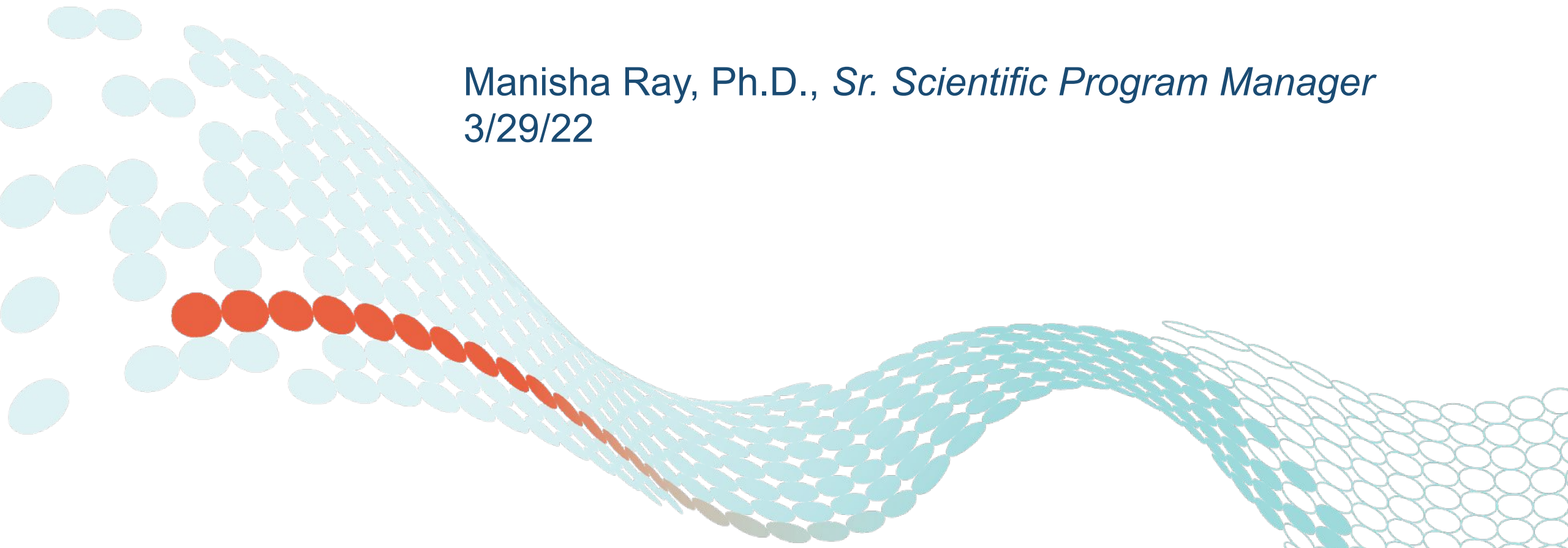


Single-Cell RNAseq with the Cancer Genomics Cloud

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





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

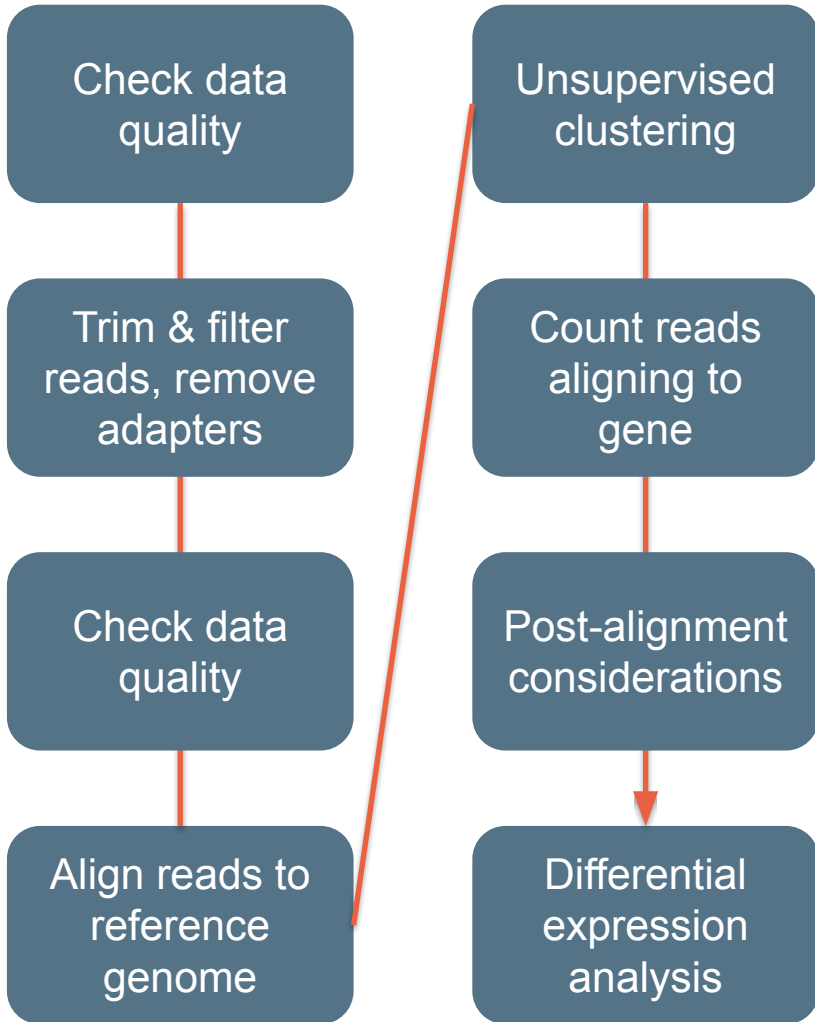
The screenshot displays the CGC Platform interface. The top navigation bar includes 'Projects', 'Data', 'Public Apps', 'Public projects', 'Developer', and 'Staff'. The main content area is divided into several sections:

- Description:** A welcome message for a new project, explaining that projects are core building blocks for scientific investigations. It lists actions users can take within a project, such as exploring public datasets, installing tools, uploading private data, and collaborating.
- Members:** A list of project members, including 'nevena_vukojcic' (OWNER) and 'manisha_ray' (ADMIN), with their respective roles and permissions.
- Analyses:** A section showing analysis tasks. A 'Data Cruncher' task is highlighted, showing a 'FAILED' status for 5 samples. Below this, a table lists completed analyses for HTAN Single Cell and HTAN Workflow.

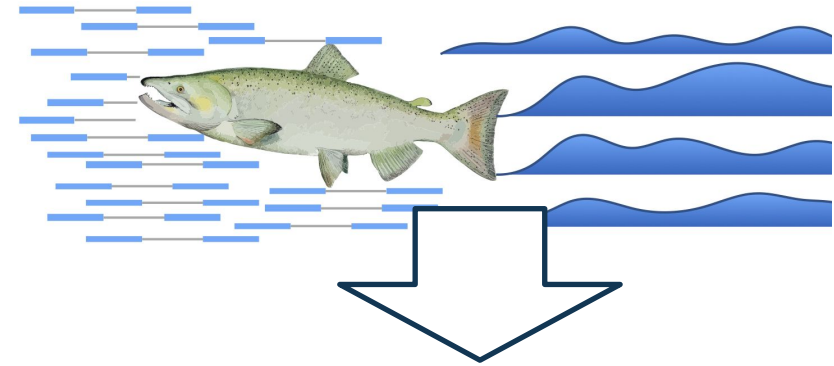
The 'Analyses' section includes a search bar, filters for extension and tags, and a table with the following columns: Name, Experimental strategy, Extension, Size, and Sample ID.

