



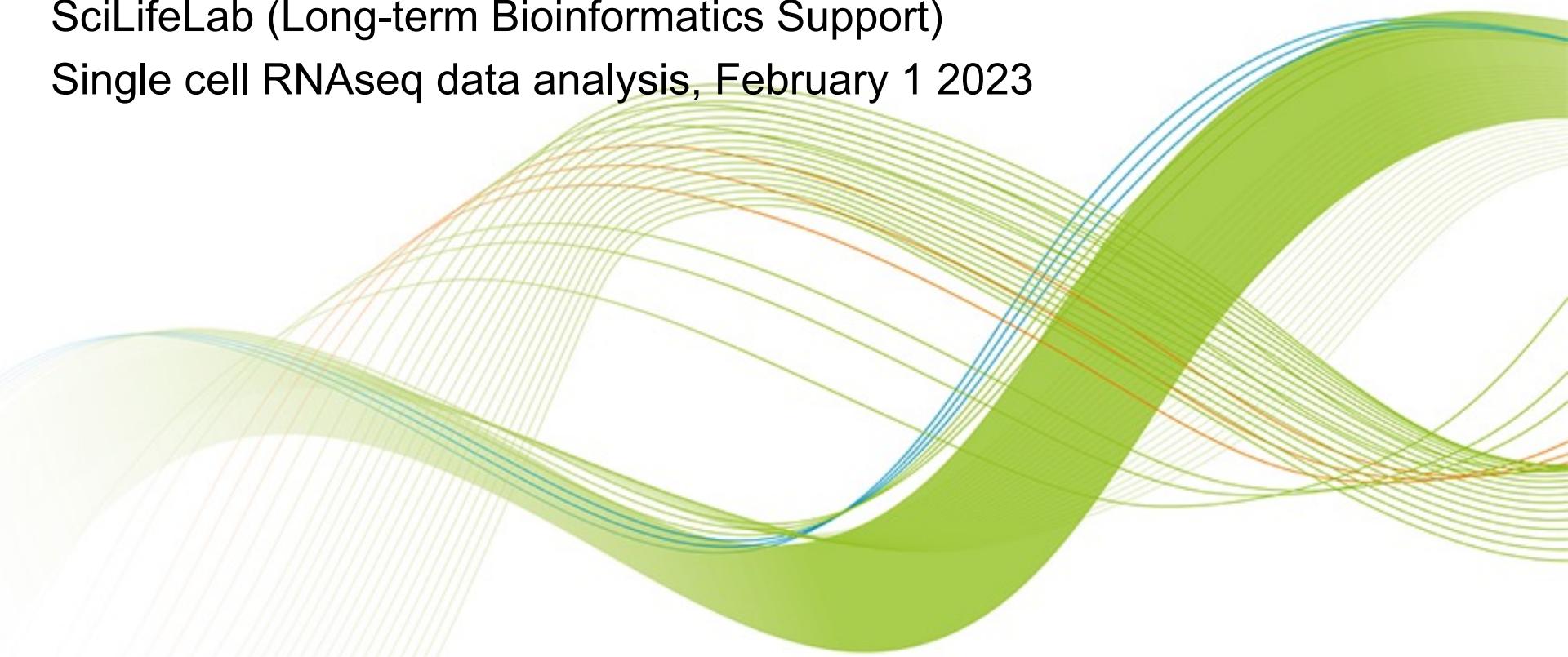
SciLifeLab

Single cell methods in epigenomics

Jakub Orzechowski Westholm

SciLifeLab (Long-term Bioinformatics Support)

Single cell RNAseq data analysis, February 1 2023

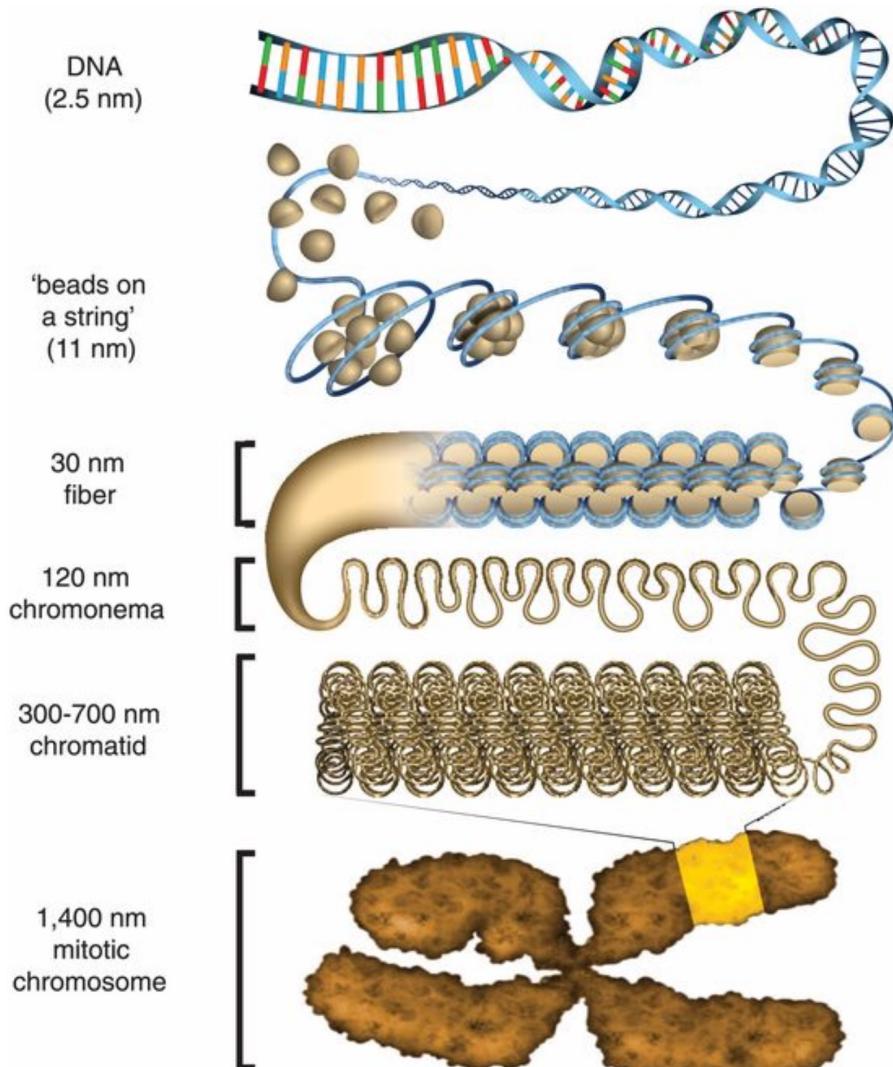


Outline

- Background on epigenomics & ATAC-seq
- Single cell ATAC-seq
- Single cell CUT & TAG
- Single cell DNA-methylation
- Multi-omics
- Broad overview
- Focus on concepts over details.

Epigenomics

- The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells
- In a field of study known as epigenomics, researchers are trying to chart the locations and understand the functions of all the chemical tags that mark the genome.

