Summer Institutes of Statistical Genetics, 2023

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

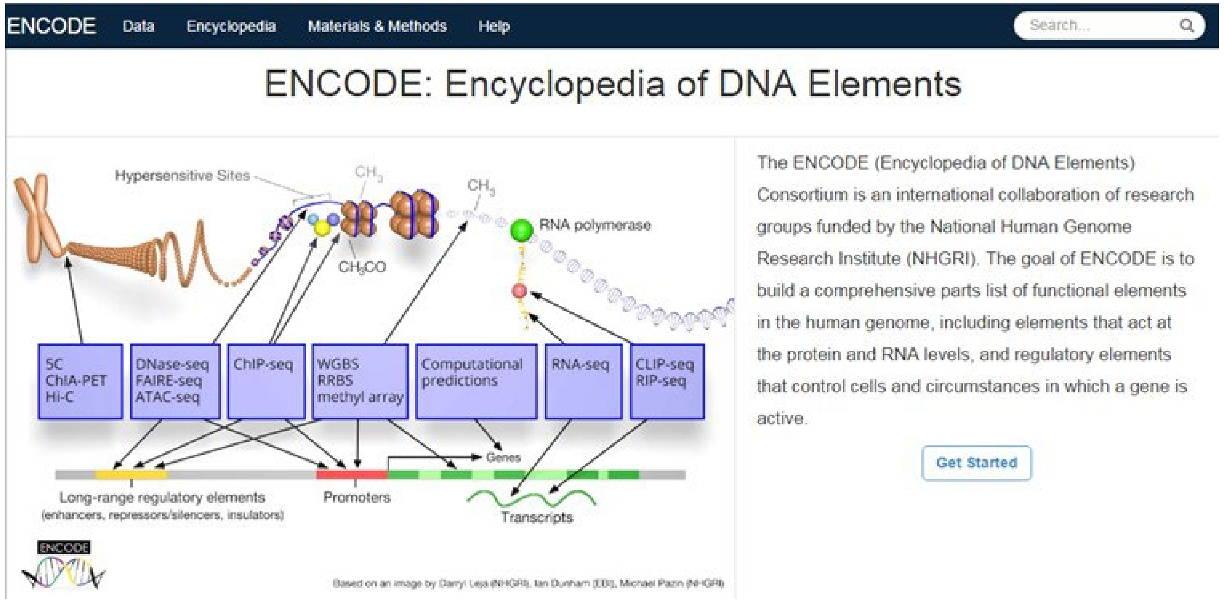
Georgia Institute of Technology

Lecture 6: EPIGENOMICS AND INTRO TO scRNAseq

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

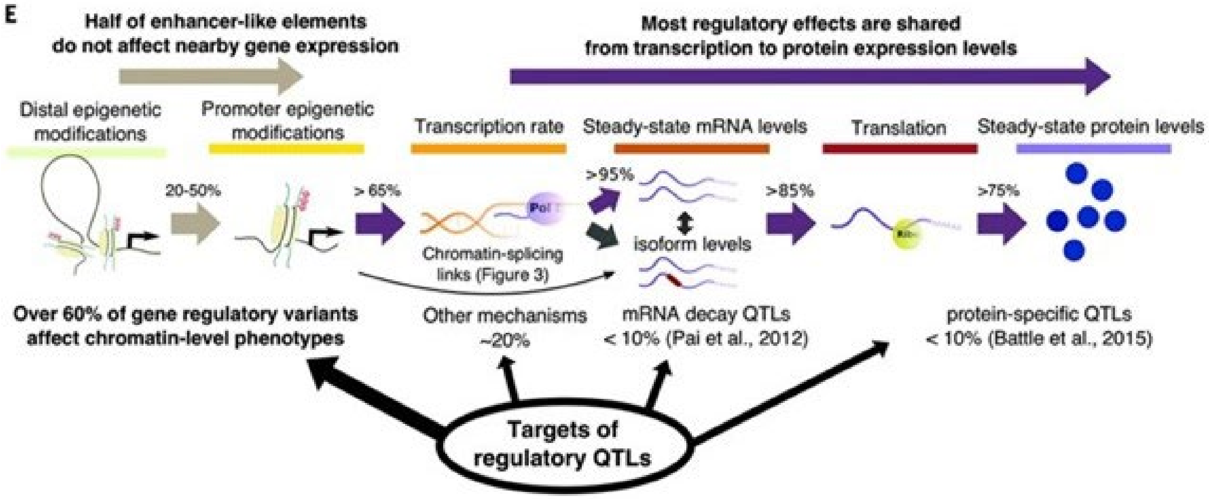
Epigenomics



ENCODE: Encyclopedia of DNA Elements

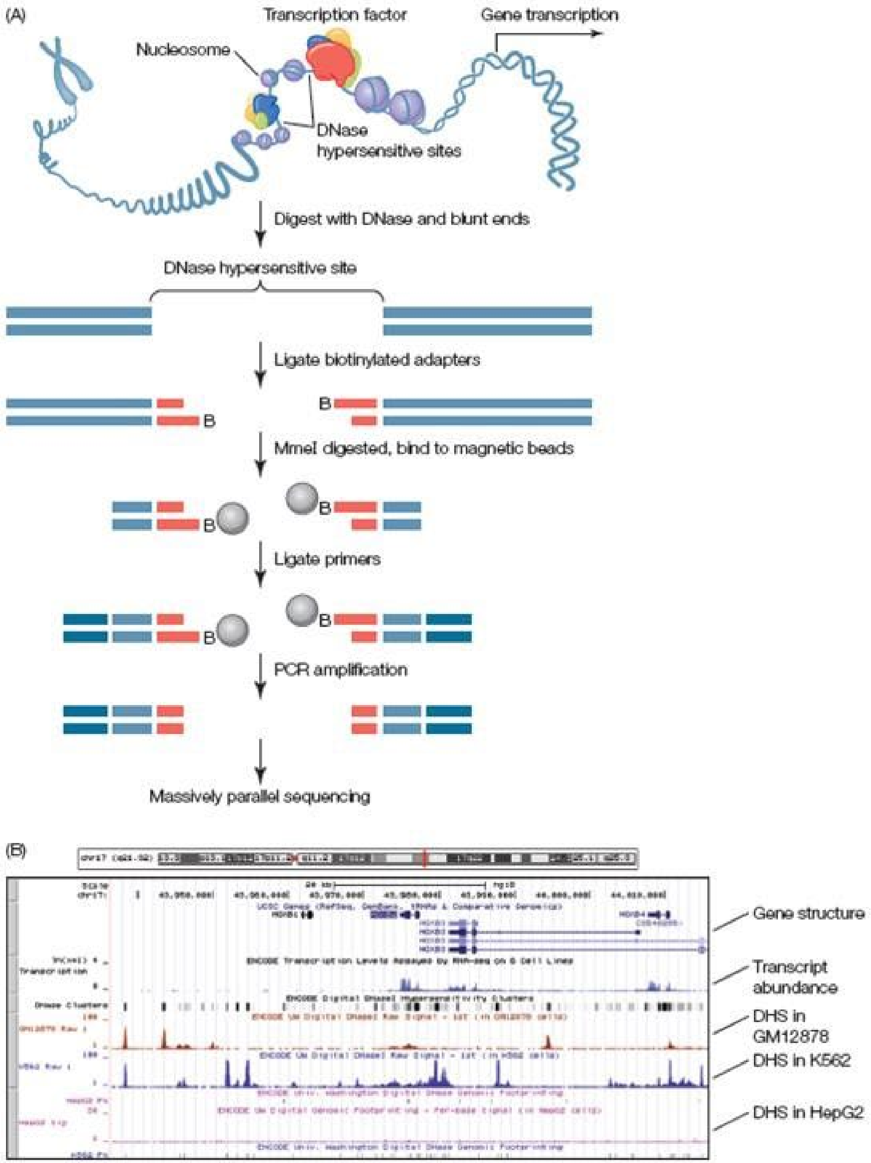
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

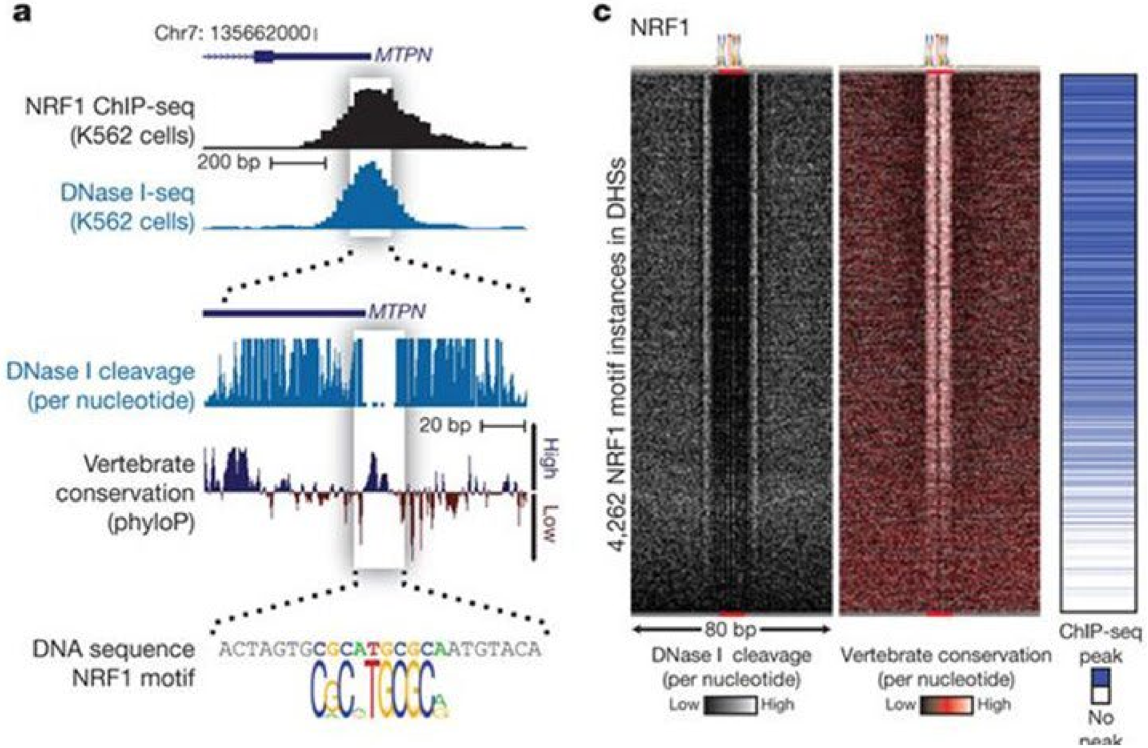
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)


E
Half of enhancer-like elements
do not affect nearby gene expression
Most regulatory effects are shared
from transcription to protein expression levels
Distal epigenetic modifications
Promoter epigenetic modifications
Transcription rate
Steady-state mRNA levels
Translation
Steady-state protein levels
20-50%
> 65%
>95%
>85%
>75%
Pol
Rib
isoform levels
Chromatin-splicing links (Figure 3)
Over 60% of gene regulatory variants affect chromatin-level phenotypes
Other mechanisms ~20%
mRNA decay QTLs < 10% (Pai et al., 2012)
protein-specific QTLs < 10% (Battle et al., 2015)
Targets of regulatory QTLs

