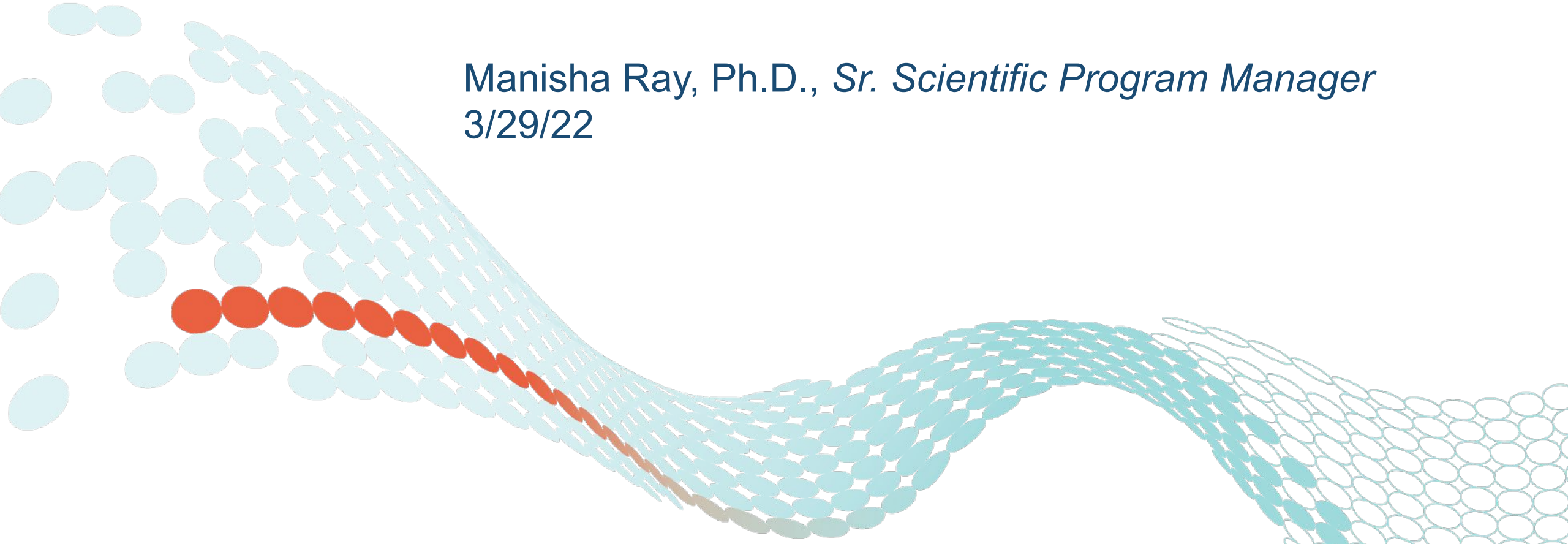
CANCER GENOMICS CLOUD
SEVEN BRIDGES

Single-Cell RNAseq with the Cancer Genomics Cloud

Manisha Ray, Ph.D., *Sr. Scientific Program Manager*
3/29/22



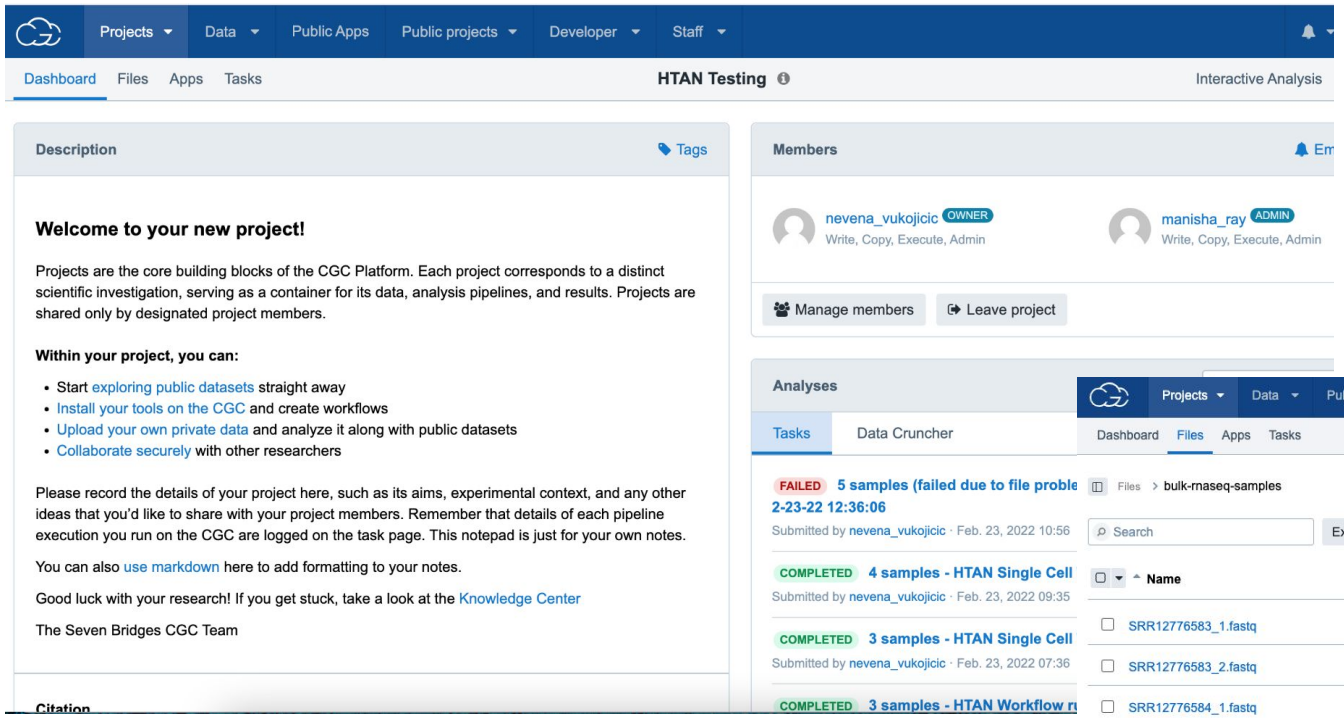
CANCER GENOMICS CLOUD
SEVEN BRIDGES

Outline

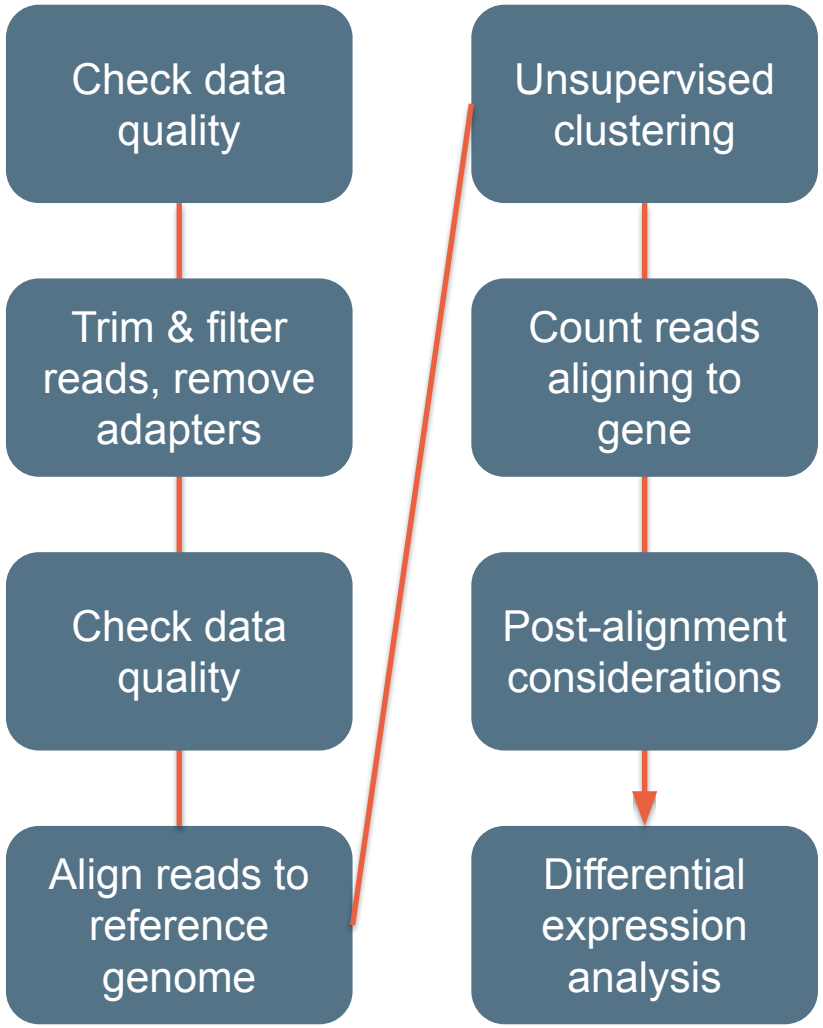
Review of previous material
Why Single-Cell Analysis?
Differences between Single-Cell and Bulk Analysis
Tools for Single-Cell Research on the CGC
Demonstration

Recap of CGC classes

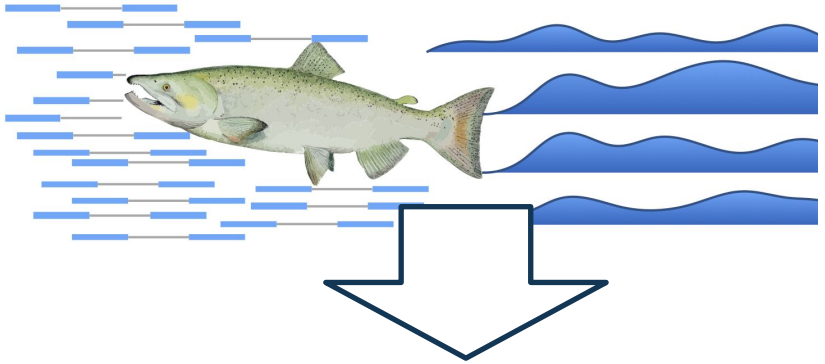
1: CGC Platform; transferring data from SRA to CGC



2: Steps of RNAseq analysis



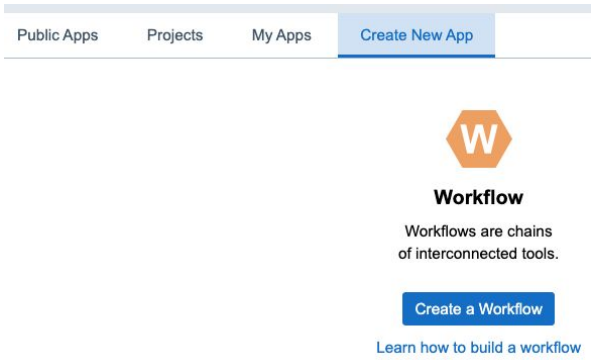
Pseudoalignment of RNA transcripts to genes using **Salmon**

