Linear Model+GLM

Topic 1. DNA-seq techniques

Why sequence a genome:

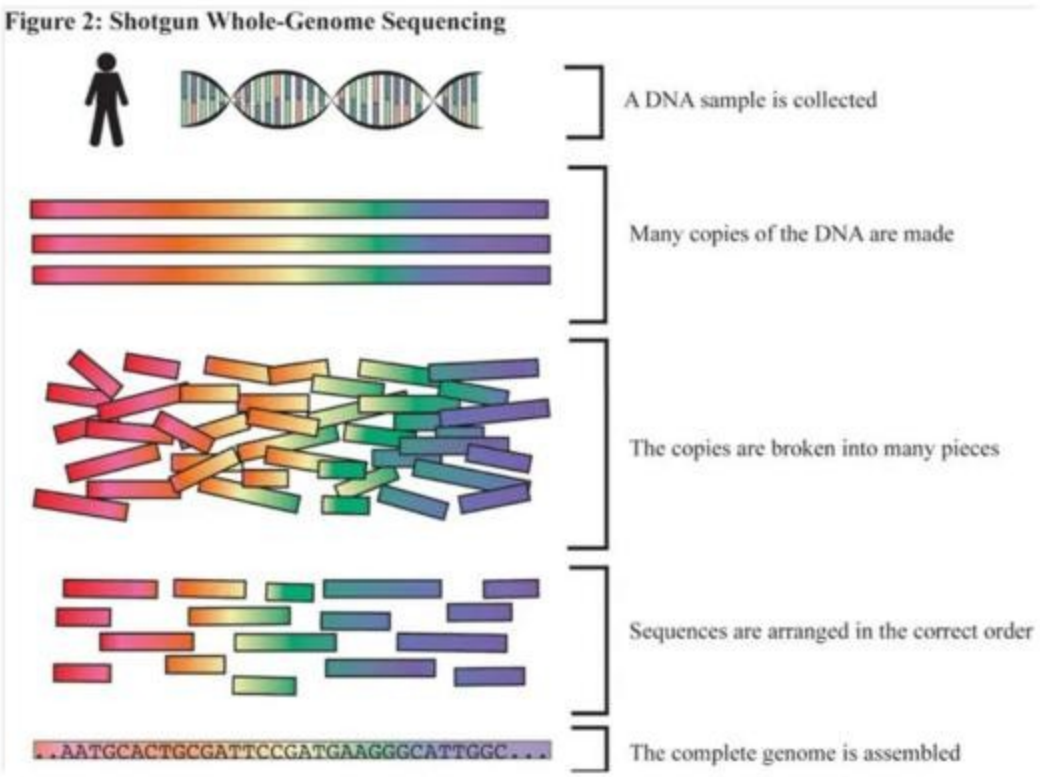
- Description of sequence of every gene valuable. Includes regulatory regions which help in understanding not only the “biochemical” activities of a cell but also ways in which they are controlled.
- Identify & characterize important inheritable disease genes or “useful” functional genes (e.g. bacterial genes for industrial use).
- To understand relationships between organisms and provide information on how they evolve.

Sequencing technologies

Perfect situation: We have a chromosome with 2 chains, we take one of them and we obtain its complete sequence in a single piece by simply passing this chain through a pore.

This does not happen.

We first obtain the DNA, make many copies of this DNA and then we break it into little pieces. Then we rearrange them in the correct order and we make an assembly.



First generation → Sanger (1977-1990)
It is still used but it is not very common.

We infer nucleotide identity using dNTPs then visualize with electrophoresis.
500-1000 pb fragments.

Here we are sequencing gene by gene, so it has a low throughput.

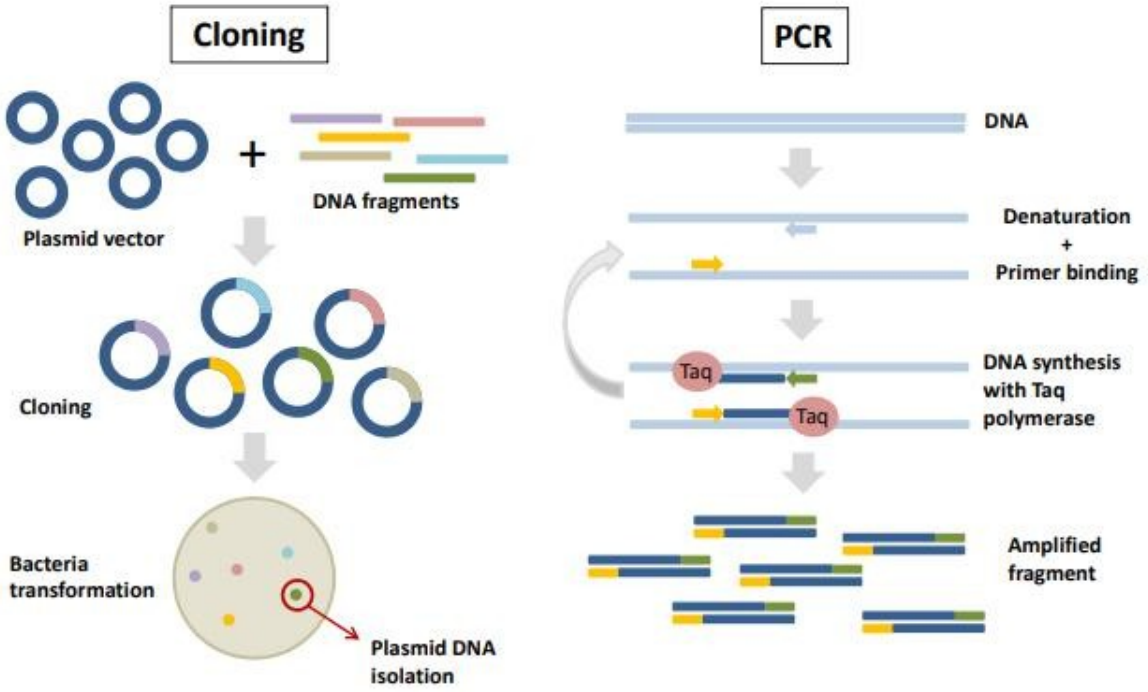
Second generation → 454, Illumina, SoliD, Ion Torrent (2005-2010)
High throughput from the parallelization of sequencing
reactions 50-500 bp fragments

Third generation → PacBio, NanoPore (2011-present)
Sequence native DNA in real time with single-molecule
resolution Tens of kb fragments

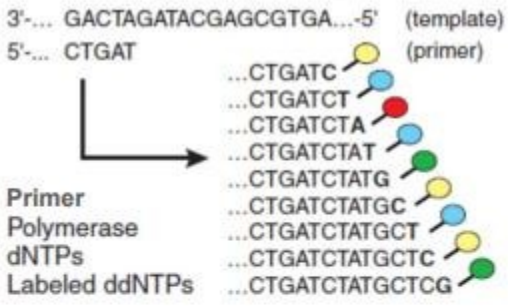
Sanger sequencing method

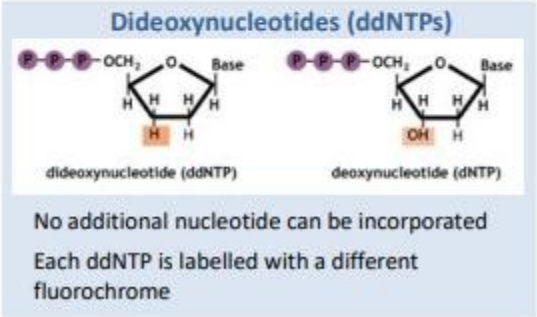
DNA Amplification: We start with a small fragment that we need to amplificate. We can do this in 2 different ways:

- **Cloning:** These DNA fragments are introduced in a plasmid, this vector is transformed in a bacteria and then we grow the bacteria. Finally we select the bacterias in a colony that contain our fragments.
- **PCR:** We need to design the primers, add the nucleotides, change the temperature…

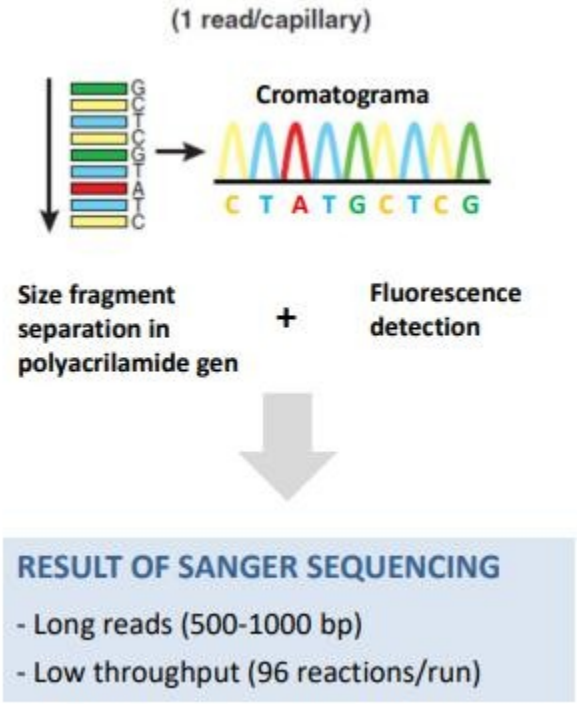


Sequencing reaction: We use labeled ddNTPs that stop the reaction of adding nucleotides.





Capillary electrophoresis: To read the results



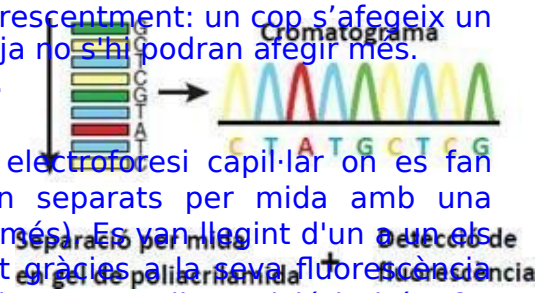
The first nucleotides have a very low signal and therefore we need to trim them.
The last nucleotides also have very low signal and we also trim them.

Mètode de sanger

1. El primer pas és crear una genoteca genòmica, fragmentem el DNA de l'organisme que volem seqüenciar i seleccionem els inserts de més o menys la mateixa mida i els insertem en un vector. Aleshores ho transformem en bactèries per tal d'amplificar. Primer pas = amplificació del DNA per clonació o PCR.

2. Extraiem el DNA de les bactèries i calentem a 90 graus per separar les dues cadenes de DNA. Refredem fins a 50 graus per permetre que s'uneixi la polimerasa i que comenci a seqüenciar fins que es trobi posi un ddNTP. Ara es torna a calentar fins a 90 graus per tornar a separar les cadenes.

- Primer
- Polimerasa: afegirà nucleòtids fins que trobi un ddNTP
- Nucleòtids
- Didesoxinucleòtids (ddNTPs) marcats fluorescentment: un cop s'afegeix un en la seqüència, s'atura la seqüenciació perquè ja no s'hi podran afegir més. Estan marcats de colors amb grups fluorocroms.



3. Després s'agafa la mostra i es fa una electroforesi capil·lar on es fan córrer els diferents fragments, que quedaran separats per mida amb una diferència d'un nucleòtid (els més petits corren més). Es van llegint d'un a un els últims nucleòtids col·locats i es van identificant gràcies a la seva fluorescència (cromatograma) per interpretar quina base hi ha en aquella posició i així refer tota la seqüència.

4. Finalment s'obté una lectura (read) per a cada reacció de seqüenciació d'entre 500-1000 pb → 1 read per cada carril.

- *Els fragments que tenen moltes bases ja costa més identificar-los perquè els pics són més difusos en el cromatograma.*
- *Com la fracció de didesoxinucleòtids és molt petita, hi ha pocs fragments en els que s'aturi la reacció al començament, per això els primers pics també són difusos*

És una tècnica costosa i laboriosa, però permet fer 96 seqüències cada vegada que fem córrer la màquina.

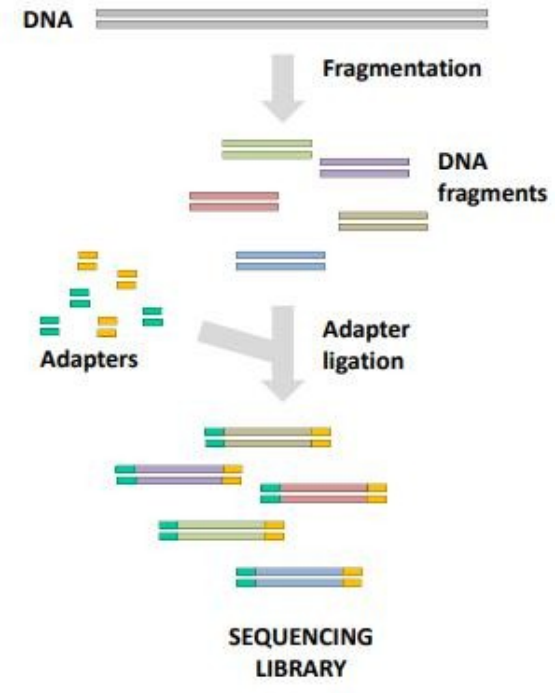
Limitations of classic sequencing techniques

- The main limitation of both these classical sequencing techniques is their low throughput, due to template preparation and in the case of Sanger Sequencing also to carry out the enzymatic reaction.
- Each run in Sanger Sequencing can sequence up to 1000 bp, and with an automated sequencer 384 sequences can be run in parallel with a throughput of 80–100 kb per hour.
- Due to its singleplex nature, Sanger Sequencing is not a hardly scalable process. In 1985, reading a single base cost $10, while in 2005, the various improvements reading 10,000 bases cost the same. However, large projects such as the Human Genome Project still required vast amounts of time and resources.
- Another limitation of First Generation Sequencing is that variants present at low frequency, such as mosaics, are difficult to detect due to high background levels.
- Finally, compared with modern technologies, the cost per base is still high

Next generation sequencing (NGS) methods can generate as much data in one day as several hundred Sanger DNA capillary sequencers.

Common characteristics of NGS methods

1. Cell-free preparation of sequencing library (fragmentation + adapters)



2. Solid-phase amplification
3. Massively parallel sequencing reaction of each DNA fragment independently
4. Direct sequencing without need of electrophoresis

Massively parallel sequencing

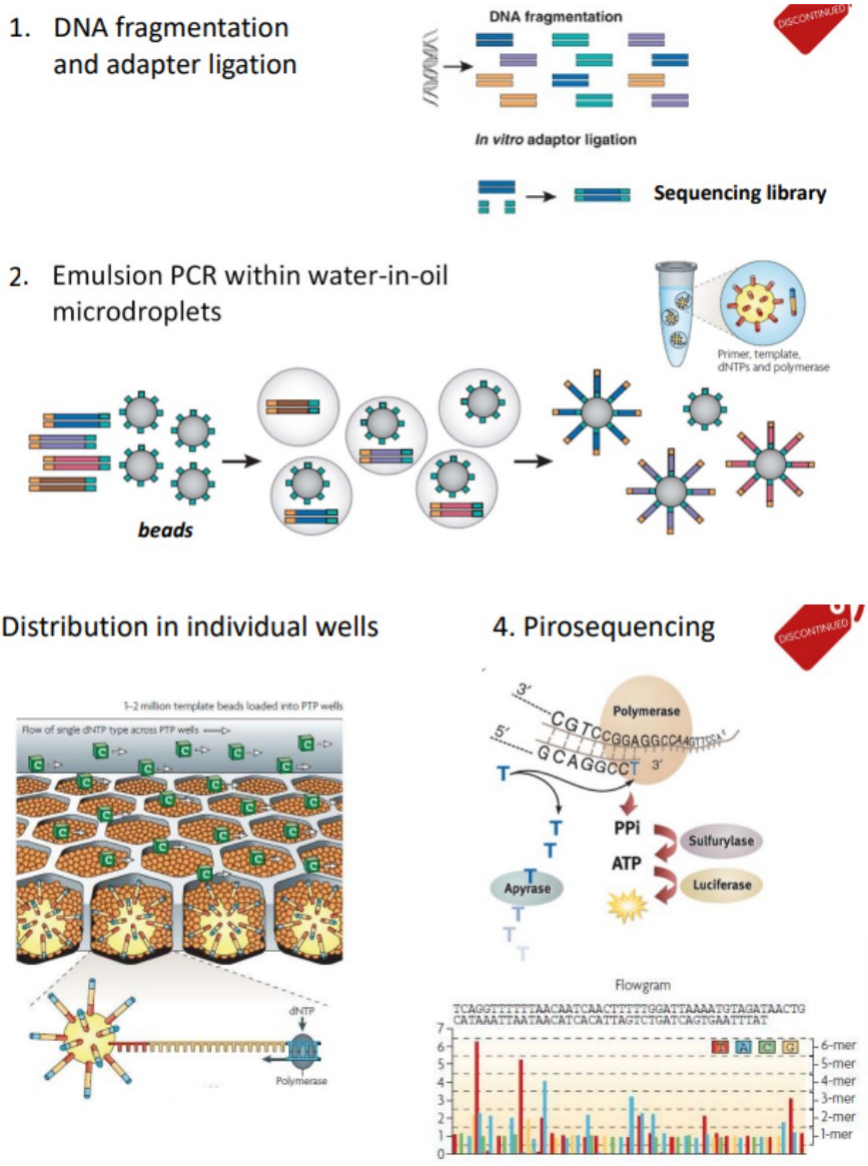
Tots aquests mètodes tenen en comú que també requereixen una genoteca genòmica però en aquest cas ja no es clona en vector. S'extrau el DNA, es fragmenta i se seleccionen els fragments d'una mida determinada. Després, s'afegeixen uns adaptadors en els extrems que no es degraden i serveixen per a la posterior amplificació en fase sòlida i seqüenciació (per PCR).

Estem seqüenciant de forma massiva, però d'un nucleòtid en un.

Roche 454 - Pyrosequencing

The difference is in the way they do the amplification.
In this case, they use an emulsion PCR within water-in-oil microdroplets.

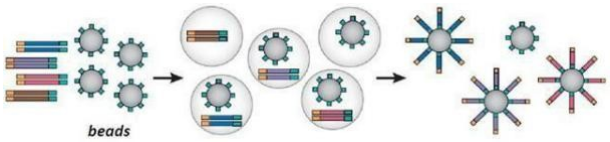
Its also different how they identify the amplification. They use pyrosequencing.
The process is a little bit slow because you need to add one type of base at a time.
The problem when we have homopolymers.



454/ Roche – Piroseqüenciació

1. Es parteix del DNA genòmic fragmentat, seleccionat per mida, que presenta els adaptadors lligats.

2. L'amplificació es fa per PCR en una emulsió de microgotes d'aigua dins una fase oliosa. S'afegeixen les seqüències de la genoteca dins de l'aigua i unes boletes (*beads*) que contenen els adaptadors que serviran de *primers* per amplificar i seqüenciar, s'afegeix també polimerasa i nucleòtids. Després es barreja la solució en oli fins formar unes micel·les d'aigua que tenen una mida determinada perquè en cada una hi vagi a parar: polimerasa, boletes, una de les seqüències i nucleòtids (elements necessaris perquè es porti a terme la seqüenciació). Com la formació de les micel·les és aleatòria, no en totes elles hi haurà els elements necessaris, però com es formaran moltes gotetes, ja seran suficients per permetre seqüenciar. **És important que en cada gota hi vagi un sol fragment**.



3. Finalment es distribueix cada gota (micel·la) en un pou individual d'una placa.

4. En aquesta placa té lloc la reacció de piroseqüenciació: es van passant els diferents nucleòtids successivament (primer es passen *citocines* (C) per tots les pouets i només reaccionaran els que necessitin *C* en la seva seqüència), de manera que, aquells pous que hagin afegit aquell nucleòtid, **emetran llum** (luciferasa) que serà detectada i es farà una foto i se sabrà quins pous han afegit aquell nucleòtid. Seguidament s'ha de fer un rentat per eliminar els nucleòtids restants en el pou.

   Es va repetint l'experiment amb diversos nucleòtids fins obtenir el patró de nucleòtids que s'han anat afegint en la seqüència de cada pou. El resultat final és un flowgrama que indica, per cada pou, la quantitat de llum emesa en fer passar cada nucleòtid. Anirem fent cicles de nucleòtid per nucleòtid. Estarem fent molts flowgrames alhora (1 flowgrama per pouet).

   Si hi ha més d'un nucleòtid igual successius, quan es fa passar aquest nucleòtid sobre la placa s'afegiran **tots de cop** i en el flowgrama es detectarà un pic de lluminositat més intens/gran. Això és un problema, perquè quan hi ha un **homopolímer** en la seqüència costa molt detectar quants nucleòtids iguals s'han incorporat (no se sap exactament el nombre de nucleòtids iguals que s'ha afegit).

- FragmentacióDelDNAilligamentd'adaptador
- AmplificacióenfaseoliosaperemulsionPCR
- Distribució en pous individuals
- Piroseqüenciació → flowgrama

Ion Torrent

It also uses a PCR within water-in-oil microdroplets but in this case we do not use pyrosequencing. There is a real time sequencing by using a semiconductor plate to count proton release during DNA synthesis.
So, we are calculating the change of pH.

Normal nucleotides (not labeled) flow sequentially through the chip.
The incorporation of one nucleotide released one proton (pH change). This is detected by the semi-conductor plate, which converts the chemical information into digital information. No optical machines are needed (no scanning, fluorescence, laser excitation, …).

It still has the homopolymer problem.

És una variant de la tècnica del 454: la diferència és que la placa on se seqüència es detecta canvis de pH i no s'utilitzen nucleòtids marcats.

1. Necessitem una genoteca de seqüenciació.
2. Amplificació del DNA per PCR en emulsió per *beads* (com tècnica 454): la diferència és que la placa on se secuencia està conectada a un *chip* semiconductor que detecta canvis de pH (detecta protons). S'afegeix un nucleòtid (citocina, per exemple) sobre la placa (és a dir a tots als *beads*), en els pous on s'incorpori el nucleòtid s'alliberarà un protó que provocarà un canvi de pH que ho detectarà el *chip*.

> *En aquest cas no cal fer fotos després d'afegir cada nucleòtid perquè la informació química es converteix en digital automàticament, de manera que el procediment és més ràpid i menys costós.*

Altre cop es dona el problema de l'homopolímer. Doncs quan s'afegeixi un nucleòtid s'incorporarà tants cops com estigui repetit i la placa detectarà un canvi de pH major, però no podrà quantificar el nombre exacte de nucleòtids iguals incorporats.

Amb això s'aconseguia seqüenciar un genoma ràpid. En aquella època si es volia una seqüenciació més precisa, que resolés els gaps incerts dels homopolímers, s'utilitzava Sanger.

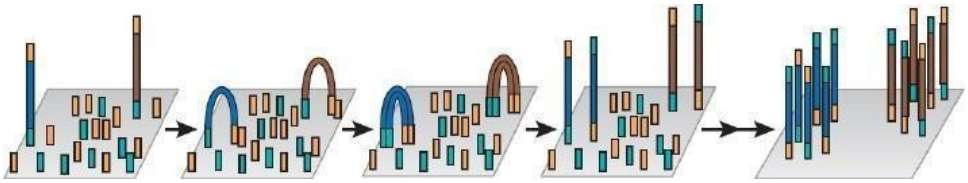
- FragmentacióDelDNAilligamentd'adaptador
- Amplificacióenfaseoliosaperemulsion-PCR
- Distribucióenpousd'unchipsemiconductor
- Piroseqüenciació →flowgrama

Illumina

The Illumina sequencing workflow is composed of 4 basic steps:

- **Sample Prep → DNA fragmentation and adapter ligation**
- **Cluster Generation**
- **Sequencing by synthesis**
- **Data analysis**

1. Obtenir la genoteca amb adaptadors en els dos extrems.
2. Amplificació en fase sòlida i generarem *clústers* mitjançant *bridge PCR*: amplifiquem en una placa on hi haurà els adaptadors enganxats que ens faran de *primers.* Es distribueixen tot de seqüències de la genoteca de forma homogènia per tota la placa. L'amplificació tindrà lloc al voltant dels *primers* que estaran enganxats a les plaques, les còpies es crearan al voltant de l'adaptador i s'aniran formant *clústers*, es formaran com ponts gràcies a l'amplificació al voltant del *primer.*



3. Circulació de dNTPs **terminadors reversibles fluorescents** i la incorporació d'una única base en cada cicle: quan s'afegeix un nucleòtid, la seqüenciació no pot continuar, així que s'afegeix un de sol en cada seqüència i com que està marcat amb un color determinat, sabrem quin nucleòtid s'haurà incorporat. Després es neteja tot (eliminem la terminació i el fluorocrom) i es tornen a afegir nucleòtids. Es va repetint el procediment fins completar totes les seqüències. Cal remarcar que la base s'incorpora en cada clúster, no en una còpia de DNA en si sola, sinó en tot el conjunt del *bridge PCR*.

4. La lectura de la identitat de cada base d'un clúster té lloc a partir d'imatges seqüencials preses després de cada incorporació.

   Els dNTPs terminadors són reversibles, un cop ja s'han afegit i s'ha fet la foto (resultat del primer cicle), modifiquem els nucleòtids perquè pugui continuar la reacció, per tant, es pot afegir un altre nucleòtid (s'ha eliminat el terminador).

Com en cada pas només s'afegeix un sol nucleòtid, amb aquesta tècnica no es dona el problema de l'homopolímer, ja que es podrà comptar quin és el número exacte de nucleòtids iguals afegits.

El problema, en aquest cas, és que el marge d'error és més gran. A més, si és desquadra la seqüenciació perquè en una de les amplificacions s'incorpora una base errònia, no es pot llegir més la seqüència.

Problem of Illumina → A lot of gaps!

- FragmentacióDelDNAi lligamentd'adaptador
- Amplificacióenfasesòlidaperbridge-PCR
- Circulació de dNTPs terminadors fluorescents
- Incorporaciód'unsol nucleòtid encada *cluster*
- Lectura per imatges seqüencials

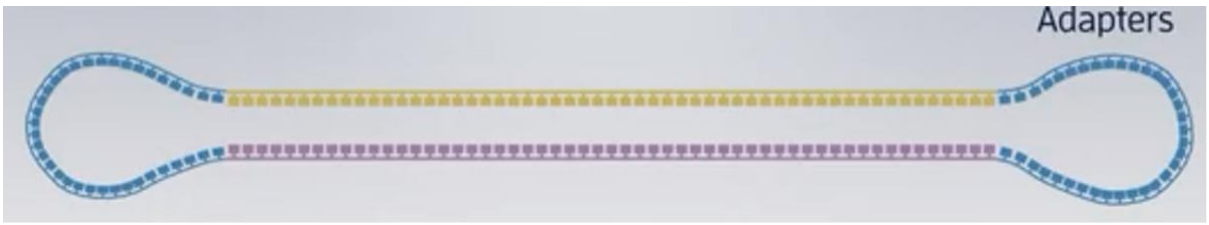
	Throughput	Length	Quality	Costs	Applications	Main sources of errors
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb	Small sample sizes, genomes/ scaffolds, InDels/SNPs, long haplotypes, low complexity regions, etc.	Polymerase/amplification, low intensities/missing termination variants, contaminant sequences
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb	Complex genomes, SNPs, structural variation, indexed samples, small RNA^{+}, mRNAs^{+}, etc.	Amplification, mixed beads, intensity thresholding, homopolymers, phasing, neighbor interference
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex genomes, counting (SAGE, CNV ChIP, small RNA), mRNAs, InDels/homopolymers, structural variation, bisulfite data, indexing, SNPs^{+}, etc.	Amplification, mixed clusters/neighbor interference, phasing, base labeling
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex small genomes, counting (SAGE, ChIP, small RNA, CNV), SNPs, mRNAs, structural variation, indexing, etc.	Amplification, mixed beads, phasing, signal decline, neighbor interference
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb	Non-amplifiable samples, counting (SAGE, ChIP, small RNA), etc.	Polymerase, low intensities/thresholding, molecule loss/termination

Challenges of NGS methods

- Increase read length
- Improve sequence accuracy
- Single-molecule sequencing (no amplification)
- De-novo assembly of complex genomes
- Sequencing of complex regions

PacBio

DNA or RNA is isolated, then a SMRTbell library is created by ligating adaptors, creating a circular template.



Then a primer + polymerase are added to the library that is placed in the instrument used for sequencing. This instrument contains a SMRTcell that contains millions of wells in which a single molecule of DNA (with the adaptors forming a circle…) is immobilized.

As the polymerase incorporates labeled nucleotides, light is emitted. Thus, nucleotide incorporation is measured in real time. 2 options:

- HiFi
- Continuous long read sequencing mode

Aquesta tècnica parteix d'una molècula única de DNA que no cal amplificar. La polimerasa treballa de forma fixa (està immobilitzada), i això permet monitoritzar i veure fluorescència d'un color determinat quan un nucleòtid s'incorpora. També s'anomenen Hifi.
La seqüenciació és en temps real (té un elevat throughput) i la longitud dels reads és molt gran (de 10-15 kb fins a 50 kb).
La taxa d'error és del 15% (molt elevada).

Nanopore

Sequencing DNA or RNA. Only nanopore can sequence RNA!

However, we normally transform the RNA to cDNA because then we will obtain a longer output. DNA is more stable

Protein nanopores are embedded into a synthetic membrane bathed in an electrophysiological solution and an ionic current is passed through the nanopores.

As molecules such as DNA or RNA move through the nanopores, they cause disruption in the current (electric base detection). This signal can be analyzed in real time to determine the sequence of bases in the strands of DNA or RNA passing through the pore.

The read length is directly related to the length of the DNA or RNA in the sample (there is no limit).
Users can influence their read length by choosing the right preparation methods for their desired experimental results.

Standard extraction methods readily achieve reads from the tens to hundreds of kilobases.

Long reads and high throughput provide a more unambiguous approach to mapping a DNA or RNA sequence, enabling much simpler assembly.

As PCR isn't necessary for nanopore sequencing, amplification bias is removed and library preparation workflows are simpler.

Aparell molt petit (de la mida d'un pendrive) i ens permet seqüenciar DNA o RNA a partir de molècules úniques (no cal amplificar).
Fem passar les seqüències per un nanopor i una corrent perpendicular a la seqüència. En funció de com canvia aquesta corrent cada cop que creua un nucleòtid, podrem obtenir nucleòtid a nucleòtid tota la seqüència. La identificació de les bases és mitjançant a les diferències en la conductivitat del DNA.
Permet llegir reads molt llarg, throughput molt alts, detecció elèctrica de les bases (no calen aparells òptics), però l'error segueix sent molt alt. S'ha utilitzat per seqüenciar l'ebola i el SARS-CoV-2.

Output

$$O = T * P * G$$

T= Tamaño de genoma
P= Profundidad
G= Número de genomas

Número de reads

$$N = (T * P * G)/R$$

T= Tamaño de genoma
P= Profundidad
G= Número de genomas
R= Tamaño de reads

Número máximo de genomas

$$Nmax = M/(T * P)$$

M= Output de secuenciación
T= Tamaño de genoma
P= Profundidad

Número de reads paired-end

$$Nr = (M/R)/2$$

M= Output de secuenciación
R= Tamaño de reads

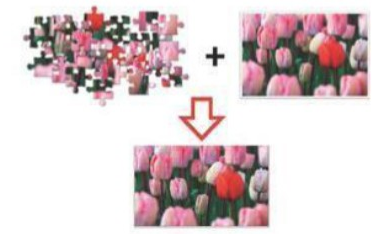
Topic 2. Applications

A les cèl·lules tenim tantes molècules de DNA com cromosomes, mentre que en els seqüenciadors obtenim milers de seqüències petites de cada cromosoma (reads). Per tant necessitem alguna eina d'assemblatge que ens permeti reconstruir el genoma original.

Podem seguir dues estratègies:

- **MAPEIG CONTRA REFERÈNCIA:** Analogia amb el puzzle i la portada

En aquest cas es disposa d'una referència del genoma (això no passa si és la primera vegada que es vol seqüenciar el genoma). És la tècnica usada quan ja es té el genoma seqüenciat i assemblat d'una espècie, i permet assemblar el genoma d'un nou individu de la mateixa espècie.
Per fer-ho, es necessita tenir el genoma de referència i els reads del genoma en qüestió. Aleshores es comparen els reads amb les seqüències més semblants del genoma de referència. Així es van posicionant tots els reads en aquells punts on s'assemblin més a la referència. Finalment es fa un consens de tots els reads alhora sense tenir en compte el genoma de referència.
L'inconvenient d'aquesta metodologia és que no permet la detecció de seqüències noves ni les reorganitzacions estructurals. Si entre un individu i un altre de la mateixa espècie hi ha seqüències diferents o fragments que estan invertits en posicions diferents, trobarem incongruències en el posicionament dels reads, així que aquests s'hauran de mapejar de novo.



- **ASSEMBLATGE DE NOVO:**

És l'única estratègia que permet seqüenciar genomes per primera vegada i finalitzar mapejos contra referència que presenten incongruències.
En aquest cas, s'ha de comparar cada read amb tota la resta de reads obtinguts (busquem solapaments). Si tenim una alta redundància, esperem que els reads solapin entre ells i tinguin seqüències comunes. De manera que es puguin anar estenent els reads i fer un consens final.

Aquesta tècnica és més complexa, lenta i es necessita molta memoria de computació.

CONCEPTES BÀSICS

- **Read**: fragment de seqüència seguida (de fins 1000pb) obtingut en una reacció de seqüenciació.
- **Contig**: conjunt de reads que s'han pogut ordenar (per mapeig contra referència o assemblatge de novo) per formar un segment continu de seqüència en base al solapament dels seus extrems.
- **Scaffold**: conjunt de contigs ordenats i orientats en el genoma en base a informació obtinguda de reads aparellats. Conté gaps o seqüències sense determinar.

Cal destacar que, generalment, en un assemblatge mai podem unificar tots els contigs perquè trobem gaps sense solapar. Per tant el que s'obté és un conjunt de contigs.

Sequence assembly

Reads are stored in large files, since we could have 100 coverage (for each position you have 100 possible bases)→ FASTq
Contig are stored in smaller files, since we have the consensus sequence → FASTA

Quality measure

We can look at 4 things to determine the quality of an assembly:

- Phred score (Q)
- Redundancy
- N50
- L50

Phred Score: There is a quality for each position of the sequence

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Q > 20 → reliable base
Nanopore was moving below Q = 20, but now it has improved up to 30.

Normally, you trim the bases that are below 20.

Redundancy: Average number of reads spanning each base of the assembly

$$R = \frac{N \cdot L}{G}$$

N = number of reads
L = average read length
G = genome size

Sometimes we do not know the size of the genome. So, it can be a problem.

When using Nanopore, reads will have different sizes and, therefore, we can not use that formula to calculate the redundancy. But we know the output, which is the product of ? and ?.

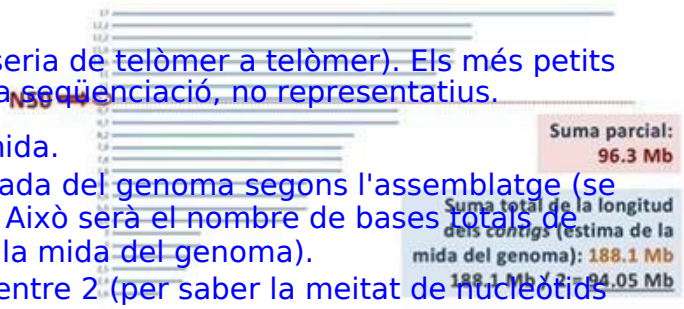
N50 CONTIG LENGTH

Es la forma de mesurar la qualitat del genoma.

Longitud L d'un *contig* tal que el 50% de bases de l'assemblatge es troben en contigs de longitud ≥ L.

Entre *contig* i *contig* és on es troben els gaps, que recordem que son terminadors de la seqüenciació.

Com més gran millor (l'essencial seria de telòmer a telòmer). Els més petits del 50% es consideren restes de la seqüenciació, no representatius.



1. S'ordenen els *contigs* per mida.
2. Es determina la mida estimada del genoma segons l'assemblatge (se suma la mida de tots els *contigs*). Això serà el nombre de bases totals de l'assemblatge (i una estimació de la mida del genoma).
3. Es divideix aquest nombre entre 2 (per saber la meitat de nucleòtids assemblats).
4. Es van sumant *contigs* (les seves longituds, de major longitud a menor) fins arribar al valor
   obtingut anteriorment en el pas 3 (suma total dels contigs / 2). *Suma parcial fins a 94.05Mb o poc més (en l'exemple)*
5. La mida de l'últim *contig* que sumem és la mida mitjana dels contigs on es troba la meitat del genoma assemblat -> aquest serà l'N50.

L50 CONTIG LENGTH

Mesura el número de contigs inclosos en el recompte del N50.

Com més petit millor --> *En l'exemple, el N50 és 9.7 Mb i el L50 és 8 contigs.*

The next level of organization is the **scaffold**. They are ordered and oriented contigs based on information from paired-end reads.

They contain gaps and there are 2 strategies to solve this:

1. Using a PCR and a set of primers. But if it is too long the gap, we can not do this. **We use this to know the gap length, not its sequence.**
2. Mixing reads from different technologies (Hybrid technologies are used to fill the gaps):

- Use short reads to make the contings
- Use long reads to fill the gaps (even if the quality is bad). We could also use the long reads and use the short reads to correct the long reads.

Topic 3. RNA-seq

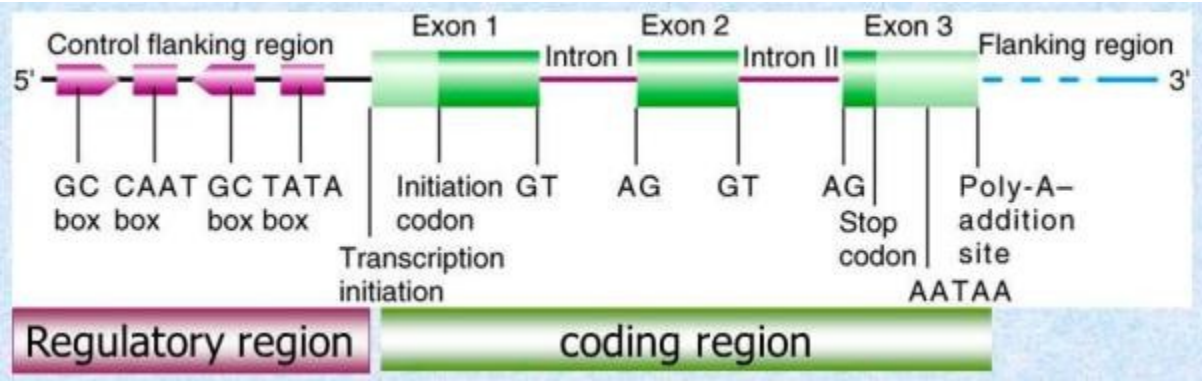
What is a gene?