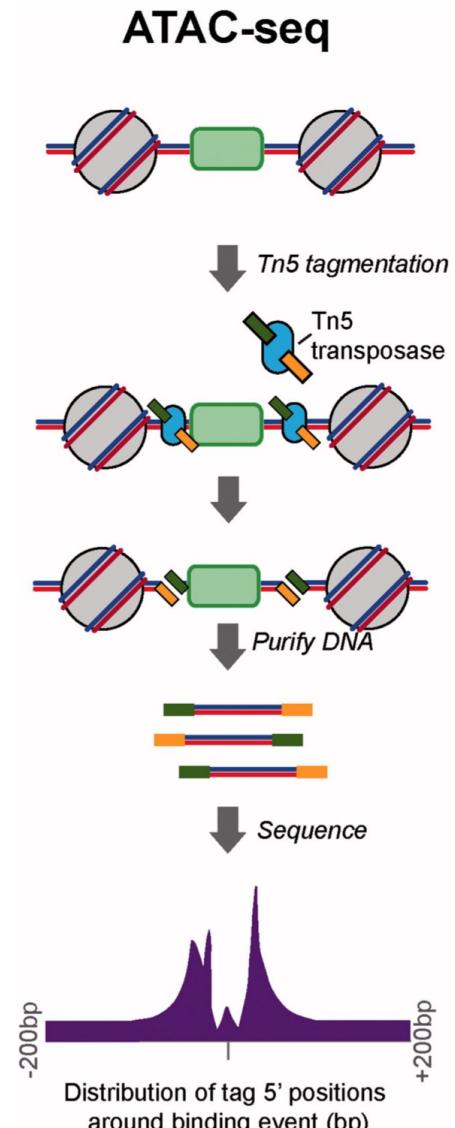


<https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet>

(Ou et al, 2017, Science)

- The epigenome is involved in many processes, e.g. development, cancer, aging,...
- Things to measure:
 - DNA methylation (e.g. bisulphite sequencing)
 - Chromatin accessibility (e.g. ATAC-seq)
 - Histone modifications (e.g. ChIP-seq, CUT&TAG)
 - Transcription factor binding
 - Other proteins
- Such methods were first developed for bulk samples, but have been adapted for single cell assays.

- Assay for Transposase-Accessible Chromatin using sequencing.
- Measures chromatin that is accessible, i.e. not bound by any big molecules or folded into compact structures.
- Basic steps:
 - Transfect cells with Tn5 transposase
 - This inserts sequencing adaptors into regions of the chromatin that it can access.
 - These adaptors can be used to purify DNA on open chromatin, and create sequencing libraries.



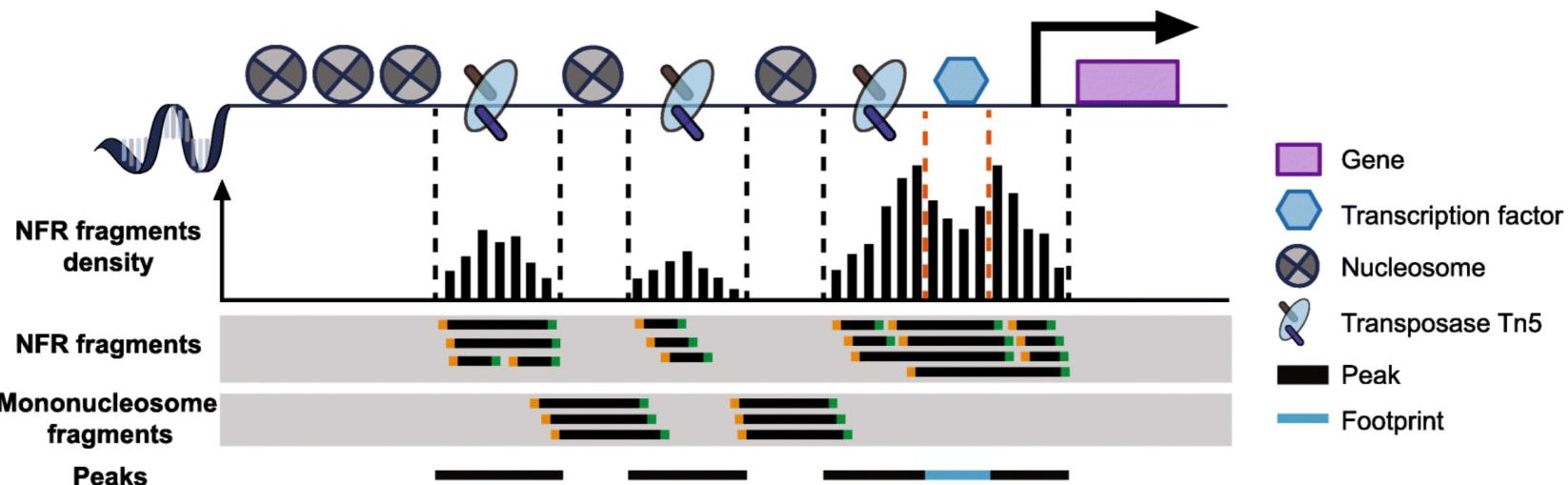
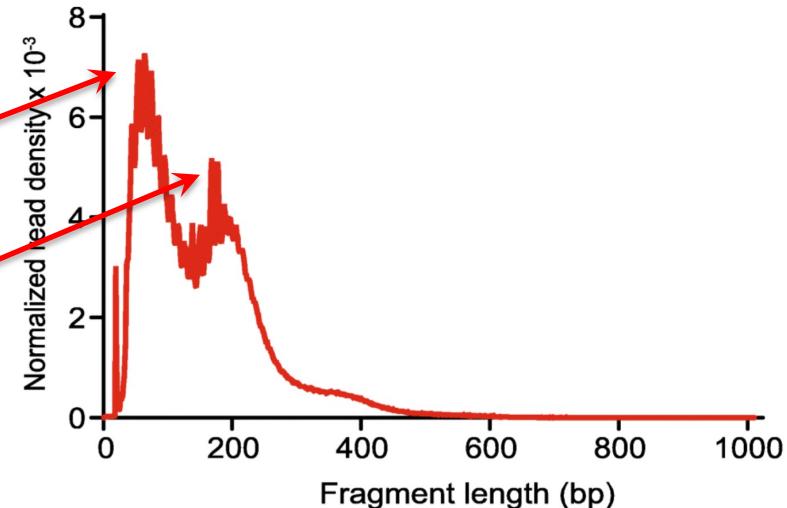
(Pugh 2015)

Why ATAC-seq?