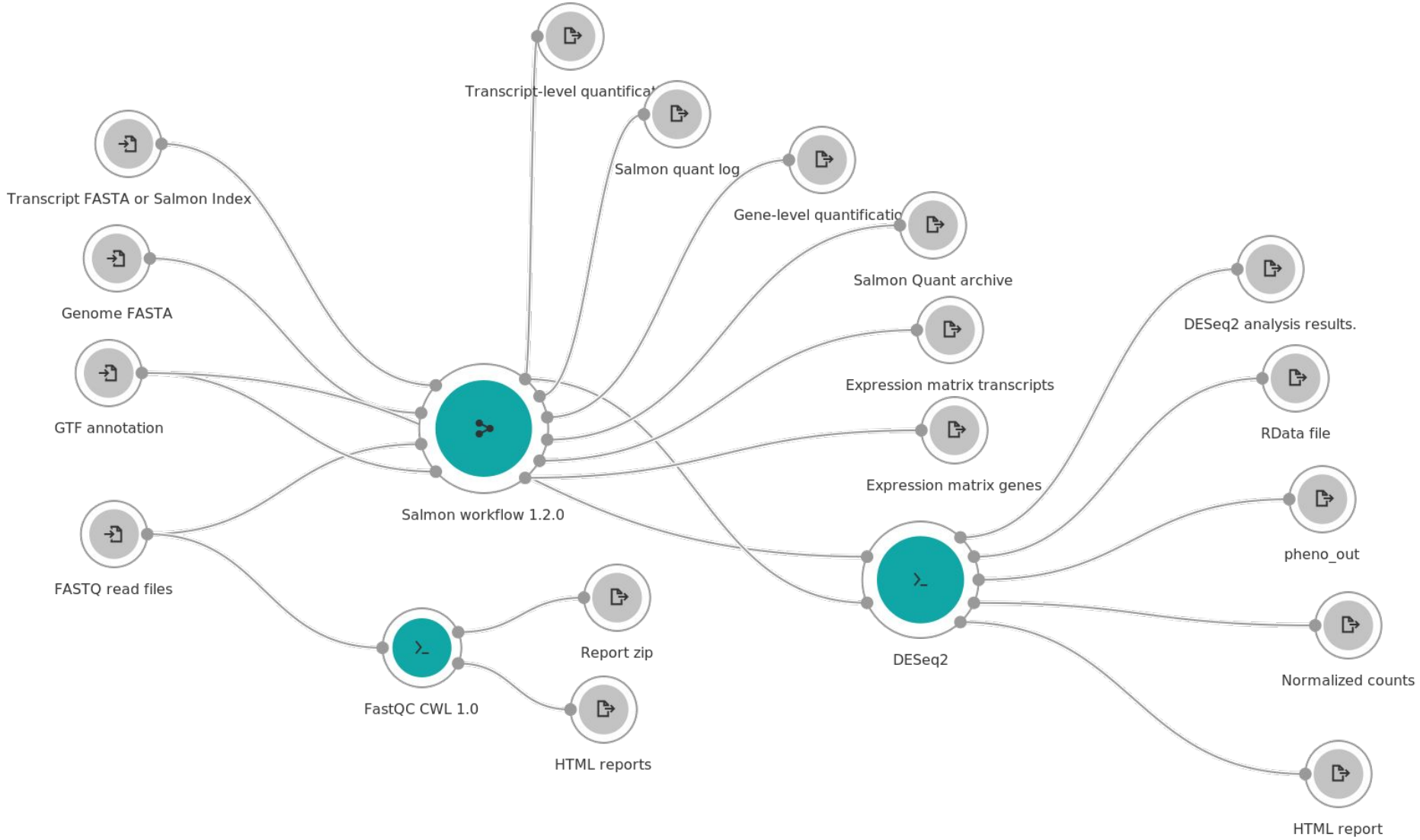


Quantification of genes using **DESeq2**

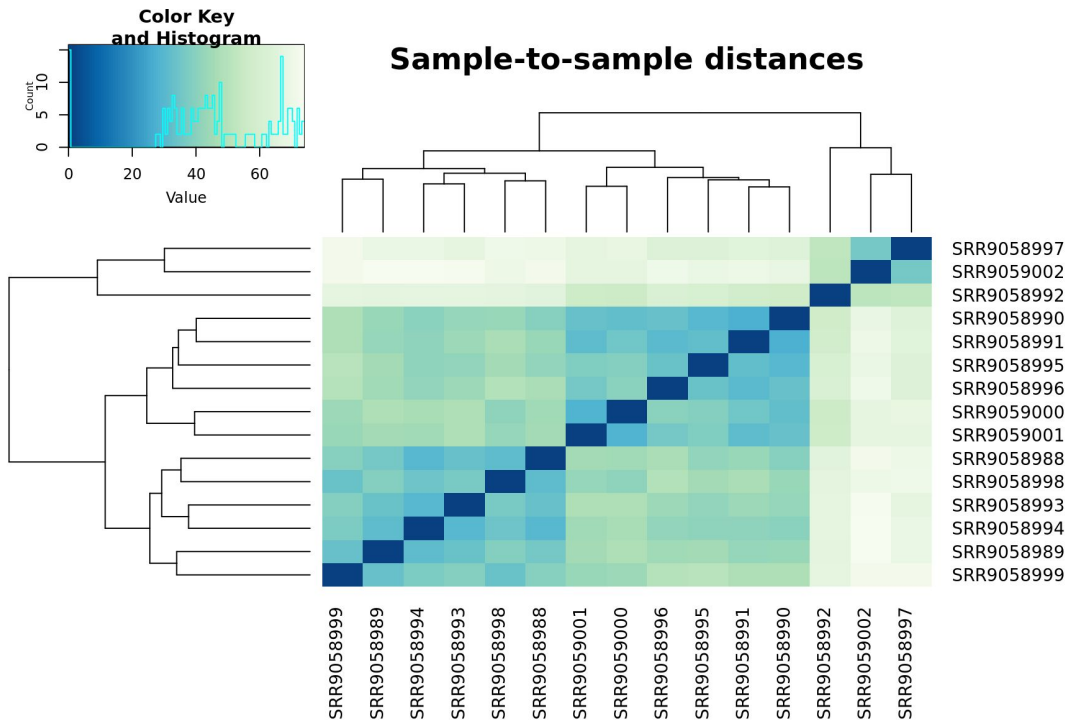
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

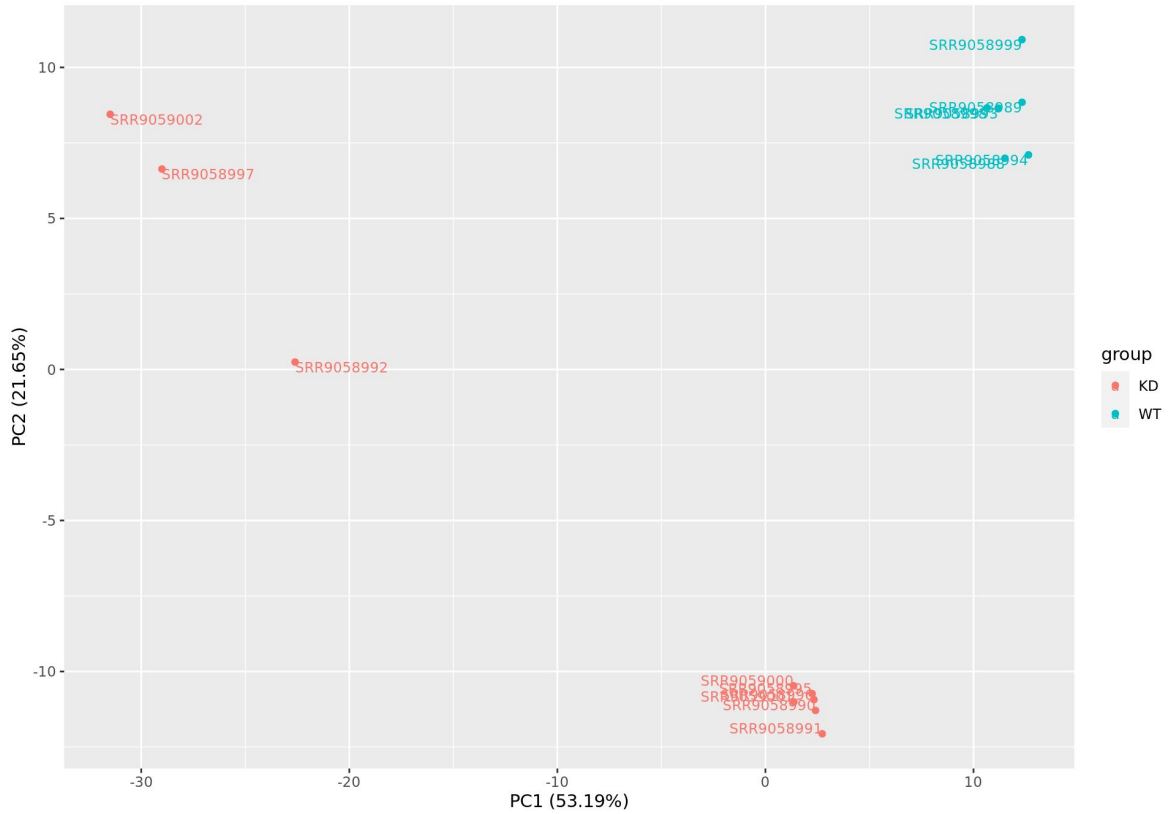
2: Running an RNAseq workflow





3: Interactive analysis of biological groups by clustering heat maps and Principle Component Analysis in RStudio



