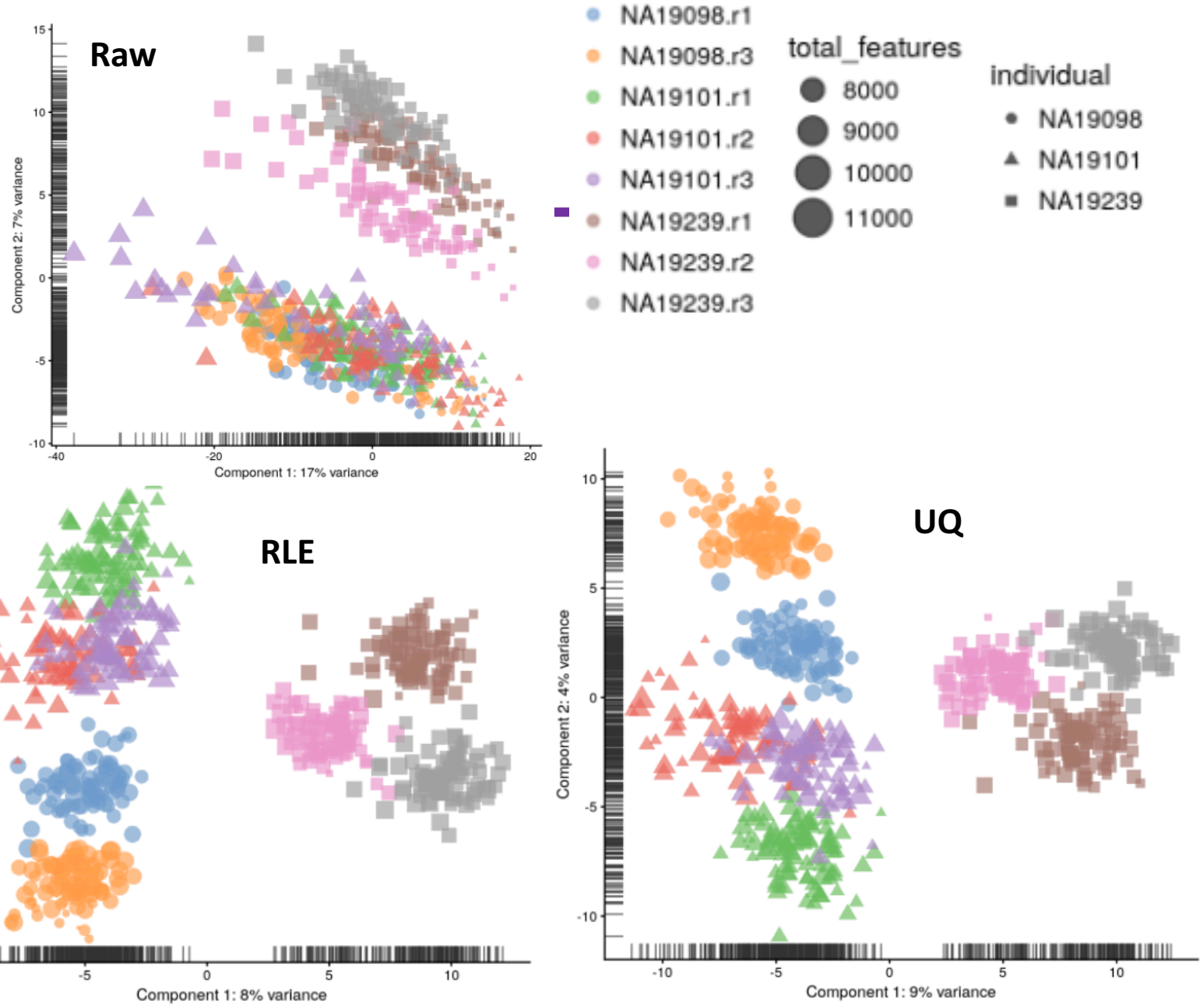
Normalization

- scRNASeq assays are prone to confounding batch effects, hence we normalize
- Many RNA-Seq normalization techniques can directly be applied
- Exploring the idea of combining or selecting from a collection of normalization or correction methods best for a specific case
- Some believe that UMI base analysis need not be normalized between samples as absolute count of molecules are reported

