UNIVERSITY OF MICHIGAN

CENTER FOR STATISTICAL GENETICS

Overview of single-cell RNA-seq

Biocomputing Bootcamp Day 5

Hyun Min Kang

University of Michigan

DNA, RNA, and protein…

(Warning: very simplified view)

- DNA encodes the same information across all the cells (if somatic mutations are ignored).

- **RNAs** are transcribed from DNA, at different levels by cell types, environment, or individuals,

- Proteins are product of RNAs, representing *genes.*

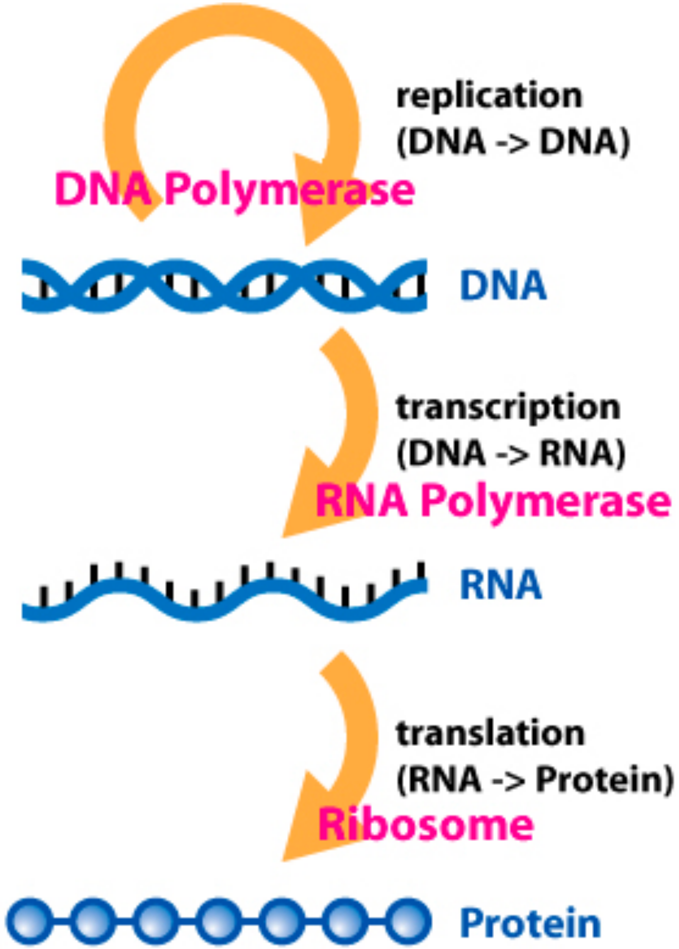
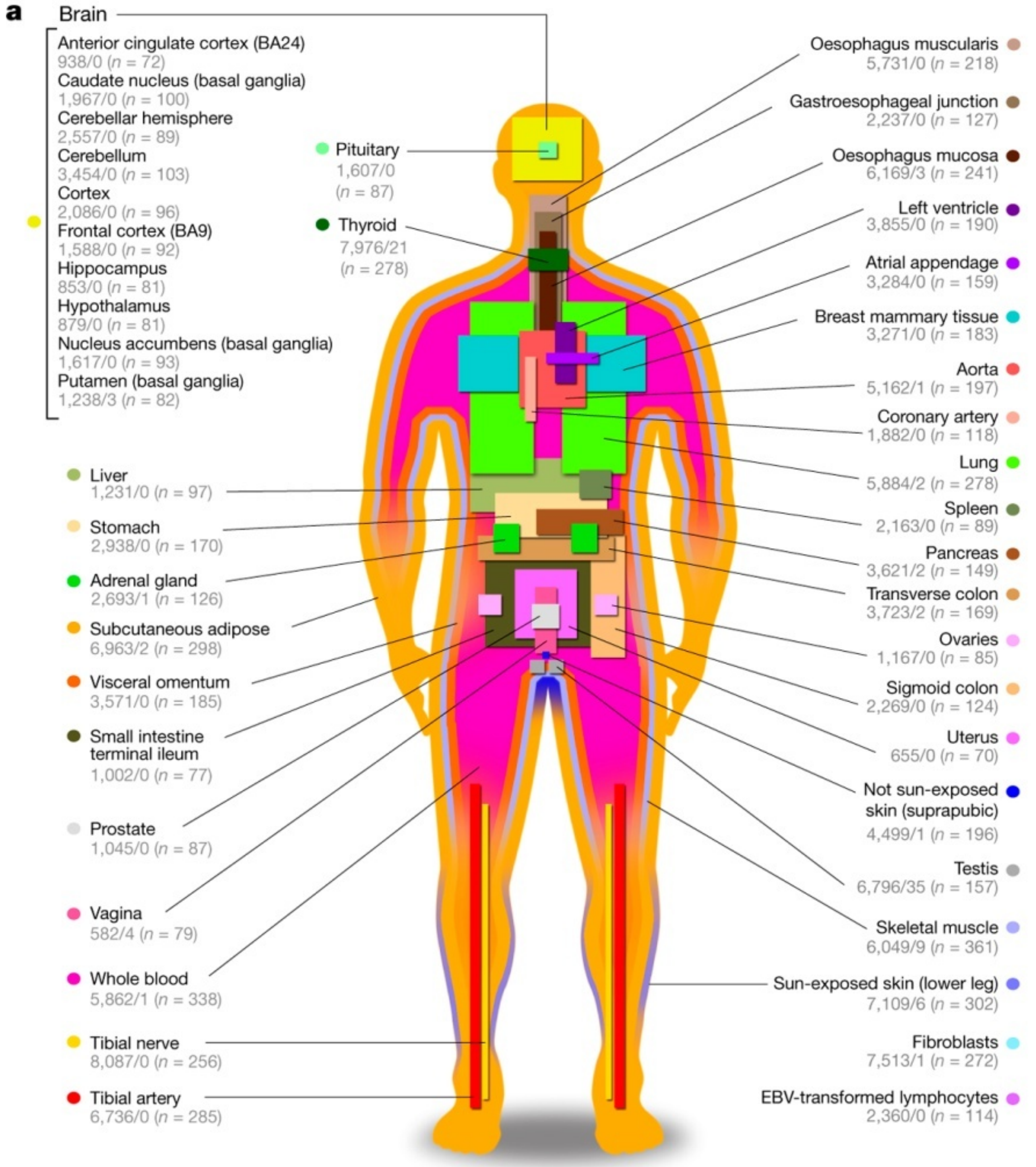