Name	Experimental strategy	Extension	Size	Sample ID
<input type="checkbox"/> SRR12776583_1.fastq	-	FASTQ	61.7 GiB	SRR12776583
<input type="checkbox"/> SRR12776583_2.fastq	-	FASTQ	61.7 GiB	SRR12776583
<input type="checkbox"/> SRR12776584_1.fastq	-	FASTQ	72.2 GiB	SRR12776584
<input type="checkbox"/> SRR12776584_2.fastq	-	FASTQ	72.2 GiB	SRR12776584
<input type="checkbox"/> SRR12776585_1.fastq	-	FASTQ	43.5 GiB	SRR12776585
<input type="checkbox"/> SRR12776585_2.fastq	-	FASTQ	43.5 GiB	SRR12776585
<input type="checkbox"/> SRR12776586_1.fastq	-	FASTQ	74.4 GiB	SRR12776586
<input type="checkbox"/> SRR12776586_2.fastq	-	FASTQ	74.4 GiB	SRR12776586
<input type="checkbox"/> SRR9058988_1.fastq	-	FASTQ	3.4 GiB	SRR9058988
<input type="checkbox"/> SRR9058988_2.fastq	-	FASTQ	3.4 GiB	SRR9058988

2: Steps of RNAseq analysis



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

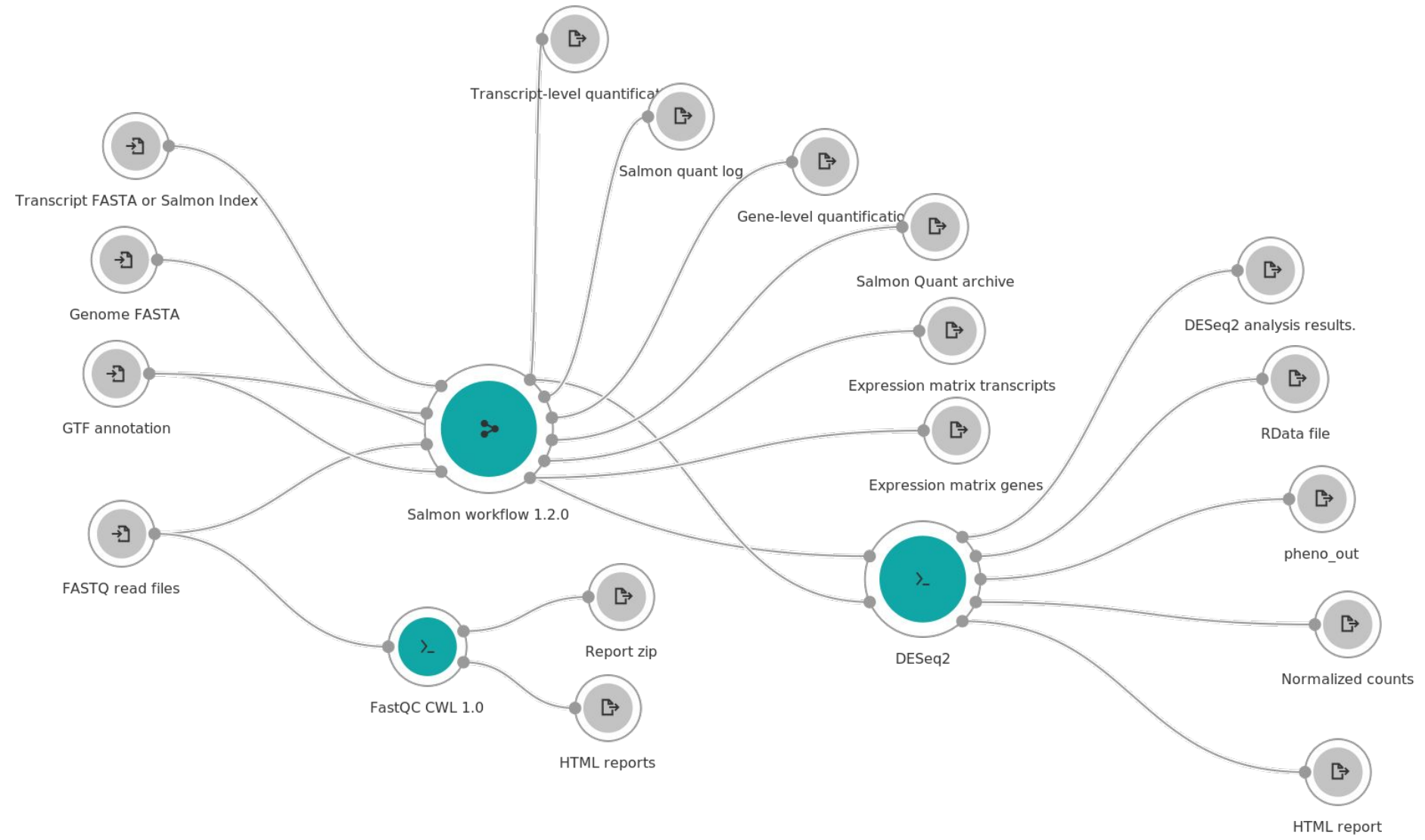


Quantification of genes using **DESeq2**

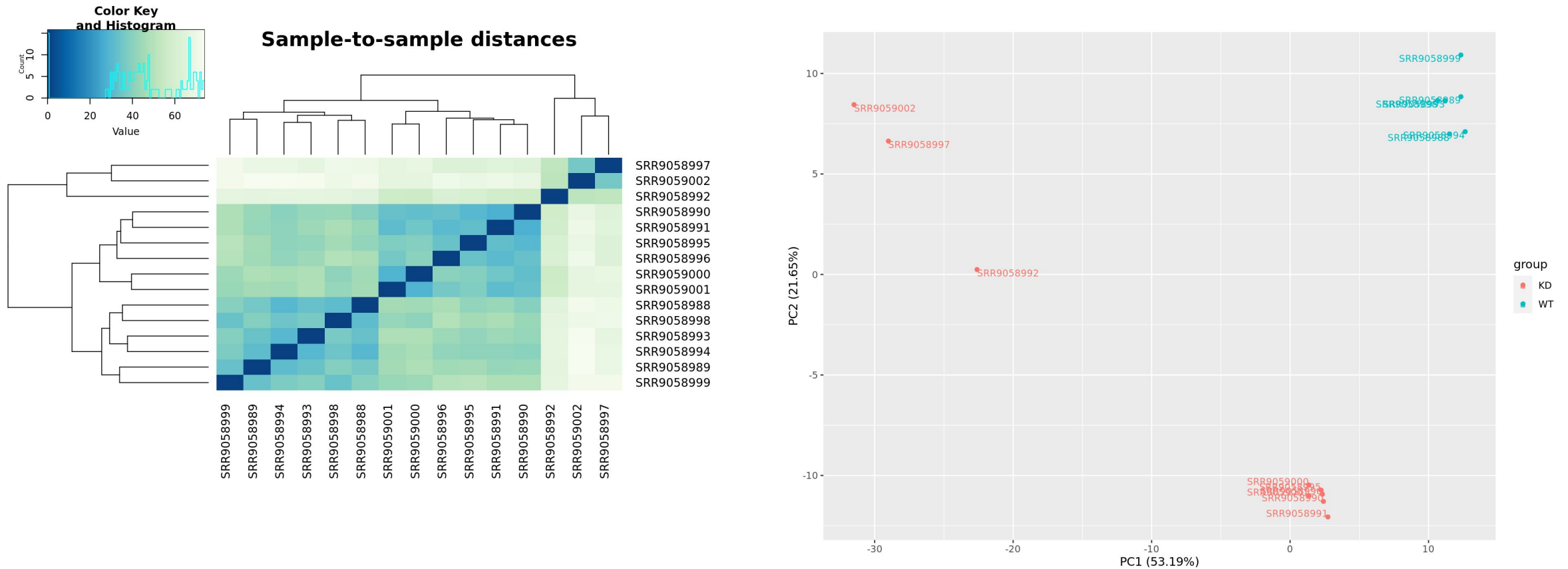
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...
...
...
GeneM	25	0	.	0

2: Running an RNAseq workflow

W
Workflow
Workflows are chains of interconnected tools.
[Create a Workflow](#)
[Learn how to build a workflow](#)



