Single Cell and Spatial Omics

Course held by prof. Tebaldi, A.Y.22/23 - University of Trento

Ilaria Cherchi
Telegram: @ilariacherchi
Github: https://github.com/ilariache/Single-cell-and-spatial-omics

June 10, 2023

Contents

Chapter 1

Evolution of single cell

1.1 The history of sequencing

The history of sequencing begins in the 1950s with the first sequenced protein (Insulin) by Fred Sanger. The term "bioinformatics" first appears in 1979. From the 2000s, the Human Genome Project allows to 'crack the code' and sequencing costs dramatically reduce going towards 2010 with third generation sequencing. Data analysis becomes a central topic in genome studies, since sequencing requires a huge data management. The majority of biological data is stored by:

- NCBI (USA)
- SIB (Switzerland)
- EBI (Europe)
- BIG (China)
- DDBJ (Japan)

1.1.1 Analysis of gene expression in single neurons (1992)

The first two single cell articles focus on the characterization of single neuron gene expression in rat hippocampus and cerebellum, respectively. They applied two similar techniques:

- **in vitro transcription**, which amplifies the original material linearly through a transcription reaction
- **PCR**, to obtain an exponential increase in transcripts

1.1.1.1 Analysis of gene expression in single live neurons (IVT)

Quotes from the paper:

"The RNA from **defined single cells** is amplified by microinjecting primer, nucleotides, and enzyme into acutely dissociated cells from a defined region of rat brain."

Table 1. Statistical Analysis of Single-Cell Transcriptional Profiles

Cell Type		Pairwise Corr.
OSN (repeat)	n = 4	0.97 ± 0.01
Single HGC	n = 6	0.86 ± 0.03
Groups of 10 HGCs	n = 5	0.92 ± 0.04
10 pg dilution	n = 6	0.84 ± 0.03
100 pg dilution	n = 7	0.95 ± 0.02
OSNs	n = 9	0.68 ± 0.05
OPCs	n = 7	0.71 ± 0.05
OSNs versus OPCs		0.57 ± 0.04

probe set	gene name	p value	mADV OSNs (±sem)	mADV OPCs (±sem)	logFC/ 95% CI ratio
OPC-enriched					
aa250191_s_at	Hes6	0.0004	-8 ±62	10839 ±3887	8.10
m89293_rc_at	Id2	0.0248	96 ±44	3093 ±2113	2.38
m60523_s_at	Id3	0.0205	435 ±342	1777 ±813	0.90
m64292_s_at	Tis21	0.0238	127 ±201	1142 ±689	1.26
u81603_s_at	Eya2	<0.0001	-60 ±12	49 ±26	0.63
x80339_s_at	Six1	0.0415	60 ±68	430 ±290	0.94
u52951_s_at	Enx1	0.0029	-7 ±17	4799 ±2244	5.29
u78103_s_at	Eed	0.0964	123 ±104	1616 ±1649	1.57
d83596_s_at	Etf	0.0031	33 ±55	854 ±381	1.73
Msa.22493.0_s_at	Rbtn1	0.0045	-254 ±131	753 ±481	3.04
y08361_s_at	RIL-pending	0.0307	-157 ±125	10679 ±8076	8.04
aa606494_s_at	Unknown 1 (5730427N09Rik)	0.0009	184 ±189	3834 ±1415	2.18
w12941_s_at	Unknown 2 (1110004C05Rik)	0.0273	-74 ±40	2673 ±1985	5.18
aa409089_rc_s_at	Unknown 3 (2410018C03Rik)	0.0036	-65 ±58	2471 ±1220	5.05
OSN-enriched					
aa270796_s_at	Oda10	<0.0001	2568 ±416	43 ±37	2.83
aa546047_s_at	Unknown 4 (AA546047)	0.0004	8598 ±2648	53 ±95	4.46
Msa.2916.0_s_at	Axcam	0.0416	1000 ±711	-10 ±20	3.94
d50311_s_at	Mef2b	0.0428	946 ±741	-100 ±47	3.81

Figure 1.1: *Tietjen I. et al. Single-cell transcriptional analysis of neuronal progenitors. Neuron. 2003 Apr 24;38(2):161-75.*

"We demonstrate this method by constructing expression profiles of single live cells from rat hippocampus".

"This profiling suggests that cells that appear to be morphologically similar may show marked differences in patterns of expression."

1.1.1.2 AMPA receptor subunits expressed by single Purkinje cells

Purkinje cells are a class of GABAergic inhibitory neurons located in the cerebellum. It was possible to determine the subset of AMPA receptor subunits expressed by single cerebellar Purkinje cells in culture by combining whole-cell patch-clamp recordings and a molecular analysis, based on PCR, of the messenger RNAs harvested into the pipette at the end of each recording.

1.1.2 Single-cell transcriptional analysis of neuronal progenitors (2003)

Single cell analysis by **microarray** in the olfactory system, which is composed of well-characterized cells in terms of shape and differentiation stages (e.g. olfactory progenitors (OPCs) and mature sensory neurons (OSNs)) $\rightarrow$ ideal system to test. The material was amplified with PCR and analyzed through an Affymetrix microarray.

> **RQ:** *Find the number of genes that were used:* Affymetrix Mu11K high-density oligonucleotide arrays containing 13027 probe sets.

Example of statistical analysis, quantification of correlation:

- technical repeat with OSN cell to tune the system

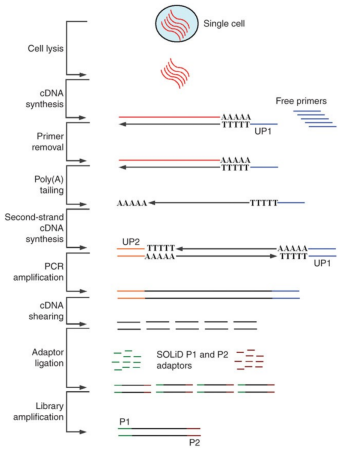

Figure 1.2: *Tang, F. et al. mRNA-Seq whole transcriptome analysis of a single cell. Nat Methods* ***6,*** *377–382 (2009)*

- measure the correlation among different cell types; variation could be due to stochasticity, biological variation (rather than technical)

Quotes from the paper:

"To uncover mechanisms involved in neuronal differentiation and diversification, we have monitored the expression profiles of **individual neurons and progenitor cells** collected from dissociated tissue or laser captured from intact brain slices."

"This technique provides a sensitive and reproducible **representation of the single-cell transcriptome**."

"In the olfactory system, hundreds of transcriptional differences were identified between olfactory progenitors (OPCs) and mature sensory neurons (OSNs), enabling us to define the large variety of signaling pathways expressed by individual progenitors at a precise developmental stage."

1.1.3 mRNA-Seq whole-transcriptome analysis of a single cell (2009)

1.1.3.1 Method

The single cell is manually picked under a microscope and lysed. Then mRNAs are reverse-transcribed into cDNAs using a poly(T) primer with anchor sequence (UP1) and unused primers are digested. Poly(A) tails are added to the first-strand cDNAs at the 3' end, and second-strand cDNAs are synthesized using poly(T) primers with another anchor sequence (UP2). Then cDNAs are evenly amplified by PCR using UP1 and UP2 primers, fragmented, and P1 and P2 adaptors are ligated to the ends.

1.1.3.2 Results

More than 100 million 35-base reads and 50-base reads were obtained from the cRNA expression profile of a single mouse blastomere. The method detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads.

Coverage plots: determine whether mRNA-Seq assay could be used to dissect the functional consequences when one of the critical genes for microRNA synthesis, *Dicer1*, was conditionally knocked out during oocyte development. The reads were found in exons with sharp boundaries at the exon-intron junction, confirming the single-exon resolution of the mRNA-Seq reads. This demonstrated that the single cell mRNA-Seq assay is accurate and has low or even no background noise.

Pros:

- increasing the cDNA length up to 3 kb by extending the incubation time for reverse transcription allows to capture full-length cDNAs for the majority of expressed genes.
- the assay can be used to discover new transcripts and alternative splicing isoforms.
- detect low-level transcripts - important for early embryo studies because some of the key transcription factors are expressed at very low levels

Cons:

- mRNAs without poly(A) tails, such as histone mRNAs, will not be detected by our mRNA-Seq assay
- for most of the mRNAs longer than 3 kb, the 5' end that is more than 3 kb away from the 3' end of the mRNA will not be detected
- the assay uses double-stranded cDNAs but cannot discriminate between sense and antisense transcripts

1.1.4 Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq (2011)

1.1.4.1 Methods

Isolated single cells were barcoded individually in a 96-well plate, then transferred into a single tube, called *single-cell tagged reverse transcription sequencing* (STRT-seq), for PCR amplification. The amplified samples were then adapted for Illumina sequencing.

The feasibility of the strategy was demonstrated by analyzing the transcriptomes of 85 single cells collected from two different mouse cell types: embryonic stem cells (ES) and embryonic fibroblasts (MEFs).

1.1.4.2 Results

Single-cell RNA-seq expression profiles were clustered to form a two-dimensional cell map onto which expression data were projected. Three levels of organization are depicted: the whole population of cells, the functionally distinct subpopulations it contains, and the single cells themselves.

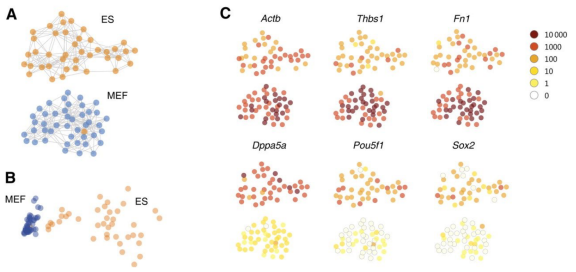

Figure 1.3: *Islam et al. Genome Res. 2011*

RQ: *How did the authors quantify the similarity between cells in panel A?*
In the two-dimensional map, more closely related cells are located near each other, so as to be able to detect and distinguish cell types based solely on expression data. In the underlying graph, the nodes represent cells and edges represent cell-to-cell similarity of expression pattern. Raw reads were normalized to TPM - obtained by counting the number of hits to each annotated gene and normalizing the data to transcripts per million. The similarity measure was computed through the Bray-Curtis distance, which handles the noise in low-expressed genes better than standard correlation.

- Bray-Curtis dissimilarity

  Is a statistic used to quantify the compositional dissimilarity between two different sites, based on counts at each site.

  $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$

  Where C_{ij} is the sum of the lesser values (see example below) for only those species in common between both sites. S_i and S_j are the total number of specimens counted at both sites.

  The Bray–Curtis dissimilarity is bounded between 0 and 1, where 0 means the two sites have the same composition (that is they share all the species), and 1 means the two sites do not share any species. It is not a distance since it does not satisfy triangle inequality, and should always be called a dissimilarity to avoid confusion.

1.1.5 CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification (2012)

Study of early *C. elegans* embryonic development at single-cell resolution. Differential distribution of transcripts between sister cells is seen as early as the two-cell stage embryo, and zygotic expression in the somatic cell lineages is enriched for transcription factors.

1.1.5.1 Methods

Individual cells are added to tubes, each with a uniquely bar-coded primer for reverse transcription. After second-strand synthesis, the reactions are pooled for IVT. The amplified RNA is then fragmented and

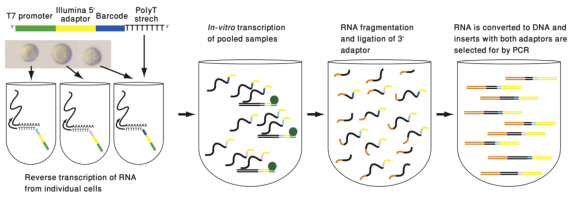

Figure 1.4: Hashimshony et al., Cell Rep,2012

purified before entry into a modified version of the Illumina directional RNA protocol, the molecules with both Illumina adaptors are selected, and the DNA library is sequenced with paired-end reads.

RQ: *Did the authors compare CEL-seq to STRT-seq? If so, how?*
CEL-Seq was applied to the same cell types by Islam et al., 2011 (STRT-seq)

1. compare the distribution of expression levels of each single-cell transcriptome across methods and cell type → CEL-Seq shows more reproducible distributions of expression, higher correlations for ES cells and distinguishes between cell types more clearly
2. evaluate the amount of detected genes → CEL-Seq detects significantly more genes in the ES cells.
3. quantify reproducibility for each method separately → with CEL-Seq, significantly lower noise was detected across biological replicates for both cell types tested.
4. compare PCA → principal component analysis based on CEL-Seq data better distinguishes between cell types than the corresponding STRT data

1.1.5.2 Results

CEL-Seq can be used to identify biological differences between closely related cells in the *C. elegans* embryo. *C. elegans* embryonic development begins with unequal cleavages producing founder cells—termed blastomeres. The AB and P1 sister blastomeres were examined 10 min after cell division → 17 genes with a mean 2-fold difference showing significantly different expression.

(A) Differential expression analysis between AB and P1
(B) The blastomeres examined in the study, with numbers of new transcripts and differential transcripts between anterior and posterior blastomeres.
(C) Gene expression levels (log10 tpm; see color scale on right) for the indicated genes; cell lineage is as in (B).
(D) Classification performance of the AB and P1 blastomeres.

Pros:

- ability to harness the power of IVT, providing both multiplexing and reproducibility
- reduced hands-on time both for the amplification and downstream processing, allowing for the preparation of dozens of samples for sequencing within 2–3 days

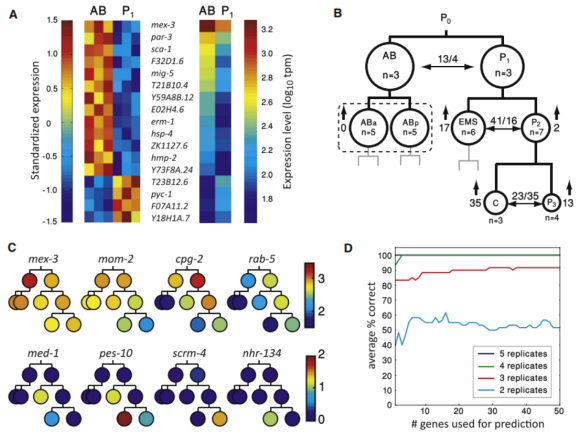

Figure 1.5: *Hashimshony et al. Cell Rep. 2012*

- commercially available kits for the amplification and sequencing library preparation → cost effective

Cons:

- mRNAs without poly(A) tails, such as histone mRNAs, will not be detected by our mRNA-Seq assay
- for most of the mRNAs longer than 3 kb, the 5' end that is more than 3 kb away from the 3' end of the mRNA will not be detected
- the assay uses double-stranded cDNAs but cannot discriminate between sense and antisense transcripts

1.1.6 SMART-seq

1.1.6.1 Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells (2012)

"Compared with existing methods, Smart-Seq has improved read coverage across transcripts, which enhances detailed analyses of alternative transcript isoforms and identification of single-nucleotide polymorphisms. The assay was applied to putative circulating tumor cells (CTCs) captured from the blood of a melanoma patient".