Image: https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Why sequence RNAs?

- Different tissues express various genes at different levels.
- RNA expression quantifies the relative abundance of each transcribed genes.
- Understanding how RNA levels change across conditions, individuals, and cell types is extremely important.

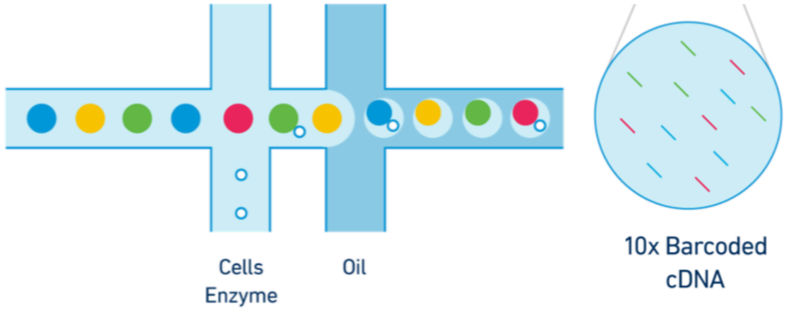


A Battle *et al.* *Nature* **550,** 204–213 (2017) doi:10.1038/nature24277

trans: 93 (FDR 10%)

Entering single cell RNA-seq…

Standard scRNA-seq analysis workflow



scRNA-seq experiment

Droplet Barcode	UMI	mRNA read (50-100bp)
AGCTGACGGCAT	TTACGCGG	ATGCGC...
AGCTGACGGCAT	TTACGCGG	AGCGTA...
AGCTGACGGCAT	AGCTTAGC	CTAGCT...
CGAAGTAGCTAG	GCCTGAAT	GTAGCC...
CGAAGTAGCTAG	GCCTGAAT	GTAGCC...

