Theory-III.pdf

Bioinformatica

Tenicas Omicas

2º Grado en Bioinformática

Escuela Superior de Comercio Internacional
Universidad Pompeu Fabra

Topic 1. Epigenetics

Epigenetics is the study of heritable phenotypic changes that do not include DNA alterations. Often involves changes in gene expression and gene activity. It explores how various factors can influence gene activity and function, leading to modifications in an organism's phenotype without altering its genetic code.

In simpler terms, epigenetics is concerned with the mechanisms that determine which genes are turned on or off in a cell or organism, influencing how genetic information is utilized. These mechanisms involve chemical modifications to the DNA molecule and its associated proteins, which can affect gene expression by either promoting or inhibiting the reading of specific genes.

They can also be inherited from one generation to another, potentially affecting the health and characteristics of offspring. The three primary types of epigenetic modifications are DNA methylation, histone modifications, and non-coding RNA molecules.

Teacher: Set of molecules and mechanisms that can perpetuate a cellular state.

Recap

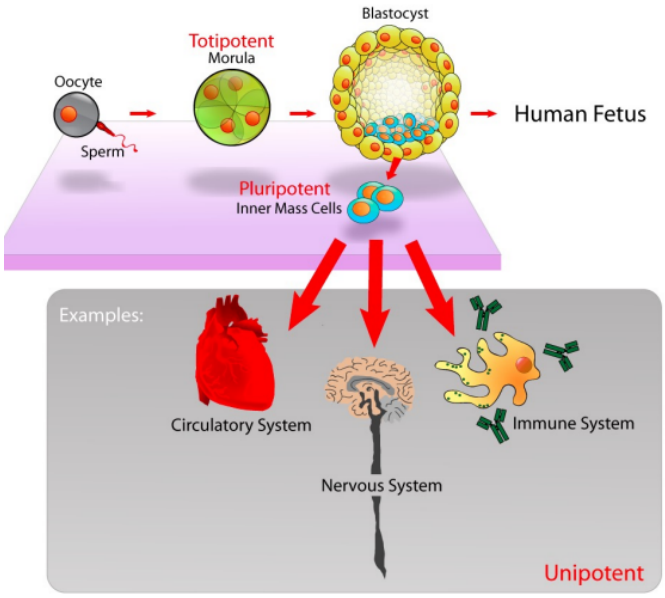
We are the result of cellular divisions that come from one single cell.

Totipotent cells: Can give rise to all cell types, including extra-embryonary (placental).

Pluripotent: Can give rise to all cell-types in the body.

Multipotent: Can give rise to more than one cell-type.

Unipotent: Can give rise to just one cell-type.



Is gene expression a heritable trait?

For epigenetic to be something heritable, it means that someone has observed that the level of expression of a gene is a character that can be inherited.

To some degree gene expression is heritable. Gene expression is controlled by elements in the genome that are encoded in the DNA and DNA can mutate and produce a hyper expression of a gene. Maybe this mutation increases the affinity of a transcription factor… This is what is called eQTL → Expression Quantitative Trait Loci → Variants in the genome that correlate to the expression of a gene.

Imagine that having a T in position 7 increases the expression of a gene. Then it is a eQTL.

So, since some of the gene expression is explained by variation in the DNA and DNA is heritable. Then, to some degree the expression of a gene is heritable.

What about the genes that are independent of the eQTLs? For example, genes that are induced by external factors such as temperature:
We put C. elegans in 2 different boxes:

- One is at 16 degrees, its optimal temperature
- The other one is at 25 degrees and, thus, it has to express a heat shock protein to avoid its death. If we put the progeny of this worm in a box that is at 16 degrees, they will still express the heat shock protein. This continues up to 14 generations and then it returns to the normal levels.

So, something that happened in the promoter of this gene has perpetuated 14 generations.

How does gene regulation work?

There are places in the genome where we have the gene and other places in which we have the regulatory elements. These regulatory elements can be active or not.

- Enhancers are specific DNA sequences that activate genes. They work by binding to specific transcription factors, which are proteins that help initiate or enhance the transcription of a gene. Transcription factors recognize and bind to the enhancer sequence, recruiting other proteins (polymerases) and forming a complex that interacts with the gene's promoter region. This interaction can influence the rate at which the gene is transcribed and the level of gene expression.

Interesting thing: Enhancers are also expressed at low levels. The transcripts are very small and fragile. But if you sequence enough, you will find them and therefore you can then find which is their sequence and locate them in the genome (since they act at long distances).

MONSTER ENERGY
GRAN PREMI DE CATALUNYA

DEL 24 AL 26
DE MAIG 2024

motoGP

MOTOGP

Una experiència per viure-la.

FIM

COMPRA LA TEVA
ENTRADA A **CIRCUIT.CAT**

Circuit de Barcelona CATALUNYA

Tenicas Omicas

Banco de apuntes de la

WUOLAH

Comparte estos flyers en tu clase y consigue más dinero y recompensas

1. Imprime esta hoja
2. Recorta por la mitad
3. Coloca en un lugar visible para que tus compis puedan escanar y acceder a apuntes
4. Llévate dinero por cada descarga de los documentos descargados a través de tu QR

Epigenetic information

- **Histone modifications:** Histones are a group of proteins that play a fundamental role in the organization and packaging of DNA within the nucleus of eukaryotic cells.

The primary function of histones is to compact DNA and facilitate its efficient storage in the cell nucleus. Without histones, the long DNA molecules would be too large to fit inside the tiny nucleus of a cell. The DNA wraps around a group of histone proteins, forming a repeating unit called a nucleosome.

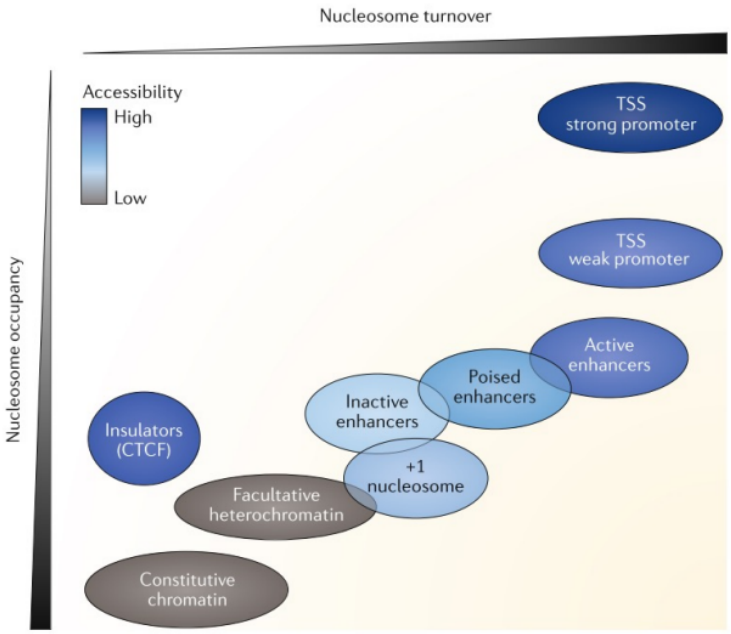
A nucleosome consists of a core particle made up of two copies each of four histone proteins: H2A, H2B, H3, and H4. These proteins form an octamer, around which approximately 147 base pairs of DNA are wound.

Histones also participate in the regulation of gene expression. The DNA wrapped around histones is less accessible to the cellular machinery responsible for gene transcription. Histones undergo various chemical modifications, such as methylation, acetylation, phosphorylation, and ubiquitination. These modifications can alter the interaction between DNA and histones, affecting chromatin structure and gene expression.

- **Nucleosome positioning:** Histones are not randomly distributed inside the nucleus. In fact, there is a correlation between gene activity. We have 2 measurements in the nucleosome position:
  - **Nucleosome occupancy:** How many nucleosomes are attached forever into a certain place of the genome. It illustrates how many nucleosomes we have in a position of the genome.
  - **Nucleosome turnover:** How much this assignment of the position of the nucleosome is present. If it is likely to be replaced by another nucleosome. If a nucleosome is never replaced/moved, it will have a turnover of 0.

In this 2 axis, we can place the different genomic elements present in the genome.

We find that if we are in a stretch of chromatin **(constitutive chromatin)**, which corresponds to a region of the chromatin that does not contain genes, it is densely packed with nucleosomes and the turnover is very low. This corresponds to **heterochromatin**.



This is typically found near centromeres and telomeres of chromosomes and contains repetitive DNA sequences.

On the other side of the spectrum we can see that TSS (Transcription Start Site) are the regions of the genome that contain less amount of nucleosomes and the nucleosome turnover is very high.

Recap

- Nucleosomes are barriers to transcription as blocks access to activators (TFs) or difficult elongation of transcripts by the polymerase.
- Positioning is particularly important at TSS.
- Loss of nucleosome upstream genes → activation of transcription

What is the mechanism by which a nucleosome is placed in one particular place or another? Chromatin remodelers have the affinity to recruit histone proteins and are responsible for altering the accessibility of DNA by moving, repositioning, or modifying nucleosomes.

Depending on the %AT, there will be more or less nucleosomes.

In places where there are a lot of transcription binding sites… there will be less nucleosomes, because it is a type of competition to get a position. If a position is full of transcription factors, a nucleosome can not be placed.

Histone modifications have a direct effect on gene expression. Histones have tails that can be modified in meaningful ways. Methylating or acetylating a histone tail will modify the enhancer ability to recruit transcription factors.

Note that histone tails are unstructured and AA are very conserved.

There are many ways in which a histone tail can be modified. There are 12 types of chemical modifications on 130 sites, so there is a high combinatorial complexity:

- **H3K4me3**: Active promoters (lysine 4 is methylated 3 times)
- **H3K27ac**: Active promoters and enhancers
- **H3K36me3**: Tx elongation
- **H3K9me3**: Heterochromatine (silences a gene forever)
- **H3K27me3**: Repressed state (if a promoter of a gene is marked like this, the gene is silent). Contrary of H3K27ac

We are not modifying the DNA sequence but they modify gene expression

- **DNA methylation of the cytosines:** Covalent transfer of methyl group to a Cytosine. This modification first originated in bacteria, and so it was also present in the first eukaryotes. Lost in other lineages

  Most of the methylation occurs in the CpG islands. So, most of the C that are methylated are followed by a G.
  We would expect equal frequency of CA, CT, CC and CG. But CG is much less represented. The reason is that when C suffers from deamination, it gets repaired into a T.

  Genes that contain CpG islands are housekeeping genes → Genes that are expressed in all cell types because they codify for proteins involved in functions present in all types of cells.
  The CpGs of these promoters tend to be demethylated. Because methylation does not allow transcription. Because methylation does not allow the recruitment of TF, polymerases…

  So, DNA methylation is a form of silencing a gene.

Imprinting: There are some genes that are only expressed either the fathers or mothers copy. This happens because one of the copies is methylated.

Methylation can also happen in the exons of the gene and this is associated with repression and activation for other genes. This is because methylation is not that important in this case because there are no TF that are trying to bind (unlike in the promoter).

Note that methylation is reversible. There is an intermediate state in which we have hydroxymethylation. If a gene is hydroxymethylated, it will be methylated in the future. So, we can look for this to know the future using TAB-seq techniques.

Who are the agents of DNA methylation? DNMTs. They are DNA methyltransferases responsible for adding methyl groups (-CH3) to the DNA molecule, leading to DNA methylation. They are responsible of:

- De novo DNA methylation: DNMT3A, DNMT3B and DNMT3L
- Maintenance DNA methylation: DNMT1 they recognize methylated C and methylate them again. Useful for Hemi-methylated DNA
- Gene silencing

Who are the agents of DNA demethylation? TET enzymes

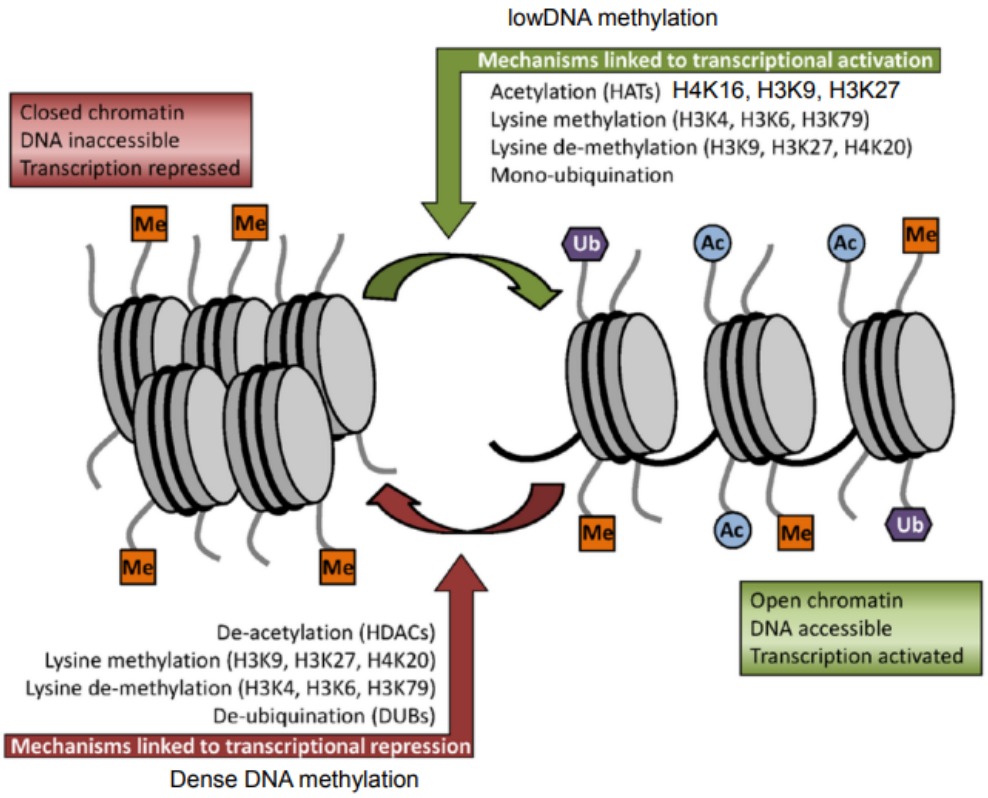
DNMT deficient organisms die at embryonic stages because:

- Aberrant expression of protein-coding genes (there is no repression)
- Genomic instability
- Massive derepression of transposable elements

Euchromatin vs heterochromatin

Euchromatin has a less condensed and more open structure. It is characterized by a dispersed and less compact arrangement of nucleosomes, allowing greater accessibility of DNA to transcriptional machinery and regulatory proteins.