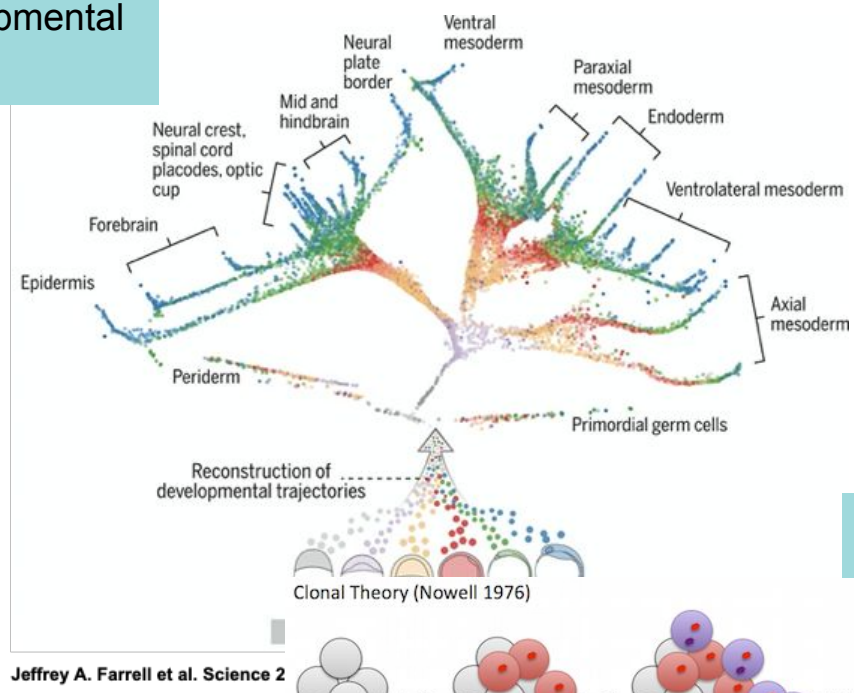


Introduction to Single-Cell Analysis

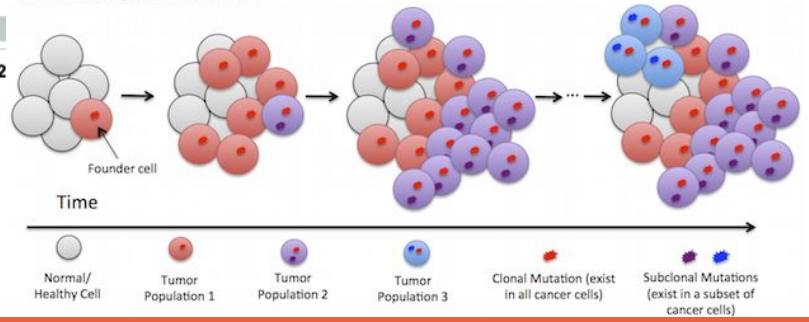
Why study single cells?

Understanding the **heterogeneity** present in **complex tissues** leads to understanding of the **emergent properties**

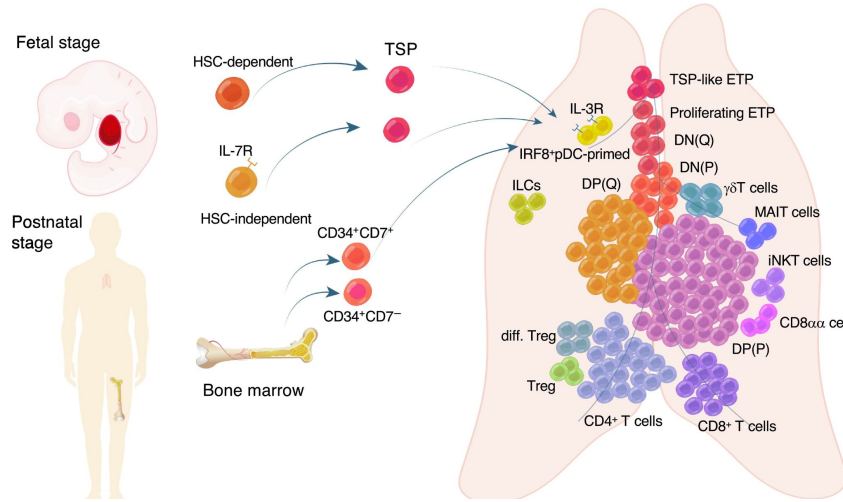
Developmental biology



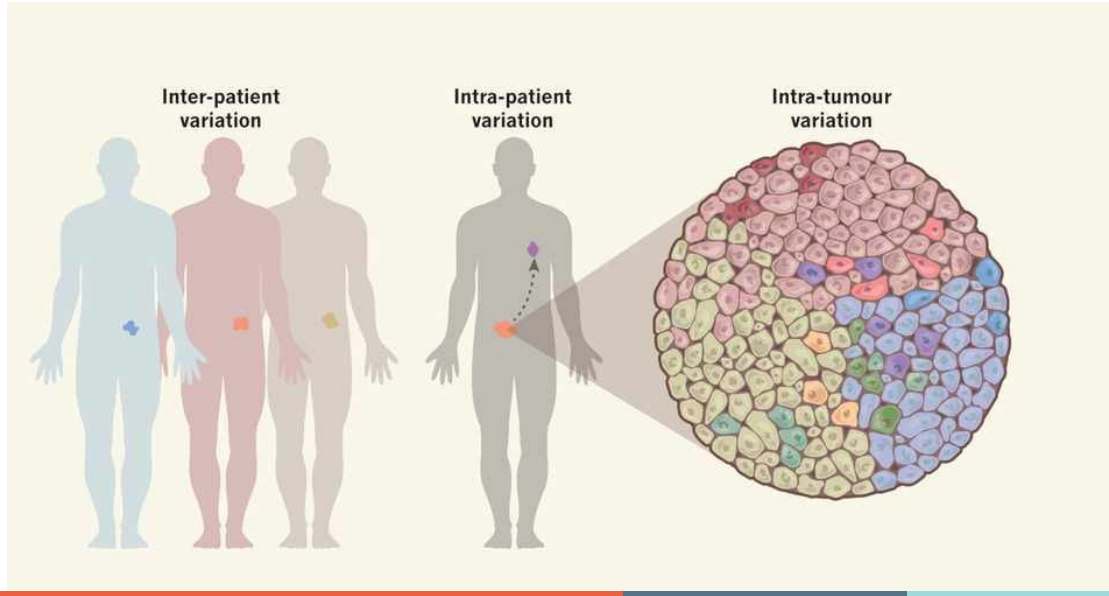
Cancer



Immunology



Trends in Immunology, Jan 2021 https://doi.org/10.1016/j.it.2020.12.004



Why do RNAseq in single cells?

Understand how the *individual* properties of cells bring about *collective* properties in tissues, organs, and systems

RNAseq allows unbiased identification of cell types

Kazer et al. Nature Medicine, Volume 26 March, 2020

Research questions only single cell RNAseq answers

Understanding the genes expressed in individual cells enables an understanding of the **cell types** present and the genes that lead to properties of those cells

- Discovery of new cell types
- Refine gene signature of known cell types
- Identification of novel marker genes
- Identification of novel treatment targets
- Understand what genes drive change in a tissue

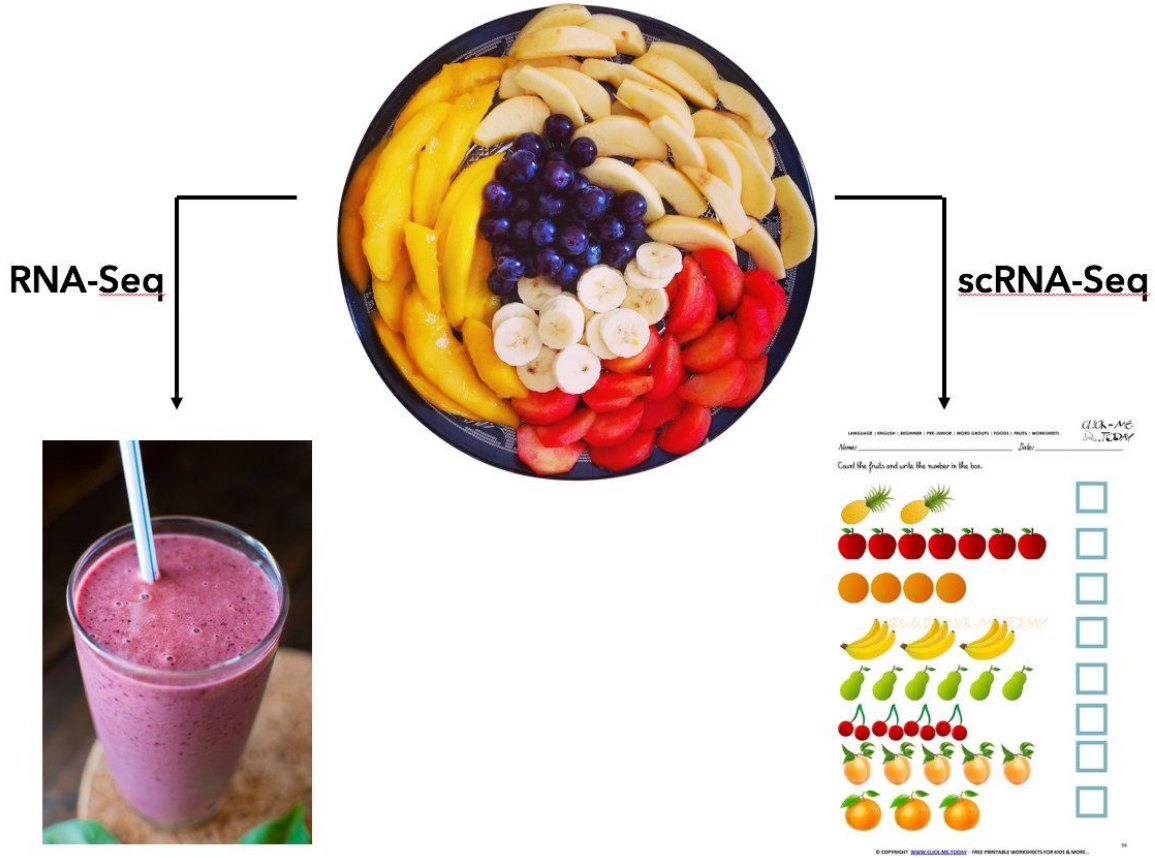
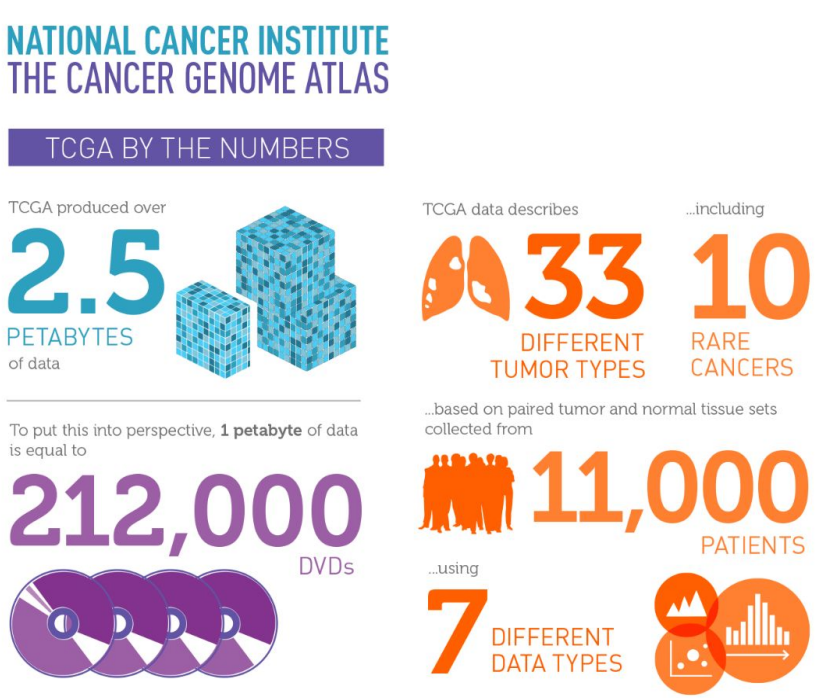