Raw sequence reads
(FASTQ)

STAR aligner

Aligned Reads

BAM

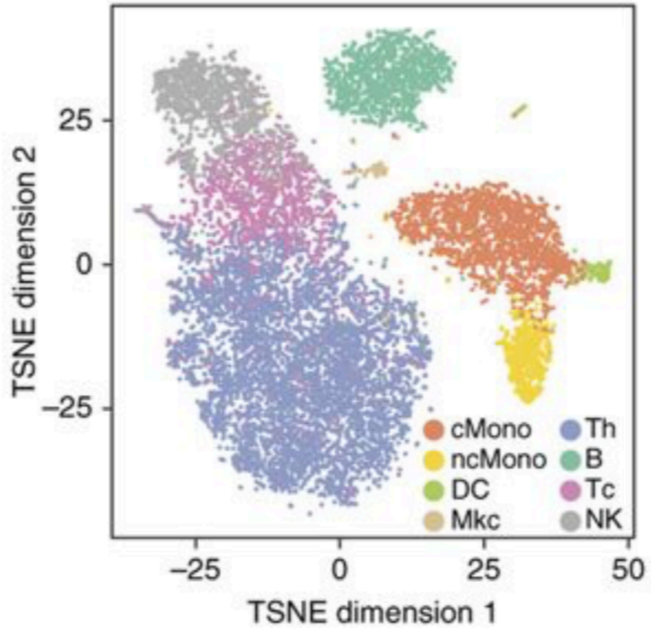
each record contains extra tags representing barcode & UMI

cellRanger, DropseqTools

Digital Expression Matrix

	Gene 1	Gene 2	...	Gene 20,000
Droplet 1	10	0		1
Droplet 2	0	1		0
...				
Droplet 5,000	1	5		0

Seurat



Digital expression matrix

Droplet Barcode	*CD3G*	*CD8A*	*CST3*	*MS4A7*	*LYZ*	*GNLY*	*S100A4*	*MS4A1*	*IL7R*
ACGTCATGCATA	0	0	0	0	1	0	3	0	3
AGTCATATACTA	0	0	1	0	3	0	0	6	2
CTAGATCGATTA	0	1	1	0	2	1	5	0	1
GCTAGTAGTTCA	0	0	22	3	24	0	16	0	0
CCGATCGATCTG	0	0	0	0	0	3	5	0	0
TGAGCTAGCTTG	1	1	0	0	1	0	9	0	0
AGATAGATCGAT	0	0	0	0	1	0	2	0	2
CGATCGQATCGT	0	0	0	0	1	0	0	0	3
TGATGCTAGCTA	0	0	0	0	0	1	3	1	0