The gene is the basic physical unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify traits. Genes are arranged, one after another (this in eukarya is nor true, only in bacteria), on structures called chromosomes.

Basic structure of a gene

We can distinguish 2 regions:

- Coding
- Non coding



Poly-A tails is a post-transcriptional modification that is very useful to extract mRNA (3%).

- 90% of the reads are rRNA

So, the first step is to extract the mRNA. We can use specialized kits or target the poly-A tail. RNA from viruses also contains a poly-A tail, so it is also useful to select the RNA from viruses.

ANOTACIÓ DE GENS

És el tercer i últim pas quan nosaltres volem obtenir una seqüència genòmica d'un organisme determinat. Un cop es tenen els fragments de genoma seqüenciats, s'intenten anotar-hi les seqüències funcionals (principalment gens).

- Anotació basada en l'anàlisi de seqüències de nucleòtids
- Descobriment de gens ab initio
- Homologia amb altres espècies
- Anotació basada en l'anàlisi de l'expressió gènica (Transcriptome)
- Seqüenciació de ESTs
- RNA-seq

We normally use:

- Similarity-based methods: Use similarity to annotated sequences like proteins, cDNAs or ESTS.
- Ab initio prediction: Likelihood based methods

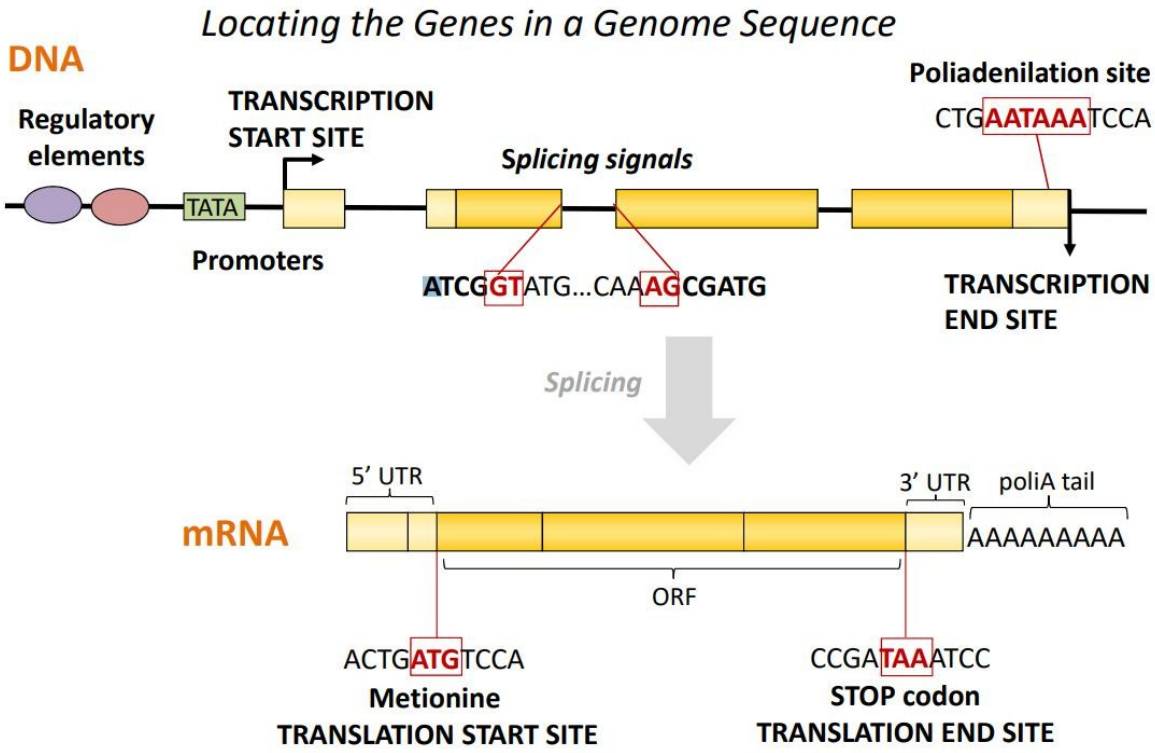
Ab initio gene prediction

To locate the genes in a Genome sequence, we can try to detect:

- Splicing signals
- Poliadenilation sites

To detect the mRNA:

- Start codon (ATG)
- STOP codon (TAA…)



What is the transcriptome?

All the transcripts of a cell, and their quantities, in a specific stage of development and a given physiological condition.

Objectives:

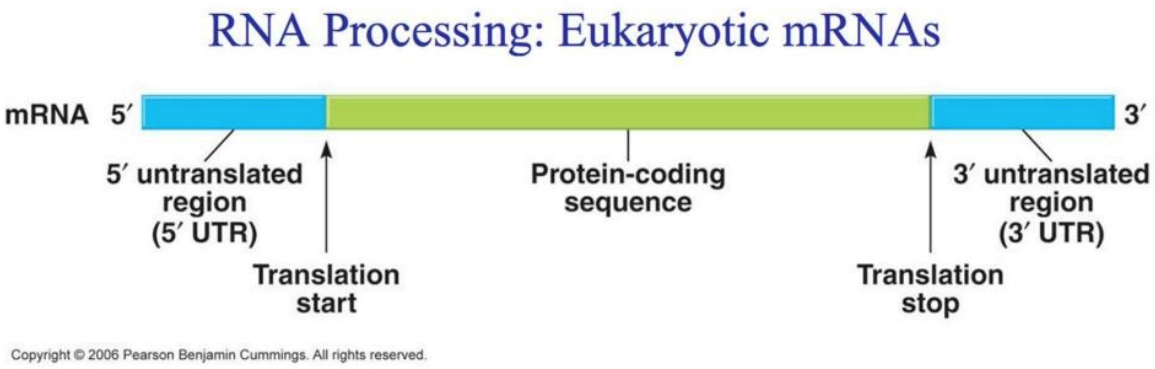
- To catalog all types of transcripts, including mRNAs, ncRNAs and sRNAs
- To determine the transcriptional structure of the genes, including the transcription start sites, 5' and 3' UTRs, the splicing patterns and other post-transcriptional modifications
- To quantify changes in the levels of expression of each transcript during development and under different physiological conditions

The study of RNAs gives information about:

- Genes and other expressed sequences of a genome
- Gene regulation and regulatory sequences
- Function of the genes and their interaction
- Functional differences between tissues and cell types
- Identification of candidate genes for any given process or disease

Eukaryotic transcription is complex

For this reason, eukaryotic gene prediction has high error rates . Gene finders generally do a poor job (<50%).



- Eukaryotic mRNAs have three main parts (Figure 13.8):
  - 5′ untranslated region (5′ UTR),
    - varies in length.
  - The coding sequence
    - specifies the amino acid sequence of the protein that will be produced during translation.
    - It varies in length according to the size of the protein that it encodes.
  - 3′ untranslated region (3′ UTR),
    - also varies in length and contains information influencing the stability of the mRNA.

MÈTODES D'ANÀLISI DEL TRANSCRIPTOMA

Existeixen diversos mètodes d'anàlisi del transcriptoma. Alguns d'ells són per estudiar la transcripció d'un únic gen, però ens centrarem en les tècniques d'estudi de tot el transcriptoma.

Mètodes d'anàlisi d'un únic gen:

- Northern Blot
- RT-PCR
- 5' i 3' RACE
- RT-PCR quantitativa (*Real time PCR*)

Mètodes d'anàlisi de tot el transcriptoma:

- Microarrays
- Seqüenciació de ESTs
- RNA-Seq

2.1. MICROARRAYS

Tècnica basada en la hibridació de RNA marcat amb fluorescència amb múltiples sondes de DNA unides a un xip sòlid (base sòlida) que ens permet estudiar l'expressió dels transcrits.

2.1.1. GENE EXPRESSION ARRAYS

És una tècnica on les sondes usades són oligonucleòtids curts (25nt. No fa falta que siguin molt llargues, ja que normalment no hi ha exons semblants), hi ha múltiples sondes situades als exons que es troben a l'extrem 3' del gen i permet quantificar l'abundància dels transcrits.

Detecció dels nivells d'expressió de tots els gens d'un genoma.

HIBRIDACIÓ D'UNA ÚNICA MOSTRA:

Es compra un xip comercial (placa) que té una sèrie de sondes amb seqüències curtes de DNA que hibriden amb gens coneguts.

Es fan hibridar fragments de cDNA marcat fluorescentment (mRNA es retrotranscriu a cDNA) amb sondes sobre un xip on, cada quadradet del xip, conté diferents sondes (totes del mateix tipus en cada quadradet, diferents a les d'un altre quadradet). Cada sonda correspon a un gen, per exemple. La hibridació farà canviar el color del quadradet degut a la fluorescència del RNA.

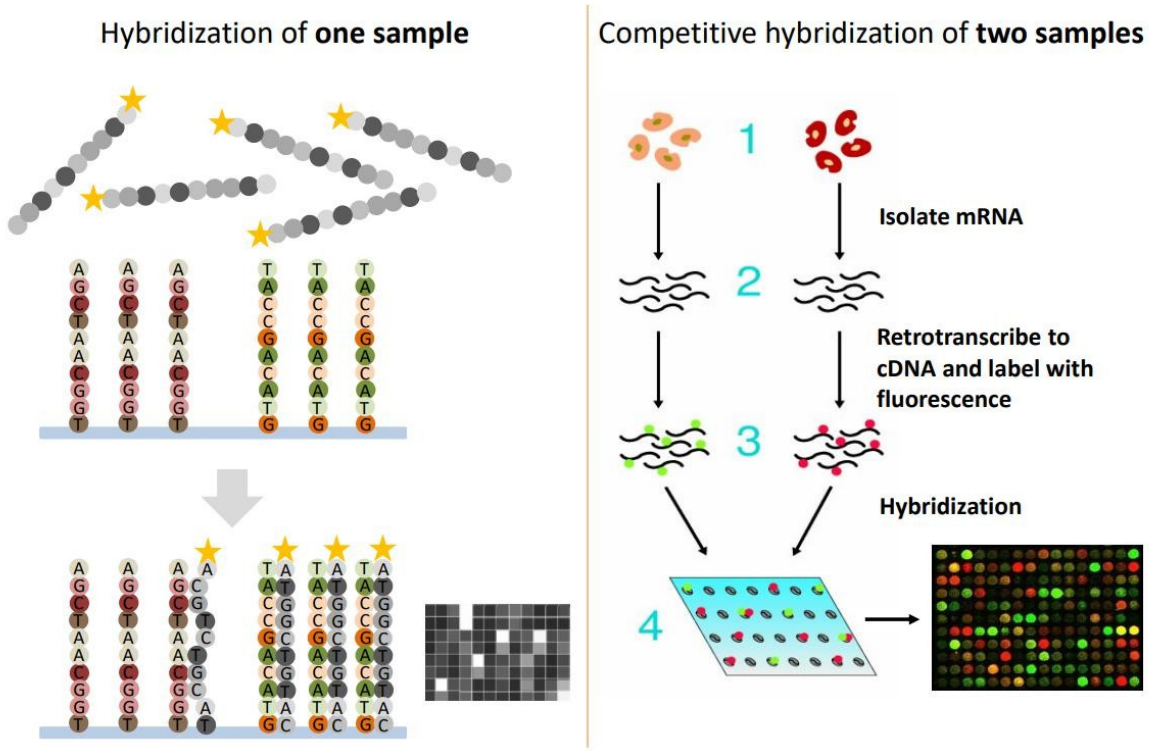
- Si total la sonda queda unida a RNA s'obté una gran fluorescència (diferents nivells de fluorescència segons la seva expressió) → Quantificable el nivell d'expressió
- Si no hi ha hibridació, queda negre.

HIBRIDACIÓ COMPETITIVA DE DUES MOSTRES:

Permet comparar l'expressió dels mateixos gens en diferents condicions (teixit sa vs teixit cancerós).

Primer s'aïlla el RNA dels dos teixits, després es retrotranscriu a cDNA i es marca amb fluorescència (colors diferents per cada teixit que es vol comparar) i es fa hibridar amb sondes que contenen determinats gens. Segons si hi ha hibridació o no, es veurà fluorescència d'un color o d'un altre.

- Expressió de la mostra vermella més alta → fluorescència vermella
- Expressió de la mostra verda més alta → fluorescència verda
- Expressió igual les dues mostres → fluorescència groga
- Si no s'expressa cap mostra → negre.



2.1.1. GENOME TAILING ARRAYS

En aquest cas s'usen oligonucleòtids llargs (60nt) que es posen solapant-se al llarg d'una regió que es vol analitzar amb més detall (tenim un especial interès en analitzar aquesta regió). Per fer-ho, es dissenyen sondes molt específiques per la regió que solapen desplaçant-se al llarg de la regió d'interès fins tenir-la tota coberta (ja que poden haver- hi seqüències repetitives).

Això és molt més car però permet la identificació de noves seqüències transcrites.

- Les sondes corresponents a exons hibridaran amb el RNA (cDNA) mentre que les que han estat dissenyades amb les seqüències dels introns no podran hibridar.
- Les sondes seran de 60 nt, el solapament entre sonda i sonda són de 50 nt (hi ha un pas de 10 nt entre sonda i sonda).

2.1.2. LIMITACIONS DELS MICROARRAYS

- Depenen del coneixement previ sobre la seqüència genòmica, per tant no es podran descobrir gens nous perquè no se n'haurà fet la sonda.
- Hi ha hibridacions de fons (degudes a hibridacions creuades) que emmascaren els resultats, degut a les inespecificitats.
- Hi ha un màxim detectable d'hibridació: quan totes les sondes disponibles hagin hibridat, per molt que hi hagi encara molt mRNA lliure, aquest no hibridarà i no es detectarà bé el nivell d'expressió (saturació).
- El cost és elevat (només en el cas dels *genome tiling arrays*).

2.1. EXPRESSED SEQUENCE TAGS (ESTs)

Tècnica basada en la seqüenciació de genoteques de cDNA.

Quan nosaltres volíem seqüenciar un genoma, nosaltres extreiem el DNA, el fragmentavem, seleccionavem els fragments d'una mida determinada i els clonavem en un vector per tal de crear la genoteca genómica.

En el cas de ESTs, nosaltres partim del RNA de les cèl·lules, el retrotranscribim a cDNA i creem una genoteca genòmica de cDNA. Aquests clons els amplifiquem fent-los creixer en cèl·lules bacterianes.

Ens interessa seqüenciar són aquells que ja han patit el *splicing* i estan madurs.

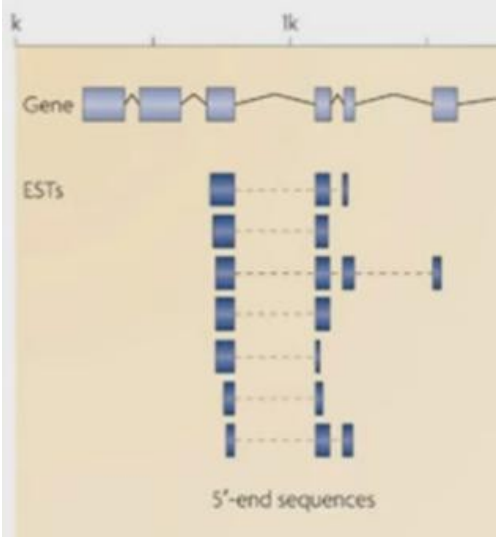
1. S'aprofita la presència de la cua de poliA per capturar els transcrits de RNA.
2. Síntesi de cDNA: es retrotranscriu el RNAs en cDNA amb una *transcriptasa inversa*.
3. Genoteca de cDNA: es fragmenta el cDNA, es filtren els fragments d'interès, es clonen en un vector, es transfereixen a cèl·lules bacterianes i es fan créixer les colònies (amplifiquem el cDNA).
4. Seqüenciació dels extrems: el que interessa és seqüenciar els fragments de la genoteca. Com estan clonats en un vector, es dissenyen *primers* en els extrems del vector per seqüenciar (amb Sanger) cap en dintre els extrems de l'insert.

Aquests fragments (que anomenarem ESTs) son seqüències que s'estaven expressant en les cèl·lules (exons) i ara

hem **d'esbrinar de quina regió del genoma provenen.**

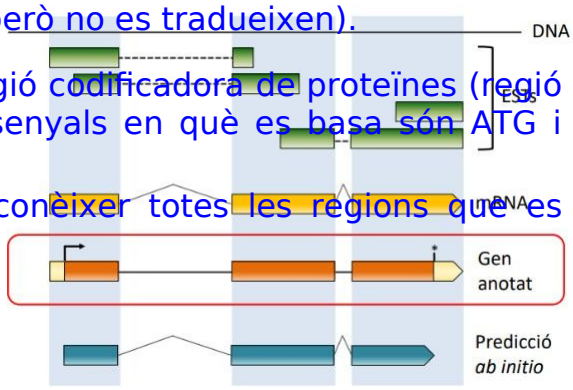
5. Mapeig dels ESTs contra l'assemblatge (genoma que tenim seqüenciat): s'alineen els fragments amb un software informàtic contra un assemblatge del genoma prèviament fet.

En el RNA ja no hi ha introns, però quan es mapeja contra un genoma que sí conté els introns, els reads de RNA (cDNA) queden dividits en fragments amb uns gaps enmig que corresponen als introns (els fragments acostumen a ser d'1kb si es seqüència amb Sanger i els exons acostumen a ser més curts i llavors cada fragment pot contenir més d'un exó). Aquesta tècnica delimita on estan els diferents exons i col·lateralment, indica també on estaran els introns perquè tot i que els extrems no són informatius, la regió interna sí.



En la imatge veiem un sol EST i com dins d'aquest hi ha 3 exons. També cal destacar que un EST no correspon a un transcrit, ja que potser el transcrit és més llarg i hem parat de seqüenciar. Per això diem que tenim informació parcial.

Les regions que s'observen en el mRNA (per tant en les ESTs) però no en les prediccions *ab initio* corresponen a les 3' UTR i les 5' UTR (regions que es transcriuen però no es tradueixen).



- La predicció *ab initio* prediu la regió codificadora de proteïnes (regió que es tradueix) perquè dues de les senyals en què es basa són ATG i codó stop.
- El mapeig de les ETSs permet conèixer totes les regions que es transcriuen (tot i que no es tradueixin).

LIMITACIONS DE LES ESTs

- Té un baix throughput perquè es basa en Sanger (no és massiva).
- Té un cost elevat perquè es basa en Sanger.
- És poc quantificable. No podrem saber ben bé si un gen s'expressa molt o poc.
- Seqüenciació parcial; les diferents isoformes (diferents formes de splicing d'un gen) són generalment indistingibles.
- Com se seqüència poc, hi ha regions dels gens que potser no queden coberts i per tant només s'estarà estudiant una part del transcrit (els ESTs a vegades no ho recobreixen tot).

2.2. RNA-SEQ

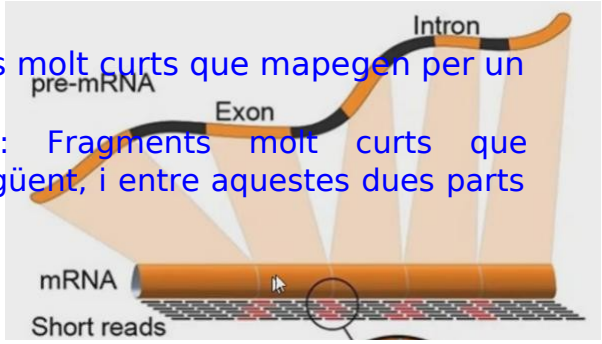
Tècnica basada en la seqüenciació massiva de cDNA.

És una tècnica molt similar als ESTs però no es basa en Sanger sinó en l'ús d'adaptadors, fet que permet seqüenciar massivament, ja que és pot seqüenciar el RNA de forma directa usant tecnologies de nova generació. Aquest mètode facilita una major precisió en la mesura dels nivells de transcrits i les seves isoformes.

1. Síntesi de cDNA: S'extreu el RNA madur (sense introns, amb cua poliA) i es passa a cDNA.
2. Genoteca de cDNA: Es fragmenta el RNA madur i, en comptes de clonar-lo, s'afegeixen adaptadors als extrems (es fa una genoteca de ESTs amb adaptadors).
3. Seqüenciació amb adaptadors: Els fragments es poden seqüenciar per *ilumina*.
4. Mapeig contra assemblatge: Un cop es té el DNA seqüenciat, es mapeja comparant-lo amb un assemblatge del qual es disposi.

El problema és que es basa en tècniques de nova generació que usa seqüències curtes (al ser tant curtes, cada read correspondrà a un exo o al final de un exo i principi del següent), amb molt d'error, cosa que fa difícil posicionar aquests petits fragments en el genoma.

Aquelles regions on s'hagin mapejat *reads* de RNA (regió obtinguda en la seqüenciació que s'ha aconseguit aparellar amb una regió de l'assemblatge), seran regions de transcripció. Hi ha de dos tipus:

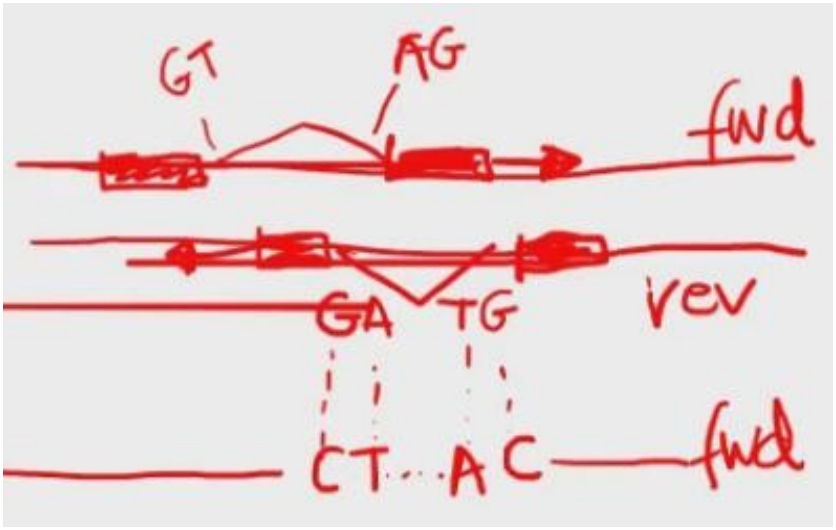


- **Exonic reads (negre)**: Fragments molt curts que mapegen per un únic exo.
- **Junction/split reads (vermell)**: Fragments molt curts que corresponen al final d'un exó i l'inici del següent, i entre aquestes dues parts quedaria inclòs un intró.

També ens haurem de fixar en el nivell d'expressió del cDNA, que ve determinada per la quantitat de reads que mapegen una regió determinada (coverage) → perfil transcripcional

També podem saber la direccionalitat dels gens donada per la junction reads gràcies a les denyals de splicing → inici intró = GT i final intró = AG

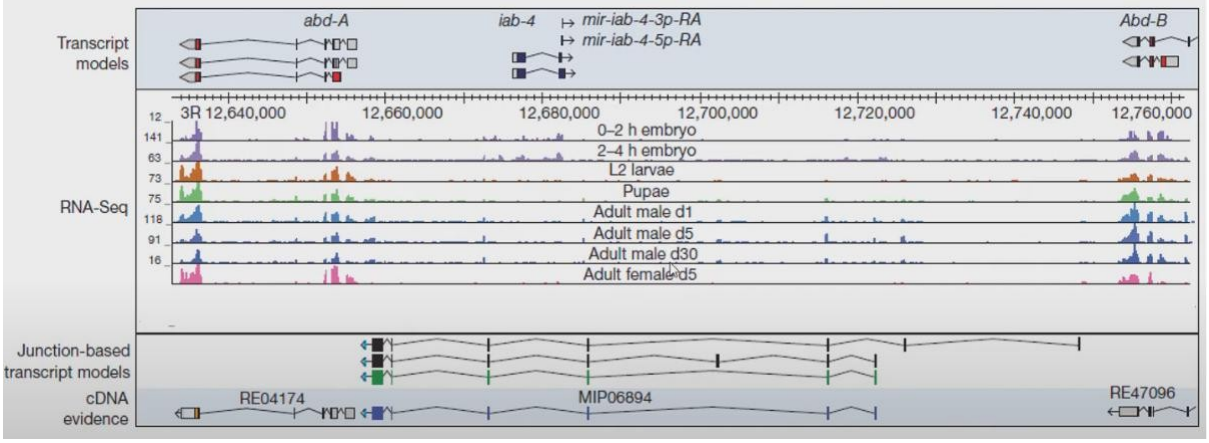
- Gen codificat en la cadena *Forward:* inici intró= GT i final intró= AG
- Gen codificat en la cadena Reverse: va de dreta a esquerra (Rv: final intró GA ← TG inici intró), però quan mapegem ho fem de la cadena forward per tant llegim el seu complementari (Fw: …CT → AC …)



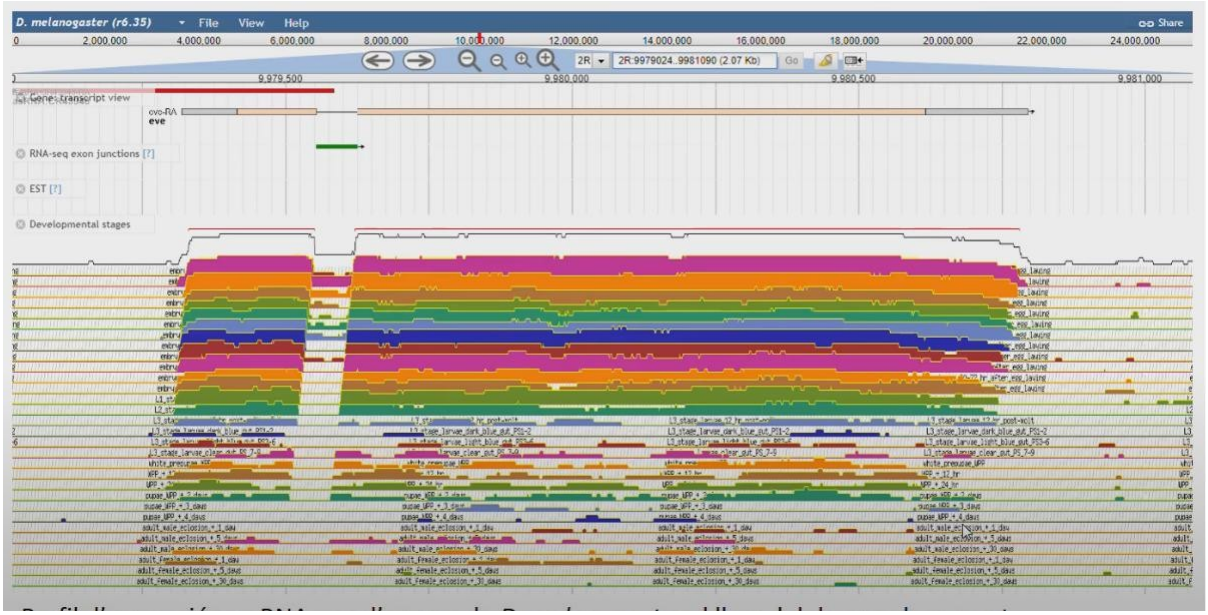
2.2.1. AVANTATGES DEL RNA-SEQ

- La seqüència genòmica no cal que estigui **anotada** prèviament.
- Permet la detecció de nous transcrits o fins i tot nous gens.
- Gran precisió en la detecció dels límits dels transcrits gràcies a fer servir *reads* molt curts.
- Permet detectar variants de splicing i principis i finals de transcripció alternatius.
- Permet detectar SNPs a les regions transcrites. Els SNPs són nucleòtids que són diferents en la seqüència heretada del pare i en la de la mare (aparellaments no WC). Això pot influir en l'expressió dels fragments (el fragment que conté la base X s'expressa molt mentre que el que conté la base Y s'expressa menys).
- Permet detectar transcripció específica de cada al·lel.
- Permet quantificar acuradament els nivells d'expressió de cada transcrit (ja que es disposa d'un ampli rang de mesures).
- Permet una gran reproducibilitat.
- Requereix molt poca quantitat de DNA inicial.

Per exemple, en la següent imatge es veu el que seria la imatge d'un navegador genòmic, on es mostra una regió molt ben estudiada d'un clúster de gens hox. Els reads de RNA seqüenciats es van mapejar contra el genoma i es va determinar el nivell d'expressió al llarg de diferents estadis del desenvolupament de la mosca. RNA-Seq va permetre detectar nous transcrits en aquesta regió.



En la imatge, cada línia és un estadi diferent del desenvolupament de la mosca. En cada estadi s'ha fet un estudi de RNA-seq per mirar l'expressió dels gens. Els diferents reads venen quantificats pels gràfics (perfil d'expressió): com més alt és el gràfic, més expressió hi ha. Es pot dir que el gen estudiat s'expressa molt en els primers estadis, mentre que disminueix la seva expressió a mesura que es desenvolupa la mosca.



1. Amb els *gene expression arrays* només podem analitzar l'expressió d'aquells gens prèviament anotats al genoma i pels quals tinguem sondes? Cert

2. Quin és el número mínim de *gene expression arrays* que hauries d'hibridar per analitzar els nivells d'expressió de 12 gens del teu interès en 84 teixits del cos humà? 84 (1 per cada teixit).

3. Els *gene expression arrays* ens permeten anotar: nivells d'expressió dels gens (NO inici ni final de la transcripció ni traducció).

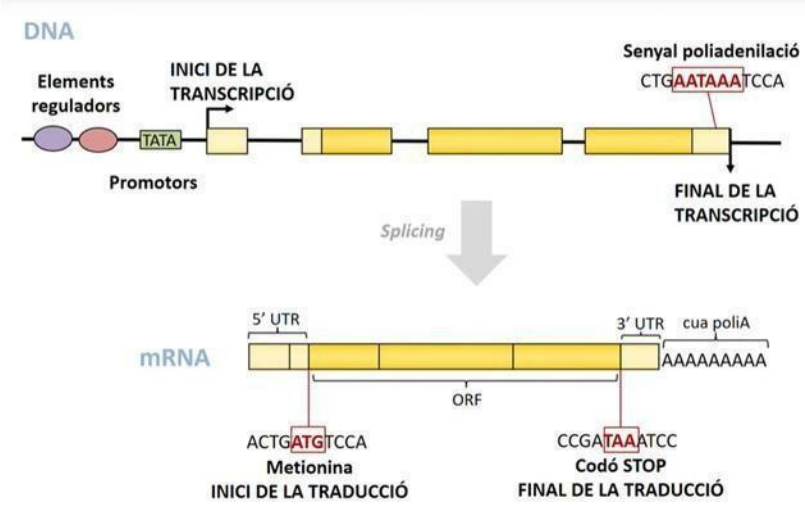
4. ESTs i RNA-seq ens permeten anotar: nivells d'expressió dels gens i l'inici i final de la transcripció i exons i introns.

Els gens estan formats per exons (caixes grogues) i els introns (línies).

Els exons tenen una part que formarà part de la proteïna (groc intens) i una part que es transcriu a mRNA i no s'elimina per *splicing* però que no es tradueixen a proteïna que són les UTR (groc clar).

El procés d'eliminar els introns del pre-mRNA és l'*splicing*, on ens quedarà el mRNA/transcrit madur amb els UTR + ORF + cua poliA.

Veiem com el 5'UTR arriba fins el codo d'inici i el 3'UTR comença amb el codo STOP

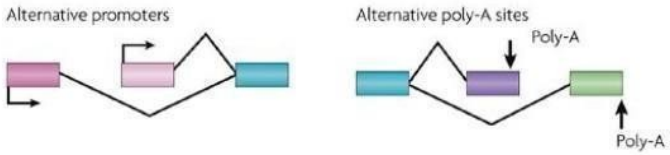


1. SPLICING ALTERNATIU

El splicing alternatiu és un fet molt important en la transcriptòmica perquè permet modificar els transcrits per generar diferents proteïnes a partir d'un únic gen. Aquest splicing es produeix en el moment en què s'eliminen els introns.

L'eliminació esperaríem que es donés des de l'inici fins al final de l'intró, on només quedessin els exons. Però a vegades això no passa exactament així → és quan té lloc el splicing alternatiu.

EXONS A L'INICI O AL FINAL



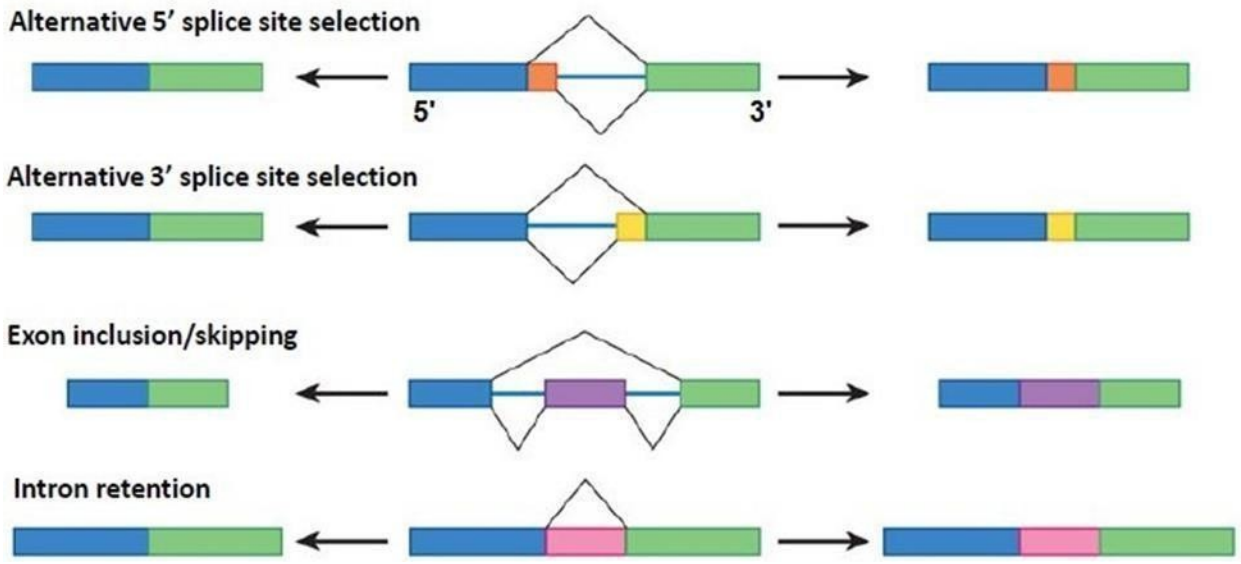
Estan afectats els exons inicials o finals.

Promotor Alternatiu: Veiem com la transcripció pot començar en llocs diferents i que per tant s'omiteix el primer exó o no. De fet, sempre s'està eliminant un exó si mirem la imatge.

Final alternatiu de la transcripció: En funció del splicing la cua poli-A es trobarà en un exó o en un altre.

Això acostuma a passar en exons que no són codificants (els extrems UTRs). En aquests casos, els transcrits comencen o acaben per punts diferents als que serien habitual.

EXONS INTERNS

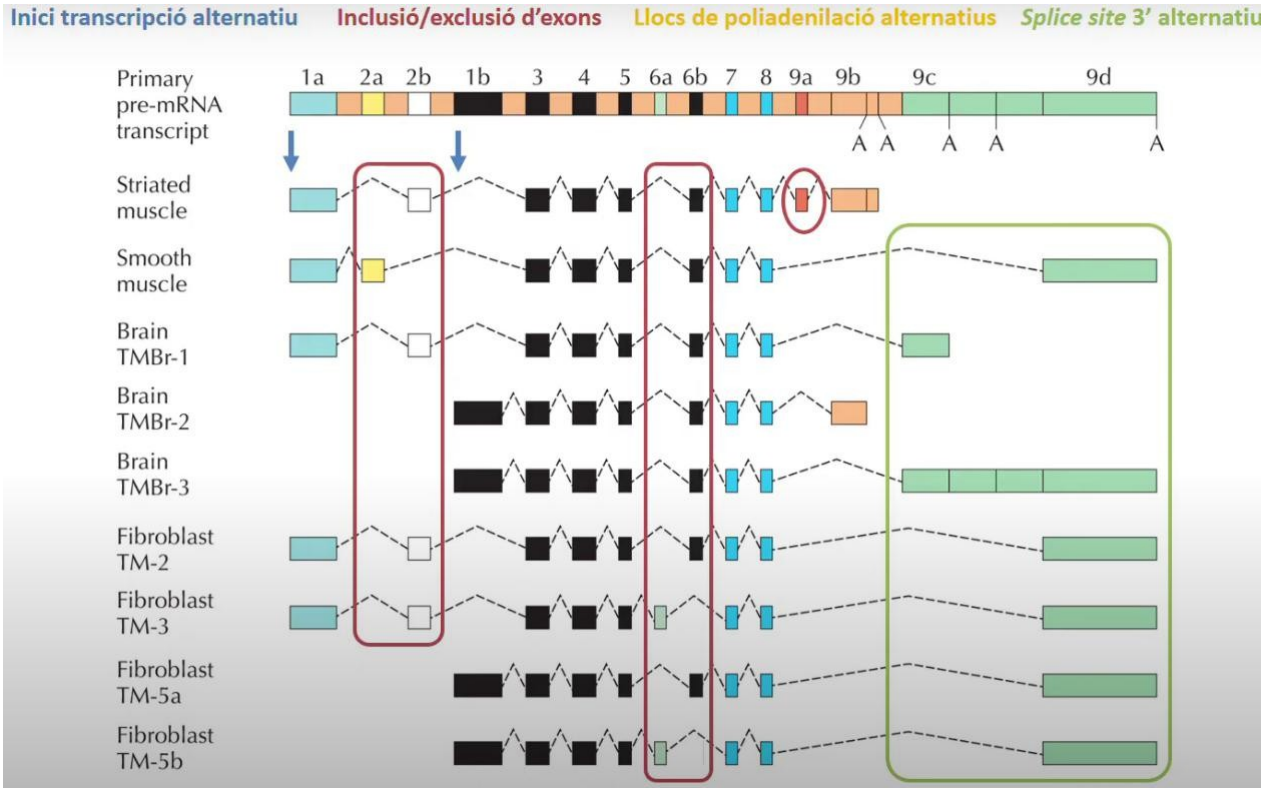


- **Splicing alternatiu en 5':** en aquest cas, la part que s'està eliminant de forma desigual és la part 5' de l'intró (queda retinguda una part de l'extrem 5' de l'intró).
- **Splicing alternatiu en 3':** en aquest cas, la part que s'està eliminant de forma desigual és la part 3' de l'intró (queda retinguda una part de l'extrem 3' de l'intró).
- ***Exon inclusion/ skipping:*** es pot incloure o no un exó.
- ***Intron retention*:** el que seria un intró, no és eliminat i es manté en el transcrit madur. La caixa rosa és un intró.