Normalization

- No. of reads per cell vary highly and library size normalization is required
- Methods like RSEM already incorporate library size and do not require normalization
- “normalization factor” is a multiplication factor, which is an estimate of the library size relative to the other cells
- TMM, UQ, RLE

Normalization

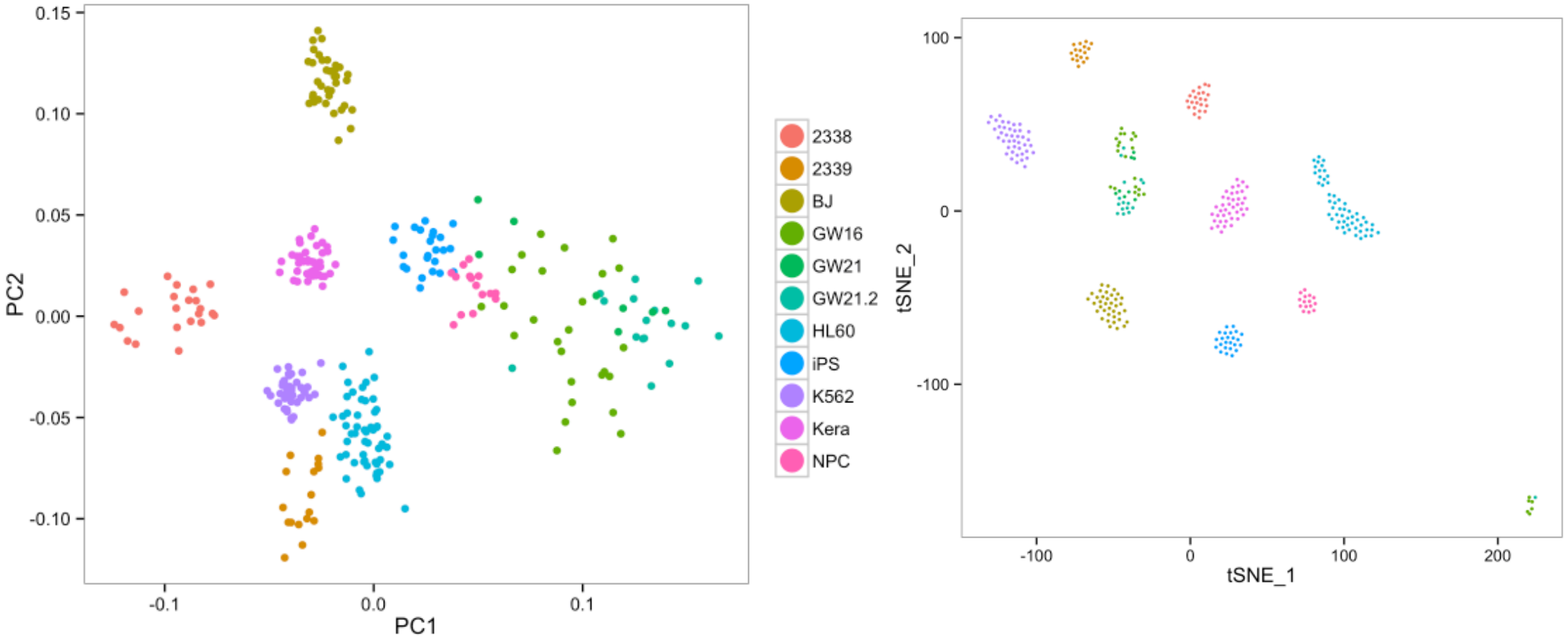


Visualization

- PCA (linear dimensionality reduction)
  - Small percentage of variance is explained by first 2 or 3 PCs

- t-SNE (non-linear dimensionality reduction) colored with k-means clustering
  - Not straight forward to determine k
  - k can be set to the number of significant PCs which can be determined by a jackstraw plot (computationally intensive)
  - Iterative process with gradual fine tuning

PCA vs. t-SNE


0.15
0.10
0.05
0.00
-0.05
-0.10
PC2
-0.1
0.0
0.1
PC1
2338
2339
BJ
GW16
GW21
GW21.2
HL60
iPS
K562
Kera
NPC
100
0
-100
tSNE_2
-100
0
100
200
tSNE_1

What is t-SNE?

