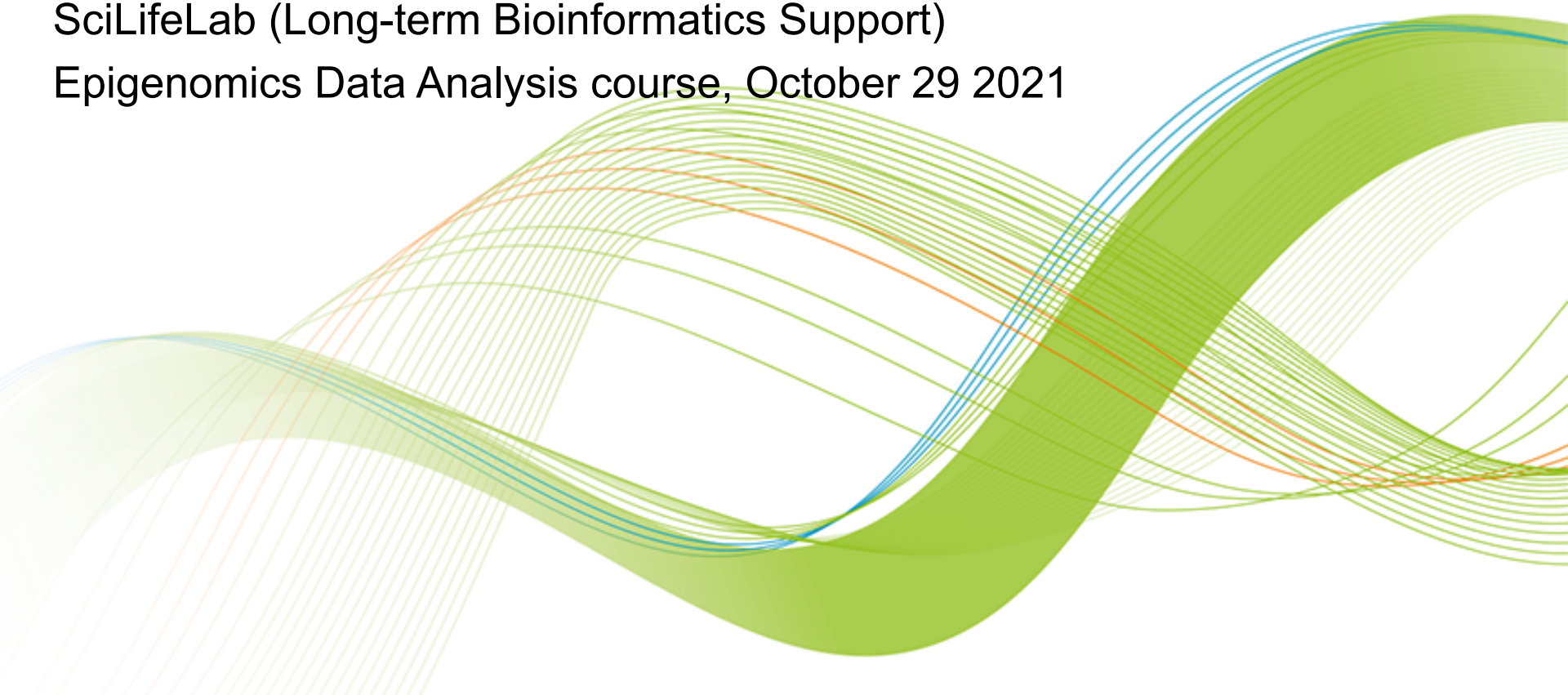
SciLifeLab

Single cell methods in epigenomics

Jakub Orzechowski Westholm

SciLifeLab (Long-term Bioinformatics Support)

Epigenomics Data Analysis course, October 29 2021



Single cell methods

SciLifeLab

- Analyzing bulk samples will give results that represent an average of the cell population, over thousands or millions of cells
- Single cell methods give results for each cell
  - Differences between individual cells (due to stochasticity, cell cycle etc.)
  - Differences between cell types
- More noise in data
- Require new analysis steps
- This is a big, and rapidly changing field. We have a whole course on single cell RNA-seq. In this talk we will give a short introduction to single cell methods in epigenomics.
  - Single cell ATAC-seq
  - (Single cell DNA methylation)