Li et al (2016) Science **352**: 600-604

DHS and TFBS: DNAse hypersensitive sites and TF Binding





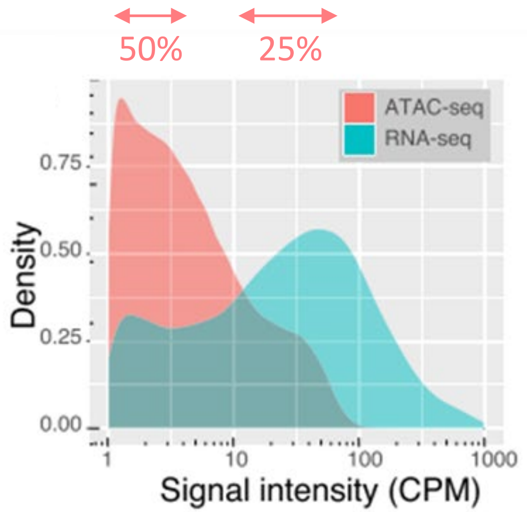
ATAC-Seq: Assay for Transposon Accessible Chromatin

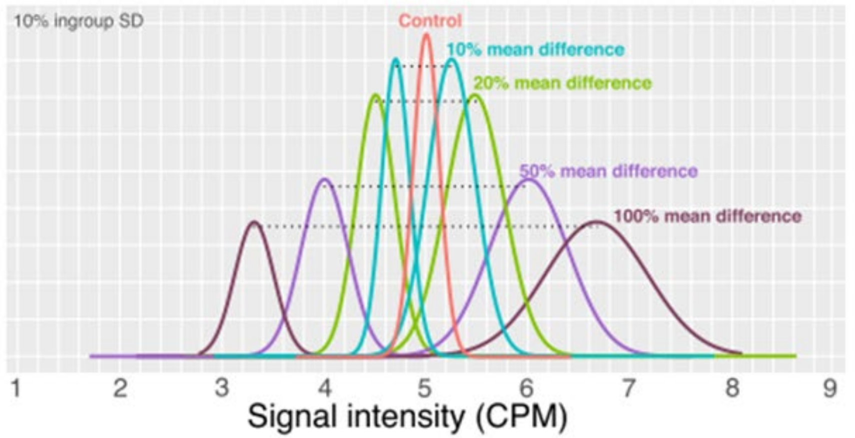
ACR = Accessible Chromatin Regions, also called OCR = Open Chromatin Regions

Peak calling generally looks for an excess (enrichment) of reads against the background, often using macs2 code

DAR analysis (Differential Accessible Region) is conceptually similar to DE, so edgeR and DEseq2 commonly used

Although bulk data is assumed to be negative-binomially distributed, single cell AC is 0, 1 or 2, and overall the frequency distribution is greatly skewed to low CPM since there are may be 50,000 ACR per cell-type





Gontarz P, et al. (2020) *Scientific Reports* **10**: 10150

ATAC-Seq Workflow

Quality Control:

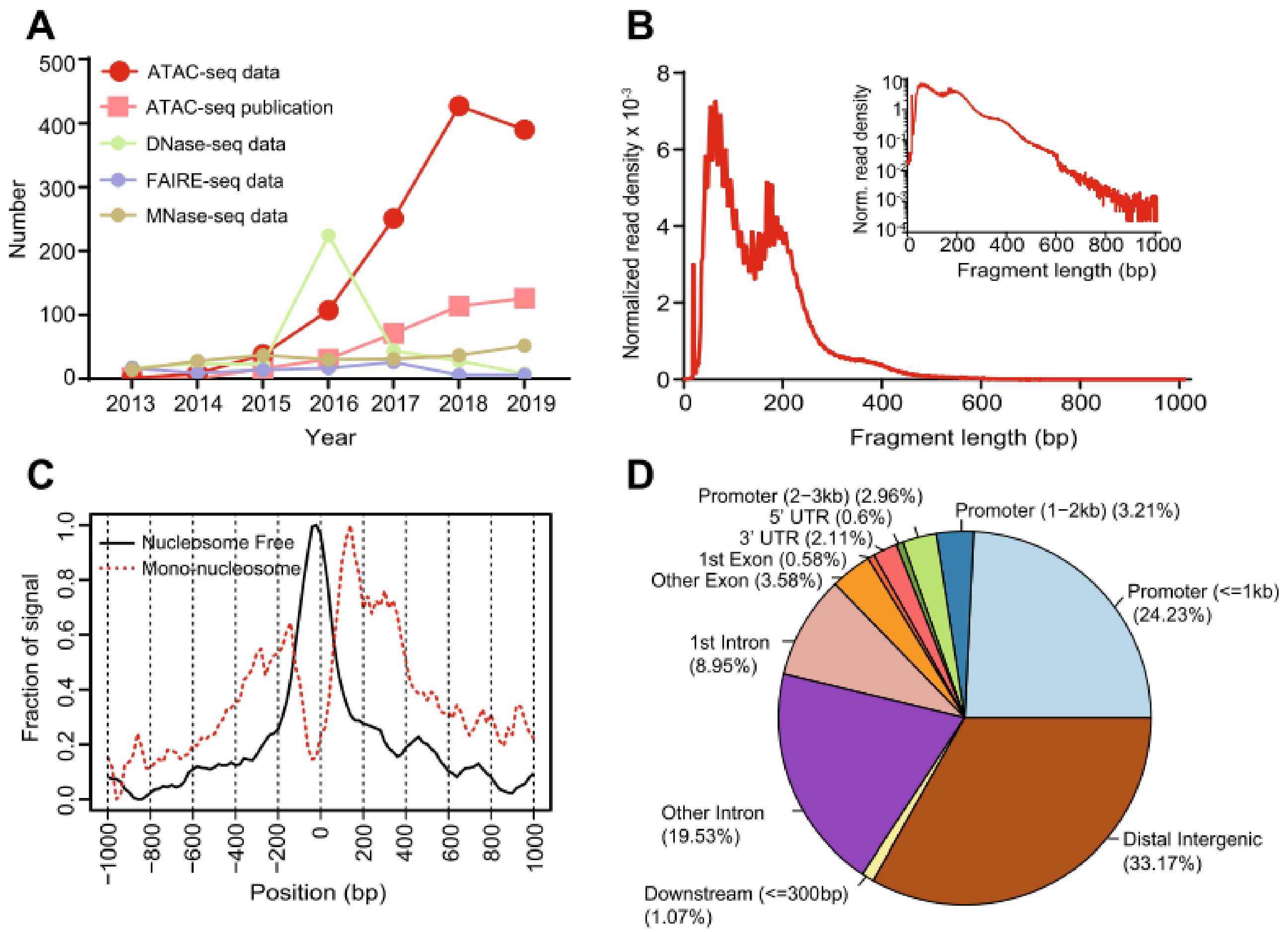
- Length distribution
- GC content
- Duplicates

- TSS enrichment
- Unique mapping

Analytical steps

- Peak calling
- DAR evaluation
- Gene annotation
- Motif detection

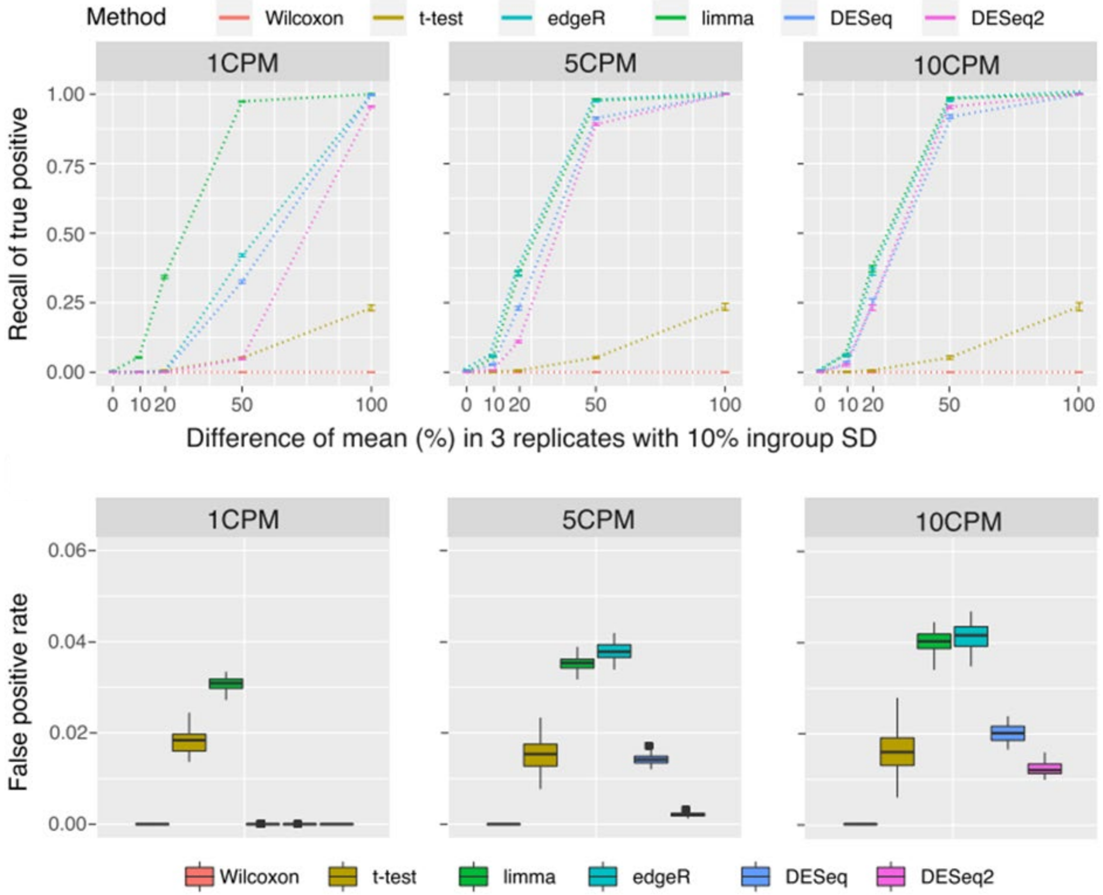
- Expression prediction
- Multi-omic integration
- Regulatory network inference



Yan F, Powell D, Curtis D, Wong N (2020) *Genome Biology* **21**: 22 “A Hitchhiker’s Guide to ATAC-seq data analysis”

Hypothesis testing for ATAC-Seq

The high percentage of zero counts per peak makes t-tests and Wilcoxon Rank Sum tests inappropriate
Gontarz et al compared limma, edgeR, and DESeq2 in a simulation study of low, moderate and high counts



DEseq2 has the best control of false positives while retaining sensitivity of limma

Note that actual power is a function of the sample (replication) size (here n=3/grp) and variance within groups.