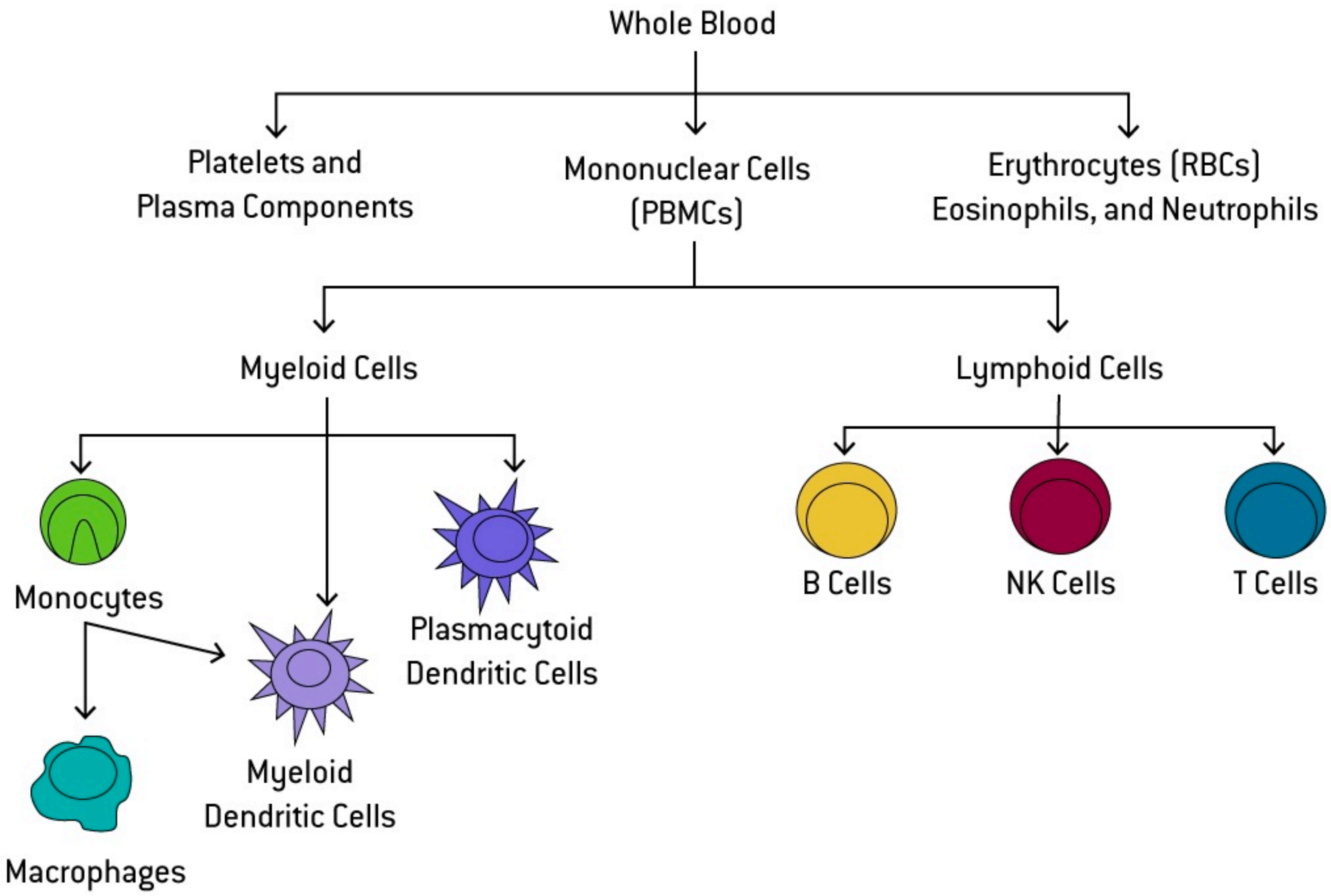
Sparse representation of digital expression

Index	Droplet Barcode
1	ACGTCATGCATA
2	AGTCATATACTA
3	CTAGATCGATTA
4	GCTAGTAGTTCA
5	CCGATCGATCTG
6	TGAGCTAGCTTG
7	AGATAGATCGAT
8	CGATCGQATCGT
9	TGATGCTAGCTA