  - Single cell ChIP-seq-like methods will be covered next, by Marek Bartosovic

Todays talk

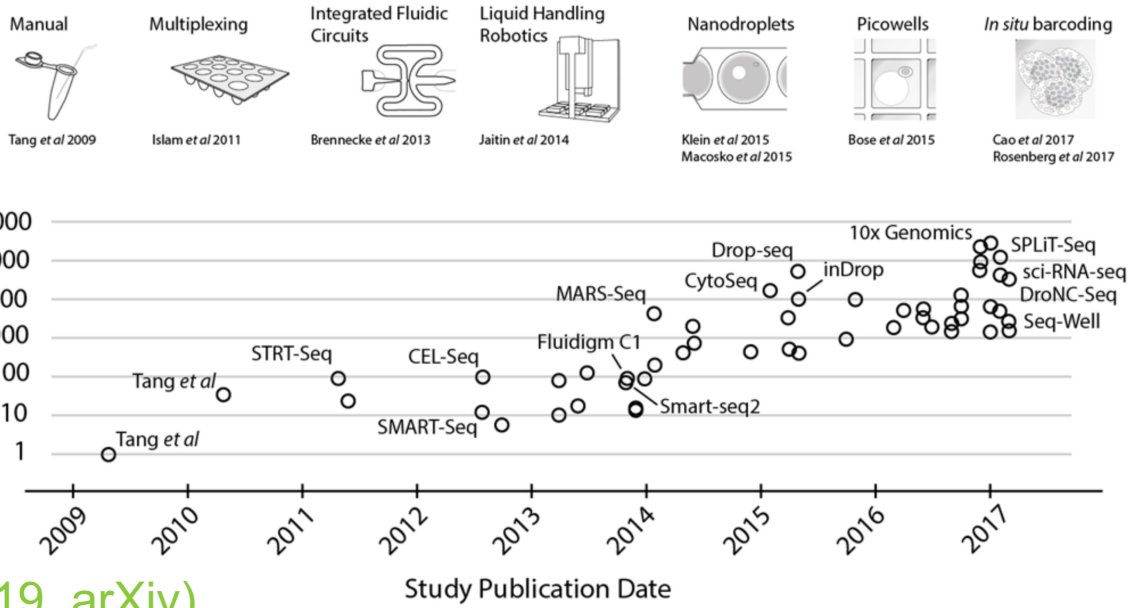
SciLifeLab

- Background on single cell methods
- Single cell ATAC-seq
- Single cell data analysis, common steps
- Single cell DNA-methylation

- Broad overview
- Focus on concepts over details.

Background

- Single cell methods started with RNA-seq
  - First paper: (Tang et al 2009, Nature Methods)
  - Output has increased a lot, from 1 cell in 2009 to millions of cells today
- Later this was adapted for ChIP-seq, ATAC-seq, DNA methylation and other assays
- Some experimental steps and analysis are similar, some are unique



(Svenson et al 2019, arXiv)

Recent example

SciLifeLab

RESEARCH ARTICLE

HUMAN GENOMICS

A human cell atlas of fetal gene expression

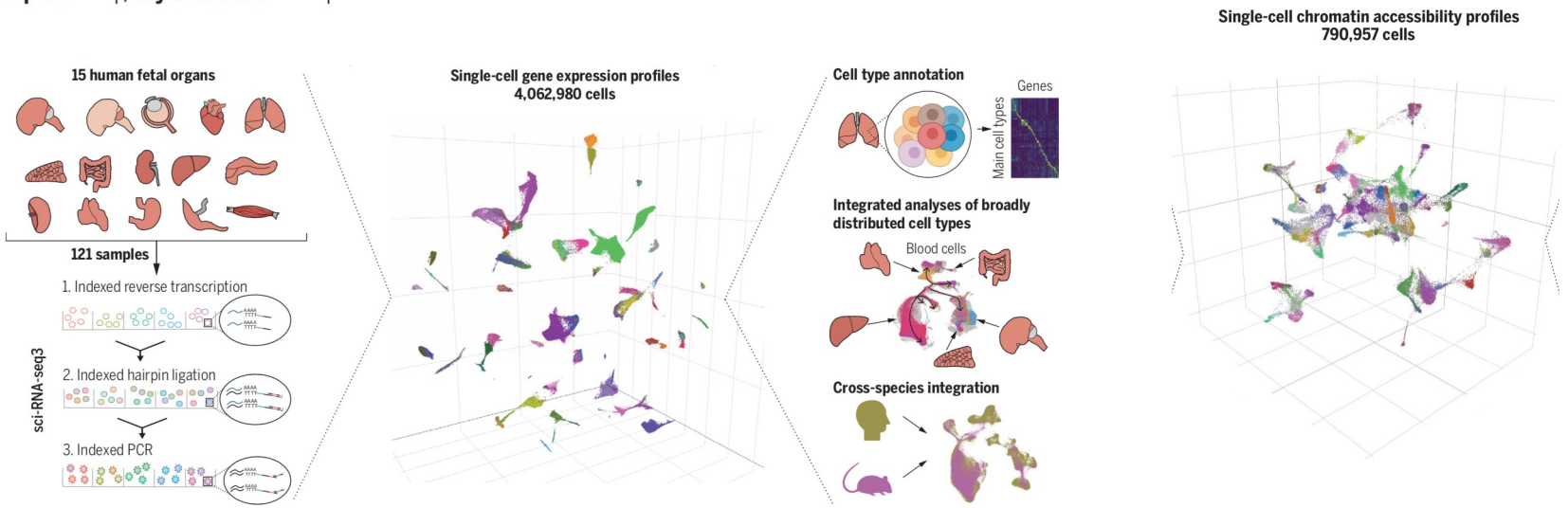
Junyue Cao[1]*, Diana R. O'Day[2], Hannah A. Pliner[3], Paul D. Kingsley[4], Mei Deng[2], Riza M. Daza[3], Michael A. Zager[3,5], Kimberly A. Aldinger[2,6], Ronnie Blecher-Gonen[1], Fan Zhang[7], Malte Spielmann[8], James Palis[4], Dan Doherty[2,3,6], Frank J. Steemers[7], Ian A. Glass[2,3,6], Cole Trapnell[1,3,10]†, Jay Shendure[1,3,10,11]†

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

A human cell atlas of fetal chromatin accessibility

Silvia Domcke*, Andrew J. Hill*, Riza M. Daza*, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell†, Darren A. Cusanovich†, Jay Shendure†



