Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Overview of Single Cell RNA-seq Analysis Workflows and Pipelines

Yongmei Zhao
Bioinformatics Manager, CCR-SF Bioinformatics Group
Advanced Biomedical and Computational Sciences (ABCS)

BTEP Training, Oct 3, 2019

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute
The Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Training Objectives

Frederick National Laboratory for Cancer Research

1. Learn the current single cell analysis technologies and protocols and resources available at CCR.
2. Capture the basic experimental design considerations
3. Get overview of single cell analysis workflow and software tools
4. Understand the typical scRNA-seq analysis methods and algorithms for:
    - Preprocessing and quality control
    - Normalization and batch correction
    - Feature selection and dimension reduction
    - Clustering analysis
    - Differential analysis between subpopulations or sample conditions
    - Result visualization
5. Get overview of downstream analysis software tools and algorithms for:
    - marker gene identification
    - cell type annotation
    - Trajectory analysis
    - multi-modal data integration analysis

scRNA-seq Data Analysis Challenges

Frederick National Laboratory for Cancer Research

- **scRNA-Seq datasets present a very high level of noise than bulk RNA-seq.**
  - This could be due to technical noises, amplification biases, batch effects, or even biological sources of variations (such as cell cycle state).
- **Gene expressions are highly variable and sparsity, large fraction of so-called 'dropout' events.**
  - Subpopulations of cells or transient states where a gene is not expressed; or technical reasons where a gene is expressed but not detected through sequencing.
- **RNA-seq experiments are inherently stochastic, technical variability are commonly observed in libraries.**
  - Sometimes the technical variability and confounding factors mask biological differences.
- **The methods developed for bulk RNA-seq may not be reused for scRNA-seq. methods for scRNA analysis are emerging and changing fast.**
  - There are more than 480 scRNA analysis software tools available as today, but yet still need well established benchmarking standard.
- **Large number of single-cells poses a large spectrum of challenges from developing more efficient aligners and clustering algorithms for efficient computing and data storage.**

Main Sources of Bias in scRNA-seq Experiments

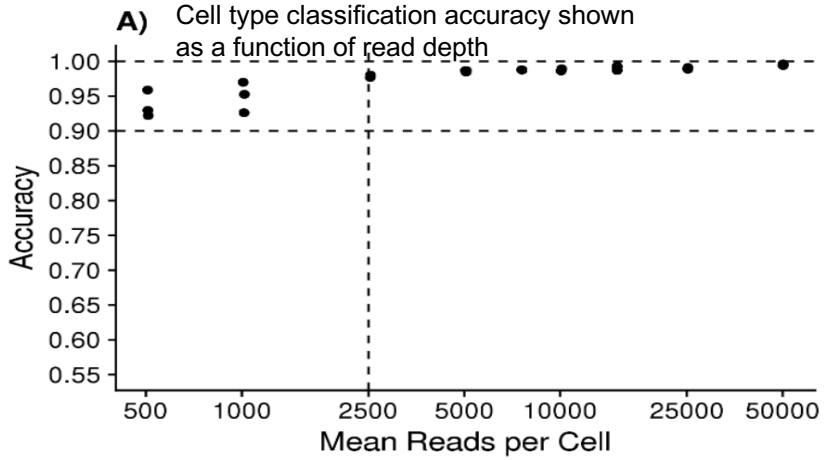
Main sources of bias in scRNA-seq experiments and solutions for limiting their impact

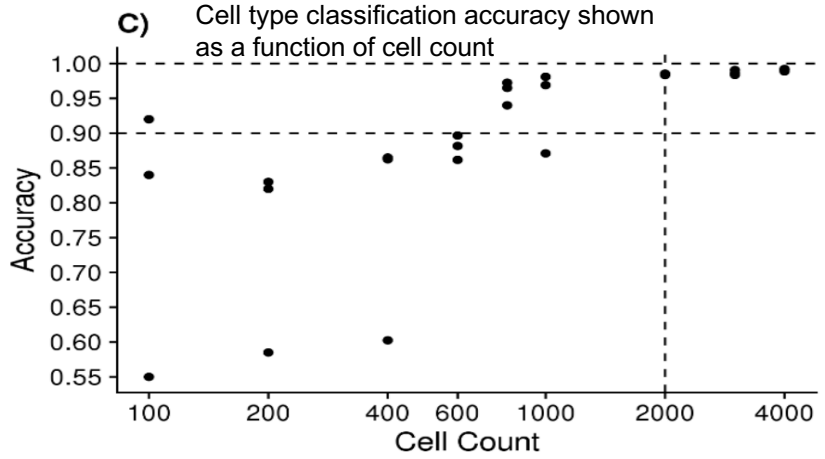
Source of bias	Type	Effect	Current solutions
RNA capture and RT efficiency	Technical	Stochastic zeroes	Spike-ins, statistical modelling
cDNA amplification	Technical	Loss of quantification accuracy	UMIs, statistical modelling
Batch effects	Technical	Introduce a signal different from the true biological signal	Statistical modelling
HVGs, transcriptional burst	Biological	Increase variance in the data	Statistical modelling
Cell-cycle stage, differentiation state, etc.	Biological	Confuse the true biological signal	Cell visualization, statistical modelling

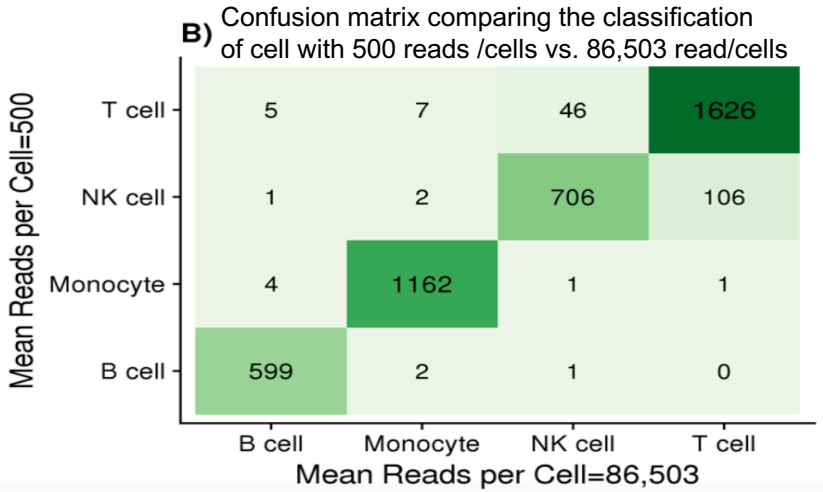
Bacher & Kendziorski, Briefings in Bioinformatics (2018)