Test performance changes a lot with sample size – Wilcoxon Rank Sum and t-tests better for n>15 and sequencing depth – recommend > 20M per sample

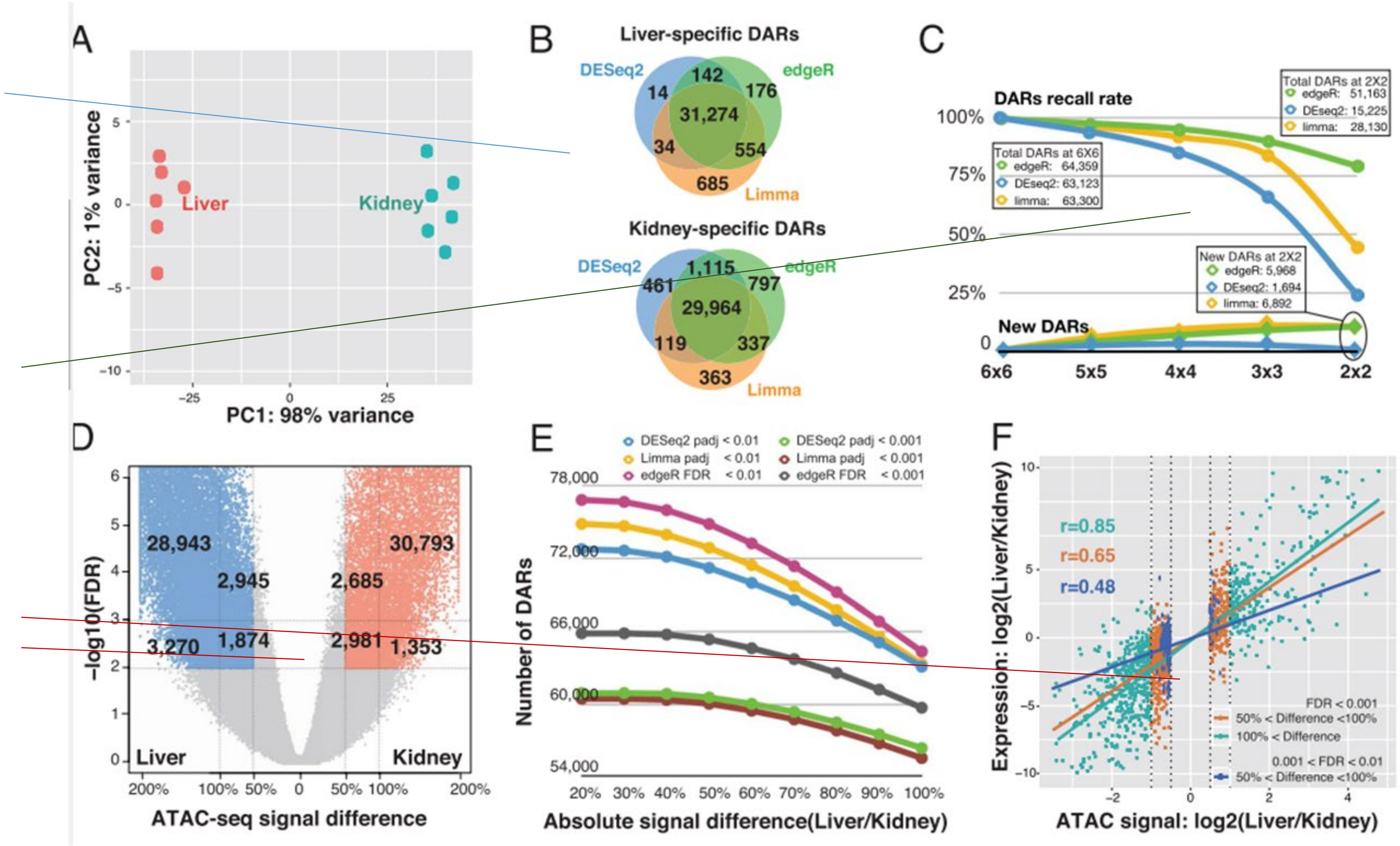
Gontarz P, et al. (2020) *Scientific Reports* **10**: 10150

ATAC-Seq method comparison on mouse kidney-liver dataset

Three methods perform comparably at stringent cutoff of $p<0.01$ and FC > 2

edgeR retains sensitivity after down-sampling to just 2 or 3 replicates

FC > 1.5 increases detection but still matches DAR to DE comparison in same samples



Gontarz P, et al. (2020) *Scientific Reports* **10**: 10150

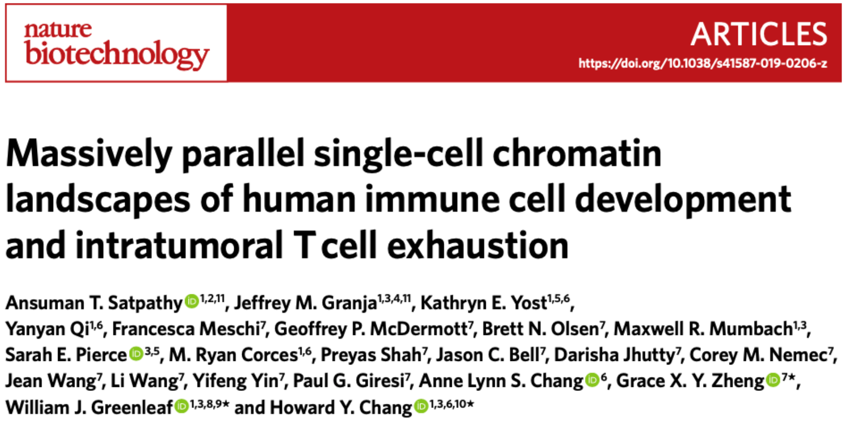
scATAC-seq

Assessment of computational methods for the analysis of single-cell ATAC-seq data
Chen, et al *Genome Biology* (2019) **20**: 241 recommends SnapATAC for single cell analysis.

Signac is the Seurat implementation.

We are finding that ArchR is easier to install and run, and meets most challenges.

ArchR

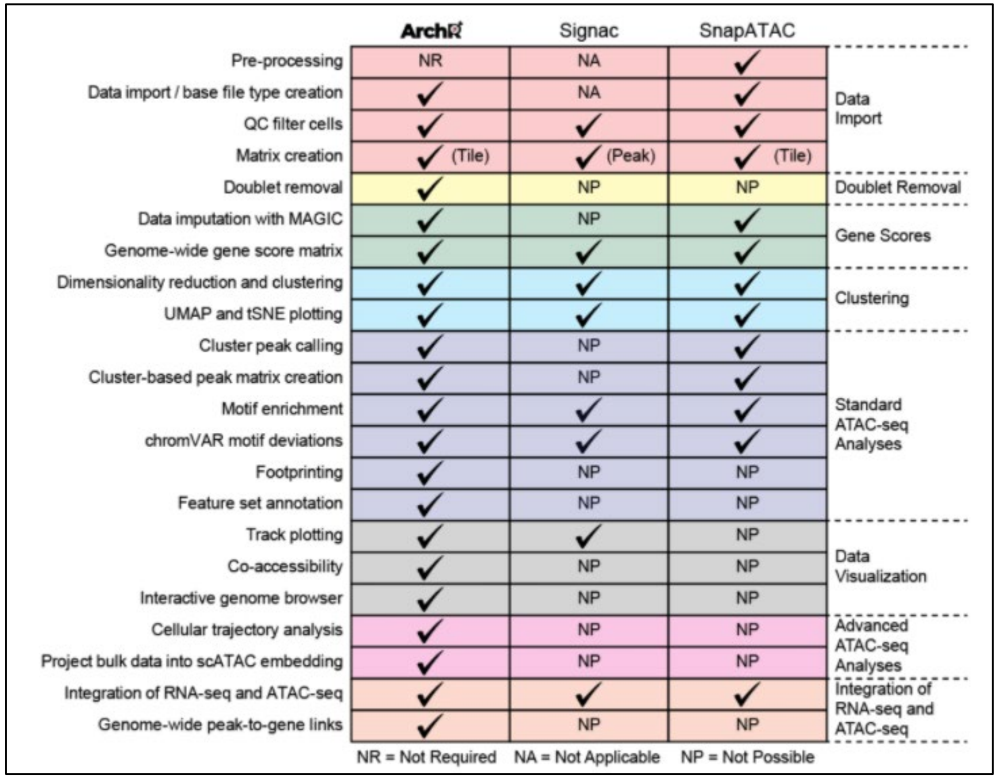


nature biotechnology | ARTICLES

https://doi.org/10.1038/s41587-019-0206-z

Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion

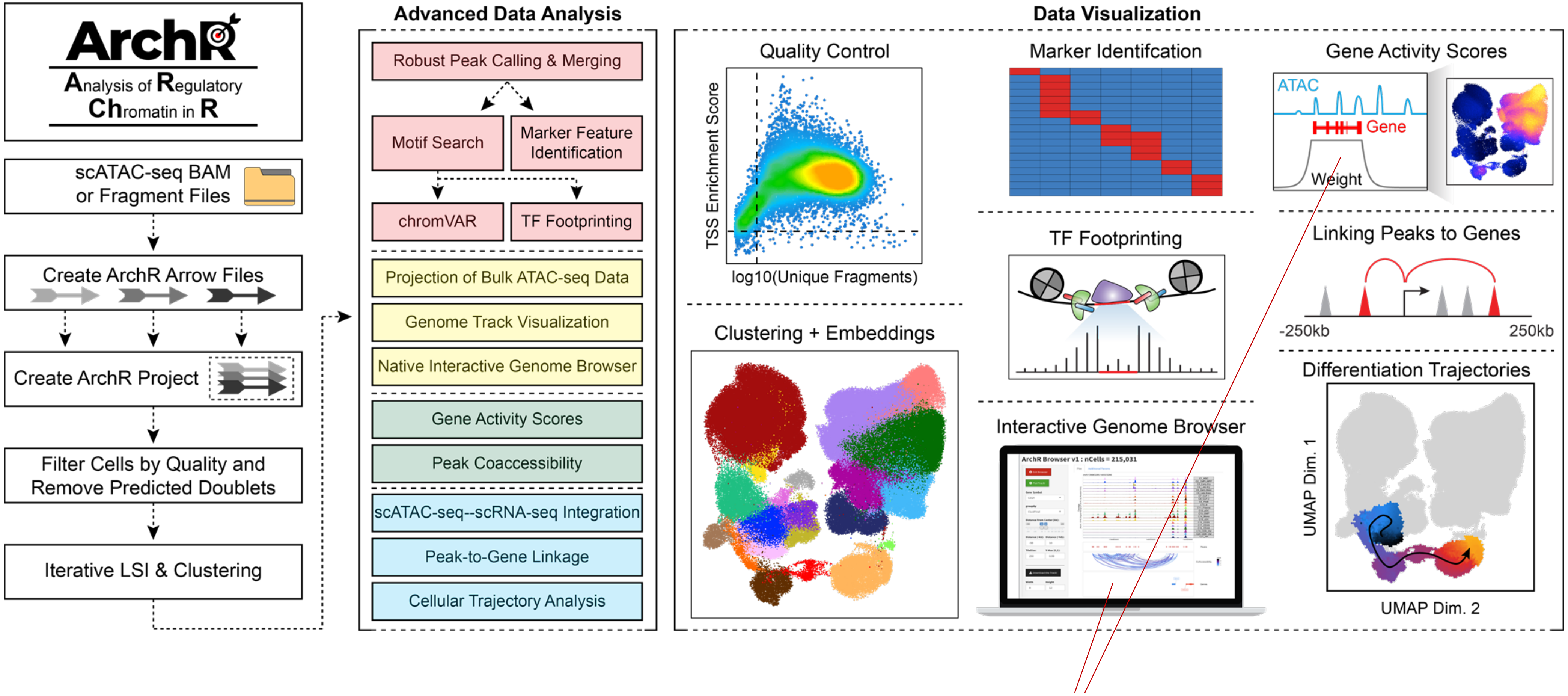
Ansuman T. Satpathy[1,2,11], Jeffrey M. Granja[1,3,4,11], Kathryn E. Yost[1,5,6], Yanyan Qi[1,6], Francesca Meschi[7], Geoffrey P. McDermott[7], Brett N. Olsen[7], Maxwell R. Mumbach[1,3], Sarah E. Pierce[3,5], M. Ryan Corces[1,6], Preyas Shah[7], Jason C. Bell[7], Darisha Jhutty[7], Corey M. Nemec[7], Jean Wang[7], Li Wang[7], Yifeng Yin[7], Paul G. Giresi[7], Anne Lynn S. Chang[6], Grace X. Y. Zheng[7]*, William J. Greenleaf[1,3,8,9]* and Howard Y. Chang[1,3,6,10]*



	ArchR	Signac	SnapATAC	
Pre-processing	NR	NA	✓	Data Import
Data import / base file type creation	✓	NA	✓	
QC filter cells	✓	✓	✓	
Matrix creation	✓ (Tile)	✓ (Peak)	✓ (Tile)	
Doublet removal	✓	NP	NP	Doublet Removal
Data imputation with MAGIC	✓	NP	✓	Gene Scores
Genome-wide gene score matrix	✓	✓	✓	
Dimensionality reduction and clustering	✓	✓	✓	Clustering
UMAP and tSNE plotting	✓	✓	✓	
Cluster peak calling	✓	NP	✓	Standard ATAC-seq Analyses
Cluster-based peak matrix creation	✓	NP	✓	
Motif enrichment	✓	✓	✓	
chromVAR motif deviations	✓	✓	✓	
Footprinting	✓	NP	NP	
Feature set annotation	✓	NP	NP	
Track plotting	✓	✓	NP	Data Visualization
Co-accessibility	✓	NP	NP	
Interactive genome browser	✓	NP	NP	
Cellular trajectory analysis	✓	NP	NP	Advanced ATAC-seq Analyses
Project bulk data into scATAC embedding	✓	NP	NP	
Integration of RNA-seq and ATAC-seq	✓	✓	✓	Integration of RNA-seq and ATAC-seq
Genome-wide peak-to-gene links	✓	NP	NP	

