

Theory-I.pdf



Bioinformatica



Tenicas Omicas



2º Grado en Bioinformática



Escuela Superior de Comercio Internacional
Universidad Pompeu Fabra



MONSIEUR'S
**¿DÍA DE CLASES
infinitas?**



**masca
y fluye**





Topic 1. DNA-seq techniques

Why sequence a genome:

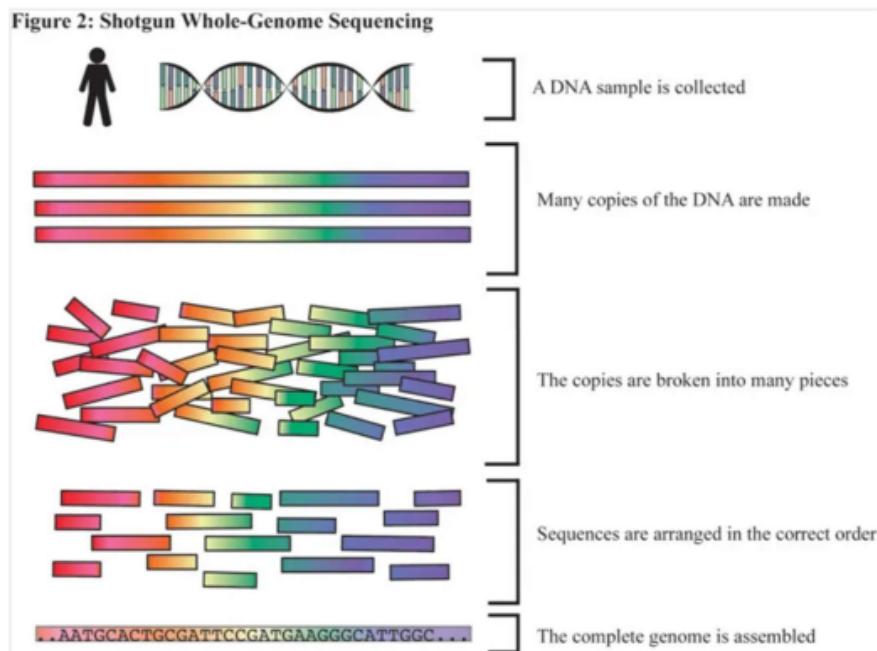
- Description of sequence of every gene valuable. Includes regulatory regions which help in understanding not only the “biochemical” activities of a cell but also ways in which they are controlled.
- Identify & characterize important inheritable disease genes or “useful” functional genes (e.g. bacterial 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

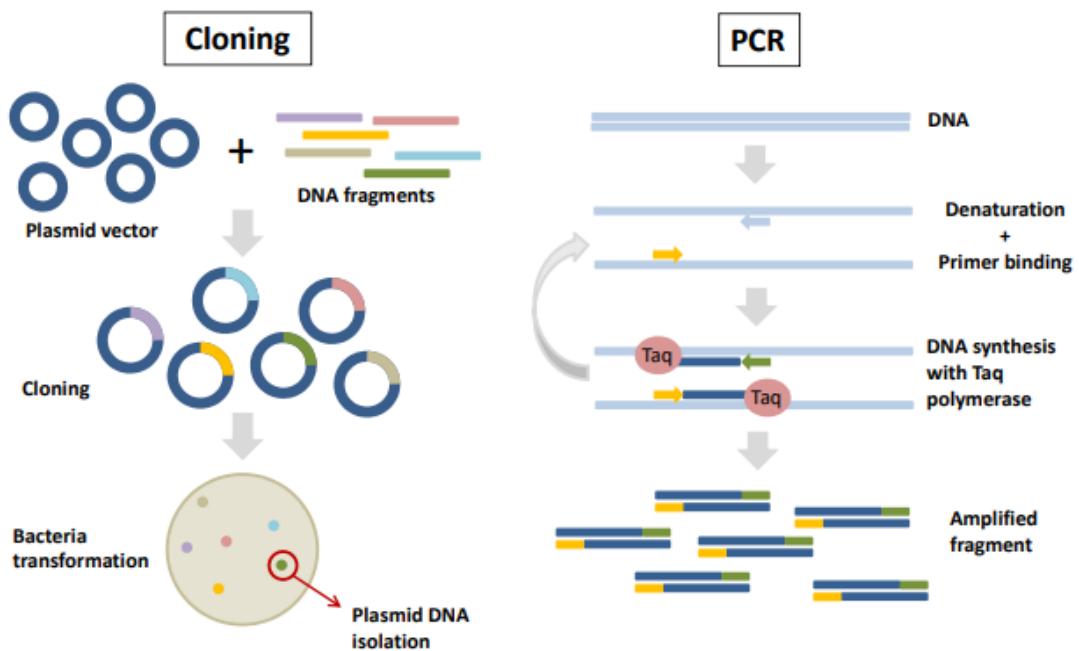
Third generation → PacBio, NanoPore (2011-present)

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 amplify. 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 bacteria in a colony that contain our fragments.
- **PCR:** We need to design the primers, add the nucleotides, change the temperature...



pipasusa

@quieromispipas



Atención: Pipa de la suerte

Instrucciones:

NIVEL 1: frotarla por el codo de tu mejor amigx para aprobar



NIVEL 2: entrar en insta para conseguir matrícula de honor



NIVEL DIOS: seguir nuestra cuenta de Insta para optar a llevarte pipas gratis

#NormalicemosLoNatural

Técnicas Omegas



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



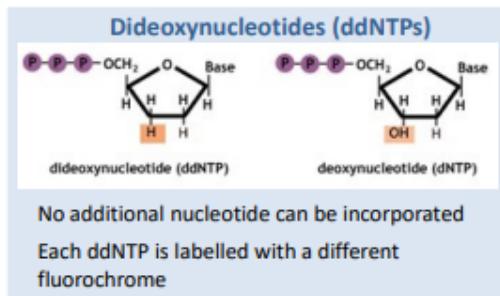
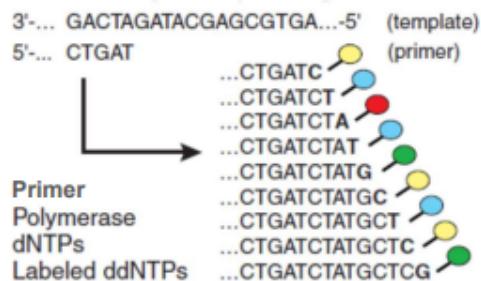
- 1** Imprime esta hoja
- 2** Recorta por la mitad
- 3** Coloca en un lugar visible para que tus compis puedan escanear y acceder a apuntes
- 4** Llévate dinero por cada descarga de los documentos descargados a través de tu QR