Clustering of single cell transcriptomes, based on expression of highly expressed genes (>100 RPKM). Confidence in clusters are indicated. Samples analyzed: human immune samples cells from putative melanoma CTCs (CTC), primary melanocytes (PM), melanoma cell lines SKMEL5 (SKMEL) and UACC257 (UACC), prostate cancer cell lines (LNCaP and PC3), bladder cancer cell line (T24) and human embryonic stem cells (ESC).

1.1.6.2 Smart-seq2 for sensitive full-length transcriptome profiling in single cells (2013)

Advantages:

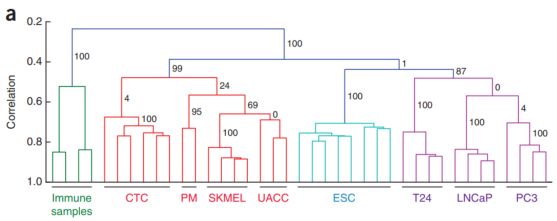

Figure 1.6: *Ramskold et al, et al. Nature Biotech 2012*

- **High coverage** across transcripts (recovery of full-length cDNAs)
- High level of mappable reads per cell (up to 10^6)

Limitations:

- lack of strand specificity
- inability to detect nonpolyadenylated (polyA) RNA
- low number of cells ($< 10^3$)

RQ: *What is the important step to obtain full length coverage?*
Template switching allows for the capture of full cDNA sequences, for which the 5' end is unknown, in a relatively sensitive manner. Poly(A)+ RNA is converted to full-length cDNA using oligo(dT) priming and SMART template switching technology. In addition, KAPA preamplification improves cDNA generation, allowing for the detection of more genes at higher GC levels and improved sensitivity and accuracy, as well as read coverage.

1.1.6.3 Single-cell RNA counting at allele and isoform resolution using Smart-seq3 (2020)

Advantages:

- High coverage across transcripts (recovery of full-length cDNAs)
- Suitable for **isoform**/allele analysis
- High level of mappable reads per cell (up to 10^6)

Limitations:

- inability to detect nonpolyadenylated (polyA) RNA.
- low number of cells ($10^3, 10^4$)

1.1.7 Method of the year (2013)

Single-cell sequencing was chosen by Nature as Method of the Year 2013.

"Technologies for single-cell amplification and sequencing are maturing. As the cost and ease of examining individual cells improves, the approach will enter the hands of more researchers as a standard tool for understanding biology at high resolution". The focus is on both RNA and DNA single cell sequencing.

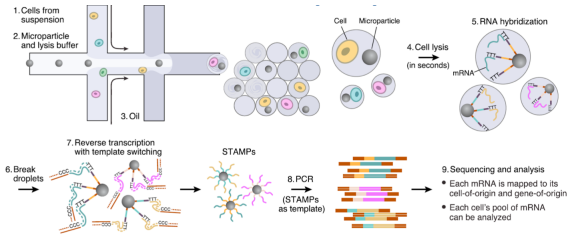

Figure 1.7: Macosko et al., Cell 2015

1.1.8 Drop-seq (2015)

A microfluidic device joins two aqueous flows. One contains cells, the other contains barcoded primer beads. After compartmentalization into discrete droplets, the cell is lysed and releases its mRNAs, which then hybridize to the primers on the microparticle surface. The droplets are broken and the microparticles collected and washed. The mRNAs are then reverse-transcribed in bulk, forming STAMPs, and template switching is used to introduce a PCR handle downstream of the synthesized cDNA.

Barcoding: sequence of primers on the microparticle. The primers on all beads contain a common sequence ("PCR handle") for PCR amplification. Each microparticle contains more than 10^8 individual primers that share the same "cell barcode" but have different **unique molecular identifiers** (UMIs), enabling mRNA transcripts to be digitally counted. A 30-bp oligo dT sequence is present at the end of all primer sequences for capture of mRNAs.

Pros:

- From hundreds to thousands cells (10^4)
- 3' end sequencing
- Use of UMI (Unique Molecular Identifiers)

1.1.8.1 UMI

Unique Molecular Identifiers are composed of a randomized nucleotide sequence (8-12 nt long) incorporated into the complementary DNA in the initial steps of RNA-seq protocol before the subsequent amplification steps. The goal of UMIs , in case of reads with identical cDNA sequence, is to distinguish between:

- amplified copies of the same mRNA molecule (same cDNA sequence, same UMI $\rightarrow$ technical duplicates, removed)
- reads from separate mRNA molecules transcribed from the same gene (same cDNA, different UMI $\rightarrow$ biological duplicates, kept)

UMIs reduce the amplification noise by allowing the removal of amplification (PCR) duplicates.

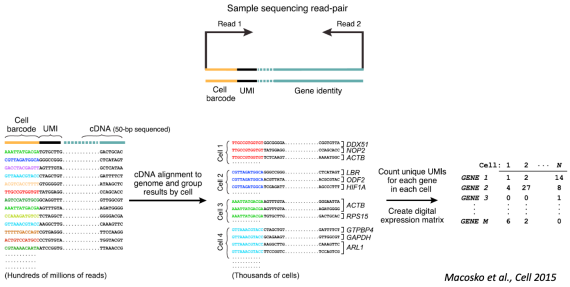

Figure 1.8: Macosko et al., Cell 2015

1.1.8.2 Cell cycle analysis

Classification of cells based on the expression of **known cell cycle genes.** A **phase-specific** score was calculated for each cell across **five** phases from gene expression levels. Classic cell cycle genes encompass CCNB1&2, MCM2-7,MCM10, ARUKA and AURKB.

1.1.8.3 Analysis of the mouse retina (45k cells)

Analysis of mouse retina, quite well known system - neurons with distinct morphological features. From 45k cells, the authors performed multidimensional scaling and reduction; they were able to identify 39 clusters with similar expression. The color of the clusters reflects neuron type, based on known marker genes.

The expression of the markers is specific, visualized as a violin plot.

RQ: *What are the computational tools used for the analysis?*

1. PCA on STAMPs and tSNE
2. combined a density clustering approach with post hoc differential expression analysis to divide 44,808 cells among 39 transcriptionally distinct clusters
3. organized the 39 cell populations into larger categories (classes) by building a dendrogram (hierarchical clustering) of similarity relationships among the 39 cell populations

1.1.8.4 inDrop device

Droplet based approach, published on the same Cell issue containing the Drop-seq paper. Each component has an inlet: RT mix, cells, DNA barcoding hydrogels (at the left of the instrument) and oil (in the middle). Cells and DNA barcoding hydrogels are then isolated in the collection outlet.

Drop-seq can be applied for evaluating the heterogeneity in differentiating ES cells.

Loadings of the components are useful to identify the expression of genes having a major impact in characterizing clusters. Example: genes connected with translation, elongation factors, cytoskeleton...

Analysis: differentiation of embryonic stem cells after leukemia inhibitory factor(LIF)withdrawal. ES cells maintained in serum exhibit well-characterized fluctuations, but are uniform compared to differentiated

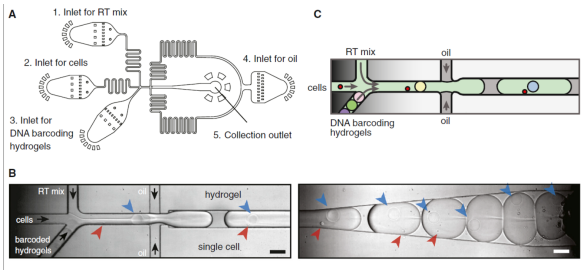

Figure 1.9: *Klein et al, Cell 2015*

cell types and thus pose a challenge for single cell sequencing. The differentiating ES cell population underwent significant changes in population structure, qualitatively seen by hierarchical clustering cells. In the first four principal components it was possible to identify:

- one rare population (6/935 cells) with very low levels of pluripotency markers and high levels of PrEn markers
- a second cell population (15/935 cells) expressed high levels of *Krt8*, *Krt18*, *S100a6, Sfn* and other markers of the epiblast lineage.
- the third population represented a seemingly uncharacterized state, marked by expression of heat shock proteins *Hsp90*, *Hspa5* and other ER components such as the disulphide isomerase *Pdia6*.

These sub-populations expressed low levels of pluripotency factors, suggesting they are biased toward differentiation or have already exited the pluripotent state. The latter population could also reflect stressed cells.

They also visualized clusters at different stages in order to determine subpopulation characterized by biomarkers. Some cells failed to express epiblast markers and a fraction of these expressed pluripotency factors at undifferentiated levels even seven days after LIF withdrawal $\rightarrow$ after 7 days, 5% (N=799) of cells overlapped with the ES cell population.

1.1.9 Human Cell Atlas (2016)

International group of researchers (2,300 members, 83 countries) use a combination of single-cell and spatial omics technologies to create **cellular reference maps** with the position, function, and characteristics of **every cell type in the human body**.

1.1.10 10X Genomics (2016-2017)

10x Technology samples a pool of ~750,000 10x Barcodes to separately index each cell's transcriptome. Next, barcoded gel beads are mixed with cells, enzyme and partitioning oil (**Gel bead emulsion**, GEM). Single cell GEMs undergo RT to generate 10x Barcoded cDNA and all generated cDNA from individual cells share a common 10x Barcode.

- *Chromium*: this Next GEM technology enables the analysis of individual biological components at scale.

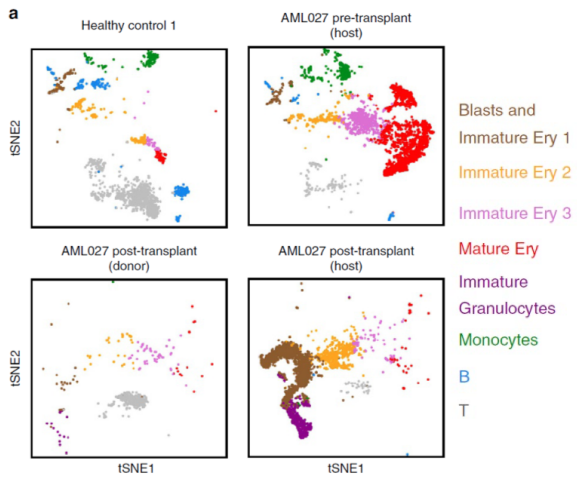

Figure 1.10: *Grace et al, Nat Comms 2017*

- *Visium*: spatial capture technology enabling the analysis of mRNA molecules within their morphological content.

Through Chromium, cell encapsulation of up to 8 samples at a time takes place in 6 minutes, with 50% cell capture efficiency.

Thanks to this technique, it is possible to detect distinct populations in peripheral blood mononuclear cells. PBMC: remove cell without nuclei (erythrocytes and platelets) and with multinuclei (granulocytes) to remain with lymphocytes (T cells, B cells, NK cells) and monocytes.

1.1.10.1 "Genotype" and single-cell expression analysis of transplant BMMCs (Bone Marrow Mononuclear Cells)

Standard treatment in AML involves transplant from a healthy donor. The analysis profiles bone marrow samples from AML patients and healthy patients and compares profiles before and after transplant. The authors built a method based on SNV identification able to distinguish cells from donor and from host. The main idea involved checking whether the transplant was modifying differentiation patterns in bone marrow; indeed profiles are quite different and this confirmed the efficacy of the treatment in restoring a healthy cell population, even though the typical signature of AML persists.

1.1.11 SPLiT-seq (2018)

- labels the cellular origin of RNA through **combinatorial barcoding**
- compatible with fixed cells or nuclei

SPLiT-seq does not require partitioning single cells into individual compartments (droplets, microwells, or wells) but relies on the cells themselves as compartments. The technique starts by passing a suspension of formaldehyde-fixed cells or nuclei through four rounds of combinatorial barcoding in 96 wells plates.

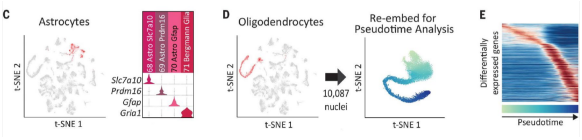

Figure 1.11: *Rosenberg et al, Science 2018*

After sequencing, each transcriptome is assembled by combining reads containing the same four-barcode combination (21M barcode combinations).

1. nuclei extraction
2. sample multiplexing
3. split-pool barcoding
4. clustering and visualization

More than 150.000 nuclei from P2 and P11 mouse brains and spinal cords were profiled in a single experiment employing more than 6 million barcode combinations obtained from 4 rounds. Each cluster was then downsampled to 1000 cells for visualization.

Astrocytes from different spots, while oligodendrocytes have a specific shape → characterize subclusters in the differentiation process. By looking at an isolated closeup, we can distinguish that immature cells and mature cells are ordered in different positions on the tSNE. The coordinate of differentiation is called ***pseudotime***. We are required to manually insert the root of the trajectory to obtain a reasonable trend.

1.1.11.1 Split-pool barcoding

In each split-pool round, fixed cells or nuclei are randomly distributed into wells, and transcripts are labeled with well-specific barcodes. Barcoded RT primers are used in the first round. Second- and third-round barcodes are appended to cDNA through ligation. A fourth barcode is added to cDNA molecules by PCR during sequencing library preparation.

Pros:

- fixation can reduce perturbations to endogenous gene expression during cell handling and makes it possible to store cells for future experiments
- **the use of nuclei bypasses the need to obtain intact single cells**
- may be used to profile single nuclei from formalin-fixed, paraffin-embedded tissue
- the use of the first-round barcode as a sample identifier makes it possible to profile a large number and variety of samples in parallel, thus minimizing batch effects
- efficient use of reagents

Critical aspect in droplet-based approaches: cells often associate with multiple barcodes by multiple beads occurring within the same droplet or heterogeneity of oligonucleotide sequences within a single bead.