- Accessible chromatin can tell us about:
 - Where promoters are
 - How active they are or if they are poised for activation
 - Where enhancers are
 - If chromatin is open around specific transcription factor binding sites
 - Spread of heterochromatin
 - ATAC-seq is simple to use, works with very little starting material (even single cells).
- Often a useful complement to RNA-seq

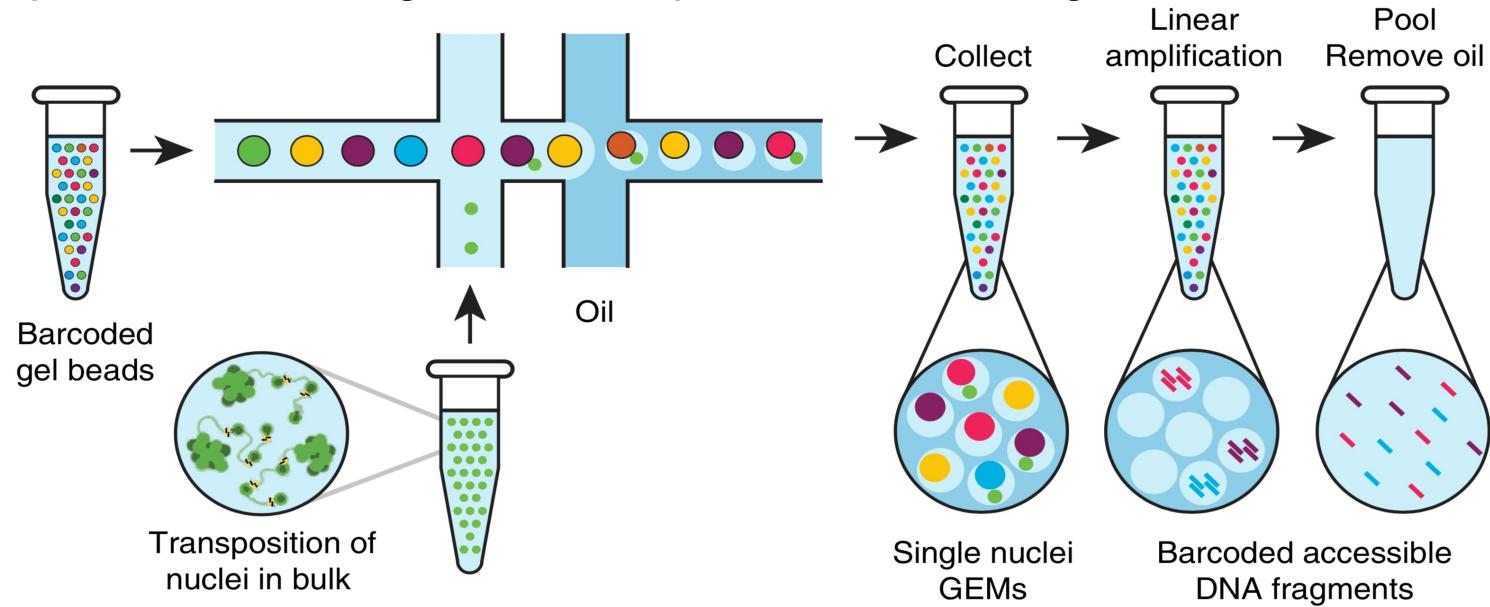
ATAC-seq data

- When mapped to the genome the sequencing reads can be used to find
 - Open chromatin (from short DNA fragments)
 - Nucleosomes (from longer DNA fragments)
- Programs, like MACS2, are used to find peaks, i.e. regions with many DNA fragments mapping.



Single cell ATAC-seq

- First paper, from Greenleaf lab: (Buenrostro et al. 2015, Nature)
- Now available as a kit from 10X Genomics
 - Each cell is attached to a bead containing a different barcode, inside an oil droplet.
 - These barcodes are attached to the DNA fragments, making it possible to assign each sequenced DNA fragment to a cell.

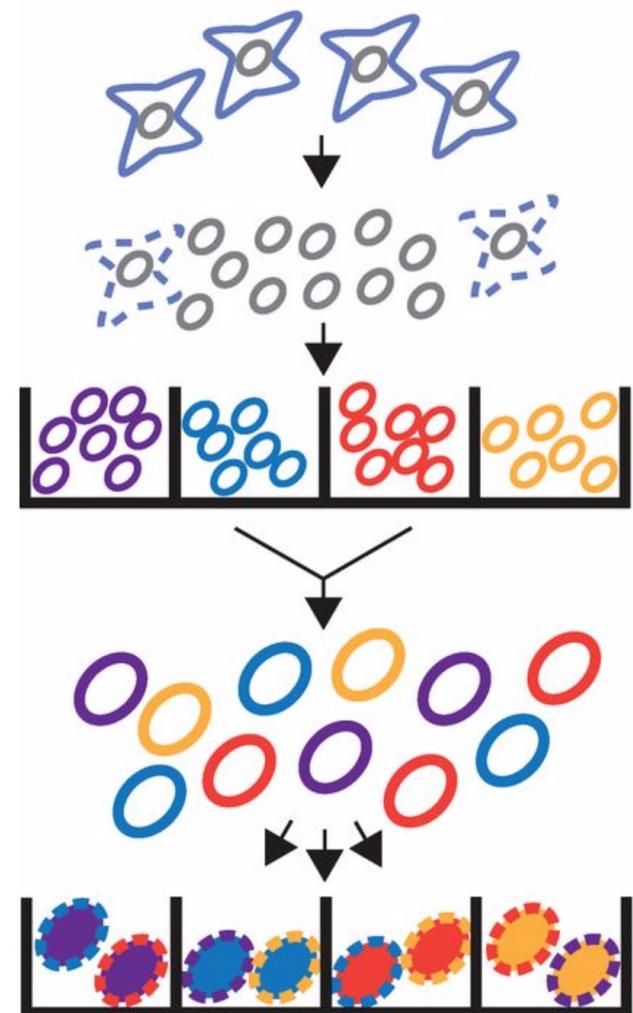


(Satpathy et al. 2019 Nature Biotechnology)