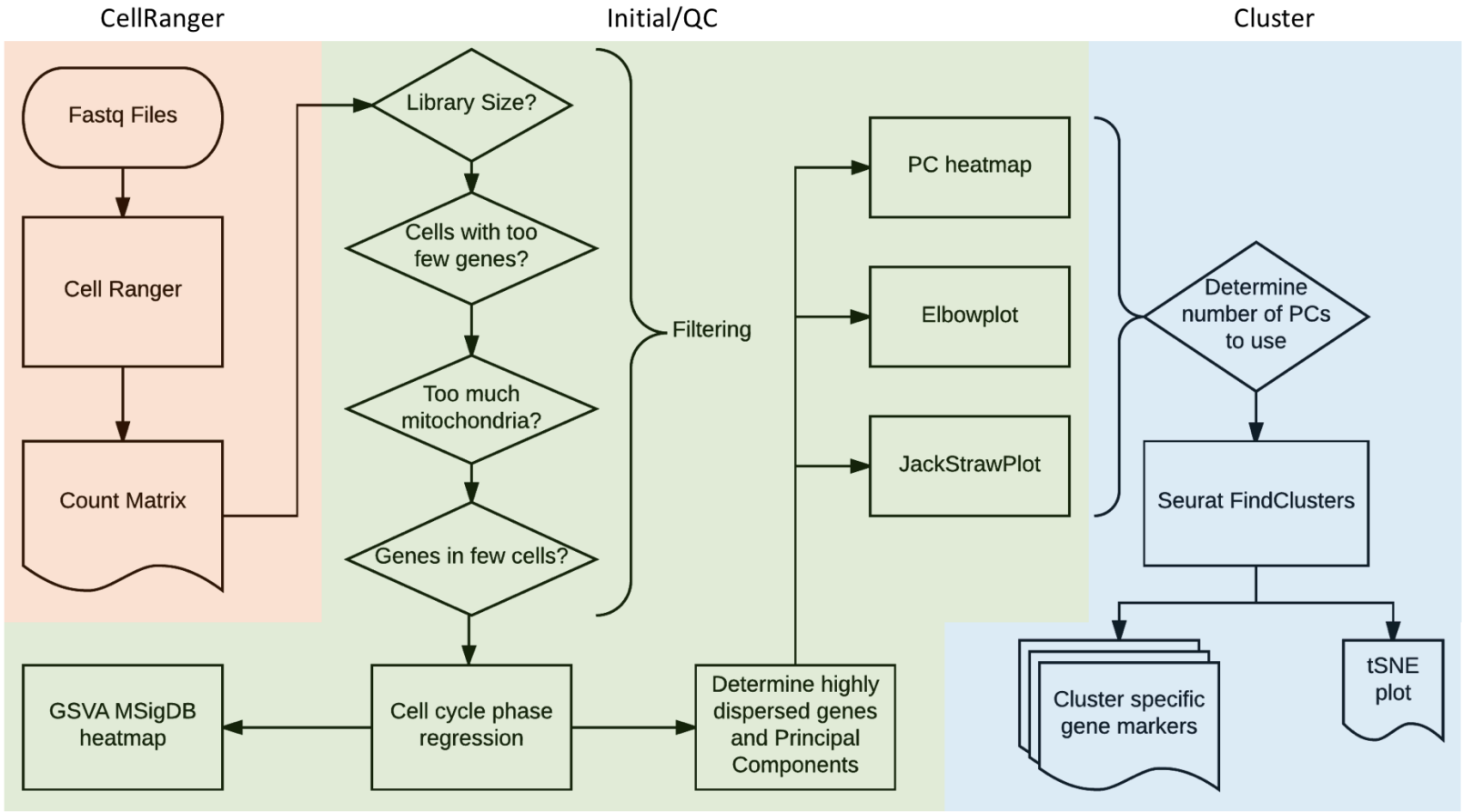
- t-Distributed Stochastic Neighbor Embedding
- “Dimensionality reduction”--> represent multi-dimensionality data in 2 or 3-D space
- PCA:
  - Linear
  - Unable to interpret complex polynomial relationship between features
  - Focuses on representing dissimilar points farther away in lower dimensions
- t-SNE:
  - Non-linear
  - Focuses on representing similar datapoints close together in lower dimensions
  - not capable of retaining both the local and global structure of the data at the same time