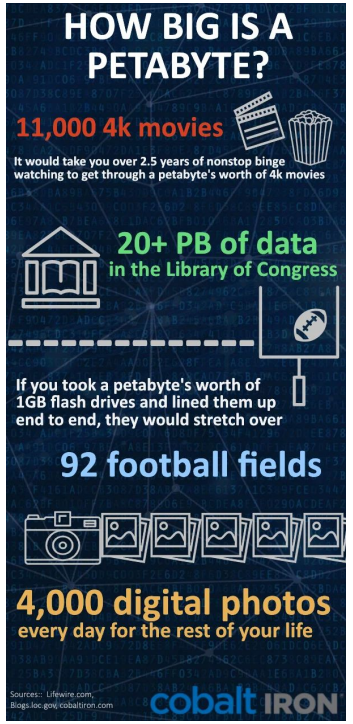
Figure by Roby Bhattacharyya; Smoothie concept by Aviv Regev

The Challenges of Single-Cell Analysis

Large and complex datasets

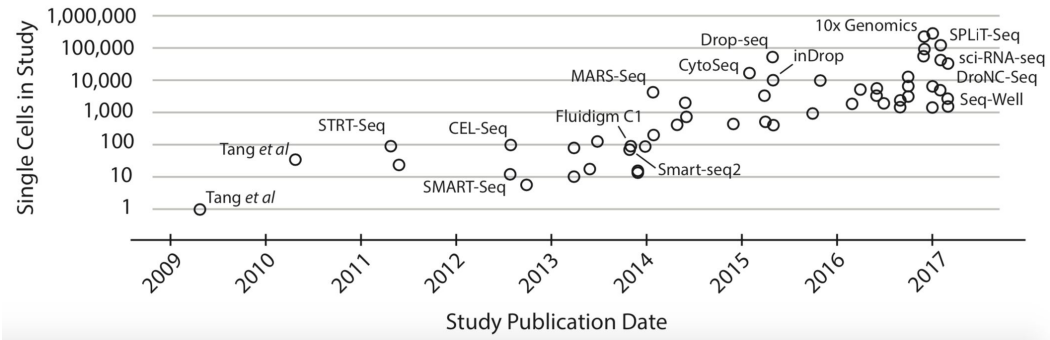
- Expensive to store
- Difficult to access
- Complex data, each cell contains an entire transcriptome
- Computationally expensive to analyze





CANCER GENOMICS CLOUD
SEVEN BRIDGES

Cloud computing makes it easier to store, access, and compute on large datasets



Svensson V, Vento-Tormo R and Teichmann SA. (2018) *Nature protocols*

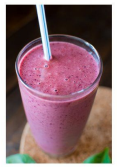
The Challenges of Single-Cell Analysis

Sparse, noisy data

- Bulk methods usually have few samples, many genes
- Single cells can lack gene expression for many reasons
  - “Zero inflated” data
- Alternative methods to traditional RNAseq required

RNAseq tools specific for single-cell in Public App Gallery

Imagine a heatmap of gene expression in a traditional bulk RNAseq experiment



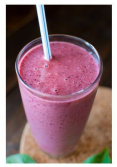
The Challenges of Single-Cell Analysis

Sparse, noisy data

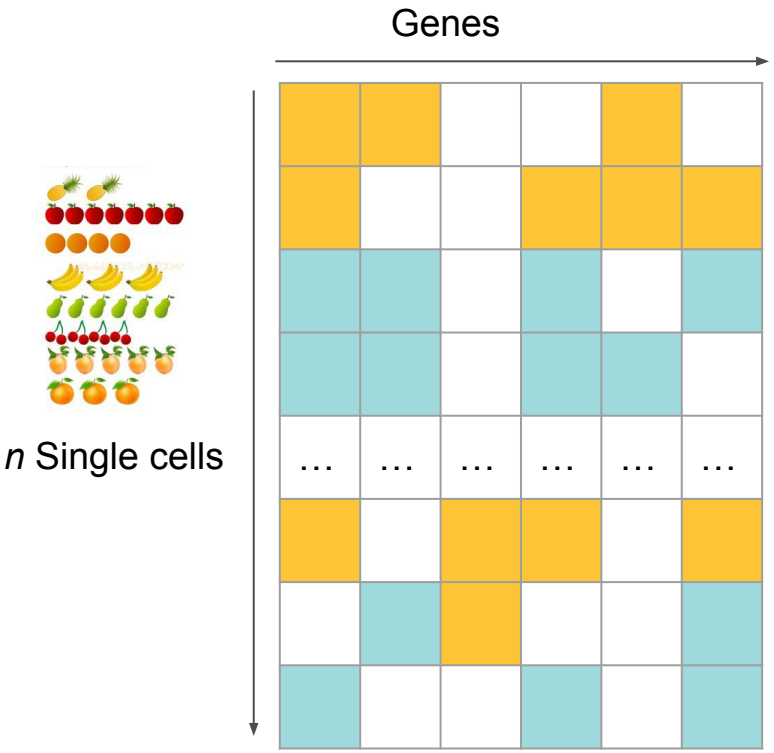
- Bulk methods usually have few samples, many genes
- Single cells can lack gene expression for many reasons
  - "Zero inflated" data
- Alternative methods to traditional RNAseq required

RNAseq tools specific for single-cell in Public App Gallery

Imagine a heatmap of gene expression in a traditional bulk RNAseq experiment



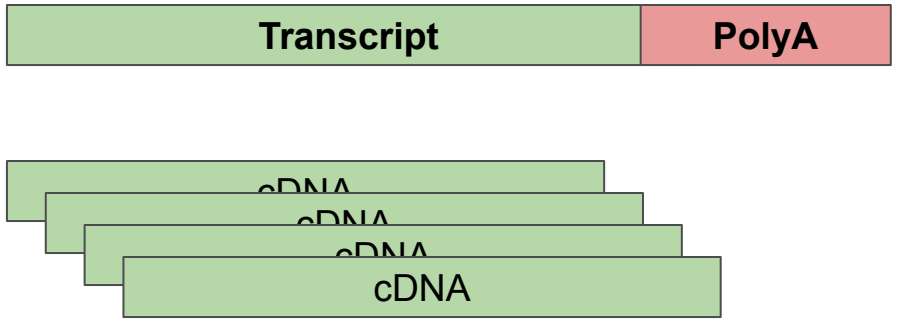
A similar heatmap of gene expression for single cells will be **sparse** - not every gene detected in every cell.



Methods for Single-Cell Analysis

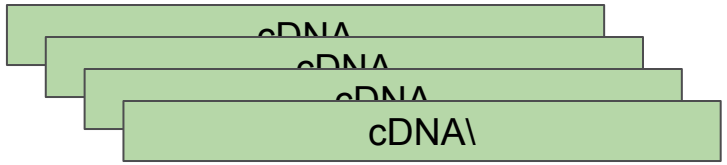
Two preparation methods for single cell RNAseq data

- **Full-length** - capture entire sequence of the transcripts
  - Useful for questions involving splicing, mutations, SNVs
  - Low multiplex ability, dependent on sequencing depth.

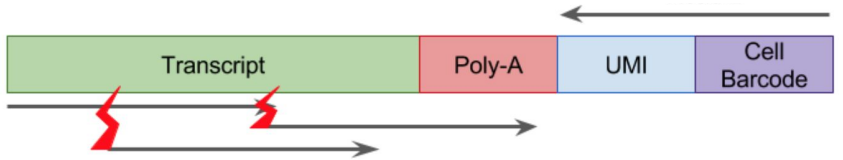


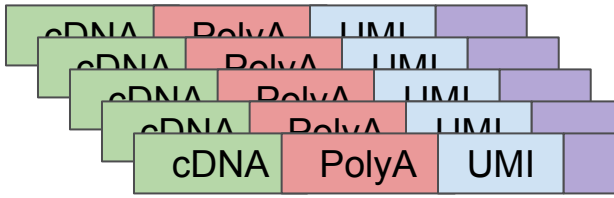
Two preparation methods for single cell RNAseq data

- **Full-length** - capture the **entire** transcript length
  - Useful for questions involving splicing, mutations, SNVs
  - Low multiplex ability, dependent on sequencing depth.



- **End counting**- capture sequences **only at the end** of transcripts
  - Adds barcodes and unique molecular identifiers (UMI)
  - High multiplex ability
  - Useful for analyzing heterogeneity of cell populations and detection of novel cell types.





General flow of single cell RNAseq analysis

Pre-processing:

- **Quality control** - detect and remove low quality cells
  - Library size
  - Number of features
  - Percent of mitochondrial genes
  - Percent of ERCC genes
- **Normalization** - remove artifacts between cells
  - Technical differences in cDNA capture
  - PCR amplification efficiency
- **Batch effect correction** - remove technical variation in data
  - Sequenced in different labs, days, operators, protocols, etc

RNAseq

- **Alignment**
  - Standard or pseudoalignment
  - Quantification - gene to cell count matrix

Data analysis:

- **Dimensionality reduction**
  - PCA - used for further processing
  - UMAP or tSNE- used for visual inspection
- **Clustering and Detection of marker genes**
  - Generate SNN graph of single cells based on PCA
  - Louvain community detection algorithm
  - Marker gene alaysis

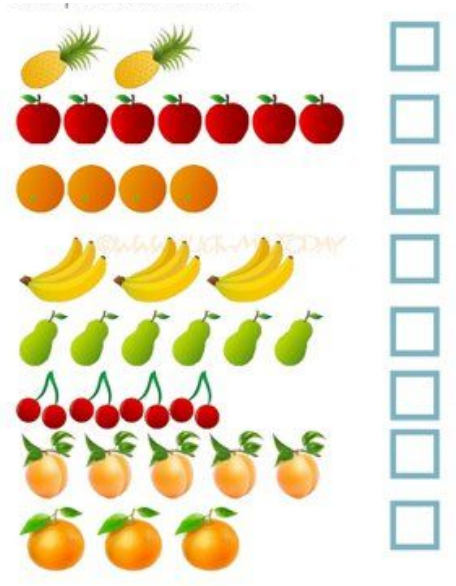
Higher level analysis

- Gene Set Enrichment Analysis
- Trajectory analysis

Downstream Single-Cell Analysis

Gene-to-Cell counts:

- Table output produced after quantification step.
- Holds information about number of reads assigned to each gene per cell.
- Rows representing gene names, columns representing names of the cells.