Banco de apuntes de la

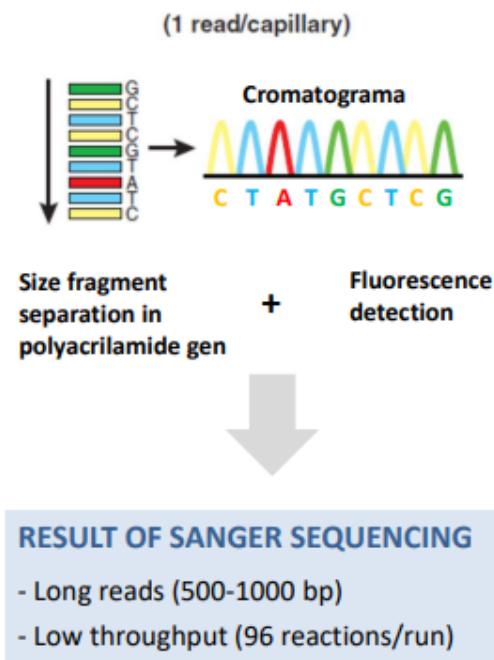
WUOLAH



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



Capillary electrophoresis: To read the results



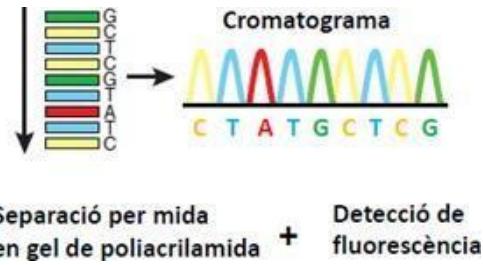
The first nucleotides have a very low signal and therefore we need to trim them.
The last nucleotides also have very low signal and we also trim them.

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 posí un ddNTP. Ara es torna a calentar fins a 90 graus per tornar a separar les cadenes.

Així doncs, podem saber la seqüència dels fragments clonats. Necessitem:

- 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 chromatograma.*
 - *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.



Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)
Calle Pelayo, 5. 08001 Barcelona

¡Únete y recibe una bebida de regalo!



NEW YORK BURGER
A fuego, but lento

NEW YORK BURGER
A fuego, but lento

Calle Pelayo, 5.
08001 Barcelona

¡Únete y recibe una bebida de regalo!



ONE WAY

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)

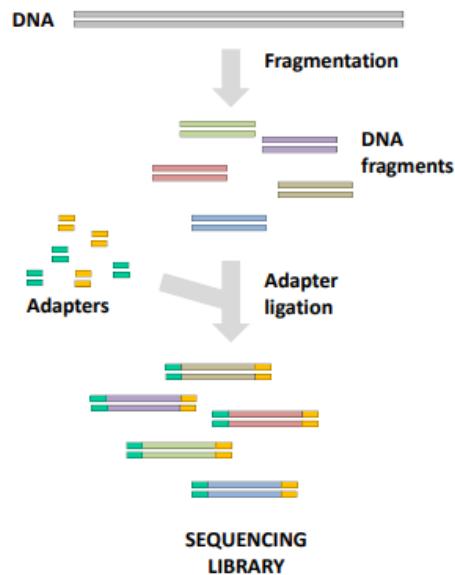
Limitations of classic sequencing techniques

- 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

WUOLAH

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).

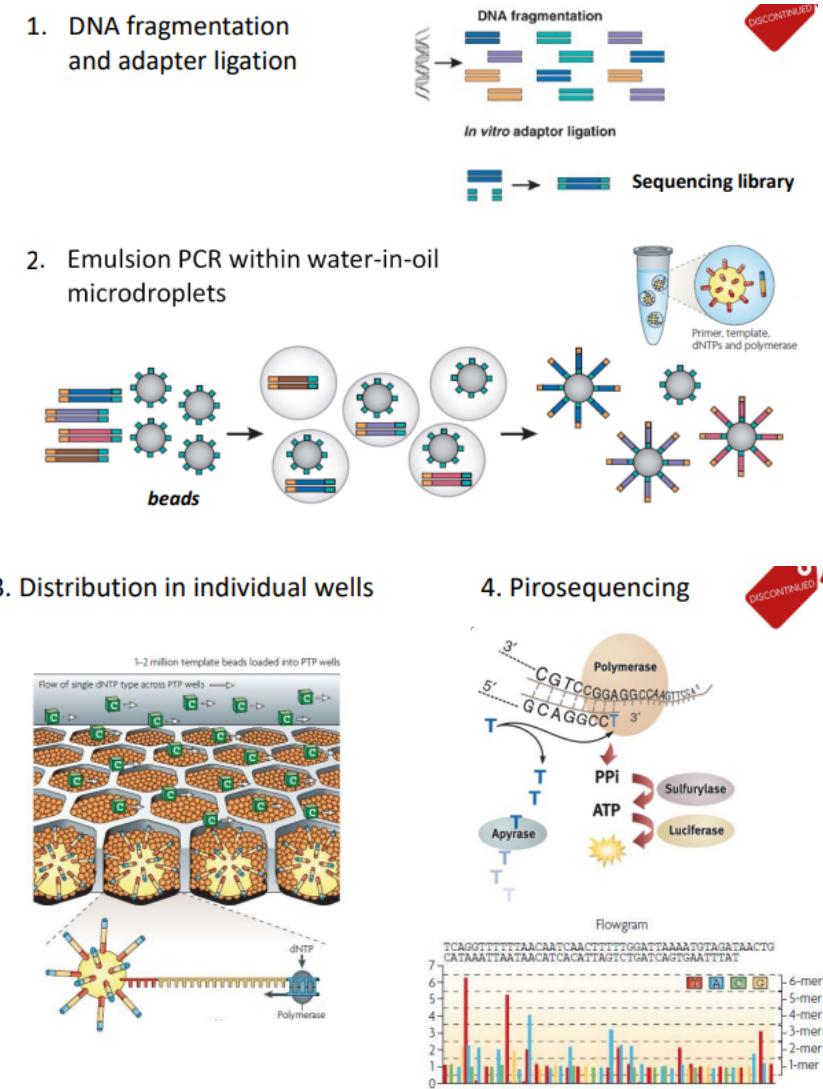
Estem seqüenciant de forma massiva, però d'un nucleòtid en un.