Things to remember about t-SNE

- Can arrive at different solutions depending on the initial “random” distributions… Hence, it is important to use the same “seed” every time
- Not capable of retaining both the local and global structure of the data at the same time
- One of the underlying algorithms used in the latest iPhone X Facial recognition software

Algorithm	FER accuracy
PCA	75.40%
LDA	75.90%
LLE	87.70%
SNE	90.60%
t-SNE	94.50%

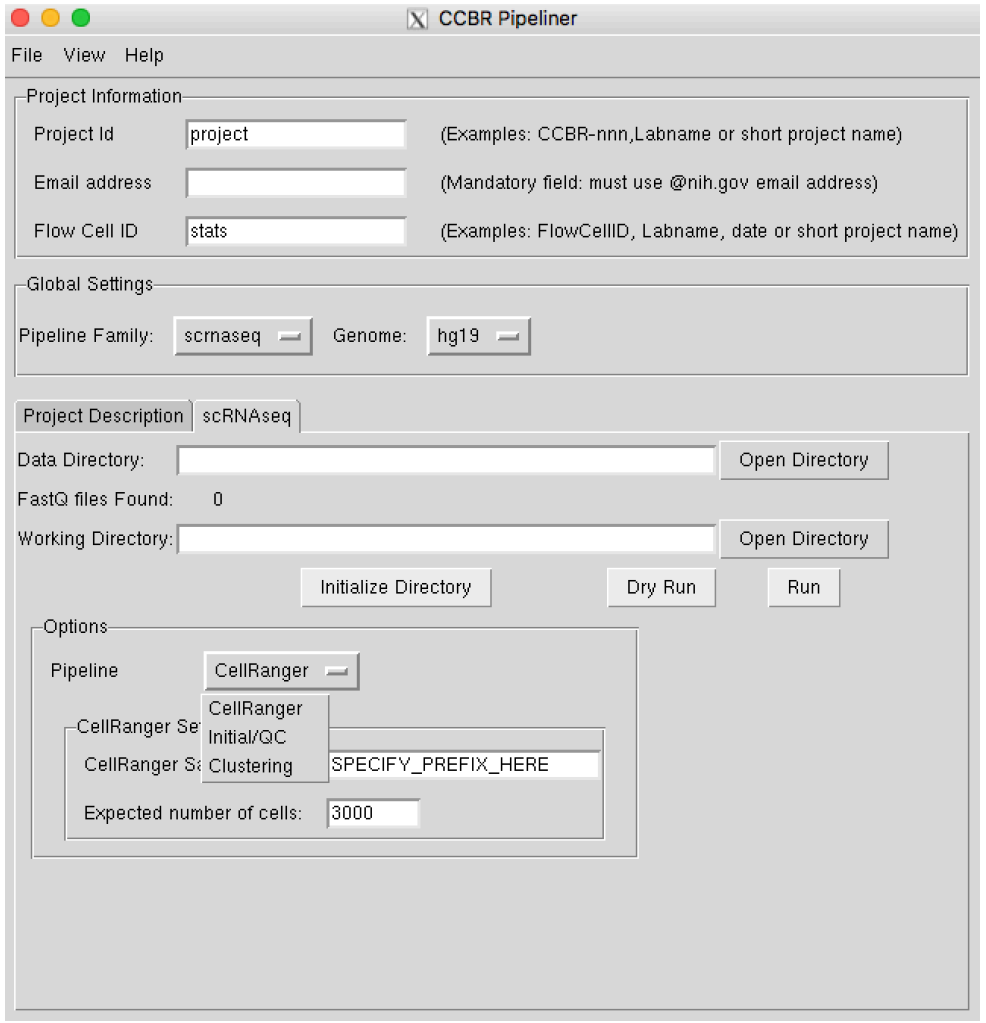
CCBR scRNASeq pipeline