Heterochromatin has a highly condensed and tightly packed structure. It is composed of densely packed nucleosomes, making the DNA less accessible to transcription factors and regulatory proteins.



Molecular agents of epigenetic information

Since we have all these agents together, it is tempting to try to make a code (the same we did with the genetic code). But this is very difficult because in general terms we can find correlations but when we move to specific terms it gets complicated.

People say that proteins that put these modifications are “writers” and the proteins that recognize these modifications are “readers”:

- But there is not reading of a code (as in triplets by ribosomes)
- A better way to describe these readers and writers is to call them as **binders** and **modifiers**, which are more descriptive terms
- The terms ‘activating’ and ‘repressive’ imply causality, and sometimes the opposite effect on transcription is observed.

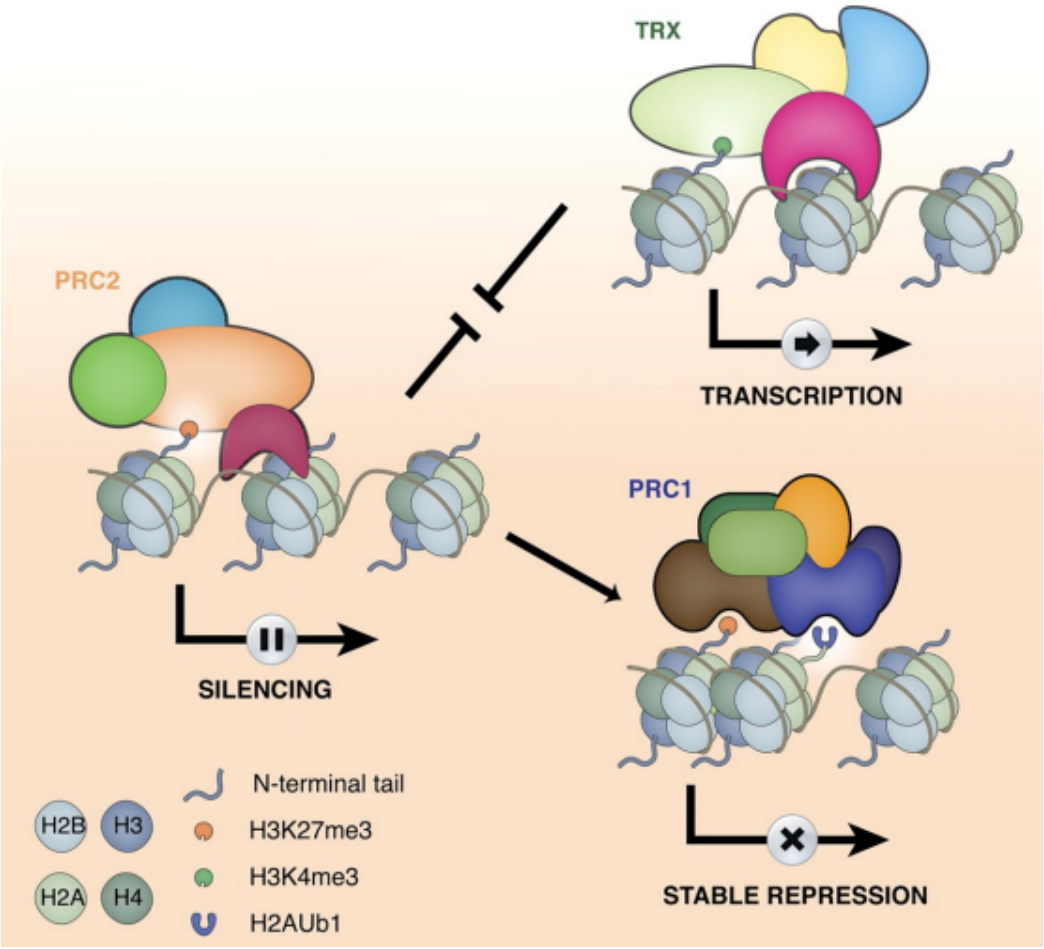
Some of the agents of the histone modifications are:

- Polycomb
- Trithorax

They are used to understand the regions that are going to be silenced or activated.

Polycomb (PRC2) is a complex of different proteins that transfer a methyl group to the H3K27. Thus, it is repressing the gene expression. PCR1 recognizes this and ubiquitinises H2A, stabilizing repression.

Trithorax (TRX) antagonizes PRC2 by tri-methylating H3K4 activating gene expression.



Cause or consequence?

Histone modifiers can recruit RNA-pol II, so one hypothesis suggests that histone modifications serve to stabilize the binding of the modifiers which usually can read the same modification they bind.

So, they create a modification, they can recruit the polymerase… So, the guy that is modifying the chromatin has the ability to recruit the transcriptional machinery.

So, methylation affects gene expression:

- Make it difficult for TF to bind DNA (~20% of TF bind less when methylated compared to methylated)
- DNMT can recruit histone deacetylases (HDAC)
- DNMT can recruit K3K9 methyltransferase

Interactions are not always intuitive

Methylation can activate gene expression by antagonizing the action of Polycomb in non-promoter CREs.

Epigenetic memory

How is epigenetic information maintained through cell division?

- Epigenetic modification need to survive DNA replication and mitosis
- Genes need to be kept silenced/activated
- How epigenetic landscape is kept through cell division?

Epigenetic barriers

Epigenetic components maintain ON/OFF states
Epigenetic barriers maintain actively somatic states.
Alterations when they are accidental can lead to disease. When programmed they constitute developmental programs.

We already know how it works for methylation. When we separate the strands to make the copies, one of them will be methylated. Thus, we will be in a hemi-methylated state. Then we just need to methylate again with DNMT1.

What about histones?

- Chromatin modifiers
- Nucleosome remodelers
- Histone chaperones

This is understudy.

Is IDENTITY reversible? Can a cell be re-programmed?

A fibroblast will be a fibroblast forever? Or can it change by being re-programmed? Yes, they can change

Induced pluripotent stem cells

If you feed a cell with these 4 transcription factors (Yamanaka factors)

- OCT4
- MYC
- SOX2
- KLF4

You are essentially reverting back the epigenetic landscape of this cell up to a pluripotent stem cell. Once you have a pluripotent stem cell, you can use it to differentiate into any tissue you want.

In the lab, you can take cells from the skin and revert them back to a pluripotent stem cell and create any tissue you want.

“Improved methods to select for iPSCs that have efficiently overcome epigenetic barriers are important to unleash the full potential of iPSC technology.”

Pioneer transcription factors

These Yamanaka factors are pioneer transcription factors. Which are a subset of TF that have the ability to bind DNA that is around nucleosomes.

By doing that, they can kick out nucleosomes from the cell. So that it does not matter which histone modifications they had before.
There are multiple models on how these pioneer TF can compete with nucleosome:

- Passive: They just compete with the nucleosome. If the concentration of the Pioneer TF is higher, it will win.
- Collateral
- Active (pioneer): TF binds to the DNA where the nucleosome is found and it replaces it

Chromosome X inactivation

1. One of the female chrX is silenced to avoid gene dosage problems.
2. The random choice for an inactivated X-chromosome (Xi) (i.e., the paternal or the maternal one) is completed at a very early phase of embryonic development.
3. Once it is decided, the copy remains silenced for life in this and all descendent cells.
4. Silencing is initiated by XIST (non-coding RNA). Recruits chromatin remodeling factors to “heterochromotize” one copy.
5. 15-20% of genes escape inactivation (recently diverged from ChrY).
6. Promoters in Xi are also methylated.

The final epigenetic element is Noncoding RNA.
DNA methylation is not the first thing that happens to inactivate chromosome X. The non-coding RNA XIST interacts with polycomb and represses the whole chromosome and later the CpG sites are methylated.

Building epigenomes

There are a number of omic techniques that are used to profile the accessibility of the chromatin on each one of these chromatin modifications:

- Chip-Seq
- ATAC-Seq
- …

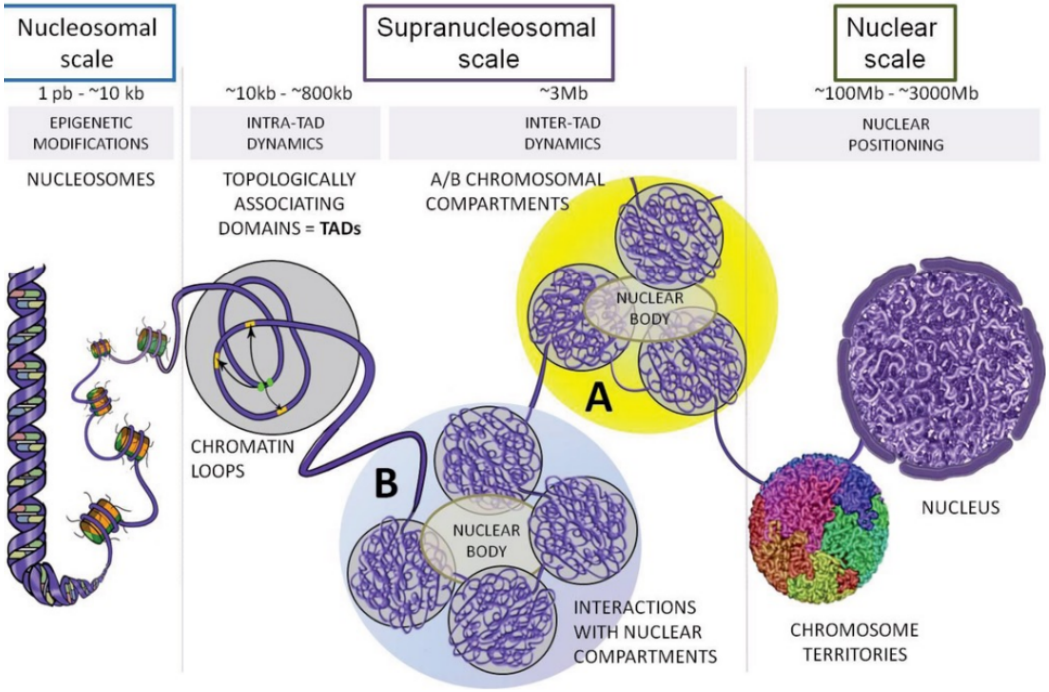
You are going to have reads that map on the genome but there will be some regions where you do not have any read.

Chip-Seq

We want to profile the places of the genome that have a k27 methylation. You will design an antibody that recognizes this specific histone modification. Then you will fragment the DNA and incubate it with the antibody.
The antibody will bind to the fragments of DNA that have this modification and you wash the rest. Then you sequence it.

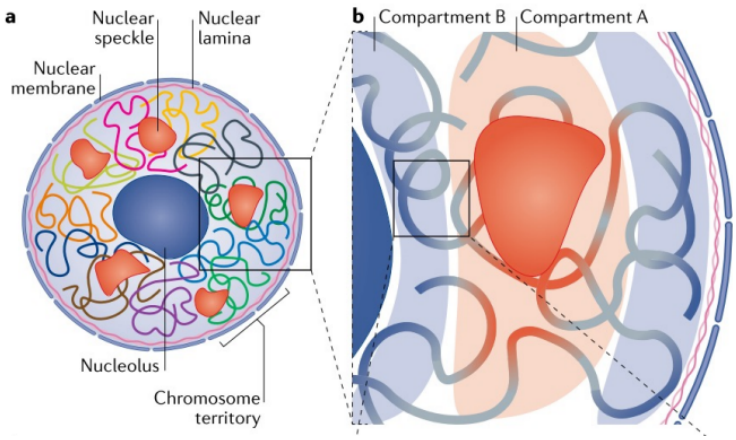
These reads will then be mapped into the genome.