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



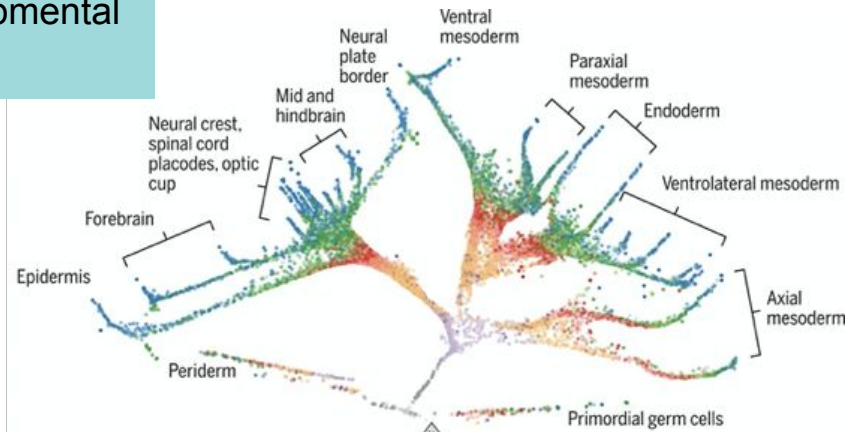


Introduction to Single-Cell Analysis

Why study single cells?

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

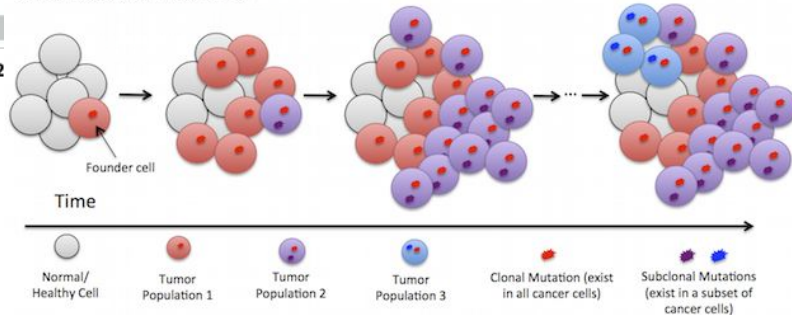
Developmental biology



Reconstruction of developmental trajectories

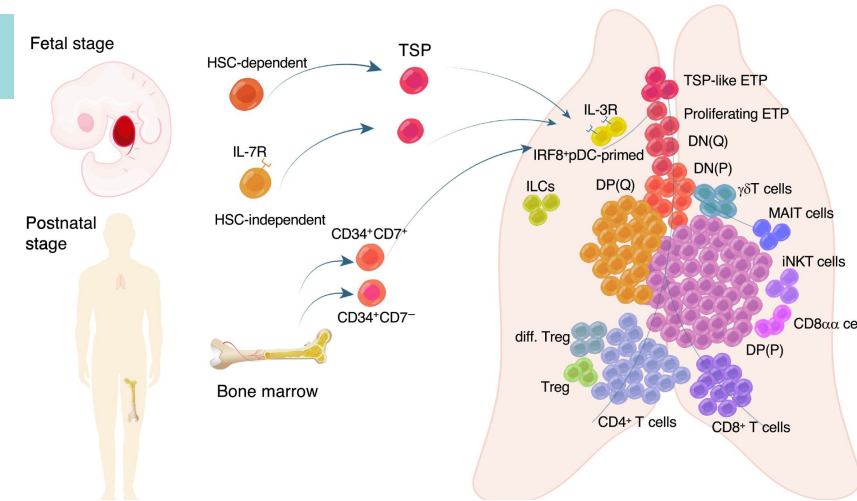
Clonal Theory (Nowell 1976)