RNA-seq: 4 million cells

ATAC-seq: 800,000 cells

Home / News / Junyue Cao Awarded The Grand Prize For Developing Four New Single Cell Genomics Techniques

NOVEMBER 20 2020

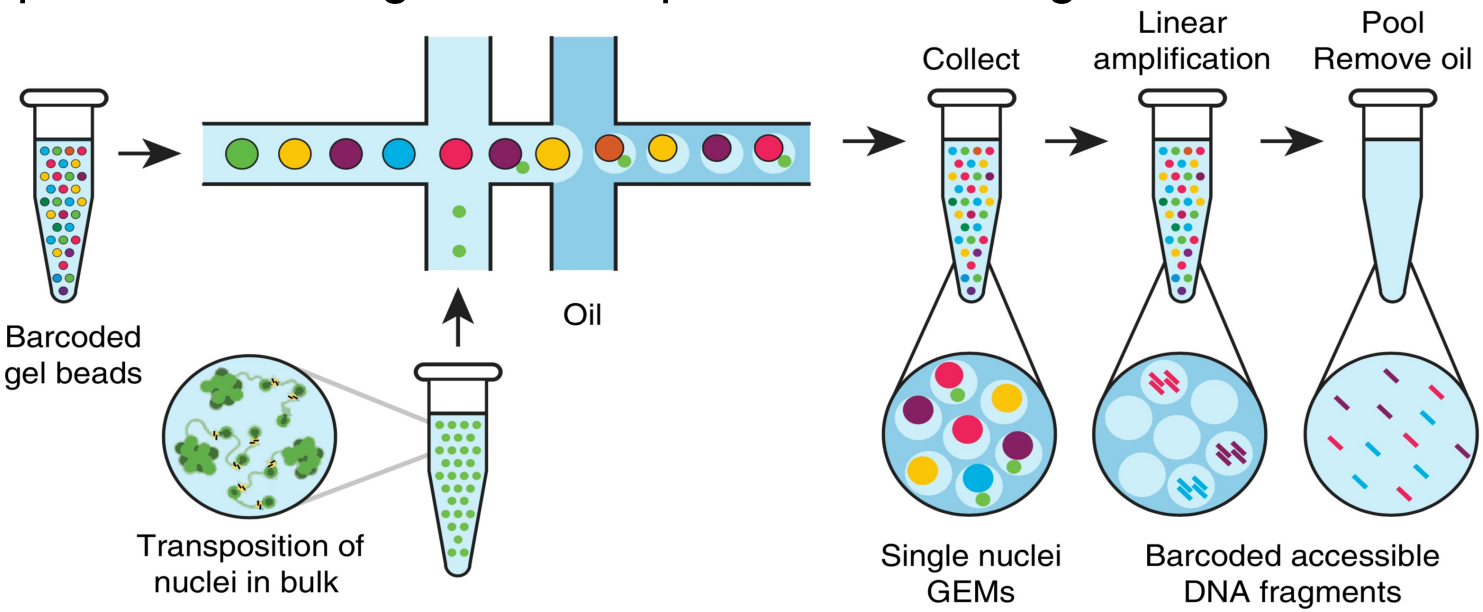
Junyue Cao awarded the Grand Prize for developing four new single-cell genomics techniques

Discuss!

- What are single cell technologies useful for? What types of questions can you answer?
- Examples of good (or bad) studies?
- Discuss in breakout rooms, for 5 minutes.
- Then give a short summary about your discussions.

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.

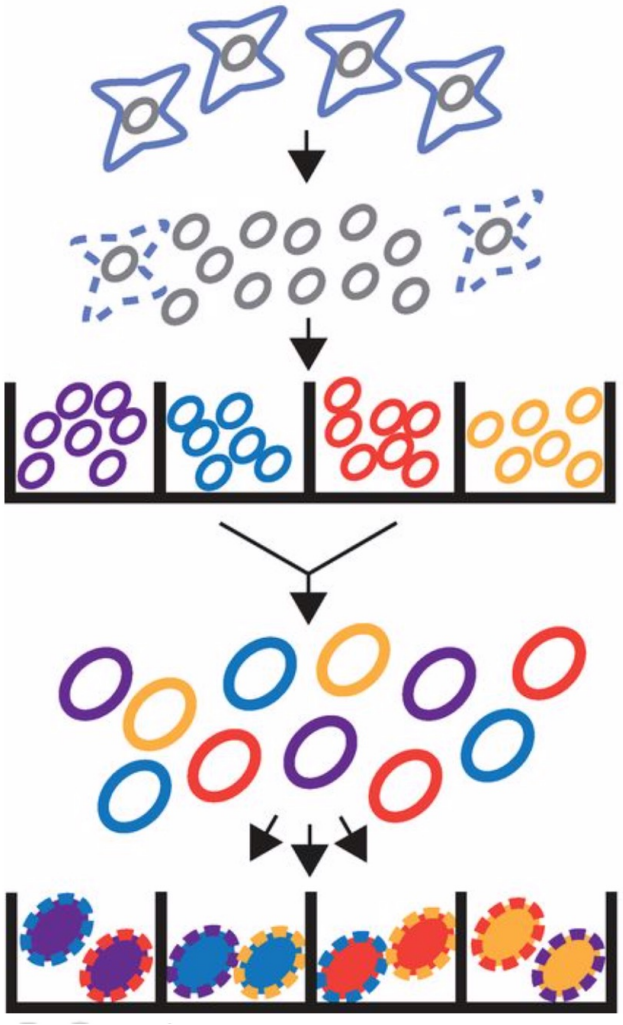


(Satpathy et al. 2019 Nature Biotechnology)

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.