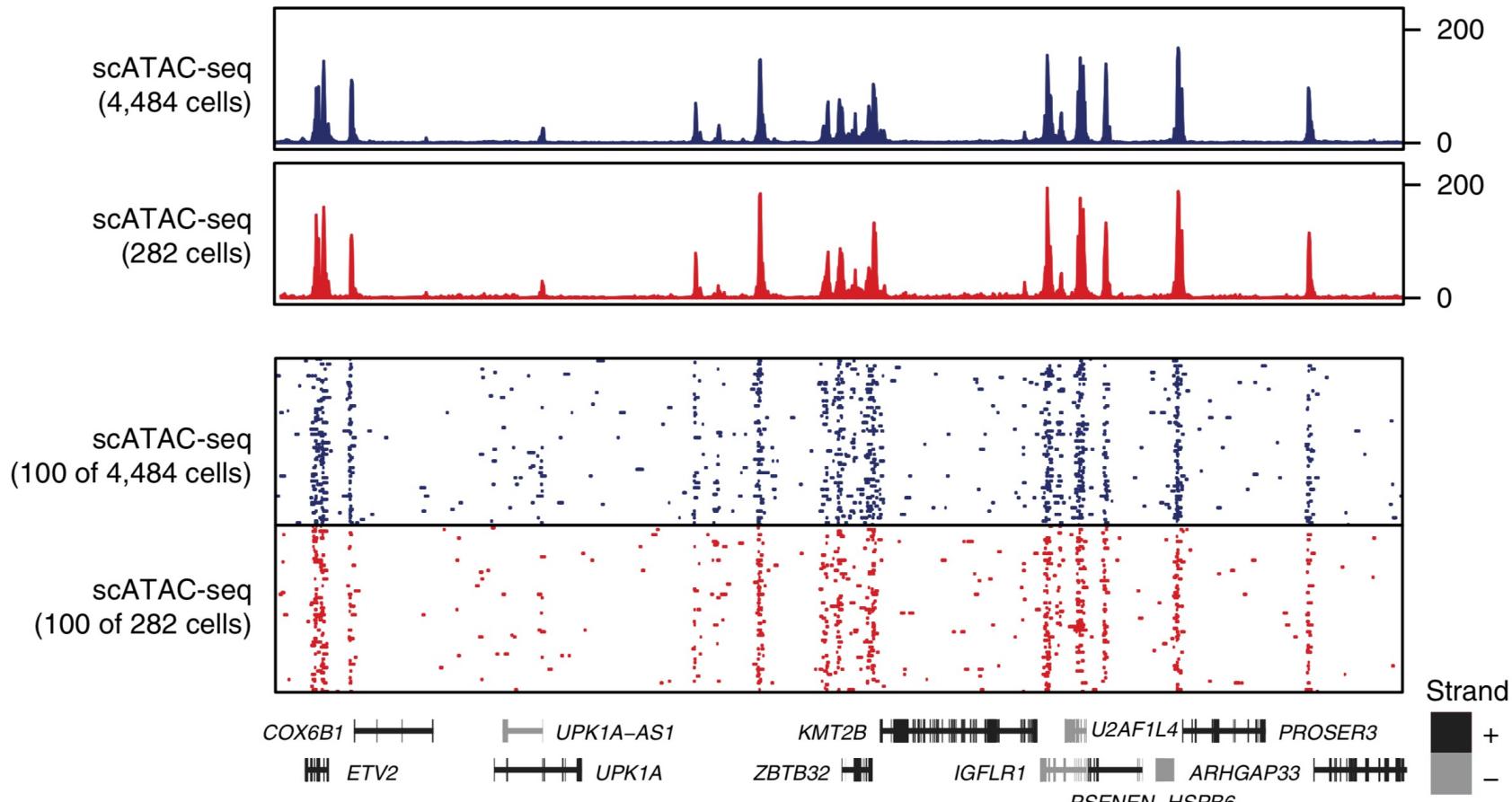
Single cell ATAC-seq II

- An alternative to the droplet based method from 10X genomics is **sci-ATAC-seq** (single-cell combinatorial-indexing with ATAC-seq).
(Cusanovich et al. 2015, Science)
 - Here, cells are split up into e.g. 96 wells, and each well has a different short barcode.
 - Cells are then pooled and re-distributed into wells again, adding another short barcode.
 - This is repeated enough times so that each cell will eventually have it's own (almost) unique combination of short barcodes.
-
- + Low cost per cell, enables high throughput
 - No commercial solution, harder to set up
 - Lower cell recovery, important when there is limited starting material

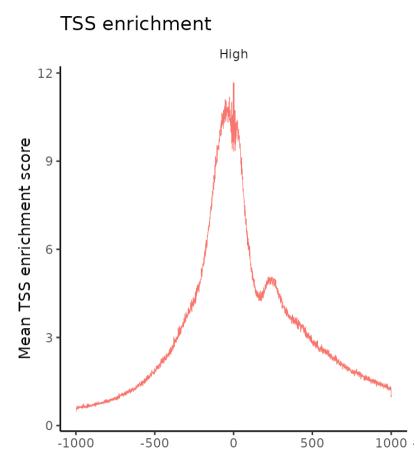


Single cell ATAC-seq data

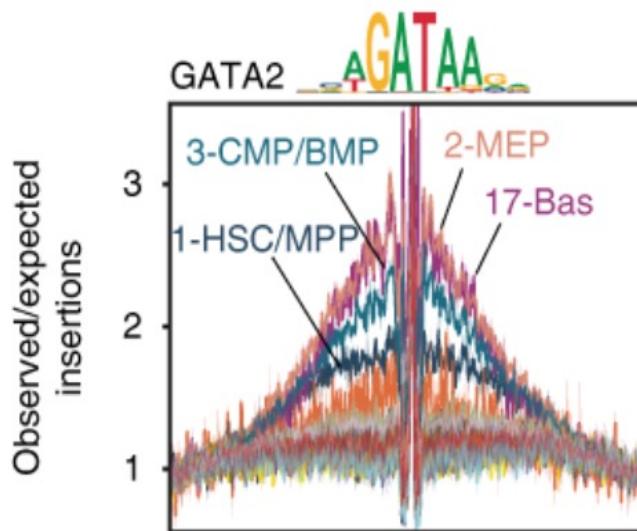
- Looking at each individual cell, scATAC-seq data are **sparse** and **noisy**.
- But combining data from lots of cells gives meaningful signals.



Combining data from lots of genomic loci can also give meaningful signals.

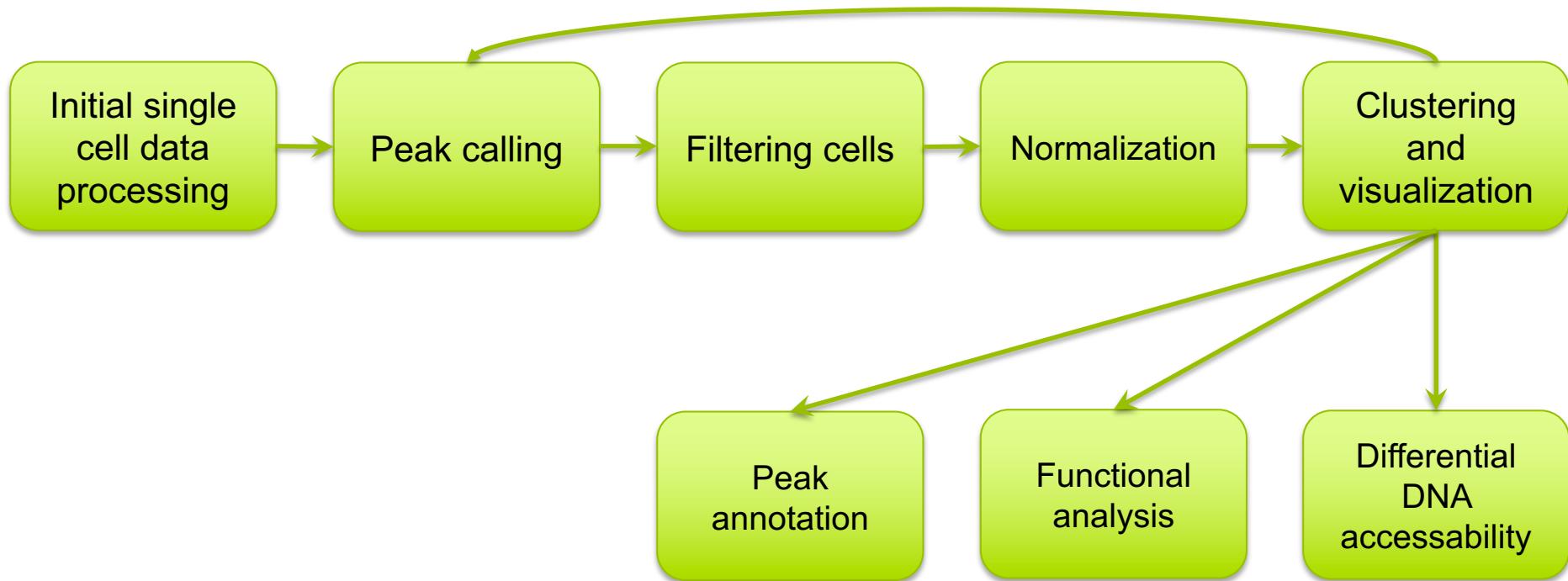


Signac tutorial



(Satpathy et al. 2019 Nature Biotechnology)

Single cell ATAC-seq data analysis



1. Initial single cell data processing



-
- De-multiplex: Using the cell specific barcodes, assign each read to a cell.
 - (Remove primer sequences.)
 - Map reads to the genome, e.g. with **BWA-MEM**.
 - Remove duplicates: If several read pairs map to exactly the same coordinates, only one is kept. Such duplicates are assumed to be PCR artifacts.
 - Filter out some bad cells already at this stage.