Jeffrey A. Farrell et al. Science 2

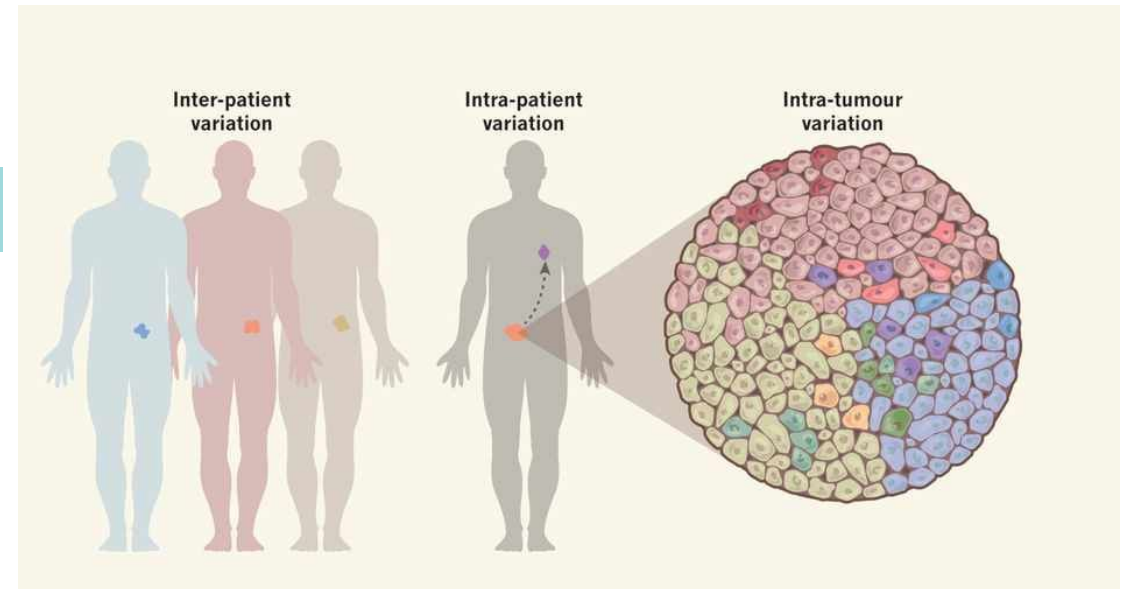


Cancer

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



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

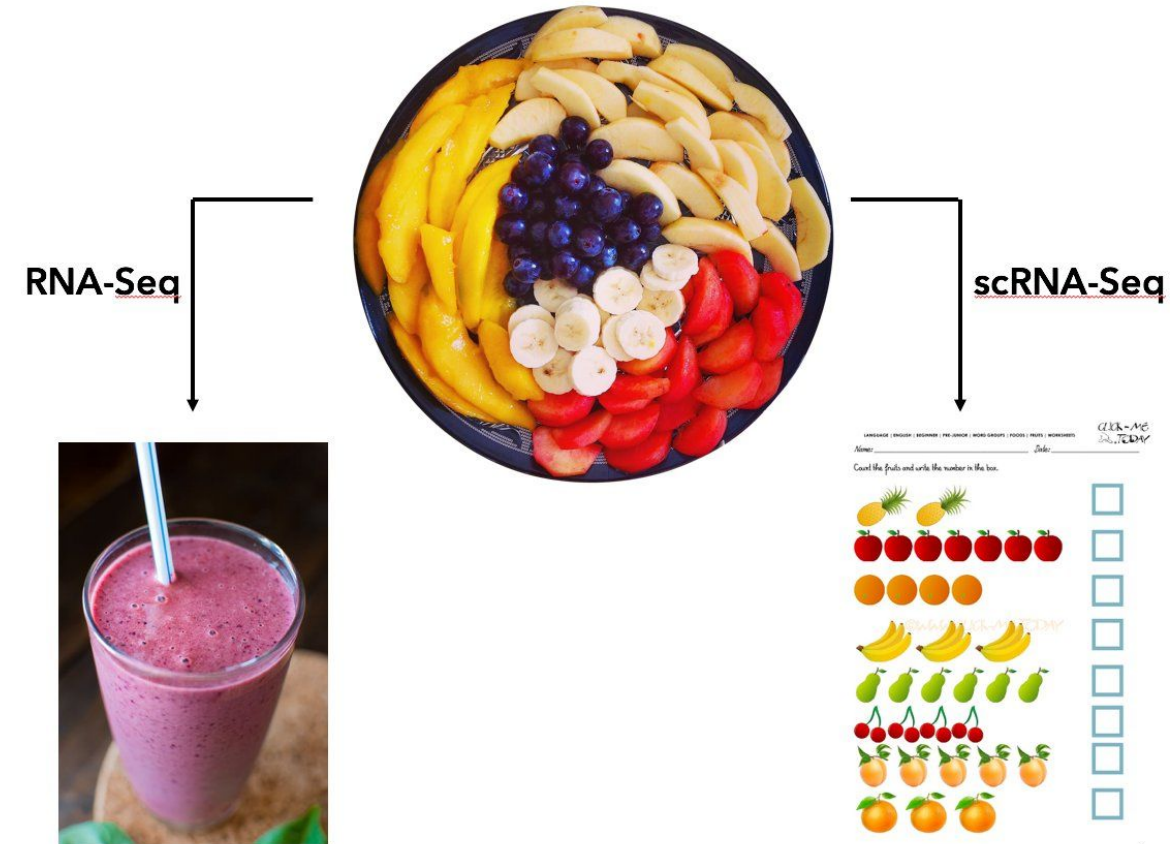


Figure by Roby Bhattacharyya; Smoothie concept by Aviv Regev

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



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

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

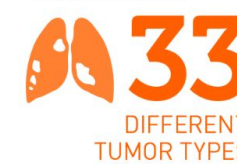
TCGA produced over



To put this into perspective, 1 petabyte of data is equal to



TCGA data describes



...including



...based on paired tumor and normal tissue sets collected from



...using



HOW BIG IS A PETABYTE?

11,000 4k movies

It would take you over 2.5 years of nonstop binge watching to get through a petabyte's worth of 4k movies

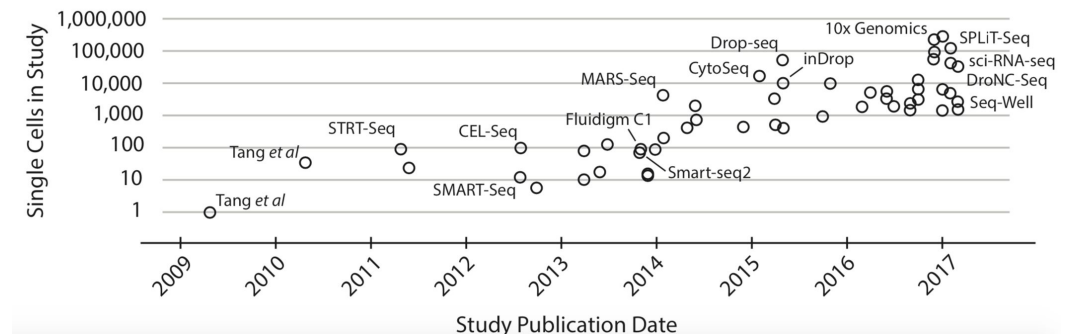
20+ PB of data
in the Library of Congress

If you took a petabyte's worth of 1GB flash drives and lined them up end to end, they would stretch over

92 football fields

4,000 digital photos
every day for the rest of your life

Sources: LifeWiz.com, Big-Science.com, cobalt IRON



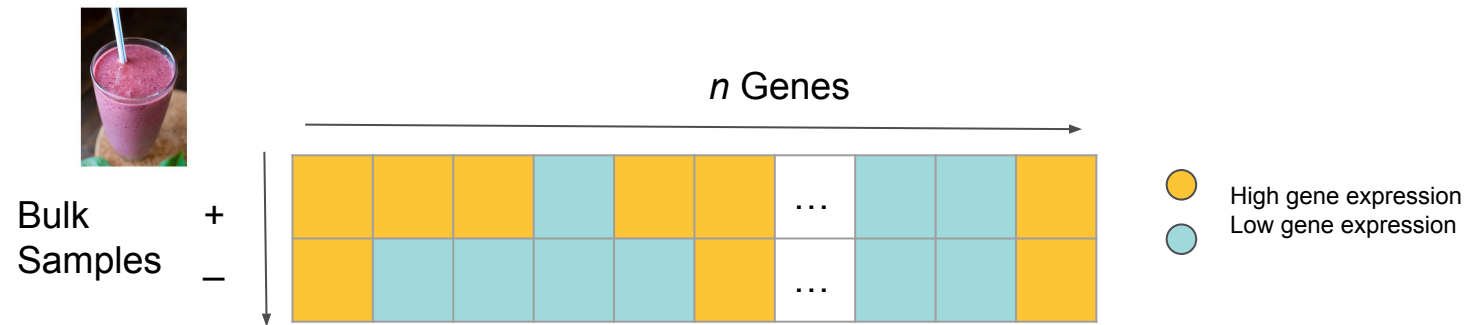
The Challenges of Single-Cell Analysis

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



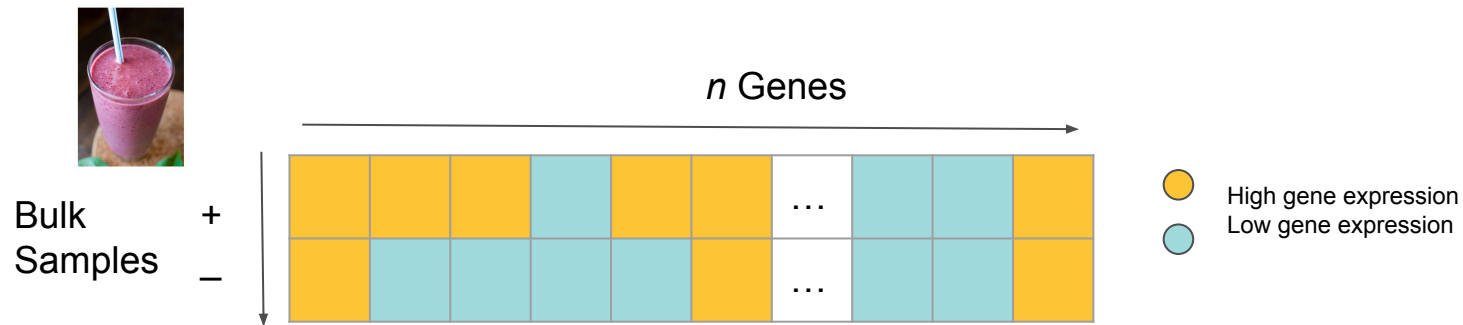
The Challenges of Single-Cell Analysis

Sparse, noisy data

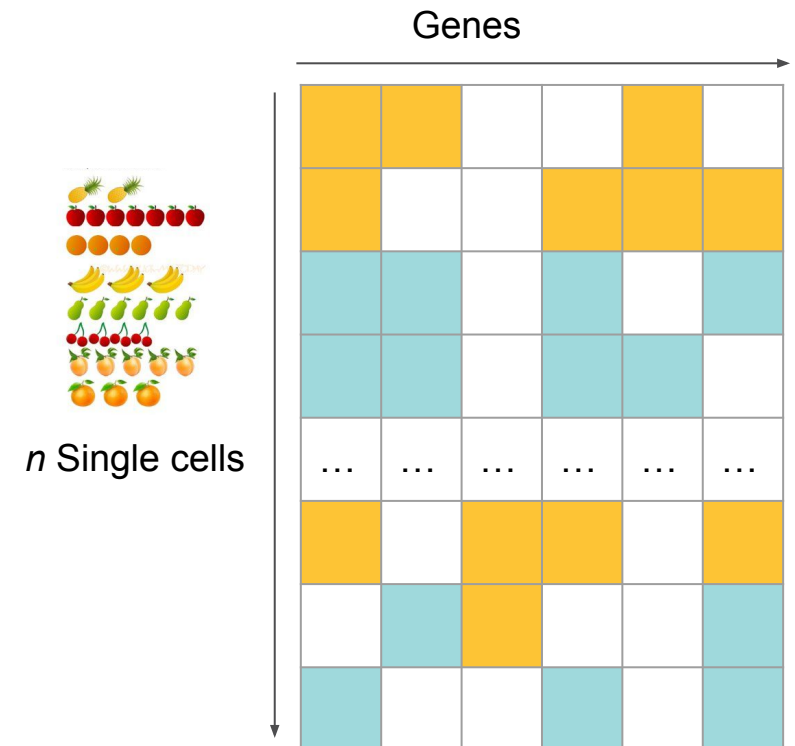
- Bulk methods usually have few samples, many genes
- Single cells can lack gene expression for many reasons
 - “Zero inflated” data
- Alternative methods to traditional RNAseq required

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.