3D GENOMIC ARCHITECTURE



The genome is organized inside the nucleus not at random. There are some scales:

- **Nuclear scale:** We can see chromosome territories. Each chromosome can be found at a certain location of the nucleus. The contacts between chromosomes tend to be preserved.
- **Supranucleosomal scale:** Each territory regions are not randomly distributed and where they are placed correlate with gene expression. We can find the genome compartments “A” and “B”.
    - **Compartment A (interior nuclear space):** The genome tends to put the parts of the DNA that are reached in genes (transcriptomically Active). Inside you can find nuclear speckles, which are nuclear domains that contain a lot of splicing machinery.
    - **Compartment B (around the center of the nucleus or below the nuclear lamina):** Inactive regions



These compartments are not static, they are dynamic during differentiation (they move). Because there are some regions that are active or inactive depending on the environment.

Some evidence that pulling DNA to the edge of the nucleus can cause silencing.

- We can see that in the lamina K9 and k27 are methylated by polycomb. Thus, these regions will become inactive.

Inside compartments A and B we can find more organization that is very important for gene regulation.

- **TADs (Topological Associating Domains):** They are regions separated by insulators and inside we can find loops. These regions of the chromatin have a higher frequency of physical interactions among themselves compared to interactions with regions outside the domain.

  TADs facilitate the interactions between gene regulatory elements, such as enhancers and promoters, within the same domain.

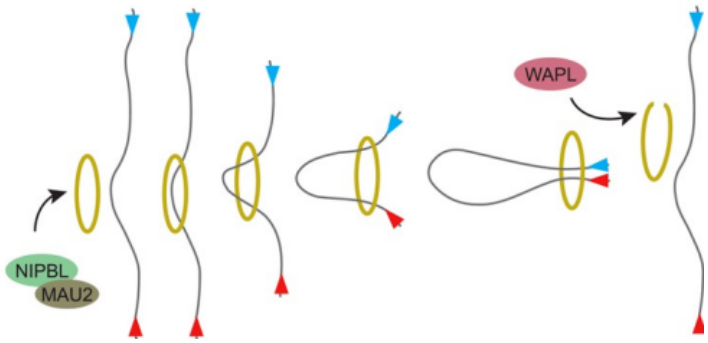
  TADs are defined by boundaries or borders that limit the physical interactions between neighboring TADs. These boundaries act as insulators (such as CTCF), preventing the spread of regulatory signals between adjacent TADs. This insulation helps maintain the independent regulation of genes within each TAD.

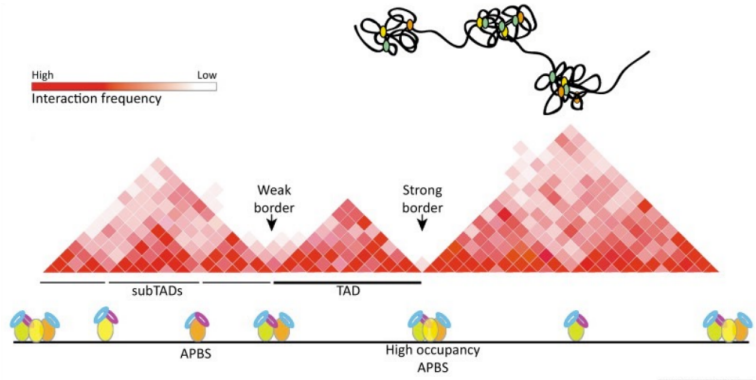
  Note that there is movement between compartment A and B, but the TAD boundaries are conserved across cell types.

  As we said, inside the TADs we have loops and FIRE (frequently interacting regions). So, inside the TADs, we have subTADs that are enriched in enhancers and super-enhancers.
  Super-enhancers are larger than enhancers and therefore they can recruit many more TF.

  Now-a-days we are working under the idea that these loops are formed following a extrusion model.

  - You take any region of the genome and when you find CTCF motives that are in the opposite direction, then you can form a loop stabilized by cohesin. So, if we can find these CTCF, we can deduce the formation of loops.





- Nucleosomal scale

Chromosome X inactivation

We said that inactivation comes from the activity of a non-coding RNA called XIST. but what about the 3D structure of chromosome X?

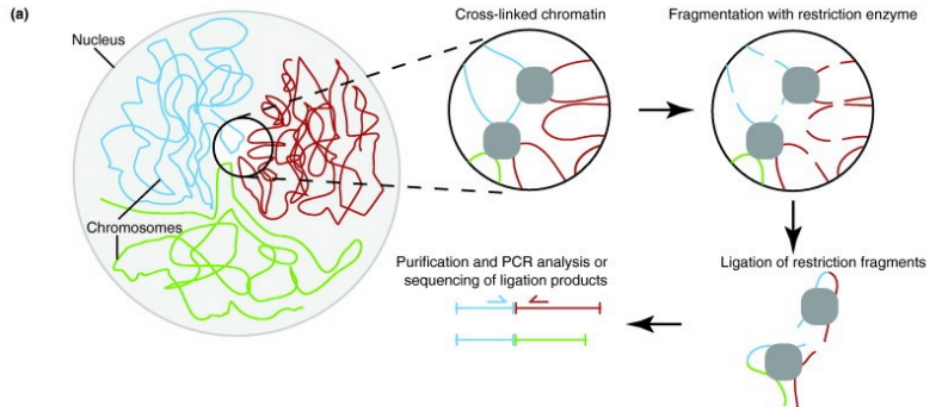
Recruitment of inactive chrX to lamina is important but not sufficient for its inactivation. So, the inactive form does not form TADs also.

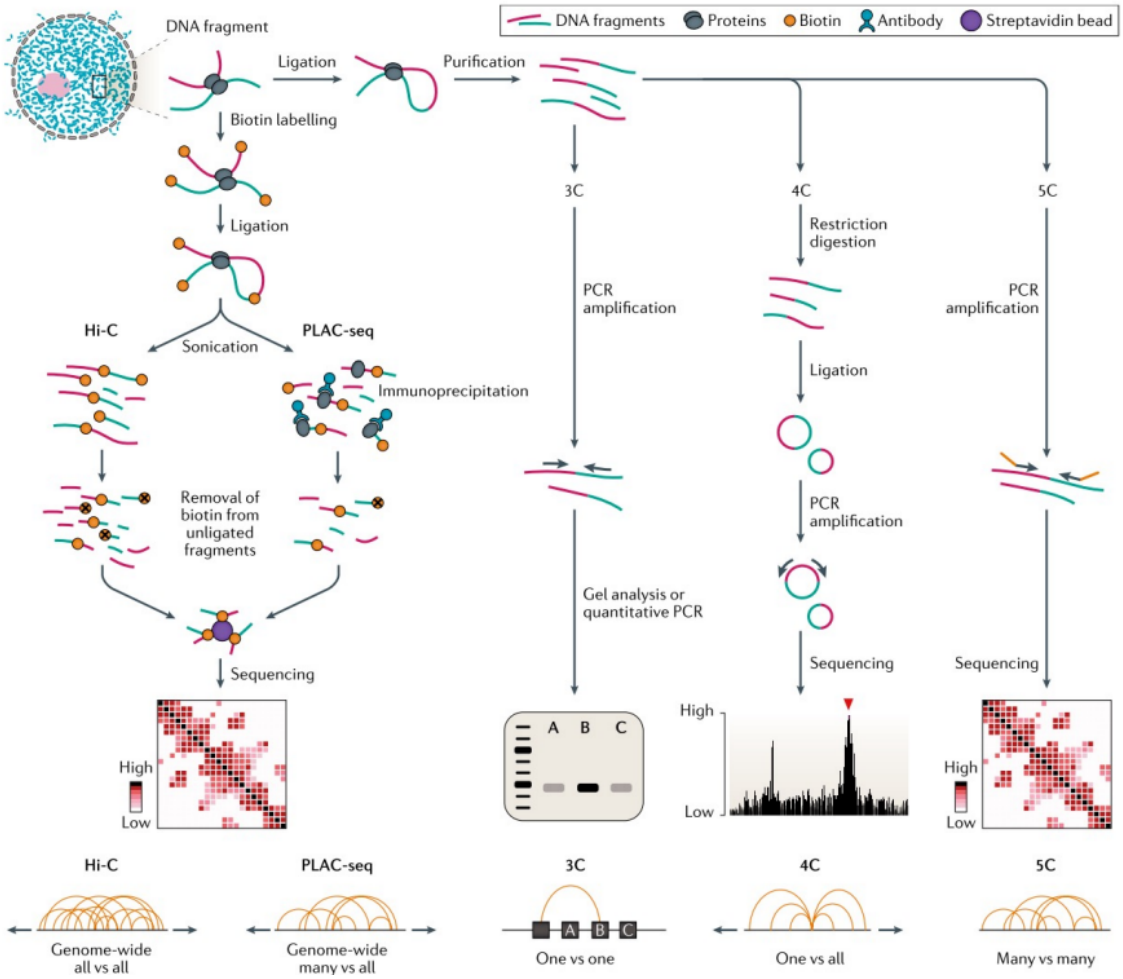
How can we evaluate experimentally and computationally interactions between areas of the genome? Which is the omic technique?

- 3C: one to one area
- 4C: one to all areas
- 5C: many to many areas
- Hi-C: all to all areas

The fundamental process to query 3D interactions is the following:

- Provided that you have an interaction between different regions mediated by a protein that we know. If I am able to cross-link (make stable the union of these proteins to DNA) and then I cut the DNA in pieces, I will be able to sequence these fragments and know the regions that were interacting.





3C

It assesses if a region of a genome is interacting with another region (a single interaction). To know this, you need to know the sequence of these regions and design primers.

In all methods we do the fixation, cross-linking, ligation and then you do different things. In 3C, we define primers. If the chimera exists, then you will find amplification.

4C

You query one region of the genome and you evaluate how many other regions of the genome interact with it.

In this case, we add an additional ligation step to circularize the molecule. Because you only know the sequence of interest (not the others). Then you use primers to amplify the unknown sequence.

Then I map all the reads and know the regions of the genome that were present in the circles.

5C

I have a megabase of the genome and I am looking for all possible regions that can interact. So, I design primers for all of them, put them all together in the mixt and make many PCRs.

I will obtain an intersection matrix. This is very expensive because you have to make all the primers.

Hi-C and PLAC-seq

It is mostly used, is cheaper, high-throughput and genomewise.

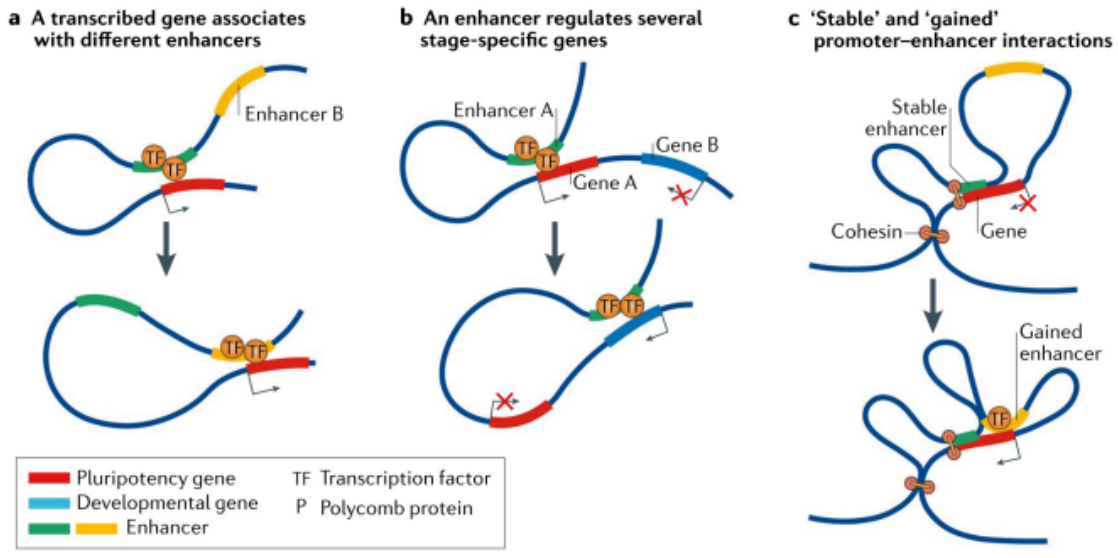
It makes an illumina library of the ligated pairs.

PLAC-Seq you select the fragments that have a particular protein bound to them. So, you need an additional immunoprecipitation step. For example, you may be interested in knowing what is interacting with the promoters (not on enhancers…), so I do an immunoprecipitation against H3K4-3me (that is found in promoters).

Enhancer/Promoter interactions

By analyzing all this data, we can see many different scenarios:

- Genes can be bound to different enhancers. So, they are regulated by multiple enhancers (not at the same time).
- One enhancer that binds to different genes. So, normally you are wrong when you say that an enhancer acts on its nearest TSS, since they can act at very long distances.
- We can also see promoter/promoter (enhancer/enhancer) interactions. Genes that are corregulated by 2 enhancers.



Now that we know that genes have a grammar in the genome, which is orchestrated by boundaries of TADs anchored by CpG island motives, no wonder that structural variation can have strong effects in gene expression even if they do not directly affect a gene.

Example: Creation of a new loop in a TAD can produce unknown interactions.