2. Peak calling

- Similar to peak calling for bulk data.
- Done on aggregated data from all cells. (There is not enough data in a single cell to call peaks.)
- If we have a rare cell type with e.g. 50 out of 2000 cells, peaks specific to this cell type can be missed when we use the aggregated data for peak calling.
 - We can go back and redo the peak calling later, only looking at specific groups of cells.
- We then count the reads from every cell in every peak:

	Cell 1	Cell 2	Cell 3	...	Cell M
Peak 1	0	1	1		0
Peak 2	0	0	0		0
Peak 3	0	0	0		1
...					
Peak N	1	0	0		0

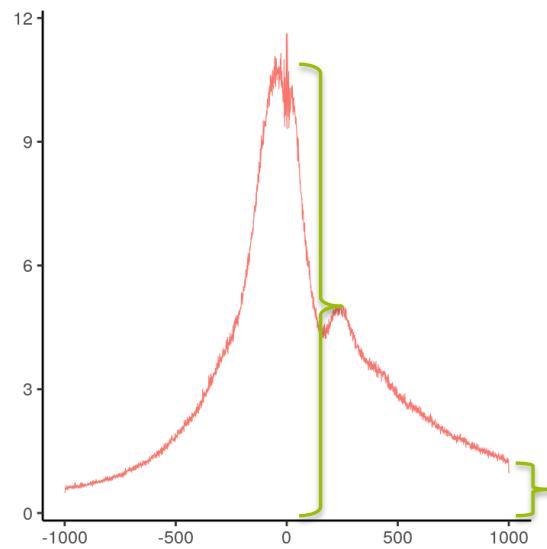
Mostly 0s 

3. Filtering cells

- There are many things that could go wrong in a single cell ATAC-seq experiment
 - No cell in a droplet
 - Several cells in a droplet
 - Dead cells
 - Few reads from a cell
 - No transfection in a cell
- Therefore we use several quality measures to identify and remove problematic cells/barcodes:
 - Number of fragments in peaks: Cells with very few reads may need to be excluded due to low sequencing depth. Cells with extremely high levels may represent doublets, nuclei clumps, or other artefacts.
 - Fraction of fragments in peaks: Cells with low values (i.e. <15-20%) often represent low-quality cells or technical artifacts that should be removed.

3. Filter cells II

- Reads in blacklist regions: The ENCODE project has provided a list of blacklist regions, i.e. regions with artefactual signal. Cells with many reads mapping to these blacklist regions (compared to reads mapping to peaks) often represent technical artifacts and should be removed.
- Transcriptional start site (TSS) enrichment score. TSS are associated with open chromatin, so a low level of chromatin enrichment would suggest poor ATAC-seq experiments.

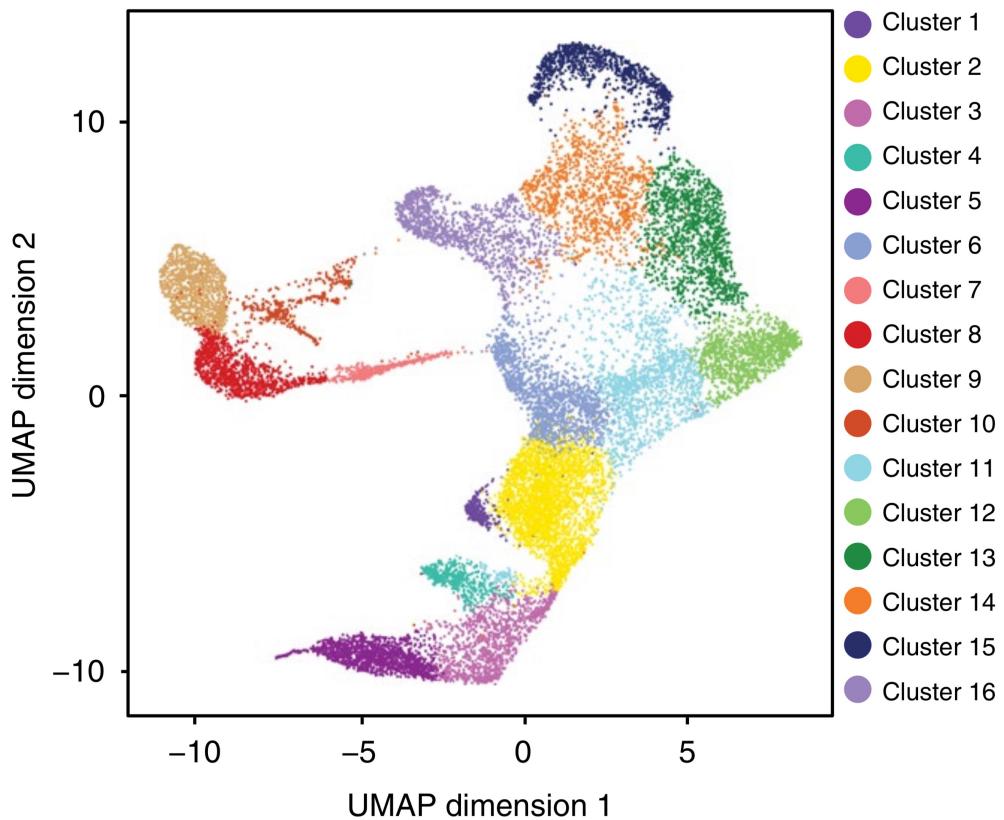


4. Normalization

- Account for different sequencing depth in different cells
- Create a simplified representation of the data, using dimension reduction (singular value decomposition). This is similar to principal component analysis (PCA).
 - The idea behind this is to reduce noise, and to select informative features to improve clustering of cells and visualization
 - Typically, the first component correlates with sequencing depth, so by removing it we get rid of artefactual signal.
 - Reducing dimensionality is often good in itself.
 - Results are often better when we select only some features (peaks)
 - Those with highest signal
 - Those with highest variability