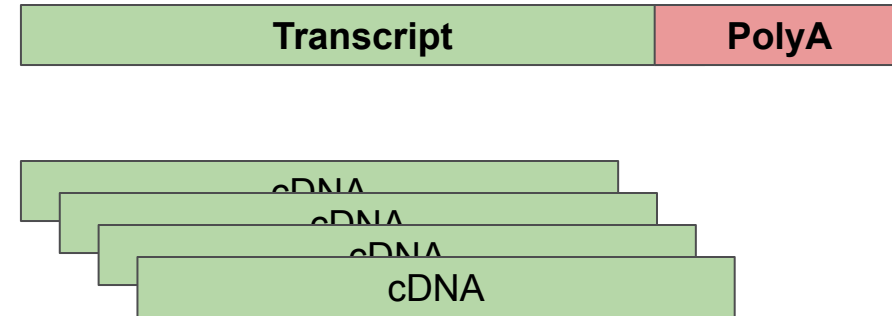




Methods for Single-Cell Analysis

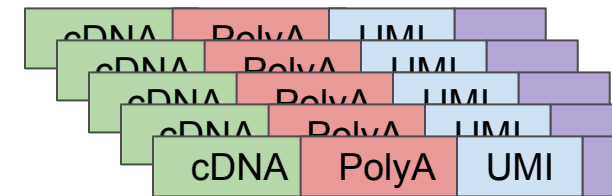
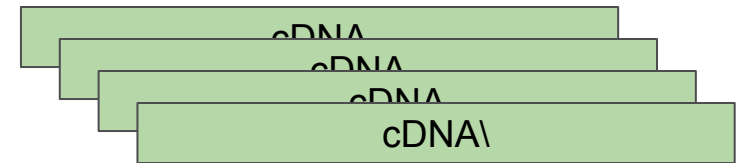
Two preparation methods for single cell RNAseq data

- **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 analysis

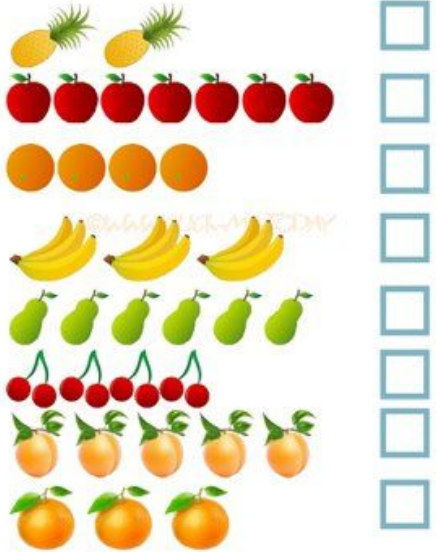
Higher level analysis

- 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]

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

*Available upon request

**Coming soon

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**]

Seurat
R CRAN





Single cell analysis with Seurat

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

Data

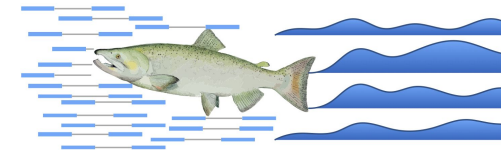
RNAseq data from the Human Cell Atlas data from Bone Marrow samples