Development: https://github.com/CCBR/scRNASeq

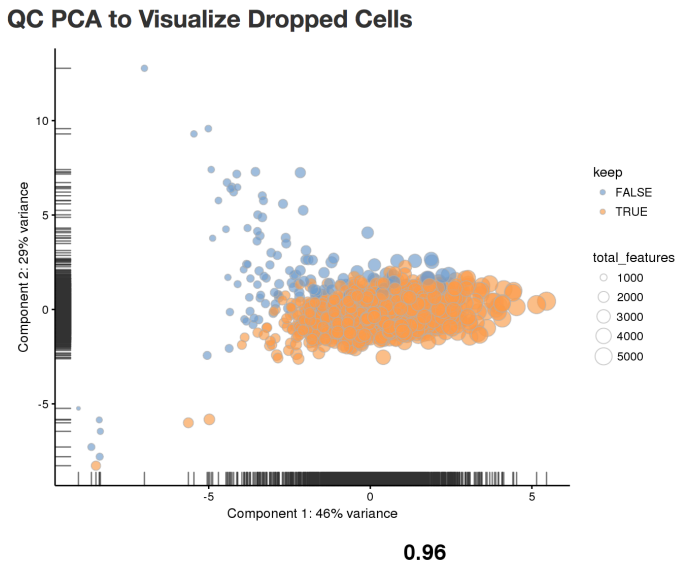
Production: https://github.com/CCBR/Pipeliner

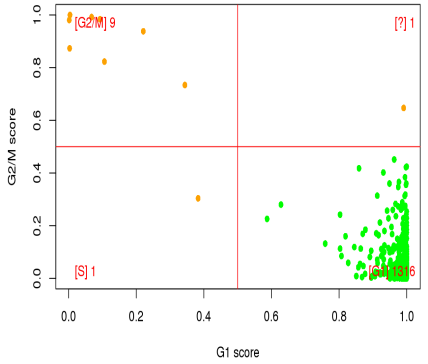
CCBR scRNASeq pipeline

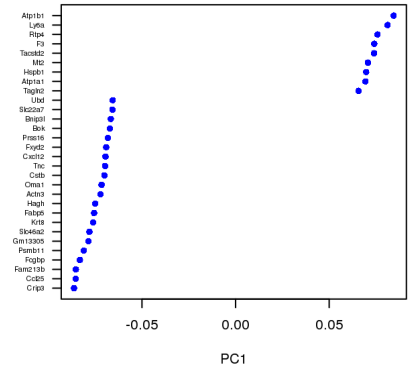


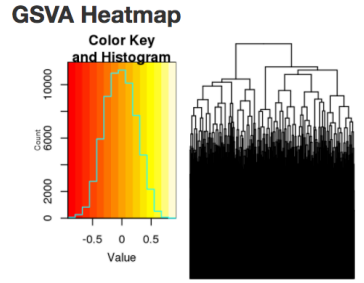
- Starting points
  - 10X genomics fastq
  - 10X genomics count matrix
- Data Filtering etc.
- Downstream analysis
  - k-means clustering
  - PCA
  - tSNE plot
  - marker gene lists
- All tasks are submitted as slurm jobs to biowulf