NR = Not Required NA = Not Applicable NP = Not Possible

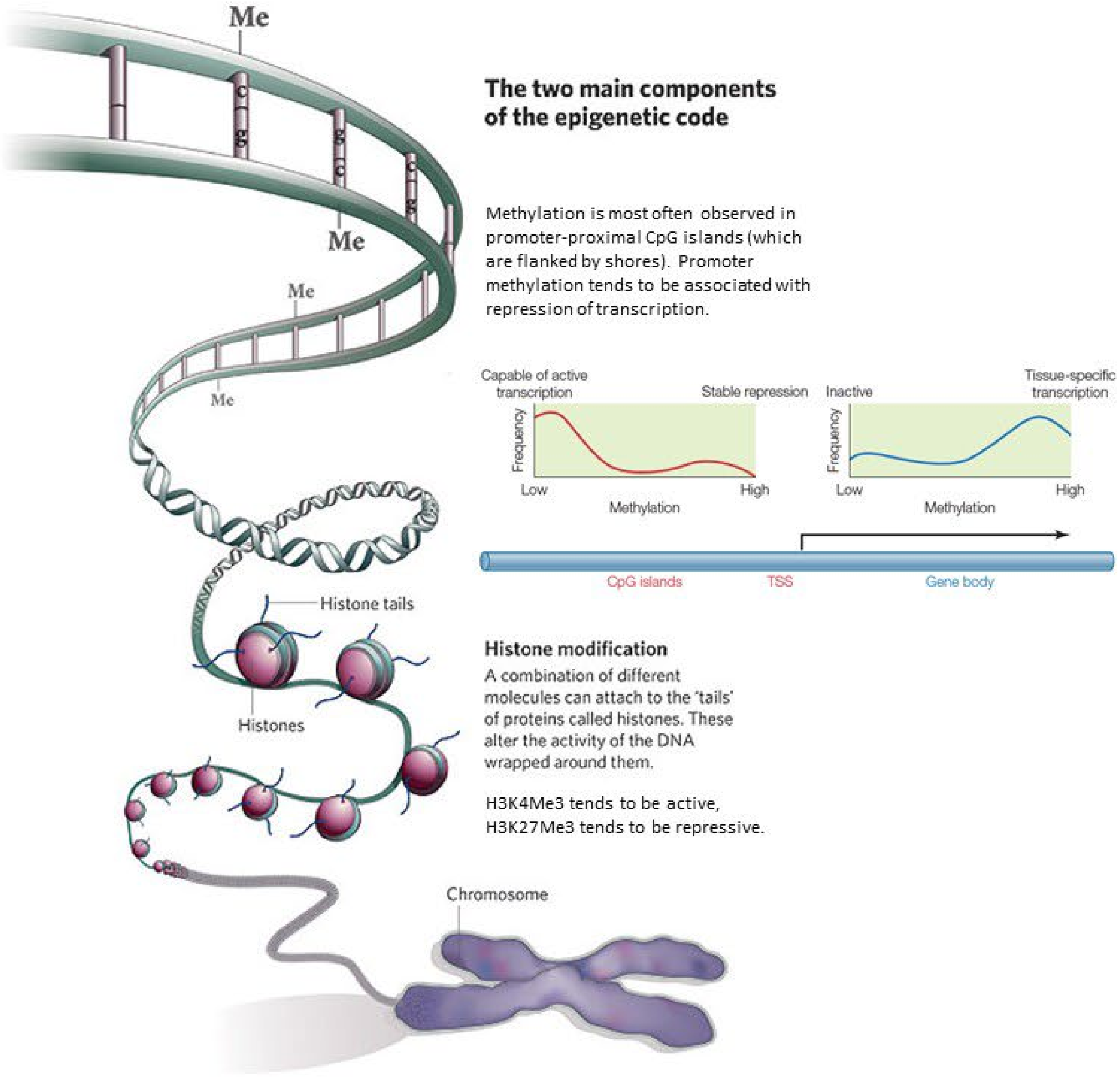
ArchR workflow


ArchR
Analysis of Regulatory Chromatin in R
scATAC-seq BAM or Fragment Files
Create ArchR Arrow Files
Create ArchR Project
Filter Cells by Quality and Remove Predicted Doublets
Iterative LSI & Clustering
Advanced Data Analysis
Robust Peak Calling & Merging
Motif Search
Marker Feature Identification
chromVAR
TF Footprinting
Projection of Bulk ATAC-seq Data
Genome Track Visualization
Native Interactive Genome Browser
Gene Activity Scores
Peak Coaccessibility
scATAC-seq--scRNA-seq Integration
Peak-to-Gene Linkage
Cellular Trajectory Analysis
Data Visualization
Quality Control
TSS Enrichment Score
log10(Unique Fragments)
Clustering + Embeddings
Marker Identifcation
TF Footprinting
Interactive Genome Browser
ArchR Browser v1 : nCells = 215,031
Gene Activity Scores
ATAC
Gene
Weight
Linking Peaks to Genes
-250kb
250kb
Differentiation Trajectories
UMAP Dim. 1
UMAP Dim. 2

Co-accessibility implies summation to predict expression

Image from ArchR Manual: https://www.archrproject.com/index.html

Three modes of epigenetic regulation



The two main components of the epigenetic code

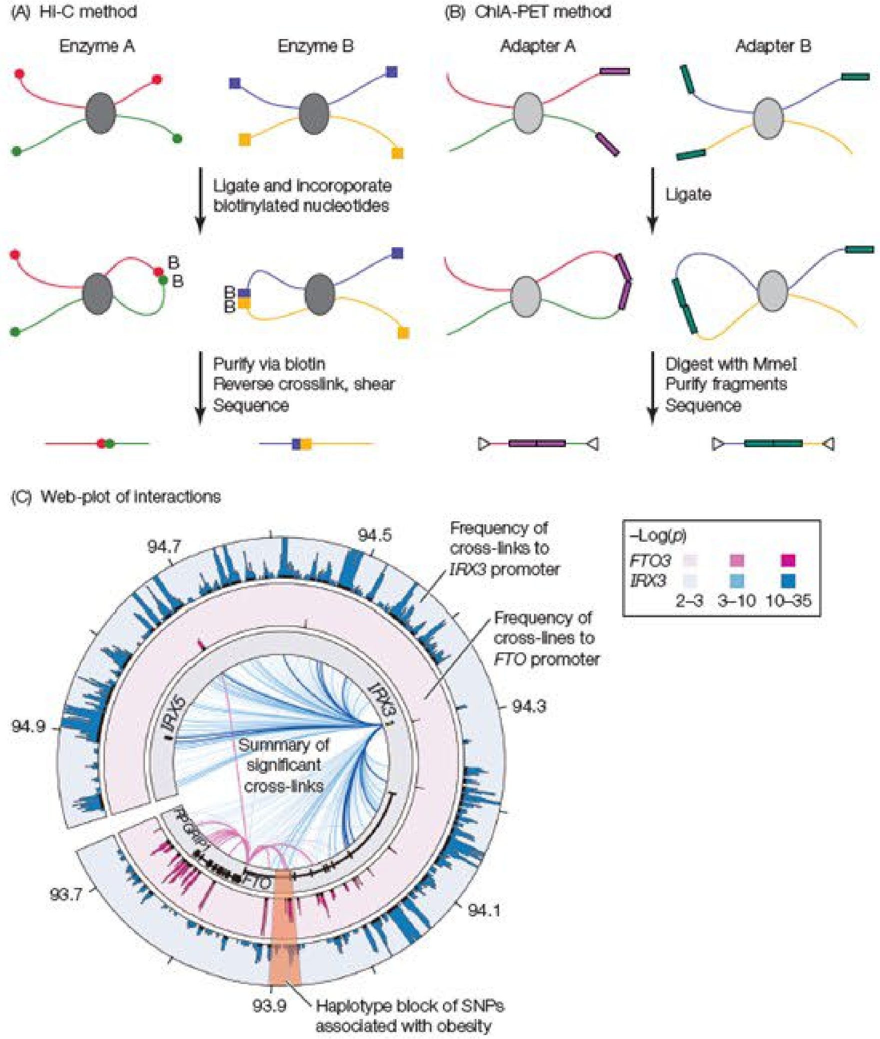
Methylation is most often observed in promoter-proximal CpG islands (which are flanked by shores). Promoter methylation tends to be associated with repression of transcription.

Histone modification

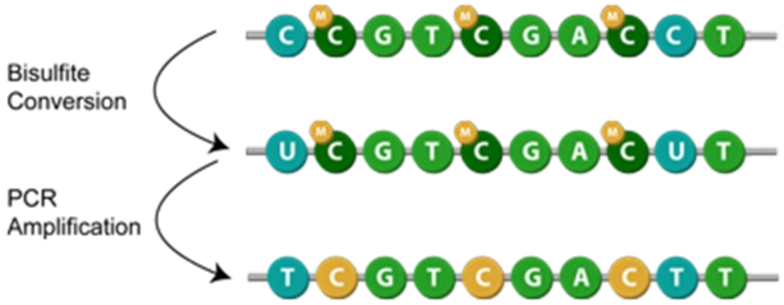
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

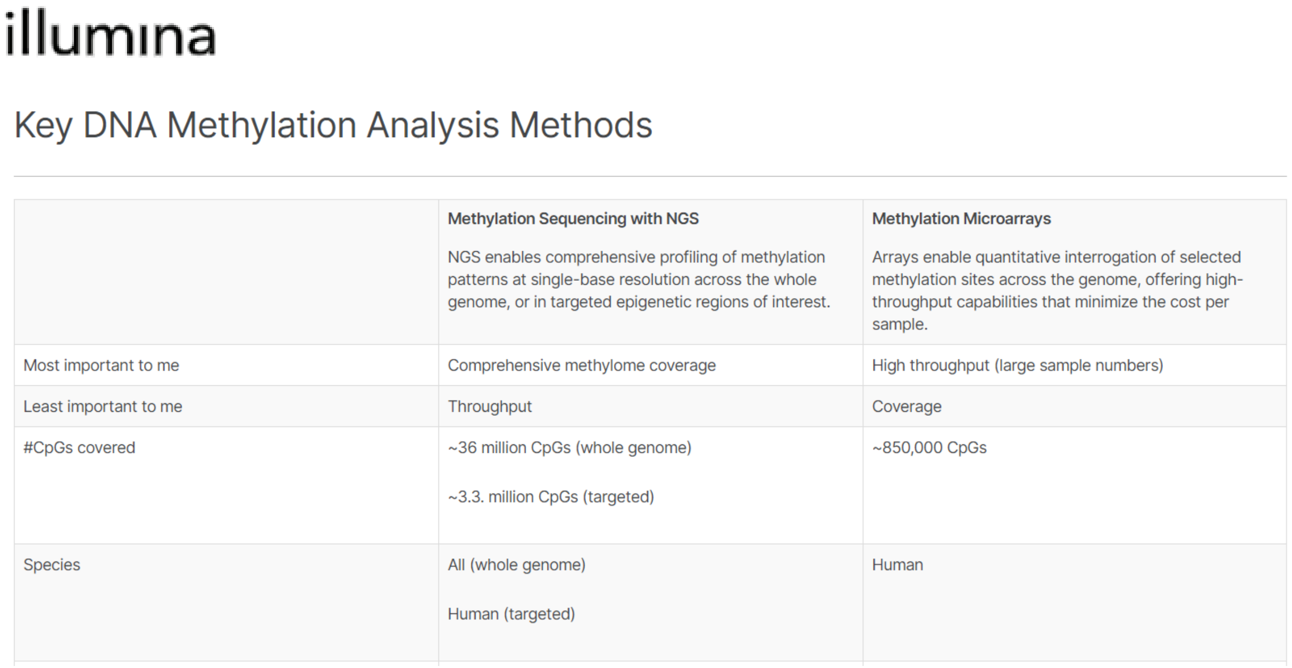
H3K4Me3 tends to be active, H3K27Me3 tends to be repressive.