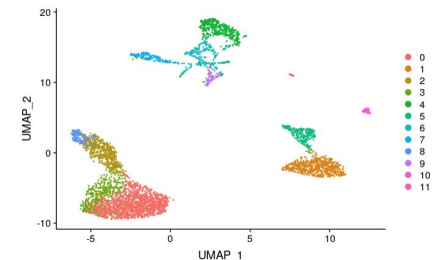
Tools

Salmon Alevin to align RNAseq data

- Optimized for single-cell data

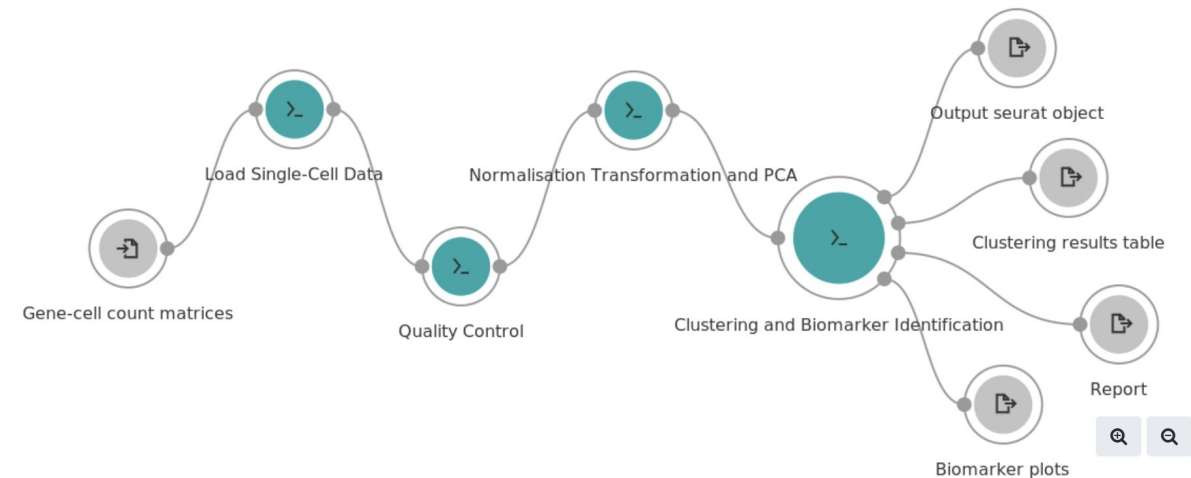


Seurat to cluster cells by similarity and identify marker genes

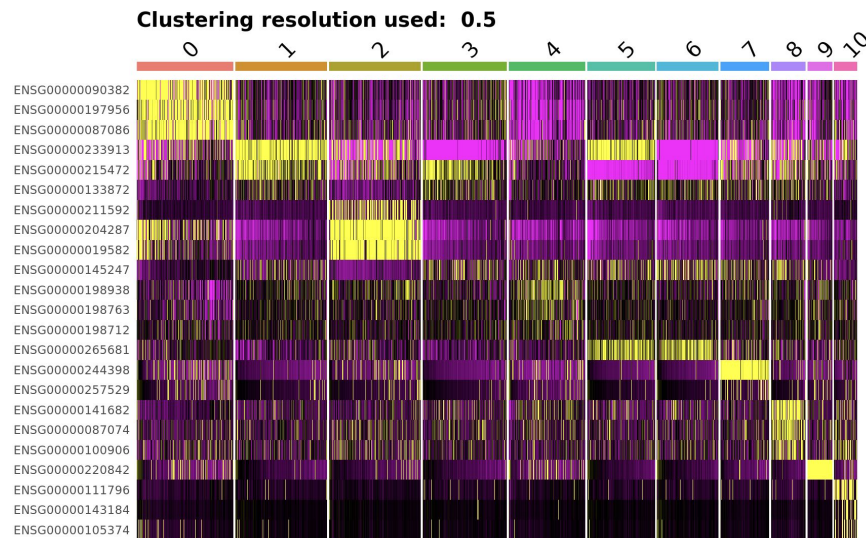
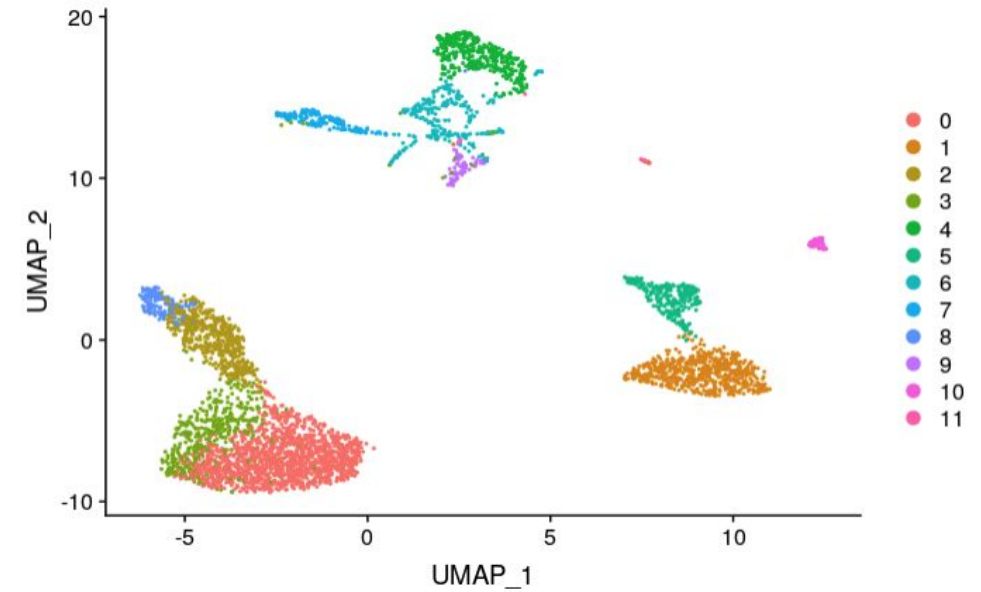
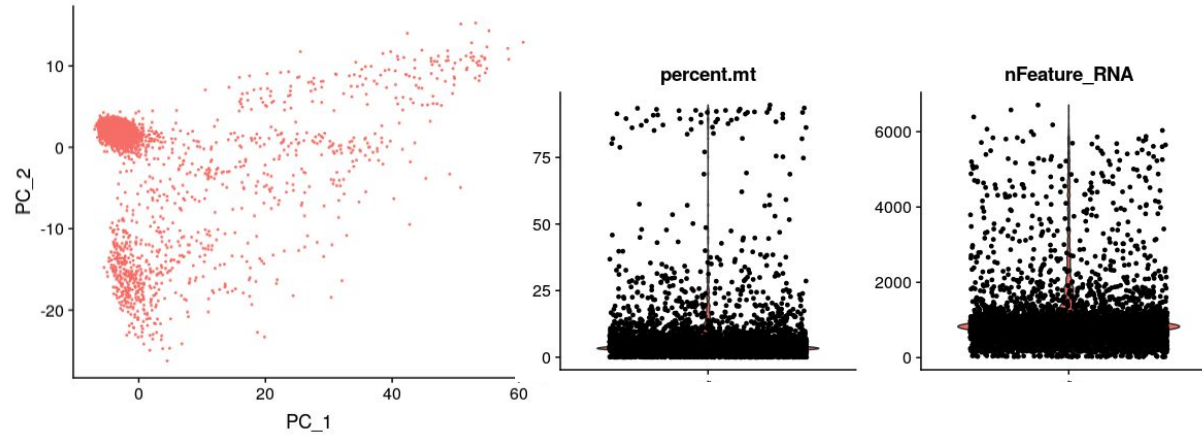


Clustering and Detection of Marker Genes: Seurat

- 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



Clustering and Detection of Marker Genes: Seurat

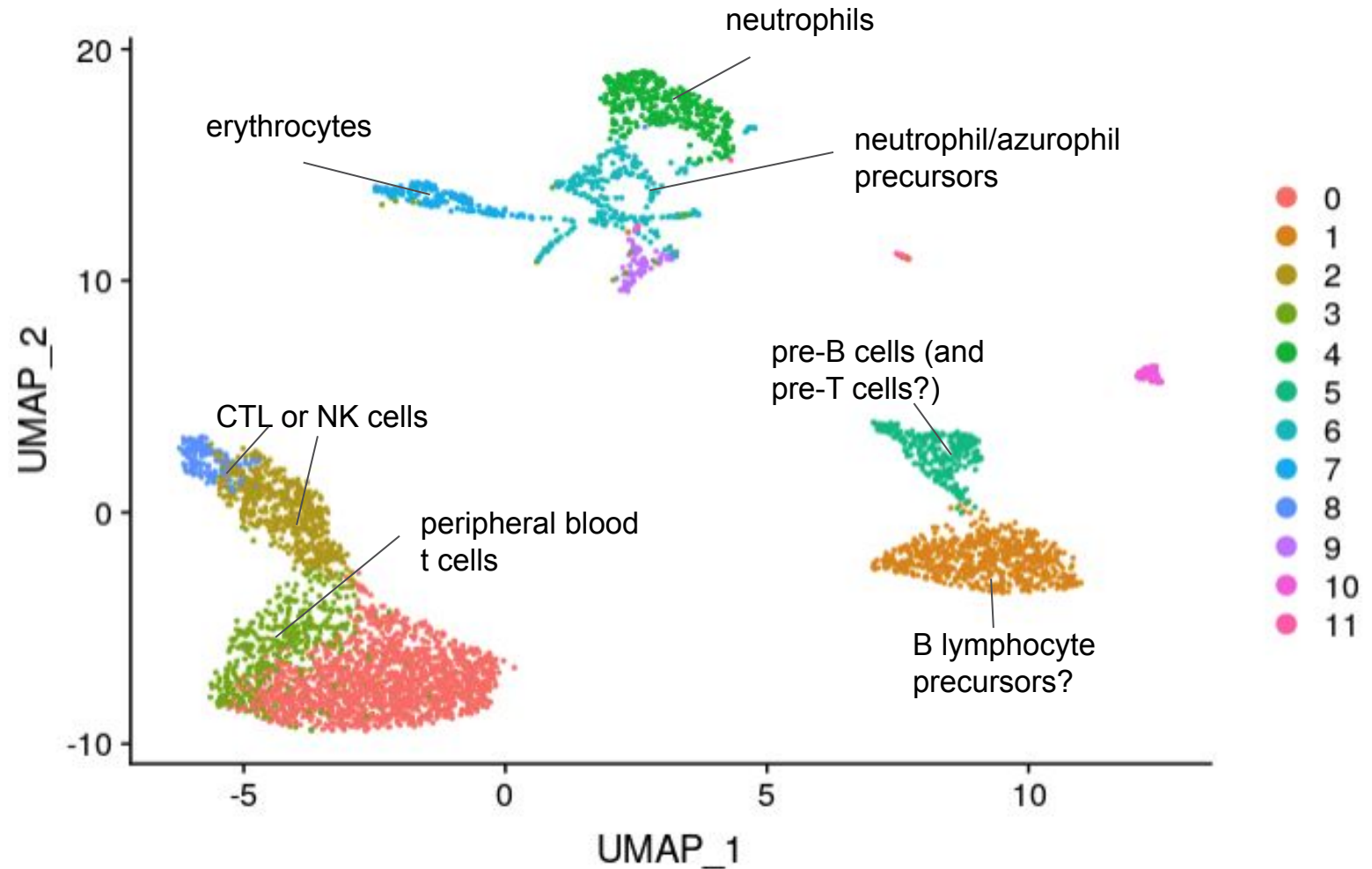


Tab 1. Top 10 differentially expressed genes for each cluster comparing to all the other clusters. Clustering resolution used: 0.5.