EXEMPLE 1:

En la figura de la següent pàgina, els introns serien els puntejats i els exons les caixetes. El gen es transcriu i després es processa de maneres diferents per generar una gran varietat de transcrits. Es pot veure que cada transcrit s'expressa en un teixit diferent: cada transcrit té un nivell d'expressió específic de teixit. Això fa que el transcriptoma sigui molt més complex en base a un únic gen del genoma.

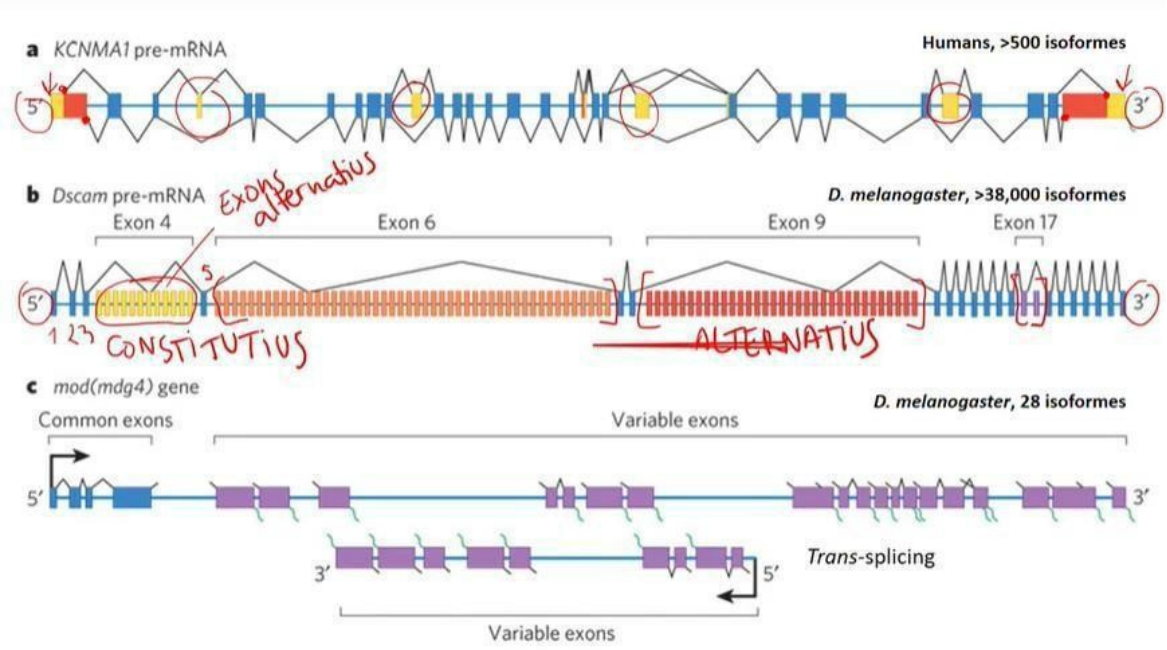
- **Inici alternatiu:** els inicis de la transcripció són diferents, alguns comencen en les caixes blaves i altres en les negres.
- **Inclusió/exclusió d'exons:** els exons del grup 2 són mútuament excloents perquè o hi ha presència (s'inclou) un exó o l'altre, però mai estan els dos. En els exons del grup 6 passa el mateix.
- **Llocs de poliadenilació alternatius:** el final dels transcrits és variable.
- **Splicing 3' alternatiu:** la caixa verda comença en llocs diferents, això vol dir que part de l'exó s'ha eliminat al fer-se splicing alternatiu de l'intró pel seu extrem 3' en diferent posició. No és un splicing 5' alternatiu per què veiem com en tots els casos les caixes blaves son iguals.



EXEMPLE 2:

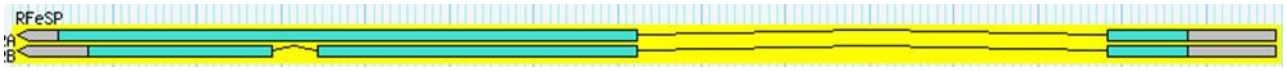
Els exons blaus són constitutius (sempre estan presents) i els no-blaus són alternatius.

- Els exons formen clústers d'exons alternatius: en cada transcrit del gen es troba només un d'aquests exons. Només veurem un exó 4 en el cas B. Només veurem un exó 6…
- Es dona *trans-splicing* segons si el transcrit s'ha generat de la cadena positiva o negativa.



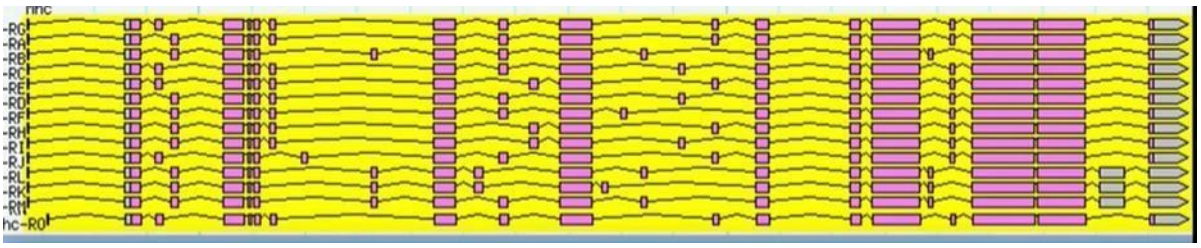
TEST SPLICING

PREGUNTA 1: Marcar mecanismes d'splicing alternatiu que generen els transcrits del gen.



- A l'inici no hi ha res.
- Al final no hi ha res (tot i que els colors siguin diferents, això només afecta la traducció).
- A la zona interna: hi ha un intró que es reté.

PREGUNTA 2:

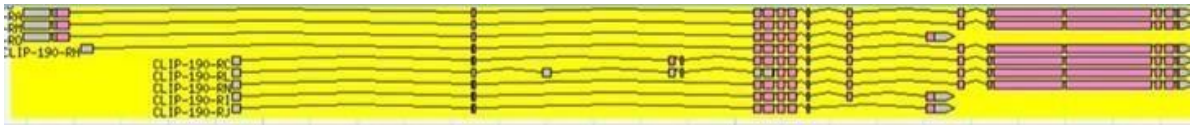


- Inici de la transcripció alternatiu.
- Exons mutuament excloent.

PREGUNTA 3:

- Inici de la transcripció alternatiu.
- Punt d'splicing 5' (alternative 5' splice site)

PREGUNTA 4:

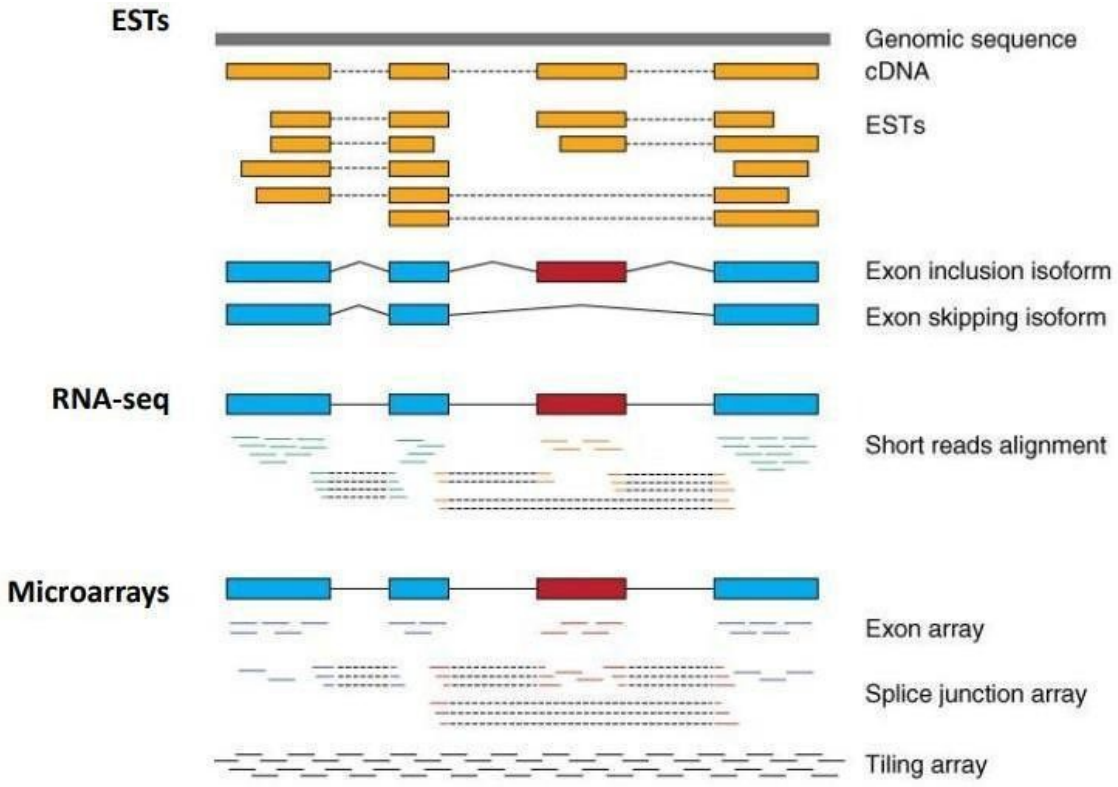


- Inici i final de la transcripció alternatiu.
- Inclusió/ exclusió d'exons.

MÈTODES D'ANÀLISI DEL *SPLICING* ALTERNATIU

Si tots els transcrits s'expressen alhora en el mateix teixit, amb un RNA-seq ho veurem tot igual perquè detectarem expressió a tot arreu. Però no podrem veure com estan compostos cadascun dels transcrits individuals. Per determinar això podem usar altres tècniques.

- **ESTs:** són *reads* prou llargs com per seqüenciar molts exons d'una tirada. Un cop seqüenciat i mapejat el transcrit, obtindrem gaps corresponents als introns. Si alguns ESTs, per a la mateixa regió, mapegen altres exons, detectarem una inclusió/exclusió d'exons.
- **Microarrays:** la majoria dels fragments estaran dintre dels exons, per tant mostraran l'expressió perquè tindrem trossos seqüenciats provinents dels diferents exons. Però per formar els transcrits, s'usen els *junction reads* (mirarem els splice junction array, ja que són els únics que ens permeten veure si hi ha splicing alternatiu o no) que s'han seqüenciat d'una tirada però mapegen en dos punts diferents: això adverteix que entremig hi ha un intró que s'ha eliminat per splicing.
- **RNA-seq:** es dissenyen sondes contra el gen que continguin trossos de dos exons (final d'un i inici del següent) i si hibrida voldrà dir que aquest transcrit s'està expressant. Es pot fer també una sonda que contingui el final d'un exó i l'inici del tercer, i si es troba hibridació es detectarà l'exclusió del segon exó. Això és conegut com *alternative splicing arrays*.



Topic 4. Metagenomics

There are a lot of omics:
- Genomics
- Metagenomics
- Metabolomics
- Interactomics

The difference between them is that the sample and output are going to be different.

Until now, we have been studying a single organism. But what happens if we sequence the genome of every organism present in a drop of water? This is **metagenomics**

We are going to be talking about 2 approaches:
- Amplicon based: We will first use a PCR to amplify. We introduce variance, which is not good.
- Shotgun metagenomics: Just sequence everything straight away

Genomics is a field of biology focused on studying all the DNA of a single organism — that is, its genome. Such work includes identifying and characterizing all the genes and functional elements in an organism's genome as well as how they interact.

Metagenomics is the study of the structure and function of entire nucleotide sequences isolated and analyzed from all the organisms (typically microbes) in a bulk sample. Metagenomics is often used to study a specific community of microorganisms, such as those residing on human skin, in the soil or in a water sample.

The great plate count anomaly

If you go to nature and take a sample and put the sample in an agar plate, you will see a few numbers of organisms growing in this plate:
- 99% unculturable.

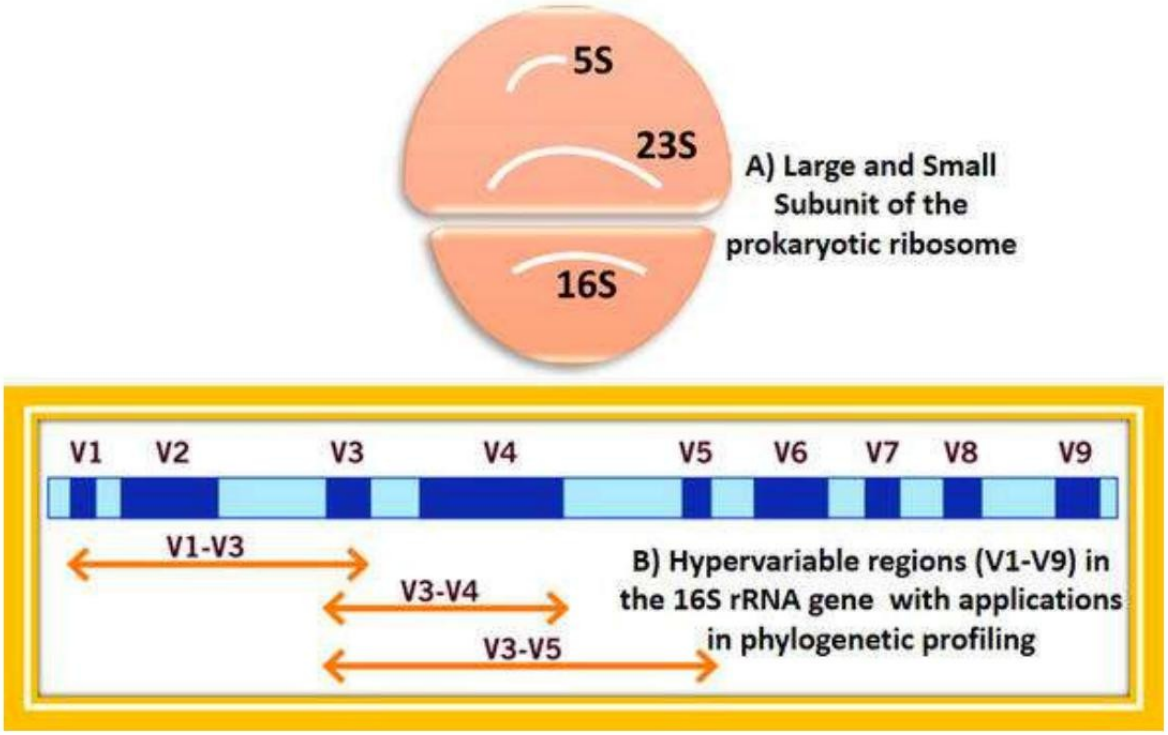
Wrong: We can't culture this organism
True: We do not know how to culture this organism, because we don't know the correct conditions.

16 S

The reads obtained from second technology sequencing techniques have an average length of 150 bp.
Genes have a larger length → 1500 bp

Thus, you can only get a portion of the gene.

There are regions of the gene that are very conserved (you find no differences between organisms) and other regions that are hypervariable (more informative).

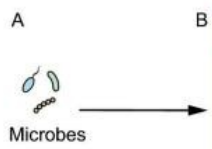


In third generation techniques, we can cover the whole gene and therefore we are reaching a different resolution.

At the beginning of COVID, we did not know anything:

- We could not use genomics to study it, since we did not know its genome.
- We could not use Amplicon to study it because we did not know the genes of COVID
- We used metagenomics

If we have microbes, we can try to culture them.



Method	Advantages	Limitations
Culturome	• High-throughput • Targeted selection • Provides microbial isolates	• Expensive • Laborious • Influenced by media and the environment

But if we have the DNA, we can do many other things:
- Amplicon
- Metagenome
- Virome
- Metatranscriptome

The first step is to extract the DNA.
- In the case of COVID, we first need to transform the RNA into cDNA.

The **Amplicon** method:
- 16S if its for bacteria
- 18S if its for eukarya

If we want to work with a particular taxon of plankton, you can identify a gene that is informative of this taxon and use it as a target. We just design the primers and make the sequencing, so we do not need to always use the 16S/18S.

Benefits:
- Very quick. If you work with nanopore, you can do real time monitoring.
- Low-biomass requirement, since it is PCR-based.
- Applicable to samples contaminated by host DNA, because you amplify a concrete gene that is informative. If you use 16S in humans, you only obtain information from viruses (not eukarya).

Limitations:
- PCR and primer biases, since primers have affinity (they are not universal). Thus, some groups will be more amplified. The only way to reach some level of equilibrium is to use more than 1 pair of primers. So, each pair is going to have a certain affinity to each group and they will compensate.
- Resolution limited to genus level because we are talking about Illumina. If we use Nanopore this is not true.
- False positive in low-biomass samples.

Amplicon is used to see what is there! For identification purposes only.
You are not looking at the metabolic profile…

Metagenome

Advantages	Limitations
• Taxonomic resolution to species or strain level • Functional potential • Uncultured microbial genome	• Expensive • Time-consuming in analysis • Host-derived contamination

You can identify the taxons but also the function of those taxons.

Regarding Host-derived contaminations, here we do not have the information from the 16S or 18S to say this is from the host or not. So, the first thing we need to do is to map all the reads to the reference genome.

Virome

Advantages	Limitations
• Can identify RNA and DNA viruses • Quick diagnosis	• Most expensive • Difficult to analysis • Severe host-derived contamination

Metatranscriptome

Advantages	Limitations
• Can identify live microbes • Can evaluate microbial activity • Transcript-level responses	• Complex sample collection and analysis • Expensive and complex in sequencing • Host mRNA and rRNA contamination

We look at the expression of the genes in our whole community.

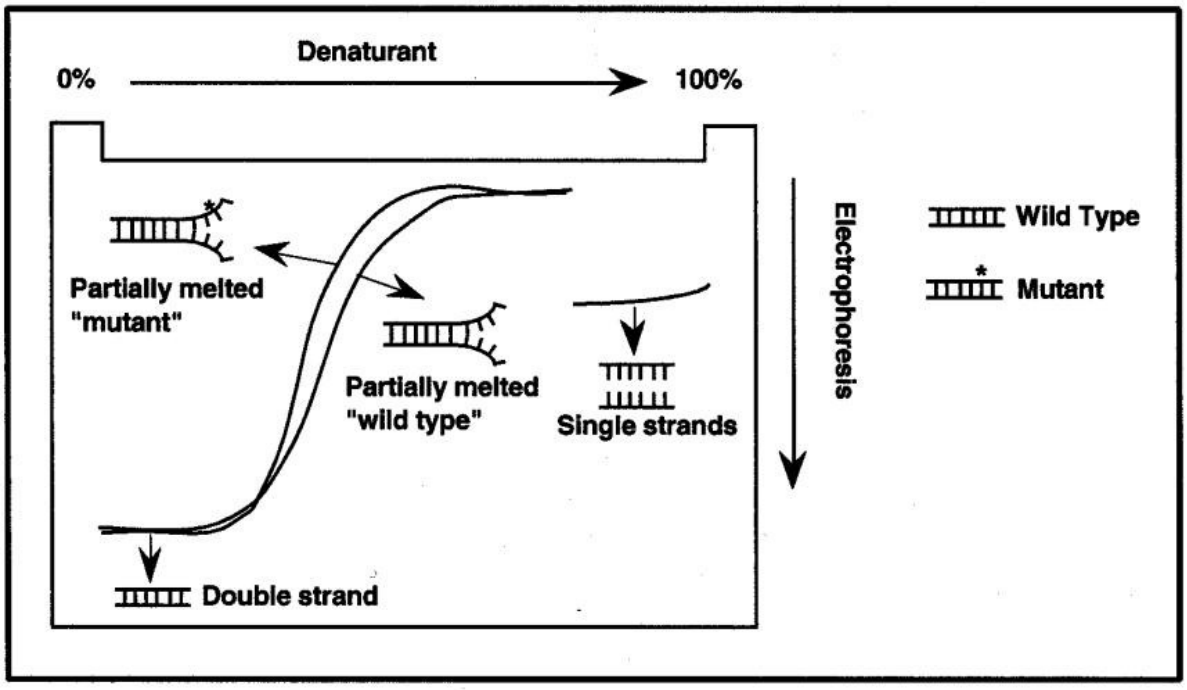
Ecology genomics

- Over the past two decades there has been an explosion in understanding of how microbes – bacteria, protists and viruses –critically influence the structure of and function of the environment
- < 1–5% of bacteria could be grown in culture and made very difficult to study the vast complexity in natural environmental assemblages
- Numbers of all microbes on Earth: between 9.2×10^29 and 31.7×10^29
- The ocean floor is home to a staggering 2.9×10^29 single-celled organisms — that's 10 million trillion microbes for every human on the planet

DGGE

- Denaturing Gradient Gel Electrophoresis
- Separated DNA of same size based on sequence differences.
- Different sequences “behave differently at different amounts of denaturing chemical (or heat; see TGGE)
- At some point 16SrRNA DNA strands completely separate.
- Complete separation of PCR amplicon is hindered by GC-clamp added to one of the PCR primers.

Since they didn’t have access to the sequencing technologies, they runned a gel.
They identified the different bands that correspond to different organisms.



Metagenomics

Metagenomics (also referred to as environmental and community genomics) is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms.

The development of metagenomics stemmed from the ineluctable evidence that as-yet uncultured microorganisms represent the vast majority of organisms in most environments on earth. This evidence was derived from analyses of 16S rRNA gene sequences amplified directly from the environment, an approach that avoided the bias imposed by culturing and led to the discovery of vast new lineages of microbial life.

Although the portrait of the microbial world was revolutionized by analysis of 16S rRNAgenes, such studies yielded only a phylogenetic description of community membership, providing little insight into the genetics, physiology, and biochemistry of the members.

Metagenomics provides a second tier of technical innovation that facilitates study of the physiology and ecology of environmental microorganisms.

QC

Quality Control is an important step. Since we are working with environmental samples, the contamination is going to be very common. You need to be sure that the results from your experiment are the ones coming from the sample.

We will look at the taxonomic diversity → Who is there?
Functional annotation → What are they doing?

When trying to find the taxonomy of our sample, we will use a DB. If our sample is not contained in any DB, we can use clustering. We just put together all the reads that are the same.

- You will obtain groups of reads that are similar.
- You can then make a taxon abundance profile

Two key approaches to profiling the microbiome

- **16S ribosomal RNA gene (amplicon based):** We have a gene that is around 1500 bp and this gene has regions that are conserved and others that are hypervariable. We need to use the hypervariable regions because the conserved are not informative. We can combine many hypervariable regions but then we will not be able to reproduce the results and compare them.

The information we obtain is:

- Which genera and species are present?
- What is the community's **predicted** functional potential? We are not going to be able to define the function, but according to the composition we can predict which is the most probable function.

Pros

- Well established
- Sequencing costs are relatively cheap (~50,000 reads/sample)
- Only amplifies what you want (no host contamination)

Cons

- Primer choice can bias results towards certain organisms
- Usually not enough resolution to identify to the strain level
- Need different primers usually for archaea & eukaryotes (18S)
- Cannot identify viruses
- No **direct** functional profiling

- **Shotgun Metagenomics:** We study everything that is present in the sample. The information we obtain is:

- Which species and strains are present?
- What is the community's functional potential?

Pros

- No primer bias
- Can identify all microbes (e.g. eukaryotes, viruses)
- Direct functional profiling

Cons

- More expensive (millions of sequences needed)
- Host/site contamination can be significant
- May not be able to sequence "rare" microbes
- Required computational resources can be restrictive
- More complex bioinformatic analyses required

Sample Multiplexing

MiSEQ: We need to combine multiple samples into a single run.

- Unique DNA barcodes can be incorporated into your amplicons to differentiate samples.

Taxonomic profiling

It's the first step. For every single read, we need to say which is the taxon.

Then we obtain the absolute and relative abundance:

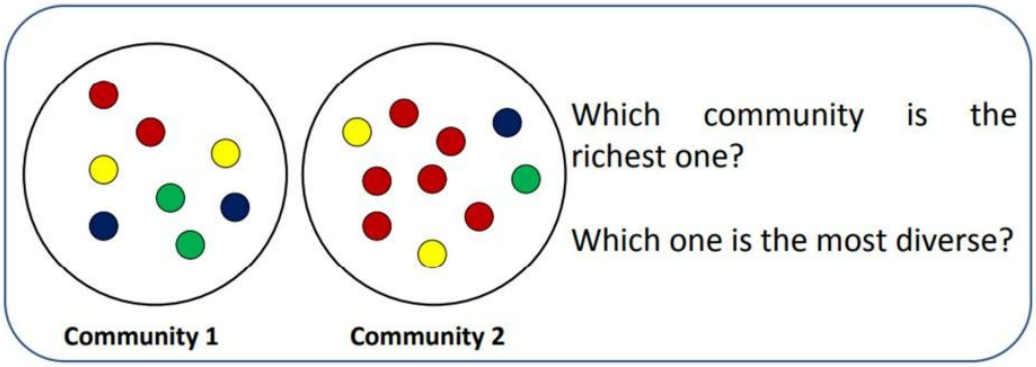
- **Absolute abundance:** Numbers represent real abundance of things being measured. Example: The actual quantity of a particular gene or organism
- **Relative abundance:** Numbers represent the proportion of things being measured within a sample. In almost all cases microbiome studies are measuring relative abundance. Allows us to compare.

Diversity index

Quantitative estimate of biological variability

We need to know which sample is more diverse and there are many ways to do this:

- **Alpha diversity:** Within a particular area, community or ecosystem.
- Richness: Number of species/taxa observed or estimated
- Evenness: Relative abundance of each taxon
- If we are talking about the Alpha diversity, it takes into account both evenness and richness.
- **Beta diversity:** Between ecosystems
- **Gamma diversity:** Overall diversity for different ecosystems within a given region



- **Richness:** total number of species within a community.
- **Evenness:** how evenly the individuals in a community are distributed over all different species. Related to **dominance.**
- **Species abundance distribution**
- **Genetic relatedness** between species detected.
- Other ecological parameters: trophic structure…

Community 1: More even
Community 2: More rich

Rarefaction

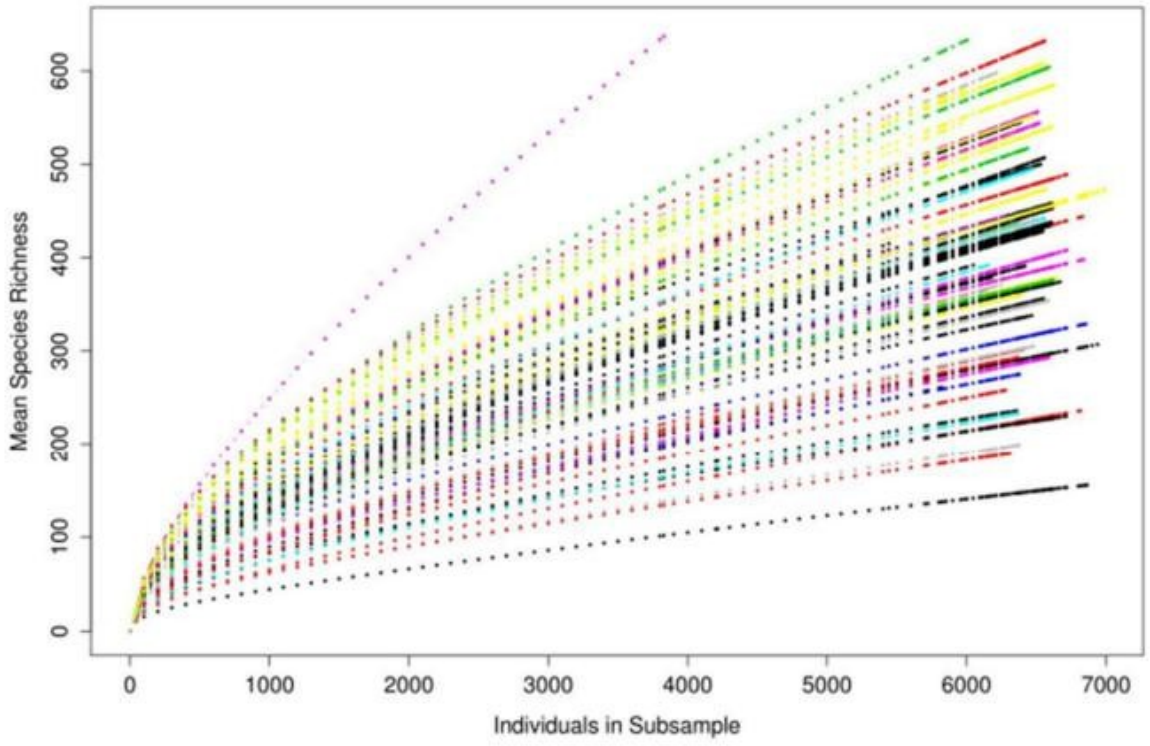
How do we know if we need more sequencing depth or if it is enough? Using the rarefaction curve.

As we multiplex samples, the sequencing depths can vary from sample to sample. Many richness and diversity measures and downstream analyses are sensitive to sampling depth:

- More reads sequenced, more species/OTUs will be found in a given environment

So, we need to rarefy the samples to the same level of sampling depth for more fair comparison

The rarefaction curve compares observed richness among sites that have been unequally sampled by calculating the number of species expected at different sized subsets for each of the sites.



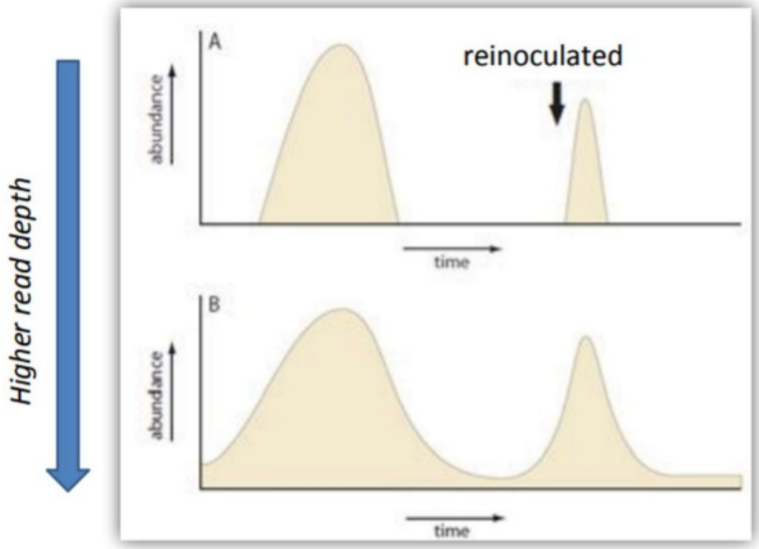
As we can see, we reach a plato. More individuals (reads) does not increase the richness.

Technical challenges

Sometimes, we can lose a lot of information. We need to increase the sequencing depth to detect low abundant organisms.

Read depth

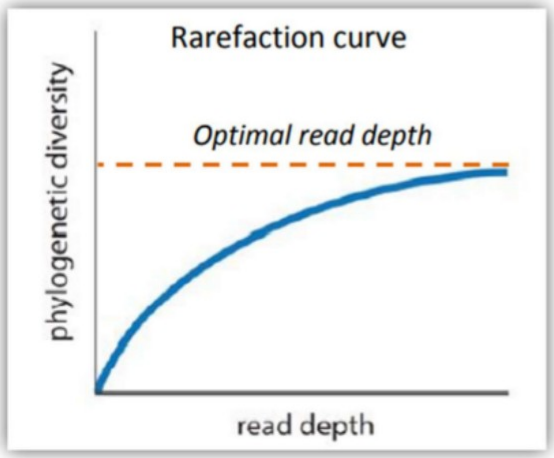
NGS makes it possible to detect organisms that exist in very low abundance within complex populations. These sub-populations can constitute a genetically diverse pool that will survive under changing environments or environmental stress.



The ability to **detect low-abundance** populations can profoundly impact the **interpretation of microbiological changes**

Sequencing errors

The primary goal is to **sequence deep enough** to distinguish low-abundance members of the population from sequencing errors. A low sequencing error rate is important, as well as strict filters to remove sequencing errors.



Sampling effect

ACGTCGCTCGATGGCTAGCTTCGCTG
ACGTCGCTCGATGGCTAGCTTCGCTG
ACGTCGCTCGATGGCTAGCTTCGCTG

ACGTCCCTCGATGGCTTGCTTCGCTG
ACGTCGCTGGATGGCTAGCTCCGCTG

Sequencing errors New species!
(False positive)

The leafy seadragon (*Phycodurus eques*) is a marine fish related to the seahorse. It is the only member of the genus *Phycodurus*. It is native to the waters bordering the Southern and Western coasts of Australia, generally living in template and shallow waters. Your research group wants to collaborate with the Genome 10K project by sequencing the genome of this species for the first time. (total score: 20 points)

Your team is considering different technologies for sequencing to decide which one will be applied in the project. Mention one "Pro" and one "Cons" for each of the six DNA-seq techniques below: (+4 points)

	Pro	Cons
Sanger	High quality	Low throughput
Roche	High Throughput	Worst quality
Illumina	Highly accessible and cheap	PCR amplification bias
Ion Torrent	No need of fluorescence or other technology devices	Bad quality
Pacific Biosciences	One single molecule	High error rate
Nanopore	Can be run with RNA	High error rate

As the budget is limited and your goal is to achieve a high-quality assembly, equivalent to the quality of the human reference genome, which sequencing technique do you recommend to your group?

454/Roche
Oxford Nanopore
A combination of the two above
The objective that you propose is not a realistic goal

Finally, your group invests a large part of the budget in generating the genomic libraries and sequencing. Now you have in your hand billions of Illumina and Pacific Biosciences sequencing reads. Explain which will be the strengths of Illumina and Pacific Biosciences data, and why it is a good idea to combine both: (+2 points)

Illumina: Provide short reads (~200 bp – 1500 bp) of better quality than a third generation technique and they are of high throughput.

Pacific Biosciences: Provide long reads. These reads can be produced very fast since they are sequenced in real time.

Why combine both: The combination of both approaches will help to identify long regions and therefore, more resolution (thanks to Pacific Biosciences) while checking small reads that are not taken into consideration and ensuring a better quality (thanks to Illumina).]

It is time for assembly and you are still discussing which assembly software you are going to use. Which assembly strategy are you going to follow? Why?

- Mapping against a reference
- De novo assembly
- Any of the two above
- Expression profiling

Since we are going to sequence a genome for the first time, we have no reference genome to compare it to (so we cannot do mapping against a reference). Also, we want to assemble a genome so there is no need to assess the expression now.

Finally, you get two separate assemblies of your sequencing data, made by two different assembly software. The first thing you do is compare basic metrics between the two. According to the values shown in the table below, which assembly looks best? Why?

	Velvet	SOAPdenovo
Number of contigs	120,479	47,571
N50 size (bp)	7,338	17,425
Longest contig (bp)	21,684	468,339

SOAPdenovo because it presents less number of contigs and we will have a less fragmented assembly (so there must be less sequencing error, because more contigs means more sequencing and therefore more errors). N50 is larger, meaning that there are longer and more continuous sequences in the assembly (more resolution). It also has the longest contig, so there is more resolution.

Considering that you have assembled 132.13 Gb of sequencing data and that the estimated genome size of the leafy seadragon is 695 Mb, calculate the redundancy (coverage):

Redundancy = (N·R)/G = 132130 Mb / 695 Mb = 190.12

N = number of reads
R = average read length
G = genome size

You decide to continue with one of the two previous assemblies. Now, to complete the assembly and form scaffolds it is essential to:

- sequence paired-end reads.
- eliminate repetitive regions.
- sequence the transcriptome by RNA-seq.
- compare contigs with a database of proteins of a nearby species.

The sequencing of a diploid species such as the leafy seadragon reveals sites in the genome where the individual has two different alleles in the form of a polymorphism. How do you think these sites can be detected?

In the Illumina reads, heterozygous sites have an intermediate coloration between the two nucleotides corresponding to the two alleles.
In the assembly, heterozygous sites have approximately half of the reads with one allele and the other half of the reads with the other allele.
In the assembly, heterozygous sites have double the redundancy (coverage) than the rest.
The sequencing and assembly of a diploid individual results finally in 2n chromosomes assembled separately, so that heterozygous positions correspond to the differences between the two chromosomes.

Mention and describe another application of DNA sequencing:

Another important application of DNA sequencing is the identification and characterization of genetic variations within individuals or populations. This process is known as genotyping or variant calling, and it has various applications in research, medicine, and agriculture.

Genotyping refers to the determination of genetic variations, such as single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations, within an individual's genome or across a population by mapping against the reference genome. It involves comparing DNA sequences from multiple samples to identify differences at specific genomic positions.

A second stage of the genome project of the leafly seadragon is related to RNA-seq. Explain how will you process RNA-seq reads what information does transcriptomic data provide you.

To get RNA-seq reads, we need to first get the mRNA transcripts and convert them to cDNA in order to sequence them (or we could directly just use Oxford Nanopore).

Then, these reads should be mapped against the genome, however, they are spliced so we need to map the different parts of the transcript against the genome. With the help of junction reads we can infer where the introns are and therefore to establish the boundaries of these exons. Taking into account these RNA reads, we will be able to determine the levels of expression of the reads.

The study of RNAs gives information about:

- Genes and other expressed sequences of a genome
- Gene regulation and regulatory sequences
- Function of the genes and their interaction
- Functional differences between tissues and cell types
- Identification of candidate genes for any given process or disease
- Gene expression for a given condition

The figure below displays EST and RNA-seq data mapped to a given genomic region. (+5 points)

How many genes does the genomic region contain? 1 gene

Do/does the gene(s) show(s) alternative splicing? Yes.