	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

Tools for Single-Cell Data on the CGC

Tool/Workflow for processing full-length data

- Smart-seq2 [Workflow]
- TraCeR [Tool]*
- MiXCR [Tool]*
- SICILIAN [Tool]

*Available upon request
**Coming soon

Tool/Workflow for processing UMI-based data

Aligners

- Cell Ranger [Toolkit]*
- Kallisto BUStools [Workflow]
- zUMIs [Tool]
- STARsolo [Tool]
- Salmon alevin [Tool]

Tertiary analysis

- Seurat [Workflow]
- Psuedobulk [Workflow]
- GSEA [Tool]
- Pairwise Differential Expression

Interactive Analysis for higher level analysis:

- Harmony: Correction of batch effect [RMarkdown]
- Seurat [RMarkdown]
- Slingshot: Trajectory inference [RMarkdown and Workflow**]
- Velocity analysis [RMarkdown and Tool**]

Single cell analysis with Seurat

Today's question - can we identify the cell types present in Bone Marrow?

Data

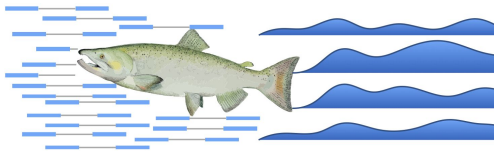
RNAseq data from the Human Cell Atlas data from Bone Marrow samples

HUMAN CELL ATLAS

Tools

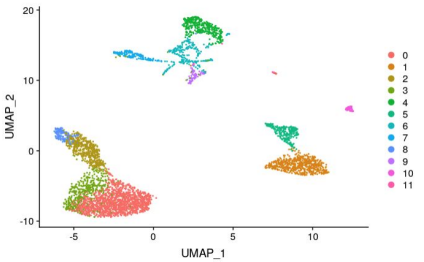
Salmon Alevin to align RNAseq data

- Optimized for single-cell data



Seurat to cluster cells by similarity and identify marker genes

Seurat R CRAN



Clustering and Detection of Marker Genes: Seurat

- Input gene count matrix
- Preprocessing:
  - Quality control
- Basic analysis
  - Normalization
  - Transformation
  - Dimensionality reduction (PCA)
- Clustering and marker identification
  - Clustering
  - Pairwise differential expression testing between clusters
  - Output tables of marker genes

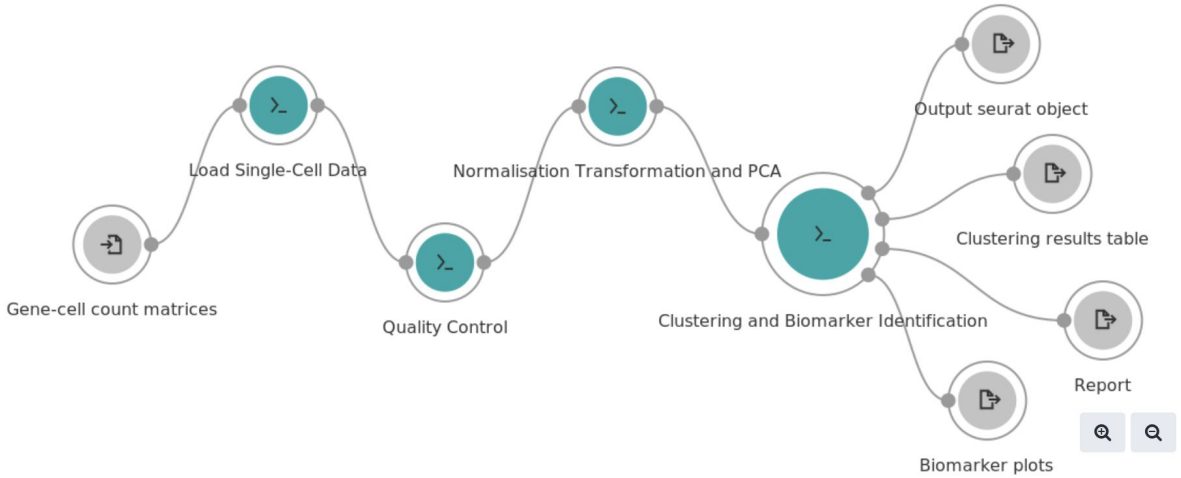
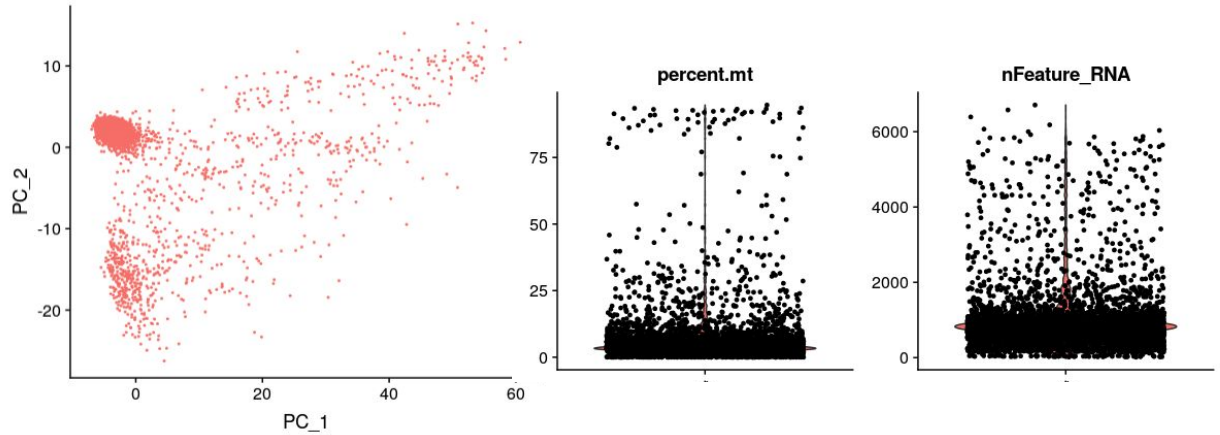
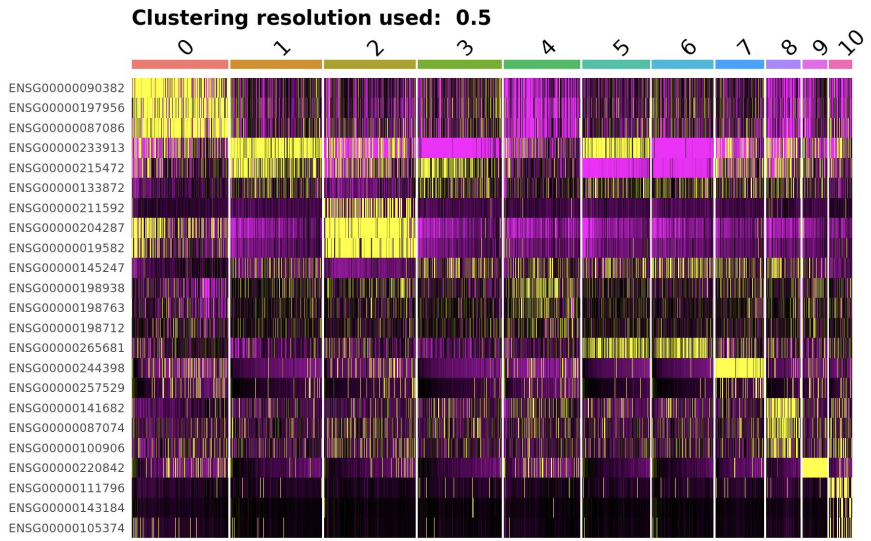
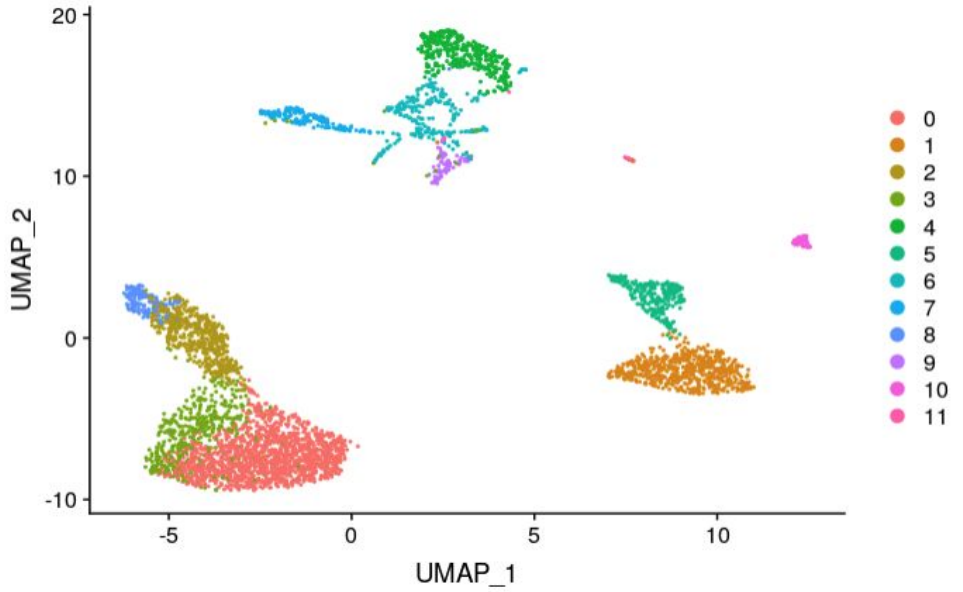