Roche 454 - Pyrosequencing

The difference is in the way they do the amplification.

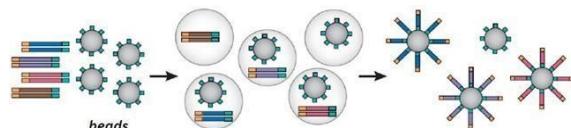
In this case, they use an emulsion PCR within water-in-oil microdroplets.

It's 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 pou que hagin afegit aquell nucleòtid, **emetran llum** (luciferasa) que serà detectada i es farà una foto i se sabrà quins pou 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).

- Fragmentació del DNA i lligament d'adaptador
- Amplificació en fase oliosa per emulsion-PCR
- Distribució en poues 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.

- Fragmentació del DNA i lligament d'adaptador
- Amplificació en fase olio-sapèremulsion-PCR
- Distribució en pous d'un chip semiconductor
- Piroseqüenciació → flowgrama

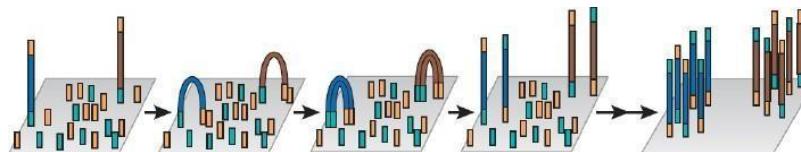


Illumina

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 gaps!

- Fragmentació del DNA i lligament d'adaptador
- Amplificació en fases sòlidiques per bridge-PCR
- Circulació de dNTPs terminadors fluorescents
- Incorporació d'un sol nucleòtid en cada cluster
- Lectura per imatges seqüencials

of

	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 alt, 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

$$N_{max} = M / (T * P)$$

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.

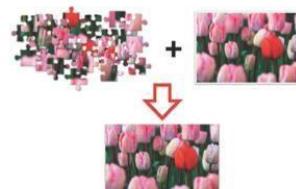
L'inconvenient d'aquesta metodologia és que no permet la detecció de seqüències noves ni les reorganitzacions estructurals. Si entre un individu i un altre de la mateixa espècie hi ha seqüències diferents o fragments que estan invertits en posicions diferents, trobarem incongruències en el posicionament 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 memòria de computació.



WUOLAH

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

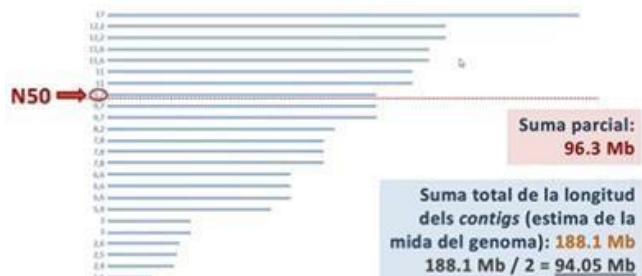
Es la forma de mesurar la qualitat del genoma.

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

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 94.05 Mb 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

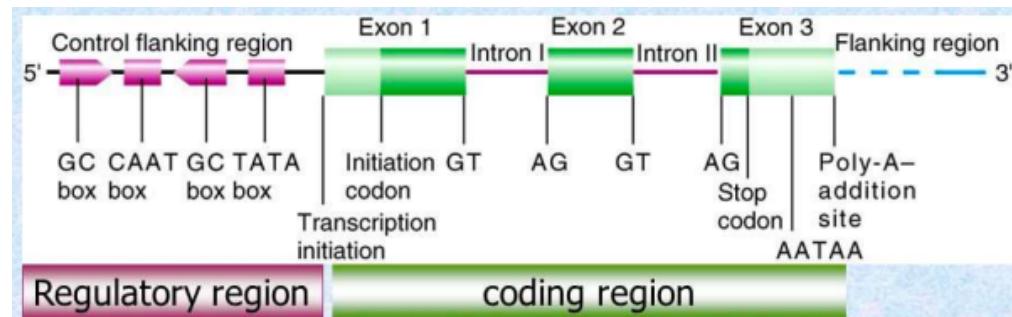
What is a gene?

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

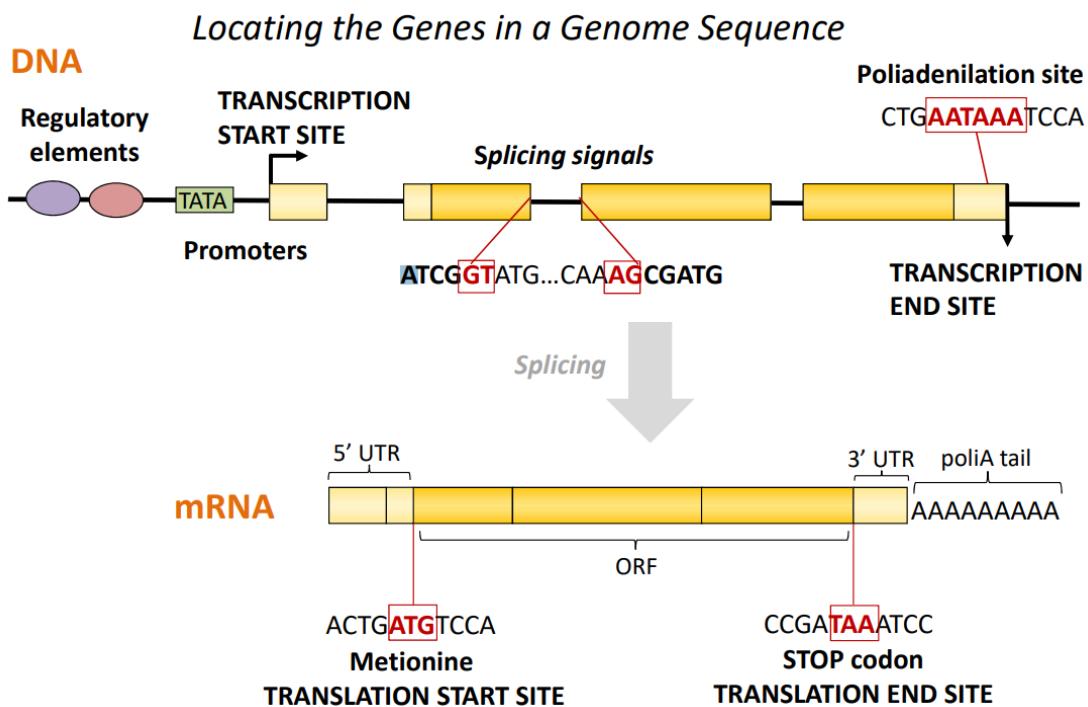
Ab initio gene prediction

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

- Splicing signals
- Poliadenylation 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