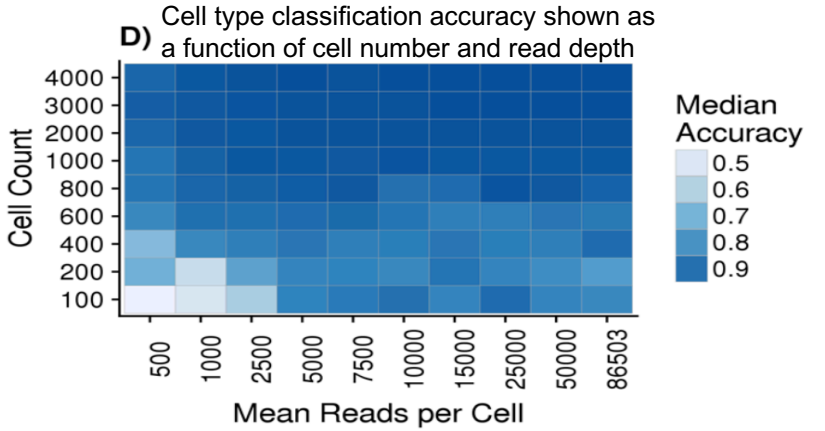
Experimental Design Considerations

- Determine the appropriate **library protocol, cell numbers, sequencing depth** of the experiment according to the project goals, sample quantity and qualities as well as platform selection.
- Cell clustering & cell-type identification benefits from large number of cells and doesn't require high sequencing depth
- Highly heterogeneity cell populations, transcription factor detection (regulatory networks) require high read depth and most sensitive protocols





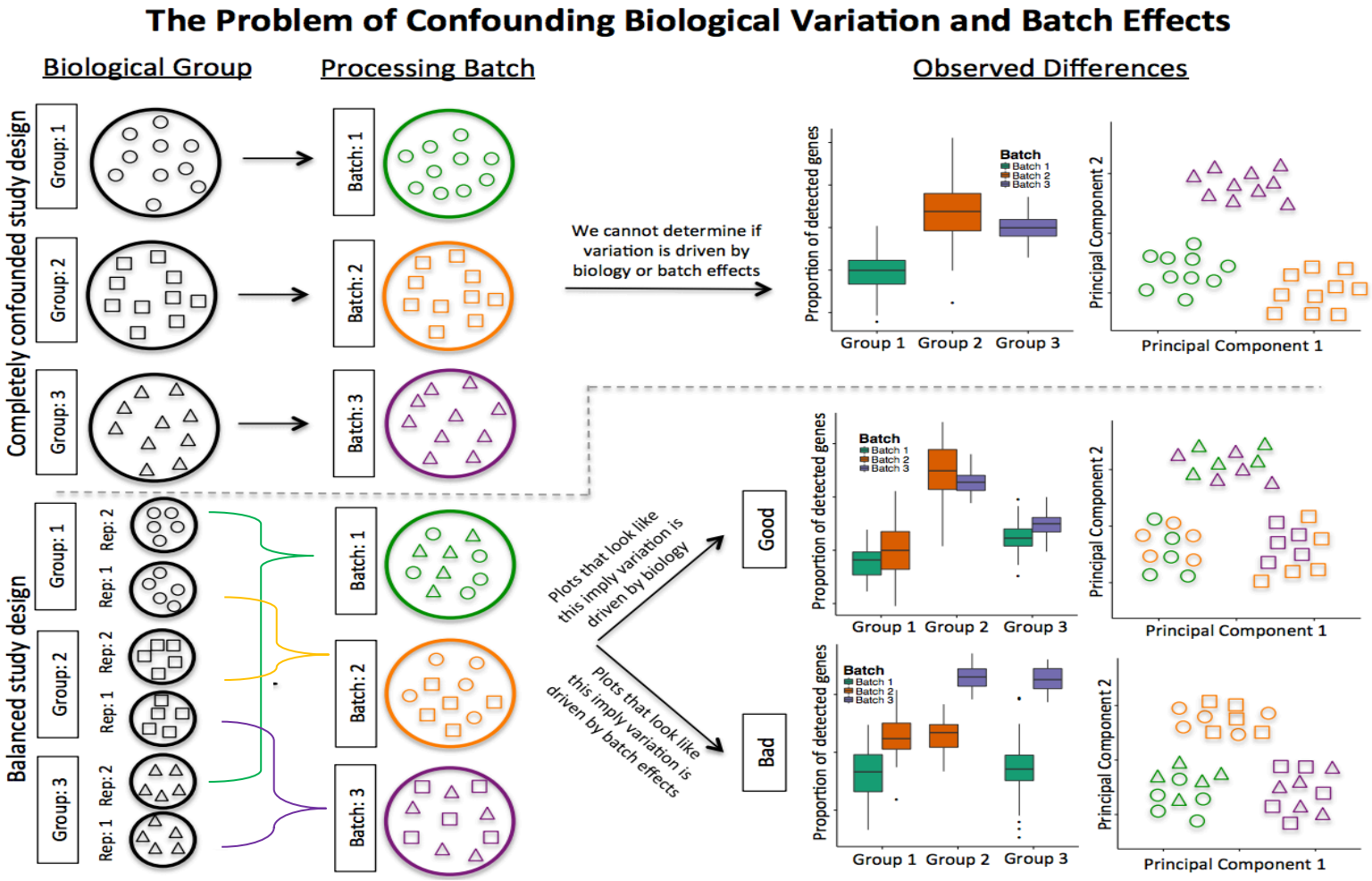




10x Genomics® | CG000148 Rev A Technical Note
https://www.10xgenomics.com

Experimental Design Considerations

- Avoid confounding biological and batch effects, consider multiple conditions are captured in the same chip if it is possible
- Include multiple biological replicates for each condition where replicates for different conditions should be performed together if it is possible
- Spike-ins may be useful for quality control and normalizing read counts, but may exhibit higher noise due to pipetting errors,. Spike-in should mimic endogenous RNA and not interfere with the measurement of endogenous RNA