Image source: https://davetang.org/muse/2017/08/01/getting-started-seurat

Clustering and Detection of Marker Genes: Seurat







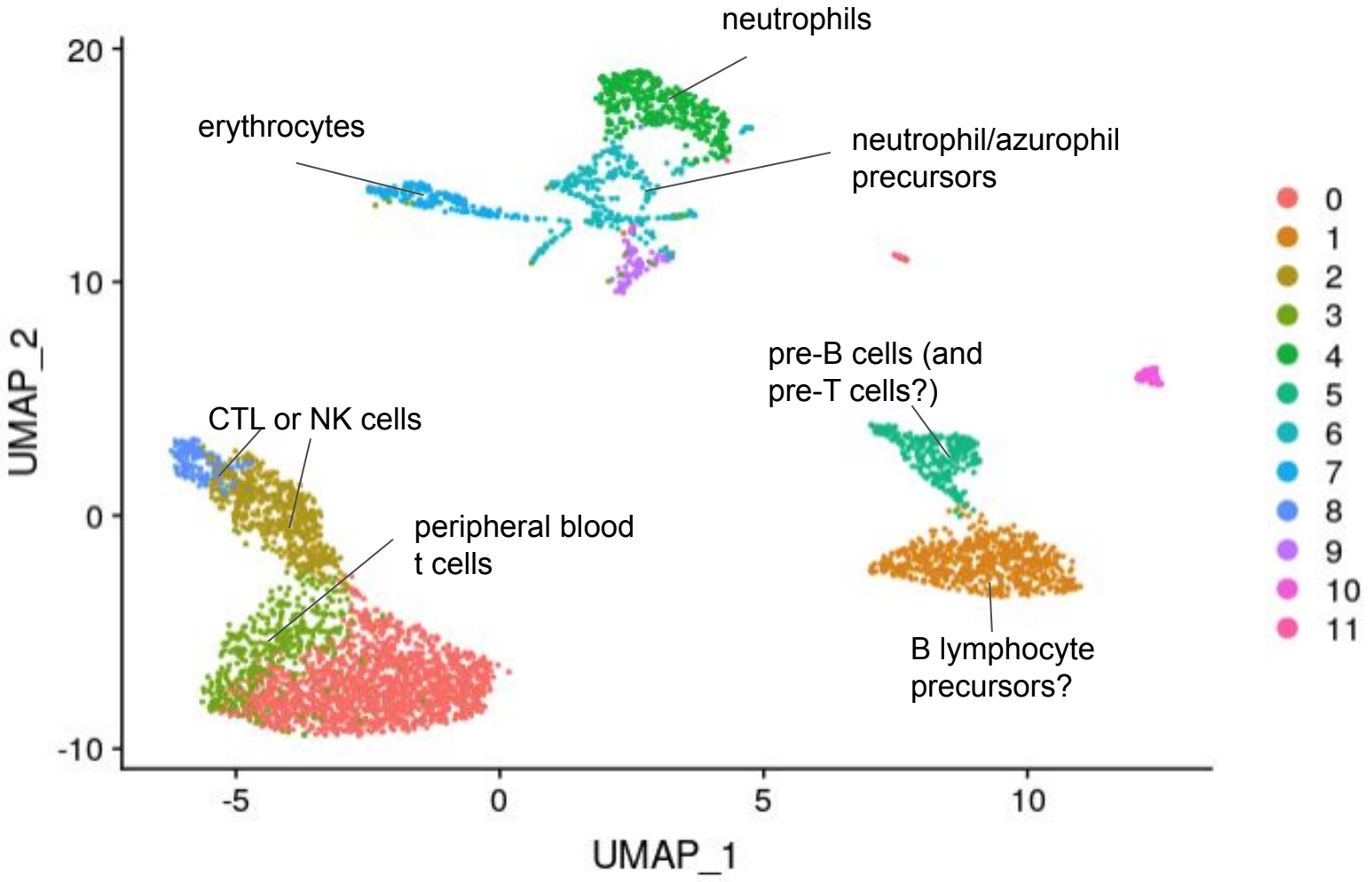
Tab 1. Top 10 differentialy expressed genes for each cluster comparing to all the other clusters. Clastering resolution used: 0.5.

	gene_id	gene_symbol	cluster	avg_logFC	p_val	p_val_adj	gene_description
21	ENSG00000112149	CD83	2	0.94	0	0	CD83 (CD83 Molecule) is a P
22	ENSG00000104894	CD37	2	0.89	3.84e-231	5.28e-227	CD37 (CD37 Molecule) is a P
23	ENSG00000162511	LAPTM5	2	0.83	1.38e-178	1.89e-174	LAPTM5 (Lysosomal Protein
24	ENSG00000157514	TSC22D3	2	0.85	1.32e-98	1.81e-94	TSC22D3 (TSC22 Domain Fa
25	ENSG00000215472	RPL17-C18orf32	3	0.67	9.45e-124	1.3e-119	RPL17-C18orf32 (RPL17-C1
26	ENSG00000133872	SARAF	3	0.35	1.12e-45	1.54e-41	SARAF (Store-Operated Calc

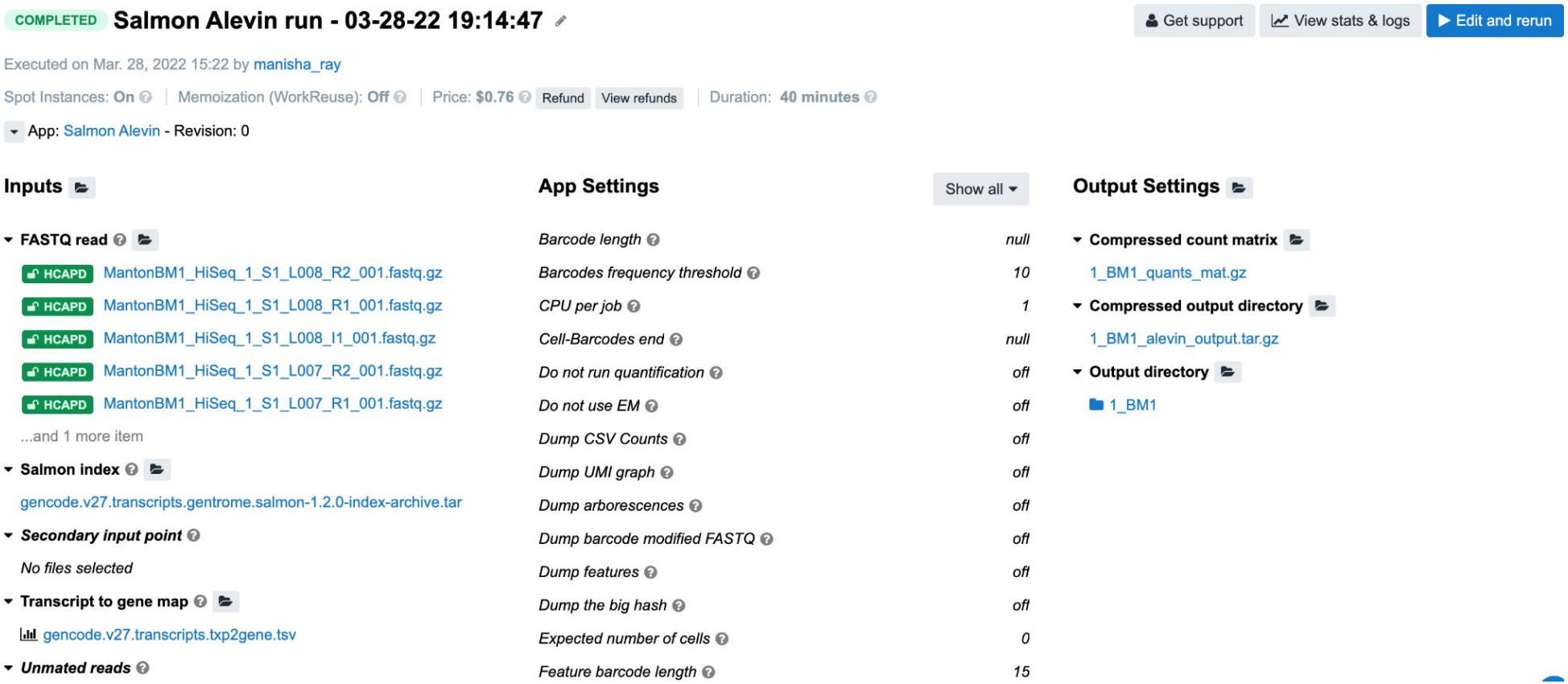
Image source: https://davetang.org/muse/2017/08/01/getting-started-seurat

Cell type clusters in Bone Marrow cells using Seurat

Use biological role of differentially expressed genes to identify cell types in each cluster



Salmon Alevin tool for aligning transcripts



Salmon Alevin tool for aligning transcripts

COMPLETED **Salmon Alevin run - 03-28-22 19:14:47**

Get support | View stats & logs | ▶ Edit and rerun

Executed on Mar. 28, 2022 15:22 by manisha_ray

Spot Instances: **On** | Memoization (WorkReuse): **Off** | Price: **$0.76** Refund View refunds | Duration: **40 minutes**

App: Salmon Alevin - Revision: 0

Inputs

- **FASTQ read**
  - HCAPD MantonBM1_HiSeq_1_S1_L008_R2_001.fastq.gz
  - HCAPD MantonBM1_HiSeq_1_S1_L008_R1_001.fastq.gz
  - HCAPD MantonBM1_HiSeq_1_S1_L008_I1_001.fastq.gz
  - HCAPD MantonBM1_HiSeq_1_S1_L007_R2_001.fastq.gz
  - HCAPD MantonBM1_HiSeq_1_S1_L007_R1_001.fastq.gz
  - ...and 1 more item
- **Salmon index**
  - gencode.v27.transcripts.gentrome.salmon-1.2.0-index-archive.tar
- ***Secondary input point***
  - *No files selected*
- **Transcript to gene map**
  - gencode.v27.transcripts.txp2gene.tsv
- ***Unmated reads***

App Settings Show all

Barcode length	*null*
Barcodes frequency threshold	*10*
CPU per job	*1*
Cell-Barcodes end	*null*
Do not run quantification	*off*
Do not use EM	*off*
Dump CSV Counts	*off*
Dump UMI graph	*off*
Dump arborescences	*off*
Dump barcode modified FASTQ	*off*
Dump features	*off*
Dump the big hash	*off*
Expected number of cells	*0*
Feature barcode length	*15*