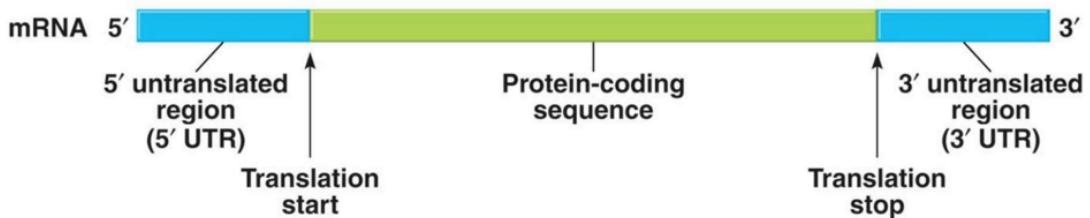
The study of RNAs gives information about:

- 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%).

RNA Processing: Eukaryotic mRNAs



Copyright © 2006 Pearson Benjamin Cummings. All rights reserved.

- 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 conegeus.

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.

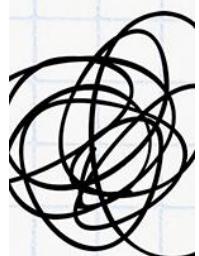
- Si total la sonda queda unida a RNA s'obté una gran fluorescència (diferents nivells de fluorescència segons la seva expressió) → Quantifiable el nivell d'expressió
- Si no hi ha hibridació, queda negre.

Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

pierdo
espacio



Necesito
concentración

ali ali ooooh
esto con 1 coin me
lo quito yo...

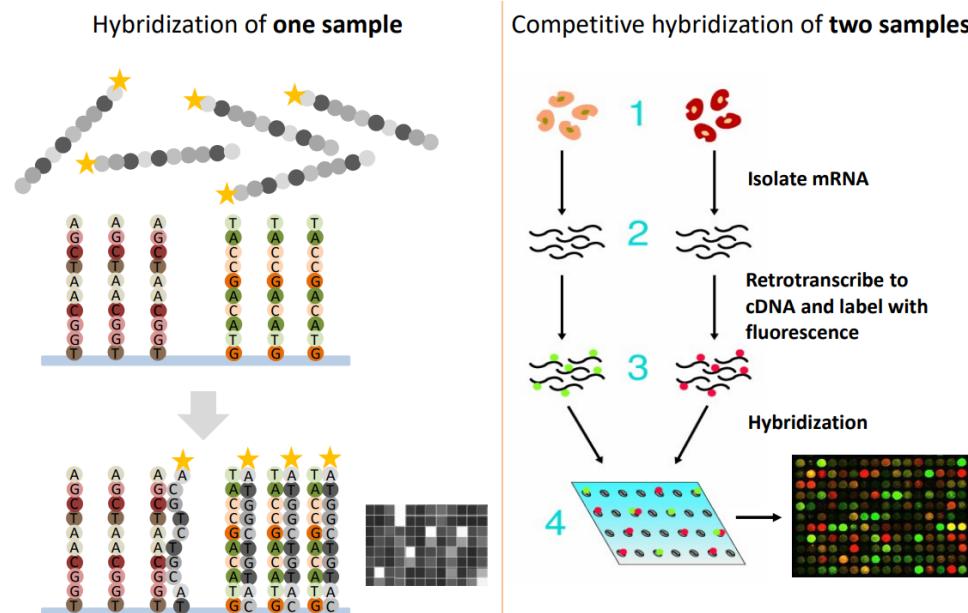
wuolah

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).

wuolah

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

- 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 emmascrenen 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 genotiques 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.

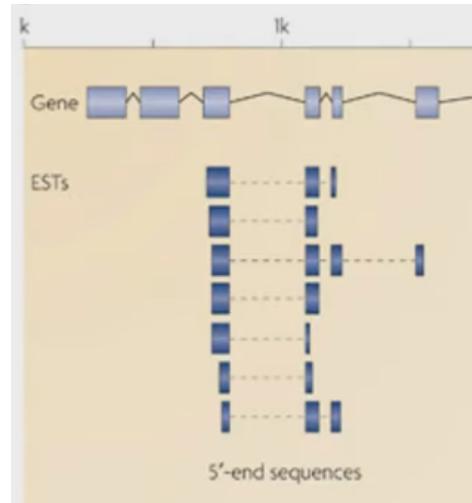
1. S'aprofita la presència de la cua de poliA per capturar els transcrits de RNA.
2. Síntesi de cDNA: es retrotranscriu el RNAs en cDNA amb una *transcriptasa inversa*.
3. Genoteca de cDNA: es fragmenta el cDNA, es filtreuen els fragments d'interès, es clonen en un vector, es transfereixen a cèl·lules 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) són 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.

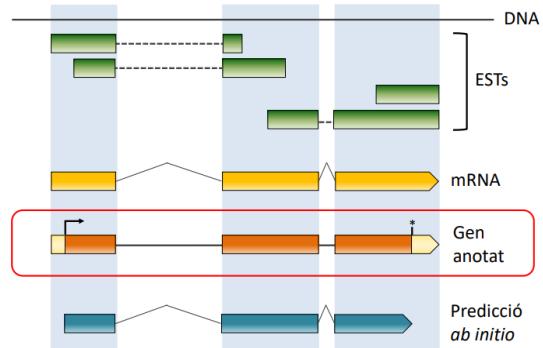
En el RNA ja no hi ha introns, però quan es mapeja contra un genoma que sí conté els introns, els reads de RNA (cDNA) queden dividits en fragments amb uns gaps enmig que corresponen als introns (els fragments acostumen a ser d'1kb si es seqüència amb Sanger i els exons acostumen a ser més curts i llavors cada fragment pot contenir més d'un exó). Aquesta tècnica delimita on estan els diferents exons i col-lateralment, indica també on estarán 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 predicción *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 ESTs 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 quantifiable. 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 quedan 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 seqüènciació massiva de cDNA.