1.2 Comparison of scRNA-seq preparation

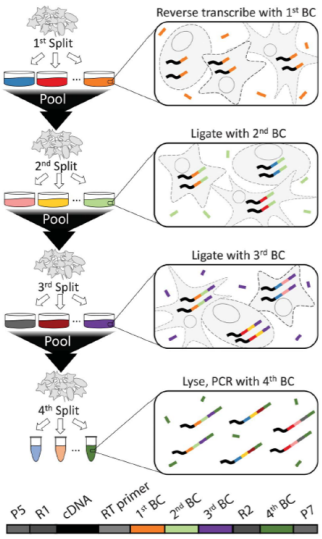

Figure 1.12: *Rosenberg et al, Science 2018*

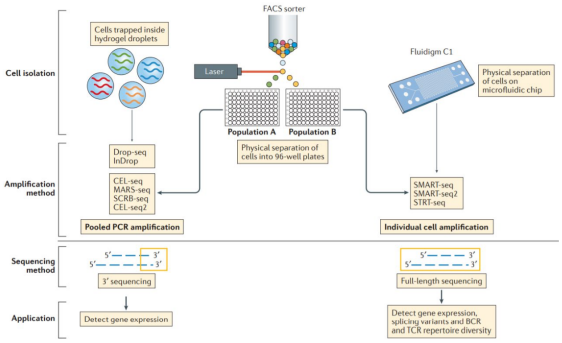

Figure 1.13: *Papalexi et al, Nat Rev Immunol. 2018*

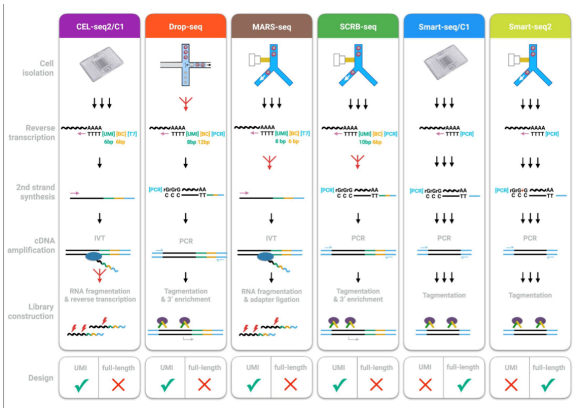

Figure 1.14: *Ziegenhaim et al, Mol Cell 2017*

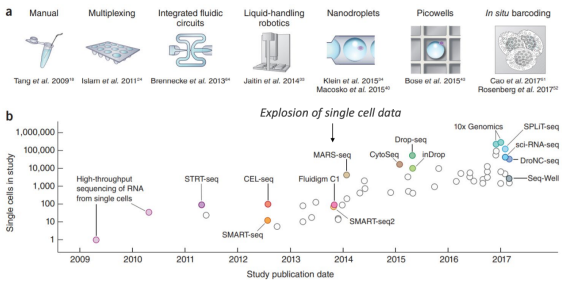

Figure 1.15: Screen Shot 2023-02-21 at 22-40-17.png

Chapter 2

scRNA-seq backstage

2.1 Challenges in single cell workflow

2.1.1 Sources of experimental variability

1. human factor
2. cell suspension preparation
3. cDNA synthesis
4. PCR
5. NGS

2.1.2 Single-cell suspension

RNA degrades very fast, so it is required to quickly transfer separated samples to the sequencing machine. A good quality single-cell suspension should have high viability, no dead or dying cells, cell debris removed and unaltered transcriptional profile.

Fresh samples can be obtained from:

- suspension cell lines e.g. blood

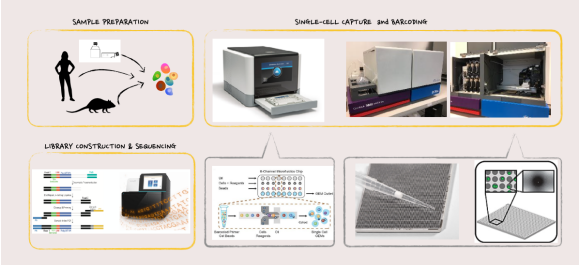

Figure 2.1

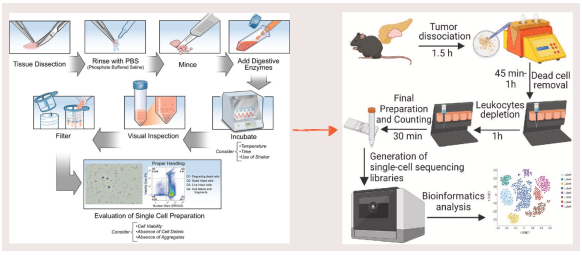

Figure 2.2: STAR Protocols 2, 100989, December 17, 2021

- adherent cell lines → trypsin digestion is usually applied for dissociation, but it is possible to choose among other enzymes e.g. Accutase or TrypLe. It is then necessary to resuspend cells to avoid aggregates. Wide-pore tips are used to avoid pression and stress induction, leading to an overall higher cell survival.
- dissociation of tissues

Cells are different in size, gene expression profile and are characterized by peculiar extracellular matrix composition, which affects dissociation techniques.

2.1.2.1 Single cell suspension from solid tissue

1. tissue dissection
2. rinse with PBS to get rid of unwanted components
3. mechanical dissociation to increase the contact area with cell and enzyme
4. add digestive enzymes (go to Tissue Atlas to choose the best option for the tissue of choice)
5. incubate considering temperature, time and use of shaker
6. remove undigested portions of tissue through a filter
7. evaluate quality of single cell preparation

This protocol is time consuming and only encompasses a small piece of the pipeline; it is possible to rely on commercial solutions, as Miltenyi Biotec automated tissue dissociator with additional dead cell removal kit (through magnetic beads) or debris removal solution.

2.1.3 Single cell sorting

While working with rare cells or genetically modified cells, it is required to apply **FACS** (Fluorescent Activated Cell Sorting), where cells are fluorescently labelled and driven by pressure in a tube where laser excites cells and a detector reads chemical properties. Finally, cells are sorted through deflector plates, included in a droplet and inserted in collection tubes. The overall procedure is long and stress-inducing, hence cells could have changes in gene expression profile (possible solution: freeze a part of the samples for final comparison).

Microfluidic sorters provide a commercial solution to the issue. They work in a sterile environment, are based on disposables, avoid cross contaminations and apply a gentle analysis. Through the *actuator,* cells are mechanically accompanied to the right channel without stress induction. Examples:

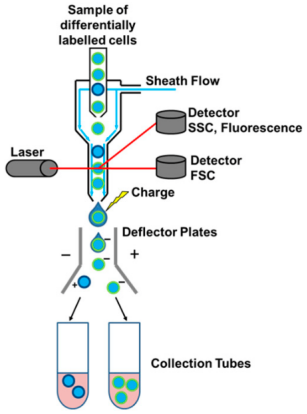

Figure 2.3: G Pfister et al., J Immunol Meth 487 (2020) 112902

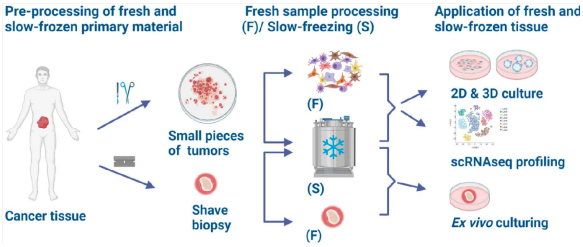

Figure 2.4: G. Restivo et al. Commun Biol 5, 1144 (2022)

- LeviCell, LevitasBio: label free system exploiting the magnetic properties of nanoparticles in cells for sorting. Drawback: small input channels, only a small number of cells is allowed.

2.1.4 Sample collection and preservation

With the advent of molecular pathology, hospitals started to collect fresh and frozen tissue in addition to FFPE-preserved tissue traditionally used in immunohistochemistry. From these sample, it is possible to isolate both nuclei and cells according to the techniques. In order to maintain samples in time, freezing is a feasible solution, as cells can be cultured again after many years.

Keep in mind that different assays require different input materials:

- protein expression: whole cells
- gene expression: whole cell nuclei
- chromatin accessibility: nuclei

Determining your required input can help you determine what collection and preservation method is most suitable for your sample type.

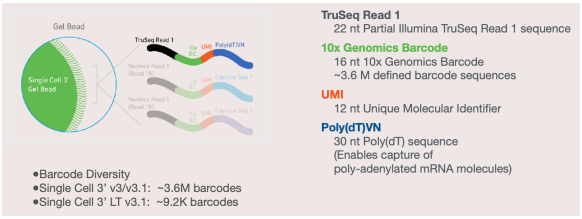

Figure 2.5: 10x Genomics

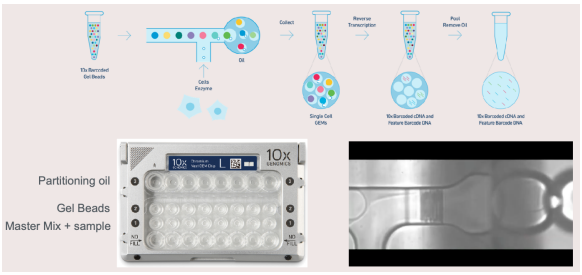

Figure 2.6: 10x Genomics

2.2 10xGenomics

Three main pillar technologies:

1. Chromium Single Cell: NGS readout
2. Visium Spatial: NGS readout
3. Xenium In Situ: microscopy-based readout

In **feature barcode technology** we aim at combining different cells in the same sequencing run through cell multiplexing, cell surface protein or CRISPR screening. Older machines run in manual mode, while the more recent Chromium Connect is automated and able to perform gene expression immune profiling.

2.2.1 Feature barcode technology

In the standard approach, we find 3.6M barcodes, while in low throughput (a bit cheaper) 9.2K barcodes. The capture sequence is directed to antibody or guide RNA for gene activity disruption.

Samples are inserted in a chip with wells.

Bio-Rad vs 10x Genomics patent war: develoment of digital PCR by Quantalife, bought by Bio-Rad.

2.2.2 Single cell gene expression workflow

1. GEM
2. GEM breakage
3. cDNA amplification and library construction

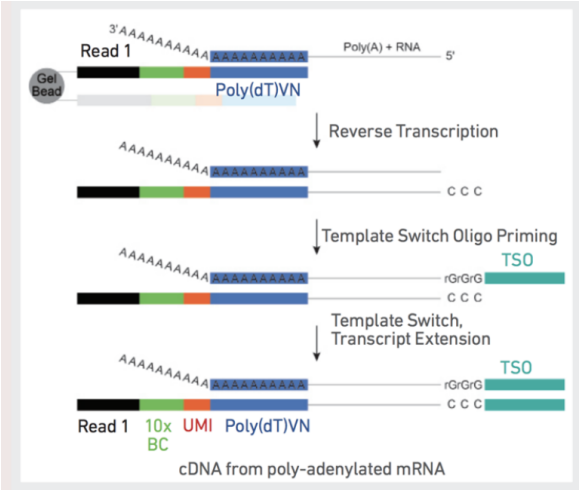

Figure 2.7: 10x Genomics

Figure 2.8: 10x Genomics

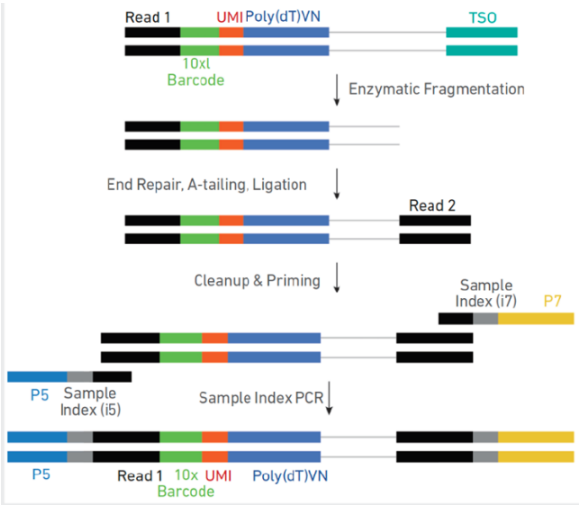

Figure 2.9: 10x Genomics

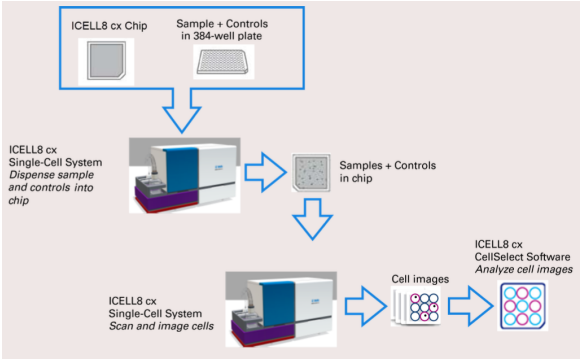

Figure 2.10: iCELL8cx cell scan workflow

2.2.3 Chip-based single cell separator

Chip-based single cell separator → iCELL8 cv (TAKARA) 72x72 wells, check the quality by inspection through a microscope. Very low volumes, we should have no air in the solution. Upper limit for nozzle is higher than 10x, so it is possible to analyze bigger cells e.g. Schwann cells or cardiomyocytes. With this technique, reagents are only dispensed in selected wells.

2.2.4 HIVE scRNAseq Solution

The HIVE collector is composed of a membrane with wells. The technology is based on pore properties and gravity sorting (more gentle approach) by the usage of vortex. It is possible to integrate sample storage in the workflow with freezing. The main advantage is the lack of specialized instrumentation, but we have a huge limit on the number of analyzable cells.

Chapter 3

Analysis and interpretation of single cell sequencing data

scRNA-sequencing is mainly applied for:

- *discrete analysis*: identify new cell types and states within tissues and organs (in physiology and disease) $\rightarrow$ clustering
- *continuous analysis*: understand development paths (e.g., embryonic development) and differentiation decisions (e.g., hematopoiesis) $\rightarrow$ trajectory reconstruction

Example from perspective paper (*Rajewsky et al, Nature 2020)*: cells in our bodies are transitioning in age-related differentiation trajectories. The idea is to investigate altered trajectories due to disease e.g. neurological, cancer, cardiological, infection diseases... in order to gain knowledge on the roots of disorders. Ideally, we should identify the initial causes through early detection and immediately intervene so as to prevent disease development.

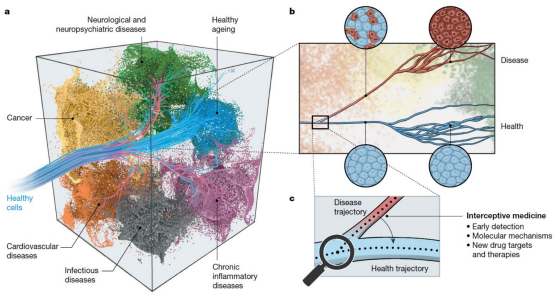

Figure 3.1: Rajewsky et al, Nature 2020

3.1 Pre-processing

3.1.1 From raw reads to count matrices

Reads are composed of:

- cDNA sequences (around 50 bp) corresponding to the original RNA fragments
- technical region: cell barcode to understand the cell of origin (e.g. 10x sequence is 12 nt long) + UMI random sequence to remove possible technical artifacts e.g. PCR duplicates

Through cDNA alignment and barcode ordering we can divide reads by cell of origin and map them to a reference genome. By counting unique UMIs for each gene in each cell we reach the *count matrix*.