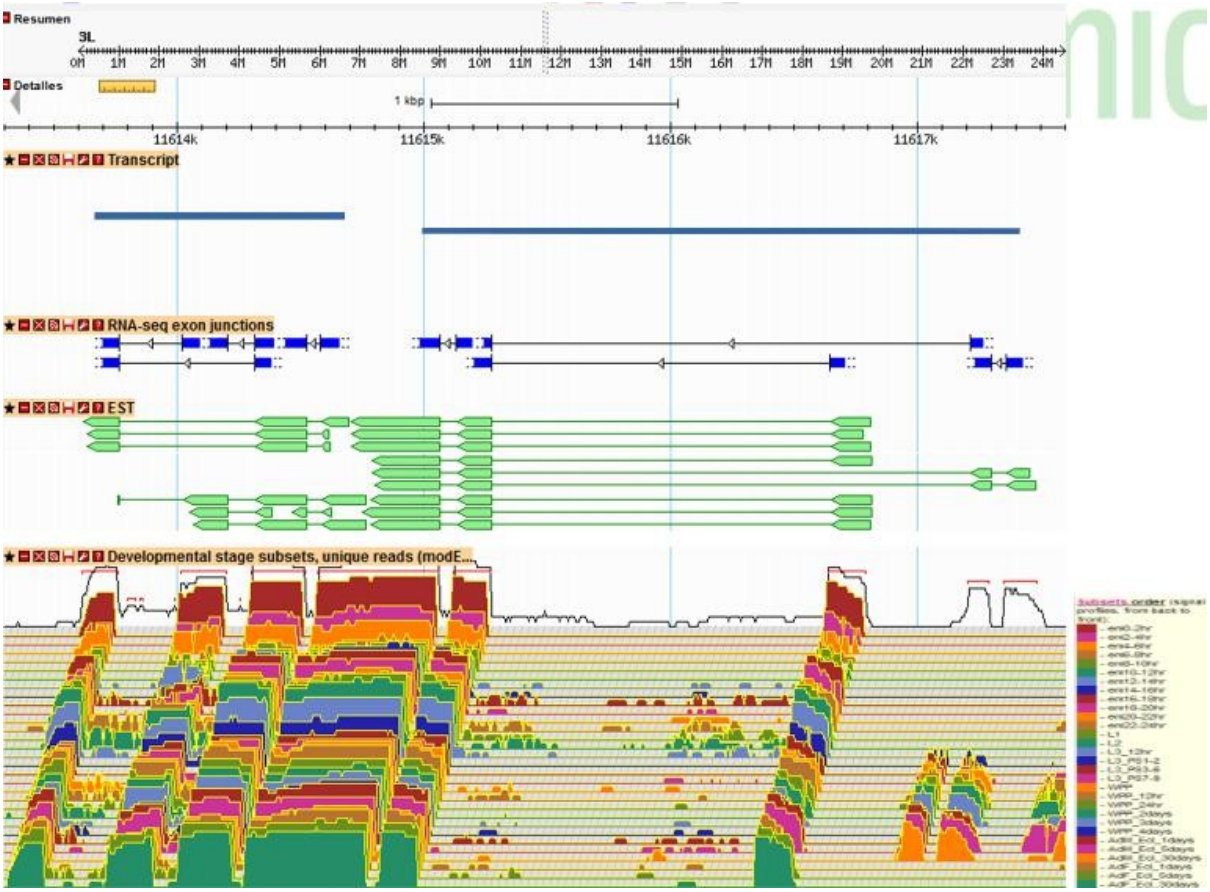
Draw all the transcripts in the reserved space within the figure.

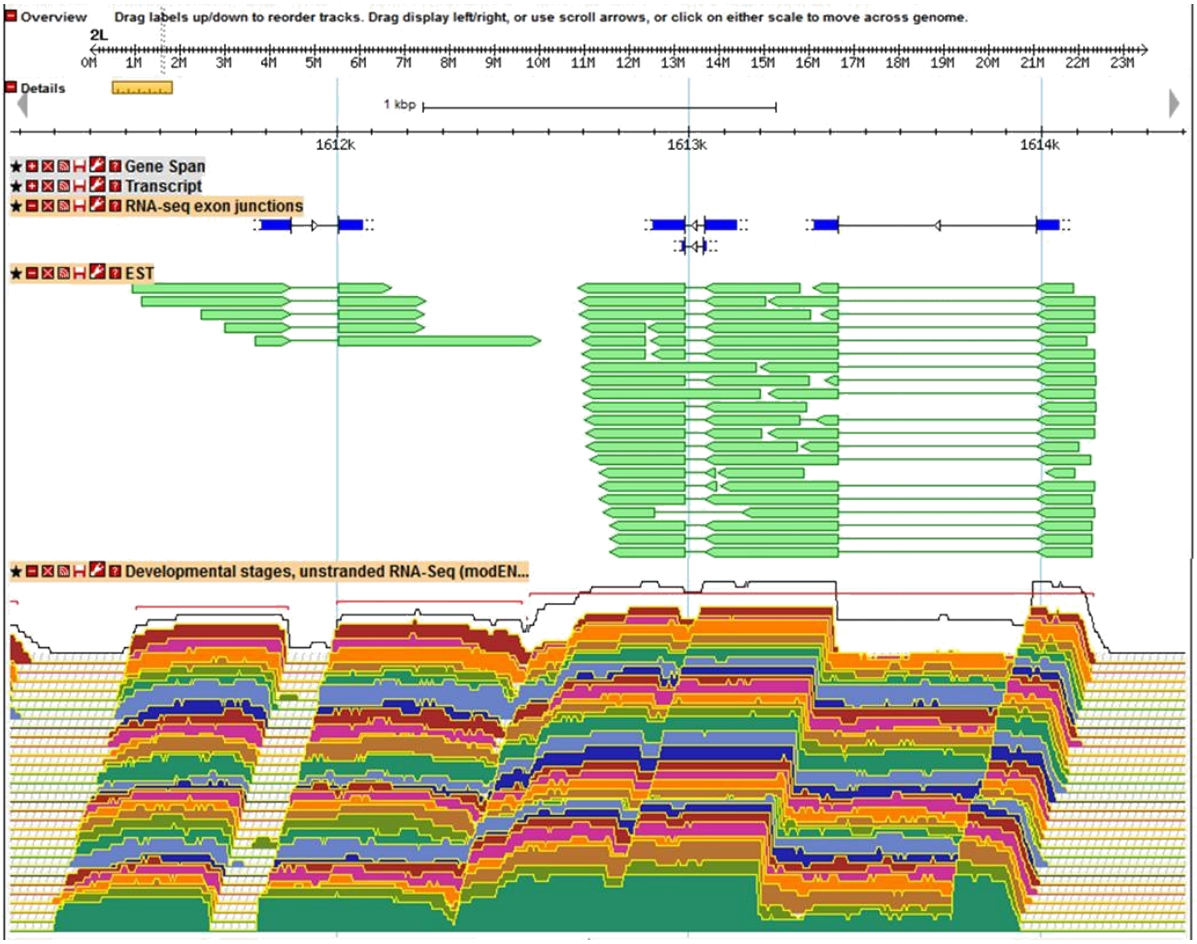
What alternative splicing mechanisms are used to generate the different transcripts? Enumerate them and mark the place where they occur in the figure. Alternative transcription start-site, skipped exon

Do the different transcripts show differential gene expression throughout development? Yes. Some transcripts are expressed only during stages of L3 until the initial AdM phase. Also, notice in some early embryo phases, the expression levels are reduced among all transcripts.

Are all the proteins encoded by the different transcripts identical? No, they are not since they are composed of different exons.

Mark in the figure the beginning and the end of the translation of each transcript. We cannot know the beginning and end of translation of the transcripts. That is because transcripts contain 5'UTR and 3'UTR regions (untranslated regions) and they are not indicated in the EST data therefore, we are unable to identify them.





You want to sequence an eukaryotic genome never sequenced before. Your budget is limited and you decide to make a whole-genome shotgun sequencing with a next-generation sequencing technique. If you could choose one sequencing technique, which one would you recommend? Why?

Illumina sequencing technology:

- Cost-effectiveness
- Established technology
- High throughput

Would it be a good idea to combine two sequencing techniques? Which ones would you combine? Why?

Yes. Illumina (short-read) sequencing + Oxford Nanopore (long-read) sequencing to solve the gaps

What do you need to form scaffolds? Briefly explain the process.

We need contigs in order to form scaffolds and reads to form contigs. (Contigs are a set of reads that overlap in their extremes to generate longer contiguous sequences and scaffolds are oriented and ordered contigs based on the information from PEM.) So, first we have the reads which create contigs overlapping within them, and based on information from Paired-end reads we can order and orient those contigs forming the scaffolds.

Paired-end mapping (PEM) is another application of DNA sequencing. Describe the aim and procedure of the PEM technique.

Paired-end mapping (PEM) is a DNA sequencing technique that involves the generation of paired-end reads from DNA fragments. The aim of PEM is to improve genome assembly. The information from paired-end reads helps in linking contigs or scaffolds, resolving repetitive regions, and improving the overall contiguity of the assembled genome.

Procedure:

- Library preparation: Fragmentation of DNA + adaptors
- Sequencing using Illumina
- Pair-end read alignment: The generated paired-end reads are then aligned to a reference genome or assembly using bioinformatics tools. Each read consists of two sequences, one from each end of the DNA fragment. By aligning the paired-end reads to the reference genome, the relative positions and orientations of the DNA fragments can be determined.

I am providing paired-end mapping data for three fosmid sequences. Do they reveal the presence of structural variants in any of the regions? Specify the type and approximate size (if possible).

READ	# HITS	BEST HIT					STRUCTURAL VARIANT?
		IDENTITY	CHR	STRAND	START	END	
F1 fwd	4	99,2%	7	+	117133465	117134193	Possible insertion ~1,2 kb
F1 rev	8	99,4%	7	−	117172022	117173660	
F2 fwd	87	96,3%	19	+	21837776	21838534	Insertion
F2 rev	182	98,2%	19	−	21868365	21870073	
F3 fwd	7	100,0%	X	+	153560230	153560952	Inversion
F3 rev	7	98,0%	X	+	153586670	153587167	

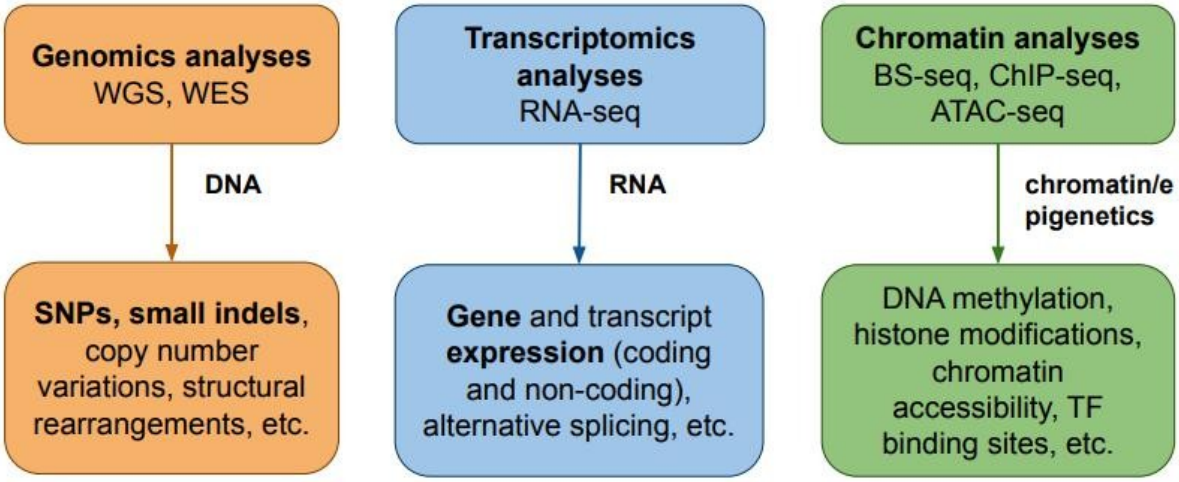
Topic 1. The OMICs data analysis process

OMICs data (NGS recap)

We are going to be talking about Illumina sequencing, so we will have short reads.

Bulk data: Refers to genetic or genomic data obtained from a population of cells or tissue as a whole. The genetic material is collected from a large number of cells, and the data obtained represents the average expression or genetic information from the entire population. We are going to work with this.
Single cell data: Refers to genetic or genomic data obtained from individual cells. The genetic material is collected from individual cells, and each cell's expression or genomic information is analyzed separately.

Things that we can do with this type of data:



Re-sequencing: We can use WGS to obtain or identify different states in the samples of our patients. We have the reference genome, we sequence our patient and we compare. We will do this!

De novo sequencing: I get a new organism with no reference sequence.

Experimental design

We are going to discuss how we assign samples to different conditions. As bioinformaticians we do not have control over this, we can just give advice to the guys in the lab.
The experimental design of the experiment is the only part that once done, we are not able to change it.

Experimental designs can include many experimental factors with multiple levels:
- Type of tissue: Normal, cancer…
- Drug doses: None, low, high.

We must also take into account additional covariates:
- Experimental batch effects
- Sex

If we have 2 illumina lanes, we should not put all the control groups in one lane and all the experimental groups in the other lane (experimental batch effect). We can alternate or randomize control and experiment.

Raw data processing: GC, mapping, quantification

Raw data processing normally involves:
- Quality control: The software used for QC of raw reads is FastQC
- Mapping the reads to the reference genome
- Quantification of the reads that are mapping (count the number of reads mapped to a gene). Of course, we will need to normalize this value, correct the different biases…

The nature of read count data

Which type of data are we generating? We are in a RNA-seq experiment, so we will get read count data.
- Number of regions/variables (genes, exons, transcripts… for example) > number of samples (individuals, for example). There are 20.000 genes per individual.
- Normally, variables are in rows and samples in columns
- Discrete, positive and skewed data (we have more 1 read count than 1000 read count, so the plot is skewed to the left)
- Mean-variance relationship. The larger the mean, the larger the variability. So, a gene that is highly expressed, it will also be very variable in comparison with a gene that is lowly expressed.
- Large dynamic range with presence of 0 counts. There are genes that do not have counts.
- The total number of sequences per sample (=library size) is not the same for all the samples.
- Maybe one sample is sequenced 20 million reads, another 40 million reads, another 1 million reads.
- If we say that the number of reads that fall in a particular gene, is the expression of this gene. If one sample is sequenced a double times another sample, we will say that it has the double of expression, which is false.

Statistical analysis: Data transformation

There are different things we can do with this data:

- **Class comparison:** We have RNA-seq samples from cases and controls for a particular disease. We are going to do a differential gene expression analysis (class comparison).
- **Class discovery:** We look at the expression in a supervised manner and we try to find groups of patients. We sequence 100 patients with different degrees of severity of a disease and then we find subcohorts that behave similarly. So, we are finding groups in our data.
- **Class prediction:** I take the 100 most variable genes between categories and I use them in a machine learning model to predict if a sample is from a patient, from a control…

Before doing this, we need to preprocess the data using data transformation (normalization if we talk wrong).

- Data transformation is correct because normalizing means transforming data to a normal distribution. But here we are just making data comparable, we are not obtaining normal data.

We have many different categories of methods for doing this:

- Scaling factors
- Variance stabilizing
- Unwanted variation removal

Normalization

Normalization is a process designed to identify and correct technical (and biological) biases removing the least possible biological signal. This step is technology and platform-dependant.

When we are doing an experiment we have some noise (technical and biological) and the signal. For example, when blocking Sex factor in the model, Sex is a biological reason why we observe the differences, but it is not the signal we want to detect. Thus, we need to correct it.

Some biases may be controlled by a proper experimental design or a good experimental protocol. It may happen that we did not do a good experimental design and we need to correct it. Moreover, there are sources of variability that we can not control, like technical biases.

Normalization aims to correct systematic uncontrollable biases such as those induced by the sequencing process.

- **Within-sample** normalization enables comparisons of features (genes, transcripts, exons) from the same sample → Gene length, sequence composition (GC content. sequences with high GC content will map better)

We need to do some normalization to make different genes comparable.

- **Between-sample** normalization enabling comparisons of features from different samples (we will talk about this). We want to compare the same gene across different samples with different conditions → Sequencing depth (total number of sequenced and mapped reads), sequence composition due to PCR-amplification, etc.

Overall strategy: We choose an appropriate baseline, and express sample counts relative to it.

Many different normalization methods. Some of them are part of commonly used approaches for differential gene expression analysis (edgeR, DESeq2). They make different assumptions and have different advantages and limitations.

- In some cases, the software that performs the analysis requires a particular normalization and it will perform it in the same step.
- So, there is a relationship between normalization and downstream analysis.

In experiment A, longer genes (in terms of exon length) will get more reads just by chance than small genes.

In experiment B with a high number of mapped reads (larger library size), a gene will get more reads than in an experiment with a small number of mapped reads.

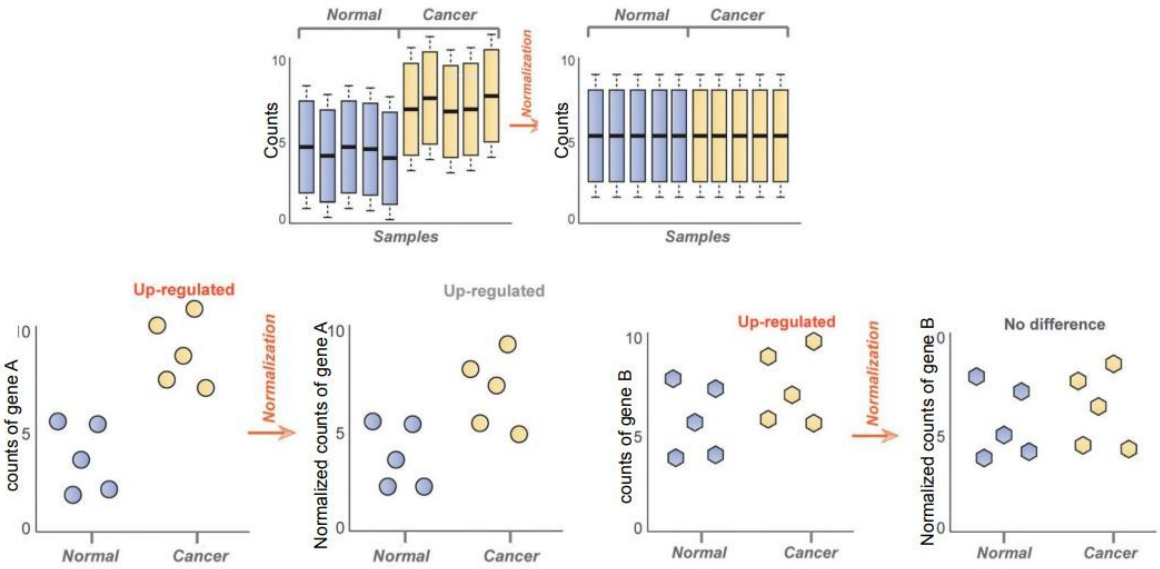
There are different metrics that take into account this facts to get normalized measurements of gene expression. Instead of saying a read is expressed 15 reads or counts, we say 15 CPM or RPKM…

Transformation methods. Scaling factors

Scaling factors is just dividing our counts by the source of variability we want to correct.

- Gene length
- Library size

We will be removing false positives and false negatives if we normalize.



Examples:

- CPM (Counts Per Million)
- RPKM (Reads Per Kilobase of exon model per Million mapped reads)
- TPM (Transcripts per million)
- Quantile normalization
- TMM (Trimmed Mean of M-values, edgeR)
- RLE (DESeq2)

$$CPM_{ij} = \frac{y_{ij}}{N_j} \cdot 10^6$$

$$RPKM_{ij} = \frac{y_{ij}}{N_j \cdot \frac{L_i}{10^3}} \cdot 10^6 \quad TPM_{ij} = \frac{RPKM_{ij}}{\sum_{i=1}^{G} RPKM_{ij}}$$

$y_{i\,j}$ = Gene expression in terms of counts that has gene i in sample j

N_j = Library size of sample j
L_i = Gene i length

So, **CPM** normalizes per the library size. This is a useful measurement when we want to compare a gene in 2 technical replicates

- **Technical replicate:** I have a blood-sample from the teacher and I extract the RNA. I split this sample in 3 (3 technical replicates) and then I do 3 different sequencing experiments, each one in a different lane. The variability is due to the technique that we are using.
- **Biological replicate:** I want to study the effect of a particular product. So, I make 10 people taste my drug and get a drug sample from each of them. So, the condition is the same for each person, but each person is different. So, the source of variability is each person (biological replicate).

So, if we have 2 technical replicates, the best measure to compare them is not to use counts (different technical replicates may have different library sizes) but CPMs.

In CPMs we are assuming that read counts are proportional to the library size.

What happens if we have 2 different genes? Imagine I want to compare gene expression in gene A and gene B (effect of gene length) in 2 different samples (effect of library size): In this case we also need to correct by gene length and library size, so we use **RPKMs**.

In RPKMs we are assuming that read counts are proportional to the library size and gene length.

We may also find **FPKM** = Fragments Per Kilobase of exon models per Million mapped reads. Paired-end experiments produce two reads per fragment (not necessarily both reads will be mappable, avoiding double-counting some fragments).

If the experiment was done perfectly, FPKM = 2 · RPKM.

A neuron does not have the same number of transcripts as other cells.
If we want to compare the expression of a gene in 2 different cells, the total number of transcripts that we have in the cells is going to be different (proportional to the number of counts that we get).

Even in the same cell, we may have some points in the cell cycle that has more transcripts than others. So, we have to correct for the total number of transcripts that we have in the cells to make the RPKMs comparable between different cell types, different tissues, different cell cycle moments…
In this case, we use **TPMs**. We are normalizing for the total number of transcripts we have in the cell.

Assumption: read counts are proportional to expression level, gene length (only in RPKM/TPM) and sequencing depth (same RNAs in equal proportion).

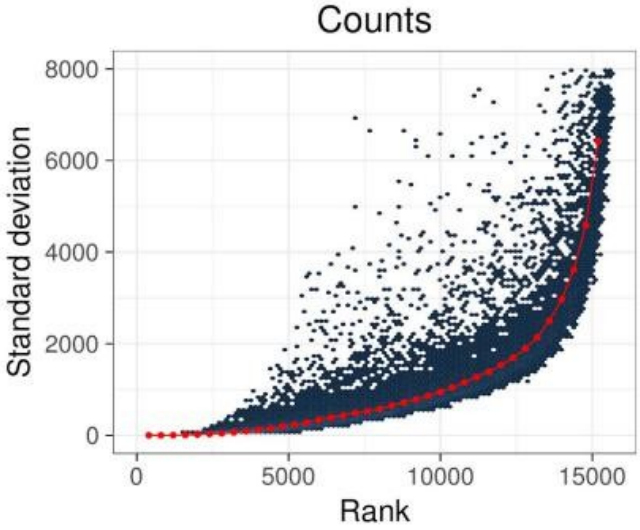
CPM and RPKM assume that the absolute amount of total RNA in each cell is similar across different cell types or experimental perturbations, which is not always the case (Loven, 2012). TPM corrects for that.

RPKM and TPM are commonly used for visualization purposes, but do not have good statistical properties for DGE analyses.

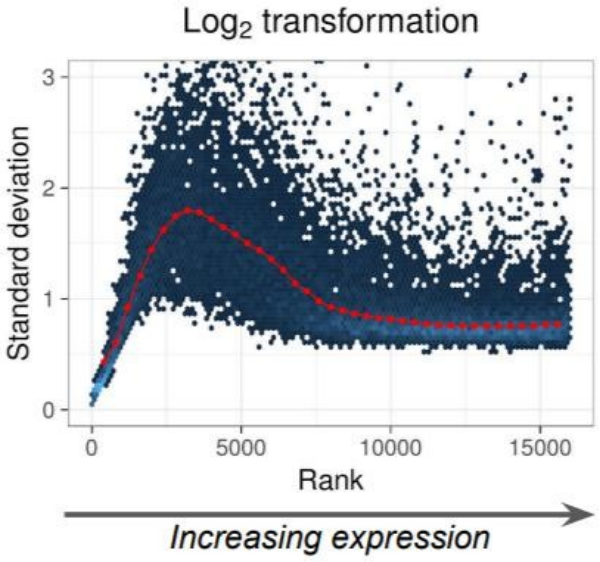
Transformation methods: variance stabilizing

Variance stabilizing means breaking the relationship we have in the data.

Here we have the SD for each gene (points) vs the mean expression (Rank).
- The genes that are lowly expressed have a low variability.
- The genes that are highly expressed have a high variability



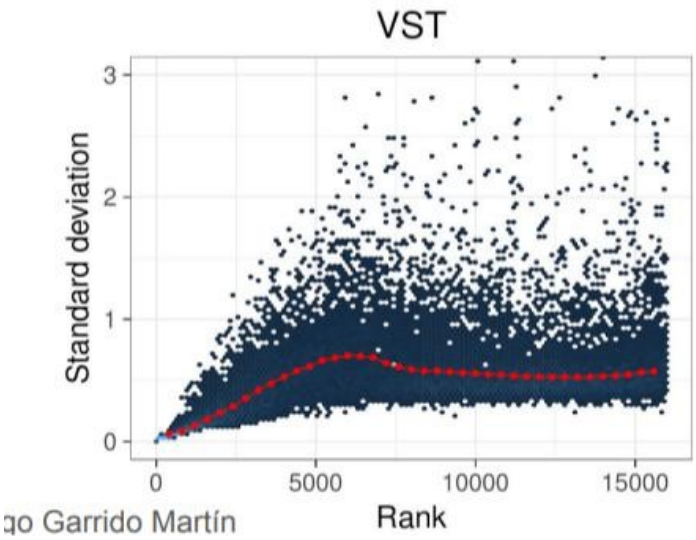
One simple way to break this is applying a log transformation and we will obtain the data less skewed.



This will inflate the variability of genes that have low expression.

Another transformation used in the DESeq2 package is the VST transformation, which is essential for the techniques we will use afterwards:

- Differential Gene Expression
- Visualization
- ML



Transformation methods: Unwanted variation removal

We have to use this when we don't do a good experimental design.

I am analyzing telomeres and I should have considered Age. The solution is to apply methods afterwards to try to guess which unknown are affecting my data.

Examples:

- PEER (Probabilistic Estimation of Expression Residuals)
- RUV (Remove Unwanted Variation)
- SVA (Surrogate Variable Analysis)

Exploratory data analysis: Dimensionality reduction

As we said, there are 2 steps prior to the statistical analysis that we want to do:

- Data transformation
- Exploratory data analysis
- Dimensionality reduction: PCA, MDS, tSNE, UMAP, biplots
- Clustering: K-means, hierarchical, DBSCAN

Dimensionality reduction can be applied for many different tasks. The idea is that I have many variables (20.000 genes) and we want to summarize this by losing the least variability as possible.

By using PCA, multidimensional scaling (MDS), UMAP, tSNE… we remove many variables without losing a lot of information. We only use this for data visualization, not for other tasks.

Why do we need data visualization based on dimensionality reductions? Basically for batch effects (identify which factors or sources of variability that we do not want and thus we need to correct in our models) and outlier identification.

- Imagine that we want to guess if including Sex in our model is relevant or not.

Exploratory data analysis: Clustering

We have an input data:

- N samples by P genes

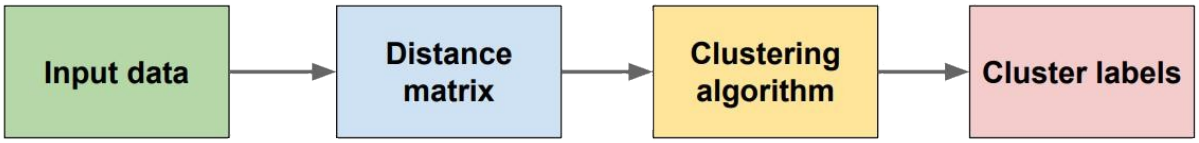
We can calculate the distances between genes or samples and obtain a distance matrix for the genes or for the samples.

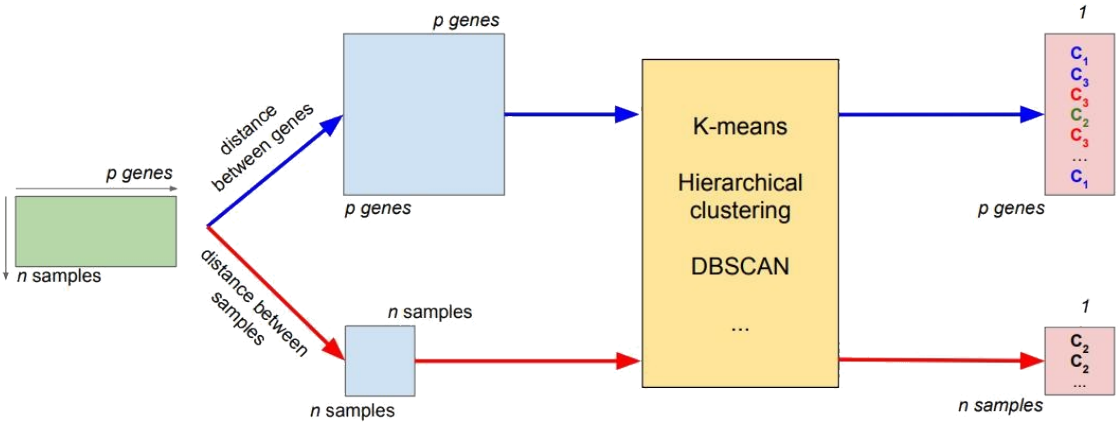
Then using any clustering algorithm, we are going to get labels for the samples.

- For example, samples 1, 6 and 7 correspond to cluster 1.

Again, we are going to use clustering to detect outliers and to find factors that may be associated with our expression and we do not want this effect.

- If there is a sample that clusters very far from the others, then it may be an outlier
- If we observe clusters that correspond to the Age or other variables we did not include, then we will have to correct it.





We have transformed our data, we have done exploratory data analysis (identifying outliers and factors of interest) and then we focus on the actual task of our analysis which is class comparison, discovery or prediction.

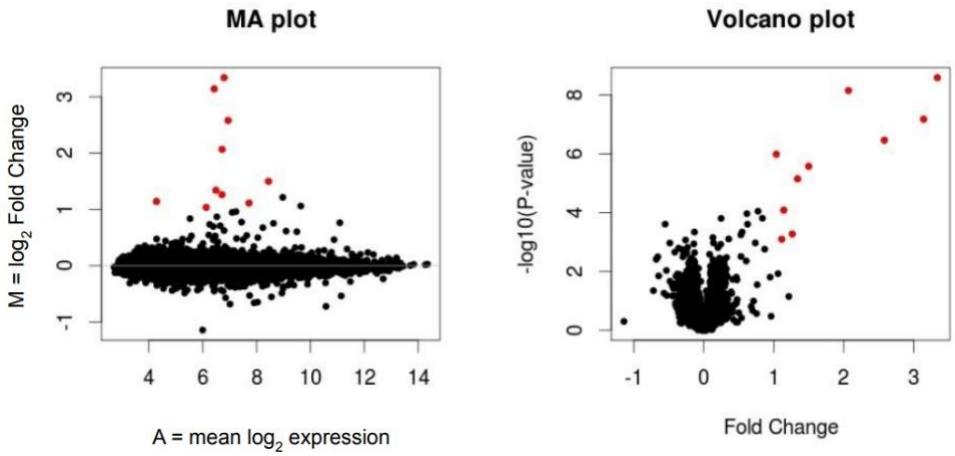
Statistical analysis: Class Comparison

We will focus more on class comparison, in particular differential gene expression analysis.

How are we going to compare conditions and test hypotheses? We are going to see if 2 groups are different or the same regarding their mean, for instance using T-test or other statistical tests.

All the statistical tests that we know can be implemented through linear models. So, we will be describing the linear models and how we can test the hypothesis using linear models.
In particular, for the data in which mean and variance is related we will use generalized linear models (GLMs).

We will also do multiple testing, because we are testing 20.000 genes.



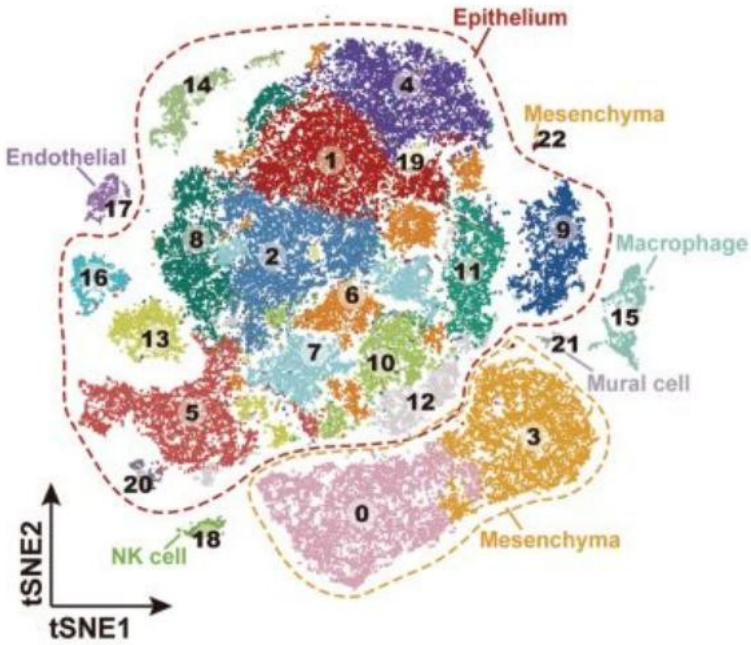
Statistical analysis: Class discovery

When we do some clustering we are assigning labels and we can discover subcohorts of our data that behave similarly depending on the expression of the genes.

- For example, we have selected 100 genes that are relevant for the disease and we may cluster our patients based on that. Identifying patients that have expression of a particular subset of the genes.

Identification of expression patterns

- Dimensionality reduction
- Clustering



Statistical analysis: Class prediction

We have some genes with a measured expression and we want to use their expression to predict prognosis or if our patient is a control…

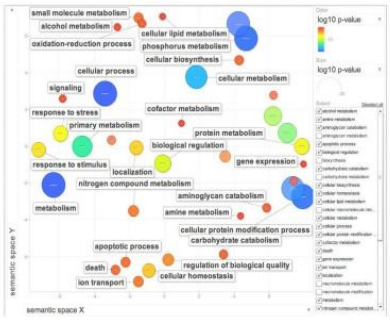
Biological significance analysis

We have addressed our question, we have done some statistical testing after correcting our data from the different sources of variation that we do not want and we have obtained a list of genes with their P-values…

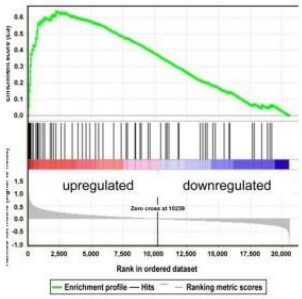
Now we need to know which is the biological significance of the result we have obtained.

The biological significance assessment can be summarized in this workflow.

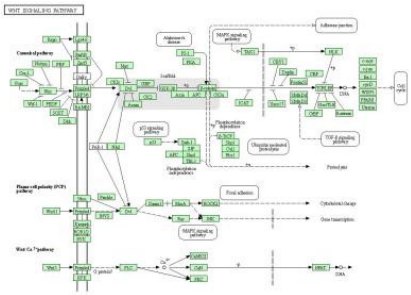
Fisher's exact test (hypergeometric test)
Kolmogorov-Smirnov goodness-of-fit test/permutations



Gene Ontology enrichment



Gene Set Enrichment Analysis



Pathway analysis

When we do a DGE analysis, we do Cancer vs Control and we get a list of genes and a P-value.

- Gene A, has X P-value
- Gene B, has Y P-value
- …

We do multiple testing correction and we get:

- These 1000 genes are significantly differentially expressed
- 200 out of the 1000 genes are more expressed in patients than in controls

What should we do with these 200 genes? We can look at the functional DB (Gene Ontology and KEGG for pathways) each of those genes, but it will take too much time.

So, we have to perform different statistical tests to identify whether our list of genes is over-enriched in some functions with respect to a group of other genes.

We are looking at overrepresentation of some functions in these 200 genes.

Bioconductor

This is the framework we are going to be using to analyze our data or other omics data.

Bioconductor is a software project for the analysis of biomedical and genomic data.

- Provide access to powerful statistical and graphical methods for the analysis of OMICs data.
- Facilitate the integration of biological metadata (Entrez, PubMed, ENSEMBL, GO) in the analysis of experimental data. You can do a GO enrichment directly from bioconductor, for example.
- Allow the rapid development of extensible, interoperable (this is good and bad if you need to do another thing), and scalable software.
- Promote high-quality documentation and reproducible research.
- Provide training in computational and statistical methods.

Packages

- **Software packages:** provide implementations of specialized statistical and graphical methods.
- **Data packages:**
- Annotation Data: provide mappings between different molecule types and annotation databases. May be organized by platforms, organisms or databases
- Experimental data: code, data, and documentation for specific experiments or projects.
- **Workflow packages:** Simple and functional workflows illustrating basic analysis pipelines for different data types and problems

ExpressionSet class

The expressionSet class was introduced to store information from expression microarray experiments, and later it was extended to other OMIC data types.