Output Settings

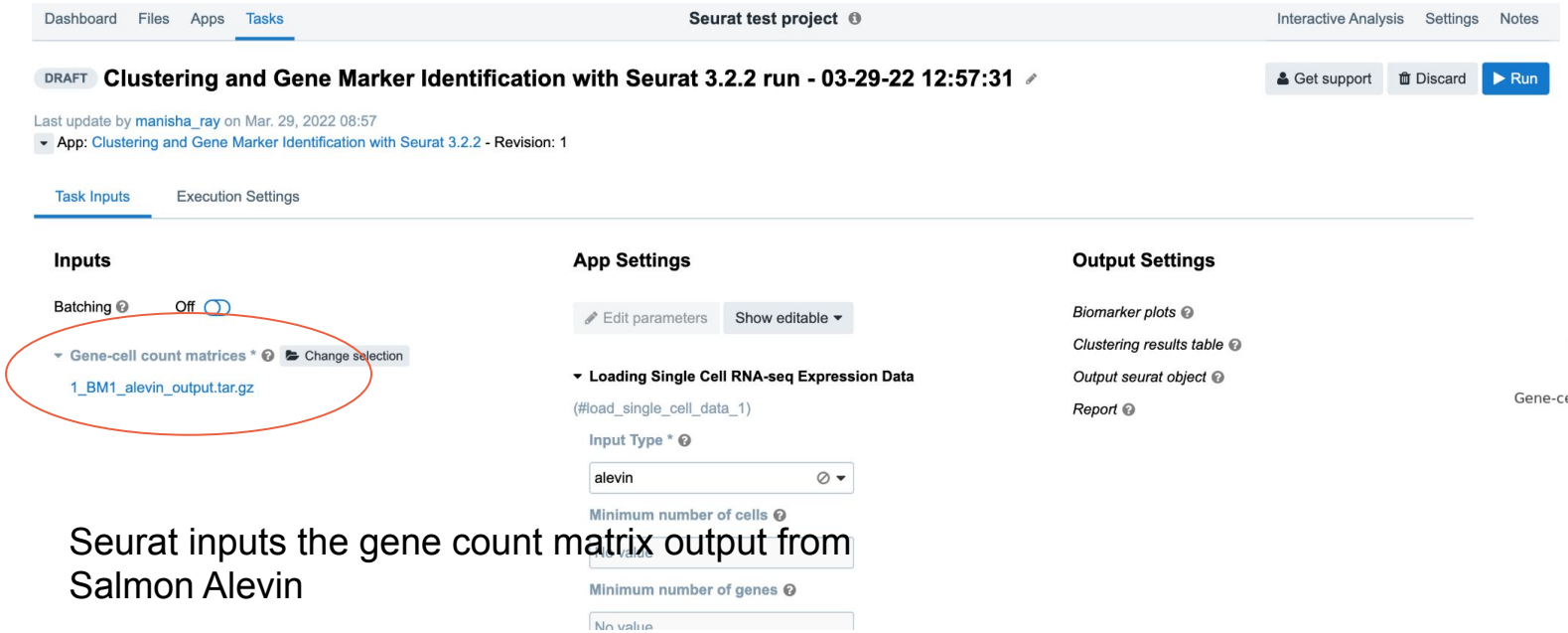
- **Compressed count matrix**
  - 1_BM1_quants_mat.gz
- **Compressed output directory**
  - 1_BM1_alevin_output.tar.gz
- **Output directory**
  - 1_BM1

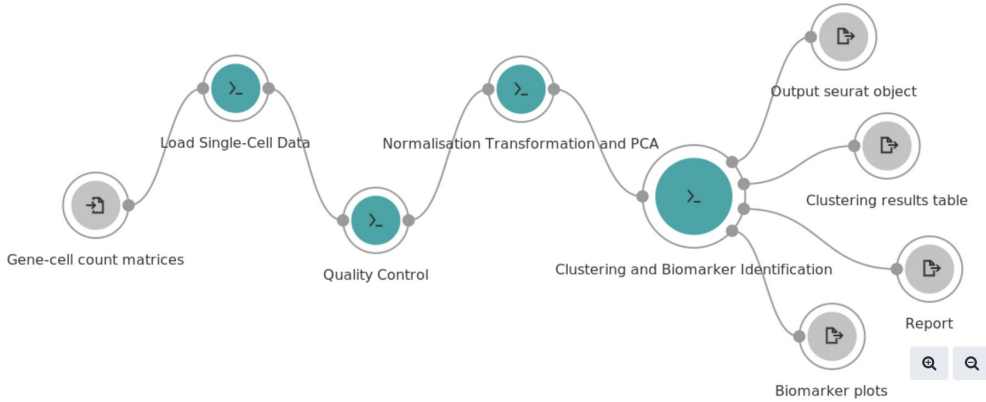
Input: fastq files of single cells

Output: matrix of gene counts per cell

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

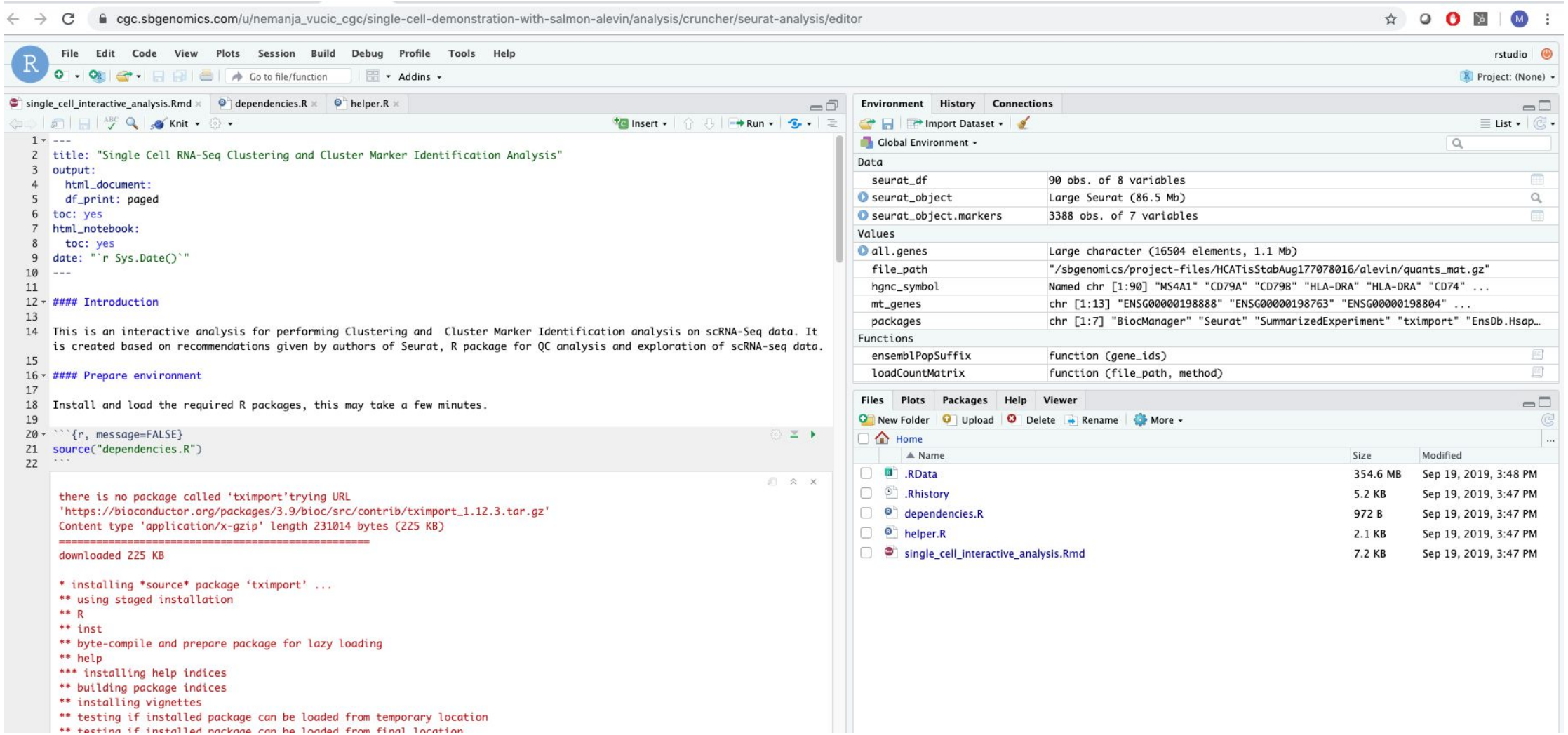
Run Seurat as a workflow





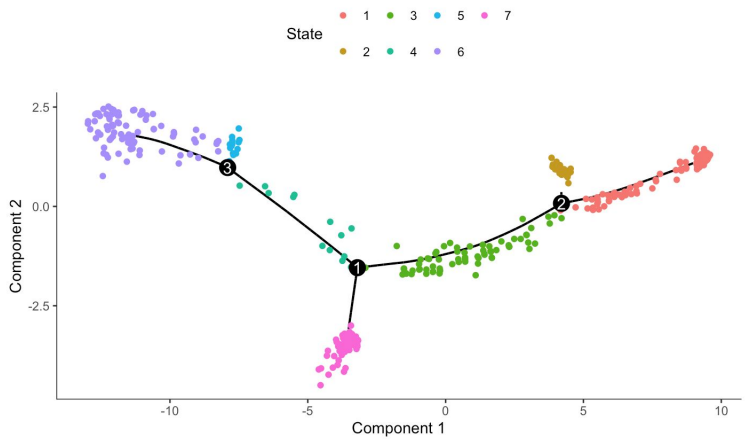
Seurat inputs the gene count matrix output from Salmon Alevin