Raw sequencing data output depend on the sequencing platform, which is usually Illumina, and are encoded in FASTQ format, which reports a title (`@title`), sequence and a quality score for each base - PHRED quality scores encoded as ASCII characters(ASCII 33-126). The title can contain additional technical information as cell barcodes and UMI, useful for downstream analysis. Unique molecular identifier reduce the amplification noise by allowing (almost) complete de-duplication of sequenced fragments. It is possible to trim a portion of the read or discard the read itself according to quality PHRED scores.

The available alignment methods can be divided into two main categories:

- **splice-aware alignment tools:** map sequences derived from non-contiguous genomic regions, e.g. spliced sequence modules (exons) that are joined together to form spliced RNAs e.g. TopHat, HISAT2, STAR. STARsolo was developed for scRNA-seq, and is able to deal with cell barcodes and UMIs. The alignment output is saved as SAM/BAM file, which specifies where the read aligns and how the nucleotides in the read match up with the nucleotides in the target sequence (genome or transcriptome).
- **pseudo-alignment tools:** the output file provides a set of target sequences that the read is compatible with. Advantage: provide accurate expression estimates very quickly and using little memory e.g. Kallisto, SALMON. Example: kallisto | bustools provides a workflow for association of reads with cell of origin, collapsing of reads according to UMIs and generation of a cell x gene matrix.

Many of the tools developed for bulk RNA-seq released versions suitable for single cell specificity. Transcription factors are usually not detected in single cell, so it is required to perform enrichment to analyze them.

Reads could map to more than one location as the result of homolog genes or transcribed transposable elements; an easy solution to this is to increase read length.

An example of analysis pipeline is CellRanger (10X Genomics), which processes Chromium single-cell data to align reads (with STAR), generate count matrices, and other secondary analyses on Linux systems.

Differently from bulk count data, scRNA matrices report gene counts as rows per cell (columns). The high presence of zeros could be due to true lack of expression or technical lack of detection.

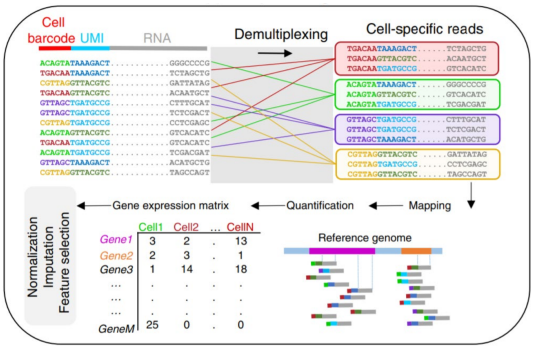

Figure 3.2: *Lafzi A et al., Nature Protocols 2018*

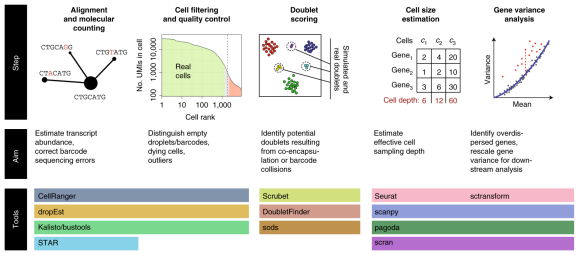

Figure 3.3: Kharchenko, P.V. *Nat Methods* (2021).

3.1.2 Quality control

Cells with a low number of UMIs or reads can be filtered out (unless linked to known biological features). The threshold can be based on a quantile or on a distribution. Another check can be applied on the number of mitochondrial genes; there are around ~ 40 mitochondrial genes and we should not see many mitochondrial genes in single cell sequencing in general, as they are contained in mitochondria and not in cytoplasm. Dying cells usually have high mitochondrial reads due to the breakage of mitochondria and can hence be filtered out.

1. remove low quality cells or empty droplets/dying cells. This is performed by isolating low UMIs (i.e. low number of genes detected) and identifying extensive mitochondrial contamination (dying cells).
2. detect and remove doublets: experimental evaluation of doubles can be applied to measure the probability of doublets (2 or more cells included in a single droplet). In silico simulation e.g. DoubletFinder (R) or Scrublet (Python) is performed as following:
    1. simulate thousands of doublets by adding together two randomly chosen profiles
    2. for each original cell, compute the density of simulated doublets in the surrounding neighborhood, and the density of other original cells in the neighborhood.
    3. return the ratio between the two densities as a "doublet score" for each cell

3.1.3 Normalization

The aim is to remove systematic/technical differences in sequencing coverage between cells. The simplest approach to achieve this is library size normalization, where we divide for the total sum of counts across all genes for each cell. This method is based on the questionable assumption that each cell should have the same number of reads; we could have issues with highly expressed genes. In addition, this method does not normalize well highly expressed genes; by binning genes in different classes according to average expression, it was seen that the higher the n of reads, the higher the number of of genes also after correction. Hence, the same normalization approach for each gene is not the same solution. We could improve by following one of these approaches:

1. **uniform scaling per cell**: identify "size factors" for individual cells, applied to all genes
    - sing bulk RNA-seq methods (TMM in edgeR, DESeq2) (problem with excess of 0s)
    - using spike-ins RNAs (e.g., ERCC) (assumption: same amount of spike-in in each cell) (BASiCS, Vallejos et al, PLOS Comp Biol 2014)
    - pooling cells with similar library sizes and estimate pool-based size factors (scran, Lun et al, Genome Biology 2016)
2. **Gene or gene-group scaling factors**: different scaling factors for groups of genes with high, medium and low abundance
    - SCnorm (Bacher et al, Nature Methods 2017)
    - Sctransform (Hafemeister et al, Genome Biology 2019)

Keep in mind that normalization choices likely affect results! As an example, we can analyze the t-SNE plots of mouse lung epithelial cells (3 embryo stages + adult cells) under various normalization methods. Simple Norm is the library size normalization, BASiCS, GRM, SAMstrt require spike-in RNAs and Linnorm, Scnorm, scran do not require spike-in RNAs.

Good practice entails applying different methods and choose a consensus, as normalization deeply impacts downstream analysis.

3.2 Analysis

3.2.1 Dimensionality reduction

Problem: a typical scRNA-seq data table consists of ~20,000 genes as features and 10,000 to up to a million cells as observations. We can reduce the number of features to reduce data complexity and ease the visualization of cells. This approach is based on the notion that not all genes are important to classify cells (*gene/feature selection*) and several genes are correlated in expression (*redundancy*).

Gene (feature) selection:

- select genes that contain useful information about the biology of the system
- remove genes that contain random noise or do not vary among cells
- reduce the size of the data, improve computational efficiency

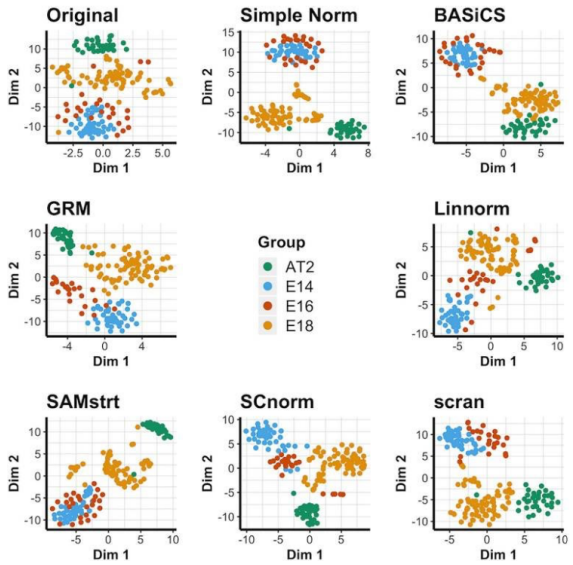

Figure 3.4: *Lytal N, Ran D, An L. Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. Front Genet. 2020*

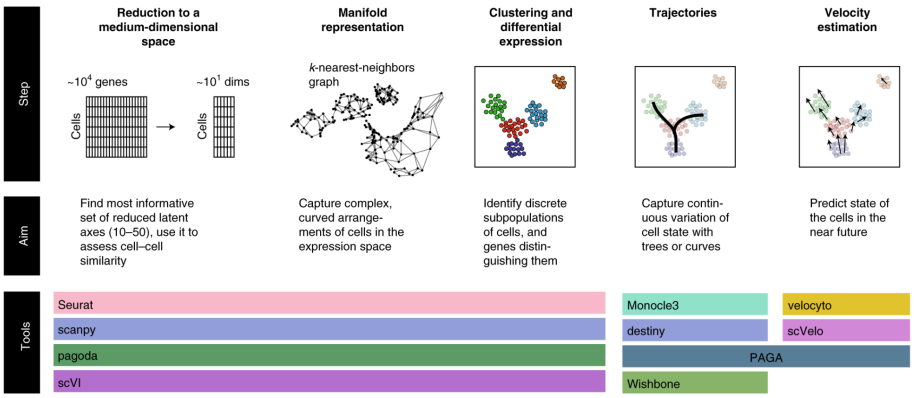

Figure 3.5: Kharchenko, P.V. *Nat Methods* (2021)

The simplest approach is to select the most variable genes based on their expression across the population of cells. However, the underlying assumption is that biological variability is higher than technical.

Dimensionality reduction:

- reduce the number of separate dimensions in the data
- help downstream analyses (clustering, trajectory inference . . .)
- help the visualization of data in 2D plots

Multiple methodologies are available, with different advantages and limitations. In Principal Component Analysis the expression of each gene is a dimension and the goal is to center data and express the variation associated with each dimension, such that cells are represented on the line that captures more variation.

- PCA: Principal Component Analysis (linear)
  - pros: highly interpretable and computationally efficient
  - cons: inappropriate for scRNA-seq visualization (non-linear data, excess of 0s).

PCA is not the optimal way to visualize single cell data, as groups tend to be more mixed and the overall result is less interpretable, due to the high presence of zeroes in the matrix. To solve the issue, two main methods were developed for dimensionality reduction visualization:

- **t-SNE**: t-Stochastic Neighbourhood Embedding (non-linear, graph-based)
  - pros: able to retain the local structures in low dimensions
  - cons: stochastic method, long time to run with high n, global structure of data is not preserved

The main issue of t-SNE is that it does not scale well with data (works best with 2 or 3 dimensions which should be established beforehand) and cannot maintain well the structure of the dataset (randomness).

- **UMAP**: Uniform Manifold Approximation and Projection (non-linear, graph- based)
  - pros: computationally efficient, better preservation of global structure
  - cons: low interpretability, stochastic method

From 2018, the gold standard is UMAP, as it manages to preserve the global structure and is computationally efficient (compromise among PCA and t-SNE). Each run could still give different results due to stochasticity, but due to the fact that the initial steps are non-random there is more stability with respect to t-SNE.

3.2.2 Clustering

For characterizing cell states it is required to organize cells according to their identity, which could be shaped by environmental stimuli, cell development, cell cycle or spatial context. Through single cell, it is possible to detect such changes through discrete analysis or continuous approaches (developmental trajectory or circle path).

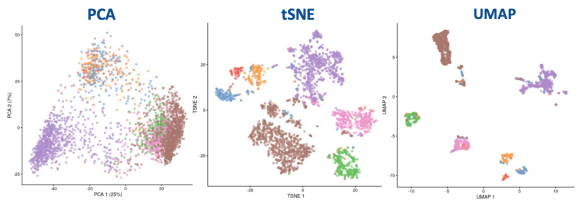

Figure 3.6

Clustering approaches developed for data sciences can be applied to single cell data e.g. hierarchical clustering, K-means or graph-based approaches (most used as they are particularly suited for high dimensional data). The aim is to identify highly interconnected communities within the population. Algorithms identify the edges that, when removed, result in a higher difference among groups. It is possible to weight edges according to cell similarity e.g.Euclidean distance or according to their neighbourhood if they share neighbouring points.

In **graph-based approaches**, each node is a cell connected to its nearest neighbors in the high-dimensional space. Edges are weighted based on the similarity between the 2 cells. The strategy focuses on identifying "communities" of cells that are more connected to cells in the same community than they are to cells of different communities. A popular approach is the Louvain method for community detection.

Once clusters are identified, the aim is to characterize clusters according to differential expression. Issue: heterogeneity in gene expression distribution makes it hard to use parametric methods $\rightarrow$ non-parametric choices are a solution. In order to identify genes with different expression among clusters of cells, hence to distinguish cluster markers, we can apply:

- non-parametric tests (Wilcoxon Rank Sum test) $\rightarrow$ problematic with tied values (0s)
- methods inherited from bulk RNA-seq (edgeR, DESeq2, based on negative binomial models)

It is possible to transform negative binomial to zero inflated negative binomial to account for the high presence of zeroes in the matrix.

- methods developed for single cell (dealing with the excess of 0s)

In Seurat package we have the FindAllMarkers function that applies Wilcoxon by default, but it is possible to use other methods e.g. ROC curve with marker genes.

3.2.2.1 t-test differential expression

With the t-test, we must apply the assumption of Normal distribution i.e. we should see a low variability in differentially expressed genes. The null hypothesis states that the mean of the control and the treatment is equal, in other words, the distribution of the mean should be normal. We compute the t-test as:

$$t = \frac{\bar{X_T} - \bar{X_C}}{\sqrt{\frac{\text{var}_T}{\text{n}_T} + \frac{\text{var}_C}{\text{n}_C}}} \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

The p-value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (e.g. t-statistic) would be the same as or more extreme than the actual observed results. It can be seen graphically as the sum of tail areas in the two-tailed test.

3.2.2.2 Non-parametric tests

In non-parametric tests, expression values are ranked from lowest to highest without making assumptions on the distribution of the data. The sum of the ranks for each group is used to check the difference among clusters, which is associated with a pvalue; if the rank sums are very close, the pvalue will not be significant.

Wilcoxon Rank-Sum test aka **Mann-Whitney U test** convert observed expression values to ranks, and test whether the distribution of ranks for one group are significantly different from the distribution of ranks for the other group. In this case there is no assumption on the distribution of data, but it could be problematic with tied values (0s).

3.2.2.3 Multiple test correction methods

High-throughput results usually have multiple comparison problem, due to expression variation (thousands of genes) and pathway enrichment (hundreds of pathways). The obtained raw pvalues can be corrected to minimize false positives. For instance, with a pvalue threshold of 0.05 in a setting of 10k genes, we expect that 500 genes are false positive, generating a very high false discovery rate. For reducing FP in selection, it is possible to apply multiple test correction methods. For instance, with Bonferroni correction we can increase the original p-values to reduce false positives - do not swap the order of the original p-values.