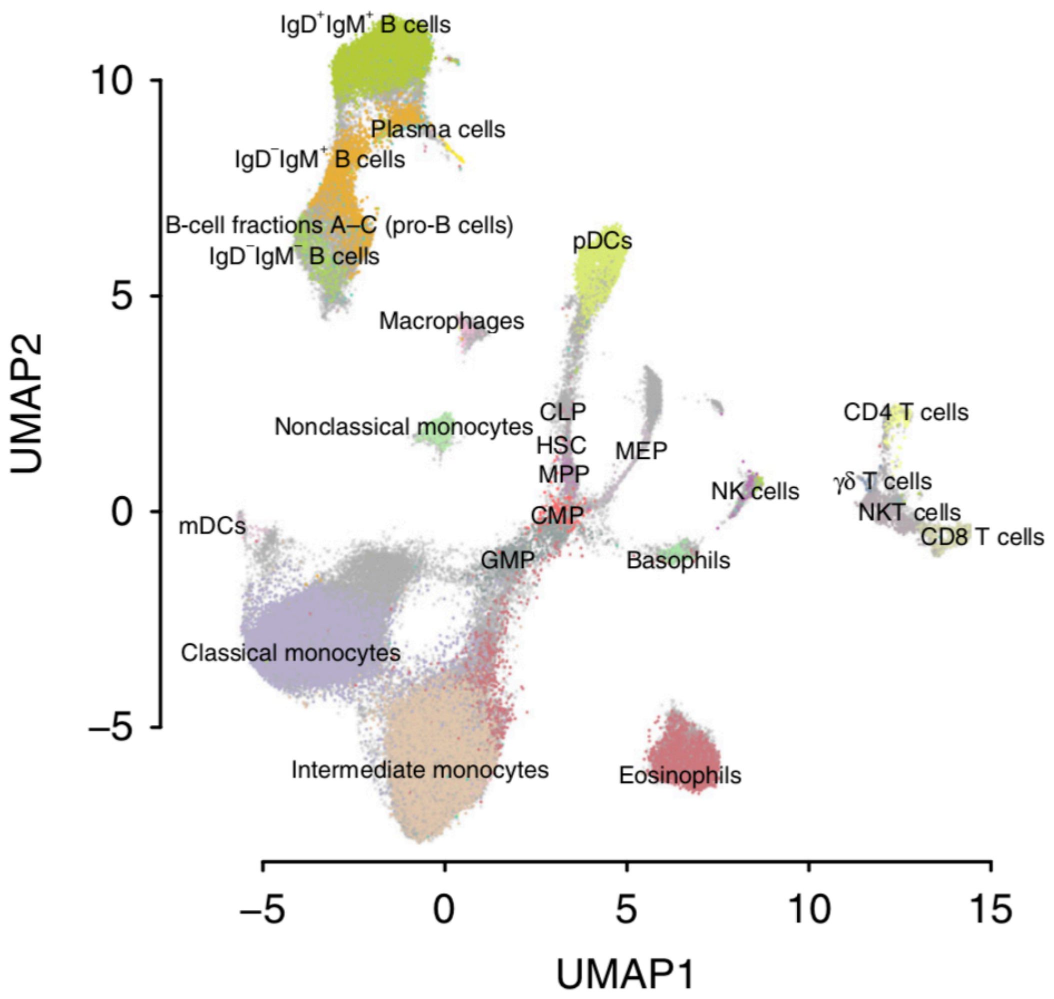
Index	Genes
1	*CD3G*
2	*CD8A*
3	*CST3*
4	*MS4A7*
5	*LYZ*
6	*GNLY*
7	*S100A4*
8	*MS4A1*
9	*IL7R*