Its an object that has different matrices linked:

- Matrix expression data → It's our expression data, for instance genes and samples.
- Matrix with the sample metadata in the columns (pData function)
- Matrix with the feature metadata in the rows (fData function)

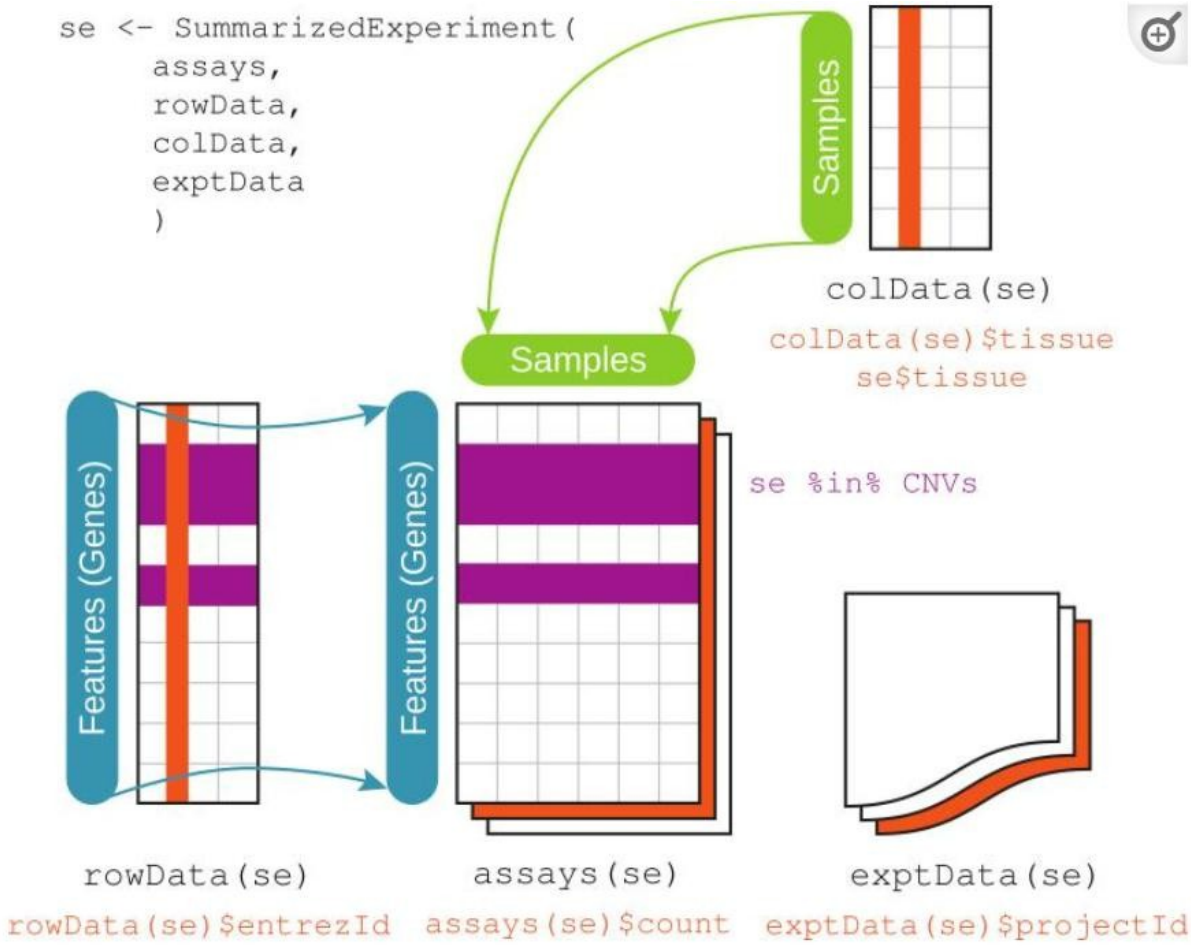
The matrix of expression data is stored in the exprs slot and you can access that with the exprs function. The phenotypic data can be accessed using pData and this gives us a data frame, which is information about the samples. The feature data is accessible with fData function and this is information about genes or probe sets.

There are some advantages to using this. For example, if we subset 3 samples, they will be subseted in all matrices.

SummarizedExperiment class

It's very similar to the other class, but there are 2 key differences. Again, the structure is based on the 3 matrices but:

- We can have more than one assay on the same samples and genes. So, we can have gene expression, open chromatin on the same genes of the same sample… We can have different assays and therefore we have many more matrices + the 2 metadata matrices.
- Also, we can work with these features (rows) as a GRanges object.
- In the case of ExpressionSet class, the rows are just names.
- In this case, they are ranges, meaning that they are objects and therefore they have properties.



Its assays component is one or several rectangular arrays of equivalent row and column dimensions. Rows correspond to features, and columns to samples. The component rowData stores metadata about the features, including their genomic ranges. The colData component keeps track of sample-level covariate data.

Why should we use Bioconductor? Because many statisticians, bioinformaticians, and computer scientists have spent time writing **methods and algorithms specifically for biological data.**

Linear models for differential gene expression analysis

Experimental design principles

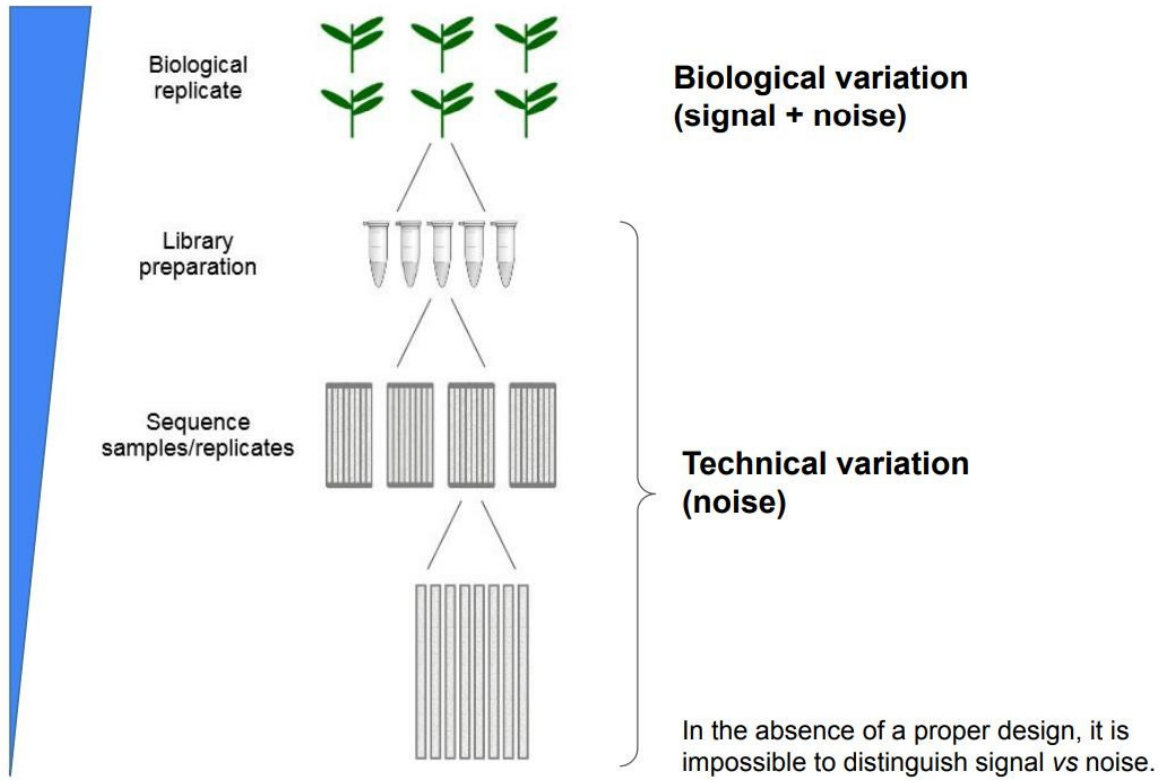
We are interested in the effect of a treatment on experimental units (plants). Many different replicates are assigned randomly to one of two treatments (randomization) then grouped and placed in each field (blocking).

- **Replication**: Measurements are usually subject to uncertainty. Repeated measurements help identify the sources of variation, to better estimate the true effects of treatments and strengthen the experiment's validity. For each field, we have many plants (not just one)
- **Randomization**: It is the process of assigning individuals at random to different groups in an experiment. It reduces bias from sources that have not been accounted for in the experimental design.
- **Blocking**: Experimental units are grouped into homogeneous clusters in an attempt to improve the comparison of treatments by randomly allocating the treatments within each cluster or 'block'. Blocking reduces known but not relevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study.

Sources of variation in NGS experiments

We have biological replicates, which have biological variation (signal + noise).
We make a library preparation and we sequence. Here we will have technical variation (noise).

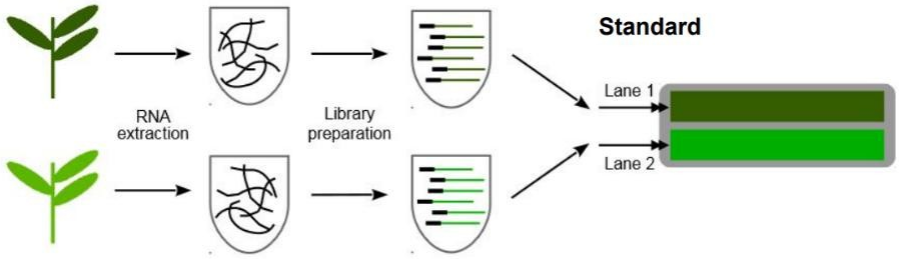
In order to be able to distinguish this signal vs noise, we need to have a proper design.



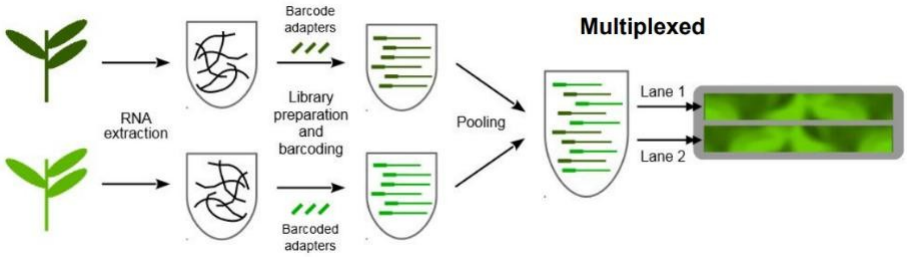
Standard vs multiplex design

Here we have an extra concept that applies for RNA-seq. When we are talking about sequencing, we have our flow cells (each one has a different number of lanes).

We could do some standard design in which we sequence each sample (control and treatment) in a particular lane. We do not have technical replicates.



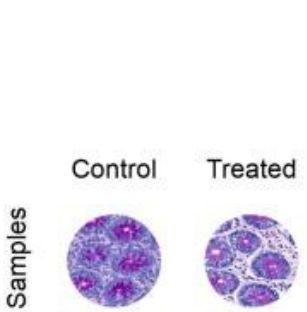
We could also have multiplex designs. We mixt both labeled primers in each lane, meaning that we are sequencing both samples in both lanes. So we have 2 different technical replicates,

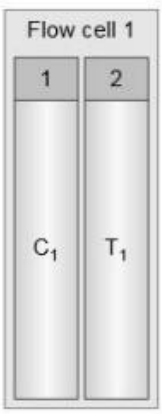


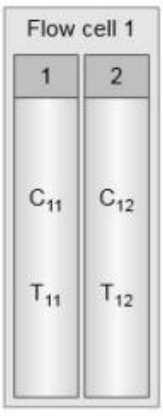
Examples

A biologically unreplicated, unblocked design with RNA isolated from subjects within each group (C1, T1) and loaded into individual lanes (e.g., the RNA from the control subject is sequenced in lane 1).

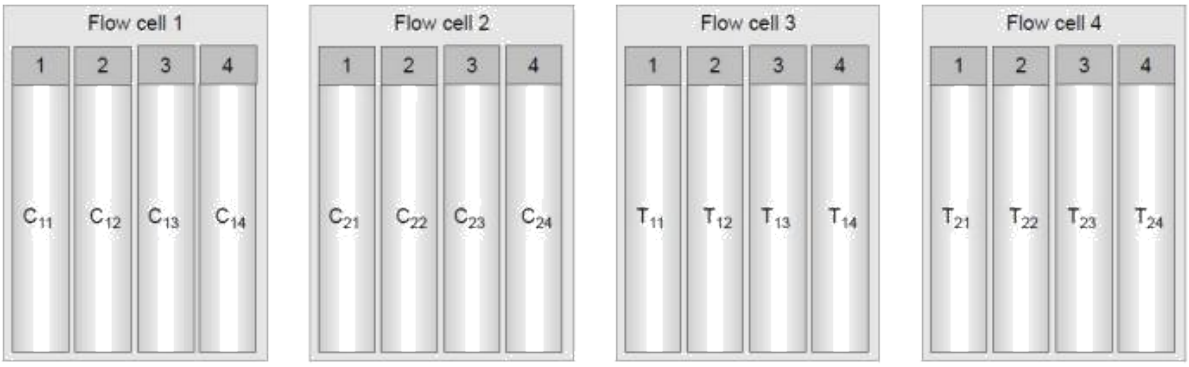
A biologically unreplicated block design with C1 split (barcoded) into two technical replicates (C11, C12) and T1 split into two technical replicates (T11, T12) and input to lanes 1 and 2.







A biologically replicated unblocked design with two treatment groups (C, T), two subjects per treatment group (C1, C2; T1, T2), four technical replicates of each (e.g., C11, C12; C13, C14) and input to four flow cells

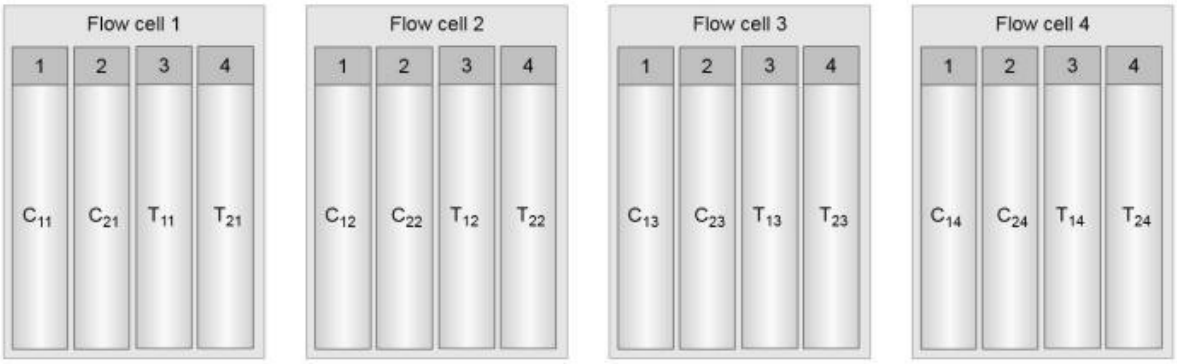


Here we have 2 biological replicates → two samples from two patients that are controls and two samples from two patients that are treatments

4 technical replicates for each lane.

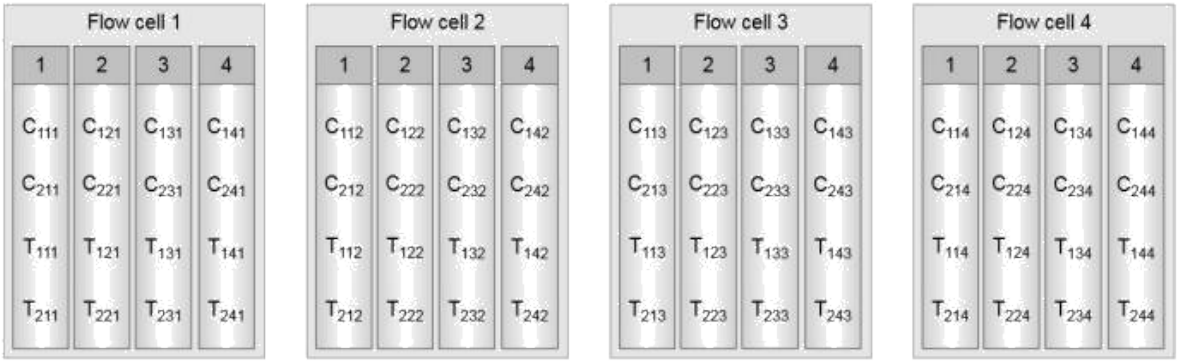
But we do not have blocking!

A biologically replicated block design. Blocking without multiplexing for two treatment groups (C, T) with two subject per treatment group (C1, C2; T1, T2), four technical replicates of each (e.g., C11, C12; C13, C14) and input to four flow cells.



Here we have biological/technical replicates and blocks → Flow cell 1 has both control and treatments (2 technical replicates from controls and 2 technical replicates from treatment).

A biologically replicated block design with multiplexing: each subject (e.g. C1) split (barcoded) into four technical replicates (e.g., C11, . . . , C14) and input to four lanes in each flow cell.



We skip exploratory data analysis because we will see it in the seminar.

Linear models for differential gene expression

Our goal is to model the relationship between a response variable, which is gene expression (quantitative or continuous variable), and a set of explanatory variables X1 , … , Xp-1 that can be quantitative or **categorical**. In the case of differential gene expression, we have categorical explanatory variables.

We usually do that to understand the relationship between Y and the X's or predict future values of Y.

In practice, besides X1 , … , Xp-1 there are always other variables, Z1 , … , Zp-1 that also explain Y but are not known, or are known but not measured. In exploratory data analysis we are going to decide which of the variables should be included.

The value of the response variable Y (gene expression) is obtained as a result of combining a deterministic signal (the part of Y that can be explained by the X's) and a random noise (the unexplained part).

Example 1.

Consider a study comparing two diets in mice: "standard diet" vs "high-fat diet".

If we assume that the weight of the animals (response variable) is a linear function of the diet (explanatory variable), we can write the following linear model:

$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$ for i in 1, …, n

where:

- Y_i is the weight of the ith experimental unit
- β_0 is the intercept
- β_1 coeficient
- X_i is the diet of the ith experimental unit (a factor with two levels: 0 when the mouse receives the standard diet and 1 when it receives the high-fat diet)
- ε_i is the random noise of the ith experimental unit

Aims

- Estimate the mean weight of mice having each diet
- Compare it across diets

Design → We have 3 mice that are from group 0 and 3 mice that are from group 1.

Example 2.

Consider a study on the effect of estrogen on the expression of the genes in ER+ breast cancer cells over time. 8 samples:

- 4 treated with estrogen (T)
- 4 untreated (C).

Gene expression measured at 2 time points:

- 10h (2T+2C)
- 48h (2T+2C).

Assuming linear effects we can write for each gene:

$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \varepsilon_i$ for i in 1 … 8

- Y_i is the expression of the gene in the ith sample
- X_{1i} is the treatment of the ith sample (a factor with two levels: 0 - untreated and 1 - treated with estrogen)
- X_{2i} is the time point of the ith sample (a numerical value, i.e. 10, 48; but also a second factor with 0/1 levels)
- ε_i is the random noise of the ith sample

Aims:

- Compare the effect of estrogen addition on gene expression
- Evaluate how this effect differs after 10h and 48h

From experimental designs to linear models

Examples (1) and (2) have different experimental designs but
- Both can be represented using linear models
- This representation will guide the estimation of the effects, and the comparison between diets or treatments

Linear models provide a convenient setting to describe experimental designs and to analyze data that has been obtained accordingly.

The general linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \ldots + \beta_{(p-1)} X_{i(p-1)} + \varepsilon_i \quad \text{for } i = 1, \ldots, n$$

Let's forget about the dummy variables (0 and 1) and let's focus on continuous explanatory variables.

Linear models assume a linear relationship between the response variable and the explanatory variables and we have model parameters that are unique and unknown (β's). So, we are going to take a sample and estimate these population parameters.

The reason we go through LM instead of doing T-tests, or other hypothesis testing is that LM is the backend that can be used to solve any of these tests.

So many statistical tests can be re-written as linear models.

A note on numerical vs categorical explanatory variables

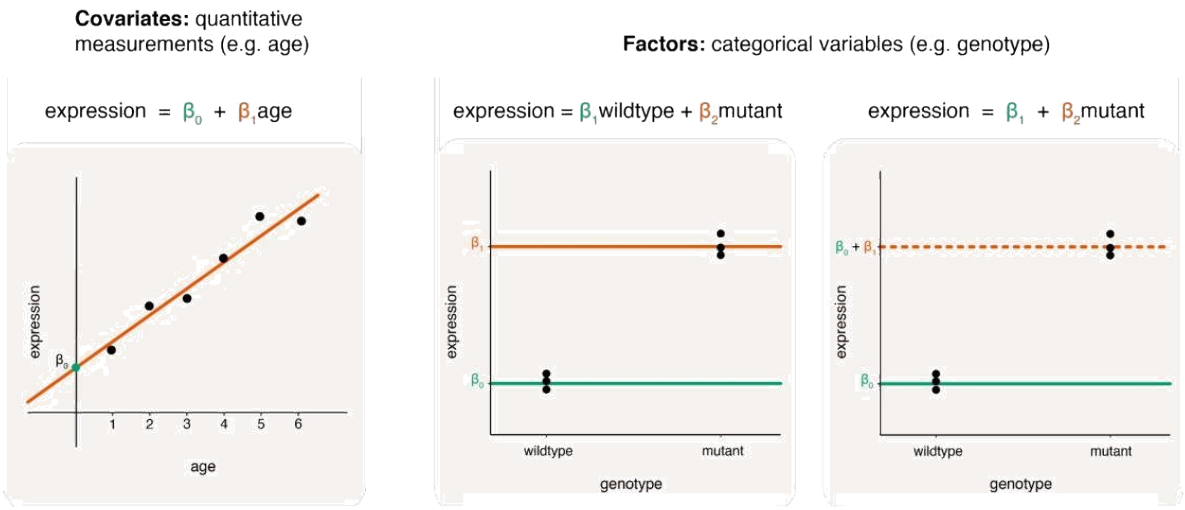
In our case we have categorical variables, the genotype for example:
- WT
- Mutant

Linear Model+GLM
1. Modelling the read counts of a gene across replicates using a Negative Binomial model assumes that:
b. The variance is larger or equal than the mean.

2. Which of the following is true about REVIGO?
e. It enables similarity-based reduc-on of GO term lists and representa-on in 2D.

We could represent the linear model in 2 different ways:
- Expression is the linear combination of the effect of WT plus the effect of the mutant, each one with its corresponding coefficient.
- This is called the "means model" because the estimates of the coefficients correspond to the actual mean of the group.

- Expression is an intercept term plus a parameter that multiplies the value of the mutant. This is called a “means reference model”.
- The basal level corresponds to the mean value of the reference condition and the other estimate is built on top of that reference.

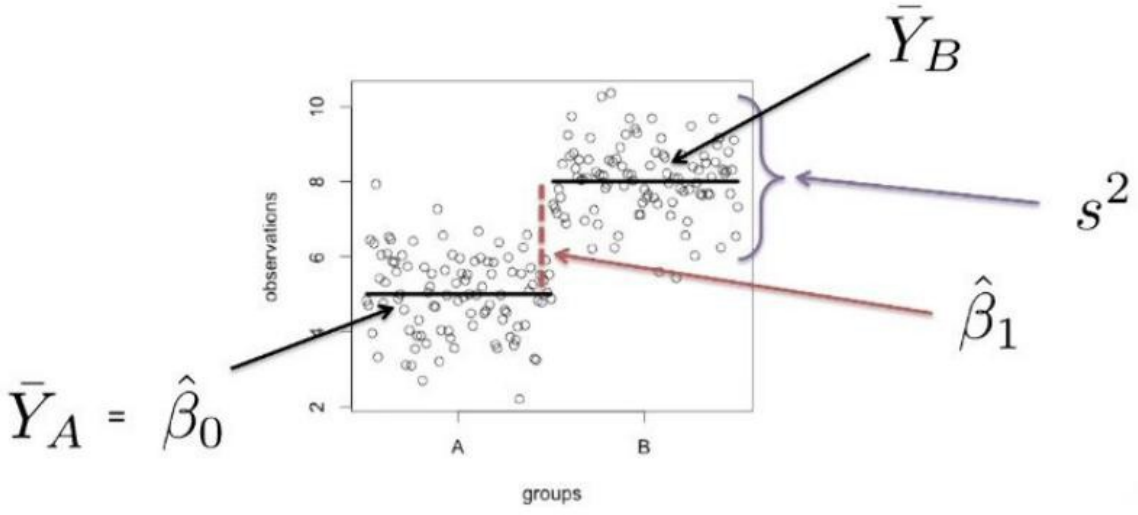


Comparison of 2 groups: t-test

In this case, one of the coefficients is the mean of one of the groups and the other coefficient is the mean of the other group.

So, in this situation, doing a linear model is like doing a T test.

A t-test is an statistical test used to determine if there is a significance difference between the mean of 2 groups.



$$\bar{Y}_A = 1 \times \hat{\beta}_0 + 0 \times \hat{\beta}_1$$

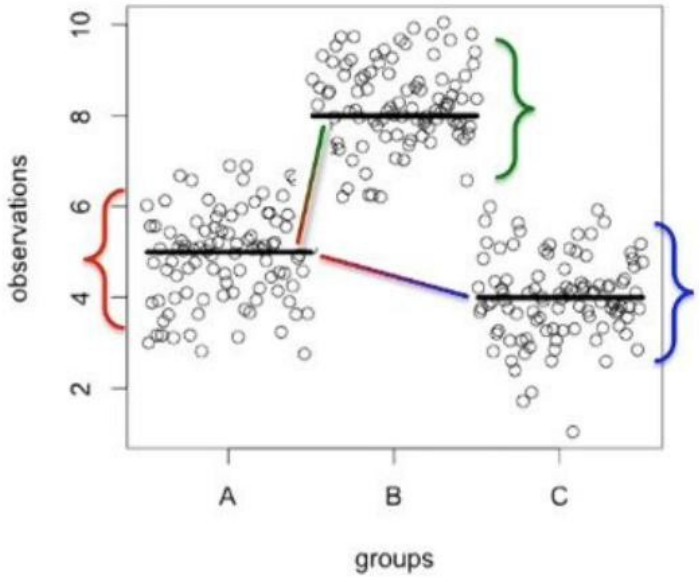
$$\bar{Y}_B = 1 \times \hat{\beta}_0 + 1 \times \hat{\beta}_1$$

Comparison of 3 groups: ANOVA

Here we have a basal level (any of the 3 conditions, in this case A).

Then, with respect to the basal level:

- The mean of condition B is the Basal level + another coefficient
- The mean of condition C is the Basal level + another coefficient



$$\bar{Y}_A = \hat{\beta}_0$$
$$\bar{Y}_B = \hat{\beta}_0 + \hat{\beta}_1$$
$$\bar{Y}_C = \hat{\beta}_0 + \hat{\beta}_2$$

Estimation by OLS (Ordinary least squares)

We need to estimate the value of the unknown βj parameters. The default criteria for most applications is to minimize the residual sum of squares:

$$RSS = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

This is known as the ordinary least squares (OLS) criterion.

OLS is selected because it simplifies computations and it allows us to take advantage of the properties of the normal distribution to do inferences about βj, which are the means.

The values that minimize the RSS are the OLS estimates, and are denoted by j

Matrix notation for linear models

Linear models can be written using explicit expressions, although matrix notation is often preferred:

- More compact
- More efficient computation

Examples:

- Linear model $$Y = X\beta + \varepsilon$$
- RSS $$(Y - X\beta)^T (Y - X\beta)$$

Matrix notation (2 groups)

$$Y = X\beta + \varepsilon$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

The matrix X is called the design matrix.
Selecting a design matrix is a critical step in linear modeling, as it determines the model's parametrization

Linear Model+GLM
3. Which of the following is false regarding usual representations of DGE analysis results?
e. The best way to summarize the results of a DGE analysis is represen-ng boxplots for each gene

4. Which of the following is true regarding the voom-limma approach?
c. None of the answers is correct. (a. limma learns the mean-variance trend and adjusts it in the context of standard linear models (implemented in voom), d. It tipically requires as input TPMs or RPKMs, e. It does not require the defini-on of design and contrast matrices.)

Matrix notation (3 groups)

$$Y = X\beta + \varepsilon$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Computers prefer matrices

1. Programming languages use highly efficient linear algebra libraries
2. Many of them perform matrix operations in parallel

Also for representation purposes.

Fitting linear models

Our goal with linear models is making inferences. Meaning that we obtain our sample, we make linear models, we obtain the coefficients and then we can infer other results

A linear model can be fitted (= OLS estimates can be obtained) by solving the normal equations

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Error estimates for the model coefficients can also be obtained

$$\text{se}(\hat{\beta}_j) = \sqrt{s^2 (X^T X)^{-1}_{jj}}$$

Hypothesis testing in linear models

If the following assumptions hold:

- $E(\varepsilon_i) = 0$, i = 1, ..., n — **Linearity**
- $Var(\varepsilon_i) = \sigma^2$, i = 1, ..., n — **Homoscedasticity**
- $\varepsilon_i \sim Normal(0, \sigma^2)$ — **Normality**
- ε_i and ε_j are *independent* — **Independence**

where σ^2 is an additional parameter, unique but unknown to us (can be estimated by the sample variance, s^2)

Assumptions:

- Expected value of the Errors is 0
- Variance is unknown but constant
- Normality
- Every observation is independent

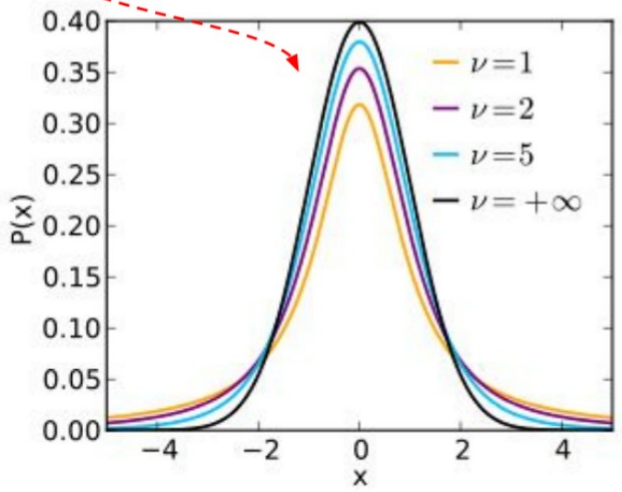
Then, we can build a test statistic that helps us to decide if B1 coefficient is equal to 0 or not (assess significance).

$$\frac{\text{signal}}{\text{noise}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim \text{t-Student}_{(\nu=n-2)}$$

e.g. for β_1

$$\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \sim \text{t-Student}_{(\nu=n-2)}$$

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$



p-value

p-value < α, accept H_0

p-value < α, reject H_0

(Usually α = 0.05)

Design and contrast matrices

The coefficient corresponds to the mean of the first condition and a second coefficient that corresponds to the mean of the second coefficient.

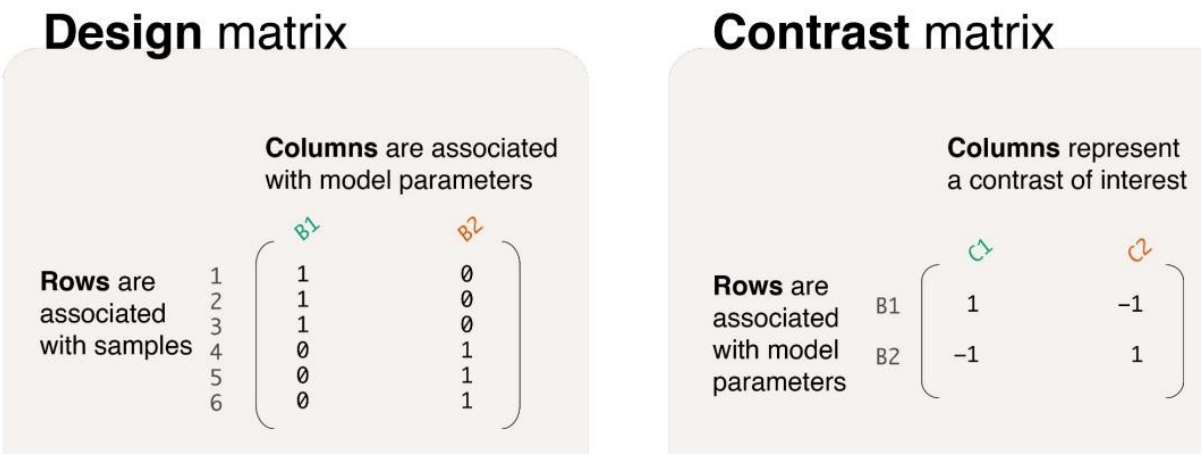
In this case, we are not interested in saying that the mean of the WT is equal to 0 or not. Our goal is to test if the difference between the 2 conditions is equal to 0 or not.

Design matrices are an encoding of the experimental design and are used in the estimation process of model parameters. The design matrix has columns associated with the parameters and rows associated with samples. If the estimated parameters are not of direct interest, a contrast matrix can be used to calculate contrasts of the parameters.

Here we have a design matrix that correspond to the mouse experiment (controls vs treatments). For each sample, we define if it is a control or a treatment.

- We get as many columns as model parameters
- We get as many rows as samples

But as we said, we are not interested in testing if this coefficient is different from 0 (because we are working with categorical variables).



What is relevant is to know the difference between the means (contrast matrix).

- The mean of the mutant is larger than the mean of the WT, this is important.

In a contrast matrix, the rows are the parameters (dummy variables) and the columns are the contrasts (each of the combinations that we want to test).

- The first contrast is B1 - B2
- Second contrast B2 - B1
- We could have as many contrasts as we want

Linear Model+GLM

5. In a principal component analysis (PCA) of gene expression data from RNA-seq of cancer patients and healthy controls:

a. All the answers are correct (b. Points represent samples and principal components are linear combinations of gene expression values, d. The distance betweeen two points represents the overall difference in gene expression values between two samples, e. The informa-on about sample groups (cancer, controls) is not used to calculate the principal components.)

6. Overdispersion of read counts is due to the fact that:

c. The variability among biological replicates is larger than in a Poisson model

Examples

Now we consider an example of gene expression on healthy and sick mice, each in triplicate. Healthy and sick mice are classified using a `group` factor which contains two levels, `HEALTHY` and `SICK`.

```
##   expression  mouse   group
## 1       2.38 MOUSE1 HEALTHY
## 2       2.85 MOUSE2 HEALTHY
## 3       3.60 MOUSE3 HEALTHY
## 4       4.06 MOUSE4    SICK
```

So, we have a single categorical variable that we can encode in many different ways.

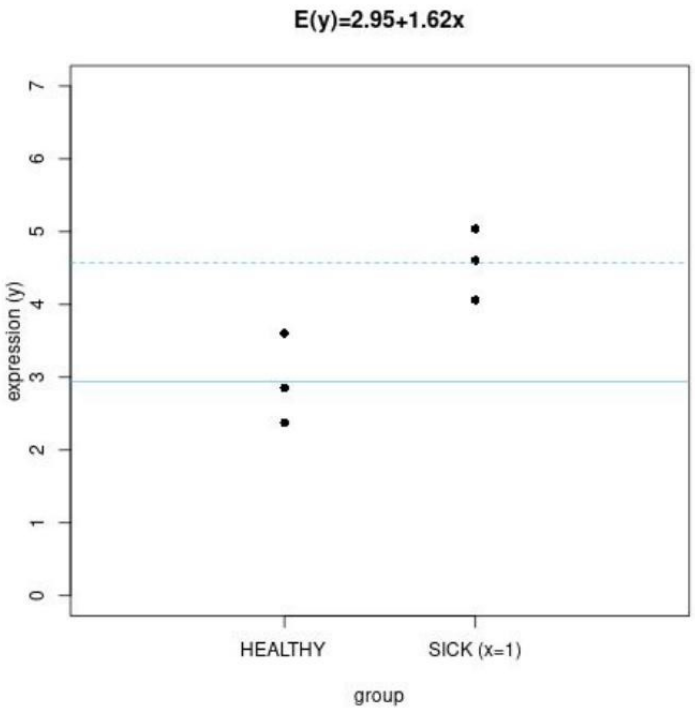
Mean-reference model (with intercept)

```
model.matrix(~group)
```

```
##   (Intercept) groupSICK
## 1           1         0
## 2           1         0
## 3           1         0
## 4           1         1
## 5           1         1
## 6           1         1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

```
fit <- lm(expression~group)
fit
```

```
##
## Call:
## lm(formula = expression ~ group)
##
## Coefficients:
## (Intercept)    groupSICK
##        2.95         1.62
```



2.95 represents the basal mean. So the mean of the sick cohort is 1.62 higher

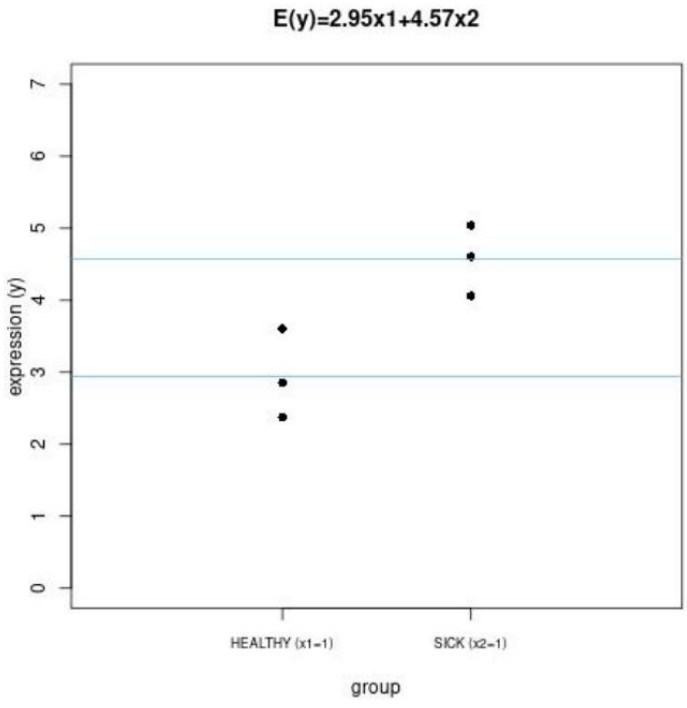
Means model (without intercept)

```
model.matrix(~0+group)
```

```
##   groupHEALTHY groupSICK
## 1            1         0
## 2            1         0
## 3            1         0
## 4            0         1
## 5            0         1
## 6            0         1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

```
fit <- lm(expression~0+group)
fit
```

```
##
## Call:
## lm(formula = expression ~ 0 + group)
##
## Coefficients:
## groupHEALTHY     groupSICK
##         2.95          4.57
```



Equivalent model, different parametrization

Contrast matrix

Here we want to test if the difference between sick and healthy is 0 or not.
In this case we are doing B2-B1

```
design <- model.matrix(~0+group)
makeContrasts(groupSICK-groupHEALTHY, levels=colnames(design))
```

```
##               Contrasts
## Levels         groupSICK - groupHEALTHY
##   groupHEALTHY                       -1
##   groupSICK                           1
```

The *makeContrast* function simply creates the contrast of (-1, 1) which subtracts the first parameter estimate (mean expression of healthy) from the second parameter estimate (mean expression of sick). Using the parameter estimates estimated earlier, the contrast calculates -2.95 plus +4.57 which equals 1.62. In other words, we expect gene expression of sick mice to be upregulated by 1.62 units relative to healthy mice. Notice how this is the same value as the second parameter estimate in the mean-reference model, since that model is directly parameterised for the difference between sick and healthy mice.