- **Bonferroni correction**: multiply the original p-value for the number of tests performed (M). It controls the probability to have at least 1 false positive result (controlling the Family Wise Error Rate FWER). It is a very stringent and conservative correction, especially with large numbers of tests $\rightarrow$ possibly leads to false negatives.

  adjusted P value = (original P value) $\times$ (number of tests)

- **Benjamini Hochberg correction**: most popular multiple test correction method, introduced in 1995. It was designed to control the False Discovery Rate (FDR), the expected proportion of "false positives". It is a stepwise method, as it requires to sort all original p-values in increasing order and determine the rank (1,2,3, ...)

  adjusted P value = (original P value) $\times$ $(\frac{\text{number of tests}}{\text{P value rank}})$

3.2.3 Cell type annotation

For annotating clusters with known cell types we can proceed with:

- manual approach, based on personal knowledge
- automatic approach, based on:
  - databases of marker genes e.g. Panglaodb or CellMarker
  - cell type expression data e.g. bulk RNA-seq
  - labeled scRNA-seq datasets

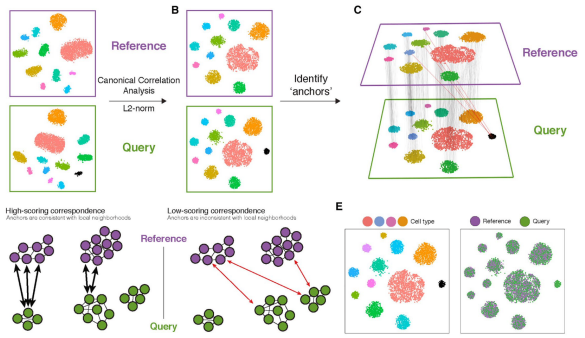

Figure 3.7: *Stuart et al, Comprehensive Integration of Single-Cell Data, Cell 2019*

3.2.3.1 Annotation based on marker genes (scType)

Score based on the scaled expression of marker genes (z-scores)

1. Calculation of marker specificity scores (weights)
2. Use of positive&negative markers

3.2.3.2 Reference based annotation (SingleR)

SingleR is a protocol for cell type annotation by reference to transcriptomes of pure cell types(score based on Spearman correlation)

3.2.4 Integration of single cell datasets

The goal is to remove batch effects and confounders. Through the Seurat package we can perform **Canonical Correlation Analysis** (CCA), which is able to capture correlated sources of variation between two datasets.

If we are trying to classify cells according to gene expression, it is required to use strictly gene expression markers. We do not want data to be clustered according to platform or batch; we can apply different integration approaches to solve the issue.

Anchor based approaches: reference and query dataset should be integrated. We identify among the two datasets that are assumed to be in the same state (similar to each other in a reduced space), which are then used as anchors. The other cells will be corrected according to a measure expressing how different the cells are from anchors and from the dataset they belong to. This method works well while integrating similar system, as most cells are assumed to align.

Canonical Correlation Analysis is a linear method for dimensionality reduction which can be compared to PCA, as it is also a linear combination. Differently from PCA, the components will maximize the correlation between two dataset. We can visualize the output of the two methods on the same dataset in Figure 3.7 .

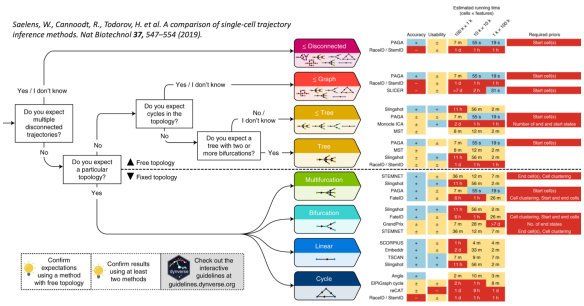

Figure 3.8

3.3 Trajectory analysis

Instead of dividing cells into discrete clusters, we can place cells along a continuous path representing the evolution of a process (e.g. differentiation). The main assumption entails capturing all the snapshots of the process as a continuous spectrum of states (with intermediates). The meaningfulness of the reconstructed trajectory must be evaluated carefully, as this method will always lead to a result i.e. any dataset can be forced into a trajectory.

One common approach is to build a minimum spanning tree over cells, minimizing the total sum of the branches, or over cell clusters while working with a high number of cells. In most methods, the choice of the root is performed manually (usually based on markers and known biological information). Next, each cell's *pseudotime* is calculated as its distance along the tree to the root.

3.3.0.1 Monocle (2014)

Monocle R package provides tools for analyzing single-cell expression experiments. In 2014, Cole Trapnell introduced the strategy of ordering single cells in pseudotime, placing them along a trajectory corresponding to a biological process such as cell differentiation.

Dataset of differentiation myoblasts (SMART-seq) in pseudotime. The trajectory captures the transition from proliferating cells to differentiated cells through interstitial mesenchymal cells. The method has a high computational cost, so cannot be applied to huge datasets.

3.3.0.2 Monocle 2 (2017)

Monocle 2 is a method for finding trajectories through **reversed graph embedding**. Each cell is represented as a point in high dimensional space (x), where each dimension corresponds to the expression level of an ordering gene. Data are projected onto a lower-dimensional space by dimension-reduction methods such as PCA.

Monocle 2 constructs a spanning tree on a set of *centroids* (diamonds) chosen automatically using k-means clustering. Cells are shifted toward the nearest tree vertex, vertex positions are updated to 'fit' cells, a new spanning tree is learned, and the process is iterated until the tree and cells converge. The user then selects a tip as the 'root', and each cell's pseudotime is calculated as its *geodesic distance* along the tree to the root.

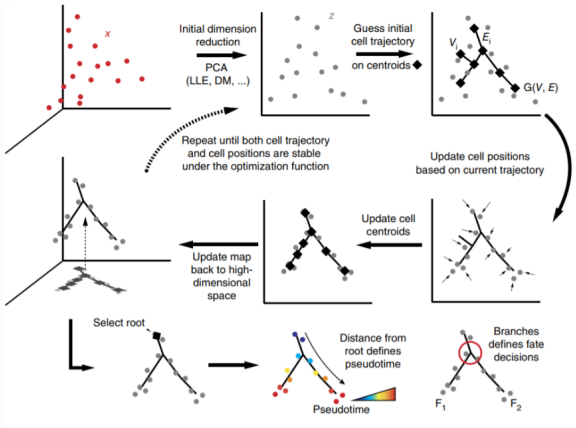

Figure 3.9: *Qiu X et al., Nature Methods 2017*

3.3.0.3 Comparing dimensionality reduction and TA choices

Datasets with increasing number of cells are analyzed with different approaches e.g. PCA, UMAP, Monocle2,... We can identify which methods scale well with increasing numbers of cells.

3.3.0.4 Velocyto (2018)

RNA velocity: the time derivative of the gene expression state ($\frac{ds}{dt} = u - \gamma s$) - based on the proportion of unspliced vs spliced reads in single-cell RNAseq - predicts the future state of individual cells on a timescale of hours. In other words, the balance between unspliced and spliced mRNAs is predictive of cellular state progression, where:

- $u > \gamma s$ induction
- $u = \gamma s$ steady state
- $u < \gamma s$ repression

A read partially mapping an intron and exon is assumed to be *unspliced*. The whole method is based on the concept that the ratio of spliced and unspliced reads can aid us in making hypotheses on the future development of the cell. If the amount of unspliced reads is higher than spliced we expect that we will reach a higher expression level in the cell → transcription induction. Viceversa, a higher amount of spliced reads characterizes repression of transcription.

RNA velocity is the time derivative of gene expression state based on the proportion of unspliced vs spliced reads in scRNA-seq.

Limitation: same constants for all genes, incorrect assumption. In addition, timescale could greatly vary in different processes e.g. hematopoiesis is not recapitulated well.

→ scVelo

! bias with 3'end approaches, not many introns → bias on splicing events capturing

3.3.0.5 scVelo (2020)

The aim is to infer gene-specific rates of transcription, splicing and degradation. scVelo provides a dynamical model, as it generalizes RNA velocity to systems with transient cell states, common in development and in response to perturbations.

3.4 Main pitfalls/challenges in current scRNA-Seq

- Incomplete detection ('drop- out') of lowly expressed genes. Drop- out can blur fine-scale distinctions. Tentative solution: imputation methods.
- Amplification bias (tentative solution: use of UMI)
- Stochastic gene expression
- Background noise
- Bias due to cell size (smaller cells are more efficiently incorporated in droplets than large cells) and other factors (cell cycle)
- Analysis restricted to polyA transcripts, no method for total RNA sequencing in single cells (solutions will be available soon)
- dealing with high dimensionality
  - the number of cells included in average scRNA-seq is growing more than Moore's law. In the plot we can analyze the number of CPUs in the evolution of processors. The angular coefficient of the growth is higher than the increase in computational availability of processors $\rightarrow$ issue
- absence of standard methodologies
- high variability of results, depending on methodologies and parameters

A lot of methods are based on graph-approaches. Example: changing resolution deeply affects the results e.g. granularity of clustering.

Chapter 4

Single cell multimodal omics

All bulk omics are reaching a single-cell resolution, currently with variable degrees of success. Some examples of trans-omics encompass genome, transcriptome, proteome and metabolome. For instance, single-cell applies both to *epigenetic* (DNA and histone modifications) and *epitranscriptomics* (the ensemble of all RNA modifications).

Single cell multimodal omics was classified as method of the year 2019 by Nature.

4.0.1 Single-cell multiomic profiling

Linking measurements from different omics layers has the potential to reveal regulatory and functional mechanisms underlying cell behaviour in healthy development and disease. In Figure 4.1, we can observe the possible connections among omics techniques and their interplay in determining cell phenotype and functional potential as well as molecular cell identity.

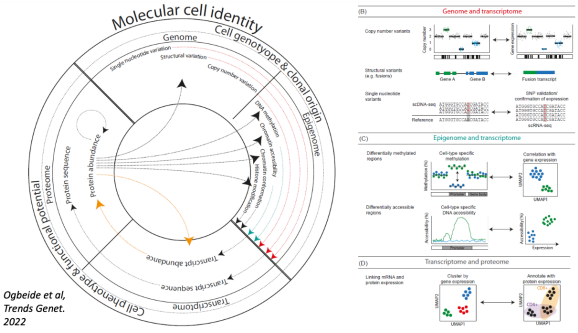

Figure 4.1

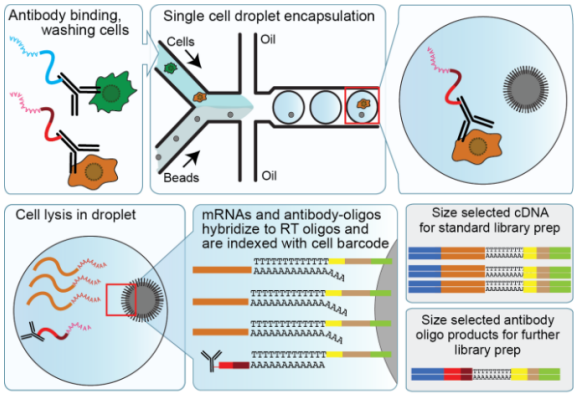

Figure 4.2: cite-seq.com

4.1 Transcriptome & proteome

Proteins determine much of cell behaviour and mRNA expression levels are a weak proxy of protein expression (post-transcriptional and translation regulation). In addition, there is no method for protein sequence amplification (measurements are dependent on antibody-based detection or mass spectrometry).

4.1.1 CITE-seq (Cellular Indexing of Transcriptomes and Epitopes) (2017)

CITE-seq uses **DNA-barcoded antibodies** to convert detection of proteins into a quantitative, sequenceable readout. Antibody-bound oligos act as synthetic transcripts that are captured during most large-scale oligo dT-based scRNA-seq library preparation protocols (e.g. 10x, Drop-seq). Cells labelled with panels of antibodies, each tagged with a specific ADT (antibody derived tag), are **captured in parallel with the mRNA from the same cell** following lysis.

Pros: immuno-phenotyping of cells allows a potentially limitless number of markers unbiased transcriptome analysis.

Cons: specific antibodies necessary, for surface protein markers

4.1.2 Cell Hashing with barcoded antibodies (2018)

Oligo-tagged antibodies against **ubiquitously expressed** surface proteins. Label cells from distinct samples, which can be **pooled** (cost reduction). **Detection of doublets** between samples (or within a sample, if used without pooling).

Other techniques based on barcoded antibodies:

- 2017: **REAP-seq** (RNA Expression and Protein Sequencing assay). Peterson et al. Nature biotechnology (library of 82 barcoded antibodies)
- 2017: **Abseq** (Ultrahigh- throughput single cell protein profiling with droplet microfluidic barcoding). Shahi et al, Scientific reports (example with 2 antibodies, B cells vs T cells)

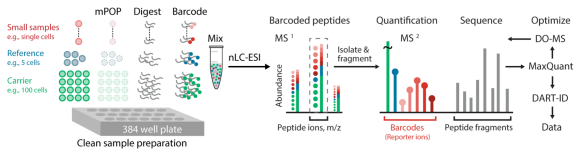

Figure 4.3: *Specht et al, Genome Biology 2021*

4.1.3 Single-cell mass-spectrometry approaches (proteome only)

Single cells are isolated in individual wells and lysed. Proteins are digested to peptides. Peptides from each single cell are **covalently labeled** (barcoded) with isobaric tandem-mass-tags (TMT).

SCoPE2 was applied to monocytes differentiated into macrophages. Quantified over 3042 proteins in 1490 single monocytes and macrophages in ten days of instrument time.

4.2 Transcriptome & genome

Somatic variation is the genomic diversification between cells, deviating from the "original" genome of the zygote. Somatic variation can be:

- *programmed*: during the maturation of lymphocytes, programmed rearrangements in B and T cell receptors are carried out e.g. rearrangements of the V(D)J regions in B and T lymphocytes to produce diverse and specific antibodies and T cell receptors.
- *spontaneous*: accumulations of SNVs or CNVs of chromosome regions during development and aging

Potentially pathogenic when variants confer a competitive advantage to the cell, leading to clonal expansion and the formation of malignant or cancerous clones.

4.2.1 DR-seq (2015)

DR-seq (gDNA-mRNA sequencing) is based on sequencing of genomic DNA and messenger RNA from the same cell. There are two different strategies for amplifying:

- PCR for gDNA amplification
- mRNA specific second-strand synthesis (polyA selection, only the 3' end is sequenced)

Coverage depth in a single cell should be 2, if we see more is produced by amplification - differently from RNA-seq. There is a positive correlation between copy number levels and average expression of all genes located in the region of interest e.g. chromosome 8.