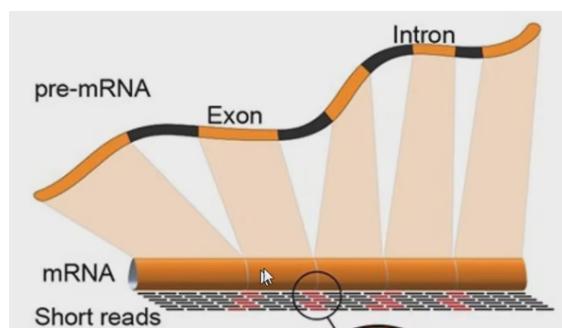
És una tècnica molt similar als ESTs però no es basa en Sanger sinó en l'ús d'adaptadors, fet que permet seqüenciar massivament, ja que és pot seqüenciar el RNA de forma directa usant tecnologies de nova generació. Aquest mètode facilita una major precisió en la mesura dels nivells de transcrits i les 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. Seqüènciació amb adaptadors: Els fragments es poden seqüenciar per *illumina*.
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 corresindrà 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 seqüènciació 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



Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)
Calle Pelayo, 5. 08001 Barcelona

¡Únete y recibe una bebida de regalo!



NEW YORK BURGER
A fuego, but lento

NEW YORK BURGER
A fuego, but lento

Calle Pelayo, 5.
08001 Barcelona

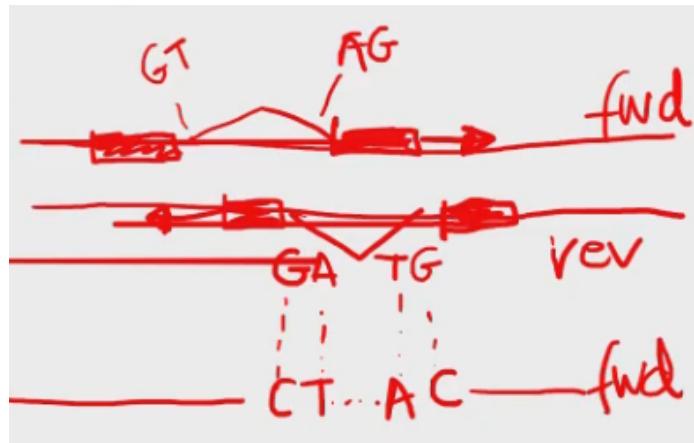


Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)



També podem saber la direccionalitat dels gens donada per la junction reads gràcies a les denials 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 ...)

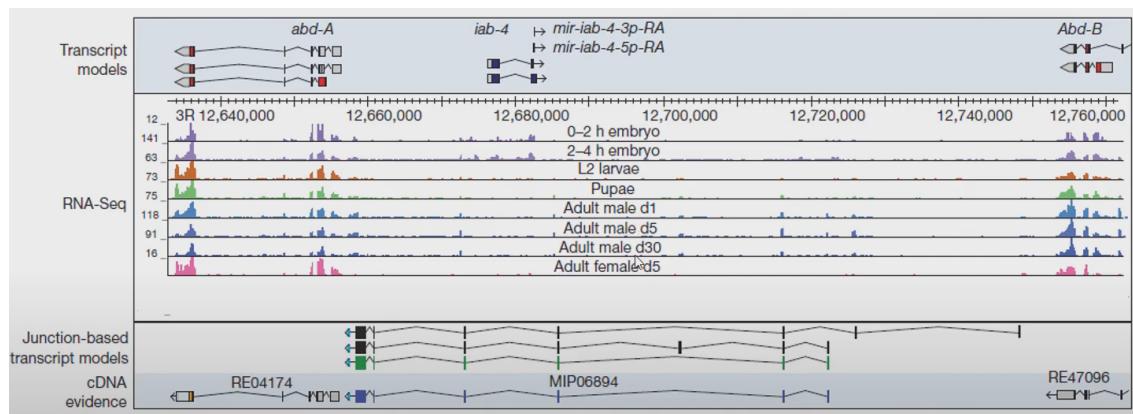


2.2.1. AVANTATGES DEL RNA-SEQ

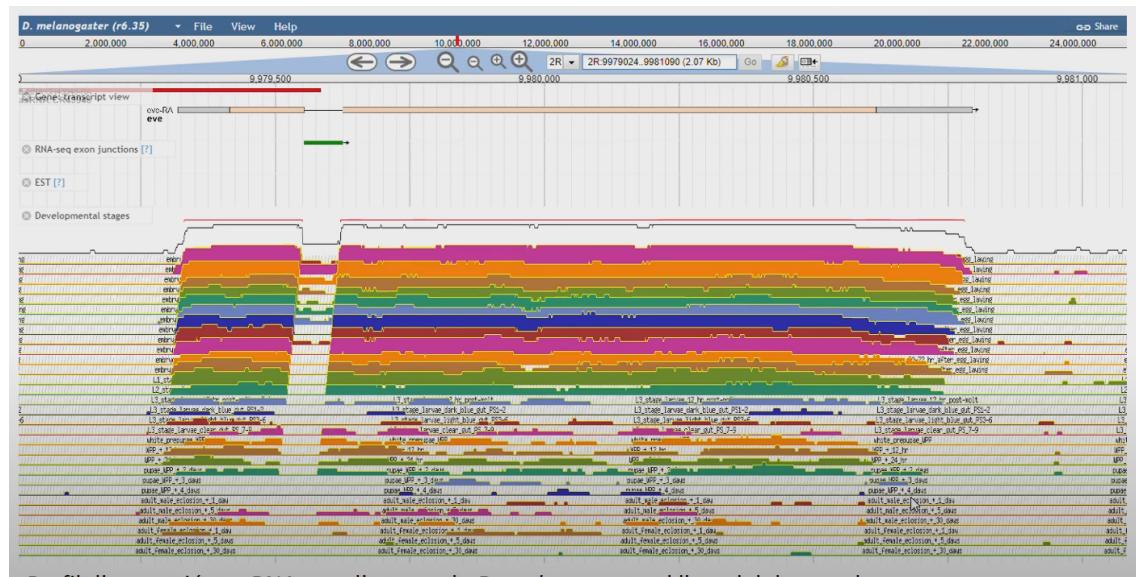
- La seqüència genòmica no cal que estigui **anotada** prèviament.
- Permet la detecció de nous transcrits o fins i tot nous gens.
- Gran precisió en la detecció dels límits dels transcrits gràcies a fer servir *reads* molt curts.
- Permet detectar variants de splicing i principis i finals de transcripció alternativa.
- 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 reproduïibilitat.
- Requereix molt poca quantitat de DNA inicial.