Example. Treatment vs control

Imagine we are measuring gene expression in 12 mice that have 4 conditions:

- Control, treatment 1, 2 and 3

Here we have the design matrix with intercept:

```
##    expression   mouse treatment
## 1        1.01  MOUSE1       CTL
## 2        1.04  MOUSE2       CTL
## 3        1.04  MOUSE3       CTL
## 4        1.99  MOUSE4         I
## 5        2.36  MOUSE5         I
## 6        2.00  MOUSE6         I
## 7        2.89  MOUSE7        II
## 8        3.12  MOUSE8        II
## 9        2.98  MOUSE9        II
## 10       5.00 MOUSE10       III
## 11       4.92 MOUSE11       III
## 12       4.78 MOUSE12       III
```

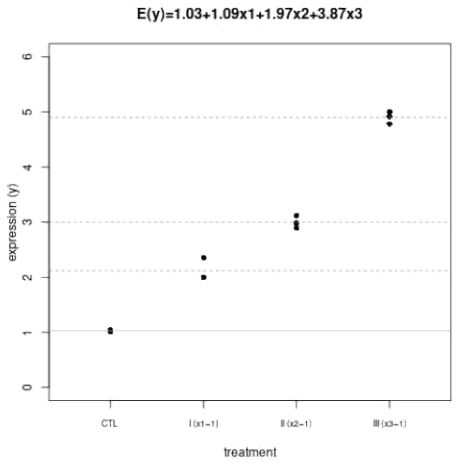
```
treatment <- relevel(treatment, ref="CTL")
```

```
model.matrix(~treatment)
```

```
##    (Intercept) treatmentI treatmentII treatmentIII
## 1            1          0           0            0
## 2            1          0           0            0
## 3            1          0           0            0
## 4            1          1           0            0
## 5            1          1           0            0
## 6            1          1           0            0
## 7            1          0           1            0
## 8            1          0           1            0
## 9            1          0           1            0
## 10           1          0           0            1
## 11           1          0           0            1
## 12           1          0           0            1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$treatment
## [1] "contr.treatment"
```

```
fit <- lm(expression~treatment)
fit
```

```
##
## Call:
## lm(formula = expression ~ treatment)
##
## Coefficients:
##  (Intercept)    treatmentI   treatmentII
##         1.03          1.09          1.97
##
##  treatmentIII
##         3.87
##
```



The fitted model for expected gene expression can then be written as $E(y) = 1.03 + 1.09x_1 + 1.97x_2 + 3.87x_3$, where the x's are indicator variables for treatment I, treatment II and treatment III, respectively. In other words, $x_1 = 1$ for treatment I, $x_2 = 1$ for treatment II, and $x_3 = 1$ for treatment III. The x's are equal to 0 elsewhere.

We can define many possible contrasts:

All pairwise comparisons

```
contrasts <- makeContrasts(
  treatmentI-treatmentCTL, treatmentII-treatmentCTL,
  treatmentIII-treatmentCTL, treatmentII-treatmentI,
  treatmentIII-treatmentI, treatmentIII-treatmentII,
  levels=colnames(design))
colnames(contrasts) <- abbreviate(colnames(contrasts))
contrasts
```

```
##              Contrasts
## Levels         tI-tC tII-tC tIII-tC tII-tI trIII-tI tIII-tII
##   treatmentCTL    -1     -1      -1      0        0        0
##   treatmentI       1      0       0     -1       -1        0
##   treatmentII      0      1       0      1        0       -1
##   treatmentIII     0      0       1      0        1        1
```

Control versus the rest

```
makeContrasts((treatmentI+treatmentII+treatmentIII)/3-treatmentCTL,
  levels=colnames(design))
```

```
##              Contrasts
## Levels         (treatmentI + treatmentII + treatmentIII)/3 - treatmentCTL
##   treatmentCTL                                                      -1.00
##   treatmentI                                                         0.33
##   treatmentII                                                        0.33
##   treatmentIII                                                       0.33
```

Study of interactions and additivity of treatments

Remember the second example.

Consider a study on the effect of estrogen on the expression of the genes in ER+ breast cancer cells over time. 8 samples:

- 4 treated with estrogen (T)
- 4 untreated (C).

Gene expression measured at 2 time points:

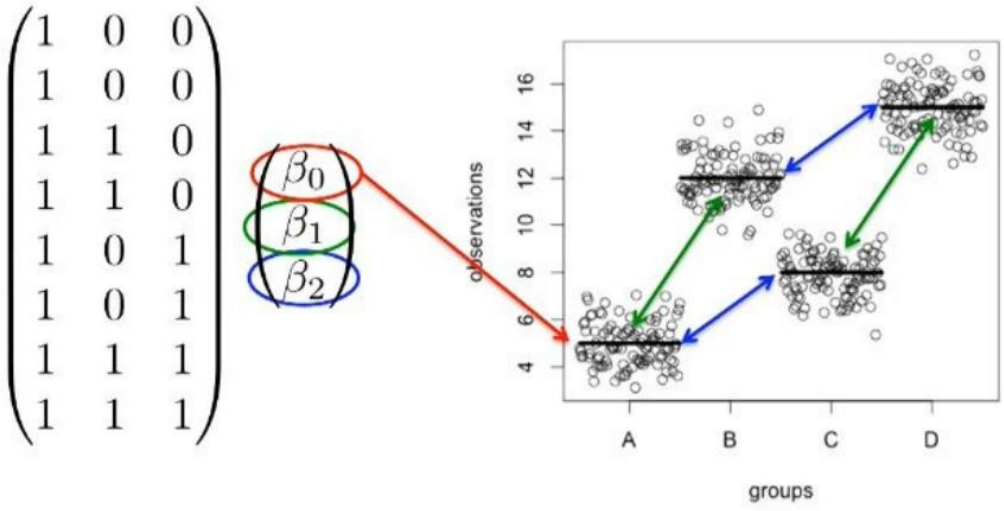
- 10h (2T+2C)
- 48h (2T+2C).

So, one factor is treatment vs control and the other factor is the measuring time. If we make a good design, we will obtain something like this:

- An intercept (which corresponds to the control and 10h)
- Control, patient, control, patient
- 10h, 48h

An important thing is that effects are additive. In the design matrix we have 4 different types of rows:

- A: 1 0 0 → Mean expression = Intercept (control and 10h). Its expression will be B0
- B: 1 1 0 → Mean expression = Intercept (B0) + treatment effect (B1)
- C: 1 0 1 → Mean expression = Intercept (B0) + time effect (B2)
- D: 1 1 1 → Mean expression = Intercept (B0) + treatment effect (B1) + time effect (B2)



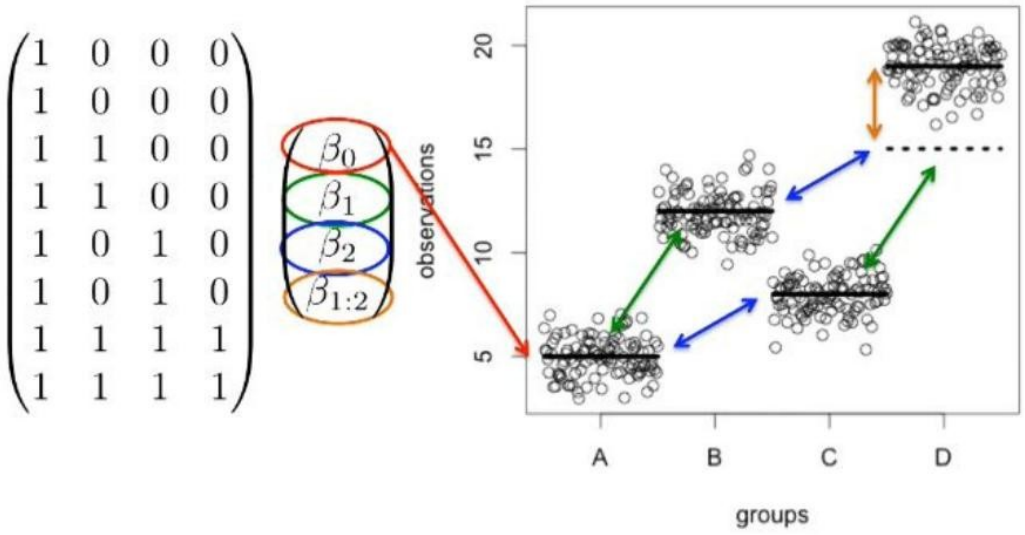
We can also try to see if we have interactions. In this case, we add an extra column in our design matrix.

The mean that we observe in our individuals is above or below our expectation just considering the additive effect? If yes, then we have an interaction

Linear Model+GLM
7. Which of the following is false regarding Generalized Linear Models (GLMs):
d. voom+limma implements a GLM for DGE analysis

8. Which of the following is false regarding biological significance assessment?
e. Both GO enrichment and GSEA require a ranked gene list

Crossed designs with interaction



Interaction shows additional to additive effects

Recap:
- We are going to move from continuous to categorical explanatory variables, which can be encoded using dummy variables (0 and 1). These explanatory variables can have an intercept or not.
- We are not going to test the coefficients of the dummy variables, we are interested in the whole categorical variable and therefore we need to create contrasts.

Mean-variance relationship of read counts

Modeling count data -intuition

We are going to work with RNA-seq data, so we have reads and counts (discrete data).

Linear models that model continuous variables, but in this case we have discrete variables. We also said that linear models assume normality, but this distribution is not normal. We also said that the variability should be constant, but in this case the mean is related to the variance.

So, we are breaking all the assumptions. Thus, we need to find a way to model this data using linear models but adapting them for reads (Generalized Linear Models).

In NGS experiments, the raw data are millions of reads which are typically aligned to genes in the genome.

Let's color every gene in a different color.

Each read that maps to a particular gene will have the same color.

Let's think of this NGS experiment as having a bag full of reads with their corresponding color. We start taking the reads out of the bag and placing them in each gene.

This can be seen as a success or failure. If we are in the green gene and we take a green read, then it's a success. We repeat this process n independent times.

- This is a binomial distribution (bernoulli)

Modeling count data - Binomial distribution

Kg is the number of sequenced reads which can be assigned to a particular gene.
This variable will follow a binomial probability distribution.

Basically, our random Kg variable can take a particular value k (0, 1, 2 …) with a given probability. This value is given by the proportion of reads of our pool that actually come from this gene.

So, by dividing the number of reads that correspond to this gene by the total pool of reads, we will obtain an estimate of the probability of getting out of the bag read from this gene.

Linear Model+GLM
9. After a clustering analysis of gene expression measured via RNA-seq on 20 individuals (10 treated with drug A, 10 treated with drug B), you observe two major clusters. The first cluster contains mostly male, drug A-treated patients, the second contains mostly female, drug B-treated patients. Which of the following is correct?
d. Sex should be included as a potential confounder in the models for DGE analysis between treatments

10. Select the right answer regarding multiple testing correction
a. Bonferroni correction is tipically more stringent than FDR

- If there are *k* mapped reads to region *g* (*successes*), then there must be *n* − *k* not aligned reads to region *g* (*failures*)
- Probability of first *k* successes, then *n* − *k* failures:

$$\underbrace{p \times p \times \cdots \times p}_{k \text{ of these}} \times \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{n-k \text{ of these}} = p^k (1-p)^{n-k}$$

- All possible ways to arrange *k* successes and *n* − *k* failures:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Hence, the total probability of *k* successes in *n* independent trials is

$$\mathrm{P}(K_g = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, 2, \ldots, n$$

Binomial distribution with parameters n and p. We write $K_g \sim \mathrm{Bin}(n, p)$

- As the number of sequenced reads becomes large and the probability *p* shrinks, the Binomial distribution can be approximated by the Poisson distribution

The binomial distribution has 2 parameters (n and p) while the poisson distribution has only one parameter (lambda).

So, Kg follows a poisson distribution.

Consider the following scenario:
Several flow cell lanes are filled with aliquots of the same prepared library (technical replicates)

- Since they are technical replicates, the concentration of a given transcript is exactly the same in each lane

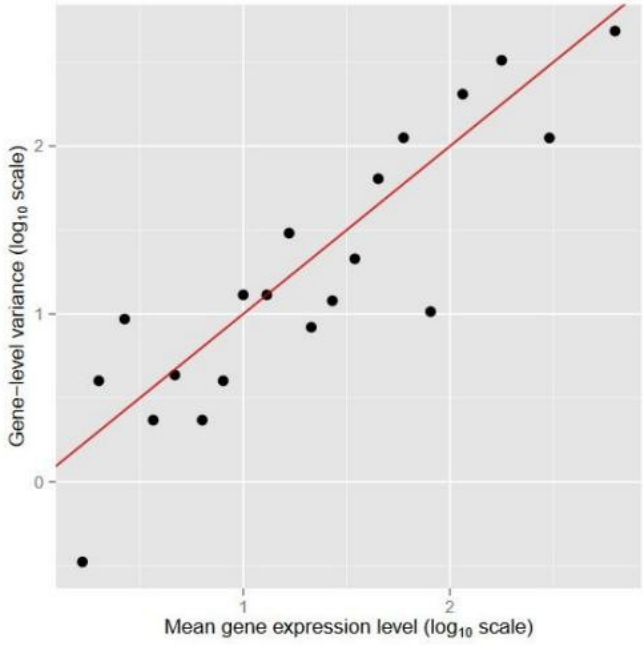
For each lane, count the number of reads obtained from the transcript. **Will the count be the same?** Of course not! Even with equal concentration, the counts will vary because of the technical process.

This theoretically unavoidable noise is called shot noise, i.e. the variance in counts that persists even if everything is identical (technical replicates)

- We assume that the shot noise follows a Poisson distribution

In the following plot, each point represents a gene. We can see a relationship between mean and variance for technical replicates. This is reflective of the poisson distribution.

- Lambda, which is the mean, is directly related to the variance.



This is the case when we get a RNA-seq analysis from an individual and we split it in 10 (technical replicates).

Overdispersion

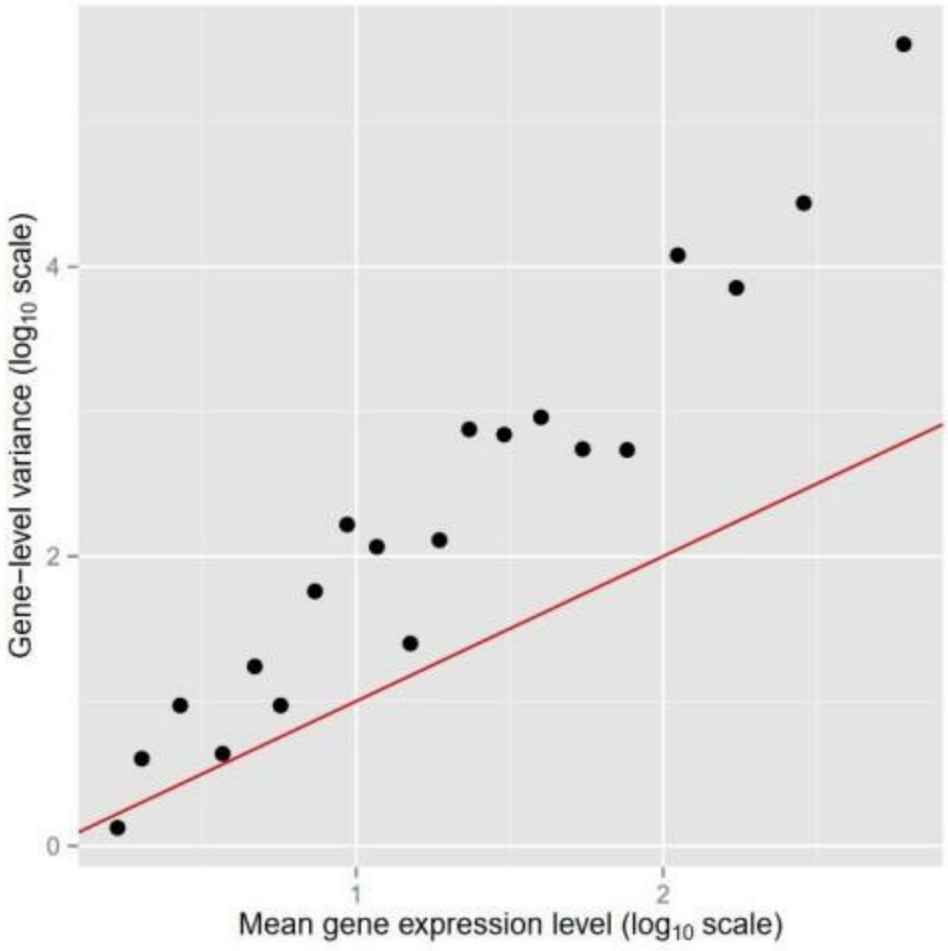
What happens if we get a sample of 20 individuals and then I sequence them? I am using the same tissue, all healthy… everything should be very similar.

I plot the mean expression of a gene in all samples vs the variance.

In this case, the variability will be larger because we are using different technical and biological procedures (each person is different).

We will see that the variability that we observe across **biological replicates** is much larger than the one that we expect in a poisson model.

While the Poisson approximation of the binomial distribution still holds for an individual sample – as n is still large and p small – the variance for the counts for gene g between samples will be often much larger than the mean



This is called overdispersion

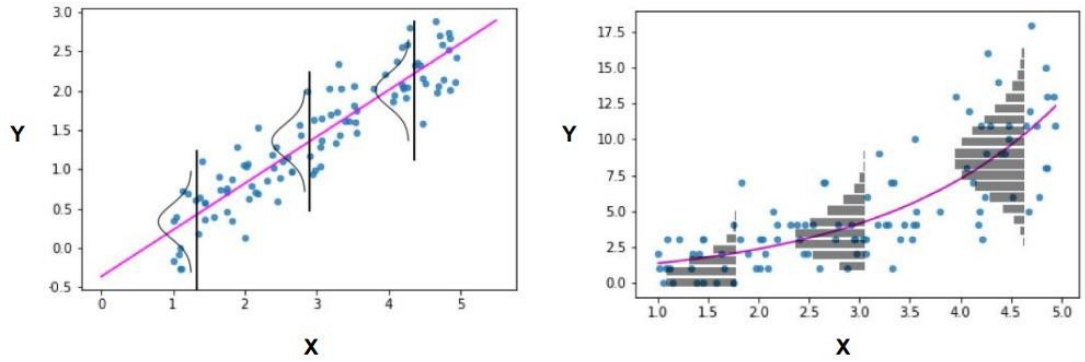
When the variance in a Poisson model is greater than the mean, the counts are said to be “overdispersed” with respect to a Poisson distribution.

To model overdispersed counts, we can use the **negative binomial distribution**.
The mean is the same as in the poisson distribution but the variance is going to be the one from the poisson distribution + an extra modifier (phi) that depends on the lambda parameter.

Generalized Linear Models (GLMs)

As we said, we can not use linear models to model our RNA-seq data because it does not follow the linear model assumptions. So, we will use GLMs.

This is a standard linear model, meaning that we are relating an independent variable with a response variable and we are assuming that the errors at every particular value of x are normally distributed. The problem is that in RNA-seq data, it does not look like this.



The variability increases with the mean.

- As we get larger gene expression, we get larger variance.

So, we need to find a linear model that can fit the negative binomial. In which the variability is smaller for smaller mean values and it increases as the mean increases. This type of situations can be modeled using GLMs.

So, with GLMs we will fit the information of the negative binomials and then do contrasts and design matrices as we already did in Linear Models. **In this case we are adapting our methods to our data.**

GLM consists of 3 elements:

1. Exponential family of probability distributions (can not have to be normal)
2. Linear predictor, which is the linear combination of explanatory variables and the coefficients (as we have seen in linear models)

$$Y_i = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \ldots + \beta_{(p-1)} X_{i(p-1)}) + \varepsilon_i \quad \text{for } i = 1, \ldots$$

3. Link function, which will be applied to our linear combination and will transform it in a way that it can account for these differences in variance.

It's a particular case of GLMs in which the link function is the identity function and we have normal data.

Family	Notation	Canonical link
Gaussian	$N(\mu, \sigma^2)$	identity: μ
Poisson	$\text{Pois}(\mu)$	$\log_e(\mu)$
Negative-Binomial	$\text{NBin}(\mu, \theta)$	$\log_e(\mu)$
Binomial	$\text{Bin}(n, \mu)/n$	$\text{logit}(\mu)$
Gamma	$G(\mu, \nu)$	μ^{-1}
Inverse-Gaussian	$IG(\mu, \nu)$	μ^2

For each exponential distribution, we have a different link function.

In normal distribution, the link is the identity function "mu" (so, no transformation is needed) and we are having standard linear regression.

For the poisson distribution, we make a ln(mu) transformation

Example: Logistic regression

We have binary data and we are trying to predict (in the other cases we used to identify associations, not to predict).

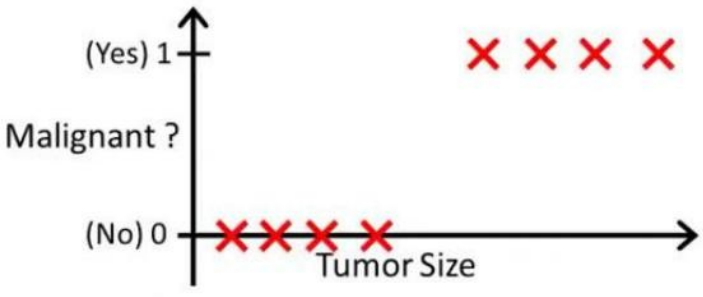
In this case we are trying to predict benign or malignant tumors according to their size:

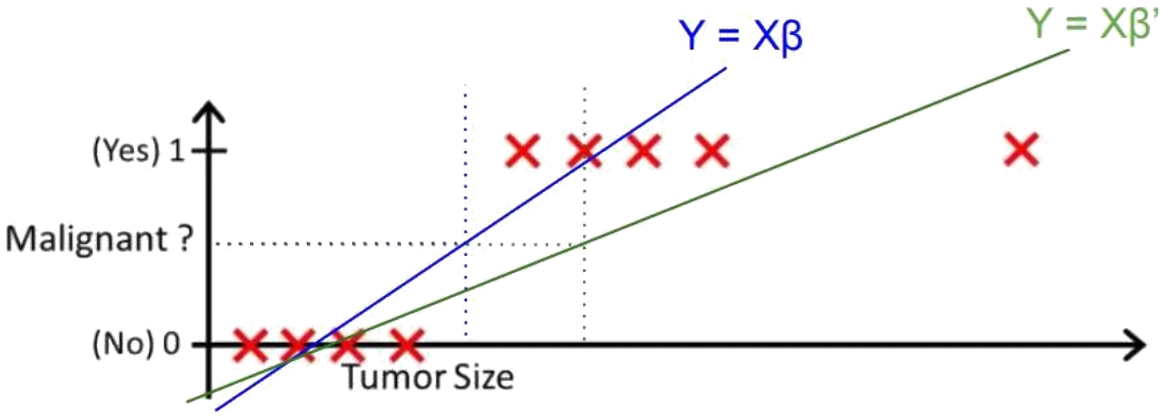
- Explanatory variable: Tumor size
- Response variable: Bening (0) or malign (1)

I could try linear regression. If the data does not have many outliers, it may work (blue).

Binary classification of tumors in benign or malignant given the tumor size

$Y \in \{0, 1\}$ 0: 'negative class' (e.g. benign tumor)
1: 'positive class' (e.g. malignant tumor)





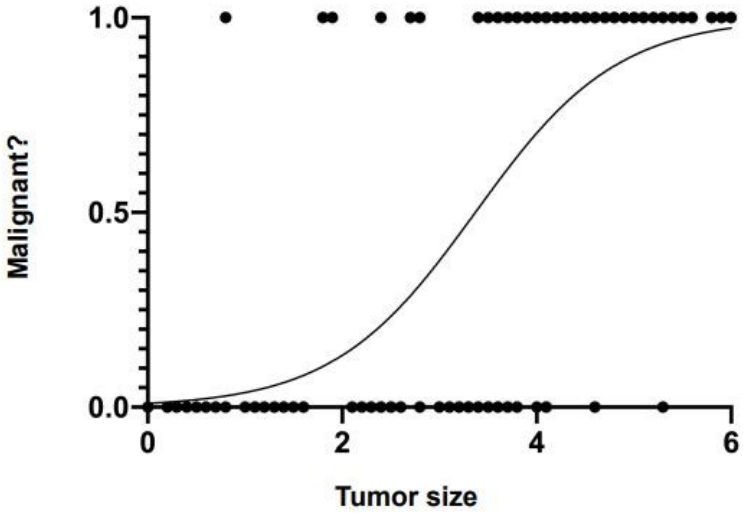
But we know that we should do logistic regression with binary data. Because if we have an outlier, our fit will be displaced towards the outlier and the predictions will be wrong.

If we use logistic regression, we will apply a transformation.

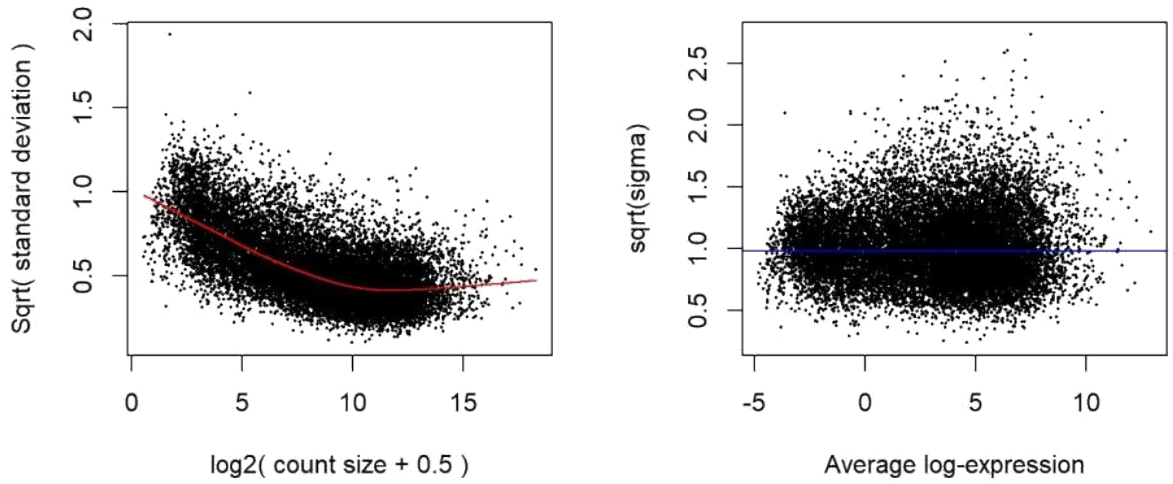
$$Y = g^{-1}(X\beta) + \varepsilon$$

where $g^{-1}(z) = \frac{1}{1+e^{-z}}$ is the *logistic* or *sigmoid* function.

logit = g



- Rather than fitting GLMs, *voom* learns the mean-variance trend and adjusts it in the context of standard linear models (e.g. as implemented in *limma*)



We have another alternative to do the analysis.
Here we are adapting our data to our methods.
Voom performs a variance stabilizing. So, it will remove the mean variance relationship from the data. From this transformed data we can apply standard linear models (LIMA).

At the end, we will obtain a P-value and fold-change (that corresponds to each contrast) for each gene.
- Fold change corresponds to the level of expression

Biological significance analysis

We have a list of genes and we need to go to a list of functions. In cancers vs control we typically find that genes upregulated in cancer samples seem to be related to with the cell cycle or DNA replication (which makes sense).

So, if we are able to understand which are the functions that are enriched among our list of genes, we will be able to understand better the biology behind the comparison of the 2 conditions.

We have an extra problem: We said that our data has a variance mean relationship, its negative binomial… So we need to use GLMs or use voom to transform our data. But then we get a huge list of 10.000 genes, so we are doing 10.000 tests.

What happens when we are doing so many hypothesis tests? There is a high likelihood of finding false positives. We have a multiple testing correlation problem.

So, differential gene expression is a problem of linear models and multiple testing correction.

Multiple testing problem

In every test we have the following matrix:

- Rows is the reality
- Columns is the decision we are taking

	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

The **null hypothesis** states that there is no significant difference or relationship between variables, while the **alternative hypothesis** asserts that there is a significant difference or relationship.

If the calculated p-value (the probability of observing the data under the null hypothesis) is lower than the significance level, the null hypothesis is rejected.

Alpha = Probability of Type I error
Beta = Probability of Type II error

Power = 1 - Beta

Type I error: False positive. Rejecting the null hypothesis when it is actually true (there are no differences between the groups). Which means getting a significant P-value (below 0.05) and we say that a gene is significantly differentially expressed between our 2 conditions.

In all omics studies we will have this problem of high throughput analysis.
Also, when doing a functional analysis we are going to find enrichment in different functions (we have many GO terms).

As the number of tests increases the chance of observing at least one false positive increases too.

Probability of being right: The probability not rejecting H0 when it is true is 1- alpha.
Where alpha is the Type I error.

$$P(\text{not reject } H_0 \mid H_0 \text{ is true}) = 1 - \alpha$$

We are saying that there are no differences between the groups when actually there are no differences (we are making the right decision).
If alpha = 0.05, we are going to be right 95% of the time.

But if we do this for "m" tests and assume that they are independent → (1-alpha)^m

The probability of being wrong is the complementary → 1 - (1-alpha)^m

This value is much larger than the value of alpha that we set at the beginning.

So, when we are making more tests we are increasing the probability of making Type I errors. So, we will have to be more strict when setting this alpha level (probability of Type 1 error).

Multiple testing adjustments

There are different possible solutions for this problem, which try to control the overall type I error rate.

- **Familywise error rate (FWER)** controlling procedures control the probability of at least one Type I error.
- Example: Bonferroni correction → Basically we are being more strict with the value of alpha.
- We are dividing alpha by the number of tests that we are doing.
- Another way of saying this is multiplying the p-value by m

We are correcting the value of alpha, but we are sacrificing statistical power.
This may lead to a large Type II error (low statistical power), meaning that we will have many false negatives.

- **False discovery rate (FDR)** controlling procedures provide less stringent control of Type I errors (they control that the expected proportion of Type I errors is below a given threshold) than FWER procedures. Thus, FDR provides greater power (smaller Type II error), at the cost of an increased number of Type I errors.
- Example: Benjamini-Hochberg FDR

Corrections depend on m. More tests → more stringent correction

If we apply Bonferroni everything is going to be correct but we will miss many things.
If we use FDR, we will have some errors, but we won't miss so many things.

FDR example

Sort P-values of all tests in decreasing order

Rank	Category	(Nominal) P-value
1	*Transcriptional*	0.001
2	*regulation*	0.002
3	*Transcription factor*	0.003
4	*Initiation of*	0.0031
5	*transcription*	0.005
...	*Nuclear localization*	...
	Chromatin modification	
52	...	0.97
53	*Cytoplasmic localization*	0.99
	Translation	

Adjusted P-value = P-value · Number of tests divided by the rank of the P-value in the sorted list.

Rank	Category	(Nominal) P-value	Adjusted P-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053
...	...	...	...
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004
53	*Translation*	0.99	0.99 x 53/53 = 0.99

The FDR corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.040
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040	0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053	0.053
...	...	...	...	...
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004	0.99
53	*Translation*	0.99	0.99 x 53/53 = 0.99	0.99

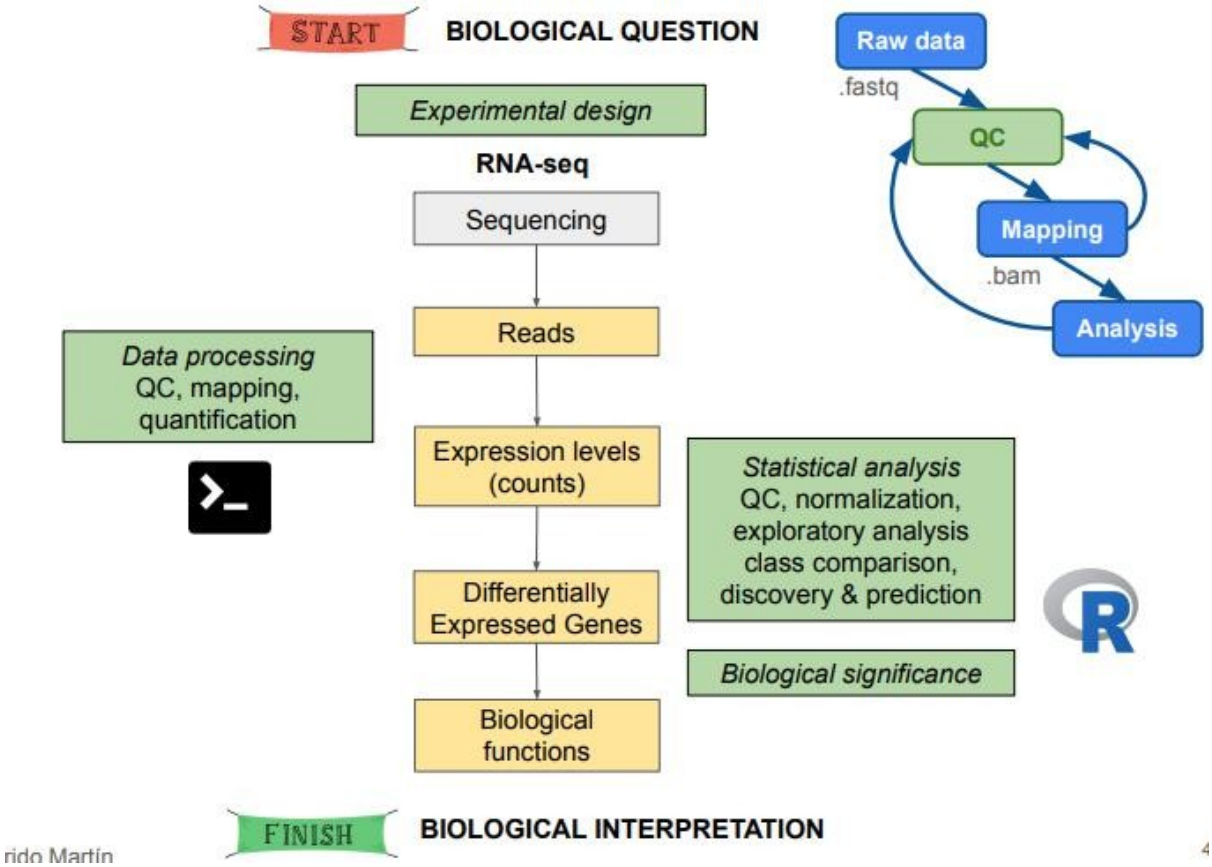
The P-value threshold is the highest ranking P-value for which the corresponding FDR is below the desired significance threshold (0.05).

P-value threshold for FDR < 0.05

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.040
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040	0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053	0.053
...	...	...	...	...
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004	0.99
53	*Translation*	0.99	0.99 x 53/53 = 0.99	0.99

Red: non-significant
Green: significant at FDR < 0.05

Biological significance analysis. Usual pipeline



We have said that the outcome of the differential gene expression analysis is a gene list with P-values and fold changes. We want to search the functions of these genes.

In the end, everything is about testing. We will obtain P-values saying if some functions are statistically enriched in my gene set.

We are going to describe:

- GO enrichment analysis or KEGG enrichment analysis
- Gene set enrichment analysis (GSEA)
- Some visualizations that are useful

So, we will obtain a gene list, we will make some statistical testing and we obtain overrepresented functions.

This involves 2 statistical tests:
- Fisher's exact test
- KS goodness of fit test

Once we have our list of genes, we could:
- Select some candidates for validations (the ones that have a lower P-value)
- Report all the list of genes, which is useless
- Make gene sets → group of genes that have some shared characteristics (functional, structural, pathway, location…). The question we are going to ask is if my list of DE genes is enriched or depleted in a given gene set.

To do this we need:
- List of genes relevant to the condition studied
- Shared functional vocabulary (GO)
- Systematic link between genes and functions (DB of annotations like GO, KEGG…)
- An appropriate statistical analysis.

GO

- Defines concepts/classes used to describe gene function, and relationships between these concepts.
- Controlled vocabulary
- 3 main categories:
- Biological Process (BP)
- Molecular Function (MF)
- Cellular Component (CC)

We will not only use the functions that are assigned from the GO terms but also the links between the different functions.

KEGG

Pathway database. We can determine if different genes are in the same pathway, so we can group them.

In the case of novel organisms, this study is very difficult because we do not have a good annotation of the functions.

An appropriate statistical analysis

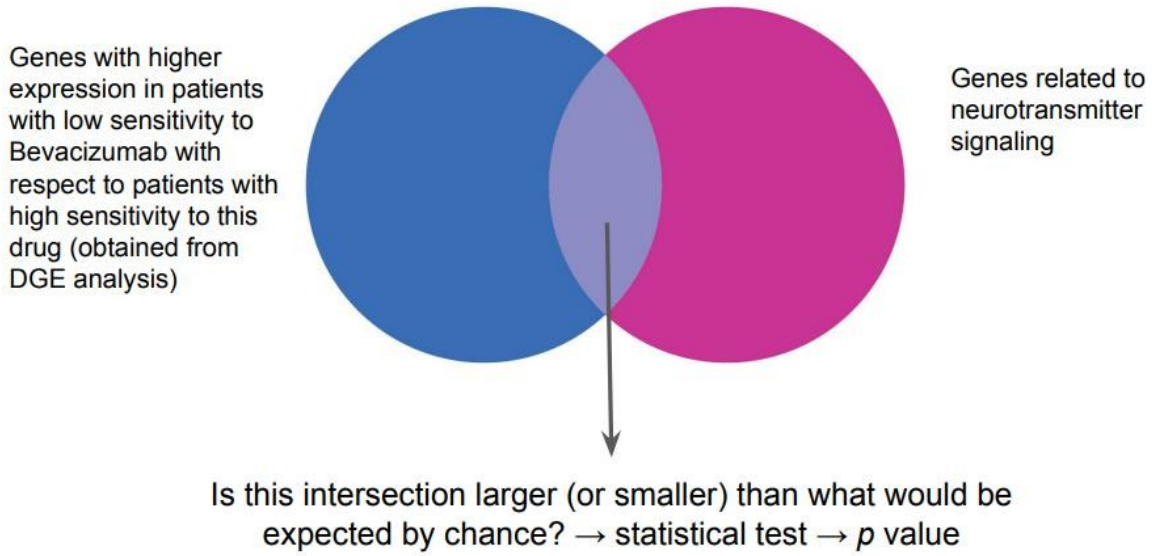
We were saying that we are looking for overrepresentation of some particular gene sets in our list.

Example: We have a DGE study in which we have brain cancer patients with different drug sensitivities:

- There are patients that have low sensitivity to this drug
- Other have a large sensitivity

So, we do a DGE on RNA-seq from the patients and we get a list of DE genes.