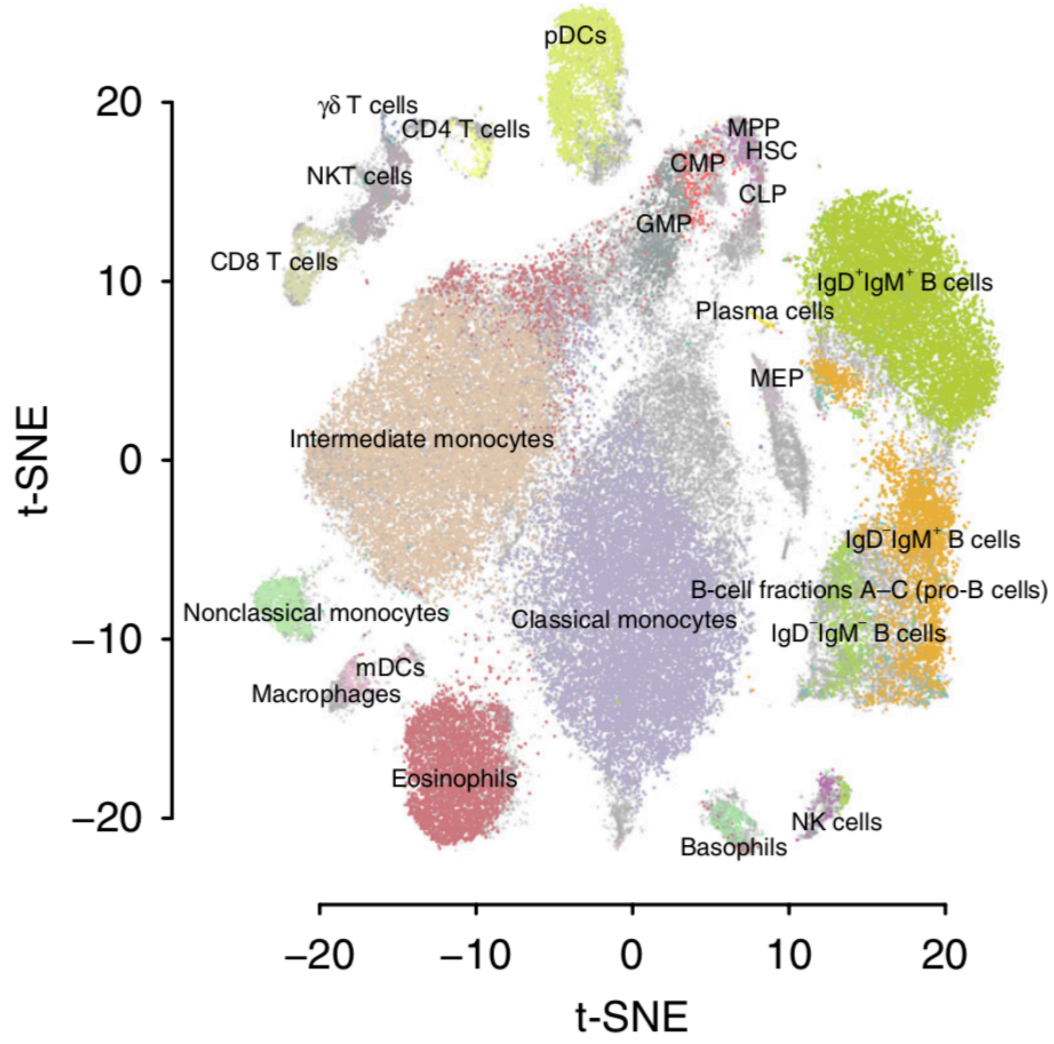
iGene	iBarcode	Count
1	6	1
2	3	1
2	6	1
3	2	1
3	3	1
3	4	22
4	3	3
5	1	1
5	2	3
5	3	2
5	4	24
...	...	...

Cell types in PBMCs


Whole Blood
Platelets and Plasma Components
Mononuclear Cells (PBMCs)
Erythrocytes (RBCs) Eosinophils, and Neutrophils
Myeloid Cells
Lymphoid Cells
Monocytes
Plasmacytoid Dendritic Cells
Macrophages
Myeloid Dendritic Cells
B Cells
NK Cells
T Cells

Manifold learning of single cells





Goal of today

- Learn how to read large data files using python and R
- Learn how to summarize the data and ask questions on them.
- Learn how to summarize the data visually.
- Learn how to apply existing methods on a large dataset.
- Learn how to perform statistical tests on a large dataset.