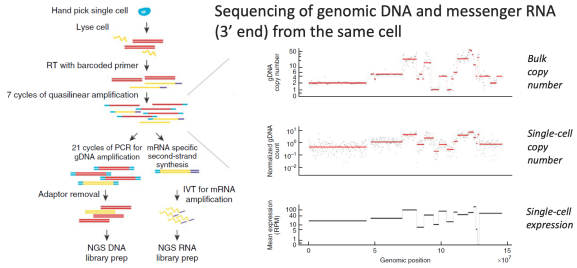

Figure 4.4

4.2.2 G&T-seq

Parallel sequencing of single-cell genomes and transcriptomes occurs through physical separation of genomic DNA and poly(A)+ mRNA after cell lysis, followed by separate amplification, library preparation and sequencing. Full-length mRNA sequencing can be obtained following Smart-seq2 protocol.

Workflow:

1. isolate cells
2. lysis
3. using oligo-dT probes isolate mRNA content and divide it in different plates with respect to DNA
    1. RNA: whole transcriptome amplifications and full length sequencing
    2. DNA: whole genome amplification and detection

The method was applied to differentiated neurons derived form trisomy 21 induced pluripotent stem cells and control disomy 21 iPSCs. By visual inspection, we can clearly see a gain in chromosome 21 with respect to control, even though some error is identified. The positive result is that chromosome 21 genes are more expressed with respect to control. Transcriptomics variation was also observed in other autosomes, which could be due to both biological and technical variation.

In theory these approaches can be used to detect fusions, while SNVs are harder to be found at low coverage. There could be location biases, but this is common for different amplification techniques.

4.2.3 Physical separation of the nucleus and cytoplasm

In this case there is no separation of DNA and RNA, but rather of the nucleus (DNA-seq) and cytoplasm (RNA-seq). Cell separation is performed by sorting, but the core of the technique a gentle lysis and separation in two plates. Full length mRNA sequencing is comparable with SMART-seq2, while for DNA tagmentation is performed.

4.2.4 Genotyping of Transcriptomes (GoT) (2019)

GoT integrates *genotyping* (detection of somatic mutations gene of interest) with single cell RNA sequencing. The main advantage of this technique is the identification of malignant cells associated with

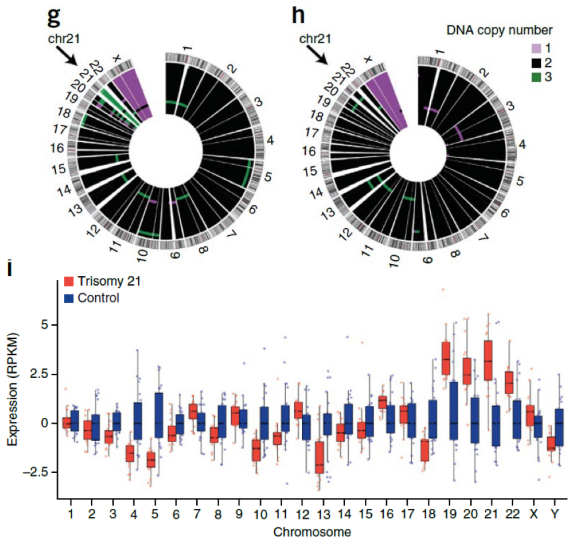

Figure 4.5: *Macaulay et al, Nature Methods 2015*

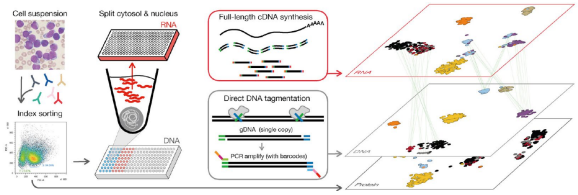

Figure 4.6: *Zachariadis et al, Molecular Cell 2020*

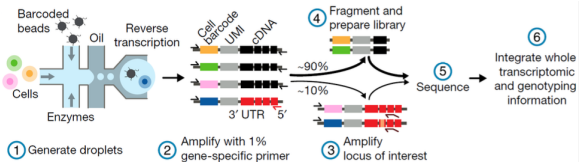

Figure 4.7: *Nam et al, Nature 2019*

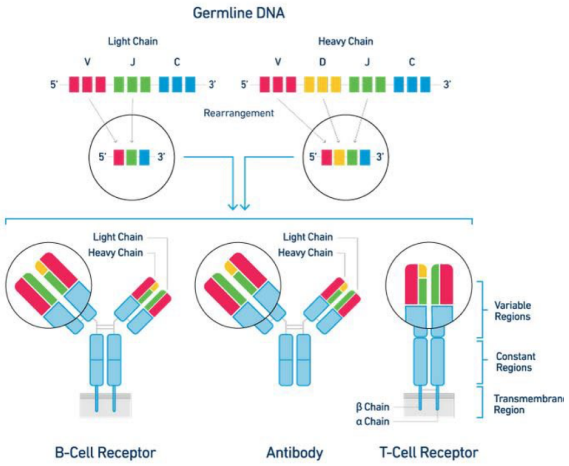

Figure 4.8: *https://www.10xgenomics.com*

a specific mutation(challenging in the absence of specific surface markers). Additionally, this method works with droplet-based methods, greatly enhancing the throughput.

We can connect the presence of a mutation with markers and cell abundance. The real advantage is the amplification of the locus of interest reducing the dropout rate.

4.2.5 Adaptive immune response

Both lymphocytes express B- and T-cell receptors. The soluble part in both receptor is an antibody able to recognize antigens. The transmembrane domain is common, while the top part is a variable region. Looking at gDNA of each cell, we have a V(D)J region where random stochastic rearrangement occur, producing variability in the amino acid region of the receptor. Once one of these cells meets a pathogen, we have the activation of the response and increase in antibody targeting a specific antigen. Through the sequence of the VDJ region, we can identify the abundance of the pathogen e.g. COVID.

4.2.5.1 Immune profiling protocol

The most used platform (10x) has an **immune profiling protocol** implementing this technique through transcriptome 5' end sequencing for paired T- and B-cell receptors. After barcoding, libraries are split for the targeted analysis of B/T cell receptor transcripts, (based on primers specific for the sequence of the constant region) and the analysis of all other transcripts (standard 10x 5' sequencing).

Sequencing of TCR alpha and beta chain allows cell grouping according to shared sequences. It is possible to predict the pathogen to which the receptor will react e.g. the clonal expansion is recognizing a sequence of a particular peptide of cytomegalovirus.

4.3 Transcriptome & epigenome

The compactness of DNA is regulated by protein modifications that act on chromatin accessibility and therefore have an effect on gene transcription. Epigenetic modifications contribute to cellular heterogeneity and regulation of gene expression during development, lineage determination and response to dynamic stimuli. In single cell epigenomics we can study:

- DNA methylation
- chromatin accessibilty

4.3.1 DNA methylation

DNA methylation was the first discovered epigenetic mark, which consists on an addition of a methyl group on cytosine residues of the dinucleotide **CpG.** Methylation is implicated in repression of transcriptional activity. **Bisulphite sequencing** is a technique based on treating DNA with sodium bisulphite, which allows the conversion of cytosines to uraciles through alkylation in the case of unmethylated cytosine - while methylated cytosine will remain unchanged. By analyzing mismatches with a reference genome sequence, we can distinguish methylated from unmethylated cytosines.

4.3.1.1 scM&T-seq (2016)

Single-cell methylome & transcriptome is a derivation of G&T-seq, based on the physical separation of DNA and RNA:

- Smart-seq2 on mRNA
- Bisulphite sequencing on DNA

The technique was applied to mouse embryonic stem cells in the first publication of the technique. Several single cell genome and transcriptome methods can be extended to study the methylome by applying bisulphite sequencing.

4.3.2 Chromatin accessibility

4.3.2.1 ATAC-seq (2013)

ATAC-seq (Assay for Transposable Accessible Chromatin) is designed to identify open chromatin regions in the genome by high-throughput sequencing. The technique is based on the activity of a key enzyme, **Tn5 Transposase**, which performs a "cut and paste" procedure and can be engineered into a hyperactive mutant form. Tn5 transposases preferentially insert into open chromatin sites, cut and add two sequencing primers (t*agmentation*). The obtained fragments will be nucleosome free or nucleosome containing. The technique is very sensitive i.e. works with small amounts of material and faster than alternative methods.

Key features to identify a successful ATAC-seq experiment:

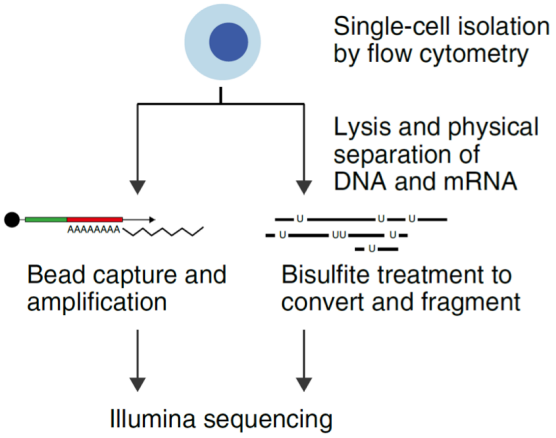

Figure 4.9: *Angermueller et al, Nature Methods 2016*

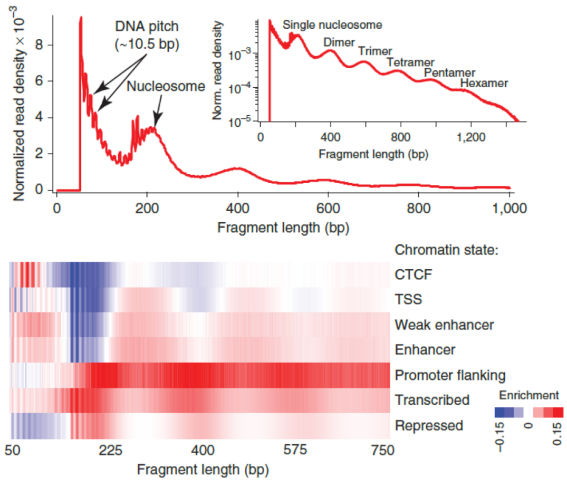

Figure 4.10: *Buenrostro et al, Nature Methods 2013*

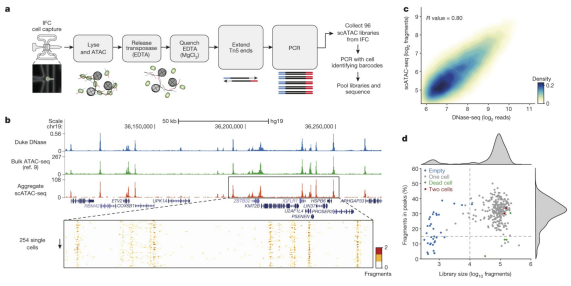

Figure 4.11: *Buenrostro et al, Nature 2015*

- fragment size peaks should follow a clear periodicity of ~200 bp length, corresponding to multiples of nucleosome-wrapped sequence size
- at the genome-wide level, fragments will be enriched in open chromatin regions:
  - promoter flanking regions
  - transcribed genes

4.3.2.2 Single cell ATAC-seq (2015)

ATAC-seq was adapted to single cell resolution with a fluidigm device (not yet droplet-based, lower throughput). We can analyze a snapshot of a genome portion comparing fragment peaks generated from bulk ATAC-seq to aggregate single cell ATAC-seq peaks (obtained from 254 cells) and verify that there is a high correlation among peaks, which are located close to the TSS. It is possible to map fragments on a single cell profile, which will have a range from 0 to 2 fragments per position. The protocol can be adapted to any single cell approach e.g. droplet-based or combinatorial indexing. In 2018, 10x developed 10x single-cell ATAC-seq with a similar workflow; the main difference is the usage of nuclei instead of cells and the capture procedure without UMI (directly designed for transposase adapters).

"Why always chromosome 19?"

4.3.2.3 Peak calling

In order to reduce the dimensions, one of the first steps in the analysis is peak calling, which can be performed by Cell Ranger on 10x data. The number of transpositions events is evaluated in pseudo-bulk, followed by a smoothing procedure and signal threshold evaluation to discriminate actual signal from noise e.g. zero negative binomial model. The outputs of this step are:

1. indexed fragment file collecting all the reads, even if they do not map to a peak, which can be used for quality control.
2. large sparse matrix, where each row is a genomic range corresponding to a peak and the column the number of fragments per cell

The main challenges in comparison to scRNA are:

- more sparse data

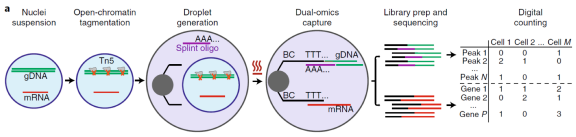

Figure 4.12: Chen et al, Nature Biotechnology 2019

- near-binary data (open vs closed, we are not really interested on how much)
- non-fixed feature set (variable features according to number of peaks per cell)
- order of magnitude more features

4.3.2.4 Enrichment

After peak calling, it is possible to perform DNA sequence enrichment on motif to identify transcription factor binding sites. Regulatory relationship among enhancers and promoters can be studied through co-accessibility networks and gene variants can be studied through genetic variant enrichment.

While working with Seurat, the best approach is to apply **Signac**, a Seurat extension for the analysis of ATAC-seq data.

4.3.3 SNARE-seq (2019)

Single-Nucleus chromatin Accessibility and mRNA Expression is a high-throughput sequencing of the transcriptome and chromatin accessibility in the same nucleus. In this setting, there is no need for probabilistic mapping of single-cell clusters from separate analyses. The technique was applied to droplet-based method, then adopted by 10x (Multiome profiling).

Capturing beads contain two adapters, one for ployA selection for mRNA and another complementary to Tn5. The advantage is that probes have the same barcode, allowing easy mapping of both libraries and digital counting matrix with common columns.

In the original publication the method was applied to ~5000 nuclei from 5 mice collected from neonatal cerebral cortex. The clusters have been annotated to known groups of neuronal cells based on gene expression (reference) and then mapped to the UMAP obtained from chromatin data analysis. The obtained result is more or less consistent with the expectation and allows the study of biological implications of chromatin regulation on gene expression. Alternatively, it is possible to use both approaches simultaneously to identify clusters.