WUOLAH

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.



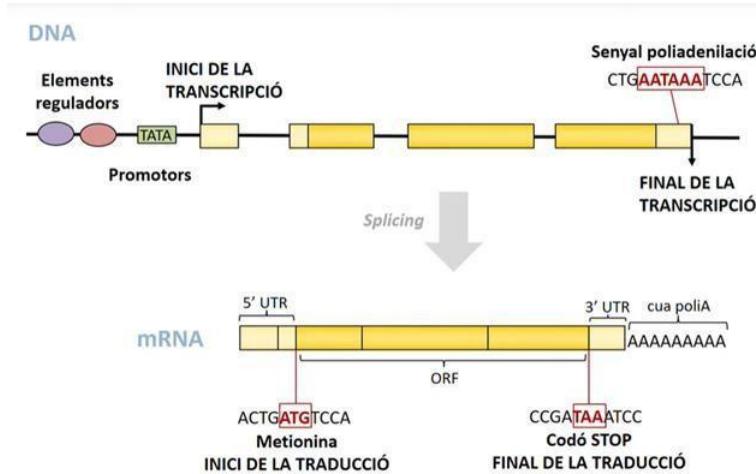
1. Amb els *gene expression arrays* només podem analitzar l'expressió d'aquells gens prèviament anotats al genoma i pels quals tinguem sondes? Cert
2. Quin és el número mínim de *gene expression arrays* que hauries d'hibridar per analitzar els nivells d'expressió de 12 gens del teu interès en 84 teixits 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

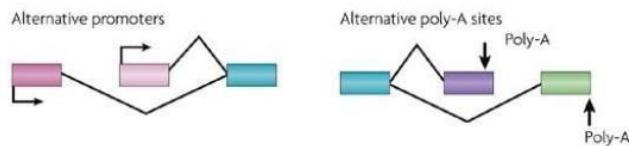


1. SPLICING ALTERNATIU

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



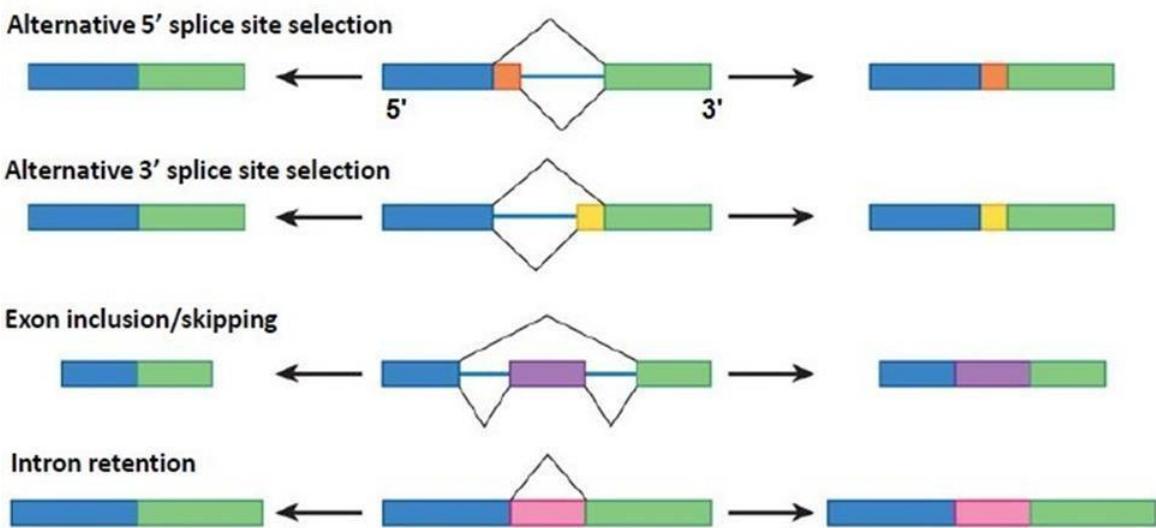
Estan afectats els exons inicials o finals.