Continue interactively analyzing Seurat in RStudio

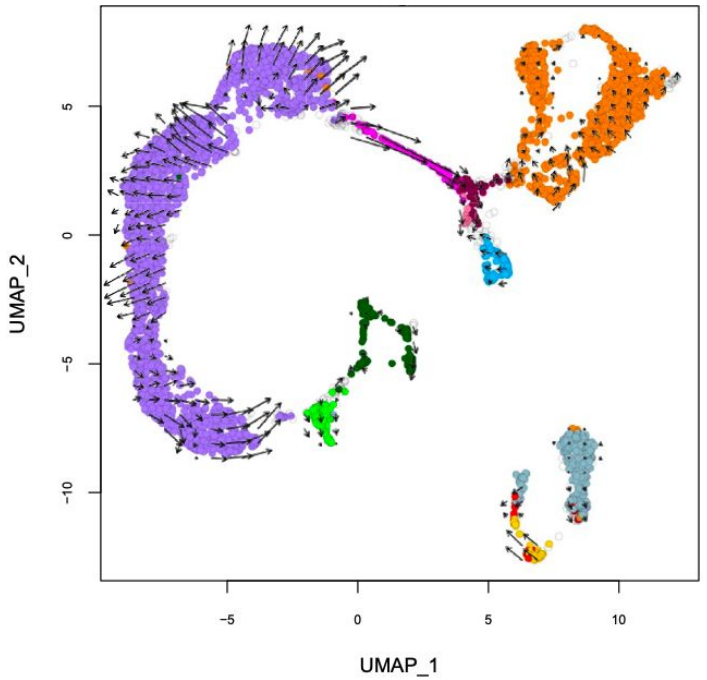


Even higher level analysis of single-cells

Developmental trajectory inference with Monocle

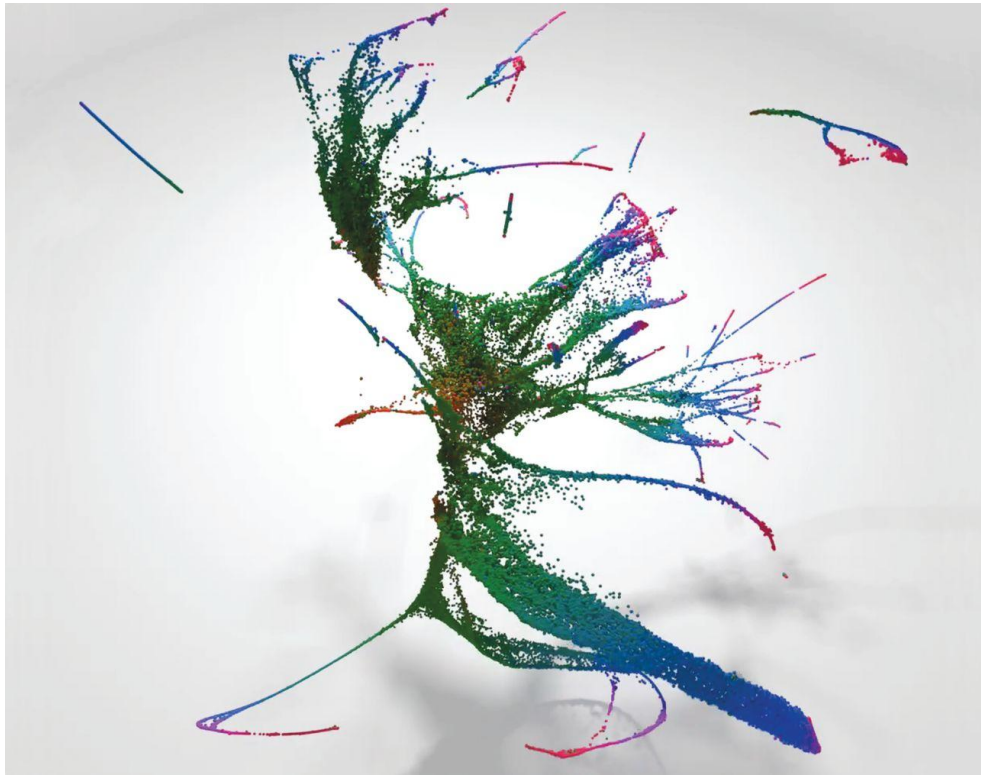


Velocity plot analysis of transcriptional dynamics



Anderson et al; *Scientific Reports* (2020) 10:19173
https://doi.org/10.1038/s41598-020-76157-4

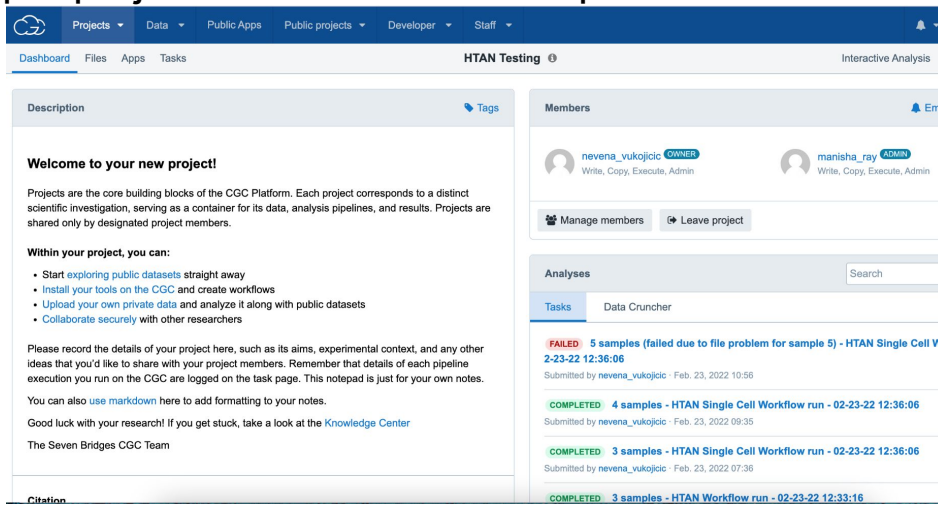
3D UMAP clustering of entire *C. elegans* at single-cell resolution



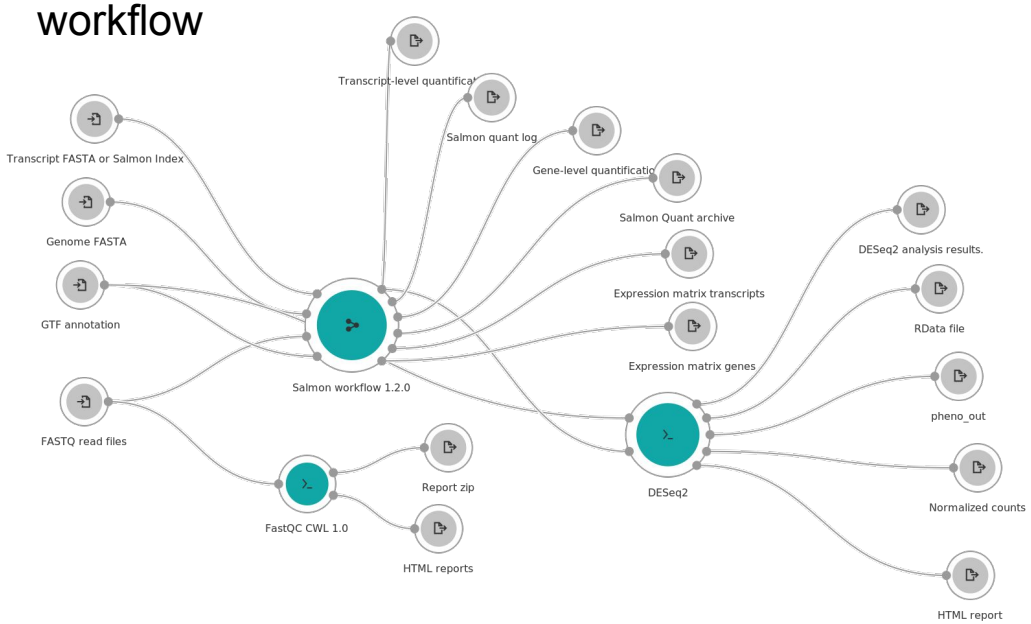
Jonathan S. Packer et al. Science 2019;365:eaax1971

Summary

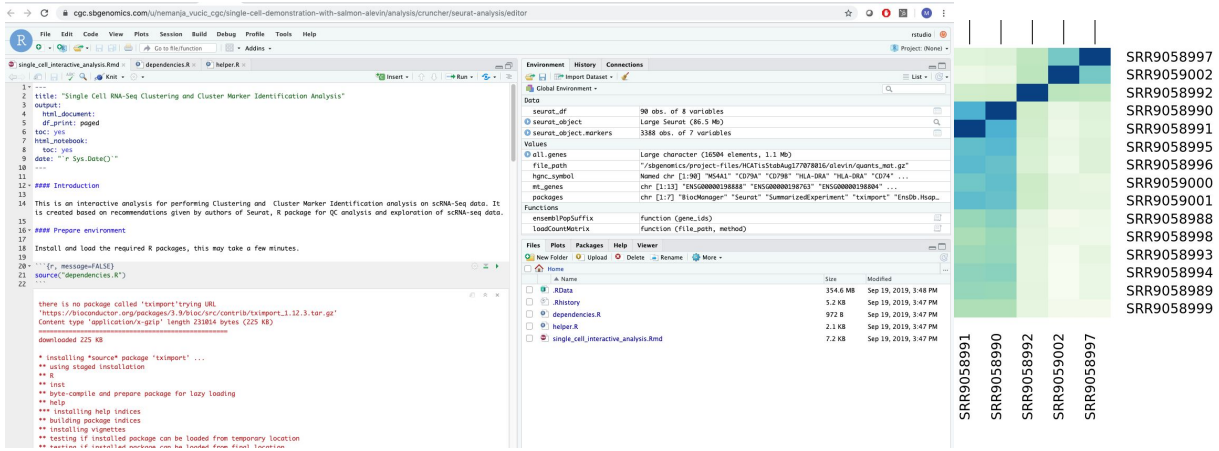
Set up a project on the CGC and import data



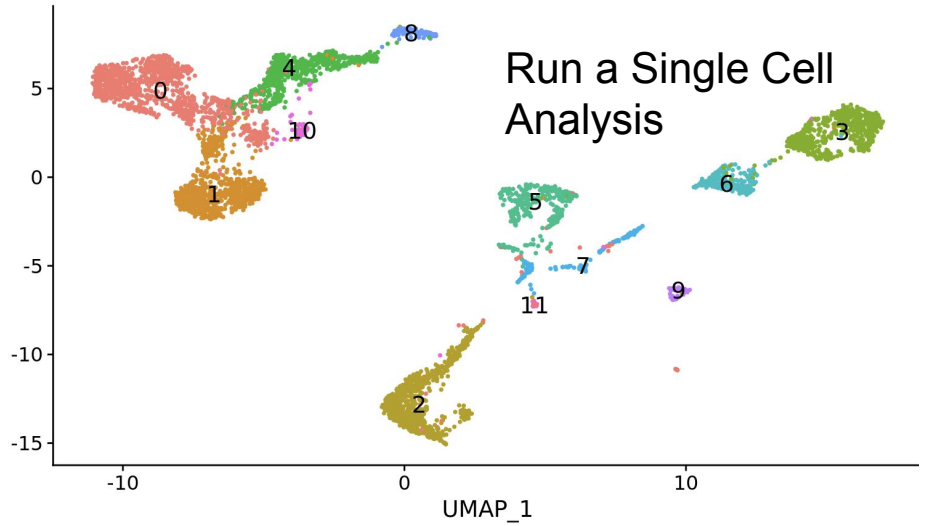
Run an RNAseq workflow



Differential Gene Expression using Data Cruncher with R



Run a Single Cell Analysis



Have questions? Contact us via email or attend office hours

Attend Office Hours every week:

- **10:00 am ET Tuesday**
- **2:00 pm ET Thursday**

Manisha Ray

manisha.ray@sevenbridges.com

Zelia Worman

zelia.worman@sevenbridges.com

Phil Webster

phil.webster@sevenbridges.com

CANCER GENOMICS CLOUD
SEVEN BRIDGES

https://www.cancergenomicscloud.org

Acknowledgements:
Dr. Jeffrey Grover
Dr. Min Zhang
Dr. Nadia Atallah Lanman

Demo