T2. Single-Cell Genomics

Advantages Single-Cell OMICS compared to doing bulk:

- Determine heterogeneity within cell types: We can understand the heterogeneity in our tissue or bulk. If we do RNA-seq from the skin, we will have a value of expression of each gene, but we have 100 different cell-types. Using Single-Cell, we will have a measure of expression for each cell type.
- Study rare cell types obscured in bulk tissue
- Determine trajectories of differentiation. Cells can be in different states.
- Identify cell type specific effects under any comparison (treatments, diseases, evolution, etc)

The transcriptome of mammalian cells consists of 10^5–10^6 individual messenger RNA (mRNA) molecules.
These messages represent some 4,000–12,000 different genes per cell

Huge number of measurements → Parallelization strategies of Illumina sequencing:

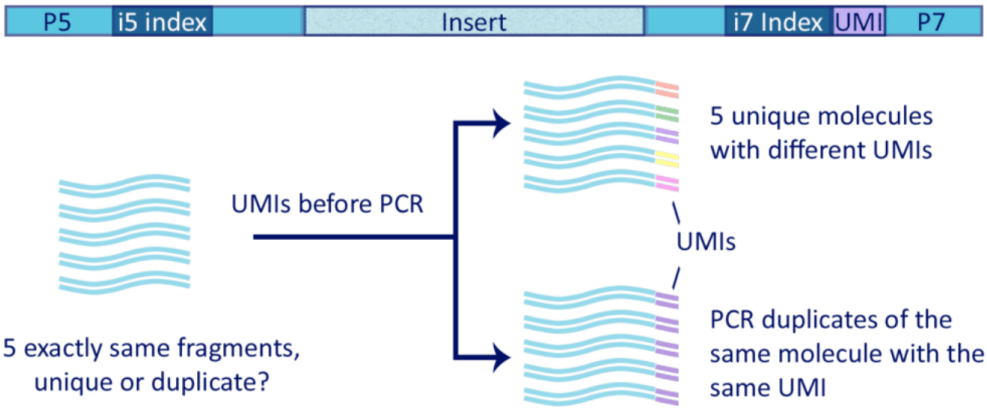
1. Microfluidics
2. Droplets
3. Molecular barcoding

Unique Molecule Identifiers (UMI)

Since we are querying individual cells, the amount of RNA is tiny. 1 cell ~10pg of RNA and typically you need ng for sequencing assays → Amplification needed by PCR.

PCR implies the loss of complexity in our sample and the creation of duplicates. To remove these duplicates, we can use UMI.

Amplification biases requires measuring unique molecular identifiers (UMI)

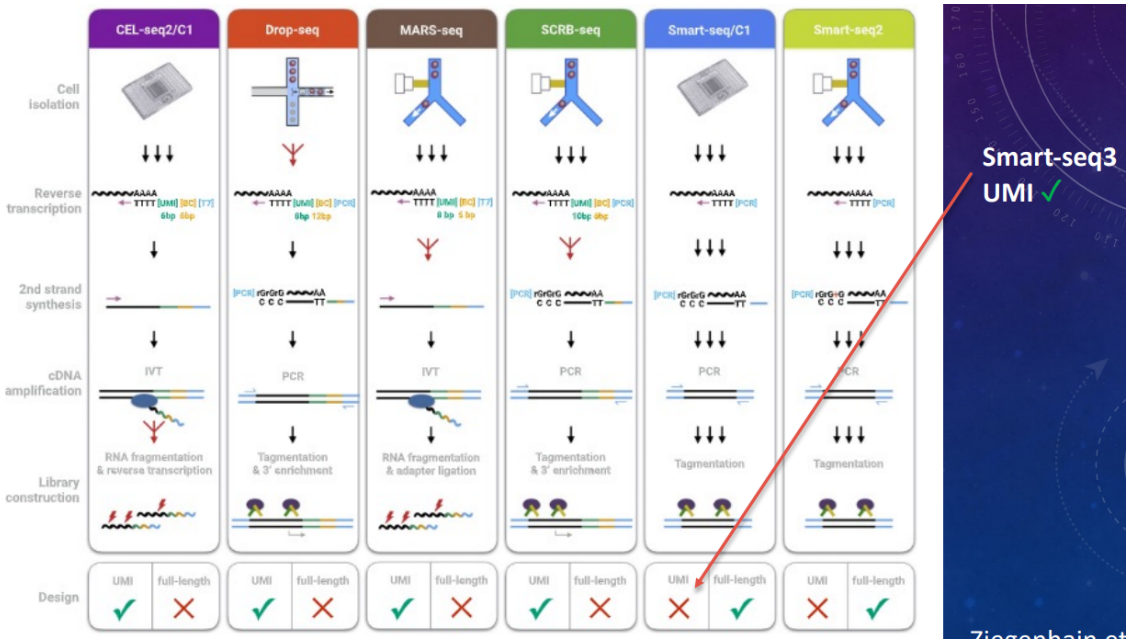


Before PCR, to each one of the reads I add an adapter that has a random combination of 8 nucleotides (the probability of finding the exact combination of 8 nucleotides is really small).

Once we do the PCR, we can collapse by this identifier and know if the distribution of the reads is equal or if there is a read that has been amplified much more than others.

Single-Cell RNA sequencing methods

These are some methods that you can use to make single-cell. Some of them use UMI and make a full amplification of the transcript or only a fraction of the transcript because they do Poly-A enrichment (for example).



Microfluidics (SMART-SEQ)

We try to dissociate the tissue to get all the cells in a suspension. Then, you put each cell one by one inside a really small tube and they are separated in different wells. We can modify the velocity of the separation to have more or less precision:

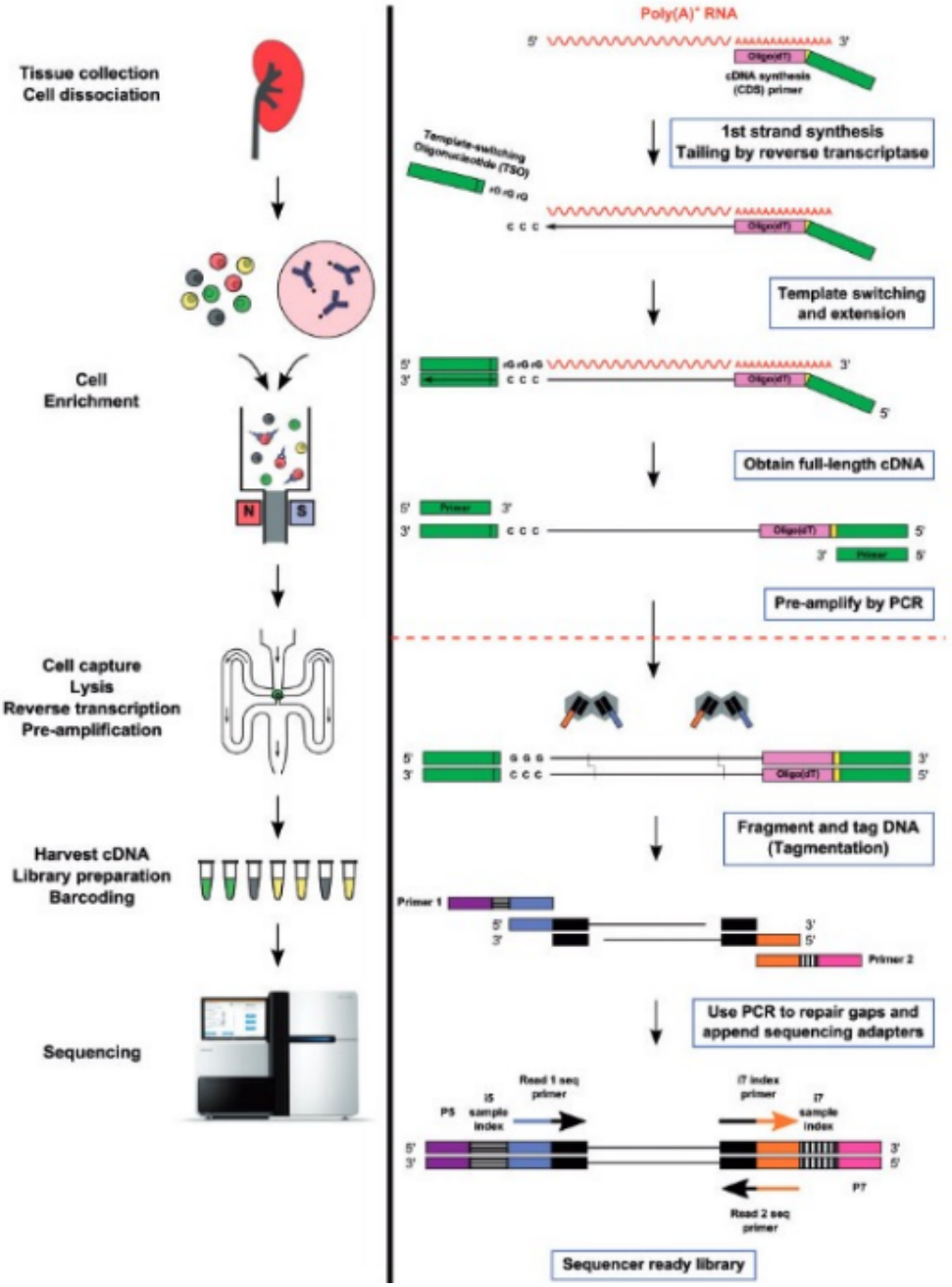
- Fast, you may obtain more droplets
- You want single-cell

Reactions for a typical amplification of an RNA-Seq library:

- Lysis of the cell
- Add adapters (UMI)
- Construct the library
- Barcoding of the reads that are in the same well. So, all the reads from the same well have the same barcode

So, I will be able to classify each read with its corresponding cell.
Within each barcode, I can fusionate by UMI.

It is expensive and has low-throughput!



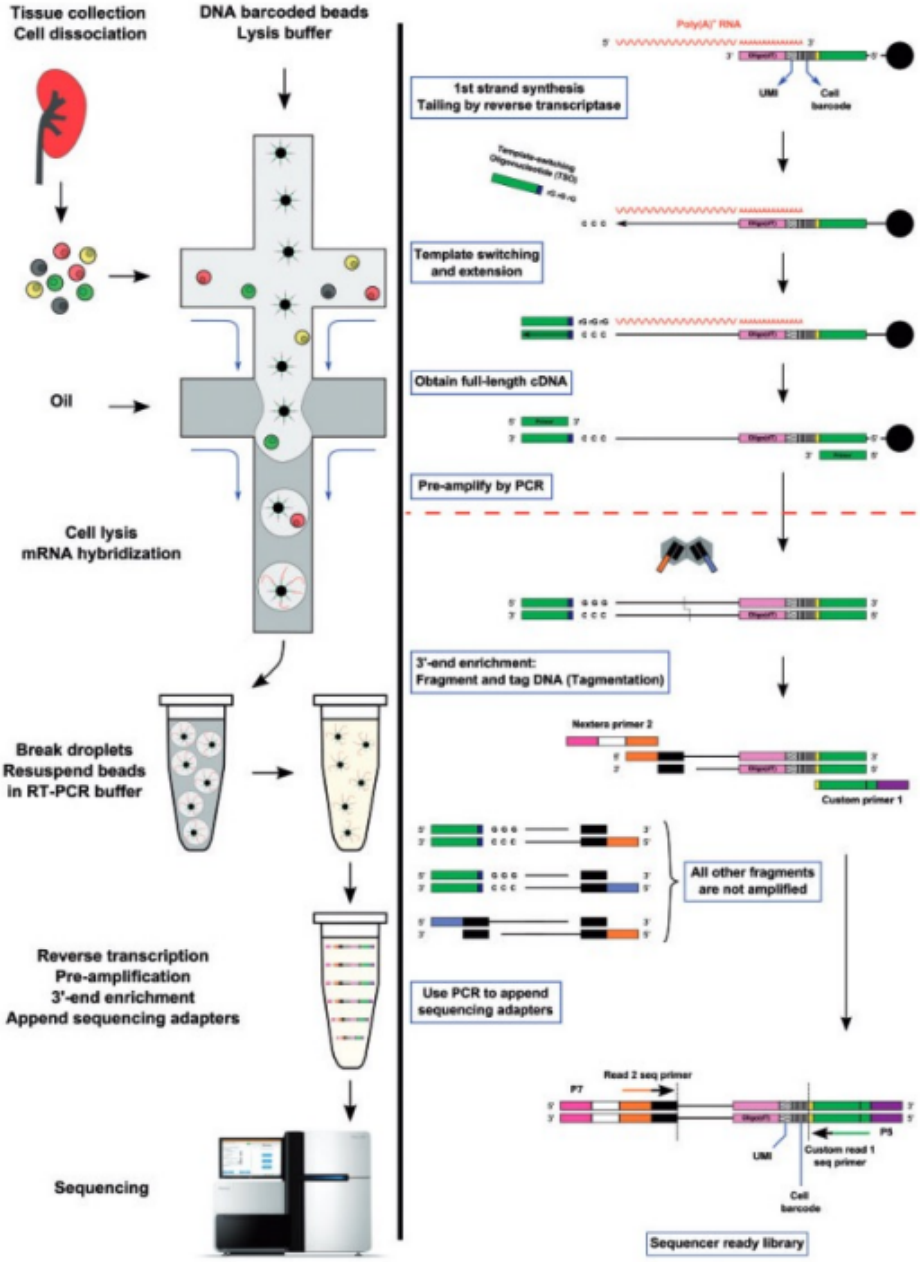
DROPLETS (DROP-SEQ) → Expensive

In this case, the reaction (amplification, putting the enzymes of the library…) does not occur in tubes but inside beads in one oil droplet.