- A fraction of them will be upregulated for patients that have low sensitivity (blue circle). So, the blue ball is a subset of genes that are upregulated according to our linear model from the RNA-seq data.
- We have another list of genes related to neurotransmitter signaling.



We want to know if genes that are involved in drug sensitivity in brain cancer are related to neurotransmitter signaling?

Basically we are asking if the intersection is larger or smaller than what we would be expecting by chance → statistical test

H0: The intersection is not significant
H1: The intersection is significant because there is some relationship between this lists of genes (intersection is larger or smaller than what we would expect by chance)

There are 2 ways of doing this:

- If we have just a gene list → before we have set a threshold on the P-value and fold change. So, over all the differentially expressed genes we will only get the ones that have a FDR lower than 0.05 (we are accounting for multiple testing and we get the ones that are significant) and we force them to have a logFold-change > 2 (4 times more or less expressed than the other condition).

  We just have a gene list with their names. In this case the question of interest is if there is any gene set enriched or depleted in the gene list. So, we are looking for overrepresentation in the list.
  To know this we will use Fisher's Exact test.

  In this condition we typically have a second list of genes (called the universe) that we are going to use to compare.

- All the genes that we have tested in our experiment, since they are related to the samples we are analyzing

- Another option is when we have a ranked list instead (by differential expression Fold Change) of just the names. We normally use the P-value to decide if genes are significant or not (and then we forget about the P-value), but we may have another score that is related to each gene (Fold change in DGE, but we can use other scores). When we have this rank list we can use this information that the score has.

  In this case, the question is if there is any gene set ranked high or low in my ranked list of genes?

  The statistical test we will use is the KS goodness of fit test (GSEA).

For these 2 cases we will have 2 solutions.

First case

We can take GO as the source of gene sets to test. In the first case, we will perform the analysis that we will explain for every GO term we are considering in the whole DB.

So, if we are doing GO enrichment, we have 1 million terms related to biological processes and for each of them we will perform the following analysis. For instance, we restrict our attention to a single gene set (genes annotated with the GO biological process term DNA-templated transcription, elongation).

Then we want to see if our differentially expressed genes are within or not this gene set. So, we will build a contingency table.

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	12	3	15
NOT IN Transcription Elongation	3	12	15
Total	15	15	30

→ Fisher's Exact test

↓

p value

With this table, we can perform a Fisher's test and obtain a P-value, saying how related the 2 variables are:

- Variable of being in this gene set
- Variable of not being in this gene set

This will be repeated 1000000 if we have 1000000 terms. We will do the same with the KEGG database.

Since we are testing many terms, we will have to correct for multiple testing (FDR).

So, with the initial gene list with the names we will obtain a list of functions.

Second case

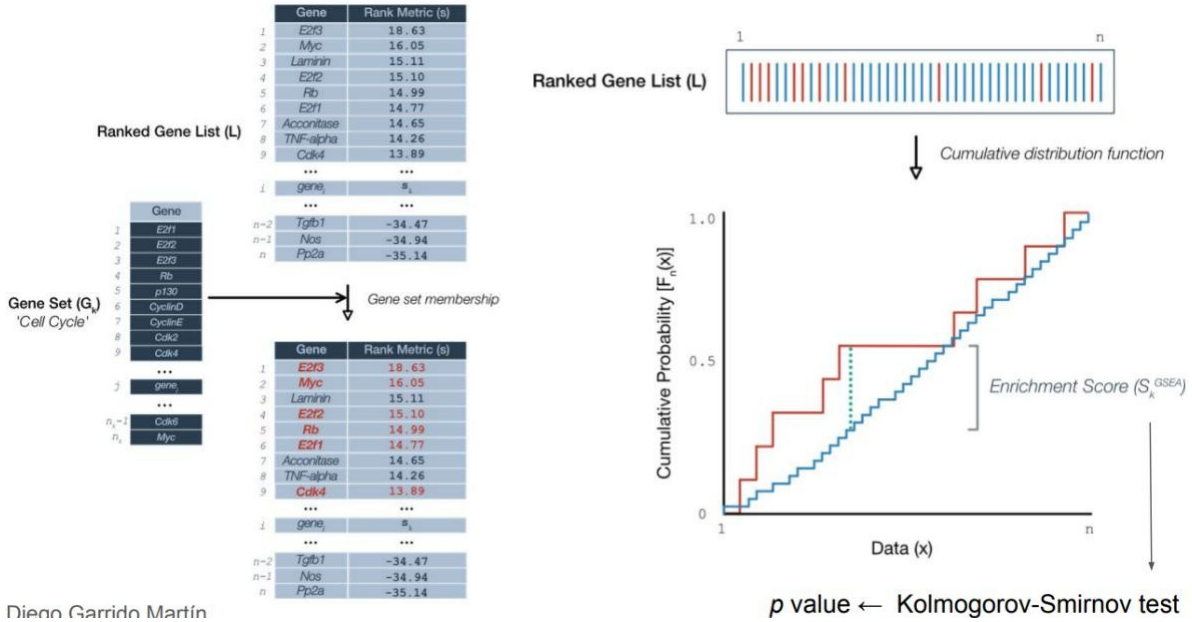
We have our gene list with a ranking metric like the logFold-Change for gene expression.

We have the list of differentially expressed genes with their rank and then we want to see if there is a gene set that is very highly ranked in the gene list (we color them in red).

With this information, we can compute a cumulative distribution function, just by summing. So, the genes that are not in my gene set will be found in the diagonal. But maybe when we look at a particular gene set that is found in many genes with high values, we observe a big distance between the diagonal and these cumulative functions (or the other way around).

We can use the maximum distance between the 2 cumulative distributions as a test statistic. So, computing the P-value we will see if it is significantly different or not (the gene set is spread over all the gene list).

.



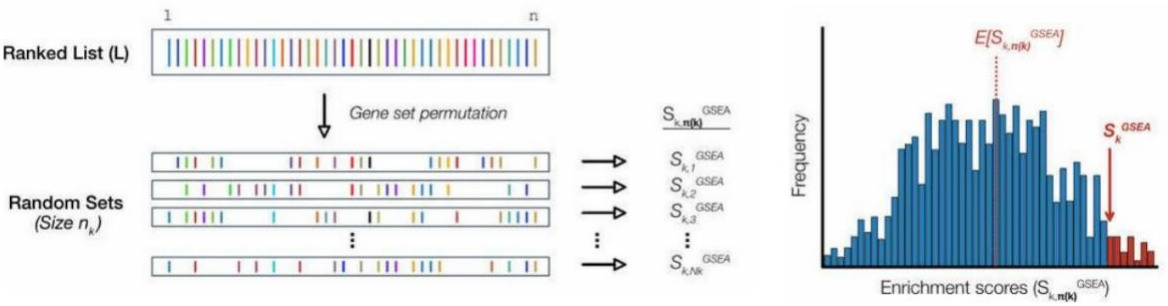
But there is still one problem: It may happen that the KS statistic is large, but the differences occur in the middle (not meaningful). We don't want our test to identify this situation. We are only interested in differences that occur at the top or bottom.

To address this, we will use a modified version of the test statistic to emphasize differences at extreme ends. For this case we do not have a known distribution, so we have to permute.

When testing, there are parametric and non-parametric tests. Among the non-parametric tests, the one that is valid for any kind of data is the permutation test.

- You don't have to assume anything

I permute the data and calculate a new statistic many times. For these statistics I build my own distribution including the one I computed with the real set. I will count how many times my permuted data reaches a value higher than the reference.



Describe the structure of the summarizedExperiment Bioconductor class. Which are its main differences with respect to the expressionSet class?

A researcher wants to perform an RNA-seq experiment, followed by a differential gene expression (DGE) analysis, to compare gene expression between lung cancer patients and healthy controls. Draw a possible workflow with the steps to carry out such study, starting from the biological question, until the interpretation of the results.

First, the researcher should ask them self the biological question: “Are there any genes which are differentially expressed between lung cancer patients with respect to healthy ones?”. After that, the researcher should carry out and RNA-seq experiment, where it would be necessary to recollect samples from both types of patients, sequence them (having previously converted them to cDNA since where are collected RNA), applying some quality control, mapping them against the genome, counting reads and applying any correction regarding normalization, some duplicated information, outliers, etc. Then, we should get the matrix of TPM and try to find which genes are significantly expressed. We should calculate the logFC and finally, plot them in a Volcano plot. Genes with a very low p-value and with a logFC whose absolute value is large enough will be the genes that will be differentially expressed among patients.

Among the following metrics: read counts, CPM, RPKM and TPM, select the most appropriate one to compare the expression of a given gene between two technical replicates. Justify your answer.

CPM. CPM normalizes library size but not length. RPKM and TKM work with the length and knowing we are working with the same sequence is not useful.

Explain why raw read count data should not be directly modeled using standard (i.e. Normal) linear models. Enumerate two alternative strategies for this purpose.

Raw read count data is not normalized. That means that some genes were sequenced more than others and show more counts and others are just longer and will obviously have more reads. Therefore, we should apply some corrections. We can use scaling factors such as FPKM (fragments per kilobase per million) or CPM (counts per million).

Reason why lowly expressed genes across all samples should be removed prior to differential gene expression analysis.

Lowly expressed genes suppose some memory space for the computer to have, so for our computer in order to better work efficiently we should remove all genes that have lowly expressed and save memory space. Moreover, a differential gene expression analysis is centered on studying genes that are differentially expressed among samples. If some gene

is lowly expressed between all individuals in the experiment, it is not making a difference and that means that the gene is just not involved in the process we are studying.

Describe the problem of overdispersion of read count data.

Read count data presents the problem that there is some mean-variance relationship and sometimes there is overdispersion (the variability among biological replicates is larger or equal than the mean). That supposes a problem because we would not be properly estimating in our model. That is why it is useful to use Negative Binomial models to avoid this and even have an overdispersion parameter (though sometimes it could be worse to have to estimate a parameter).

Which are the main differences between overrepresentation analysis (e.g. GO enrichment) and Gene Set Enrichment Analysis (GSEA)?

GO receives a set of genes and results the % of genes in the set involved in the pathway, while GSEA should receive all the genes and by comparing those genes with the ones involved in the pathway, the enrichment number/index will be increasing or decreasing and will be finally resulting in a number which will determine the significance of the genes related to the pathway.

Describe what you could do to identify a batch effect in your expression data.

To identify batch effect in our expression data, we could use PCA and heatmap representations. These approaches offer the possibility to check how is our data aggroupated and how it is related to the different independent variables that we may be considering in our study. If the heatmap presents a certain variable that seems to associate with our explanatory variables, it is important to consider these problematic variables as confounding ones and include them in the study.

A researcher wants to study gene expression in Alzheimer's disease (AD), mild cognitive impairment (MCI) and healthy (H) conditions. He performs RNA-seq on three individuals per condition, followed by a DGE analysis (voom + limma, models without intercept term) to compare gene expression between all the conditions pairwise. Draw the corresponding design and contrast matrices.

Design matrix

	H	AD	MCI
Indv1	1	0	0

Indv2	1	0	0
Indiv3	1	0	0
Indiv4	0	1	0
Indiv5	0	1	0
Indiv6	0	1	0
Indiv7	0	0	1
Indiv8	0	0	1
Indiv9	0	0	1

Contrast matrix

	AD-H	MCI-H	MCI-AD
H	-1	-1	0
AD	1	0	-1
MCI	0	1	1

(AD+MCI)/2 -H

In the previous study, after performing DGE analysis and adjusting for multiple testing via FDR, for the contrast AD – H, the researcher got the following results (only the top 6 genes are shown):

	logFC	AveExpr	t	P.Value	adj.P.Val
ENSG00000179299	-3.031999	3.641797	-6.773810	1.281663e-05	0.0074911
ENSG00000088827	3.396428	4.619954	5.742075	6.672291e-05	0.0097080
ENSG00000134755	4.011486	4.157974	5.197041	1.693045e-04	0.0339791
ENSG00000278195	-2.156150	2.609283	-5.065256	2.133572e-04	0.0907918
ENSG00000111335	2.210268	7.502138	4.840876	3.180007e-04	0.1299791
ENSG00000140443	-3.641796	6.203476	-4.794777	3.454539e-04	0.2397918

How many significant genes are there at 5% FDR? How many significant genes are over- and under-expressed in AD with respect to H? Which plot would you use to summarize the information contained in this table? Justify your answers.

There are 3 genes that are significant (we need to select those whose adj.P.val <0.05%, in this case, the first 3 ones).

To know if the genes are over or under expressed in AD with respect to H, we must check the logFC. If it is positive, they are over-expressed and if not, they are under-expressed. In this case, from the 3 selected significant genes, 2 are over-expressed

(*ENSG00000088827 and ENSG00000134755*) and the other one (*ENSG00000179299*) is under-expressed. (If we were considering all 6 genes, then there would be 3 under-expressed and 3 over-expressed).

The ideal plot to summarize this information would be a Volcano-plot. In the X-axis we would display the logFC and in the Y-axis we would represent the p-values. Values on the top and far away from the center would indicate significant genes that are over/under-expressed.

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

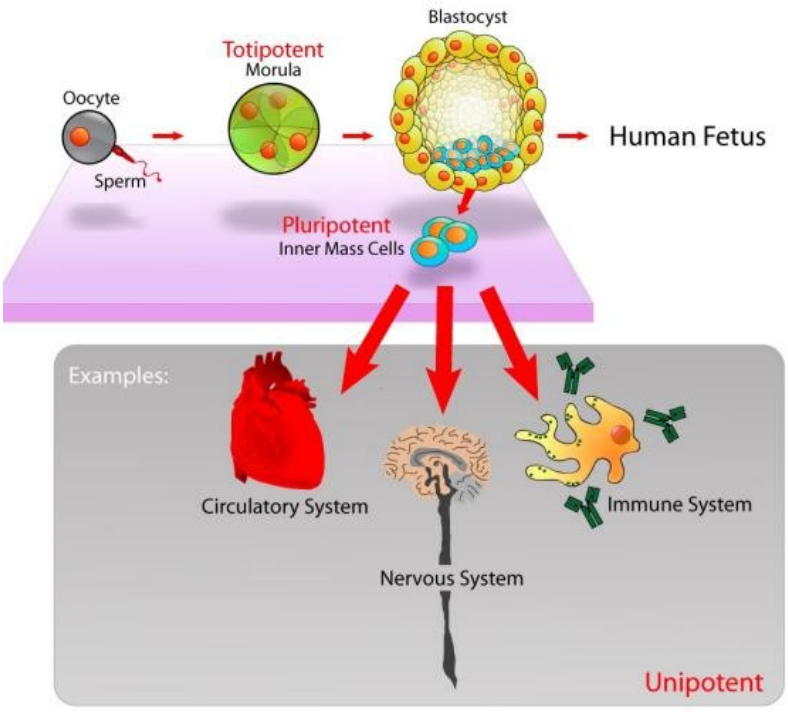
We are the result of cellular divisions that come from one single cell.

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).

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.

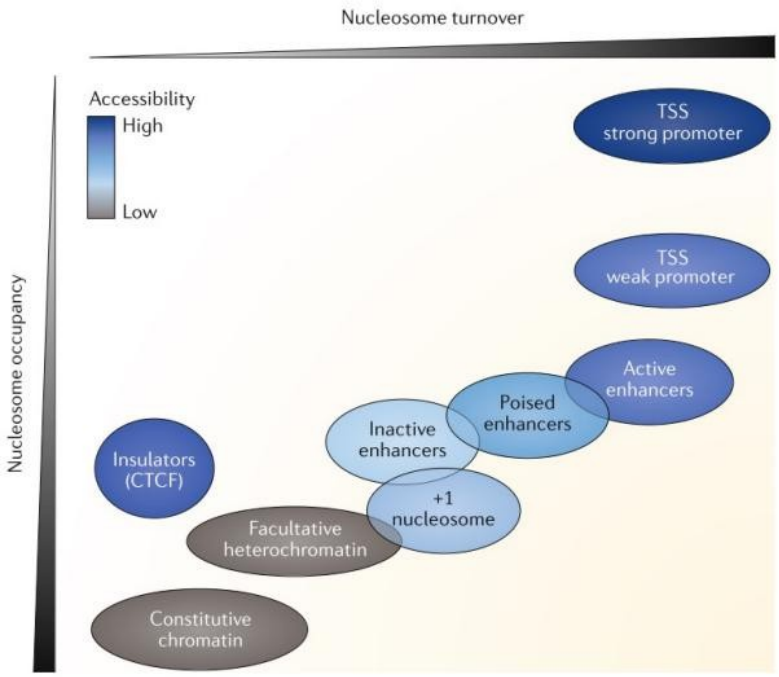
  A nucleosome consists of a core particle made up of two copies each of four histone proteins: H2A, H2B, H3, and H4. These proteins form an octamer, around which approximately 147 base pairs of DNA are wound.

  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

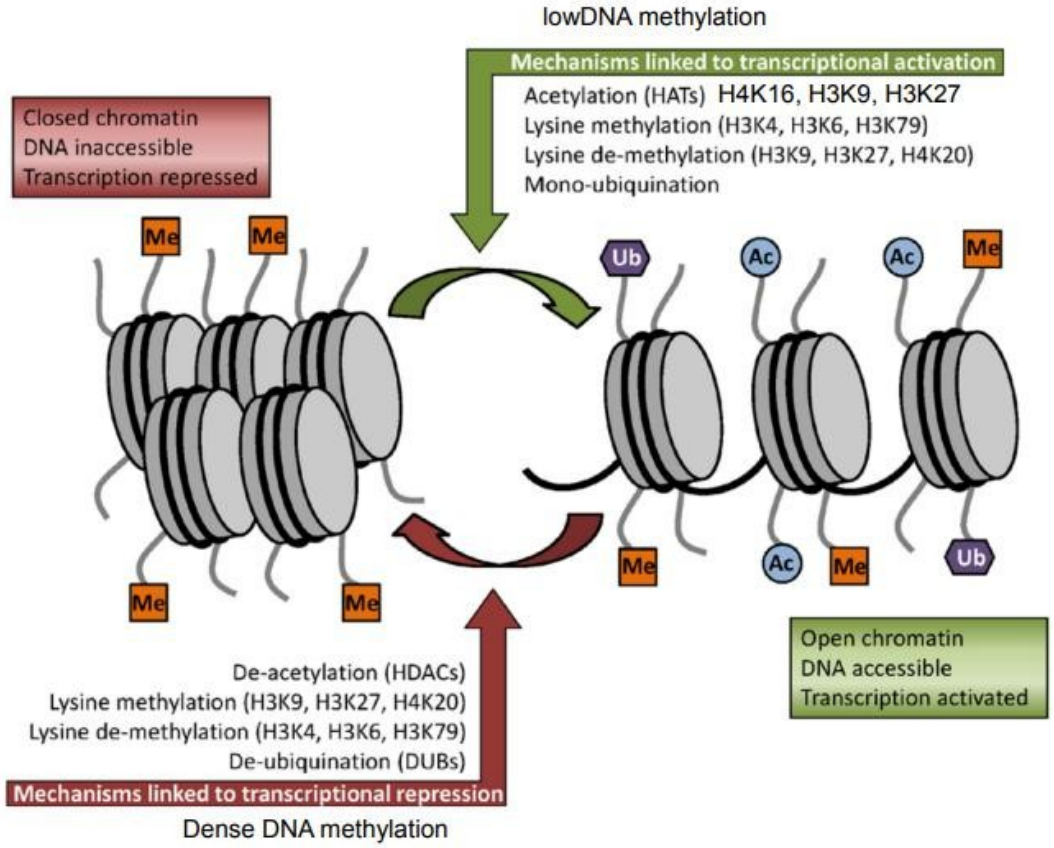
DNMT deficient organisms die at embryonic stages because:

- 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.

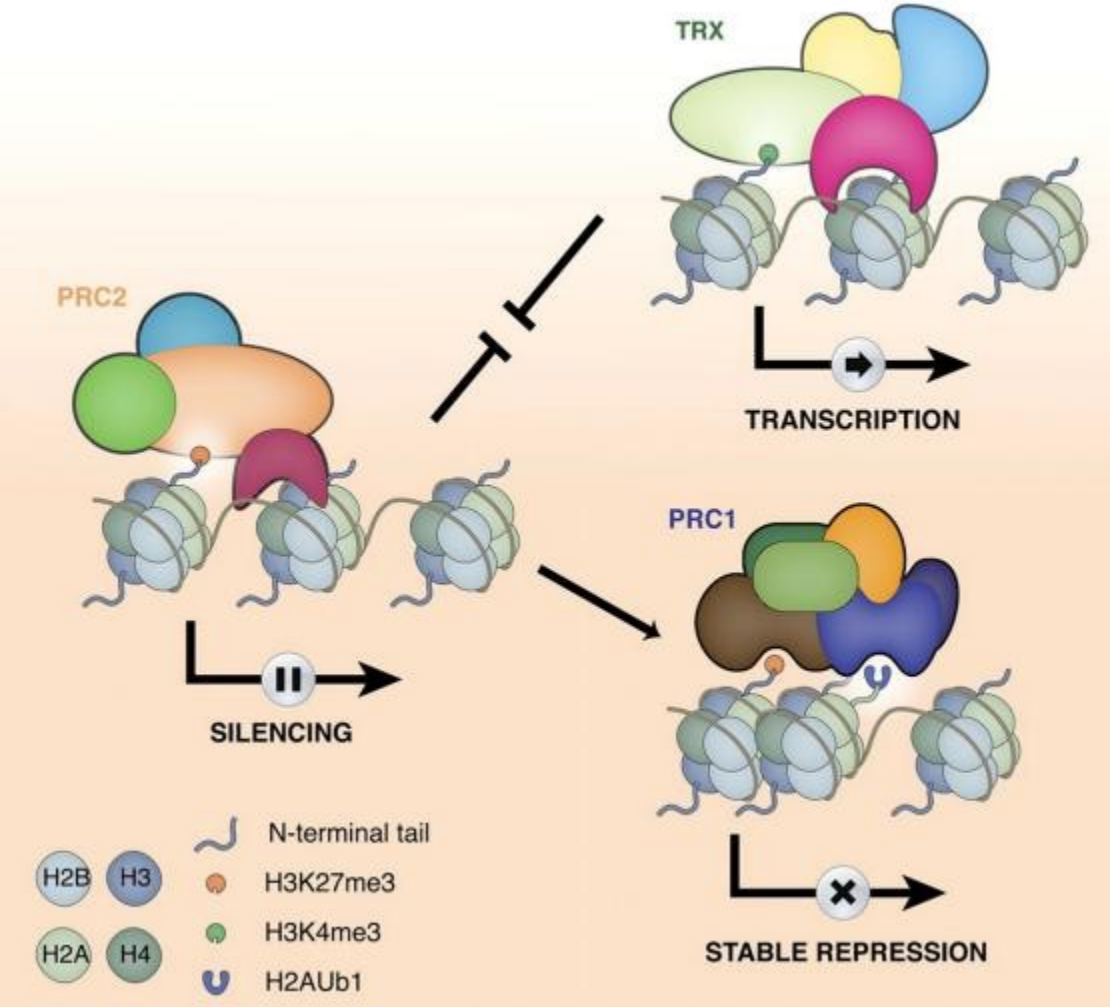
Some of the agents of the histone modifications are:

- 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
- …

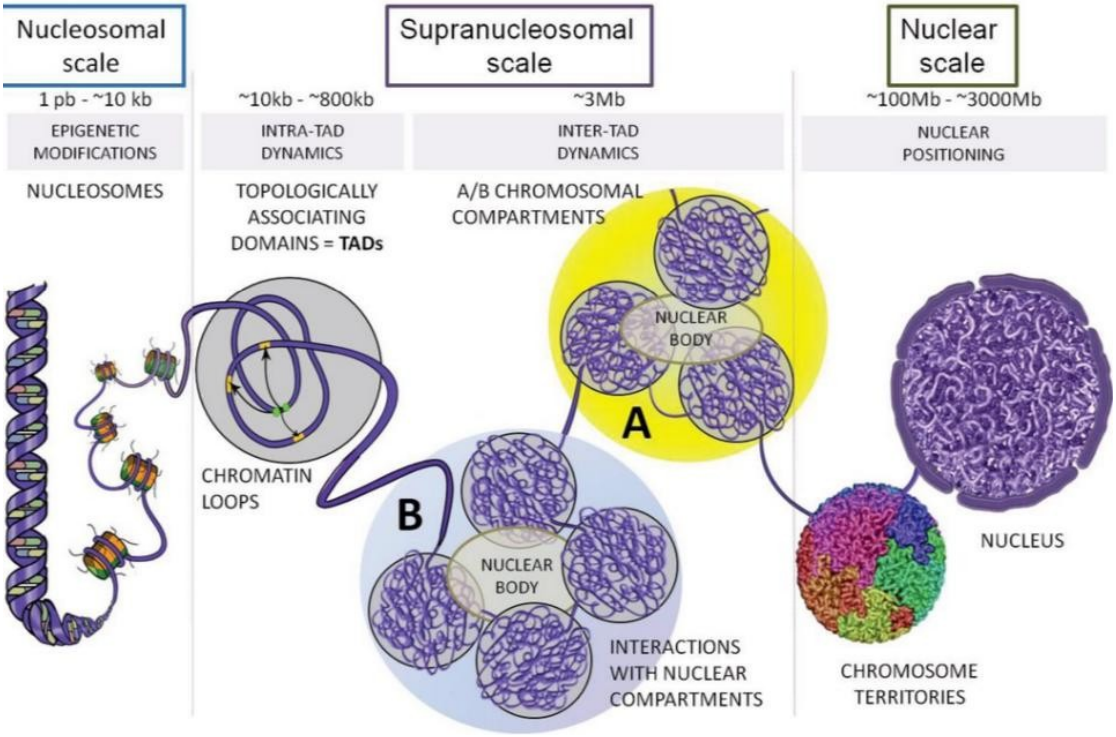
You are going to have reads that map on the genome but there will be some regions where you do not have any read.

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.

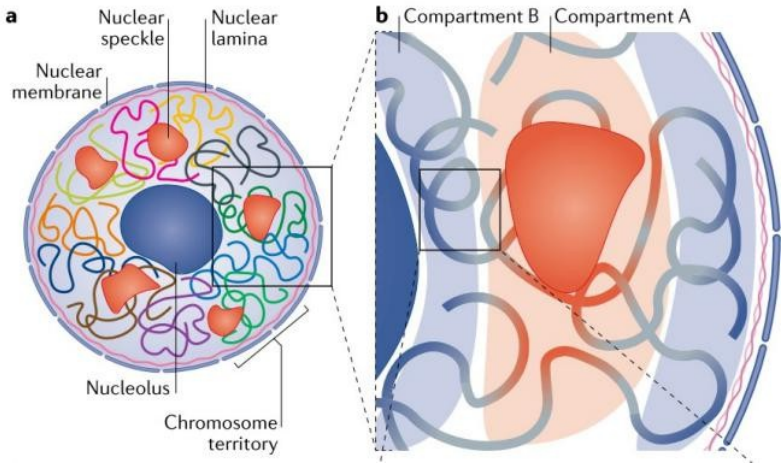
These reads will then be mapped into the genome.

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.

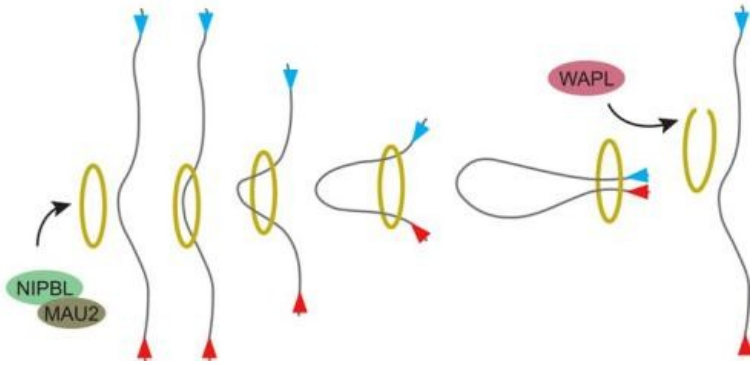
  TADs are defined by boundaries or borders that limit the physical interactions between neighboring TADs. These boundaries act as insulators (such as CTCF), preventing the spread of regulatory signals between adjacent TADs. This insulation helps maintain the independent regulation of genes within each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

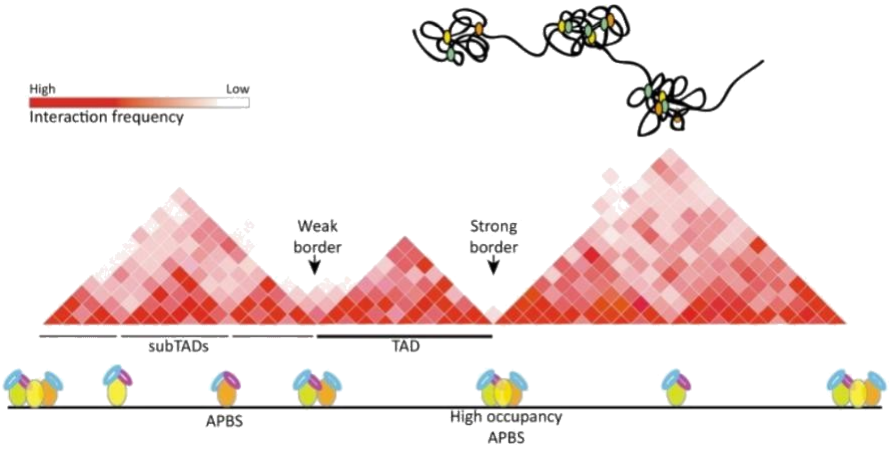
  Note that there is movement between compartment A and B, but the TAD boundaries are conserved across cell types.

  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?

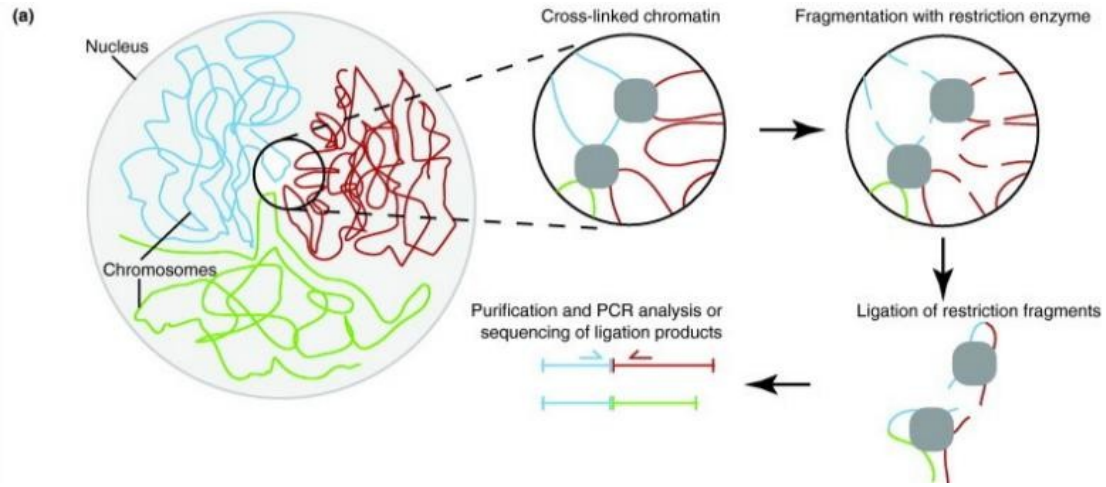
Recruitment of inactive chrX to lamina is important but not sufficient for its inactivation. So, the inactive form does not form TADs also.

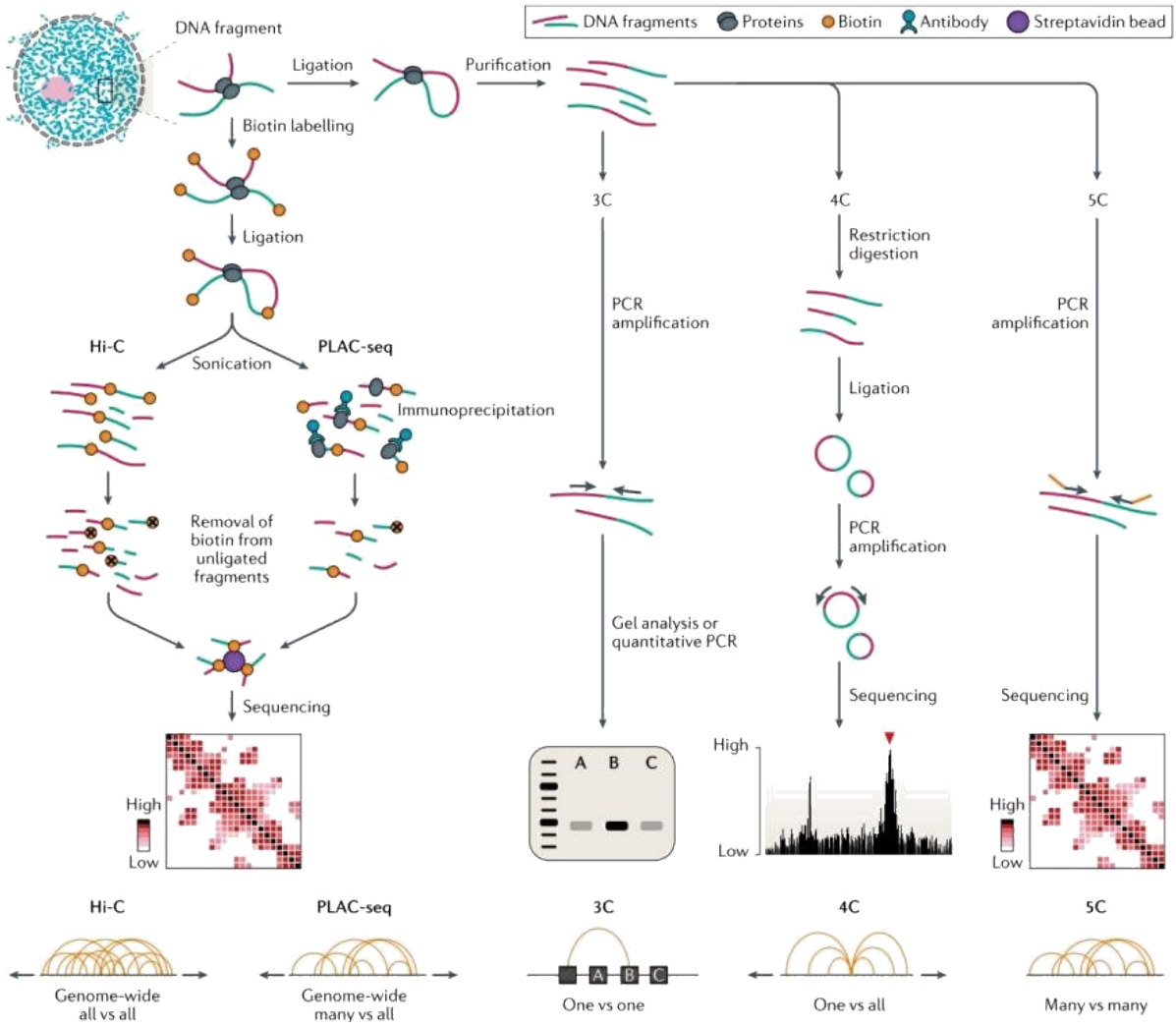
How can we evaluate experimentally and computationally interactions between areas of the genome? Which is the omic technique?

- 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.

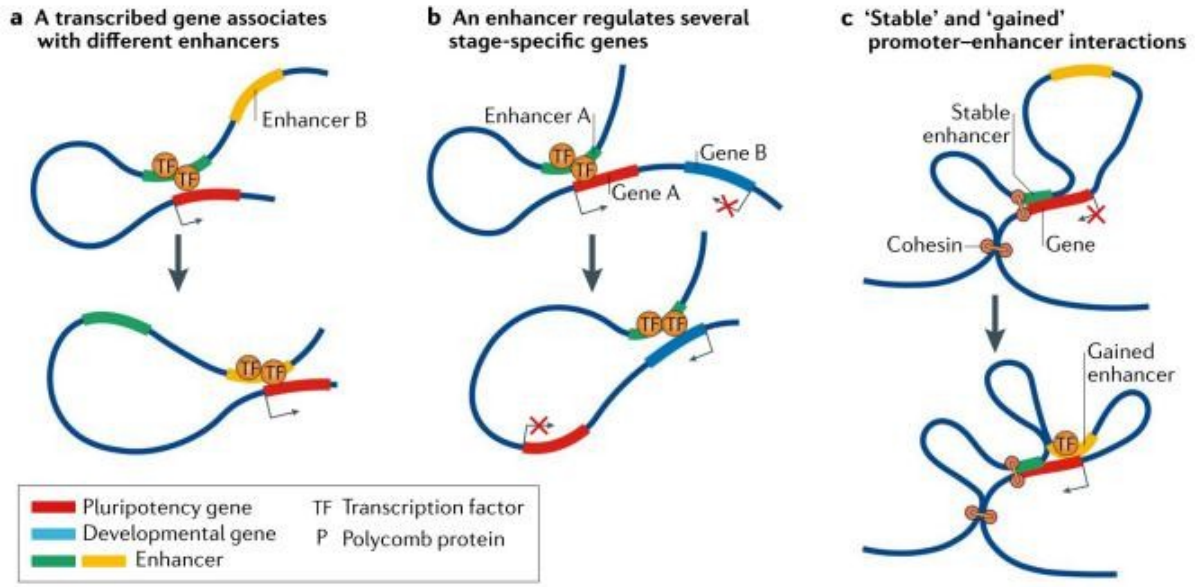
It makes an illumina library of the ligated pairs.