Chromatin conformation



Two modes of Methylation profiling





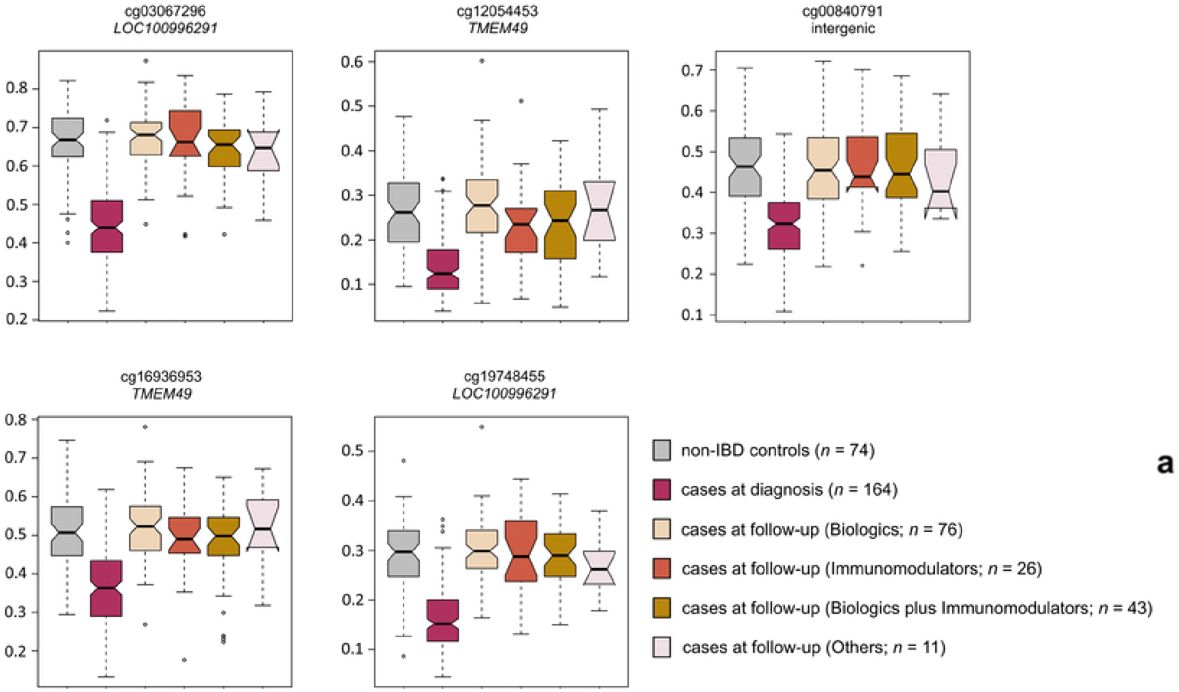
illumina

Key DNA Methylation Analysis Methods

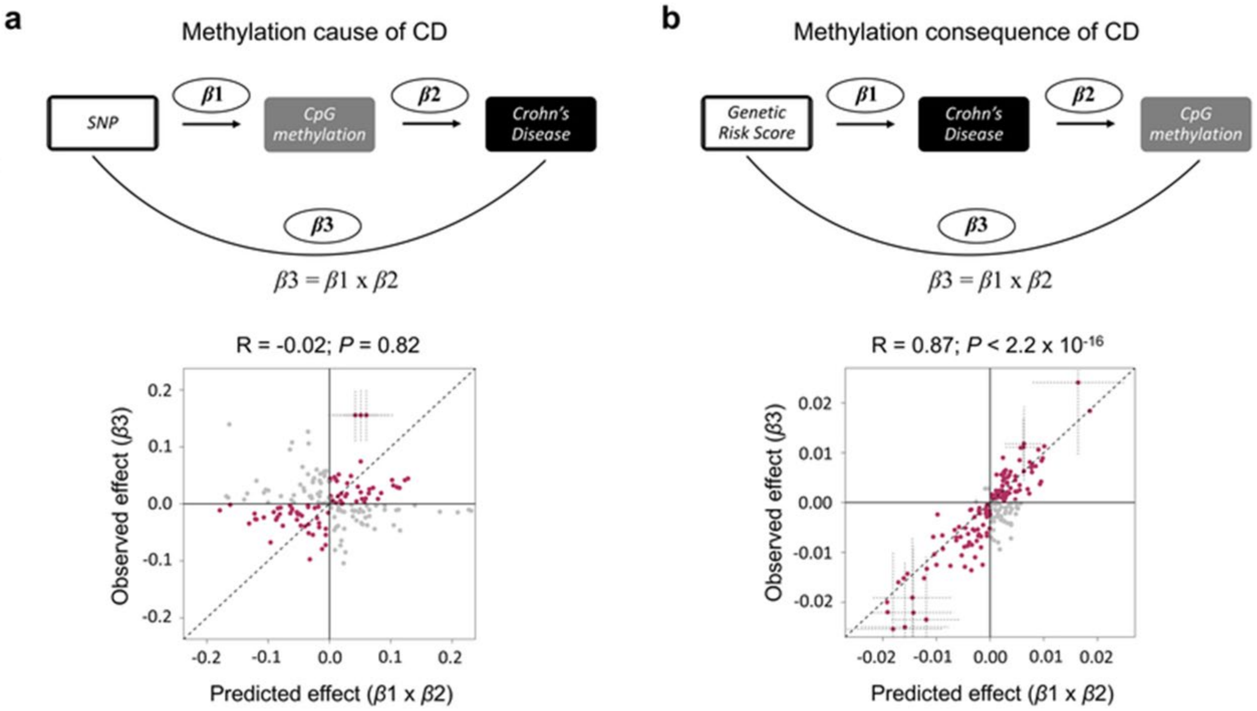
	Methylation Sequencing with NGS NGS enables comprehensive profiling of methylation patterns at single-base resolution across the whole genome, or in targeted epigenetic regions of interest.	**Methylation Microarrays** Arrays enable quantitative interrogation of selected methylation sites across the genome, offering high-throughput capabilities that minimize the cost per sample.
Most important to me	Comprehensive methylome coverage	High throughput (large sample numbers)
Least important to me	Throughput	Coverage
#CpGs covered	~36 million CpGs (whole genome) ~3.3. million CpGs (targeted)	~850,000 CpGs
Species	All (whole genome) Human (targeted)	Human

Software: <span>CpGassoc</span> by Karen Conneely et al *Bioinformatics* (2012) **28(9)**:1280-1281

Causality and Methylation



Methylation more likely a cause than a consequence of disease (in this context)



Blood methylation is actually a signature of inflammation at diagnosis, and reverts with time irrespective of treatment regimen

Somineni HK, et al (2022) *Gastroenterology* **156**: 2254-2265.e3

Some (concise) definitions

GWAS: Genome-wide association study – search for SNPs significantly associated with a trait (eSNPs)

TWAS: Transcriptome-wide association study – search for predicted transcripts significantly associated with a trait

EpiWAS: Epigenome-wide association study – search for epigenetic marks significantly associated with a trait
(EWAS also used, but earlier used to refer to Environment-wide association study)

eQTL: a SNP which influences the abundance of a transcript. Cis-eQTL act locally (~ within ± 500kb)

eGene: a gene whose transcript abundance is regulated by a locally-acting SNP

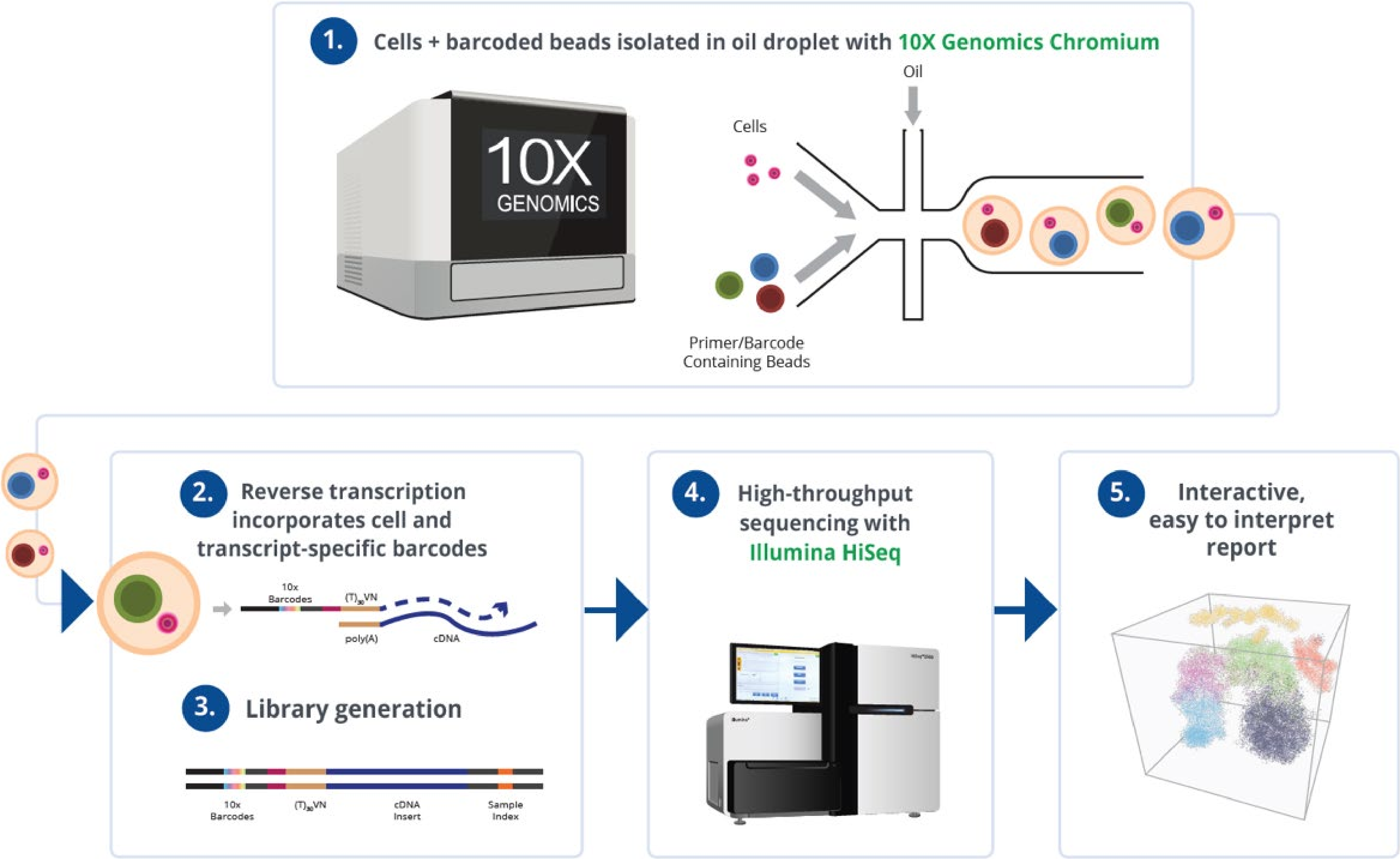
meQTL: a genotype which is associated with the degree of methylation at a CpG site