Instead of playing with the rate or velocity of the microfluidics, we isolate the cells by putting each cell in contact with an oil droplet so that each cell is isolated in a lipidic phase.

Each lipid droplet contains one bead, which contains a barcode attached. Thus, when the RNA binds to that bead, it gets linked to that particular barcode and then a PCR is used to amplify.

- Does not distinguish isoforms



MOLECULAR BARCODING (SPLIT-SEQ)

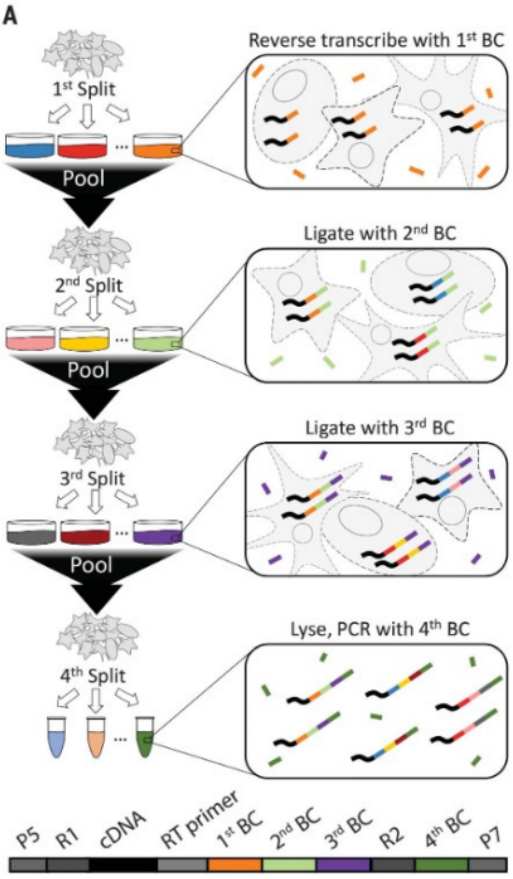
Requires no customized equipment, so it's not expensive!

It is based on sequential barcoding (combinatorial).

1. I have my pool of cells and I split in 10 eppendorfs my suspension.
2. In each eppendorf I put a different barcode.
3. Then I take each one of these 10 tubes and split them in 10 more and add 10 more barcodes.
4. I repeat this 2 more times.

Four rounds of combinatorial barcoding can yield 21,233,664 barcode combinations (three rounds of barcoding in 96-well plates followed by a fourth round with 24 PCR reactions).

We are randomly splitting and we will obtain random combinations. So, the RNA in each cell will have a different barcode.



Dissociation Matters

The dissociation of a tissue is not trivial. Some tissues are easier to dissociate than others, so:

- Can introduce a bias (different cell types easier to catch).
- If difficult to dissociate (e.g. neurons), one can use nuclei. Similar size, less bias (this is why blood is very easy to dissociate, since all cells are round, homogeneous…).

Instead of dissociating cells, we can use the nucleus (the expression of genes takes place in the nucleus). Note that each cell has a nucleus with similar shapes and thus they are easy to dissociate.

- You can also extract nuclei from frozen tissue, but not entire cells whose membranes break when frozen.

The problem is that in the nucleus, the RNA is not exactly the same as in the cytosol in which the RNA is processed.

- But: Nuclear transcripts comprise 20%–50% of all the RNA in the cell and include immature and unspliced RNA molecules containing introns.
- Intronic reads might account for 75% of all reads.

So, we will quantify the expression just using the RNA of the nucleus, even if the RNA is still not mature. We will also count reads of the introns.
Experiments suggest that quantification of expression in the nucleus correlates very well with the one from the cytosol.

Less cells with more transcriptome vs more cells with less transcriptome

We can not pretend to have the full transcriptome of a cell. We can get up to 2000 genes per cell. But we have so many cells.

We can:

- Spend more money on sequencing and getting a better representation of the transcriptome of a small number of cells
- Have a shallow coverage but for millions of cells.

We can separate two cell types, hepatocytes and blood cells (for example) based on the whole representation of the transcriptome of a few cells or a smaller representation of the transcriptome with more cells (accepting much more variation in the expression levels).

Not clear answer: Seems that shallow transcriptome coverage but many cells sufficient for common tasks: cluster identification, and PCA. At the cost of less accuracy in gene expression estimates

“The optimal allocation is to sequence at a depth of around one read per cell per gene”

Typically count matrices are 0 inflated.

- Only 10-20% of molecules present are actually observed in typical scRNA-seq experiments.
- Many technical reasons might explain why a present molecule is not observed Lost during sample collection, damaged during cell dissociation, failed to be amplified or sequenced. Some molecules have better chances (e.g. better RNA stability, cell location, sequence content), so there is a bias.

At the end, the value that you observe in the count matrix depends on the combination of 2 factors:

- Variation in expression level among cells (each gene has its own biological variance)
- The imperfect measurement process. How likely is that the gene is expressed and I catch the expression of the gene

Observation model = Expression model + Measurement model

Dropouts

A failure to detect a molecule.
Not all 0s are dropouts (some genes are truly not expressed)
Dropouts also affect non-0 observations → They can be present in cells in the matrix that are not 0. Maybe there is a 5 in the cell matrix, but there should be a 15 because there are 10 dropouts.

Workflow single-cell

1. **Quality control, mapping, and quantification:**

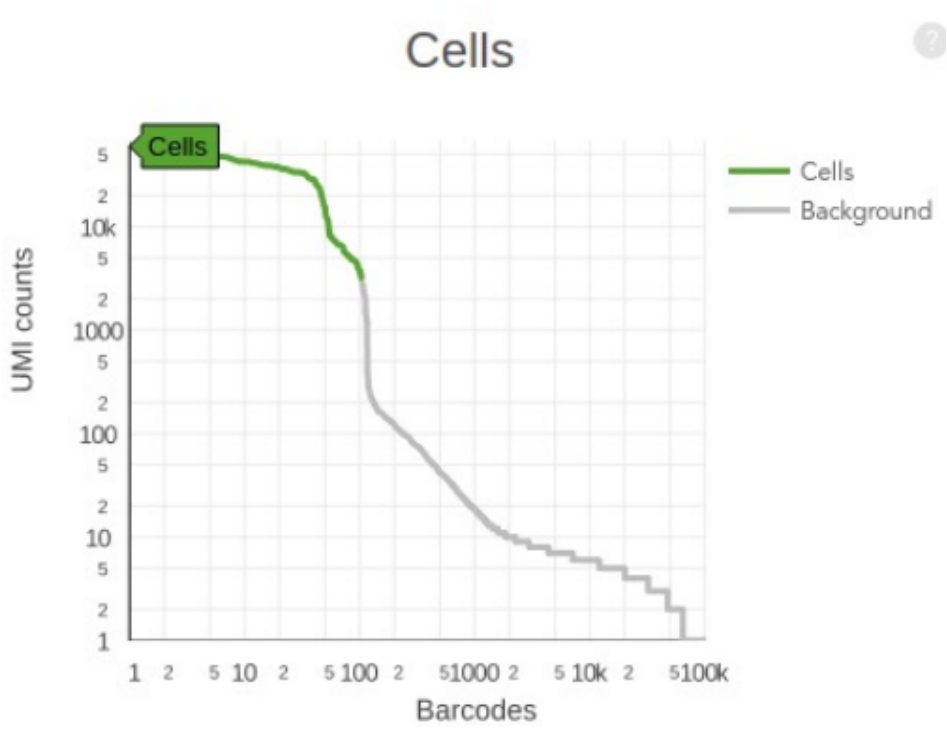
QC: Aims to decide what is a cell and what is noise.

This is the QC report that you would get when using 10x genomics.

- CellRanger (mapp, demultiplex, UMI counting) → Does the mapping, count matrix (gene expression) and tells you how many cells you have with a number of UMI.

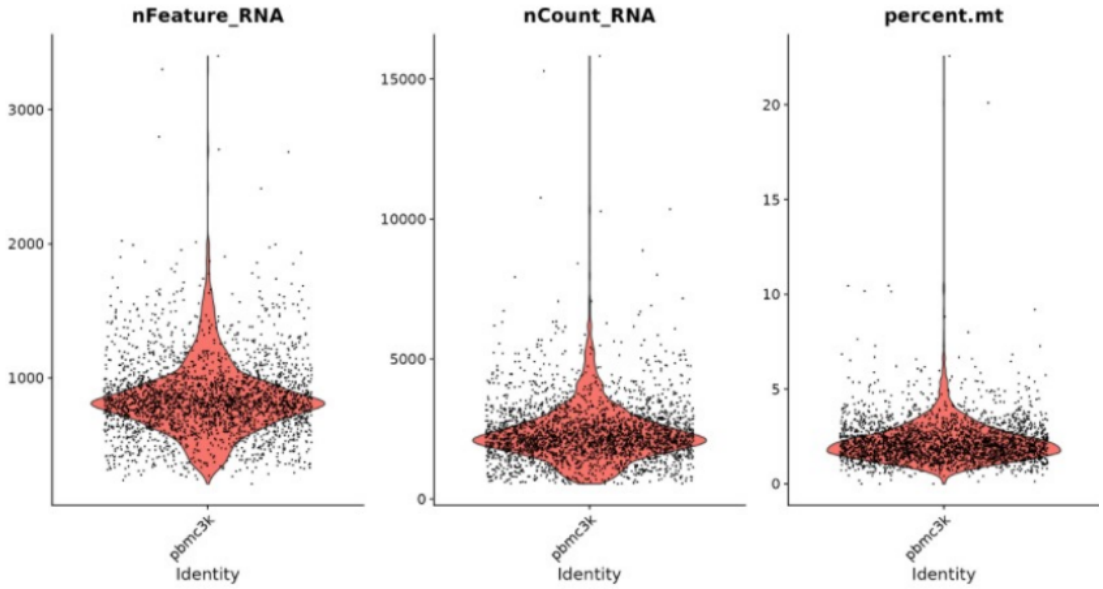
In this case, up to 1000 cells have at least 1000 UMIs.

- Knee plot to find threshold for what is a cell or not based on #UMI



Problem with small cells with low RNA content!!

EmptyDrops consider the distribution of environmental RNA to identify true cells (those that deviate from the ambient solution).



NEW YORK BURGER
A fuego, but lento
Calle Pelayo, 5.
08001 Barcelona
¡Únete y recibe una bebida de regalo!
ONE WAY
Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)

2. **Feature selection (variable genes) and dimensionality reduction**

Once you have all the big table of expression, we decide which genes are variable across cells. The gene is not equally expressed in all cells.

We use the top variable genes for dimensionality reduction. You do a PCA on the expression across all cells of the top most variable genes.

We are transforming our expression matrix of millions of columns into a new matrix that contains 20 columns of principal components.

To get the top variable genes, you essentially quantify the coefficient of variation of the genes across cells and you fit a model, because you have more variation if the gene is more expressed. The variance is much smaller in km than in cm to give you an idea.

So, we fit a model in which the variance is dependent on the mean. Then, everything that is above some SD you consider to be a variable gene.

Once you obtain the top variable genes you do a PCA (determine the dimensionality) and obtain an elbow of % of variance to choose which PC explains the variance.

3. **Visualization**

If we want to visualize the cells in the multidimensional space, it is going to be difficult. We are going to use 2D and 3D embedding methods to summarize the 20 dimensions to 2 or 3.

Non-Linear dimensional Reduction (UMPA/TSNE) → Useful for visualization of relationships between cells.

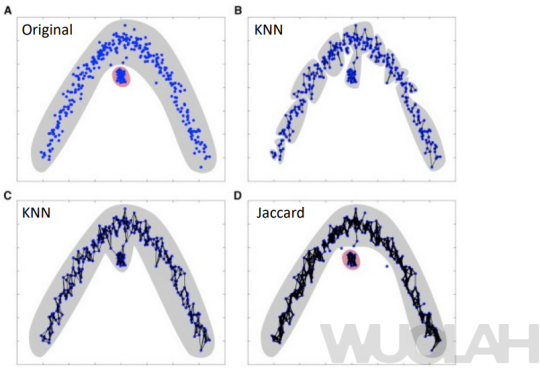
UMAP preserves the spatial structure better. So, it will make more sense from the biological point of view later on. In TSNE, the positions of the clusters are stochastic.

4. **Clustering**

Once we have the 20 dimensions, forget about the embedding (it is only for visualization) since the actual information is in the 20 dimensions, we cluster the cells according to these dimensions.

Example SEURAT.

- KNN graph based on the euclidean distance in PCA space
- We can create as many micro-clusters as we want, but then it is our job to know if this has biological meaning or not.