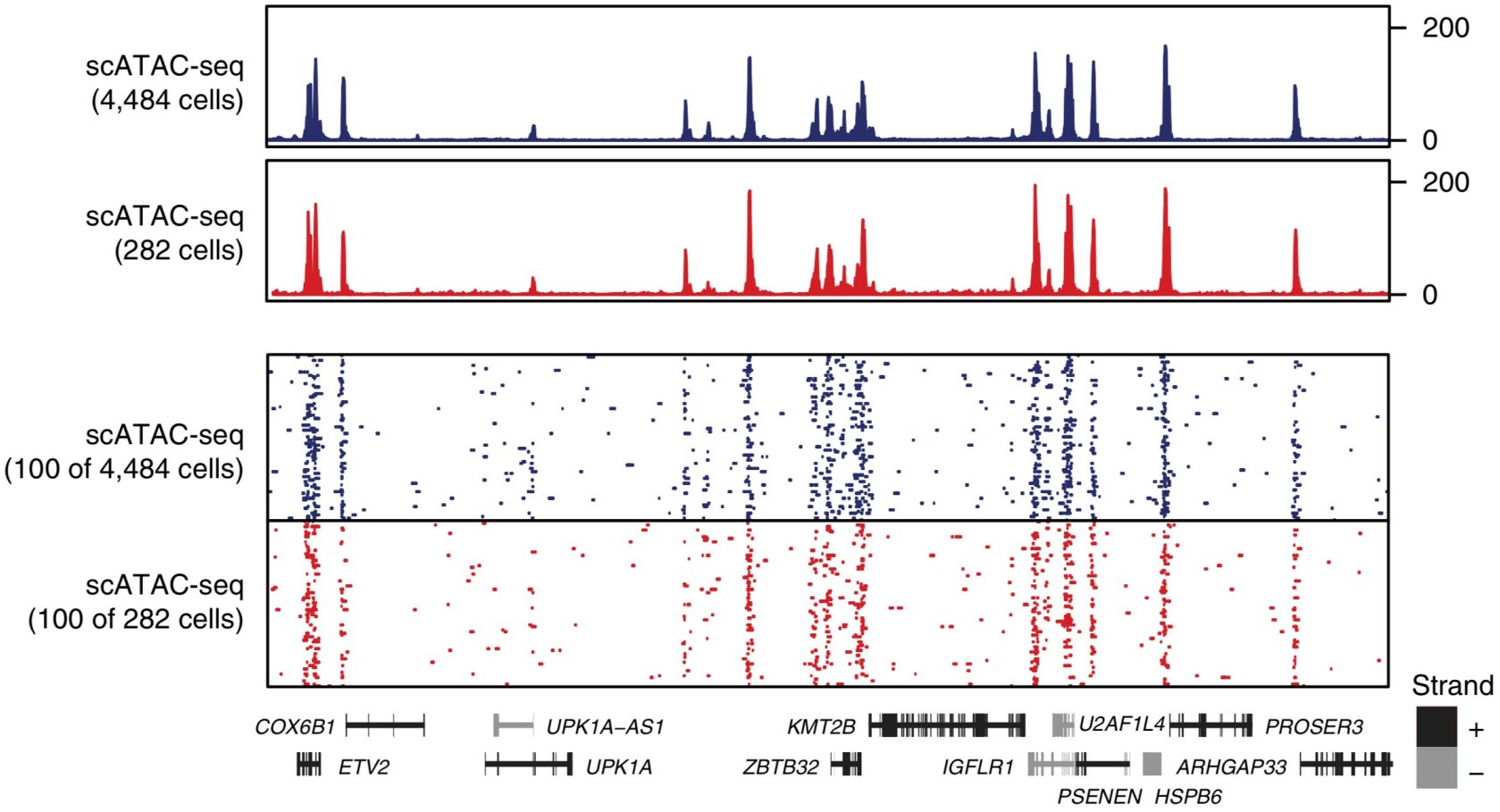
+ Low cost per cell, enables high throughput
- No commercial solution, so a bit harder to set up