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

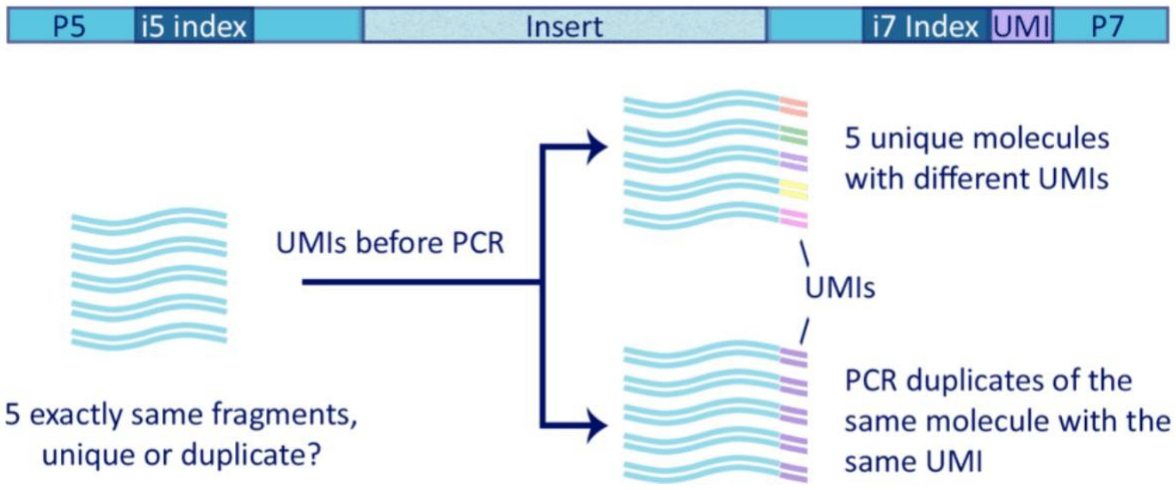
Huge number of measurements → Parallelization strategies of Illumina sequencing:
1. Microfluidics
2. Droplets
3. Molecular barcoding

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)

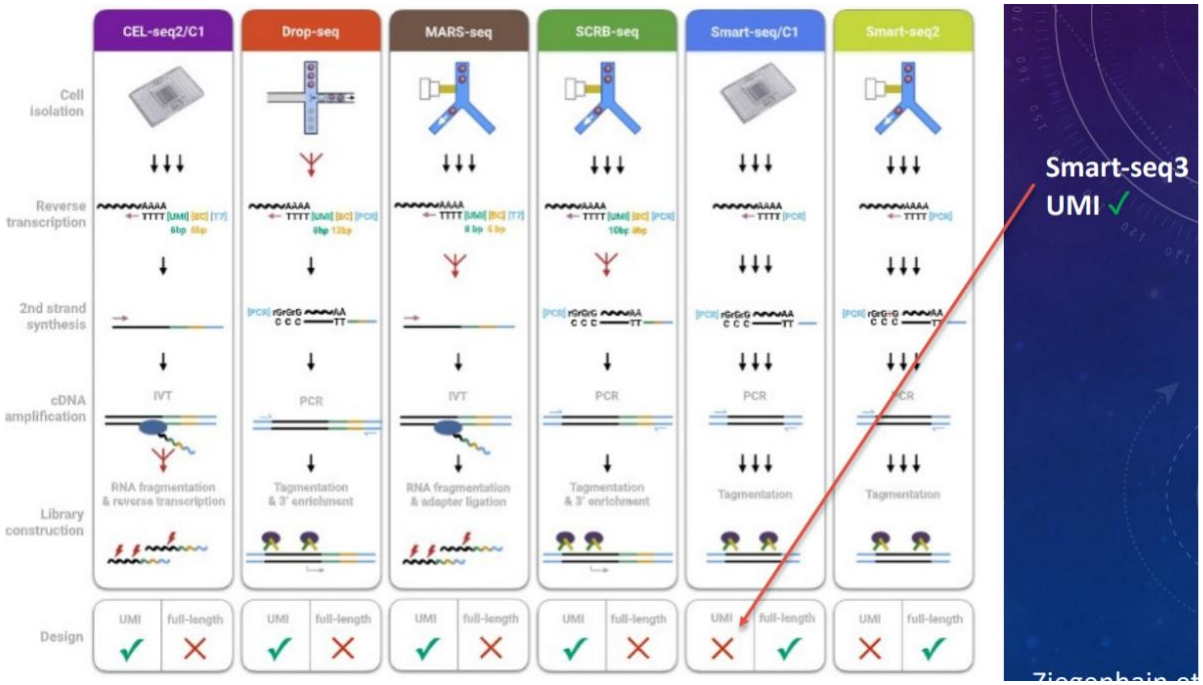


Before PCR, to each one of the reads I add an adapter that has a random combination of 8 nucleotides (the probability of finding the exact combination of 8 nucleotides is really small).

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:

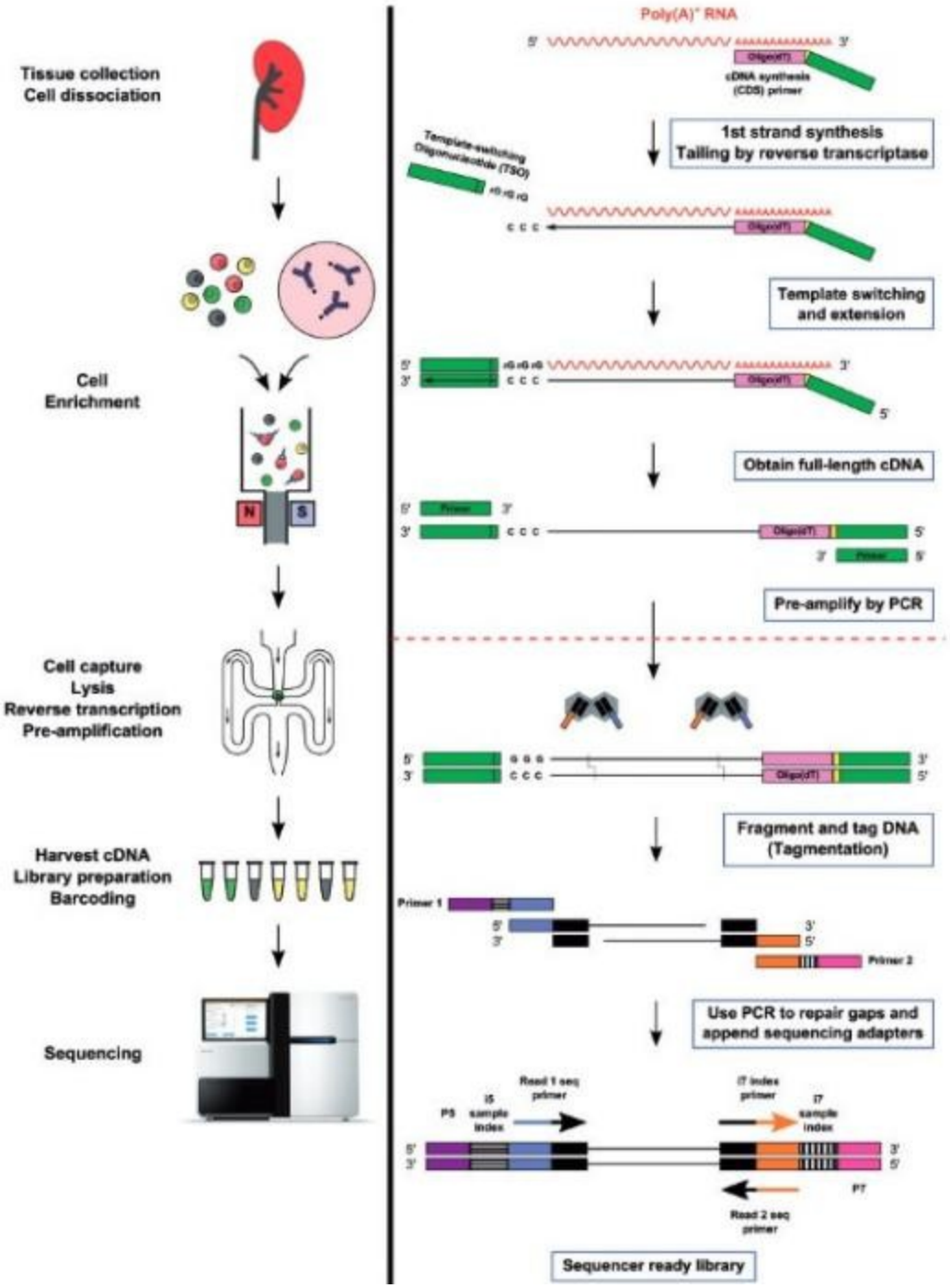
- Fast, you may obtain more droplets
- You want single-cell

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!



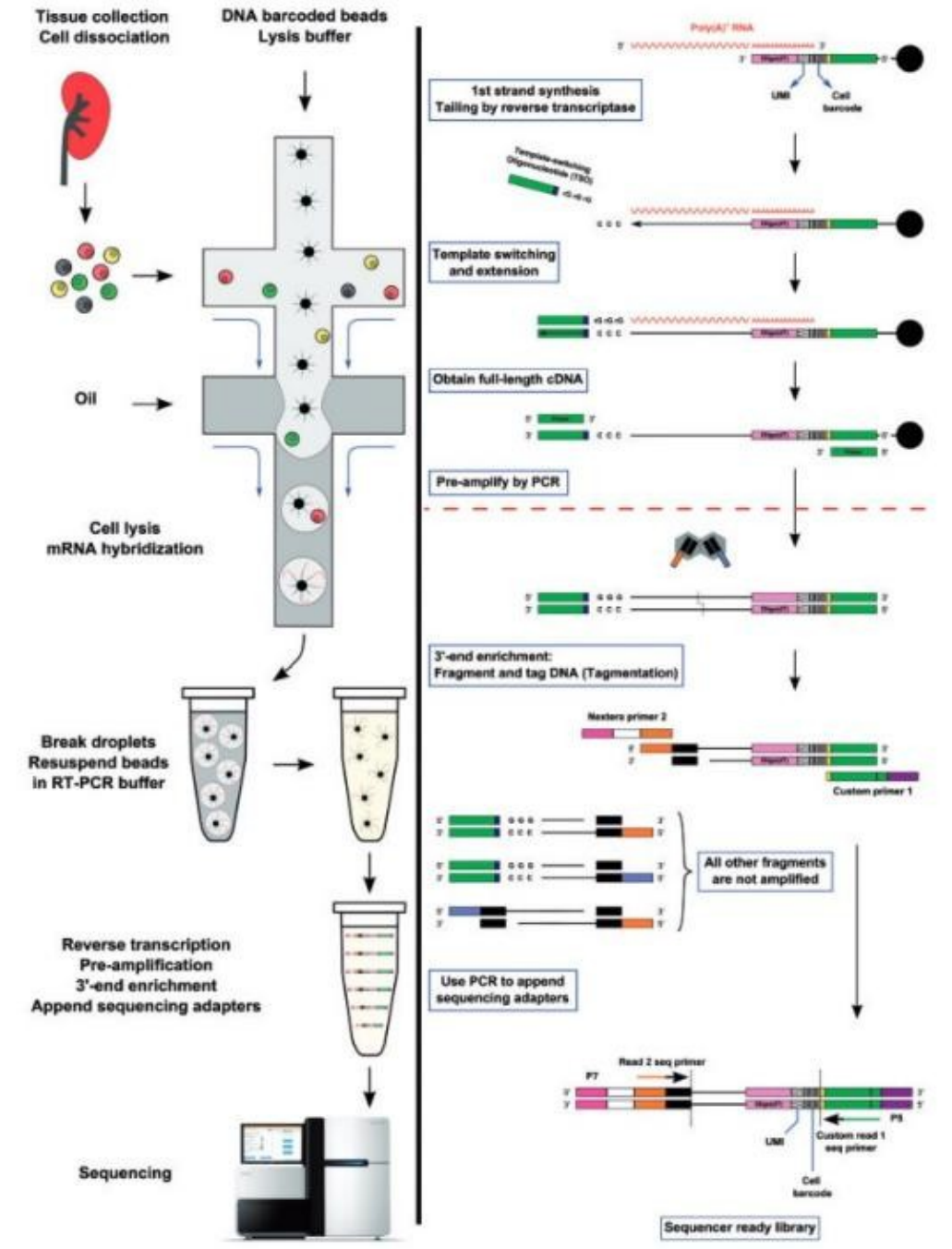
DROPLETS (DROP-SEQ) → Expensive

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)

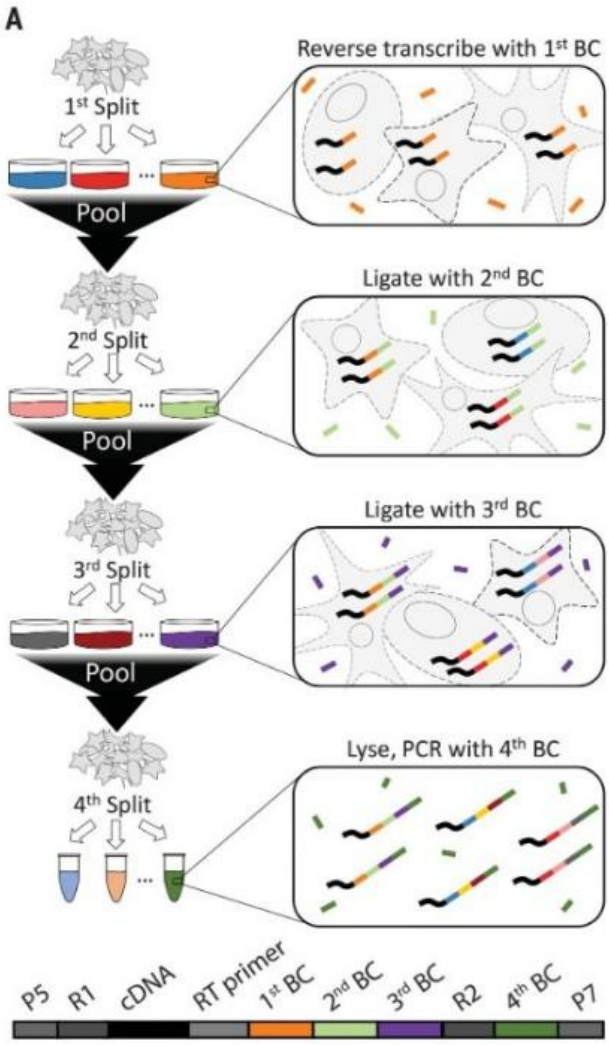
Requires no customized equipment, so it's not expensive!

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:**

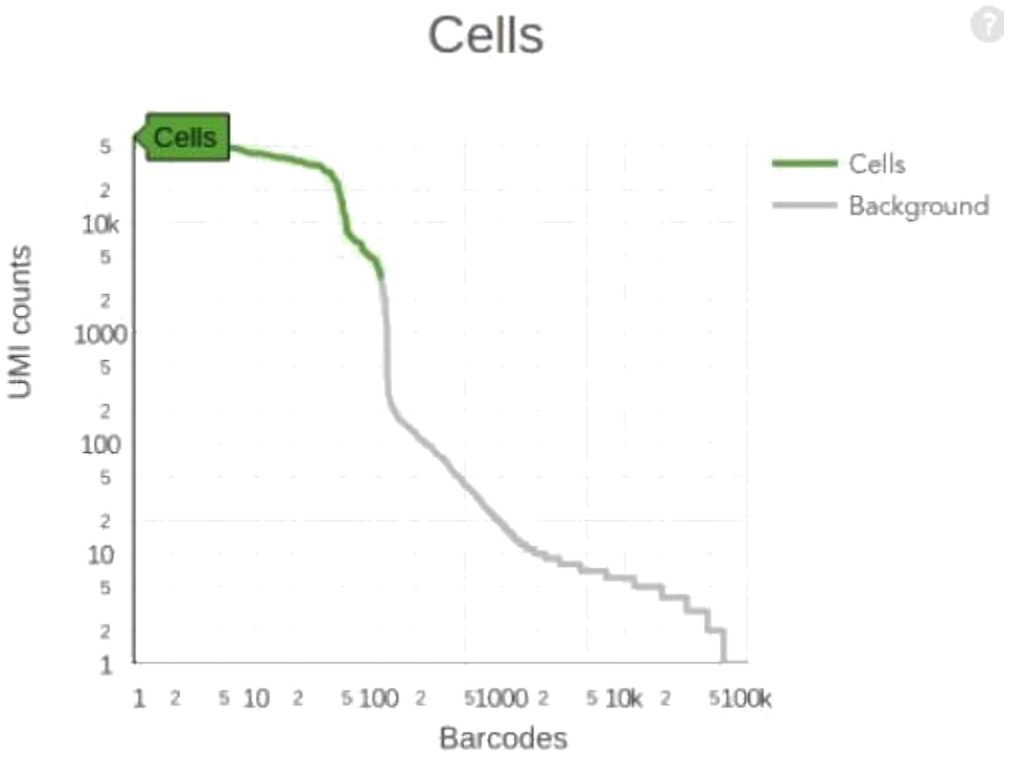
QC: Aims to decide what is a cell and what is noise.

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.

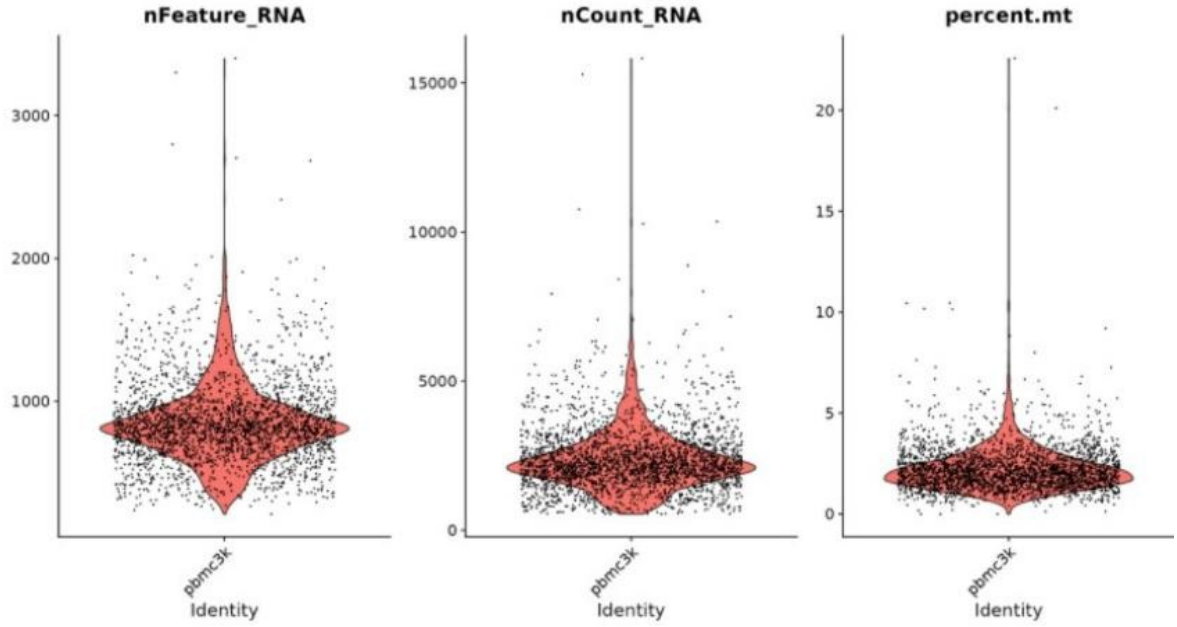
In this case, up to 1000 cells have at least 1000 UMIs.

- 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).



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.

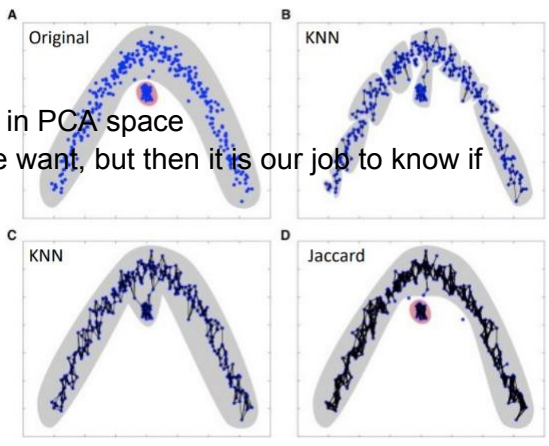
Non-Linear dimensional Reduction (UMPA/TSNE) → Useful for visualization of relationships between cells.

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.

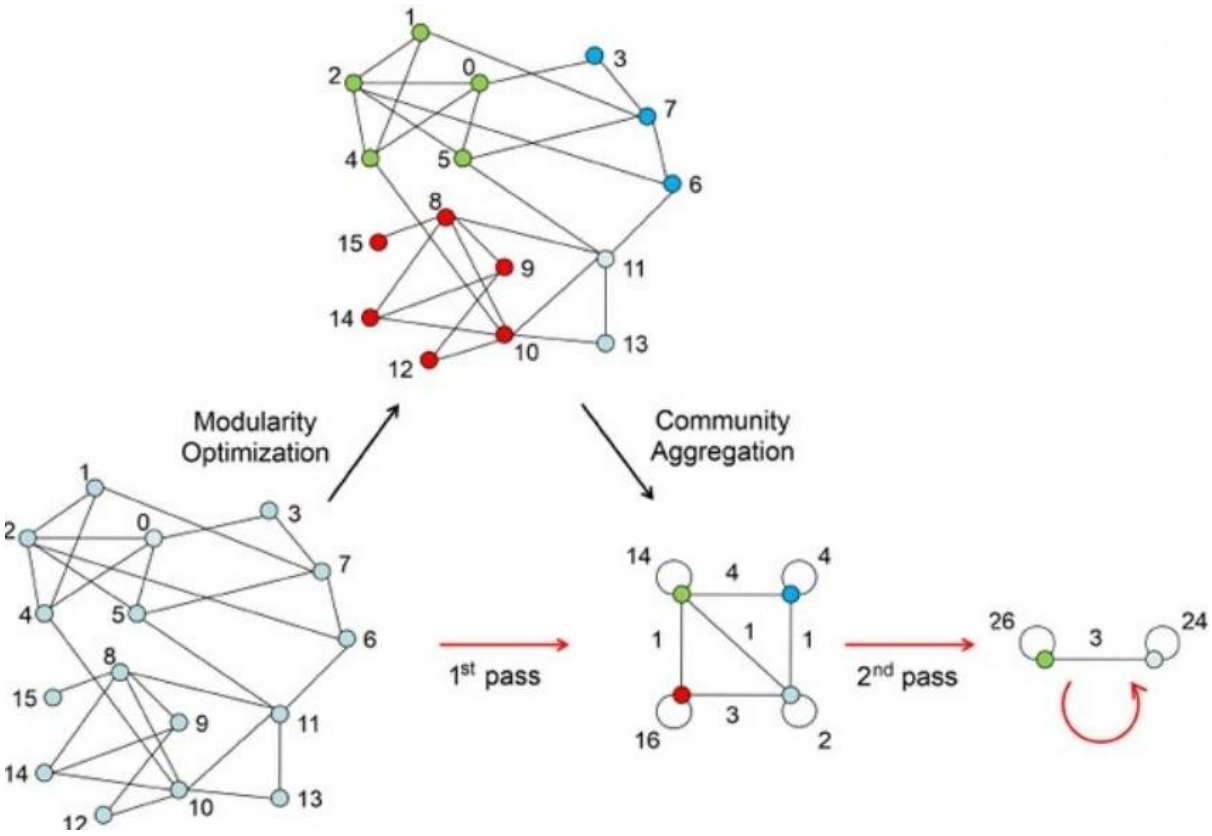
To solve this problem, we use the Jaccard distance → How many of my neighbors I am sharing with the node I am contacting.

- 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

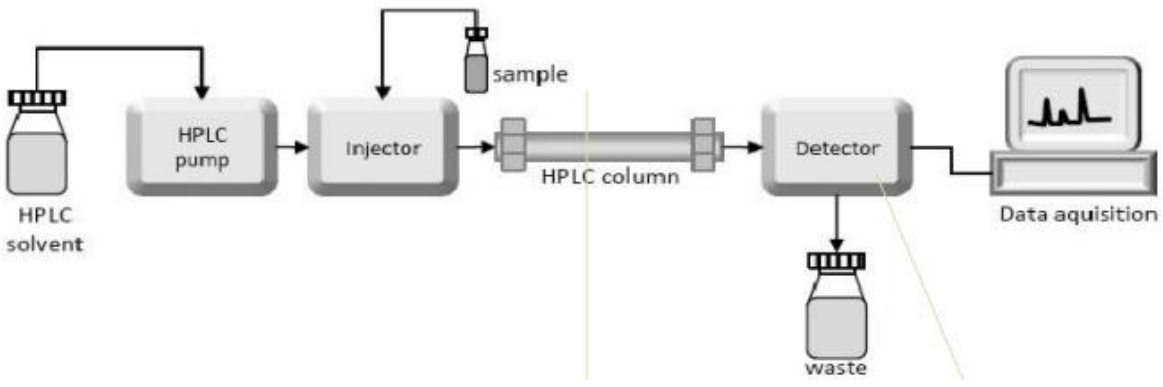
So, we first separate using the different isoelectric points of the proteins. To do this, we use Ampholytes, which establish a pH gradient across the gel that allows proteins to migrate to their isoelectric point during the first dimension of electrophoresis.

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

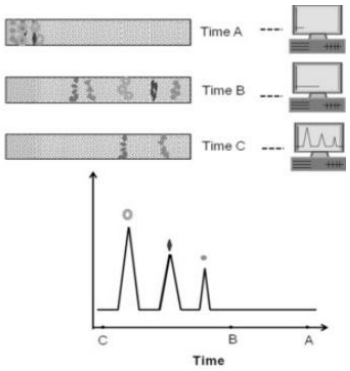
Note that there are some proteins that are preserved better during the time than other proteins. Thus, if you want to know which protein of the brain lasts longer after someone has died, you can do a 2D gel and see the results in 2h, 10h, 50h…

HPLC

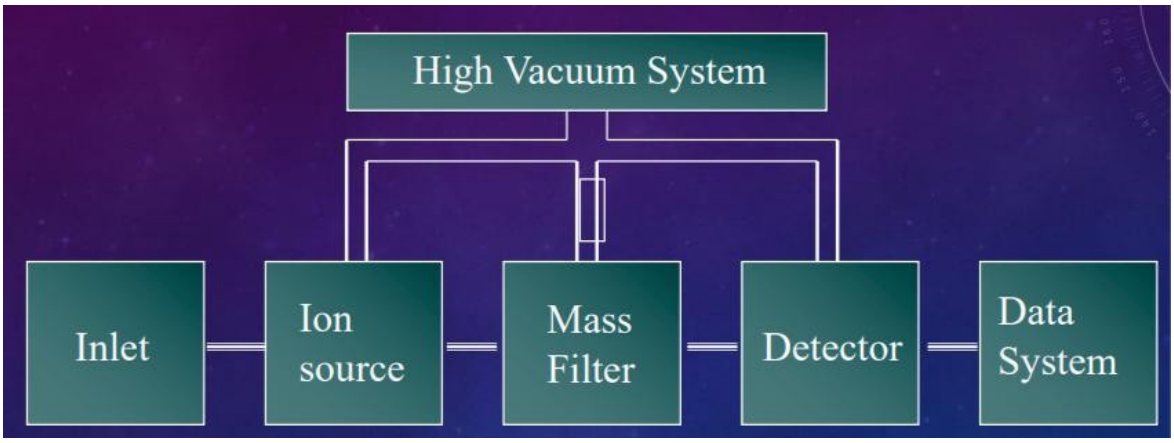
We are separating based on chemical properties (ionic charge, hydrophobicity). We throw the sample into the column and each component will be differently adsorbed in the HPLC column filled with granular solid material.



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

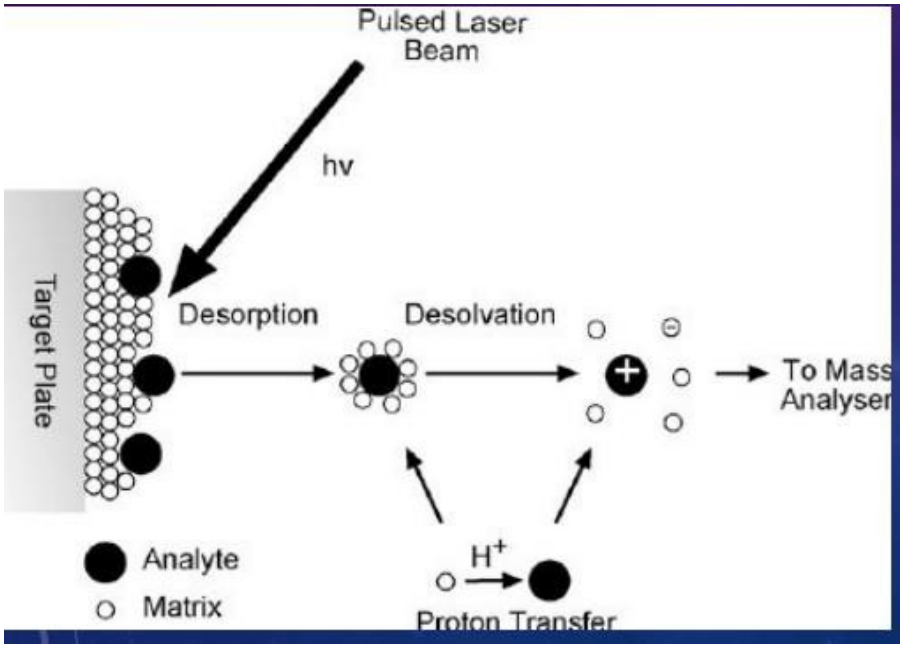
Detector: Will detect how fast the protein comes, calculating the mass/charge ratio
- Multi-channel plate
- Ion cyclotron

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)

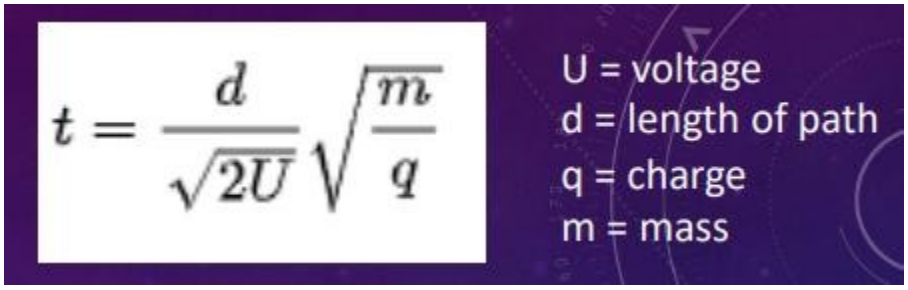
The time each molecule takes to reach the detector it's proportional to its mass.

- 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.

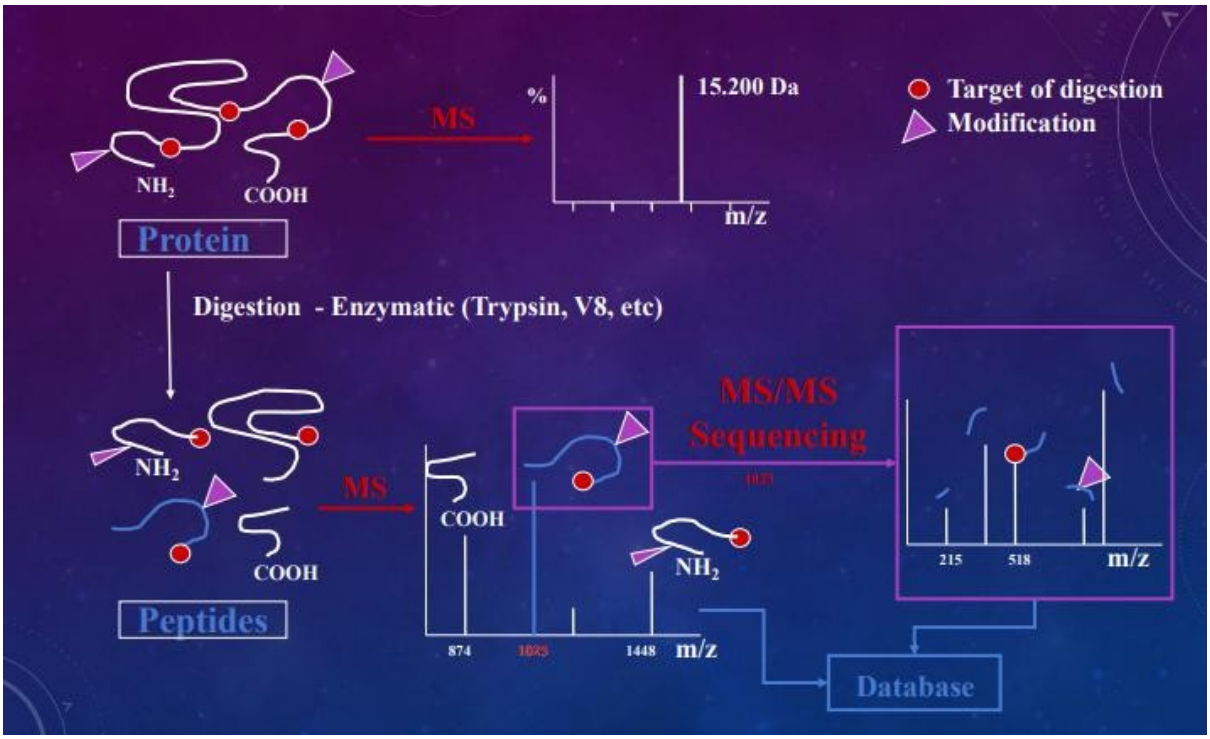
Fragmentation or digestion

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.

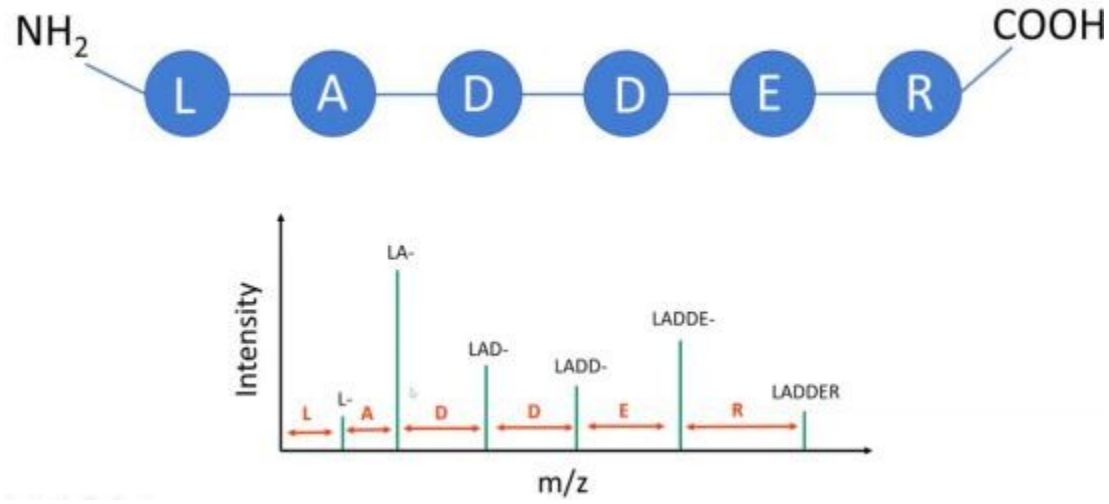
Also, membrane proteins (hydrophobic and large) will never be charged. If I have cytosol proteins, they will charge much better. So, depending on the nature of the protein, we will have different results.
If I cut them in peptides, I will not have this problem.

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.

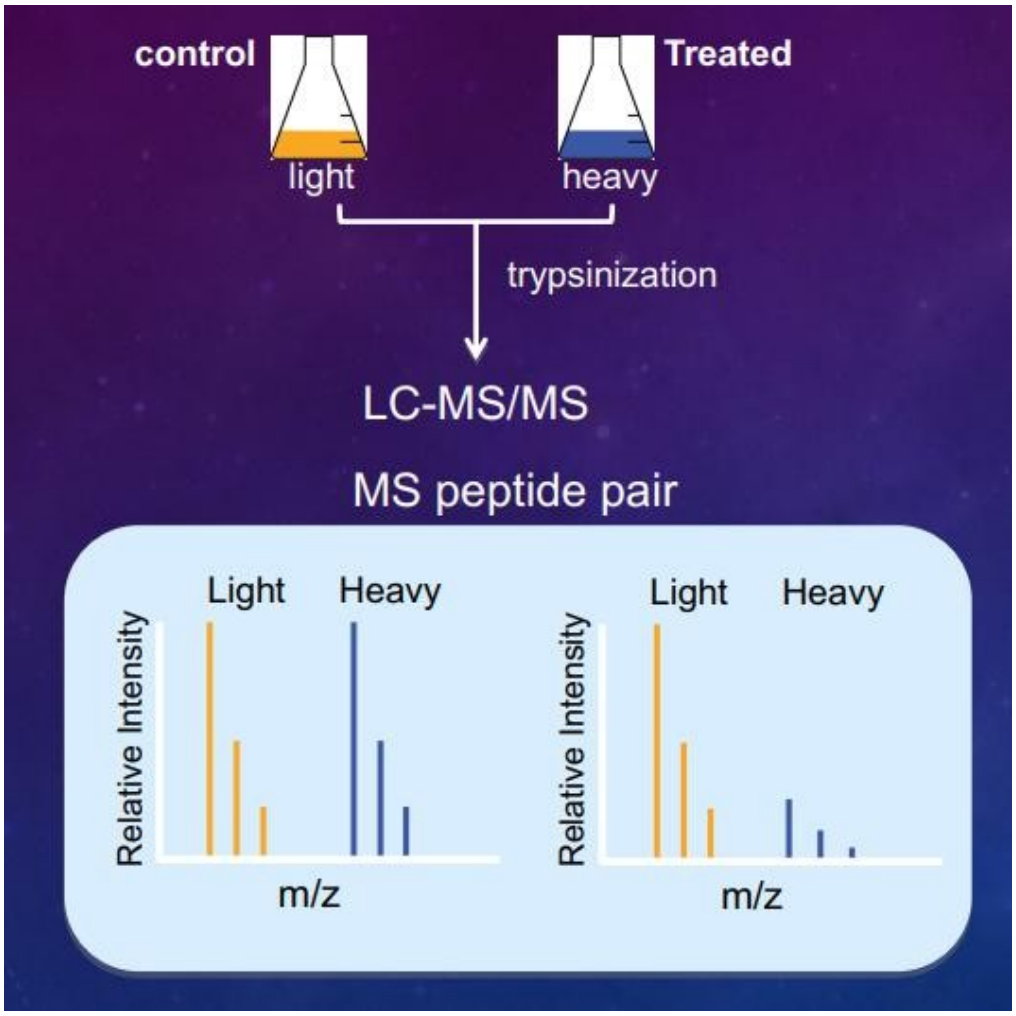
- If the DB is small, you won't find proteins because maybe they are not contained in the DB
- If the DB is very big, you won’t find it because of the false discovery rate (they say it's the protein but it is not).

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.