Important: For clustering cells, it is not only important the proximity of the cells but also the mutual proximity of the neighbors.

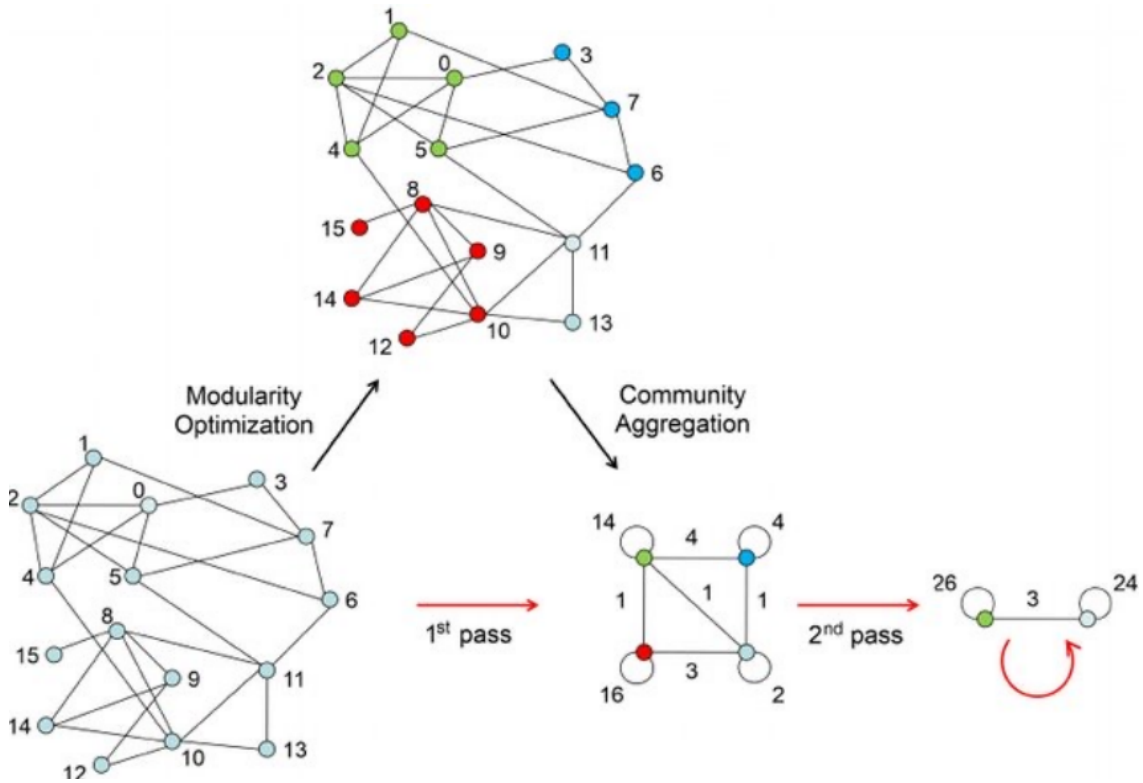
To solve this problem, we use the Jaccard distance → How many of my neighbors I am sharing with the node I am contacting.

- Refine weights based on Jaccard similarity with near neighbors
    - I.e. Scale distance by a measure of how many neighbors two cells share

On this distance matrix we apply the Louvain algorithm → Clustering algorithm that tries to find clusters out of a graph based on the distance, for example the Jaccard corrected distance.

- Louvain community detection method is then used to find a partition of the graph that maximizes modularity

1. We start saying that each cell is a community, we have not assigned any clusters.
2. Select randomly one node
3. Calculate what would happen in terms of global modularity if this node makes a cluster with another one. If the modularity increases, then we keep this cluster.
4. I do this for all combinations.
    a. Calculate the modularity if node 1 leaves cluster green
    b. Calculate the modularity if it goes to cluster red
5. At the end, instead of having 15 communities, we will have 4 communities.



We are maximizing the modularity score. So, by doing the modularity analysis we are deciding what clusters are meaningful.

Modularity (Q) is a measure to evaluate the quality of community structure within a network. Modularity measures the degree to which a network can be divided into distinct communities, with densely connected nodes within communities and sparsely connected nodes between communities. It compares the number of edges within communities to the expected number of edges if the network were randomly connected.

A higher positive modularity value indicates a stronger community structure, whereas a negative or close-to-zero value suggests a weak or random community structure.

LOUVAIN Algorithm
Phase 1: Modularity optimization

- Each node a community
- For each node move it to each other community and keep the one with maximum DeltaQ
- If all DeltaQ < 0 keep the community as it is.

Phase 2: Community aggregation

Repeat Phase 1 and 2 until delta Q = 0, meaning that I have achieved the maximum modularity.

5. **Marker genes**

Once we have the clusters, these are putative cell types. So, we need to identify which is each cell type. We do this using marker genes.

- I go back to the full matrix of genes
- Make a statistical comparison (Lima, Wilcoxon Test, T-test…). Statistical test to compare counts between conditions. The conditions are cells of cluster 1 against the cells of the other clusters.
- We find which are the genes that are Top highly enriched in each cluster.
- Imagine that the top gene marker in cluster 1 is insulin. Then, maybe the cells correspond to the pancreas.

6. **Trajectories (pseudotime)**

We can organize the cells based on differentiation trajectories, calculating the pseudotimes.

7. **Integration with other datasets (even with other data modalities)**

Finally, we can use multimodal data integration.

ATAC-seq

We will talk about chromatin
What is the ATAC-Seq technology
Look at differentially expressed regions
Gene activity → Measure that summarize the regulatory activity that surrounds genes
Compute different modalities of data

We have been seeing that chromatin is structured around nucleosomes. Nucleosomes are not randomly distributed around the genome and there is a correlation between the presence of nucleosomes and gene activity.

So, by looking at the presence of nucleosomes we can know how accessible the gene is and therefore we can know if TF can bind or not to regulate the transcription.

How can we profile genome accessibility genome wide? We use a number of techniques such as DNA-seq, RNA-seq, DNA-sensitive-seq and ATAC-seq.

They are all based on the following principle: If the genome region is accessible, I can cut it (DNAsa) and sequence it with NGS. Then we just map the reads on the reference genome and we can see which are the regions that are accessible and therefore there are no nucleosomes.

Alternatively, you can do the complete opposite. Precipitate the nucleosomes and find the regions that are not accessible.

Today, it is very popular ATAC-seq because:

- It requires much less input DNA. As we remember, in single cell we have small quantities of DNA, thus it is logical to choose this technique.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a widely used technique in genomics research to study the accessibility of chromatin regions in the genome. It provides insights into the regulatory regions, such as promoters and enhancers, that control gene expression. ATAC-seq combines the principles of chromatin immunoprecipitation (ChIP) and DNA sequencing to identify open chromatin regions.

Here's a step-by-step overview of how ATAC-seq works:

1. **Cell Lysis:** The first step is to lyse the cells of interest, typically by using a hypotonic buffer, to break open the cell membranes and release the nuclei.

2. **Transposition:** A transposase enzyme is added to the isolated nuclei. The transposase used in ATAC-seq is typically Tn5 transposase, which has been modified to have both transposase and DNA fragmentation activity. The transposase binds to accessible regions of chromatin and inserts sequencing adapters into the DNA.

3. **DNA Purification:** Following transposition, the DNA is purified using various techniques, such as phenol-chloroform extraction or commercial purification kits. This step removes proteins and other contaminants.

4. **PCR Amplification:** To increase the amount of DNA for downstream sequencing, a limited-cycle PCR amplification is performed using primers that target the sequencing adapters inserted by the transposase. This amplification step selectively enriches the DNA fragments with the inserted adapters.

5. **Library Preparation:** The PCR-amplified DNA fragments are then prepared into a sequencing library. This involves processes such as DNA fragment size selection, end repair, adapter ligation, and library amplification.

6. **Sequencing:** The prepared library is subjected to high-throughput DNA sequencing using next-generation sequencing platforms. The DNA fragments are typically sequenced using a paired-end approach, where both ends of the DNA fragments are sequenced.

7. **Data Analysis:** The generated sequencing data is analyzed to identify regions of open chromatin. This involves aligning the sequencing reads to a reference genome, removing duplicates, and calling peaks or regions with enriched read coverage. Peaks represent accessible chromatin regions, which often correspond to regulatory elements such as promoters and enhancers.

By identifying open chromatin regions, ATAC-seq allows researchers to gain insights into the regulation of gene expression and study changes in chromatin accessibility between different cell types, developmental stages, or disease conditions.

If we sequence enough, we can detect drops of coverage that represent the TF bound to the genome. This is called TF footprinting. We find coverage but not very high because for a period of time there was a TF bound there and therefore the DNA was not accessible.

We will also see something called insert size periodicity of 147bp → The distance between mapped paired-end reads shows a periodicity of 147 bp, which suggests that the transposase preferentially inserts the sequencing adapters at sites that are approximately 147 bp apart from each other within the open chromatin regions.

There are many methods to call peaks:

- Methods based on the counts → Macs2 uses a model-based approach to distinguish real signal from background noise, taking into account the local biases and characteristics of the data. It uses a Poisson distribution to model the background noise
- Methods based on HMM
- Methods based on the shape

We know that different regions of the genome, because of different reasons (GC composition), have different values of mappability.

To know exactly where the transposase has made the cut, we need to take into account that this enzyme introduces a 9 bp gap. This is important for TF footprinting, because it does really matter this 9 bp gap (motifs are 6 bp). So, we need to subtract 4 bp from the left and 5 bp from the right.

How many counts do I expect for a given region if I do ATAC-seq? If I take an exon and I do single-cell RNA-seq the number of counts will depend on the level of expression of that gene. In the case of ATAC-seq we can get a maximum of 2 counts:

- In diploid organisms we have 2 copies of DNA. So you either read one or read the other.
- We normally binarize it, meaning that it is detected or not.

Pseudo-bulks: Aggregate the expression profiles of individual cells (aggregate the counts).

When we do single cell RNA-seq, one common pipeline is to use the CellRanger (makes the mapping, summarizes the matrix of counts…). For ATAC-seq, there is another pipeline of CellRanger that also gives you the peaks, the count matrix…

- The way CellRanger calls peaks is a little bit different from Macs2. But you can still use Macs2.

We also have the knee plot to decide how many cells you keep based on the number of fragments per read.

QC

- In single cell, for each cell we will also do the periodicity.
- We can also look at the enrichment of reads in the promoters (around the TSS). We must see many cuts of the transposase.
- Fractions of reads that are associated to peaks. If something went wrong, reads are going to be scattered everywhere. Else, reads should only be found in peaks.

Dimensionality reduction

Now we need to make a dimensionality reduction to later do clustering.

We just do a PCA or better to use an analysis of words from texts (single value decomposition). You calculate 2 terms:

- Term frequency
- Inverse document frequency to normalize the data. So we are promoting peaks that are found only in a subset of cells.

Instead of obtaining PC, we obtain LS size.

Then, we can do a UMAP embedded in 2D as we did in the other case.

To do differential accessibility, we need to identify clusters (using Louvain algorithm) based on chromatin accessibility. We want to know what is specifically accessible in a cluster compared to the other clusters.

Proteomics

We know the dogma of biology: DNA → RNA → Protein → Phenotype

The proteome is the set of proteins present in a particular cell or tissue, under defined conditions and at a given time.

Genome and Complexity

The so-called C-Value Paradox refers to the observation that genome size does not uniformly increase with respect to perceived complexity of organisms.

The so-called G-value Paradox refers to the observation that number of genes does not uniformly increase with respect to perceived complexity of organisms.

Complexity is proportional to the level of regulation and splicing mechanisms of an organisms. For example, in the case of humanas:

- We have 22.000 genes
- Due to splicing, mRNA editing… we obtain a total of 200.000 transcripts
- Due to PTMs, we obtain more than 1.000.000 proteoforms.

So, at the end, due to our regulation we will have a higher complexity than other organisms that have a higher genome size or more genes.

Differences Between Protein Chemistry and Proteomics

Protein chemistry	Proteomics
• Individual proteins	• Complex mixtures
• Complete sequence analysis	• Partial sequence analysis
• Emphasis on structure and function	• Emphasis on identification by database matching
• Structural biology	• Systems biology

What can we do with proteomics?

- **Protein Mining:** Identification of many different proteins in complex samples.
- **Protein Expression Profiling:** Comparison of protein abundance level under determined conditions (i.e. search for protein candidate for biomarkers of diseases).
- **Protein Network Mapping:** Approach to look at the interaction between proteins from different systems.
- **Mapping of Protein Modifications:** Characterization of post-translational modifications and site mutation localization.

In order to characterize a sample using proteomics, we must do the following 3 steps:

- Separation and selection of target proteins.
- Make a digestion and measure of peptide masses. Since we are identifying based on the mass, not on the sequence.
- Comparison with available databases

Separation methods

We are going to isolate putative proteins using:

- 2D-SDS gel electrophoresis
- High Performance Liquid Chromatography (HPLC)

Bidimensional electrophoresis

Proteins will be separated according to their isoelectric point (first dimension) and mass (second dimension).