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



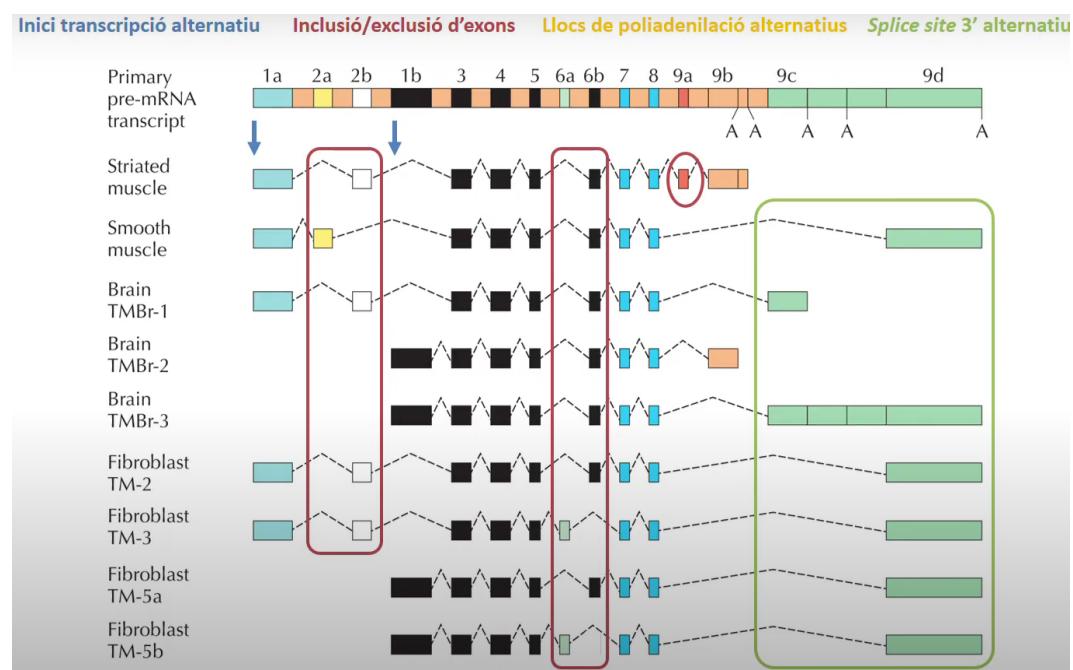
- **Splicing alternatiu en 5':** en aquest cas, la part que s'està eliminant de forma desigual és la part 5' de l'intró (queda retinguda una part de l'extrem 5' de l'intró).
- **Splicing alternatiu en 3':** en aquest cas, la part que s'està eliminant de forma desigual és la part 3' de l'intró (queda 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.

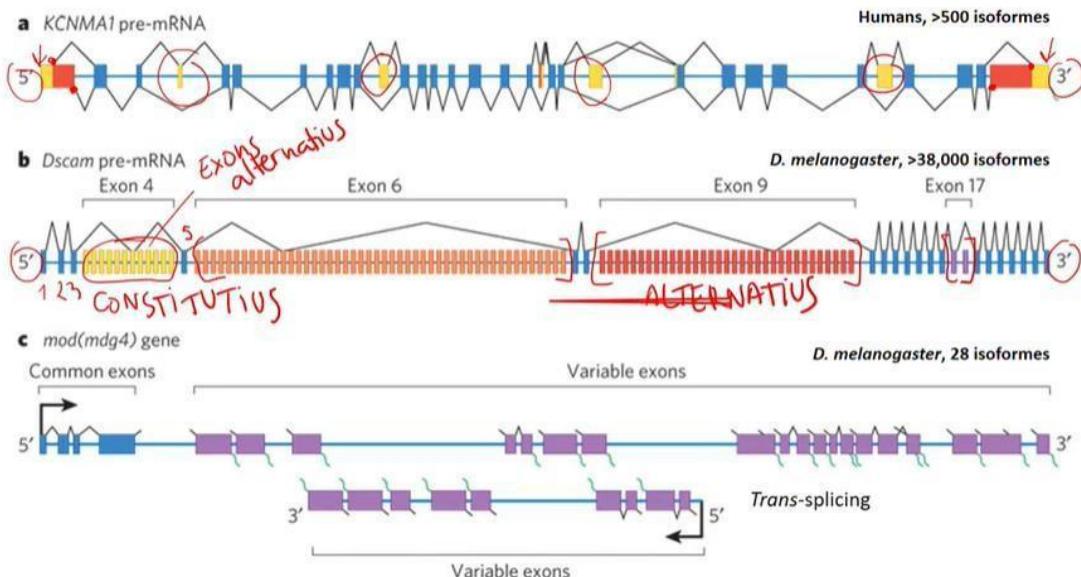
- **Inici alternatiu:** els inicis de la transcripció són diferents, alguns comencen en les caixes blaves i altres en les negres.
- **Inclusió/exclusió d'exons:** els exons del grup 2 són mútuament excluents perquè o hi ha presència (s'inclou) un exó o l'altre, però mai estan els dos. En els exons del grup 6 passa el mateix.
- **Llocs de poliadadenilació 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.



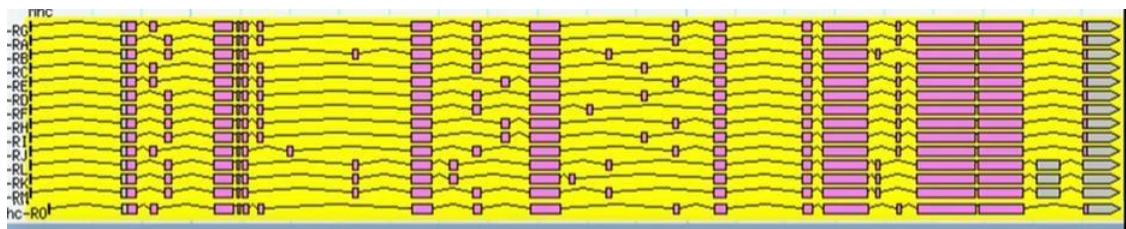
TEST SPlicing

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:



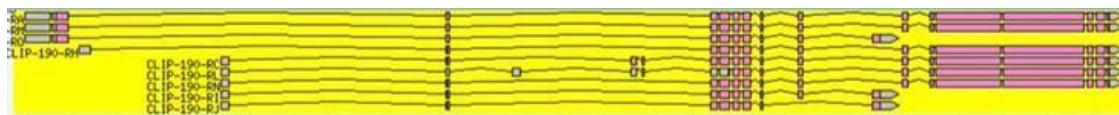
- Inici de la transcripció alternatiu.
- Exons mutuament excloent.