Single cell ATAC-seq data

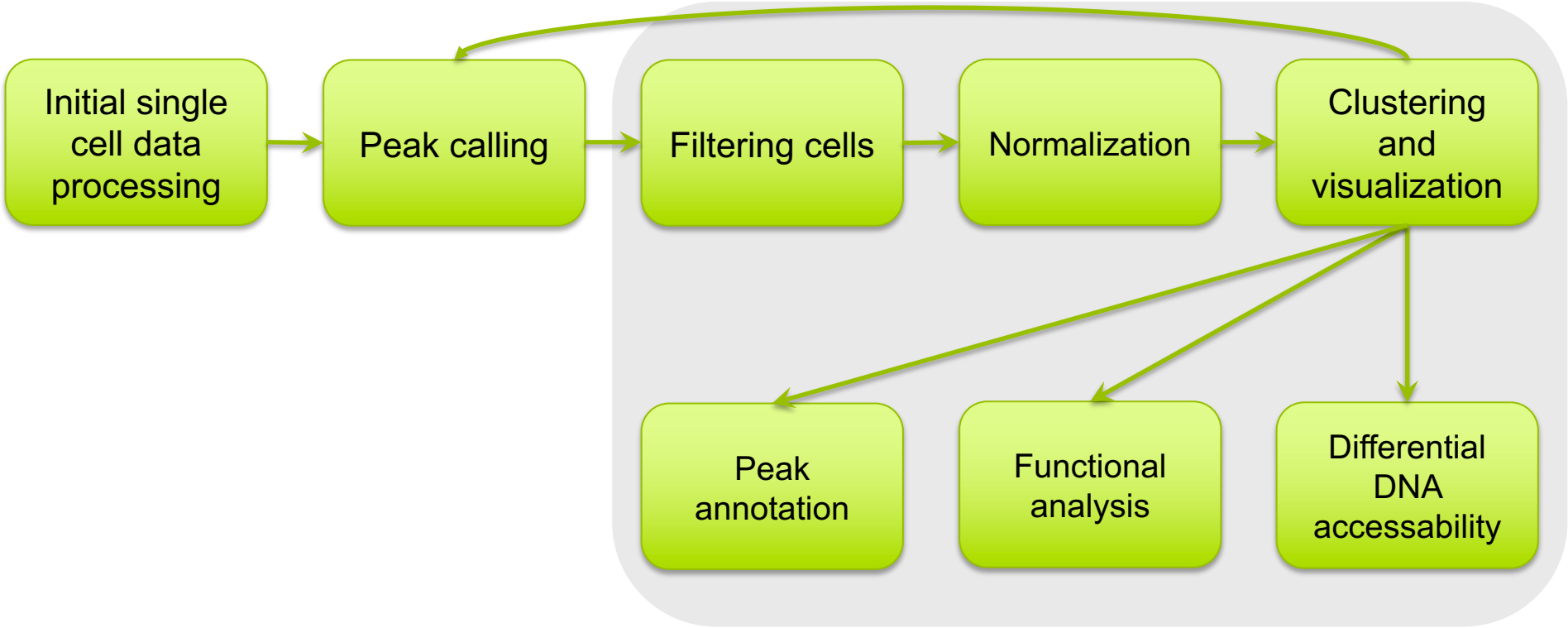
SciLifeLab

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



(Satpathy et al. 2019 Nature Biotechnology)

Single cell ATAC-seq data analysis



Exercise later today

1. Initial single cell data processing

SciLifeLab

- 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

SciLifeLab

- 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

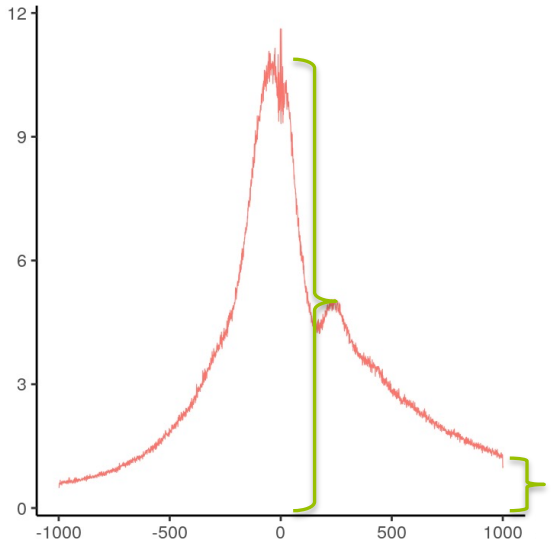
3. Filtering cells

SciLifeLab

- 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

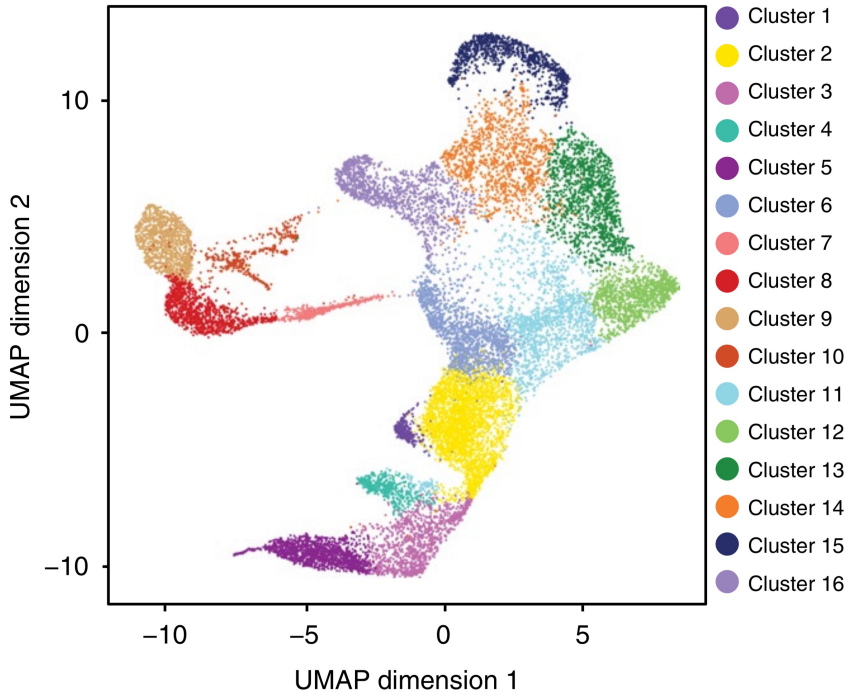
SciLifeLab

- 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

- Visualize how cells relate to each other
  - Project many dimensions down to 2.
  - Conceptually similar to PCA, but not linear
  - UMAP algorithm