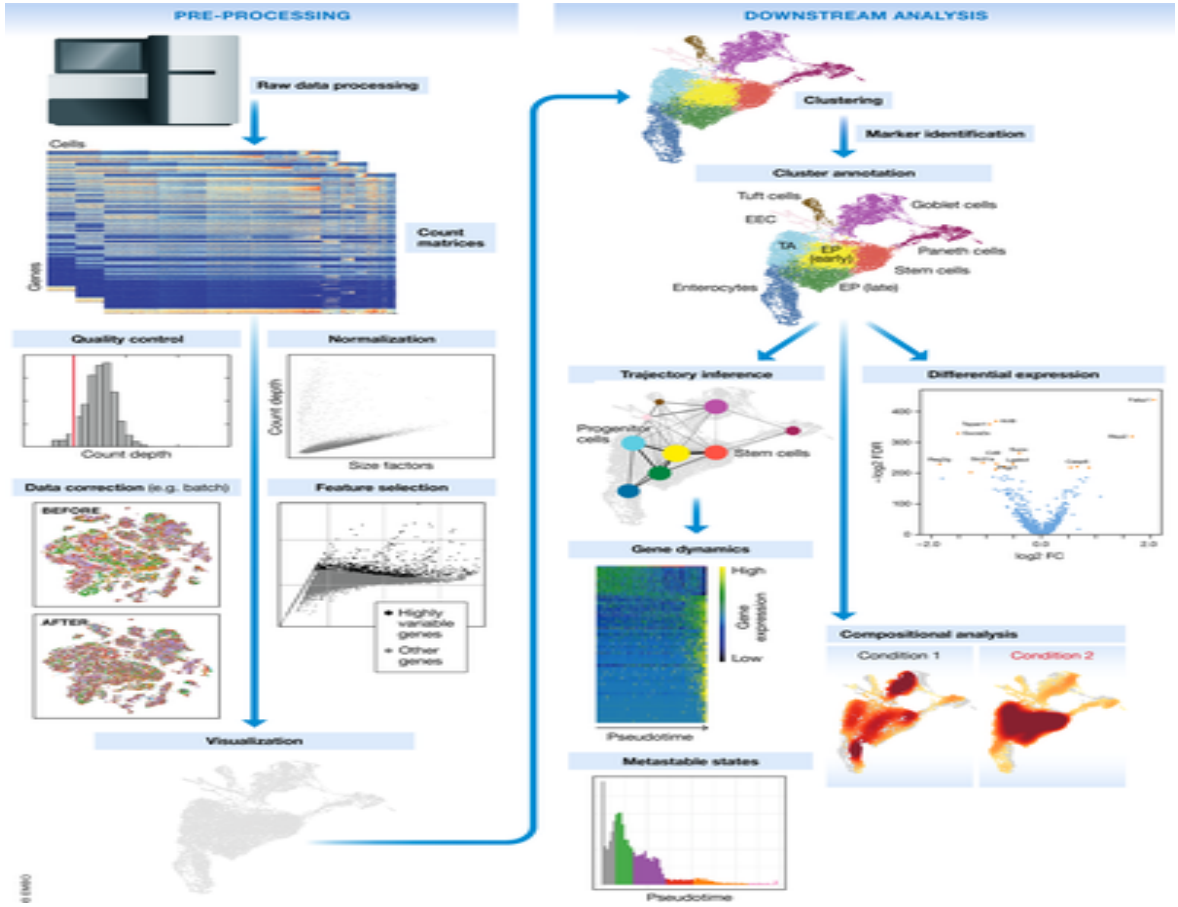
scRNA-seq Analysis Workflow and Software Tools

Frederick National Laboratory for Cancer Research

- **Upstream Analysis**
  - Preprocessing, raw count matrix generation and quality control.
  - Normalization, and noise reduction, Imputation, batch effect removal.
  - Feature selection, dimension reduction, clustering and visualization
- **Downstream Analysis**

  Cell level
  - Clustering,
  - Subpopulation identification and cell type classification
  - Pseudo-time/trajectory analysis

  Gene level
  - DE analysis and multi-group comparative analysis
  - Gene set analysis and gene network inference
- **Multi-assay data Integration**



Luecken,M., et al. Current best practices in single-cell RNA-seq analysis: a tutorial, Mol Syst Biol (2019)15:e8746

Which Method and Tool to Choose?

Frederick National Laboratory for Cancer Research

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

awesome-single-cell

List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc. Contributions welcome...

Citation

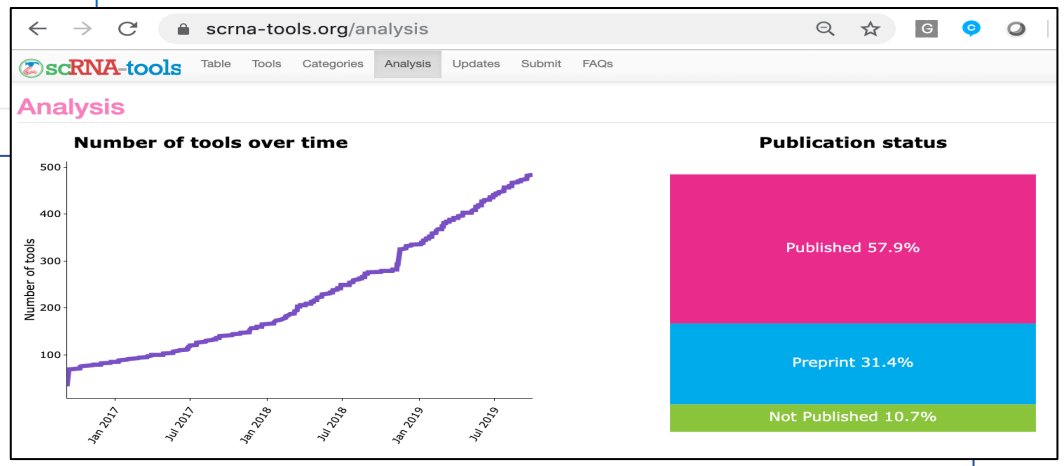
DOI 10.5281/zenodo.1117762

Software packages

RNA-seq

- anchor - [Python] - ⚓ Find bimo
- ascend - [R] - ascend is an R pacl
  statistical challenges of single cel
  a flexible framework to perform fi
  expression and a wide-range of p

487 scRNA-seq-tools in current database



scRNA-seq Analysis Workflow and Software Tools

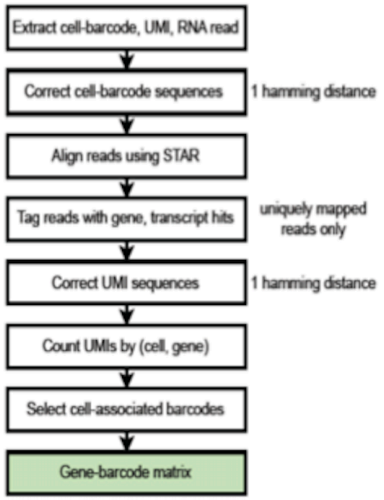
Frederick National Laboratory for Cancer Research