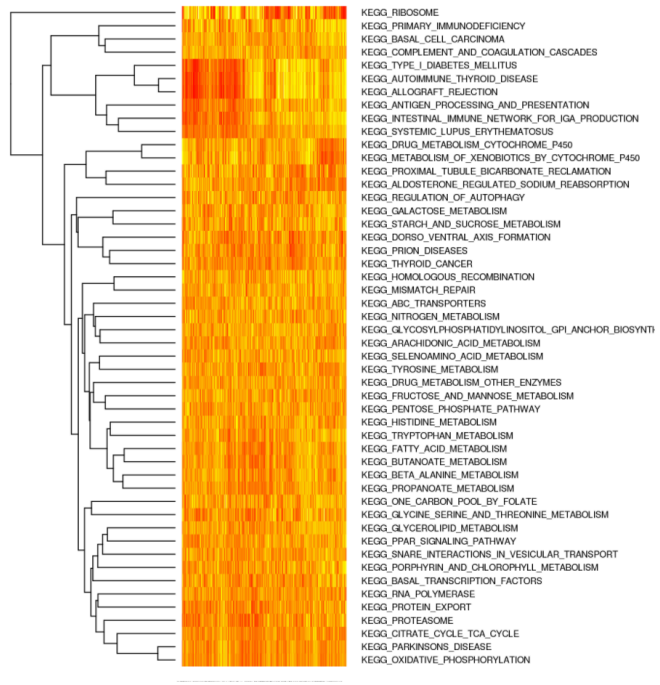
CCBR Pipeliner: example visualizations

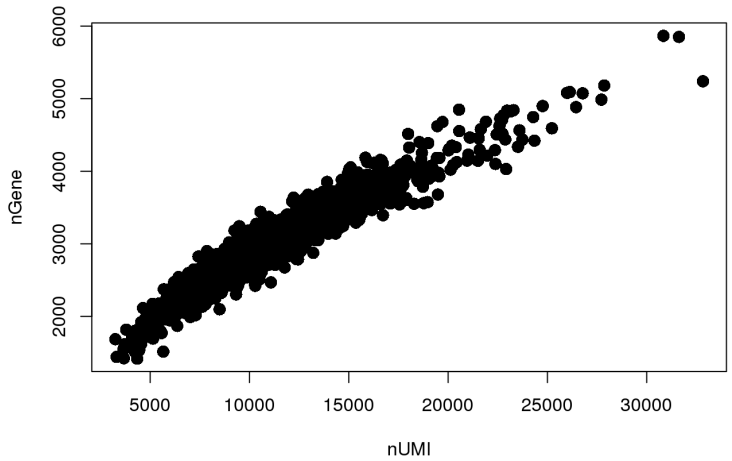


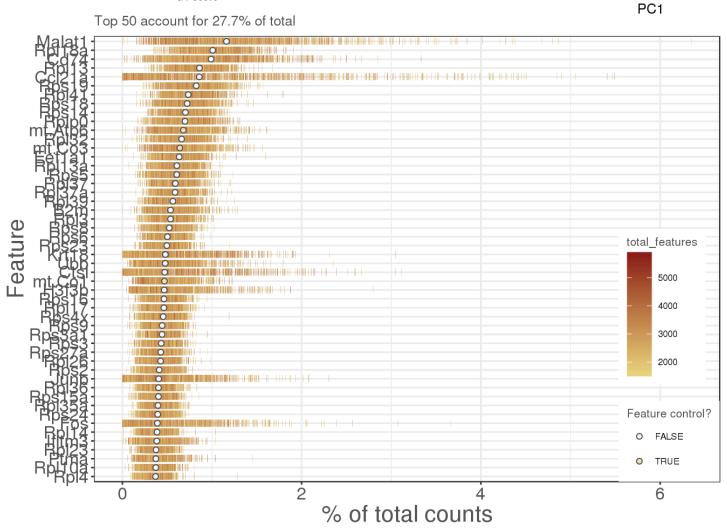












CCBR Pipeliner: What's in the "pipeline"?

- Future enhancements:
  - Comprehensive QC at a cellular level
  - Smooth merging of multiple datasets with minimal batch effect (newer version of Seurat)
  - Pseudo time analysis
  - Shiny app for interactive 3-D t-SNE for effective visualization which will allow pseudo coloring based on:
    - Cluster number
    - Gene
    - Genesets / Pathways

Take home message

- “scRNA Seq and its analysis” field is changing at a rapid pace
- Many of the things that you heard about scRNA Seq analysis today will be obsolete in a few months!

THANK YOU FOR YOUR ATTENTION

Contact Us

- E-mail
  - CCBR@mail.nih.gov
- Website
  - http://bioinformatics.cancer.gov
- Blg37/R3041