	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-C11
26	ENSG00000133872	SARAF	3	0.35	1.12e-45	1.54e-41	SARAF (Store-Operated Calc


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

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

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

Show all 

Output Settings

Compressed count matrix

[1_BM1_quants_mat.gz](#)

Compressed output directory

[1_BM1_alevin_output.tar.gz](#)

Output directory

 [1_BM1](#)

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

- 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

Show all

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

DRAFT Clustering and Gene Marker Identification with Seurat 3.2.2 run - 03-29-22 12:57:31 ✎ 👤 Get support 🗑 Discard ▶ Run

Last update by manisha_ray on Mar. 29, 2022 08:57
App: Clustering and Gene Marker Identification with Seurat 3.2.2 - Revision: 1

Task Inputs Execution Settings

Inputs

Batching ⓘ Off

Gene-cell count matrices * ⓘ 📄 Change selection

1_BM1_alevin_output.tar.gz

App Settings

✎ Edit parameters Show editable ▾

▼ **Loading Single Cell RNA-seq Expression Data**
(#load_single_cell_data_1)

Input Type * ⓘ

alevin ⓘ

Minimum number of cells ⓘ

No value

Minimum number of genes ⓘ

No value

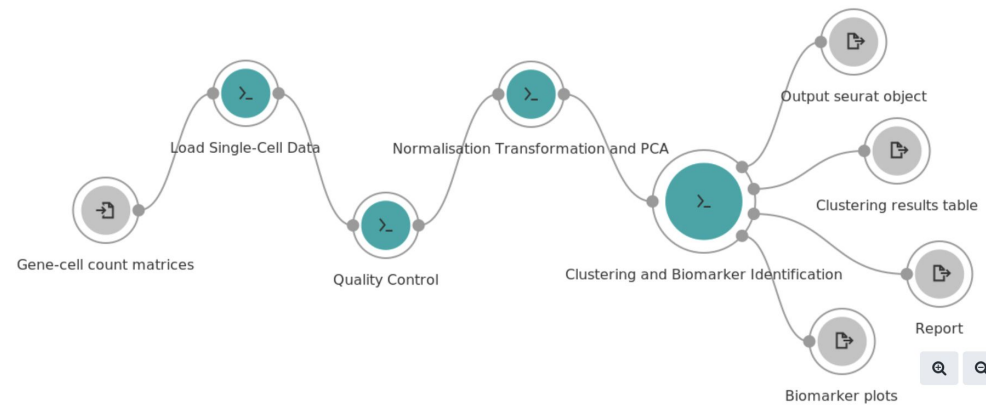
Output Settings

Biomarker plots ⓘ

Clustering results table ⓘ

Output seurat object ⓘ

Report ⓘ



Seurat inputs the gene count matrix output from Salmon Alevin

Continue interactively analyzing Seurat in RStudio

The screenshot displays the RStudio interface within a web browser. The browser address bar shows the URL: `cgc.sbgenomics.com/u/nemanja_vucic_cgc/single-cell-demonstration-with-salmon-alevin/analysis/cruncher/seurat-analysis/editor`. The RStudio window has a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help) and a toolbar. The script editor shows a file named `single_cell_interactive_analysis.Rmd` with the following content:

```
1 ---
2 title: "Single Cell RNA-Seq Clustering and Cluster Marker Identification Analysis"
3 output:
4   html_document:
5     df_print: paged
6   toc: yes
7   html_notebook:
8     toc: yes
9   date: "`r Sys.Date()`"
10 ---
11
12 #### Introduction
13
14 This is an interactive analysis for performing Clustering and Cluster Marker Identification analysis on scRNA-Seq data. It is created based on recommendations given by authors of Seurat, R package for QC analysis and exploration of scRNA-seq data.
15
16 #### Prepare environment
17
18 Install and load the required R packages, this may take a few minutes.
19
20 ```{r, message=FALSE}
21 source("dependencies.R")
22 ```
```

The Environment pane on the right shows the following data and functions:

Object	Description
seurat_df	90 obs. of 8 variables
seurat_object	Large Seurat (86.5 Mb)
seurat_object.markers	3388 obs. of 7 variables
all.genes	Large character (16504 elements, 1.1 Mb)
file_path	"/sbgenomics/project-files/HCATisStabAug177078016/alevin/quants_mat.gz"
hgnc_symbol	Named chr [1:90] "MS4A1" "CD79A" "CD79B" "HLA-DRA" "HLA-DRA" "CD74" ...
mt_genes	chr [1:13] "ENSG00000198888" "ENSG00000198763" "ENSG00000198804" ...
packages	chr [1:7] "BiocManager" "Seurat" "SummarizedExperiment" "tximport" "EnsDb.Hsap..."
ensemblPopSuffix	function (gene_ids)
loadCountMatrix	function (file_path, method)

The Files pane shows a list of files in the current directory:

Name	Size	Modified
.RData	354.6 MB	Sep 19, 2019, 3:48 PM
.Rhistory	5.2 KB	Sep 19, 2019, 3:47 PM
dependencies.R	972 B	Sep 19, 2019, 3:47 PM
helper.R	2.1 KB	Sep 19, 2019, 3:47 PM
single_cell_interactive_analysis.Rmd	7.2 KB	Sep 19, 2019, 3:47 PM

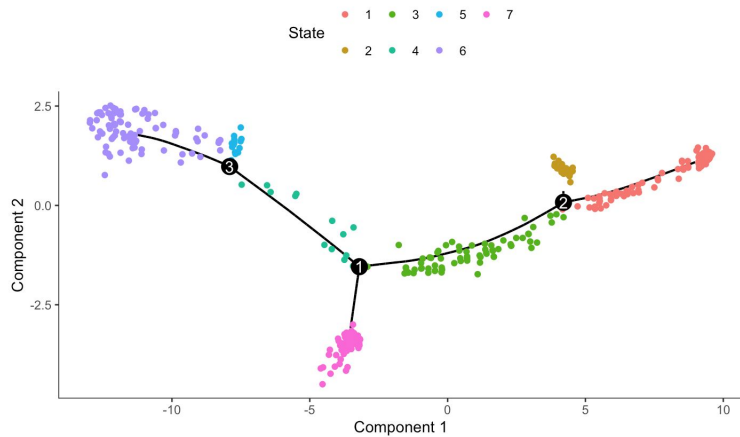
The console output shows the following messages:

```
there is no package called 'tximport' trying URL
'https://bioconductor.org/packages/3.9/bioc/src/contrib/tximport_1.12.3.tar.gz'
Content type 'application/x-gzip' length 231014 bytes (225 KB)
downloaded 225 KB

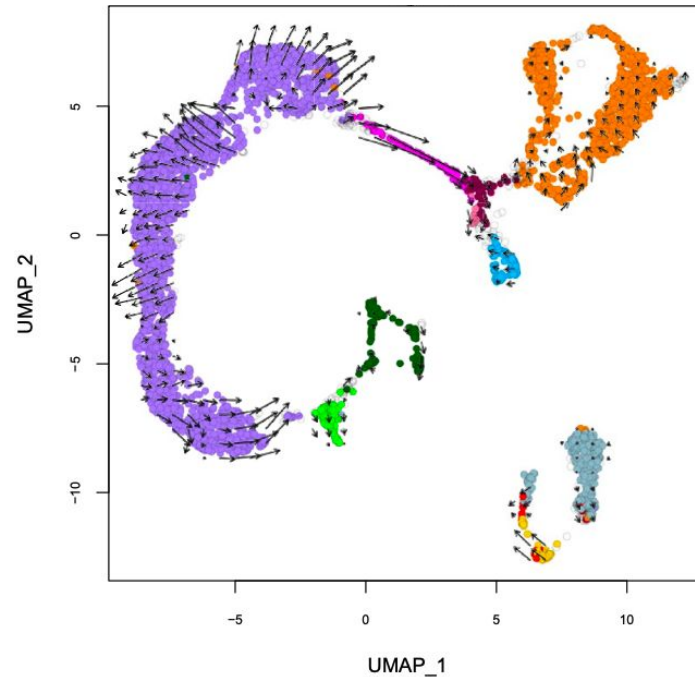
* installing *source* package 'tximport' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
```