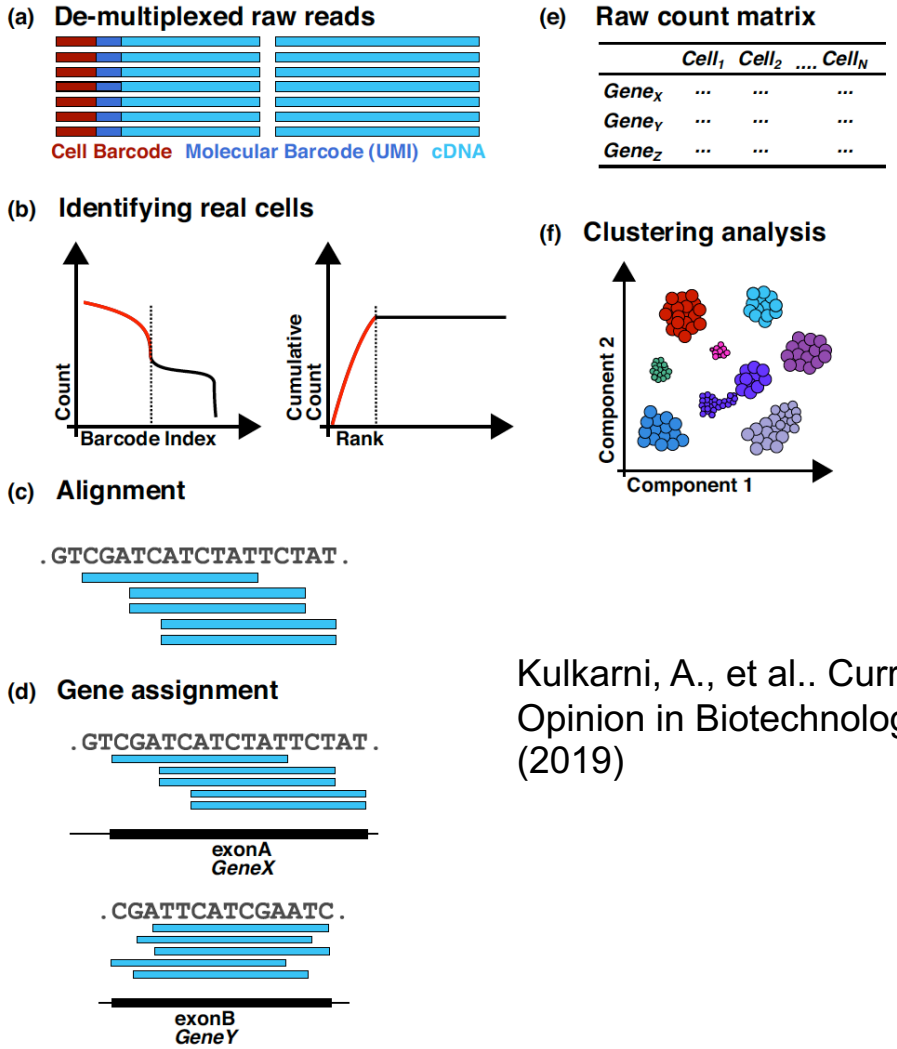
- **Upstream Analysis**
  - **Preprocessing, generate raw count matrix and perform quality control.**

Software tools:

- Cell Ranger pipeline – 10x genomics protocol
- zUMIs - is compatible with nearly all UMI based as well as non-UMI based protocols scRNA-seq protocols
- Scater - quality control and visualization
- DoubletFinder/ Scrublet– find doublets
- Seurat/Scran – cell cycle identification and regression
- Salmon/Alevin – quantification and create cell gene count matrix
- dropSeqPipe – support Drop-seq, 10x, DroNc-seq, SCRB-seq. snakemake pipeline for count matrix and QC

Cell ranger pipeline





Kulkarni, A., et al.. Current Opinion in Biotechnology (2019)

scRNA-seq Analysis Workflow and Software Tools

Frederick National Laboratory for Cancer Research

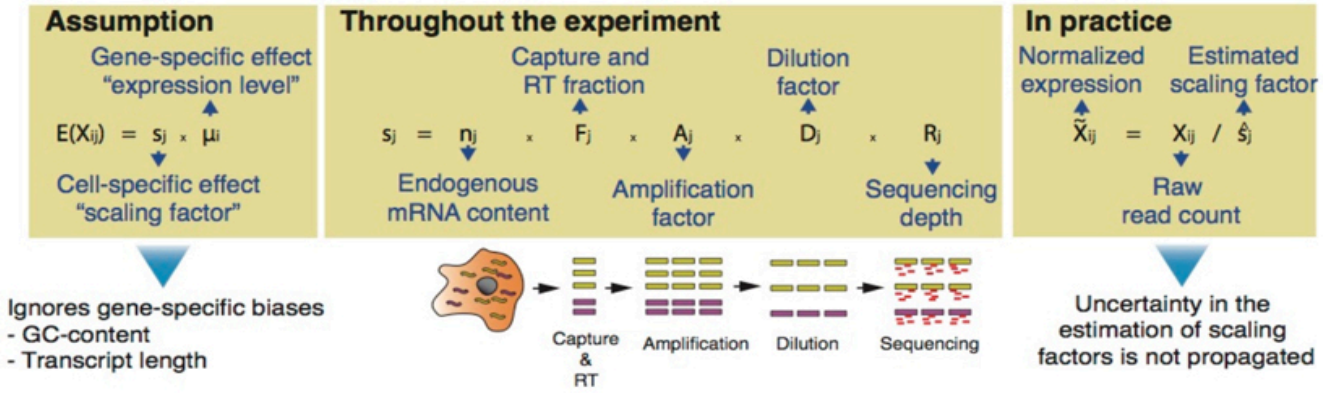
Upstream Analysis

- **Normalization, Noise reduction and batch correction**
- **Feature selection, dimension reduction, clustering, visualization**

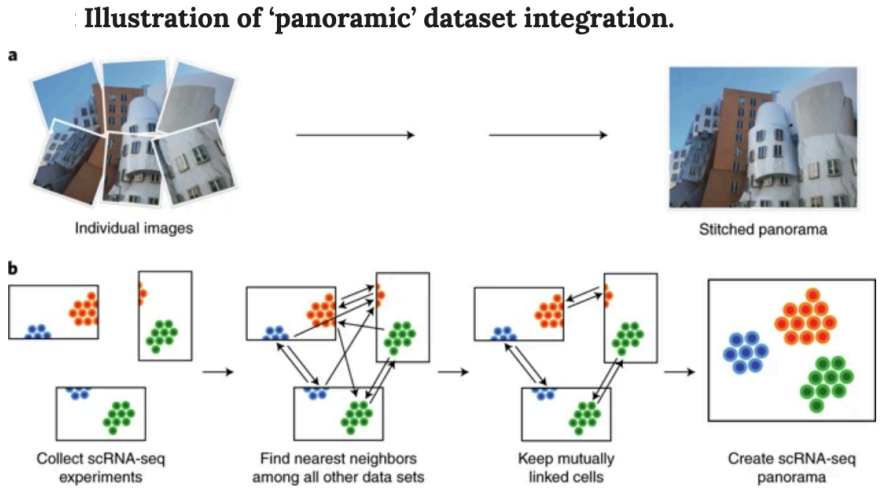
Software tools:

- Seurat – normalization, remove unwanted sources of variation, linear dimensional reduction, clustering analysis. SCTransform
- Scran – library size normalization, assignment of cell cycle phase, find HVGs, batch correction
  Scater – quality control, normalization and quantification, dimension reduction, visualization
- Cellranger – read depth sub-sampling
- DCA/ALRA – Imputation of missing values
- MNN / Harmony / scanorama – batch effect correction
- PCA/t-SNE / UMAP – Dimension reduction and visualization

Global scaling normalization, and batch effect correction for scRNA-seq datasets, aims to remove technical effect but retain biological variation



Vallejos, C.A., et al. *Nature methods* vol. 14,6 (2017)



- Searches nearest neighbors to identify shared cell types among all pairs of datasets.
- Accurately integrates heterogeneous collections of scRNA-seq data.