Methyl ß: typical measure of the degree of methylation, ranging from 0 to 1 (none to complete)

hQTL: a genotype that is associated with the intensity of a histone mark (may be acetylation or methylation)

ccQTL: a genotype that influences the level of chromatin conformation / cross-linking

Single Cell RNA-seq: Easy as 1,2, 3, … 5



1. Cells + barcoded beads isolated in oil droplet with 10X Genomics Chromium

2. Reverse transcription incorporates cell and transcript-specific barcodes

3. Library generation

4. High-throughput sequencing with Illumina HiSeq

5. Interactive, easy to interpret report

Types of Single Cell RNA-seq

1. SmartSeq2

- Essentially full-length RNA-seq applied to libraries generated from single cells
- Low throughput and relatively expensive, but comprehensive
- Commercial option is Becton-Dickinson Rhapsody^{TM}

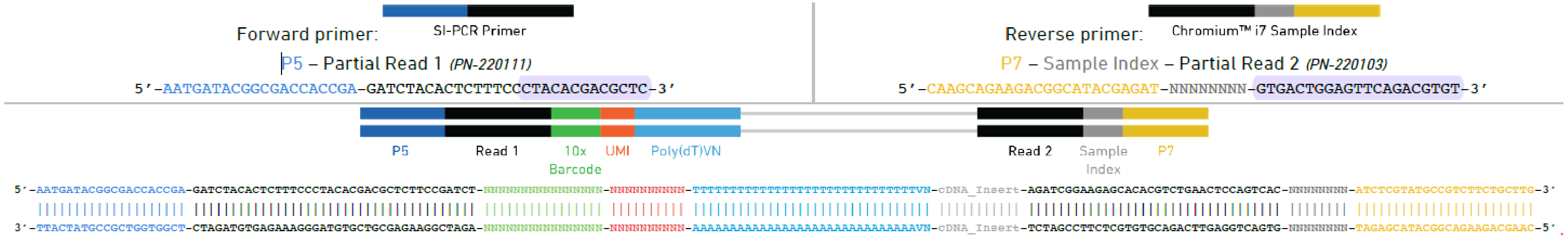
2. Droplet Sequencing

- Each cell is encapsulated in a droplet with enzymes and reagents for sequencing
- High throughput, dollars per cell, but only detects tags for each transcript
- Commercial options are 10X Genomics Chromium^{TM}, BioRad, and OneCellBio

3. sci-Seq

- Single cell Combinatorial Indexing in microtiter plates
- High throughput, very inexpensive, amenable to dual profiling with other assays
- Implemented in academic labs

Chromium Droplet Barcodes



Forward primer:

P5 – Partial Read 1 *(PN-220111)*

```
5'-AATGATACGGCGACCACCGA-GATCTACACTCTTTCCCTACACGACGCTC-3'
```

Reverse primer:

P7 – Sample Index – Partial Read 2 *(PN-220103)*

```
5'-CAAGCAGAAGACGGCATACGAGAT-NNNNNNNN-GTGACTGGAGTTCAGACGTGT-3'
```

```
5'-AATGATACGGCGACCACCGA-GATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNNNNNNNNNNNN-NNNNNNNNNN-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-cDNA_Insert-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNNNN-ATCTCGTATGCCGTCTTCTGCTTG-3'
3'-TTACTATGCCGCTGGTGGCT-CTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA-NNNNNNNNNNNNNNNN-NNNNNNNNNN-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAVN-cDNA_Insert-TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG-NNNNNNNN-TAGAGCATACGGCAGAAGACGAAC-5'
```

Sample Index is a barcode specific for the sample (individual, tissue, treatment, etc)

10X Barcode is for the cell, it tags all molecules derived from the same cell

UMI is a Unique Molecular Identifier for each actual mRNA molecule, basically controls amplification biases

Since library costs start at $1300, multiple samples can be combined in one reaction by adding a 4th type of barcode such as a BioLegend cell surface antibody, or using the person’s genotypes

In a typical cell: 50,000 reads may correspond to 10,000 UMI and 3,000 expressed genes
most transcripts may have from 1 to 5 UMI each represented by multiple reads

Read Depth, Cell number, and Expense

Sequencing is done on either a NextSeq or NovaSeq Illumina sequencer. Typical current options might be:

NextSeq lane = 400 Million 28x96 bp = 50,000 reads per cell for 8,000 cells, at a cost of ~$2,500 {30 c/cell}
S1 flow cell = 3 Billion 28x96 bp reads = 100,000 reads per cell for 30,000 cells, at a cost of ~$6,000 {20 c/cell}
S4 flow cell = 18 Billion PE reads = 50,000 reads per cell for 360,000 cells, at a cost of ~ $25,000 { 7 c/cell}

What read depth is required?

It depends on the cell-type: 50K is sufficient for many, but some require >100K

It depends on the application: if low abundance transcripts are key, you need more
if differential expression is key, you may need more
if defining novel cell types and states is key, you may need more

Five concerns about rigor and reproducibility in sc genomics

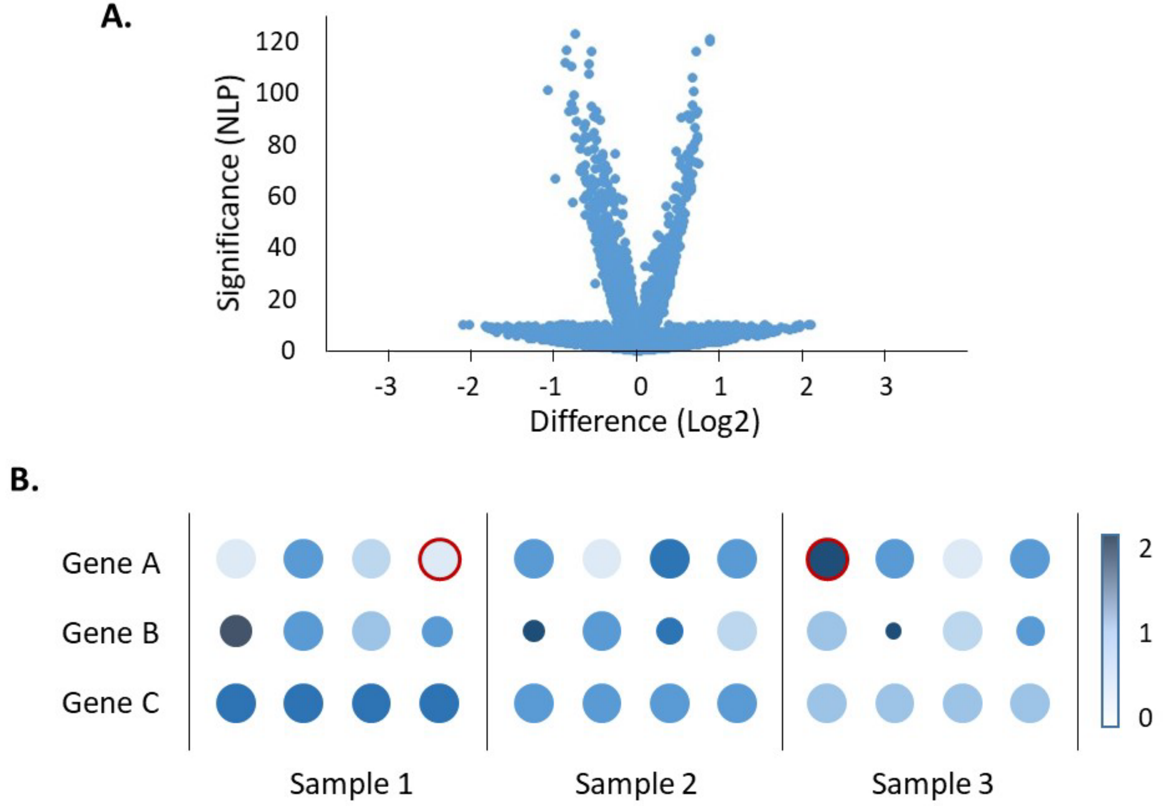
Repeatability: Few results are independently validated in new datasets

Clustering: Clusters of cell types and states are not routinely presented with support intervals

Significance testing: Individual cells are too often treated as biological not technical replicates: pseudobulk solution?

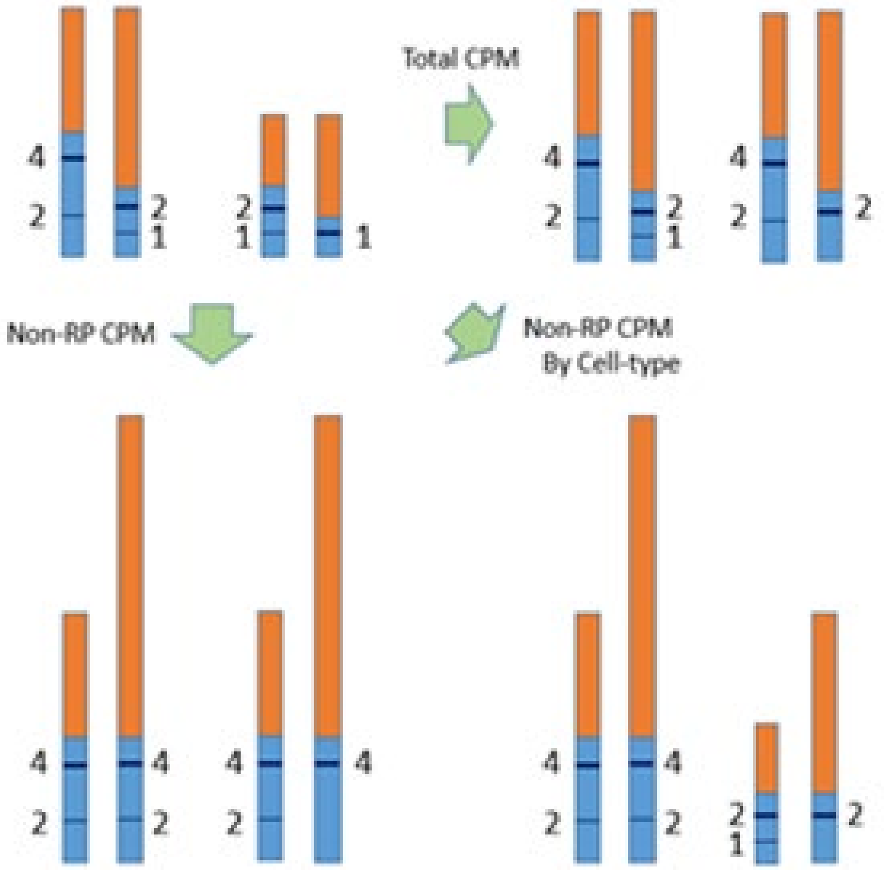
Covariate adjustment: Samples are random effects, which are rarely adjusted for

Normalization: Supervised normalization approaches are yet to be introduced



Gibson, G. (2022) *PLoS Genetics* **18**: 1010210

The normalization problem



Suppose we represent scRNA abundances for two cells of each of two cell types by these bars, with ribosomal proteins in orange and common transcripts in blue. Now focus on two genes represented by the horizontal bars, with counts shown next to them.