Isoelectric point: pH at which net charge is 0

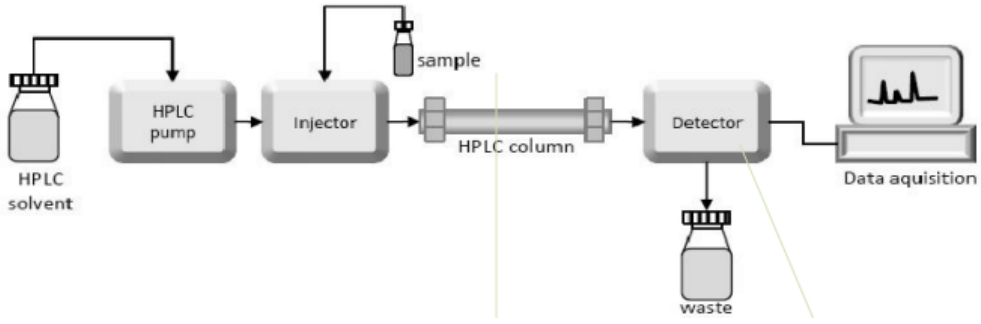
So, we first separate using the different isoelectric points of the proteins. To do this, we use Ampholytes, which establish a pH gradient across the gel that allows proteins to migrate to their isoelectric point during the first dimension of electrophoresis.

We can use softwares that makes comparisons of two gels. Thus, it allows quantification of protein expression through the spot itself.

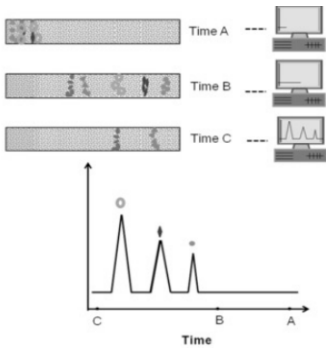
Note that there are some proteins that are preserved better during the time than other proteins. Thus, if you want to know which protein of the brain lasts longer after someone has died, you can do a 2D gel and see the results in 2h, 10h, 50h…

HPLC

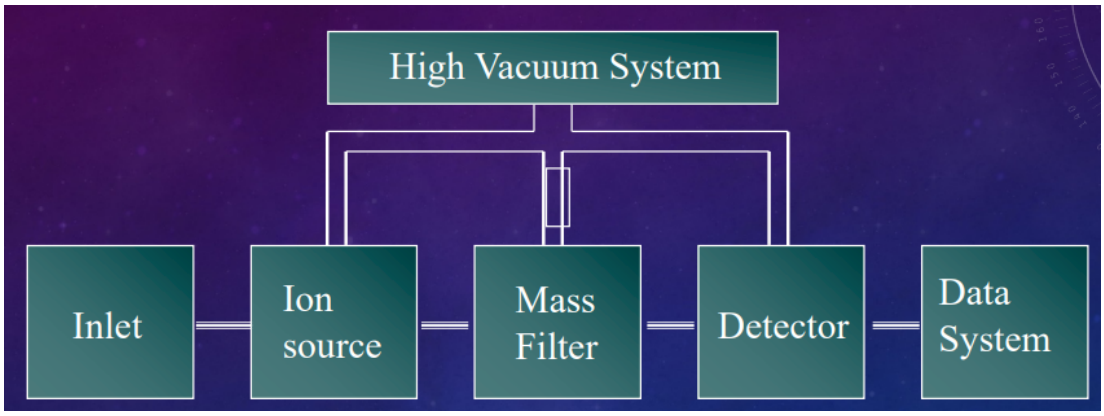
We are separating based on chemical properties (ionic charge, hydrophobicity). We throw the sample into the column and each component will be differently adsorbed in the HPLC column filled with granular solid material.



There are different types of columns and each protein will exit the column in different times or rates, based on its chemical properties.



Mass Spectrometer



Ion source: Way of charging the proteins.

- Volatile gas
- Electrospray
- MALDI

Mass Filter: Select the proteins (by mass) that you want to separate.

- TOF
- Quadrupole
- Ion trap
- Hybrids

Detector: Will detect how fast the protein comes, calculating the mass/charge ratio

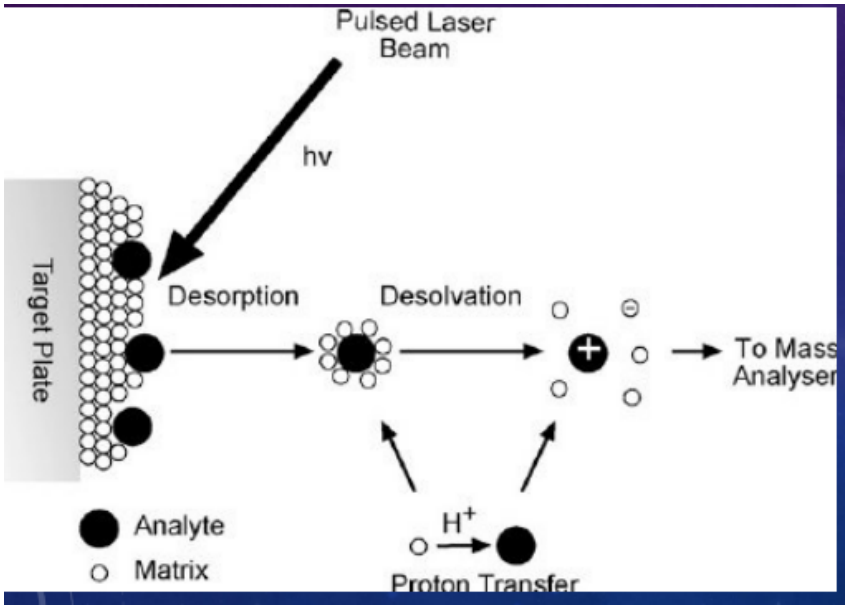
- Multi-channel plate
- Ion cyclotron

MALDI (Matrix Assisted Laser Desorption Ionization)

The molecules are ionized by proton transference in a gas phase.

Our proteins are embedded in a matrix that is going to be irradiated with a pulsed laser beam. The energy of this laser is absorbed by the matrix molecules, causing them to vaporize and release energy.

As the matrix molecules vaporize, they carry the biomolecules with them into the gas phase. The intense laser pulse desorbs and ionizes the biomolecules, creating a cloud of ions.



An electric field is applied to accelerate the ions towards a detector. The time it takes for the ions to reach the detector depends on their mass-to-charge ratio (m/z). The detector measures the arrival time of ions, allowing for the determination of their mass-to-charge ratio.

The mass-to-charge ratio data obtained from the detector is processed to create a mass spectrum. The spectrum represents the distribution of ions based on their mass-to-charge ratios, providing information about the molecular weight and composition of the biomolecules in the sample.

Electrospray

The sample is in acidic liquid and flows through a capillary under high voltage. It perfectly couples to the HPLC since the sample is in liquid phase.

Due to the high voltage, the liquid jet is exposed to a strong electrostatic field. This causes the liquid jet to disintegrate into smaller droplets.

As the droplets move through the air, the volatile solvent quickly evaporates, leaving behind charged droplets containing the biomolecules. The droplets become progressively smaller due to the evaporation process.

As the charged droplets shrink in size, the repulsion between like charges becomes stronger. This eventually leads to the droplets breaking apart into smaller droplets known as microdroplets.

This leads to the formation of ions.

TOF (Time Of Flight)

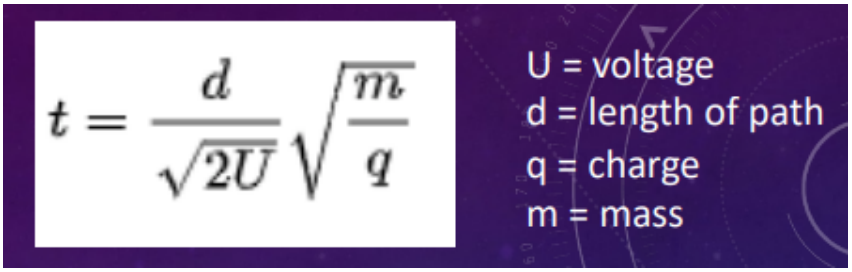
The time each molecule takes to reach the detector it's proportional to its mass.

- Smaller mass, higher velocity

It is also proportional to the distance (but the distance is the same for each molecule)
It is also inversely proportional to the charge.

- Higher charges, higher velocities

By measuring the flight time of ions and knowing the distance traveled, the mass-to-charge ratio can be calculated, allowing for the determination of the mass of the ions.



QUADRUPOLE

A quadrupole consists of four parallel metal rods, typically arranged in a square or rectangular configuration. The rods are electrically charged, with adjacent rods having opposite polarities.

The ions are then accelerated into the quadrupole region and they will oscillate (they are attracted then repelled…). Depending on the mass, the oscillation is going to be so big that the molecules will abandon the quadrupole.

So, for a given ratio of voltages between rods only ions with a certain mass/charge ratio will reach the detector. Keep changing voltage configuration to select different ranges of mass/charges values.

Ion Traps

Select specific ions and fragment them in the same space. Quadrupole requires different spaces.

Ions get traps into and oscillating saddle field
Select ions by mass/charge.
You can also fragment the ion and sequentially eject the product to the detector

The benefit of Ion Traps is that you can fragment the ions, which is the next step.
So, you fragment each protein into peptides and you detect the TOF of each peptide, obtaining the mass/charge ratio of each peptide. This will be the fingerprint of my protein.

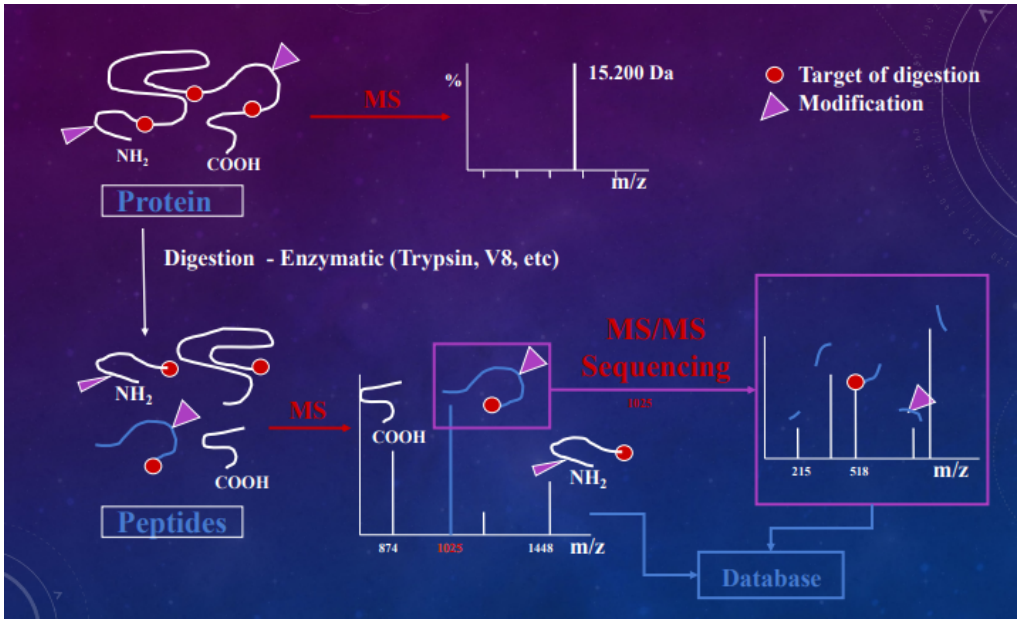
Fragmentation or digestion

If I do not fragment my protein and I launch it directly to the quadrupole and they go to the detector, my spectrum will say that my protein weights 15.2 Daltons.
But many proteins have this weight and we will not be able to identify it.

For this reason, to identify this protein, we need to fragment this protein using trypsin (for example) and then obtain the “peptide mass fingerprint” (PMF) of those peptides.

The profile obtained can be compared in a database and find which proteins have this specific profile.

I still can do another mass spectrometry (MS/MS) → So, I couple a Quadrupole or TOF to another quadrupole and I only fragment a specific peptide. I will obtain a profile of that peptide and I will know which is the sequence of the peptide.



Why do we digest proteins?
We digest the proteins because the bigger the molecule, the biggest is going to be the error. So, the smaller the mass/charge ratio, the better.

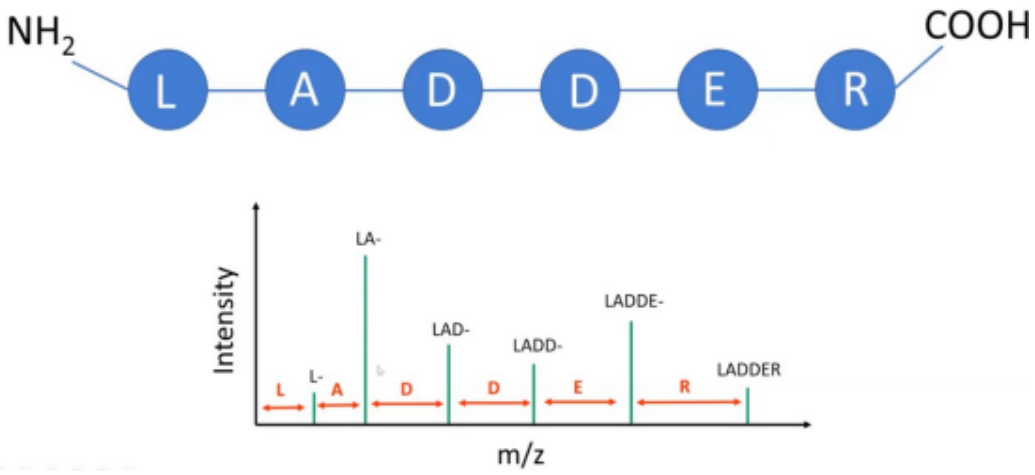
Also, membrane proteins (hydrophobic and large) will never be charged. If I have cytosol proteins, they will charge much better. So, depending on the nature of the protein, we will have different results.
If I cut them in peptides, I will not have this problem.