Hie,B et. al., Nature biotechnology (2019)

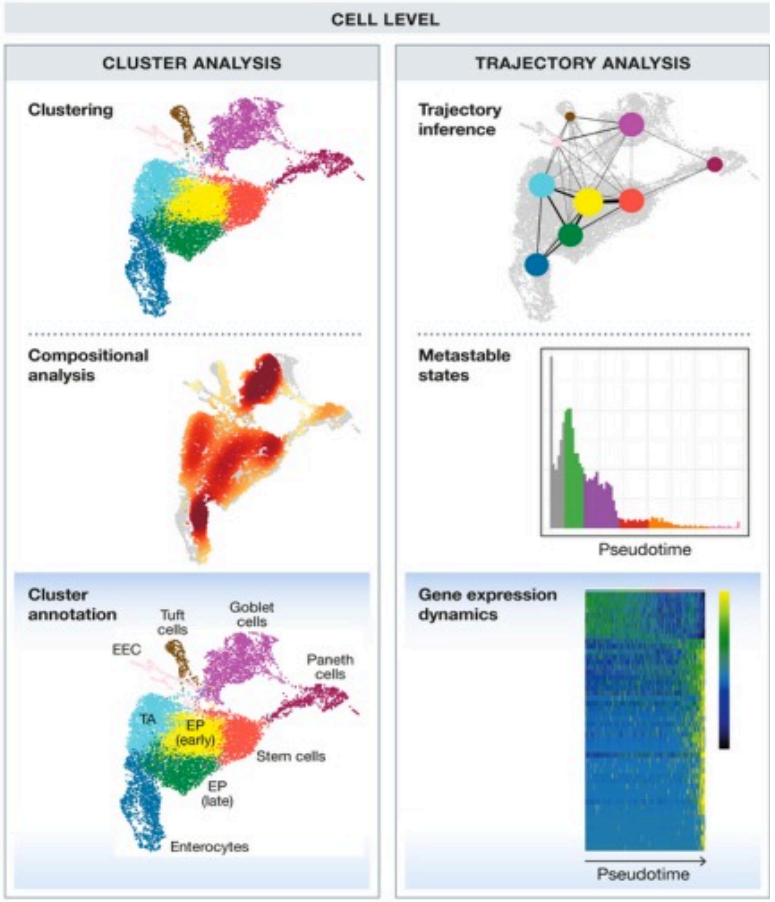
scRNA-seq Analysis Workflow and Software Tools

Frederick National Laboratory for Cancer Research

- **Downstream Analysis**
  - **Subpopulation identification and cell type classification**
  - **Trajectory analysis/ RNA velocity**

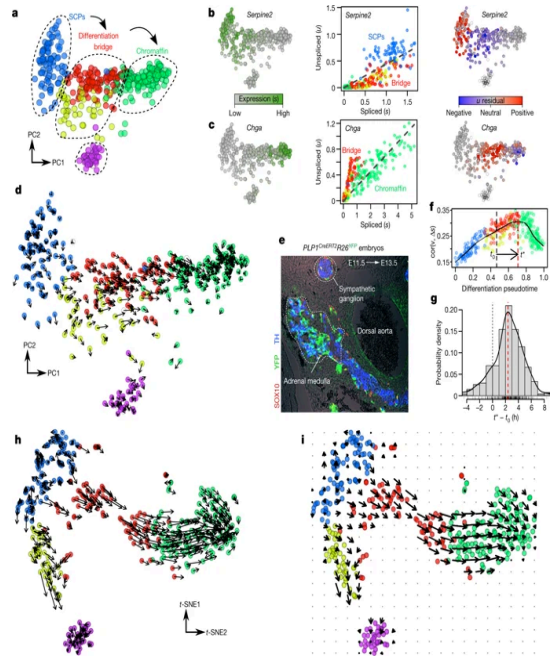
Software tools:

- Seurat –clustering, cell maker gene identification and visualization
- SingleR –cell type classification.
- Scanpy – python based gene expression analysis for single cell package
- Monocole 2/3 – clustering, classifying and counting cells, DE analysis. trajectory analysis, and track cells change over pseudo time.
- RNA Velocity – distinguishing between unspliced and spliced mRNAs to predict the future states of individual cells; analysis of developmental lineage and cellular dynamics



Luecken,M., et al. Current best practices in single-cell RNA-seq analysis: a tutorial, Mol Syst Biol (2019)

RNA velocity recapitulates dynamics of chromaffin cell differentiation.



Manno, G., et al. RNA velocity of single cells, Nature (2018)

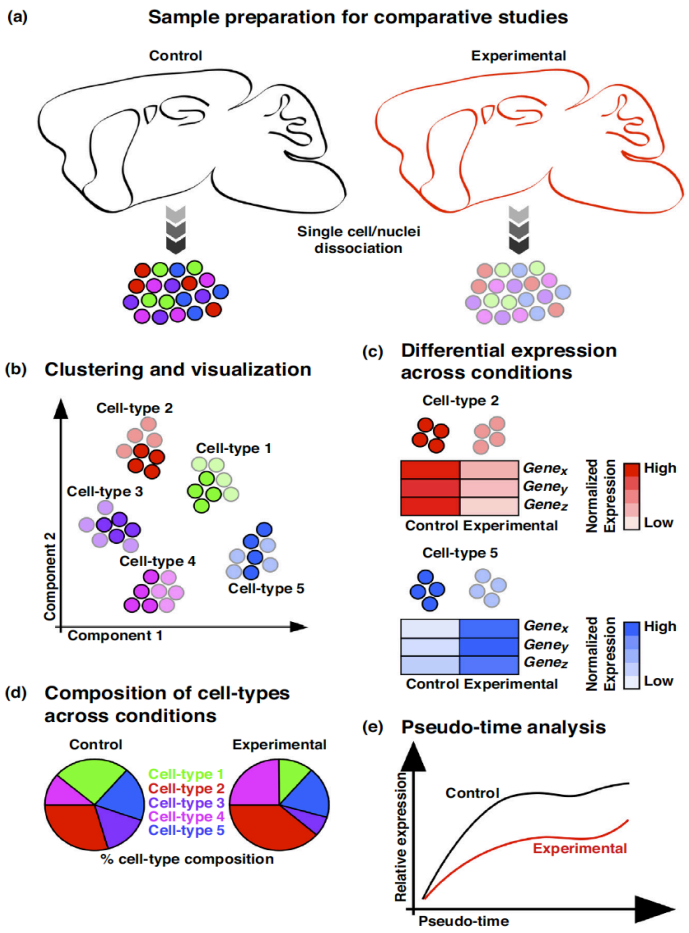
scRNA-seq Analysis Workflow and Software Tools