10x method for simultaneous detection of gene expression and chromatin state from the same cell can aid in identifying activation of gene expression even though expression data does not show relevant patterns. In addition, putative regulatory elements directly linked to a gene of interest can be inferred thanks to the correlation of gene expression with the presence of accessible peaks in clusters.

4.4 Transcriptome & "CRISPR" perturbations

CRISPR technology can be applied to inactivate regions of interest and study perturbation effects (also called *reverse genetics* approach).

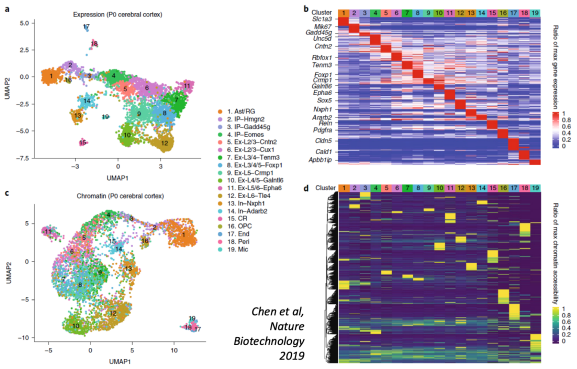

Figure 4.13: Chen et al, Nature Biotechnology 2019

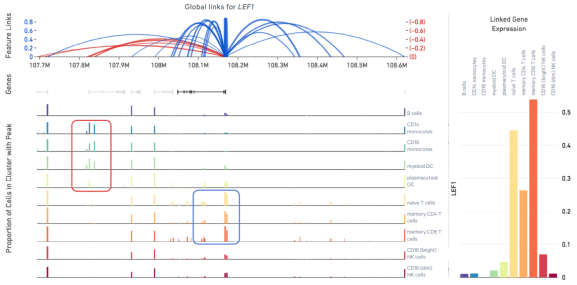

Figure 4.14: *https://www.10xgenomics.com*

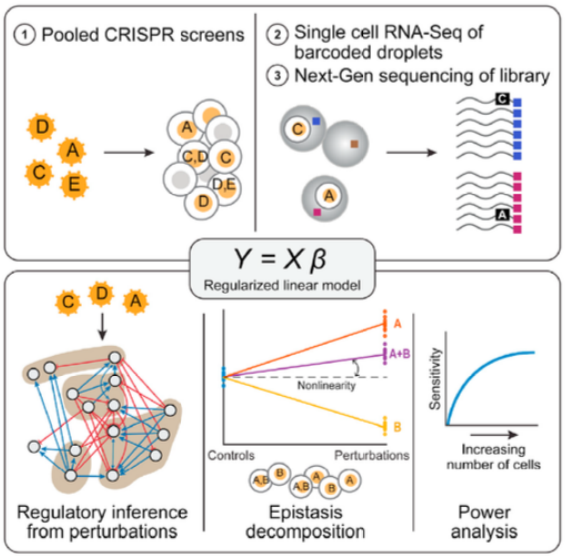

Figure 4.15: Dixit et al, Cell 2016

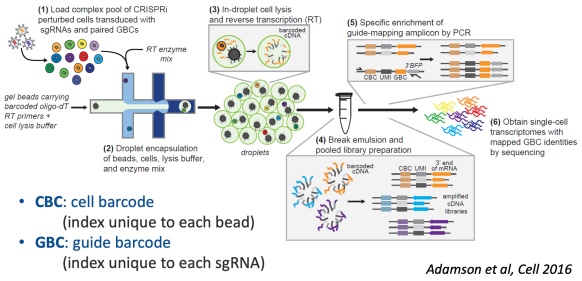

Figure 4.16

4.4.1 Perturb-seq (2016)

Perturb-seq combines single-cell RNA sequencing with CRISPR- based perturbations (inactivation of genes). The main aim is to map the transcriptional effects of genetic perturbations and identify regulatory circuits. By applying single cell, we can detect the expression level of each cell and identify which kind of CRISPR guide was present in each cell i.e. which gene was inactivated.

The mRNAs are captured with a cell barcode (**CBC**) and matched to sgRNAs and paired guide barcode (**GBC**).

4.4.2 CRISP-seq (2016)

CRISP-seq combines single-cell RNAseq with CRISPR-based perturbations (editing/inactivation of genes). It allows to discover interactions and redundancies between developmental and signaling- dependent factors. The technique exploits mathematical modeling and immune stimulation factors.

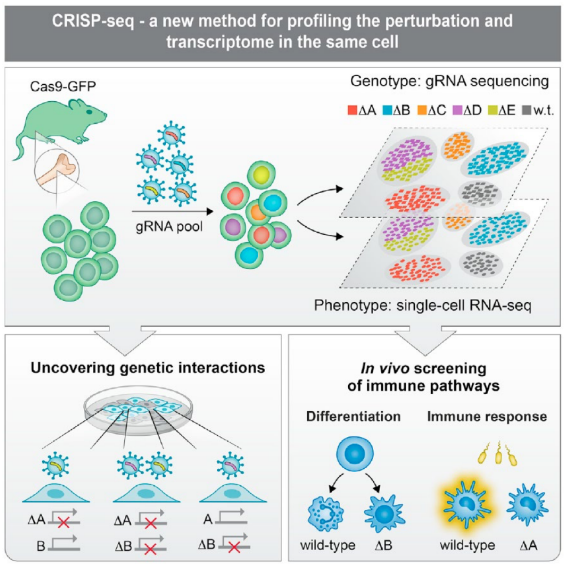

Figure 4.17: *Jaitin et al, Cell 2016*

4.4.2.1 10x CRISPR screen

In addition, 10x now provides single-cell CRISPR screens, which provide information on:

1. guide RNAs
2. whole transcriptome profiling
3. cell surface protein expresssion
4. isolation of paired immune receptor sequences

4.5 Challenges and opportunities in single-cell multimodal omics

Chromatin accessibility layer is more sparse than the genomic one, so the combination of multiple layers can aid in achieving a better separation. This instead in not an issue with tagged antibodies, but on the other hand we are working with restricted information and mainly on surface proteins. Epigenetic information on histone marks and TF binding, even though not in the same cell, can be combined to study regulatory networks.

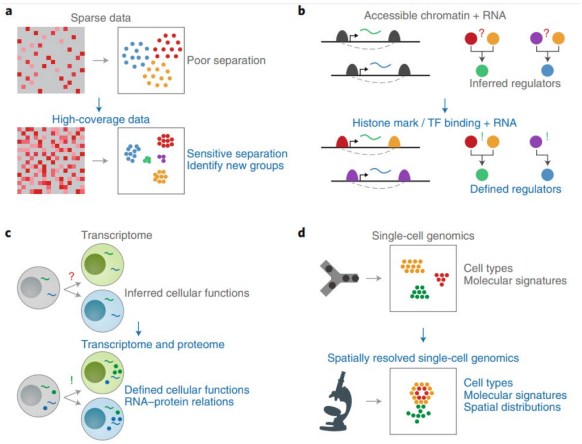

Figure 4.18: *Zhu C et al. Nat Methods. 2020*

Chapter 5

Spatial omics

5.1 Spatial Transcriptomics

A range of methods for **quantifying RNAs** in **specific locations** on histological sections *"retaining information on spatial context"*. Important criteria:

- *How many RNAs?* $\rightarrow$ throughput
- *At which resolution?* $\rightarrow$ the resolution is the unit of space that can be discriminated in the image. We cannot have a grid perfectly adapting to single cells, so we will observe pieces of different cells rather than single cells.

If we look at the evolution of main ST platforms, we observe a correspondence with single cell techniques with a 4-5 years gap.

The main ST approaches can be divided into:

1. *imaging based methods*: read out by microscopy e.g. multiplexed FISH
   - Pros: high resolution, depending on the microscope
   - Cons: low coverage or low number of RNAs, need to design target probes (not possible to have more than 10 fluorophores)
2. *sequencing based methods*: spatial barcoding, read out by sequencing - more similar to single cell approaches. The cells are placed on a plate covered with adapters (UMI + polyA sequence capturing mRNAs different according to the location).
   - Pros: high coverage
   - Cons: low resolution

5.2 Sequencing-based methods

Sequencing based spatial stranscriptomic methods all rely on the idea of barcoding the molecules in the original tissue using some in situ technique, and then to study those molecules using NGS. (The read

itself will contain both the RNA sequence and some barcode sequence that defines the position in the tissue of origin).

Sequencing-based spatial transcriptomics methods can be broadly divided into four main classes, depending on the strategy used to perform the in situ barcording. The barcoding strategy also affects the spatial resolution of the technique which can range from 55 μm spots to 0.5 μm spots.

These techiques, sorted from lowest to highest resolution are:

- **10X Visium platform** (2019), which is the optimized version of **spatial transcriptomics** (2016)
- **Deterministic barcoding in tissue** (2020)
- **Slide-seq** (2019)
- **Seq-scope** (2021)

NOTE: Resolution order and chronological order do NOT match.

These techiniques:

- Allow to analyze thousands of RNAs and allow for unbiased capture. In general though, only poly-A transcripts are analyzed.
- Do not have high-enough resolution to allow for cellular, or subcellular, level analysis (unique exception being Scope-seq).

5.2.1 Spatial transcriptomics (2016)

Spatial transcriptomics (ST) is the first sequencing based spatial technique. Distilling the procedure into its main steps:

- Create an array on which probes have been printed in about one thousand spots of 100 μm diameter, and with center-to-center distance of 200 μm. The probes are composed of cleavage site (to detach the probes from the slide), amplification and sequencing handle, spatial barcode, unique molecule identifier (UMI), mRNA capture region. Notice that *(a)* the amplification and sequencing handle is obviously shared by all probes, *(b)* the spatial barcode is shared by all probes in the same spot, *(c)* the UMI is different for all probes in a spot, but probes from different spots can have the same UMI (since the combination with the spatial barcode still keeps the probes distinct).
- Place the tissue section on top of the array and perform any required imaging step (H&E staining, FISH) to acquire an histological image on which to project the spatial data.
- Permeabilize the cells. In the original paper adult mouse olfactory bulb was used since it is a brain region with clear histological landmarks and gene expression reference data.
- The polyadenilated RNAs hybridize with the probes directly beneath.
- The hybridized probes can be removed from the array and can be used to construct a standard NGS library. The final reads will still contain the spatial barcode, therefore allowing to map the reads to the original spot on the array.

Given the fairly low resolution, multiple cells can be present in the same spot; for this reason one cannot be sure that all the transcripts sharing the same spatial barcode do indeed belong to the same cell (doublet detection and/or deconvolution is needed).

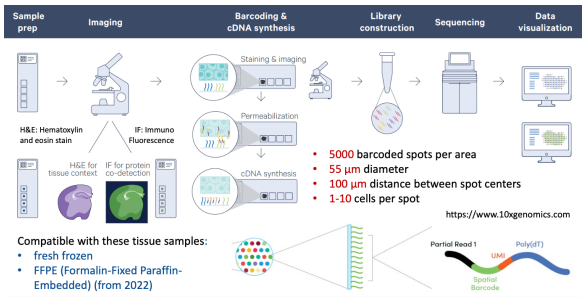

Figure 5.1

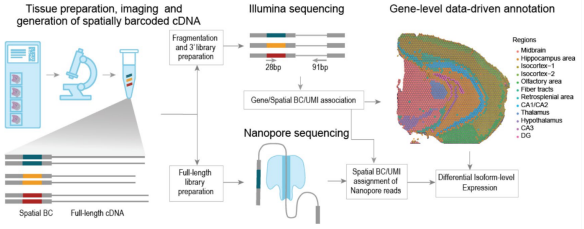

Figure 5.2: *Lebrigand et al bioRxiv 2022*

5.2.2 10x Visium platform for spatial gene expression (2019)

The 10x Visium platform works almost identically to the original spatial transcriptomics, it is simply a more optimized and streamlined procedure.

The spot size on the slide is reduced to 55 μm, the distance between spot centers is 100 μm and the number of spots is 5000. At this resolution, 1 to 10 cells could still be in the same spot. Notice that there still is space on the array which is not covered by any probe.

This technique works with fresh frozen tissues and, starting from 2022, with FFPE (Formalin-Fixed Paraffin-Embedded) tissues (which generally have lower quality since the fixation procedure can degrade RNA).

5.2.3 SiT (2022)

Spatial Isoform Transcriptomics (SiT) combines Visium platform with long read sequencing (Nanopore). The initial steps are the very same ones from the Visium platform, but then two different sequencing libraries can be prepared, a short read 3' library for Illumina sequencing, and a full read length library for Nanopore sequencing. This allows to analyze spatial distribution of isoforms (splicing variants).

Given that the first steps are identical to the standard Visium protocol, resolution and downsides are the same.

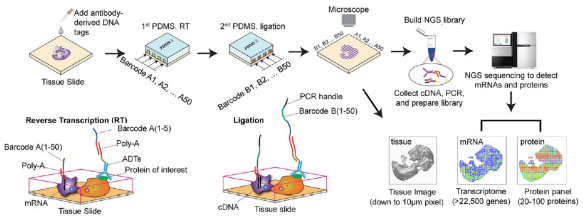

Figure 5.3: Liu et al., Cell 2020

5.2.4 DBiT-seq (2020)

Deterministic Barcoding in Tissue, or DBiT-seq, is a method that uses microfluidics to distribute the barcodes on the tissue slide, rather than having them pre-spotted on the array. Moreover the technique allows for parallel profiling of mRNA and antibody-based protein barcoding.

The procedure is the following:

- Fix the tissue on the slide
- If performing antibody-based protein barcoding, add antibodies for the protein(s) of interest. These antibodies are bound to an RNA sequence containing an antibody-derived tag (ADT, just like for CITE-seq) and a a ploy-A sequence.
- Using a microfluidic flow cell with 50 parallel channels, allow 50 different barcodes to flow over the tissue slide. These first barcodes, barcodes A, will bind to all poly-A chains (both those from mRNAs and those from antibodies).
- Using a second microfluidic flow cell with 50 parallel channels, placed orthogonally with respect to the first one, 50 barcodes are allowed to flow on the tissue slide. These barcodes, barcodes B, will bind to all barcodes A. Moreover, barcodes B contain a PCR handle. After this step, all pixels of the slides will be univocally identified by a combination of barcode A + barcode B.
- Perform imaging steps; this allows to define the position of all pixels on the slide.
- Standard sequencing library preparation.

Notice that the same sequences can be used for barcodes A and barcodes B, since they are always attached in a specific order. Channel width is what defines the size of the pixel and thus the resolution; usual channel widths are 50, 25 or 10 μm. The number of pixels is the square of the number of barcodes meaning, for instance, 2500 spots with a 50x50 grid.