Even higher level analysis of single-cells

Developmental trajectory inference with Monocle

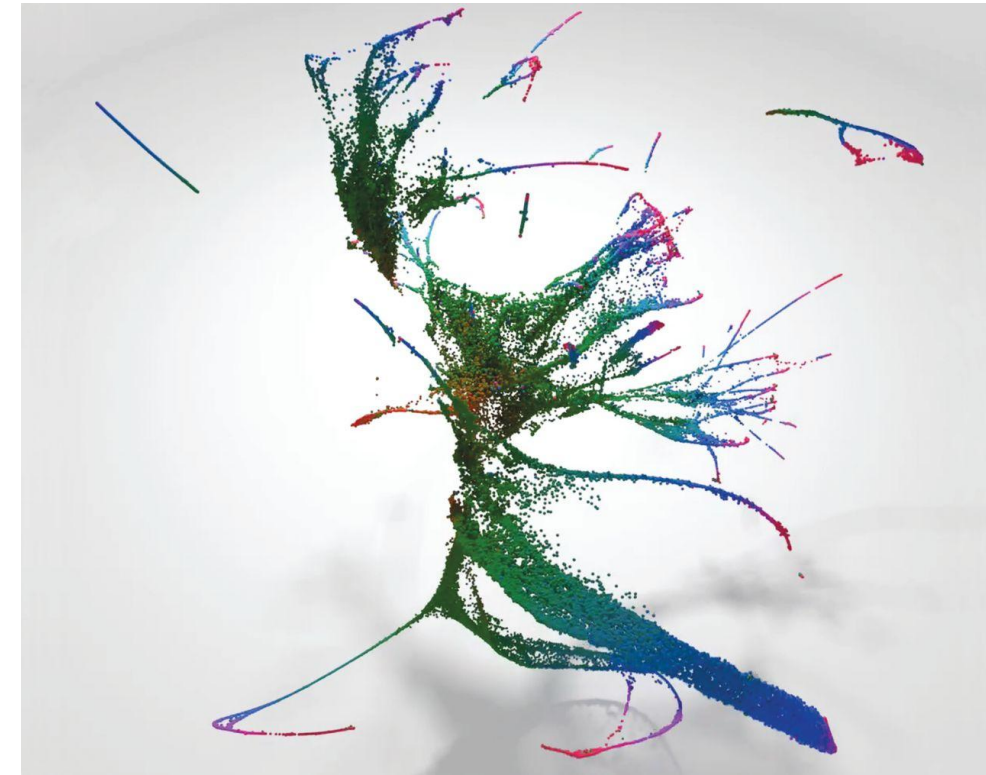


Velocity plot analysis of transcriptional dynamics



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



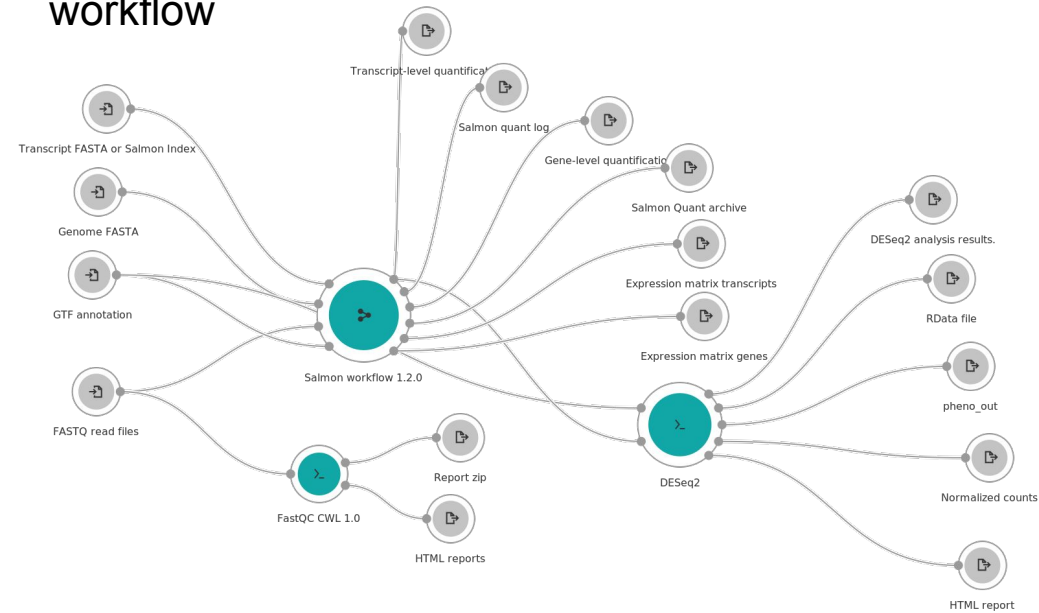
Jonathan S. Packer et al. *Science* 2019;365:eaax1971

Summary

Set up a project on the CGC and import data

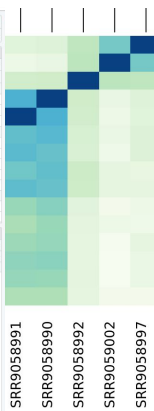
The screenshot shows the 'HTAN Testing' project page on the CGC platform. It includes a description, a list of members (nevena_vukojic and manisha_ray), and a list of analyses. The analyses section shows several runs of the 'HTAN Single Cell Workflow' and 'HTAN Workflow', with one run failing due to a file problem for sample 5.

Run an RNAseq workflow



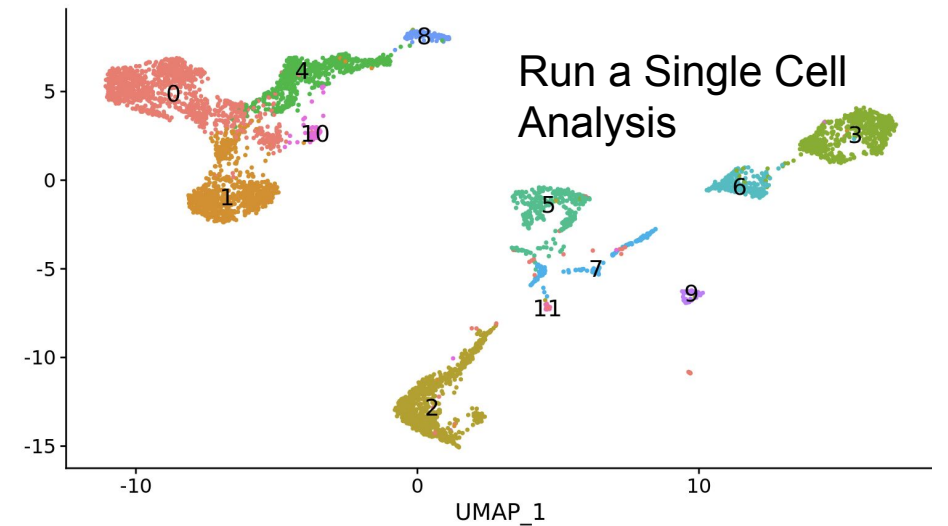
Differential Gene Expression using Data Cruncher with R

The screenshot shows an R script in RStudio. The script is titled 'Single Cell RNA-Seq Clustering and Cluster Marker Identification Analysis'. It includes comments and code for installing and loading R packages, and for performing clustering and marker identification analysis. The script is run in a Data Cruncher environment.



SRR9058997
SRR9059002
SRR9058992
SRR9058990
SRR9058991
SRR9058995
SRR9058996
SRR9059000
SRR9059001
SRR9058988
SRR9058998
SRR9058993
SRR9058994
SRR9058989
SRR9058999

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



<https://www.cancergenomicscloud.org>

Acknowledgements:

Dr. Jeffrey Grover

Dr. Min Zhang

Dr. Nadia Atallah Lanman



Demo