5. Cluster and visualize cells

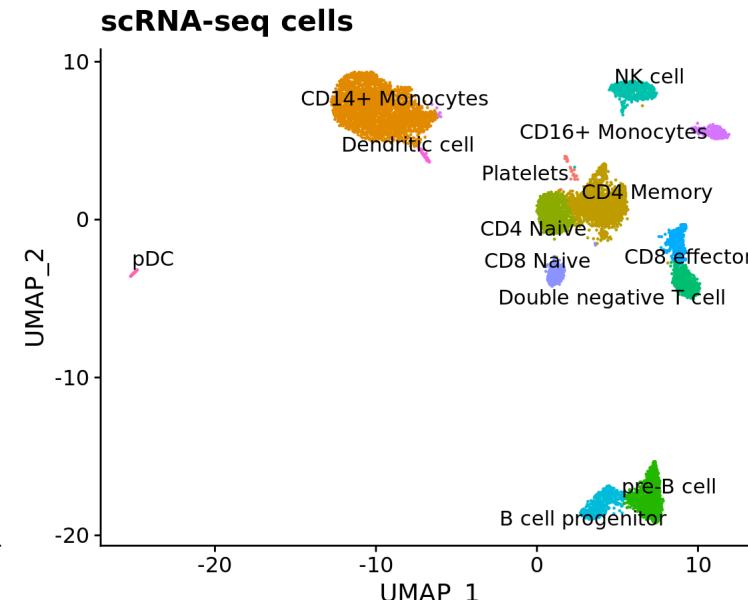
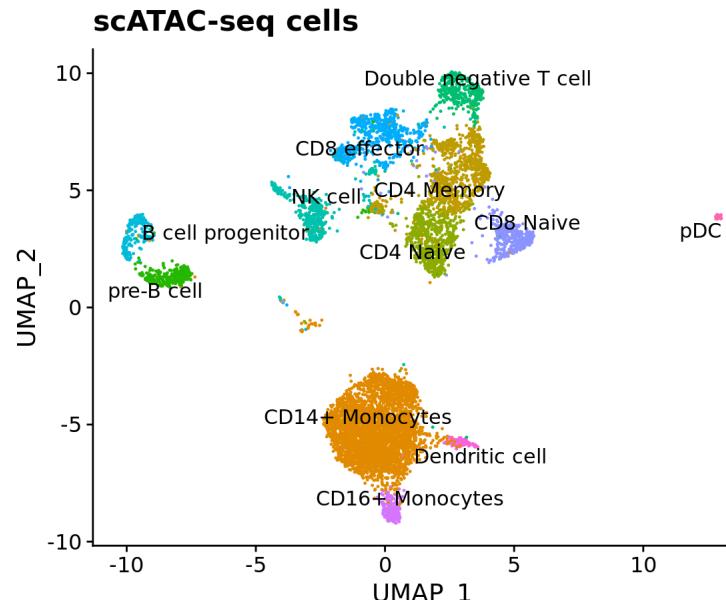
- Similar to single cell RNA seq:
 - Project many dimensions down to 2.
 - UMAP algorithm
- Clustering, to identify groups of similar cells (representing different cell types or cell states).



(Satpathy et al. 2019 Nature Biotechnology)

5. Cluster and visualize cells II

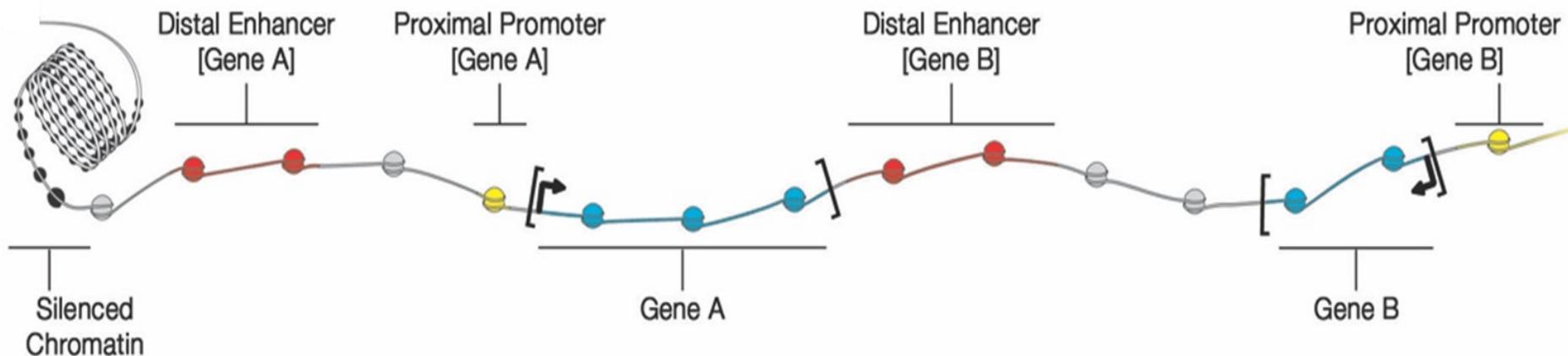
- It's often not clear which cell types etc. these clusters represent.
 - In single cell RNA-seq we can look at marker genes, unique to a specific cell type. In single cell ATAC-seq, this is harder.
 - If it's possible to get RNA-seq data from a similar set of cells, these can be annotated and then used to annotate the ATAC-seq clusters.
 - For this we can use gene activity scores (level of open chromatin around genes), as a proxy for gene expression.



Example from Seurat web site

6. Peak annotation

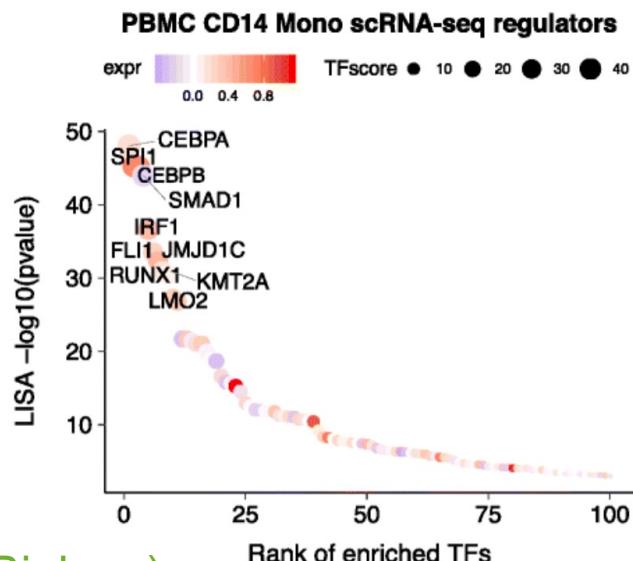
- To easier interpret the peaks, it's useful to note their location with regard to the nearest gene (or the nearest TSS).
- A region might not interact with the nearest gene, this is just a starting guess!