Frederick National Laboratory for Cancer Research

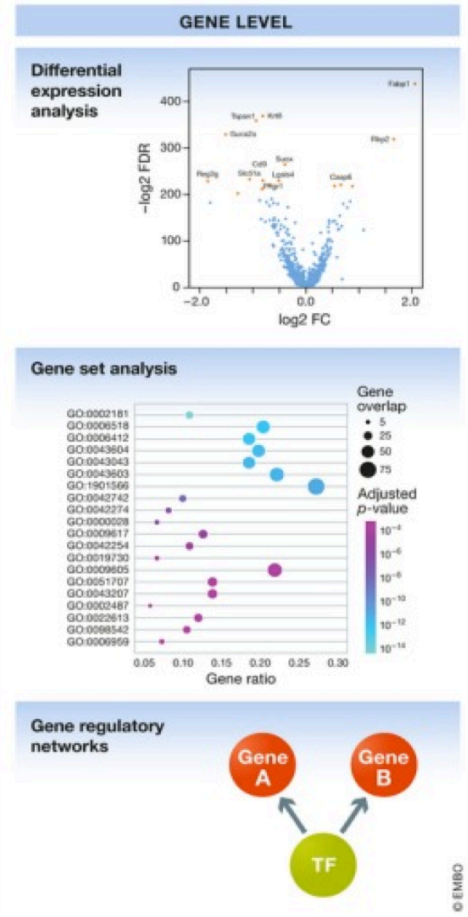
- **Downstream Analysis**
  - **DE analysis and multi-sample comparative analysis**
  - **GSA and gene regulation network inference**

Software tools:

- Seurat – normalization, integrate across conditions, identify common cell types and markers among samples.
- MAST – gene expression and differential analysis
- ClusterProfiler – analyze and visualize functional profiles (GO and KEGG) of gene and gene clusters.
- Sc. ClustViz – a shiny app which allows interactive visualization to explore the clustered data
- Partek Flow for single cell – application for interactive analysis
- Loupe browser - cluster comparison and marker gene identification



Kulkarni, A. et al., Current Opinion in Biotechnology 2019



Luecken,M., et al. Mol Syst Biol (2019)

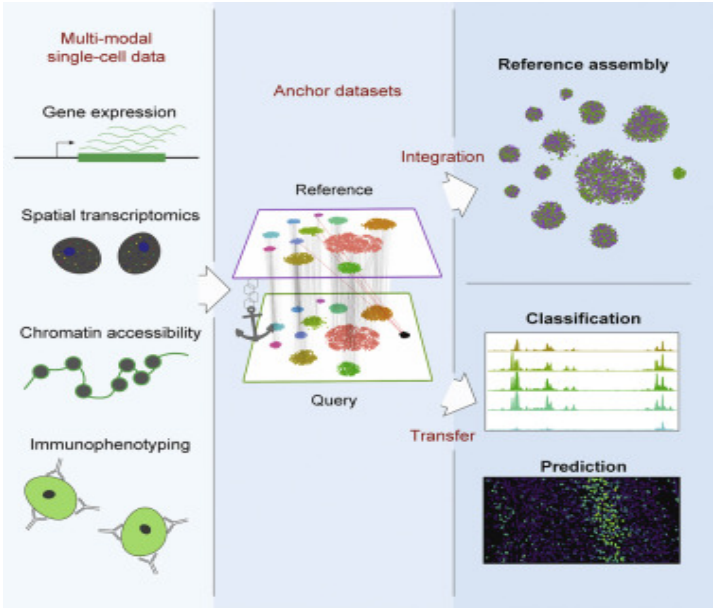
scRNA-seq Analysis Workflow and Software Tools

Frederick National Laboratory for Cancer Research

- **Downstream Analysis**
  - **Multi-assay data Integration Analysis**

Software tools:

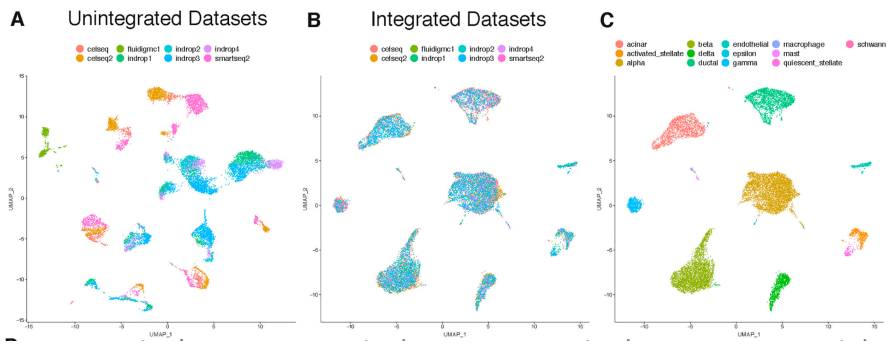
- Seurat v3 – Integration and label transfer. Have standard workflow, SCTransform, reference-based, or reciprocal PCA for multi-modal data integration
- MNN/Scanorama and scMerge- batch effect correction, harmonize datasets for integration analysis
- Loupe browser – visualize feature barcode and gene expression analysis together together or combine VDJ and gene expression analysis result together for visualization.



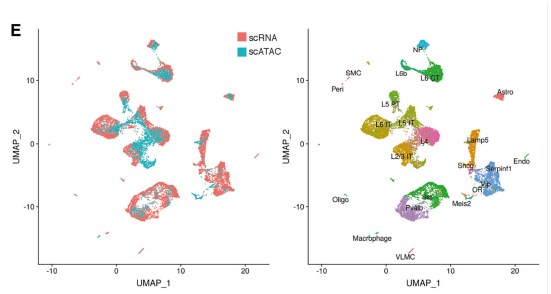
Sthart, Butler*, et al. Cell 2019*

- Seurat v3 identifies correspondences between cells in different experiments
- These “anchors” can be used to harmonize datasets into a single reference
- Reference labels and data can be projected onto query datasets
- Extends beyond RNA-seq to single-cell protein, chromatin, and spatial data