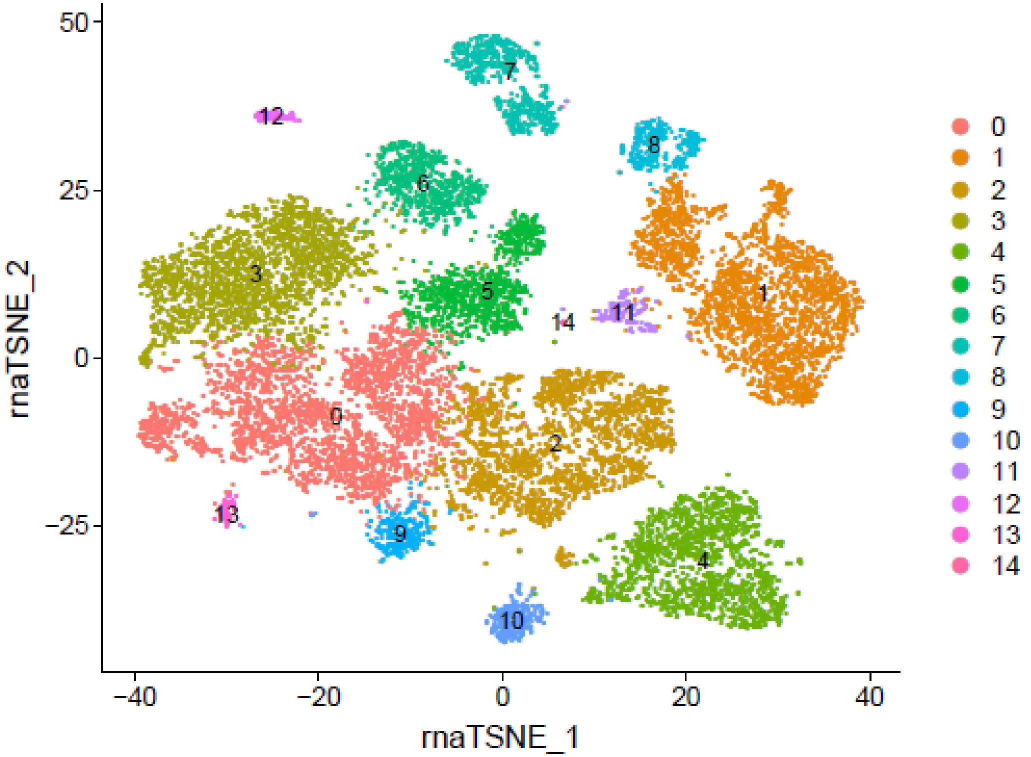
Normalizing by total cpm leads to the conclusion that there is little difference between the left and right cell types, except for the drop-out transcript, but there is high within-sample variability.

Normalizing by non-ribosomal CPM alone leads to the conclusion that all four cells are very similar.

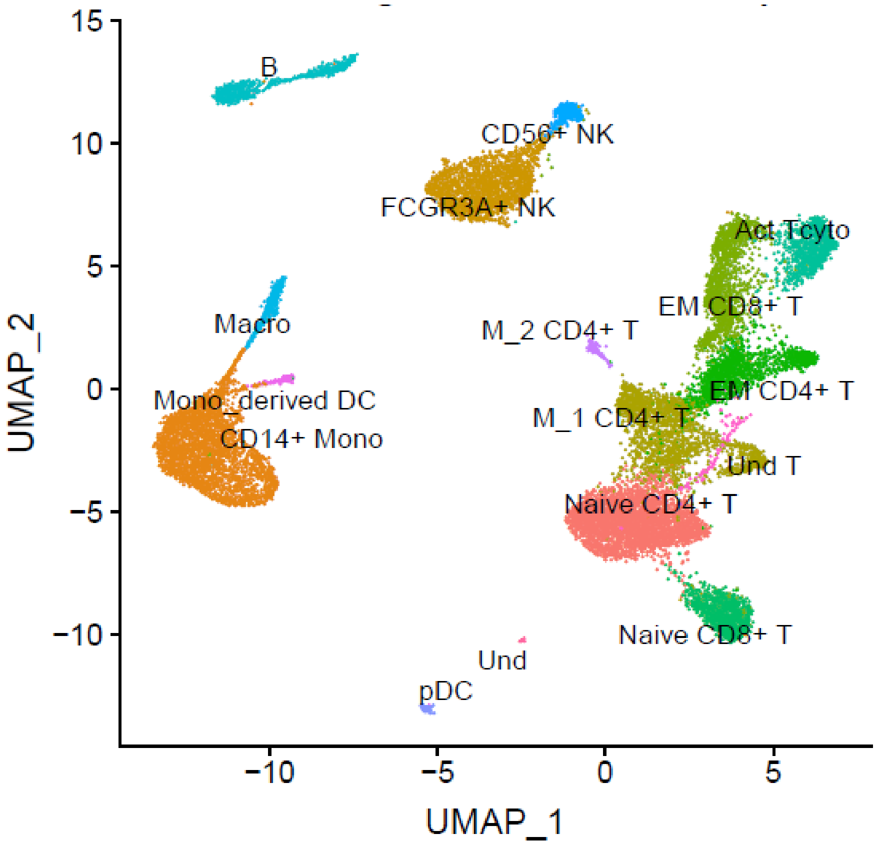
Normalizing by cell-type and non-ribosomal CPM recovers the cell-type difference in absolute abundance.

Cluster Visualization

tSNE



UMAP



https://distill.pub/2016/misread-tsne/

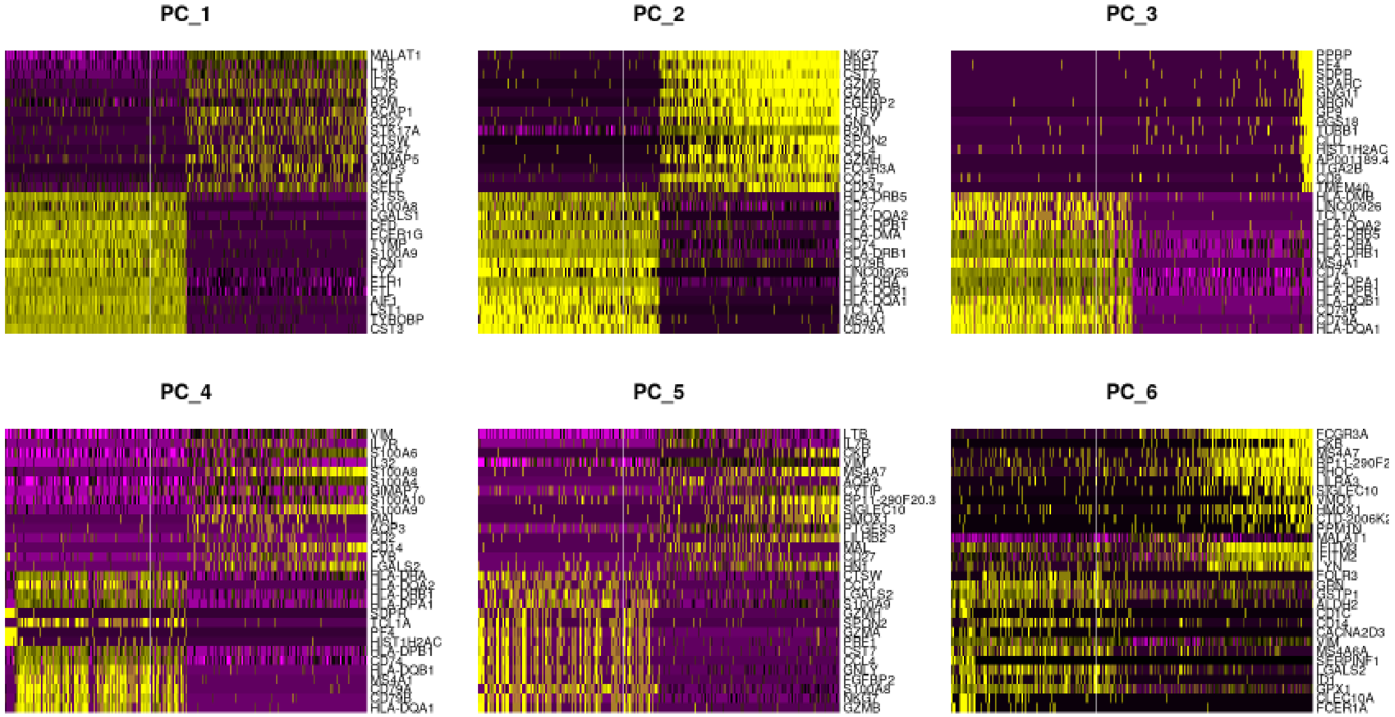
Deciding how many PC to include in CCA

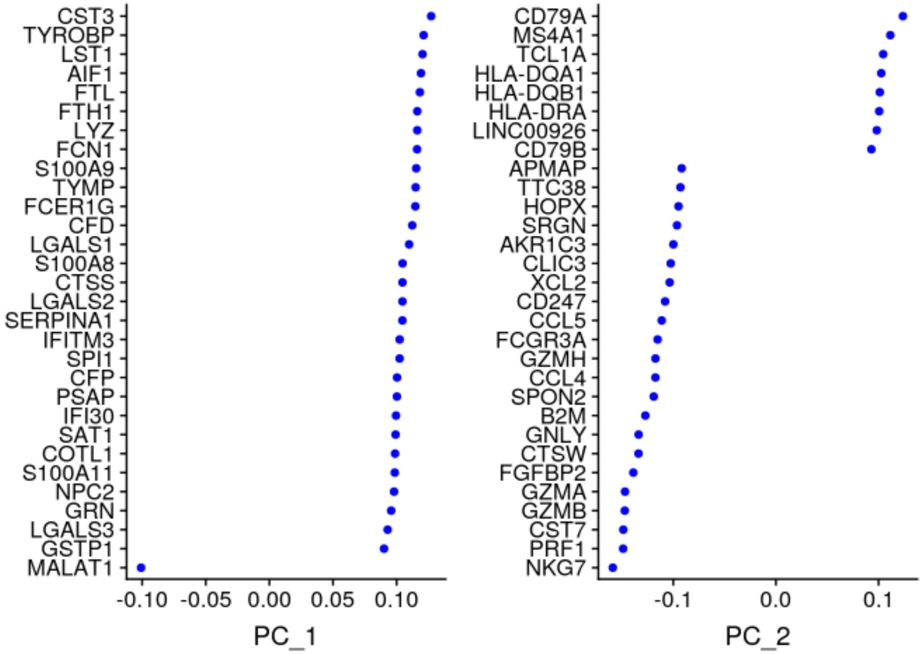
https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html

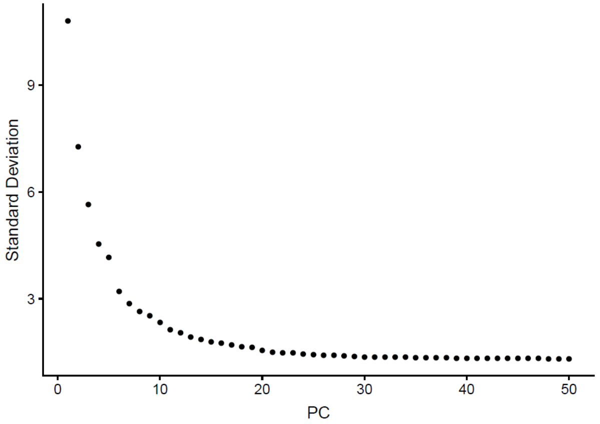
SATIJA LAB

Home News People Research Publications Seurat Join/Contact Single Cell Genomics Day