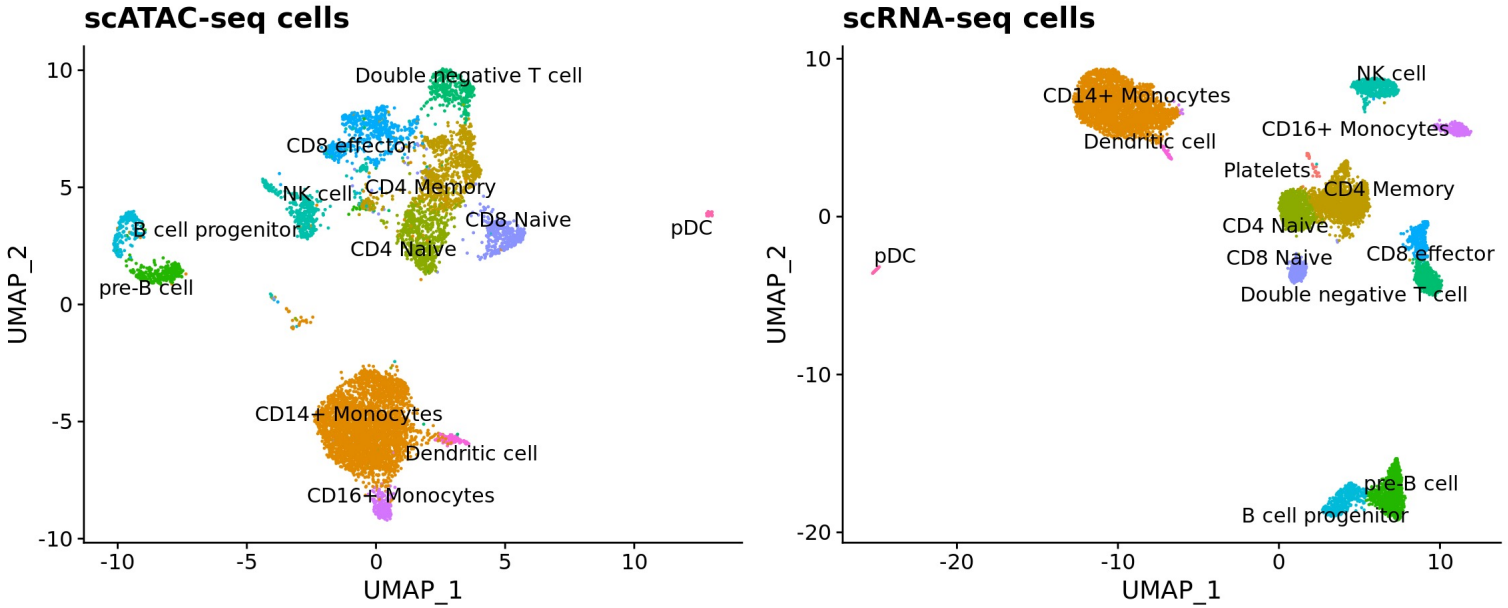
- Clustering, to identify groups of similar cells (representing different cell types or cell states).
  - There are many different clustering methods
  - Many settings for such methods
  - Trial and error



(Satpathy et al. 2019 Nature Biotechnology)

5. Cluster and visualize cells II

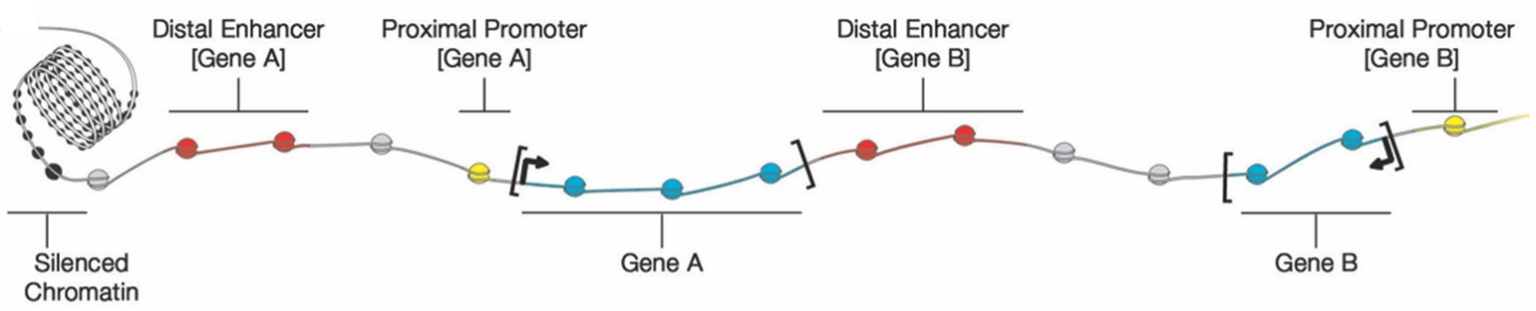
- It's often not clear which cell types etc. these clusters represent.
  - In single cell RNA-seq we can look at marker genes, unique to a specific cell type. In single cell ATAC-seq, this is harder.
  - If it's possible to get RNA-seq data from a similar set of cells, these can be annotated and then used to annotated the ATAC-seq clusters.
  - This is sometimes called label transfer, will be discussed in the next talk, about intergrating –omics data, and in the exercise.