Integration of Human Pancreatic Islet and Mouse Retinal Bipolar Cells,



Integration of scRNA-seq and scATAC-seq From mouse visual cortex

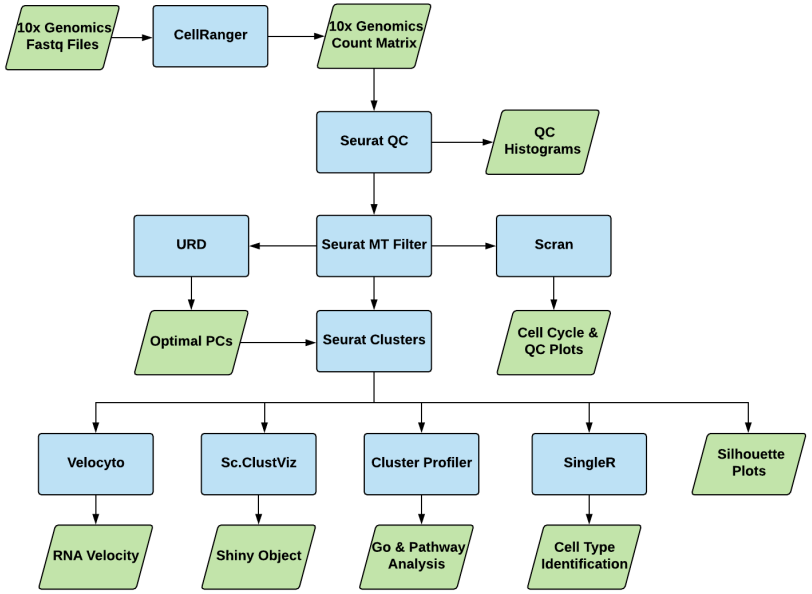


Single Cell Analysis Pipelines
- Developed by CCR bioinformatics teams

Frederick National Laboratory for Cancer Research

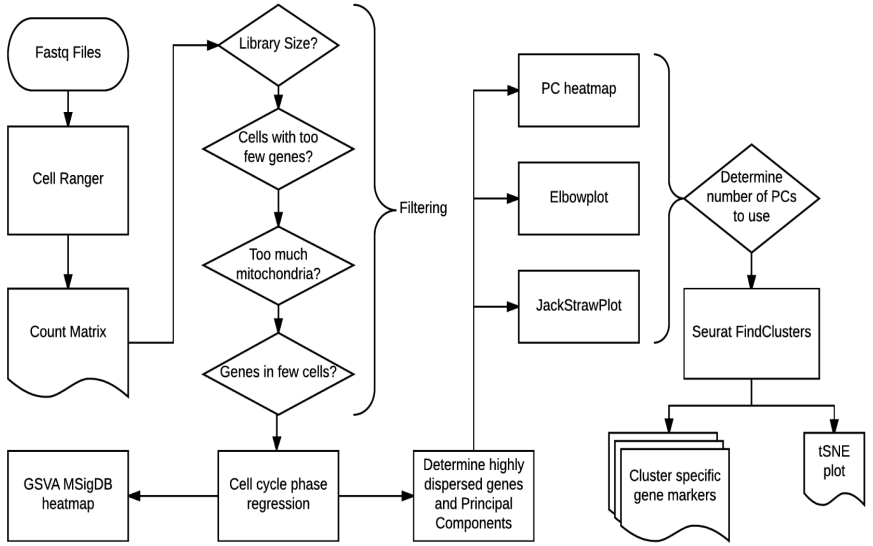
CCR-SF Single Cell Analysis Pipeline:

https://github.com/abcsFrederick/scRNA_pipeline/wiki



CCBR Single Cell RNA-seq Pipeline:

https://github.com/CCBR/scRNASeq

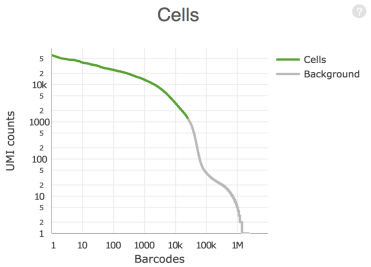


CCR-SF Single Cell Analysis Pipeline

Frederick National Laboratory for Cancer Research

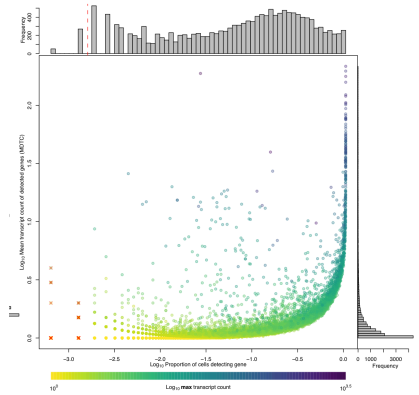
Preprocessing QC, Clustering analysis and visualization

Cellranger QC Statistics

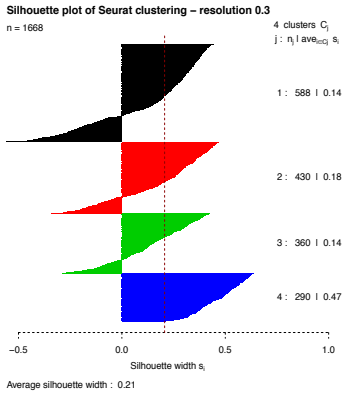


Estimated Number of Cells	26,182
Fraction Reads in Cells	72.0%
Pre-Normalization Mean Reads per Cell	88,969
Post-Normalization Mean Reads per Cell	26,396
Median Genes per Cell	966
Median UMI Counts per Cell	2,445

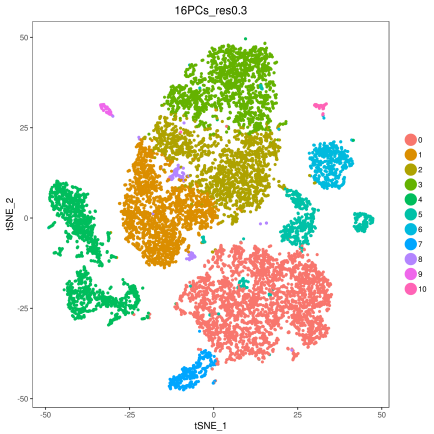
Gene Expression Plot



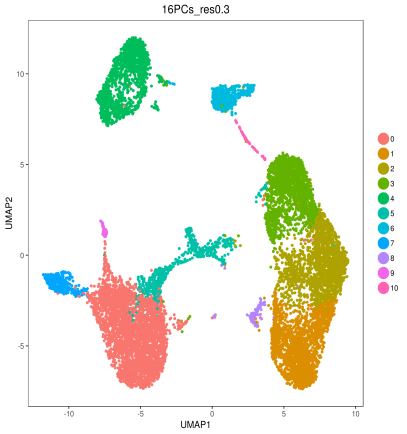
Silhouette Plot



Clustering Result TSNE



Clustering Result UMAP



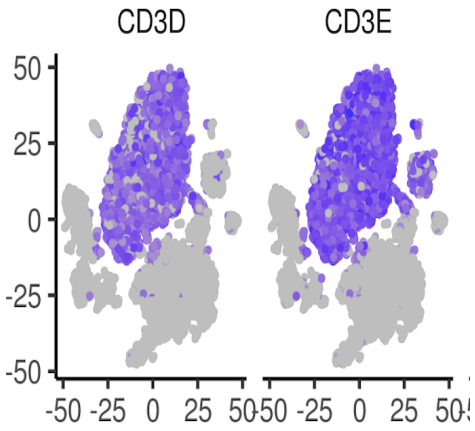
Cell Ranger is used for gene quantification. Seurat and Scran for filtering and removing unwanted source of variations. Scran for cell cycle and library size QC. Scran deconvolution, logCPM are used for library size normalization. Scran and Seurat are used for clustering. Optimal numbers of PCs are calculated in URD and the clustering results evaluated with silhouette plots . TSNE and UMAP plots are provided for clustering visualization.