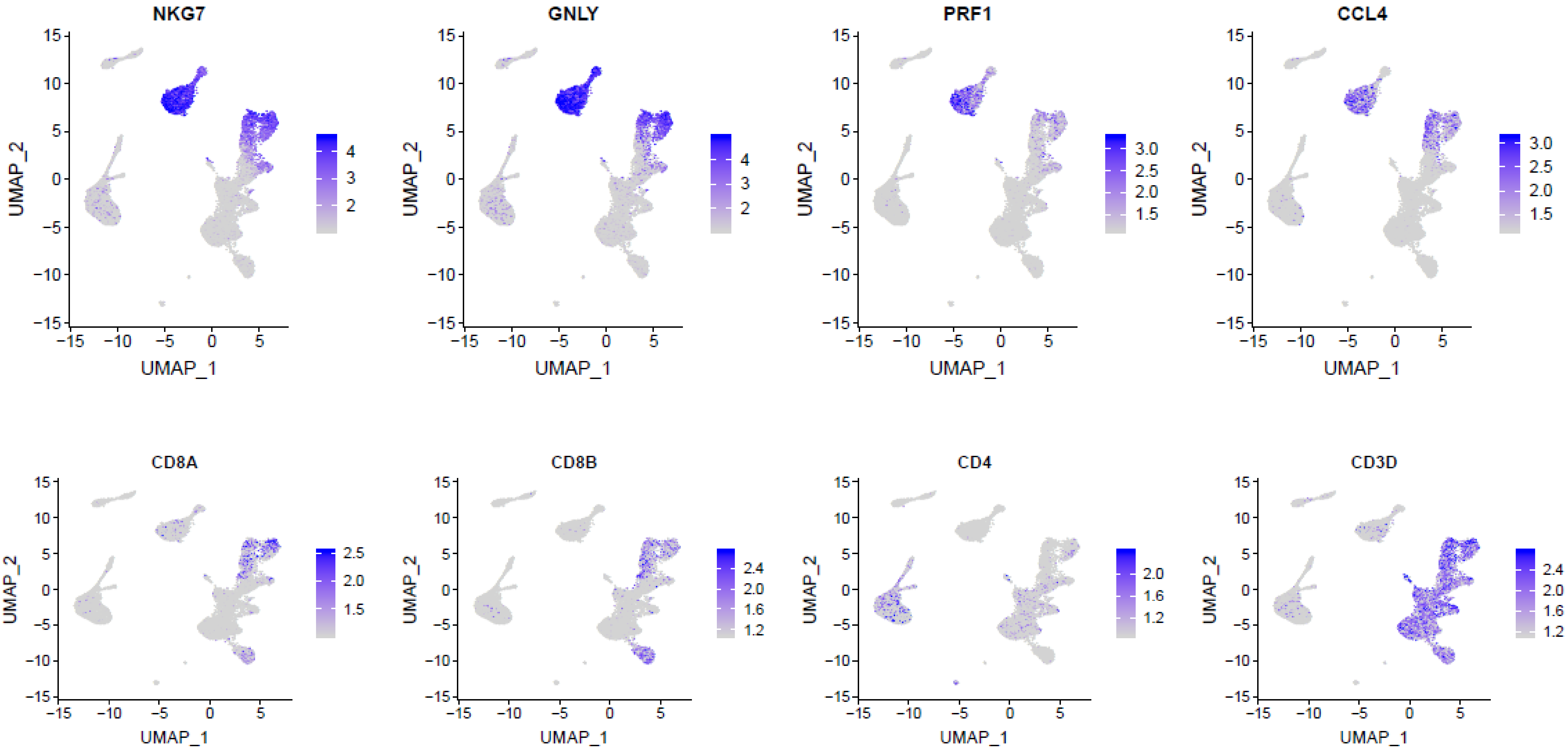
SEURAT R toolkit for single cell genomics



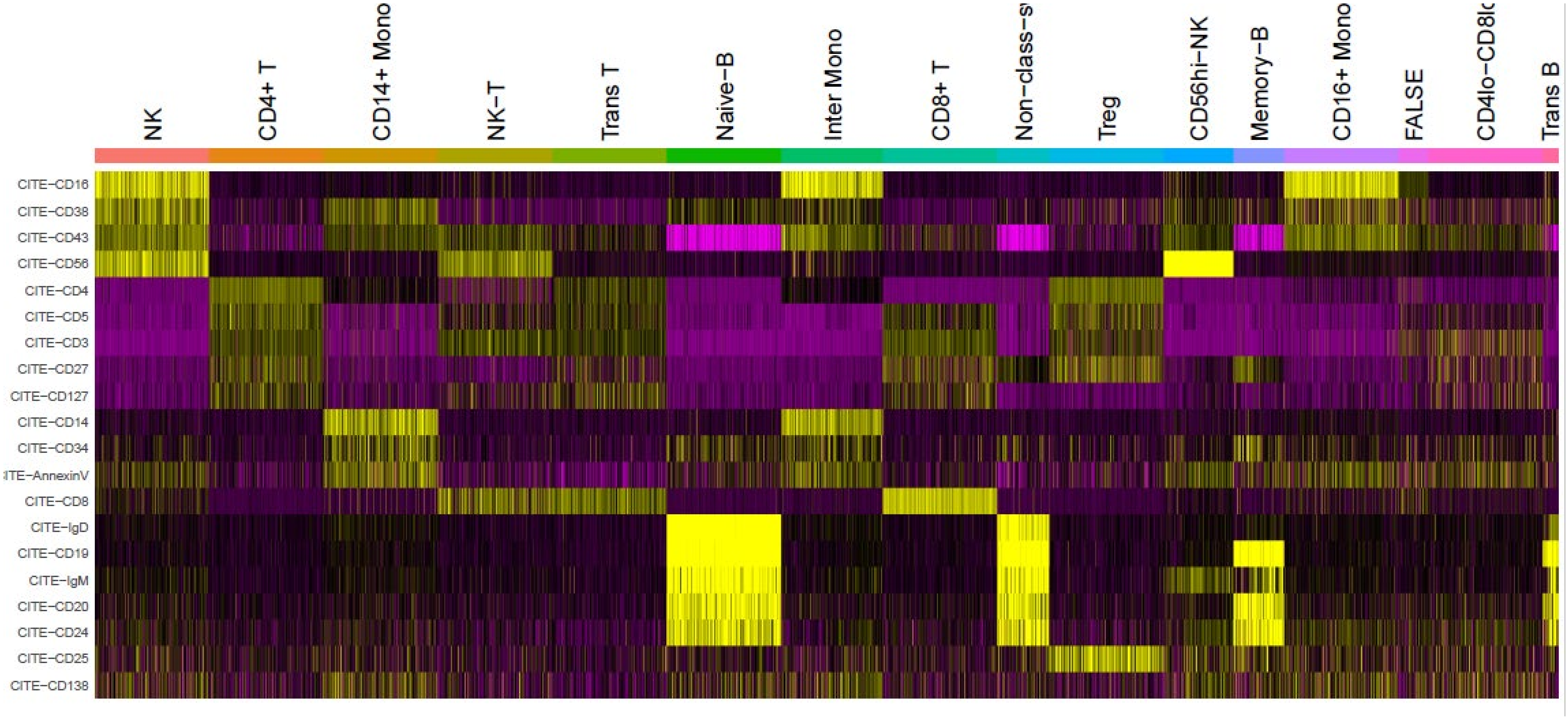




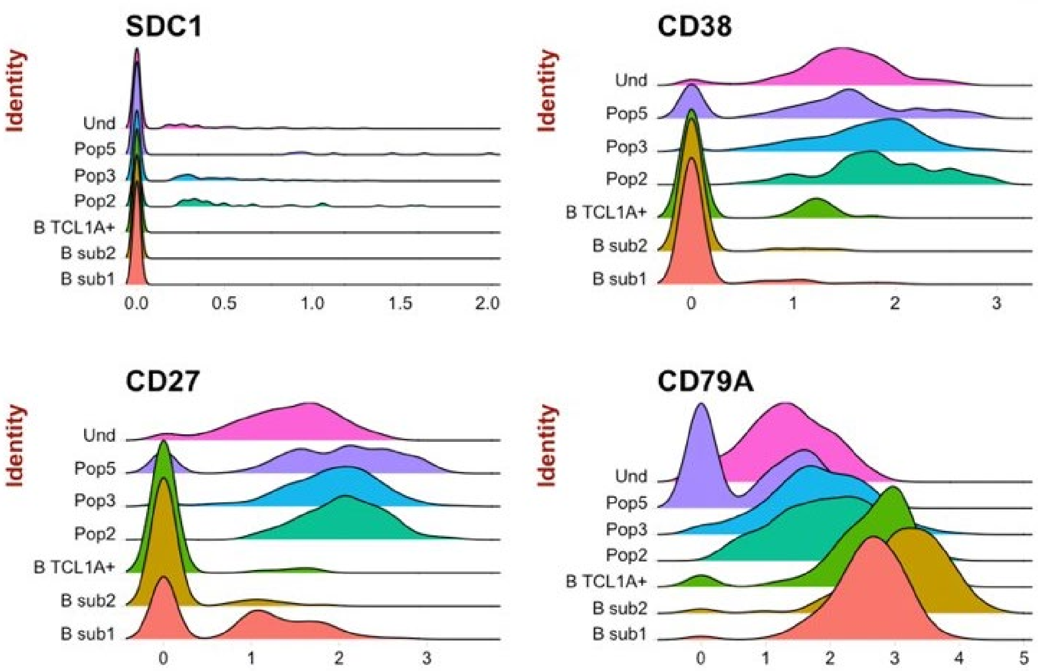
Anchoring Cell-types by Marker genes

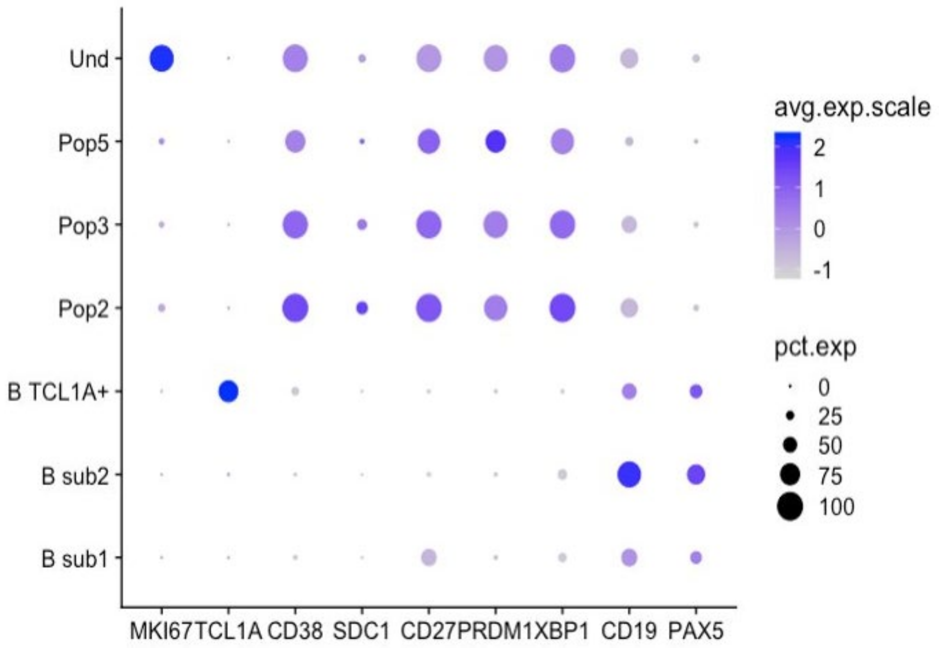


Hierarchical Clustering of Marker Genes



Differential Expression Analysis





Some extensions of Single Cell Genomics

1. Crop-Seq / Perturb-Seq
    - Microdeletion of SNPs in single cells followed by RNA-Seq
    - Requires co-transfection with Cas9 and lentivirus or plasmid expressing guide RNAs
    - Generally useful to monitor alterations of gene networks

2. CITE-Seq
    - Addition of oligonucleotide-conjugated Antibodies that bind cell surface receptors
    - Receive the same cell barcodes as the cell contents, but sequenced separately
    - Supports gating to homologize flow cytometry with scRNAseq

3. ATAC-Seq
    - Assay for Transposon-Accessible Chromatin (basically, identifying enhancers)
    - 10X Genomics now provides kits; reports of joint scRNA and scATAC appearing
    - https://www.10xgenomics.com/solutions/single-cell-atac/

4. Repertoire-Seq
    - Sequencing of the TCR (T-cell receptors) or BCR (immunoglobulins) from single cells
    - Options available from 10X and OneCellBio
    - Data analysis requires specific expertise