Example from Seurat web site

6. Peak annotation

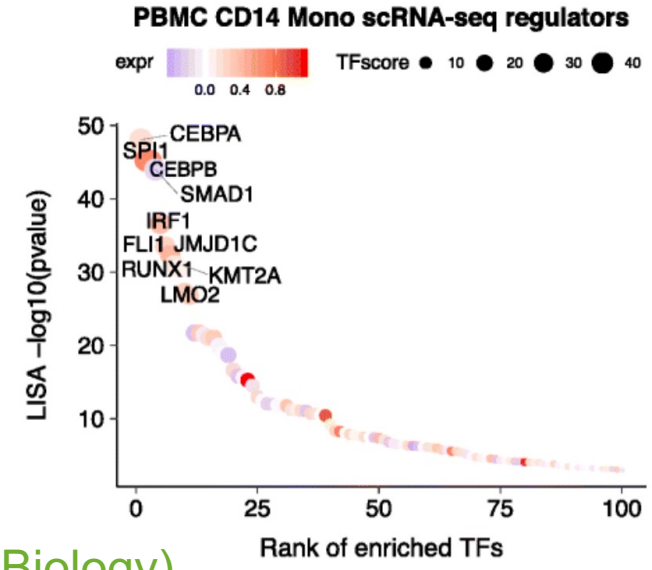
- To easier interpret the peaks, it's useful to note their location with regard to the nearest gene (or the nearest TSS).
  - Remember that a region might not interact with the nearest gene, this is just a starting guess!
- This is similar to what was discussed for ChIP-seq data.



(Hinman & Cary. 2017, eLife)

7. Functional analysis

- Like for bulk ATAC-seq the regions with open chromatin can be further analyzed, to see with 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



(Wang et al. 2020 Genome Biology)

8. Differential DNA accessibility

SciLifeLab

- 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) and differential binding (for ChIP-seq data).
- But single cell ATAC-seq data have special properties that make the analysis different from bulk analysis.
  - Sparse data (low read counts, most entries are 0)
  - Many replicates/cells.
- Examples of methods:
  - Logistic regression
  - Negative binomial generalized linear model

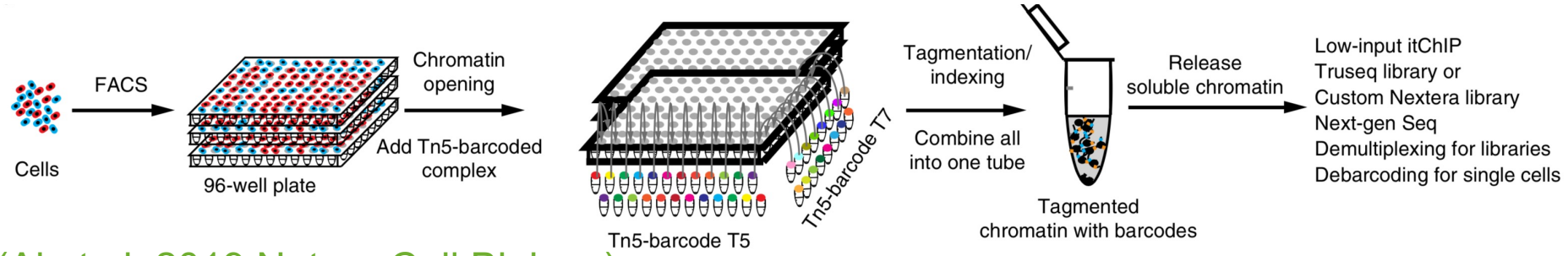
Single cell ATAC-seq analysis tools

SciLifeLab

- **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.

Single cell ChIP-seq

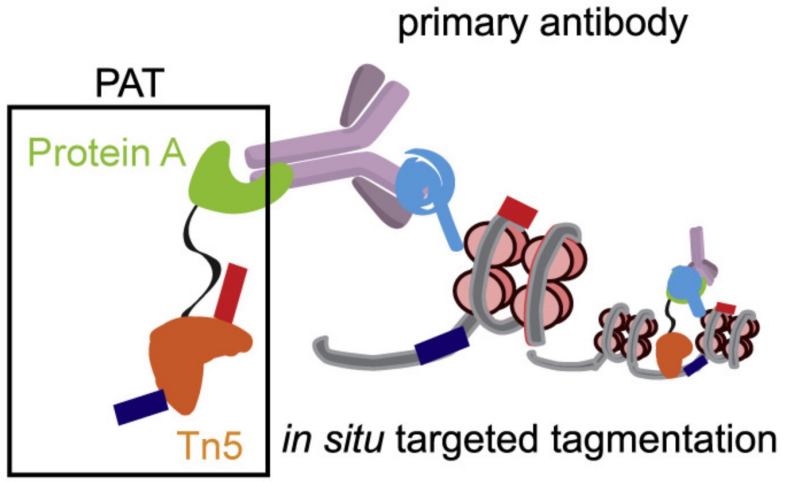
- ChIP on single cells, e.g. using droplets, is hard.
  - (Rotem et al. 2015, Nature Biotechnology) had around 800 reads/cell. Still enough to distinguish different cell types.
  - (Grosselin et al. 2019, Nature Genetics) had around 1600 reads/cell.
- Tagmentation based methods:
  - (Ai et al. 2019 Nature Cell Biology) came up with sc-itChIP-seq (single cell indexing and tagmentation ChIP-seq):
    - First use Tn5 transposase to add cell barcodes to DNA.
    - Then do ChIP in bulk.
    - 9000 reads / cell.
    - 96 well plates



(Ai et al. 2019 Nature Cell Biology)

Single cell ChIP-seq II

- 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.)