The sensitivity of measures for intact protein masses is not as good as sensitivity for peptide masses.

I will make a digestion to obtain an optimal size of 6 to 20 aa.

MS/MS

We go further in the digestion. We select a peptide and digest every peptide bond (not using trypsin). In the mass spectrometry, we will see peaks that do not correspond to individual AA fragments, but fragments of different lengths, compatible with a given sequence of AA.



Once you have the sequence of the peptide, you can go to the DB and know to which proteins contain this peptide. We do a BLAST with a NR database.
We will have a high False Discovery rates.

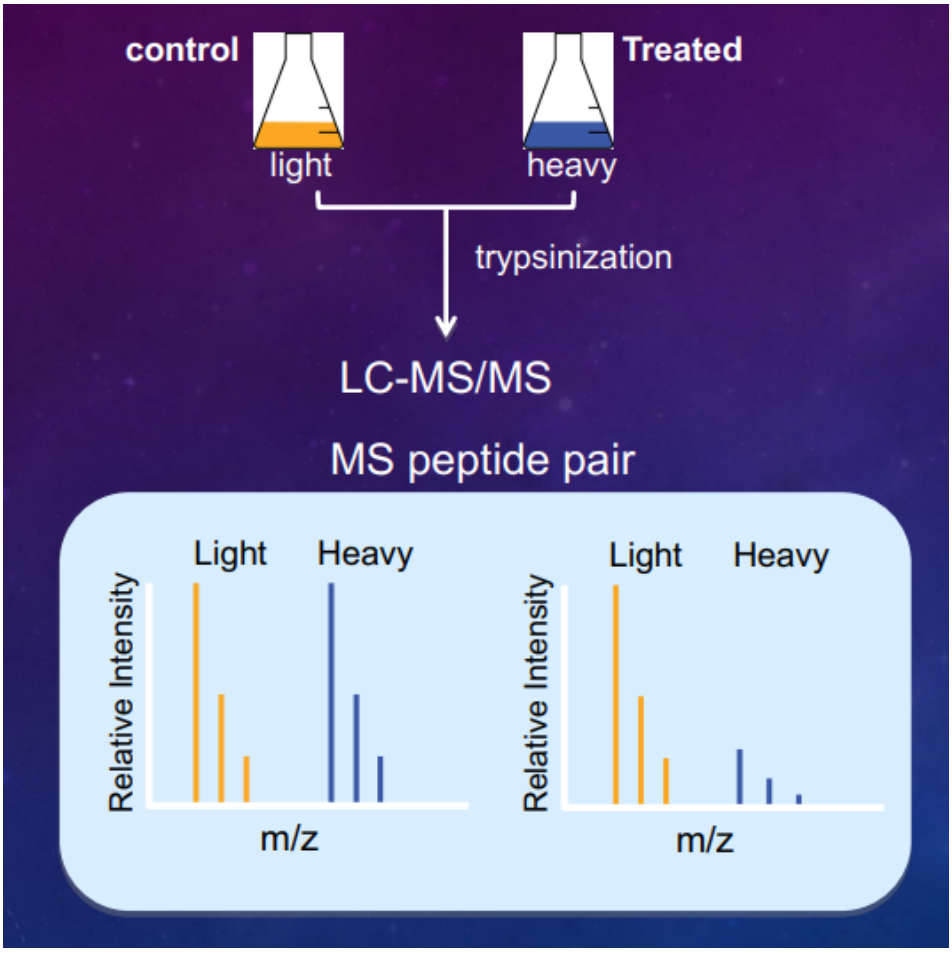
- If the DB is small, you won't find proteins because maybe they are not contained in the DB
- If the DB is very big, you won't find it because of the false discovery rate (they say it's the protein but it is not).

We can also use MS to know all the possible PTM of a residue.

Protein quantification by MS. Stable isotope labeling

Isotopes are the same element but with different numbers of neutrons.

If you have an isotope labeled sample (weight 15), you can compare it to a non-isotope labeled sample (weight 14).
When doing the quantification, we will see that the peaks are shifted to the right. Thus, you can quantify the 2 samples.



When we query the whole proteome of a sample, how is it compared to the transcriptome? So, is the transcriptome predictive of the proteome? Is there a correlation between gene expression and proteome?

Some genes correlate very well, but for many genes there is no correlation.

Proteomics is not as sensitive as transcriptomics. We can only analyze the genes that are very highly expressed.

Write a definition of epigenetics.
Epigenetics is the study of heritable phenotypic changes that do not include DNA alterations. Often involves changes that affect gene activity and expression. Phenotypic changes are the result of the environment.

How would you expect to find the promoter region of a highly transcribed genes in terms of nucleosome positioning, DNA methylation and histone modifications?
Highly transcribed genes correspond to euchromatin. Nucleosomes will have low occupancy, DNA demethylated and histone acetylated.

Which is/are the chromatin state/s most highly associated with the following histone modifications:
H3K4me3 - Active promoters
H3K27ac - Active promoters and enhancers
H3K36me3 - Transcription (Tx) elongation
H3K9me3 - Heterochromatin
H3K27me3 - Repression state

Which genomics elements contain mostly methylated CpG sites?
CG islands contain methylated CpG sites. 70 % of proximal promoters are in CG islands. 60 % of genes have GC islands

What is a TAD?
A topologically associating domain (TAD) is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD.

Describe one technique to study chromatin interactions (3-C, 4-C, 5-C, Hi-C or GAM).
ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

What is the G-value / C-value enigma and what explanation would you provide to solve it?
Lack of correlation between the number of protein-coding genes among eukaryotes and their relative biological complexity.
The reason is found in the RNA world, which is more complex and it is related to gene regulation.

Which protein property is used to separate each dimension in a 2D-SDS-PAGE electrophoresis?
Isoelectric point (pH) and weight.

Pair with arrows:
Ion Source → Electrospray, MALDI
Mass filter → TOF, Ion Trap, Quadrupole

Describe the purpose of the second MS step in an MS/MS applied to protein identification.
Ions from the MS1 spectra are selectively fragmented and analyzed by a second MS stage to generate the spectra for the ion fragments.

Name three categories of epigenetic modifications.
- DNA-Methylation
- Histone modification
- Nucleosome positioning
- Non-coding RNA

Methylation at CpG sites:
a) occurs at similar extent in all organisms.
b) is always associated to silencing of gene expression.
c) is irreversible
d) none of the above

Rewrite the following terms in hierarchical order:
Nucleosomes, A/B compartments, FIREs, TADs, chromosome territories.
Chromosome territories > A/B compartments > TADs > FIREs > Nucleosomes

Why we digest proteins to peptides before MS instead of running the whole molecule?
The problem of MS is: the higher the mass, the higher the error. So, when running the full molecule with MS, we will get a single peak (the mass of the whole protein), but that does not mean that it is trustworthy. In order to have more reliable results, it is highly recommendable to break the protein into fragments, run MS and even break those peptides and run MS again to better analyze our samples.

Explain why and how isotope labelling can be used to quantify relative protein amounts among conditions.

What is the Jaccard similarity and how do we calculate it using genomic coordinates?
The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, with a range from 0% to 100%.
We can use it to calculate two sets of binding sites.

What solution do you know it is used in single-cell genomics to deal with the problem of PCR duplicates?
Using UMI, unique molecular identifiers

If you can only work with frozen tissue, would you use single-cell or single-nuclei RNA-seq? Why?
Single - nuclei RNA-seq, You can extract nuclei from frozen tissue, but not entire cells whose membranes break when frozen.

Mention three advantages of using single-cell genomics versus bulk genomics.

- Study Rare Cell Types Obscured In bulk tissue
- Determine trajectory of differentiation
- Identify Cell Type Specific Effects Under a Comparison(treatments, diseases, evolution, etc)

What kind of information is provided by ATAC-seq?

ATAC-seq can be used to get some mixed samples directly from the environment and study the expression of the genes in it. With it, we can evaluate which genes are expressed in a certain tissue-specific cell type and use clustering approaches to understand the expression resemblances among different types.

What is the Louvain algorithm designed for in the context of single-cell?

Louvain algorithm helps to identify "neighboring" cells and therefore, cells that may form a cluster (in other words, cells that share some gene expression characteristics and consequently have similar properties and functions). Some cells interact more with some other cells than others. Therefore, it is ideal to identify which are more associated between them and which are not. Louvain algorithm calculates the QC and in each step it checks that this value is improving or not:

- $\Delta QC > 1 \rightarrow$ the last change improves the network, it should be considered
- $\Delta QC < 1 \rightarrow$ the last change does not improve the network, it should not be considered

Which region of the genome is particularly useful and used for characterizing microbiomes richness and why?

The region of the genome that is particularly useful and commonly used for characterizing microbiome richness is the 16S rRNA gene.

The 16S rRNA gene is highly conserved among bacteria and archaea, but it also contains variable regions that can be used to differentiate between different microbial species and genera. These variable regions allow researchers to classify and identify microorganisms based on their genetic sequences.

Specific regions within the 16S rRNA gene, such as V1-V9 regions, can be targeted using universal primers that are designed to amplify the gene across a wide range of microorganisms. The 16S rRNA gene sequences can be compared to reference databases to assign taxonomic classifications to the microorganisms present in the sample.

What is an OTU?

Operational taxonomic unit

Clusters of (uncultivated or unknown) organisms, grouped by DNA sequence similarity of a specific taxonomic marker gene

In a 16S rRNA next generation sequencing effort to determine microbial richness from a sample, can we consider a difference in one single nucleotide position between two sequences as evidence of the presence of two species? Give reasons for you answer.

No, a difference in one single nucleotide position between two 16S rRNA gene sequences is not sufficient evidence to conclude the presence of two distinct species.

The 16S rRNA gene exhibits natural variability even within a single species. This variation can arise due to genetic polymorphisms, strain-level differences, or sequencing errors. Therefore, a single nucleotide difference alone does not necessarily indicate the presence of two separate species.

Next-generation sequencing techniques used for 16S rRNA gene analysis can introduce sequencing errors or biases, which may result in apparent nucleotide differences that are not biologically meaningful.

How would you construct and interpret a rarefaction curve from next generation sequencing metagenomics data?

A rarefaction curve is a graphical representation used to explore the richness and diversity of species in a microbial community based on next-generation sequencing metagenomics data. It helps estimate the number of unique species detected as a function of sequencing effort or sample size. Here's how you can construct and interpret a rarefaction curve:

Data Preparation: Start with your next-generation sequencing metagenomics data, which includes information on the abundance and diversity of microbial species in your samples.

Sampling Effort: Determine the number of sequences or reads you will consider at each step of the rarefaction curve analysis. This step aims to simulate different sequencing depths or levels of sampling effort.

Subsampling: Randomly select a subset of sequences from your data, ensuring that the number of sequences chosen corresponds to the desired sampling effort or read count for that step of the rarefaction curve. Repeat this subsampling process multiple times to generate an average result.

Species Accumulation: Calculate the number of unique species or operational taxonomic units (OTUs) observed at each subsampling depth. An OTU represents a taxonomic unit used to group similar sequences, often based on a predefined sequence similarity threshold.

Plot Construction: Plot the number of observed OTUs on the y-axis and the cumulative number of sequences or reads on the x-axis. Each point on the rarefaction curve represents the average number of OTUs detected at a specific sequencing depth.

Interpretation: Analyze the rarefaction curve to gain insights into microbial richness and diversity:

a. Early Slope: Initially, the curve tends to have a steeper slope, indicating a rapid increase in the number of detected OTUs as more sequences are analyzed. This suggests that there are many rare or low-abundance species that become evident with increased sampling effort.

b. Plateau: As the sequencing depth increases, the slope of the curve gradually levels off, approaching a plateau. This plateau indicates that the majority of common or abundant species have been detected, and additional sequencing effort is less likely to reveal many new OTUs.

c. Comparison: Rarefaction curves can be compared between different samples or experimental conditions. A higher curve indicates greater microbial diversity or richness, while a lower curve suggests lower diversity.

d. Sampling Optimization: Rarefaction curves help in determining the optimal sequencing depth required to capture the majority of species diversity within a given dataset. Researchers can assess the point of diminishing returns, where additional sequencing effort provides minimal gains in detecting new species.

It's important to note that rarefaction curves provide a snapshot of species richness based on the analyzed data. The interpretation should consider potential biases introduced during sequencing, such as PCR amplification biases, primer selection, and sequencing errors. Additionally, rarefaction curves are influenced by the specific clustering or OTU definition method used.