5.2.5 Slide-seq (2019)

Slide-seq is akin to spatial transcriptomics in the sense that the tissue sample is placed on top of the probes; unlike spatial transcriptomics though, the probes are not directly spotted on the slide, but rather fixed on top of 10 μm diameter beads which are stacked on top of the slide. The probes have pretty much the same structure as those used in spatial transcriptomics, where the spatial barcode is bead specific and all the probes on a bead have different UMIs.

The procedure is the same as spatial transcriptomics but with the different slide.

Still, even with 10 μm beads, one bead could correspond to 1-3 cells.

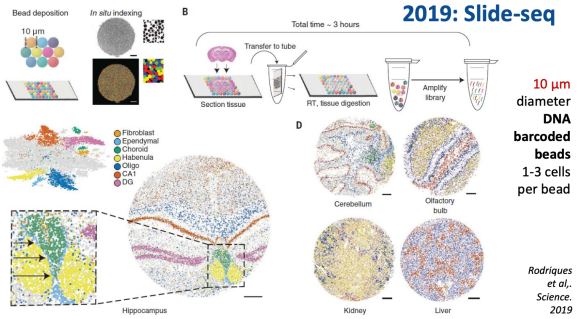

Figure 5.4

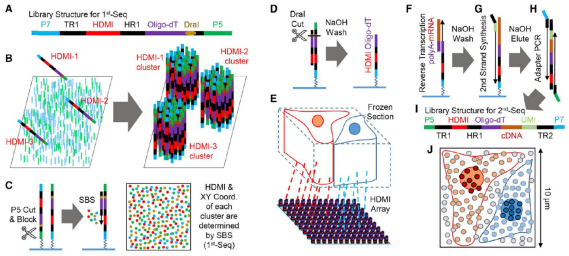

Figure 5.5: *Cho et al., Cell 2021*

5.2.6 Seq-scope (2021)

Seq-scope is a technique that allows for resolution comparable to an optical microscope (0.5-0.8 μm) by using a repurposed Illumina sequencing platform (MiSeq flow-cell in the publication).

The procedure is as follows:

- The probes, containing different high-definition map coordinate identifiers, or HDMIs, (which are basically spatial barcodes) are hybridized in random positions on the MiSeq flow-cell.
- The probes undergo the bridge amplification step, therefore forming clusters with one HDMI each.
- By using sequencing by synthesis, the position of the clusters on the flow cell and the HDMIs of each of them are defined. (Consider the clusters as analogous of the spots in other techniques, but randomly placed on the slide and with way smaller size).
- The probes are cut in order to expose an oligo-dT tail.
- The fresh frozen tissue (mouse liver sections in the original paper) is placed on top of the flow cell and the cells are permeabilized, allowing for the poly-adenylated RNAs to bind the probes.
- Imaging steps to define cell positions in the tissue.
- During the second strand synthesis a UMI is included for each molecule.
- The hybridized probes are detached from the flow cell and PCR adapters are added.
- The probes are therefore amplified and sequenced.

Given the very small size of the clusters, the technique allows for sub-cellular resolution; each cluster will contain only the RNAs from a region of the original cell.

5.3 Imaging based methods

5.3.1 In-Situ Hybridization

ISH (In Situ Hybridization) is based on nucleic acid probes synthesized, **labeled**, purified, and annealed with the specific target (by complementarity). The technique can be applied to DNA and RNA with nucleic acid probes (20-50long).

5.3.1.1 Radioactive ISH (1967)

ISH was first used to study the formation and detection of RNA-DNA hybrid molecules in cytological preparations (*Gall and Pardue, PNAS,1969*). Cells are treated in order to denature DNA and incubated with RNA to detect hybrids by **autoradiography**. Probes are able to catch radioactivity signal and are identified where the ribosomes are being produced.

5.3.1.2 FISH (1997)

- **Direct FISH detection:** fluorescent labels attached to the probe which will hybridize to a target DNA/RNA
- **Indirect FISH detection:** biotin is attached to the probe. Streptavidin linked to a fluorescent tag binds biotin with high specificity.

Fluorescence microscopy can be used to find out where the fluorescent probe is bound.

5.3.1.3 Whole mount ISH (1989)

Whole mount in situ hybridization determines the RNA expression pattern of a gene in the context of a whole embryo or embryo piece/organ (also 3D). First experiments were carried out already knowing the pattern of expression of certain genes. They selected:

1. hunchback
2. krippel
3. knirps
4. fushi tarazu

They could also look at how different was the localization between mRNA and proteins produced, with antibodies.

5.3.1.4 RNA smFISH (2002)

The release of the first reference genomes, allowed the computational design of probes. Multiple probes can be used for a single transcript for increased specificity and single molecule resolution. Each probe is labeled with different labels. The combination of "colors" forming a pseudocolor is assigned to a specific transcript.

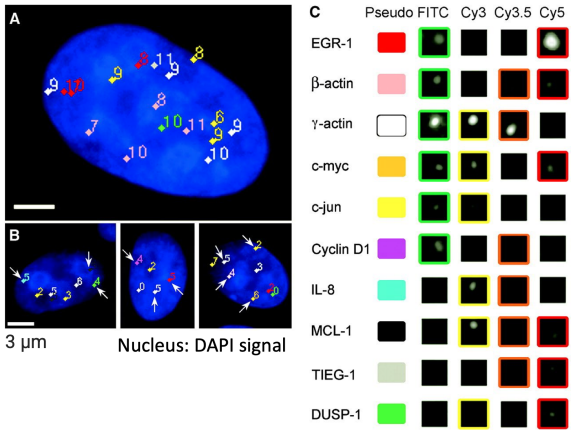

Figure 5.6: *Levsky et al, Single-Cell Gene Expression Profiling, Science 2002*

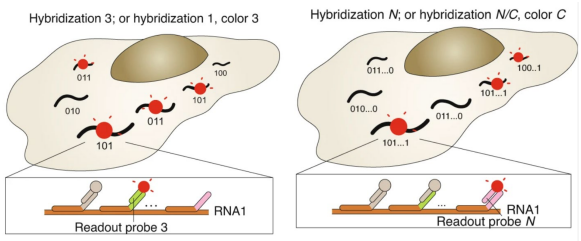

Figure 5.7: *Chen et al, Science 2015*

For smFISH, they used 4 different fluorophores for the identification of 10 genes. smFISH has a problem: the number of combinations (pseudogenes) is limited by the number of probes. This was solved by the next technique.

5.3.1.5 MERFISH (2015)

Multiplexed Error-Robust FISH (MERFISH) is based on multiple rounds of hybridization and detection of signal, with consequent removal of probe for each round. The barcode is formed by checking at each round if the signal of the probe you are expecting is present on an RNA molecule (N-bit binary code: if yes, bit signal 1, else 0). They use an indirect method with two types of probes, one type that binds to RNA and one that binds to the fluorophore.

Encoding-readout labeling strategy allows genome-scale imaging with shorter experimental duration hybridization of FISH probes to exogenously introduced readout sequences is much faster than hybridization directly to cellular RNAs. The binary coding scheme (1 color) distinguishes 2N genes with N rounds of hybridization. When C colors are used, 2NC genes can be distinguished with N hybridization rounds.

Analysis example: simultaneous measurement of 140 RNA species in a single cell (IMR90, fibroblast). MERFISH with a 16-bit MHD4 code (modified humming distance 4 code) at least four bits must be read incorrectly to change one valid codeword into another). Error detection & correction.

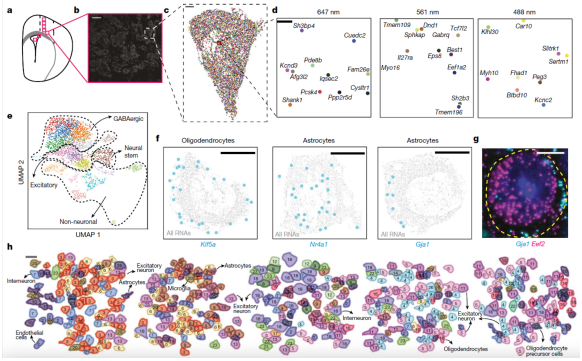

Figure 5.8

5.3.1.6 RNA seqFISH+ (2019)

80 rounds of hybridization with 3 color imaging (identified 47k mRNA). Tried in mouse brain slices and identified the position of transcripts with respect to all the others, in different types of cells. Is it also possible to then build the architecture of the tissue, in terms of what kind of cells appear close to others.

5.3.1.7 MERSCOPE platform (2021)

Commercial platform using MERFISH technology. It is possible to study spatial features at different resolution levels:

- whole section (9x7 mm) e.g. tissue organization
- wide field of view (200x200 micron) e.g. cell interaction or function
- sub-cellular (12x12 micron, equivalent to $<$ 100 nm) e.g. L2 or 3 IT Glutamatergic neuron

Vizgen MERSCOPE automatically segments the cells in the MERFISH measurement using a segmentation method called Baysor, that optimizes cell boundaries considering joint likelihood of transcriptional composition and cell morphology (size and shape of the cell). The approach can take into account **nuclear (DAPI) or cytoplasm (poly(A)) staining**, however, can also perform segmentation based on the detected molecules alone. By tuning parameters, a better visualization can be obtained e.g. alpha for transparency, palette, etc. Using the positional information of each cell, we compute spatial niches (region of tissue). A relevant aspect into account is the doublet rate i.e. mixture of cell types in the same pixel.

5.3.1.8 10x Xenium platform (2023)

Probes are designed in a slightly different way: they have two complementary arms that, if close to each other, can undergo a process of **ligation** (this allows to rule out the fact that the probe might bind to two different mRNA molecules).

5.3.1.9 AKOYA platform (2021)

The CODEX platform developed by Akoya Biosciences is based on DNA barcoded antibodies, which allow to perform multiplexed imaging. Antibodies are specific for one protein, and the addition of fluorescent probes complementary to the DNA barcode (antibody specific) allow the visible rendering of the antibody. After signal detection, chemical stripping of fluorescent oligonucleotides occurs. The difference with previous techniques is that we capture proteins instead of transcripts.

5.3.2 Summary

Improvements:

- Throughput (up to 1000s different RNAs)
- High resolution (subcellular)
- No polyA restriction (but sequence and splicing knowledge necessary to design complementary probes)

Challenges:

- optical resolution (microscope)
- density & structure of transcripts in cells (RNA granules, RNA-protein interactions could compromise probe hybridization)

5.4 Spatial omics analysis

5.4.1 Deconvolution of low-resolution multi-cell spots

The aim is to uncover cellular heterogeneity in low resolution spots (e.g. Visium) to disentangle the spatial patterns of cell types. Most methods rely on the availability of a single cell annotated reference.

- understand how many nuclei are present in a spot
- analyze cell types present in each spot

In *Li et al. Nature Communications. 2023* 18 different deconvolution methods are compared. The composition of the spot is represented as a pie chart.

The best performing method is CARD, tool published in 2022 based on matrix factorization. A single cell reference or matrix of gene expression data is required (for cell marker identification) and the location of the spots from spatial transcriptomics. The final results will report the estimated cell type proportion.

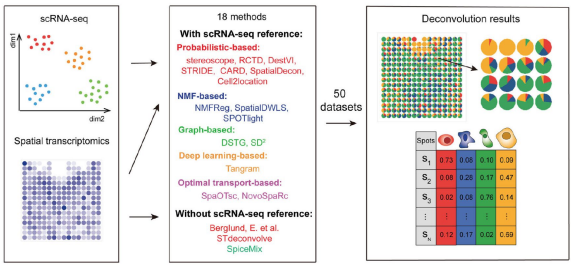

Figure 5.9: *Li et al. Nature Communications. 2023*

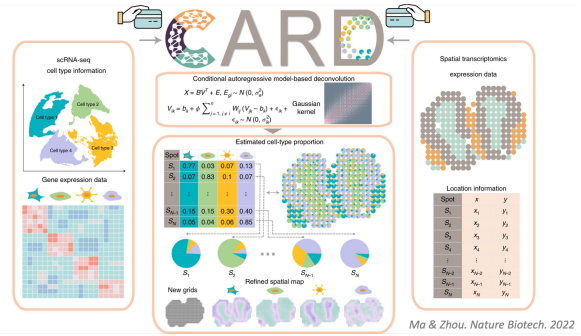

Figure 5.10

5.4.2 Cell segmentation in imaging-based methods

Analysis specific for imaging approaches. **Baysor** is applied to distinguish the boundaries of individual cells based on:

- Co-staining information
  - nuclei (for example, DAPI) - most common approach
  - cell bodies (for example, polyA)
  - cellular membranes
- Spatial expression data
  - increased spatial density of molecules within cell somas
  - transcriptional composition of local molecular neighborhoods

5.4.3 Inference of Cell-cell communication

Reconstruct cell-cell communication from single-cell and/or spatial omics. The idea behind this is studying how cells react to stimuli from their environment(in multicellular organisms, other cells) relevant for apoptosis and cell migration, essential in homeostasis and disease. Ligands (soluble proteins) bind to receptors on the cell surface, activating cascade affecting transcriptional regulation.

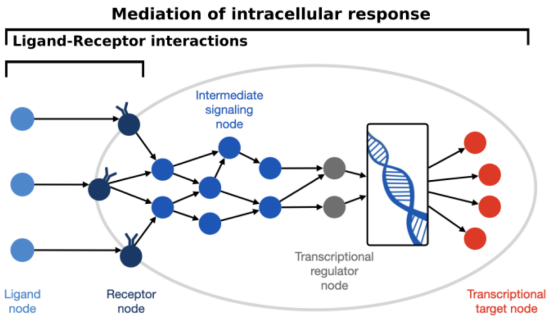

Figure 5.11

5.4.3.1 CellPhoneDB

The method infers crosstalk between pairs of cell groups (source and receiver) and is mainly based on expression of **ligands** (by the source) and **receptors** (on the receiver). Information about ligand-receptor interactions extracted from prior knowledge resources e.g. PPI-interaction, protein complexes and secreted or membrane proteins databases.

Enriched receptor–ligand interactions are identified from the average expression of a ligand (source) and a receptor (target). Assumption: expression levels match protein levels. The method can be used in multiple conditions to study how communication changes upon stimuli or in disease settings.

5.4.4 Integration between single cell and spatial omics

In order to perform a comprehensive study, it is possible to combine multimodal spatial omics e.g. DbitSEQ to combine epigenome and transcriptome.

5.4.5 Spatal omics reference atlases

- BICCN: diverse cell types in human, mouse and non-human primate **brain**
- HuBMAP: global atlas of human body
- Human Cell Atlas: map every cell type of the human body
- Spatial Omics DataBase (SODB):