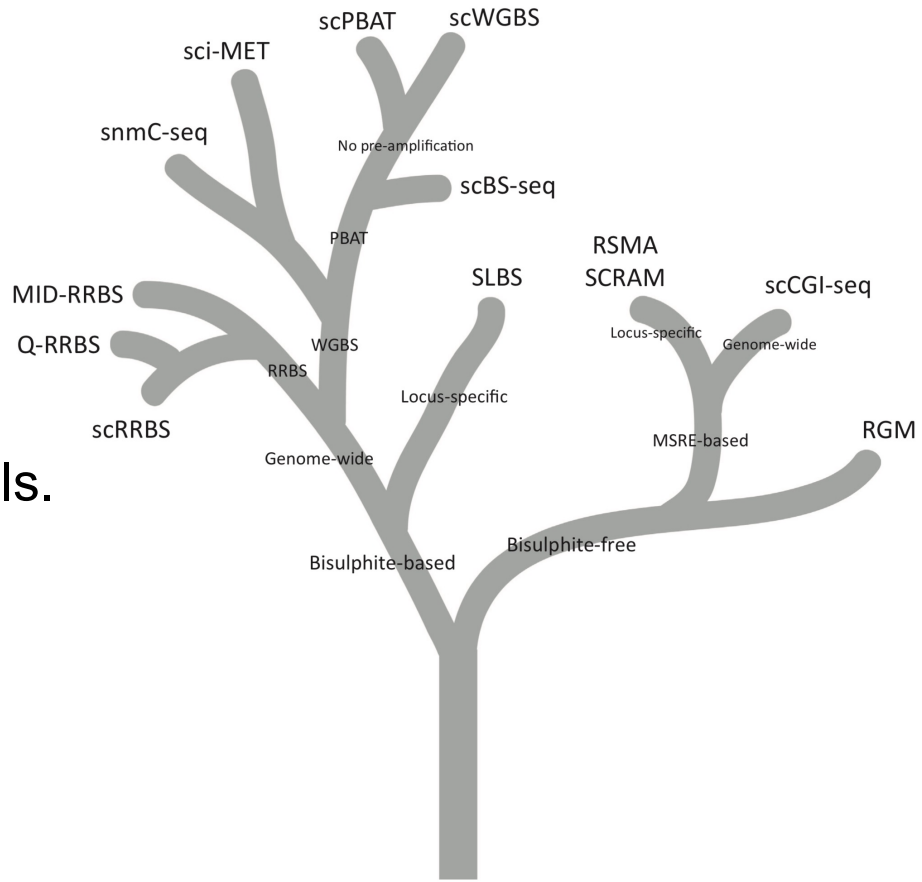
Single cell ChIP-seq III

SciLifeLab

- Data analysis for all of 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.

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

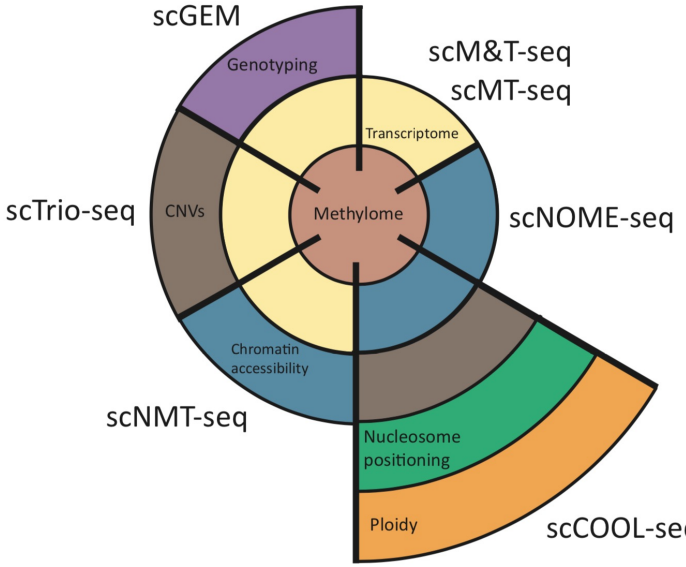


Karemaker & Vermeulen. 2018, Trends in Biotechnology

Combining assays from the same cells

SciLifeLab

- Many methods combine several assays from the same cells, e.g.
  - scRNA-seq and scATAC-seq (Chromium Single Cell Multiome from 10X genomics, SNARE-seq, and many other methods..)
  - 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



Karemaker & Vermeulen. 2018, Trends i

Summary

SciLifeLab

- Single cell ATAC-seq
  - Usually works quite well
  - Commercial kits available
  - You will have a look at the data analysis in the exercise.
- Single cell DNA methylation
  - A lot of development happening
  - Useful methods will become more widely available (already scWGBS at NGI/Scilifelab).


the End.