(Hinman & Cary. 2017, eLife)

7. Functional analysis

- Regions with open chromatin can be further analyzed, to see which transcription factors might bind there. This can give important information on which signaling pathways drive gene expression in different cells.
 - Looking for enriched motifs
 - Cross-referencing open chromatin regions against public ChIP-seq data on different TFs.
- This can be done for each cell or cluster of cells

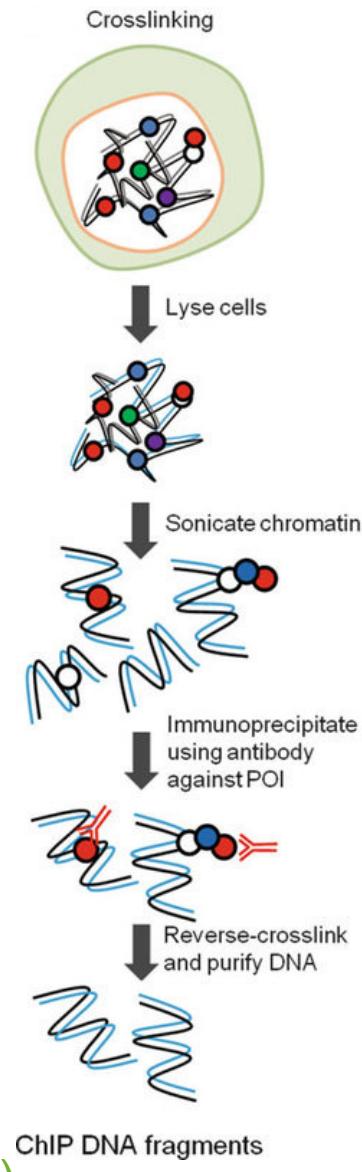


8. Differential DNA accessibility

- It's often interesting to know which chromatin regions differ in accessibility between cell types etc.
- This is a similar problem to differential gene expression (for RNA-seq data)
- Examples of methods:
 - Logistic regression
 - Negative binomial generalized linear model

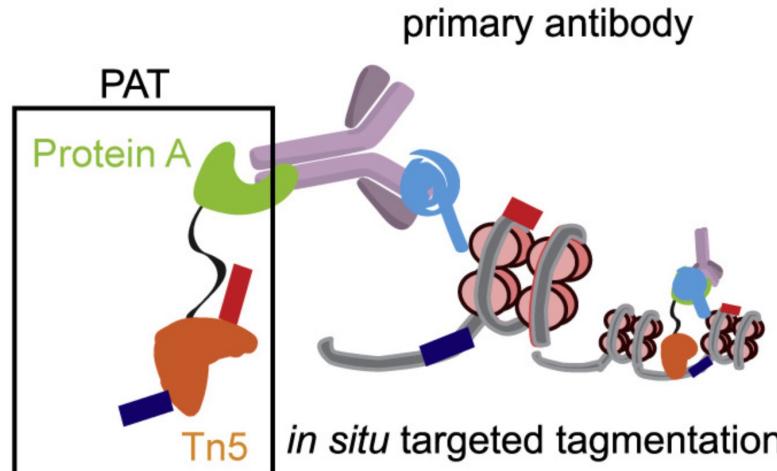
- **ATAC Cell Ranger**
 - Computational pipeline from 10X genomics, does (more or less) all of the analysis steps described here
- **Seurat/Signac**
 - R packages originally developed for single cell RNA-seq: Filtering cells, normalization, clustering, visualization, differential DNA accessibility. Data integration.
- **episcanpy**
 - Python package, originally developed for single cell RNA-seq. Similar functionality to Seurat/Signac
- **ChromVar**
 - R package, mostly useful for motif analysis. (Can do clustering, visualization, differential DNA accessibility too..)
- **Giggle**
 - Command line tool for cross-refencing genomic regions against public data sets.

- Chromatin Immuno-Precipitation, followed by sequencing
- Measures interactions between a protein of interest and DNA
- Uses an antibody towards the protein of interest to enrich for bound DNA.
- Analysis similar to ATAC-seq: finding peaks (regions with many reads mapping)
- ChIP on single cells, e.g. using droplets, is hard.
 - (Rotem et al. 2015, Nature Biotechnology) had around 800 reads/cell. Still enough to distinguish different cell types.
 - (Grosselin et al. 2019, Nature Genetics) had around 1600 reads/cell.



Single cell ChIP-seq like methods

- ChIP-free methods:
 - (Wang et al. 2019, Molecular Cell) CoBATCH
 - Antibody binds to protein of interest. → This recruits PAT complex with Tn5 → Tagmentation of DNA near protein of interest.
 - 12000 reads/cell
 - Combinatorial indexing (like for sci-ATAC-seq)
 - Quite simple protocol, no ChIP
 - (Kaya-Okur et al. 2019, Nature Communications) CUT&Tag, similar idea. (Used nanowells instead of combinatorial indexing.)



nature
biotechnology

ARTICLES

<https://doi.org/10.1038/s41587-021-00869-9>



Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues

Marek Bartosovic^{ID}¹✉, Mukund Kabbe¹ and Gonçalo Castelo-Branco^{ID}^{1,2}✉

In contrast to single-cell approaches for measuring gene expression and DNA accessibility, single-cell methods for analyzing histone modifications are limited by low sensitivity and throughput. Here, we combine the CUT&Tag technology, developed to measure bulk histone modifications, with droplet-based single-cell library preparation to produce high-quality single-cell data on chromatin modifications. We apply single-cell CUT&Tag (scCUT&Tag) to tens of thousands of cells of the mouse central nervous system and probe histone modifications characteristic of active promoters, enhancers and gene bodies (H3K4me3, H3K27ac and H3K36me3) and inactive regions (H3K27me3). These scCUT&Tag profiles were sufficient to determine cell identity and deconvolute regulatory principles such as promoter bivalency, spreading of H3K4me3 and promoter-enhancer connectivity. We also used scCUT&Tag to investigate the single-cell chromatin occupancy of transcription factor OLIG2 and the cohesin complex component RAD21. Our results indicate that analysis of histone modifications and transcription factor occupancy at single-cell resolution provides unique insights into epigenomic landscapes in the central nervous system.

-
- Data analysis for these methods is similar to single cell ATAC-seq.
 - Single cell ChIP-seq is still new, but developing fast. Throughput will likely increase a lot.

- Methyl group bound to cytosine in DNA, typically at CpG sites.
- Usually associated with repression of gene expression
- Bisulphite sequencing: converts cytosine residues to uracil, except where there is methylation
- Other approaches
 - using methylation-sensitive restriction enzymes
 - fluorescence-based

Single cell DNA methylation

- Methods
 - Whole genome vs reduced representation/targeted
 - Bisulphite vs bisulphite-free (methylation-sensitive restriction enzymes)
- Quite hard and expensive
- Data
 - Mostly 5mC
 - Thousands of cells
 - $10^4 - 10^7$ CpGs per cell
 - Not the same CpGs in all cells.
- Analysis still hard