CCR-SF Single Cell Analysis Pipeline

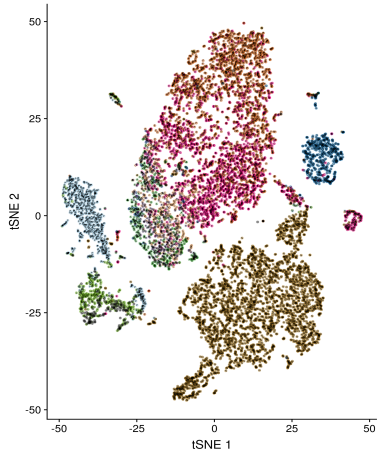
Frederick National Laboratory for Cancer Research

Marker gene identification and cell type classification

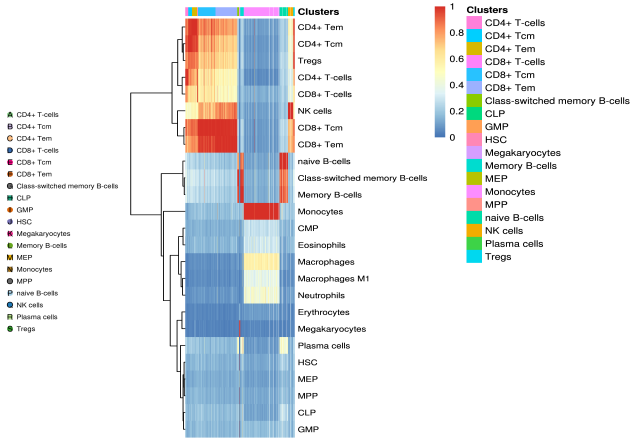
Marker Selection



Cell Type TSNE



Cell Type Heatmap



Gene Set Analysis

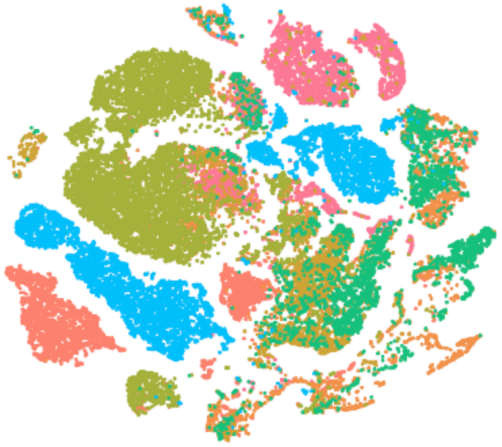
Marker genes were selected from clustering analysis results. SingleR is used to identify cell types based on marker genes in each cluster. Both TSNE and Heatmap were used to visualize the proportion of cells expressed certain marker genes in sample(s). ClusterProfilter, which generates enriched KEGG and GO results for the enriched gene set

CCR-SF Single Cell Analysis Pipeline

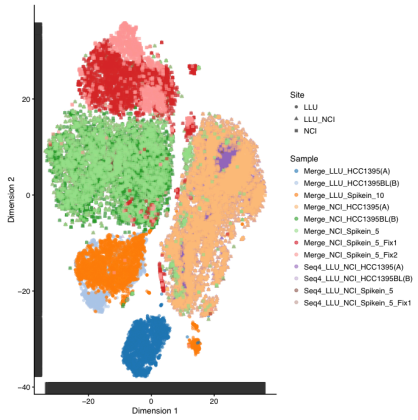
Frederick National Laboratory for Cancer Research

Multiple sample analysis, batch correction

Cellranger Aggregation

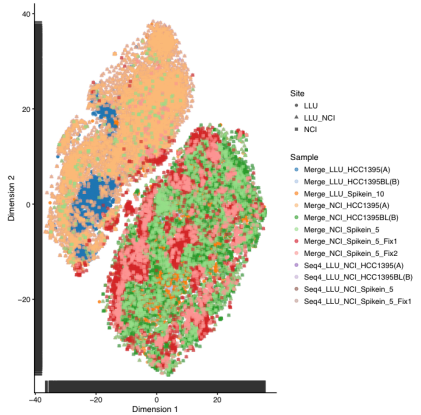


Before Batch Correct



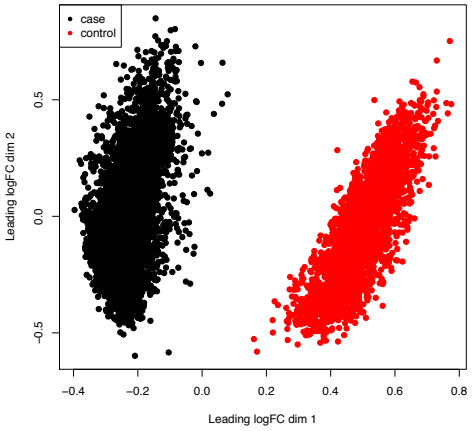
No Correction

After Batch Correction



MNN-corrected

EdgeR Sample DE Analysis



Cell Ranger is used to aggregate multiple samples and perform read depth correction. Pipeline has configure file to flag which samples are grouped, and what samples should be compared. fastMNN and scanorama are used for batch effect correction on multiple samples across batches or protocols. EdgeR and MAST are used to perform differential expression analysis between samples. Sc. ClustViz shiny objects are generated for interactive analysis of different clusters or sample conditions.

Single Cell Analysis Public Resources

Frederick National Laboratory for Cancer Research

- **Comprehensive list of single-cell software and resources**
  - https://github.com/seandavi/awesome-single-cell (Sean Davis)
  - https://www.scrna-tools.org/categories
- **scRNA-seq online training courses**
  - https://hemberg-lab.github.io/scRNA.seq.course/ (Sanger Institute)
  - https://broadinstitute.github.io/2019_scWorkshop/ (Broad Institute)
  - https://satijalab.org/seurat/ (Satija lab)
  - https://cole-trapnell-lab.github.io/monocle3/ (Trapnell lab)
  - https://github.com/SingleCellTranscriptomics (ISCB)
- **Online Databases**
  - https://www.humancellatlas.org/data-sharing

Some take home messages

- **Spend the time on experimental design considerations**
  - Pilot experiments are worth the time and effort
  - Have the person who will be running the analysis involved in the design
- **Communication between subject matter expert and bioinformatician are crucial**
  - Analysis can be informed / directed by the biological knowledge and hypothesis
  - Informatics (besides efficiently implementing complex tools) helps make sure no fundamental statistical rules are broken during analysis
- **Quality of the input sample defines the experiment**
  - Optimize sample prep, when possible
  - Meaningful insight can still be gained from ‘best you can get’ samples
- **Single cell is often a starting point (or a supporting assay)**
  - Validate observations from single cell