Article

DNA methylation atlas of the mouse brain at single-cell resolution

<https://doi.org/10.1038/s41586-020-03182-8>

Received: 30 April 2020

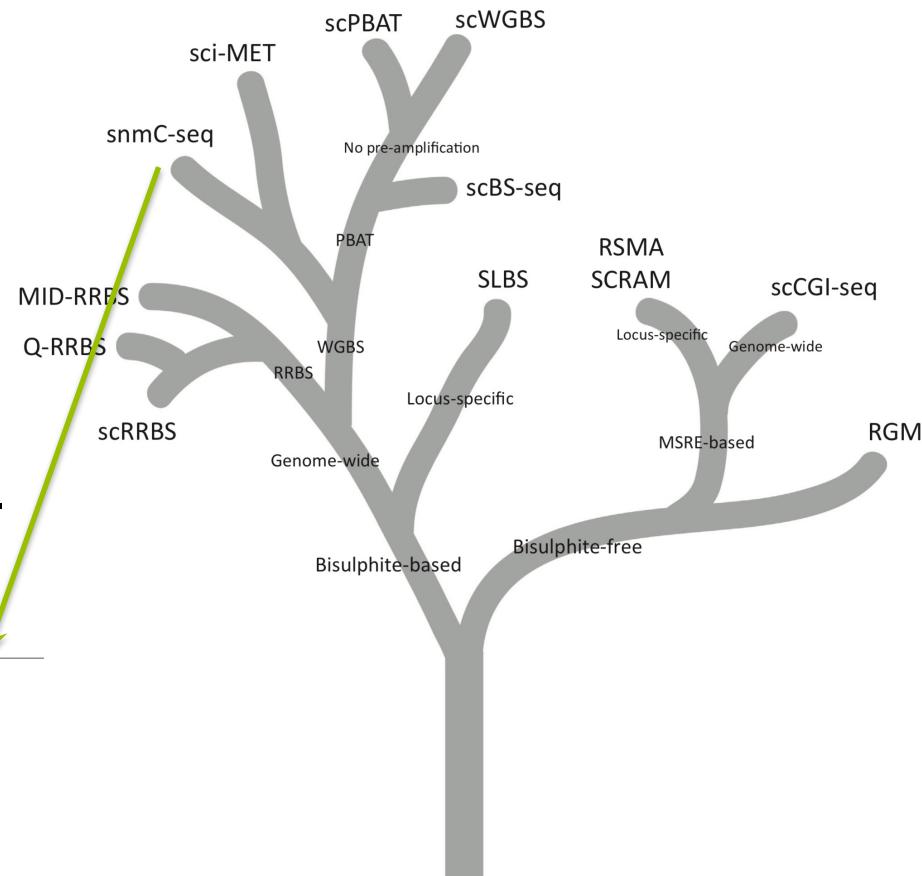
Accepted: 23 December 2020

Published online: 6 October 2021

Open access

Hanqing Liu^{1,2,16}, Jingtian Zhou^{1,3,16}, Wei Tian¹, Chongyuan Luo^{1,4}, Anna Bartlett¹, Andrew Aldridge¹, Jacinta Lucero⁵, Julia K. Osteen³, Joseph R. Nery¹, Huaming Chen¹, Angeline Rivkin¹, Rosa G. Castanon¹, Ben Clock⁶, Yang Eric Li⁷, Xiaomeng Hou^{8,9,10,11}, Olivier B. Poirion^{8,9,10,11}, Sebastian Preissl^{8,9,10,11}, Antonio Pinto-Duarte⁵, Carolyn O'Connor¹², Lara Boggeman¹², Conor Fitzpatrick¹³, Michael Nunn¹, Eran A. Mukamel¹³, Zhuzhu Zhang¹, Edward M. Callaway¹⁴, Bing Ren^{2,8,9,10,11}, Jesse R. Dixon⁶, M. Margarita Behrens⁵ & Joseph R. Ecker^{1,15,16}

>100K cells



Combining assays from the same cells

- Many methods combine several assays from the same cells, e.g.
 - scRNA-seq and scATAC-seq (10X genomics, SNARE-seq, and many more)**
 - scRNA-seq and sc-protein abundance (CITE-seq)
 - scRNA-seq and scDNA methylation
 - scRNA-seq and scDNA methylation and sc nucleosome (scNMT-seq)
 - scRNA-seq, scATAC-seq, sc-protein abundance and clonal info from mitochondrial DNA (DOGMA-seq)

Method	Molecular layers profiled				Throughput (low/medium/ high)	Special features (compared to techniques from same category)	Format	References				
	Epigenome		Genome	Transcriptome								
	Chromatin accessibility	Chromatin conformation										
scCAT-seq	x			x	+	usable fragments	well	Liu et al. (2019)				
Paired-seq	x			x	+++	throughput	well	Zhu et al. (2019)				
sc(ATAC + RNA)-seq	x			x	+	cost; simple workflow	well	Reyes et al. (2019a)				
sci-CAR	x			x	+++	acc. & RNA intersect coverage	well	Cao et al. (2018)				
SNARE-seq	x			x	+++	sensitivity	droplet	Chen et al. (2019)				
ASTAR-seq	x			x	++	price-performance ratio	microfluidics	Xing et al. (2020)				
SHARE-seq	x			x	+++	throughput, performance	well	Ma Sai, et al. (2020)				
ISSAAC-seq	x			x	+++	throughput, performance (esp. ATAC)	well/droplet	Xu et al. (2022)				
scDam&T-seq		x		x	+	protein-DNA interactions information	well	Rooijers et al. (2019)				
scNOMe-seq	x		x		+	estimates nucleosome phasing	well	Pott, (2017)				
scCOOL-seq	x		x		+	acc. & DNAme intersect coverage	well	Guo et al. (2017)				
iscCOOL-seq	x		x		++	accessibility coverage	well	Gu et al. (2019a)				
scMethyl-HIC		x	x		+	mapping rate	well	Li et al. (2019)				
sn-m3C-seq		x	x		+++	DNAme coverage	well	Lee et al. (2019)				
scNMT-seq	x		x	x	++	throughput	well	Clark et al. (2018)				
scNOMeRe-seq	x		x	x	+	DNAme coverage	well	Wang et al. (2021)				
scSIDR-seq			x	x	+	captures total RNA	well	Han et al. (2018)				
TARGET-seq			x	x	+++	cost; throughput	well	Rodriguez-Meira et al. (2019)				
RETrace		x	x		+	captures microsatellites	well	Wei and Zhang, (2020)				
scTrio-seq2		x	x	x	++	DNAme coverage	well	Bian et al. (2018)				

- Seurat/Signac supports analysis of multiomics data, e.g. from 10X
- Best to first analyze each data type separately (QC, filtering, clustering etc.)
- It's also possible to do a joint neighborhood graph to do clustering and UMAP on all data together
 - WNN: Weighted nearest neighbor
- Another option is to first cluster cells on RNA, then subcluster on e.g. ATAC-seq.

- Single cell ATAC-seq
 - Usually works quite well
 - Commercial kits available
 - Often used together with single cell RNA-seq
- Single cell ChIP-seq/cut & tag etc.
 - Data analysis for all of these methods is similar to single cell ATAC-seq.
 - Still new, but developing fast. Throughput will likely increase a lot.
- Single cell DNA methylation
 - A lot of development happening
 - Useful methods will become more widely available (already scWGBS at NGI/Scilifelab).
- Multi-comics
 - Many method available, a lot of development.
 - Seurat etc. can be used for such data

Some resources

-
- Signac website, lots of tutorials
 - <https://stuartlab.org/signac/>
 - Epigenomics data analysis course
 - <https://nbis-workshop-epigenomics.readthedocs.io/en/latest/>
 - Next occasion probably fall 2023
 - 10X genomics
 - <https://www.10xgenomics.com/products/single-cell-atac>
 - <https://www.10xgenomics.com/products/single-cell-multiome-atac-plus-gene-expression>

The End