PREGUNTA 3:



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

PREGUNTA 4:

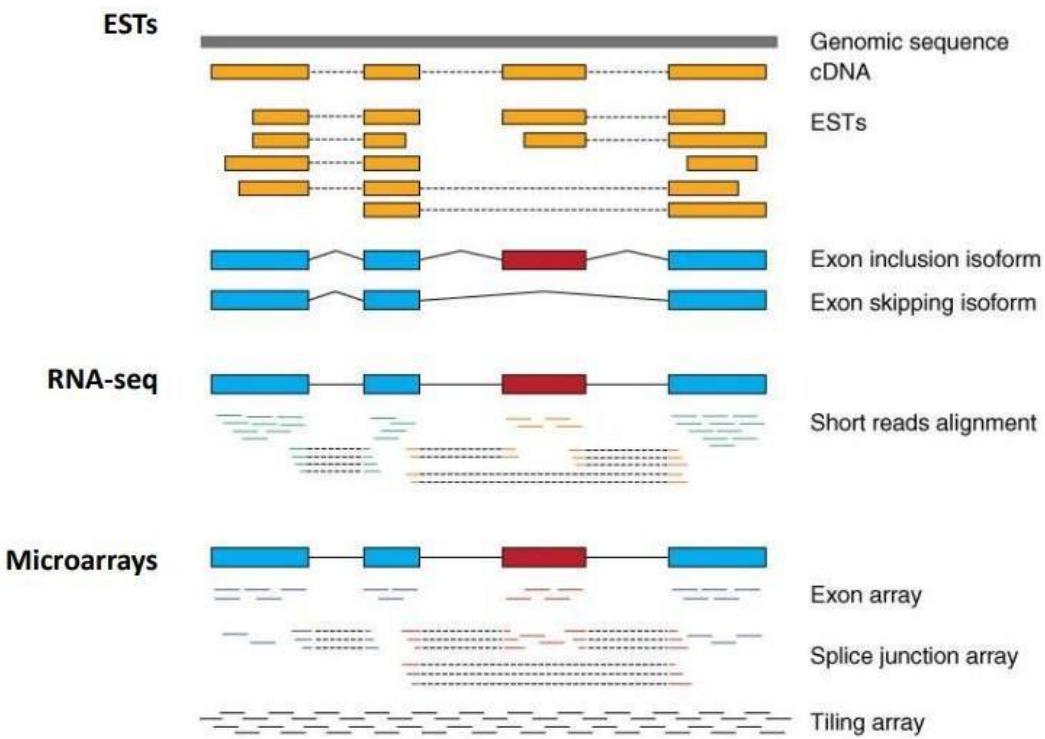


- Inici i final de la transcripció alternatiu.
- Inclusió/ exclusió d'exons.

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*.





Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)
Calle Pelayo, 5. 08001 Barcelona



NEW YORK BURGER
A fuego, but lento

NEW YORK BURGER
A fuego, but lento

Calle Pelayo, 5.
08001 Barcelona



ONE WAY
®

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)

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.

WUOLAH

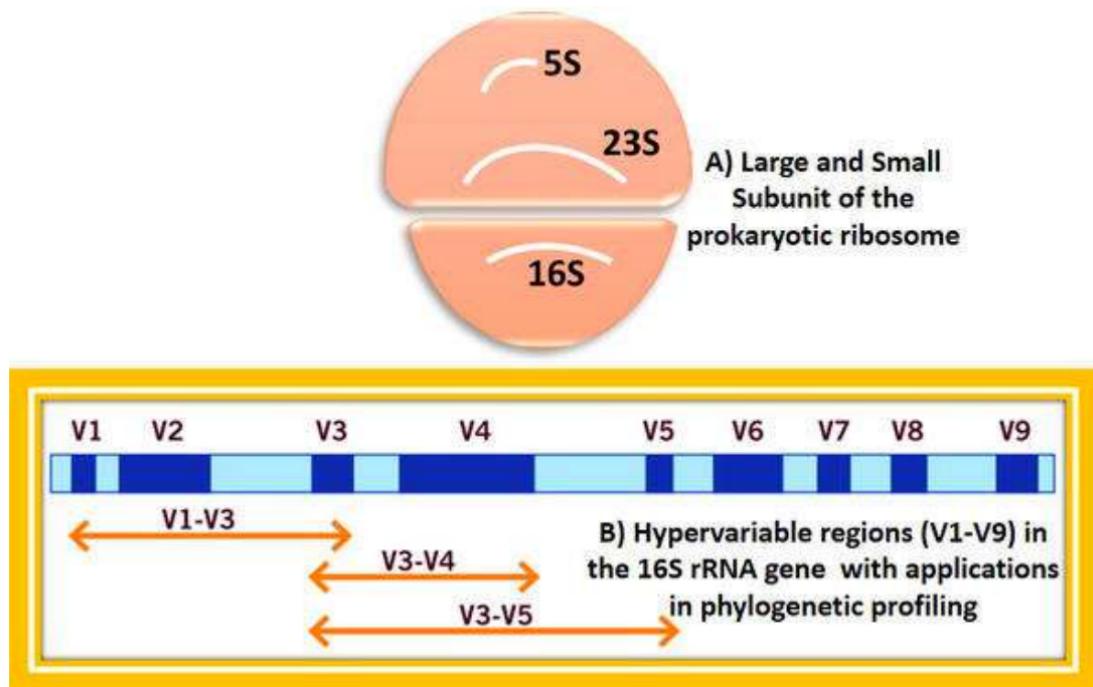
16 S

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).

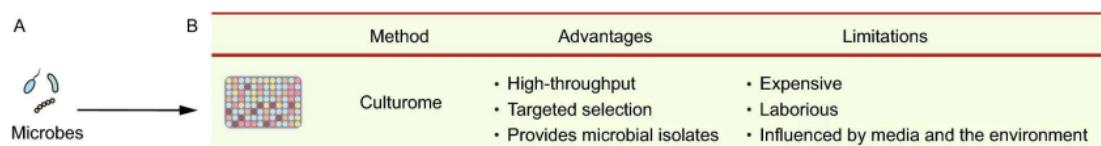


In third generation techniques, we can cover the whole gene and therefore we are reaching a different resolution.

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.



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
<ul style="list-style-type: none">• Taxonomic resolution to species or strain level• Functional potential• Uncultured microbial genome	<ul style="list-style-type: none">• 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
<ul style="list-style-type: none">• Can identify RNA and DNA viruses• Quick diagnosis	<ul style="list-style-type: none">• Most expensive• Difficult to analysis• Severe host-derived contamination

Metatranscriptome

Advantages	Limitations
<ul style="list-style-type: none">• Can identify live microbes• Can evaluate microbial activity• Transcript-level responses	<ul style="list-style-type: none">• 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.

**ONE
WAY**

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)
Calle Pelayo, 5. 08001 Barcelona

¡Únete y recibe una bebida de regalo!



**NEW
YORK
BURGER**
A fuego, but lento

**NEW
YORK
BURGER**
A fuego, but lento

Calle Pelayo, 5.
08001 Barcelona

¡Únete y recibe una bebida de regalo!



**ONE
WAY**

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)

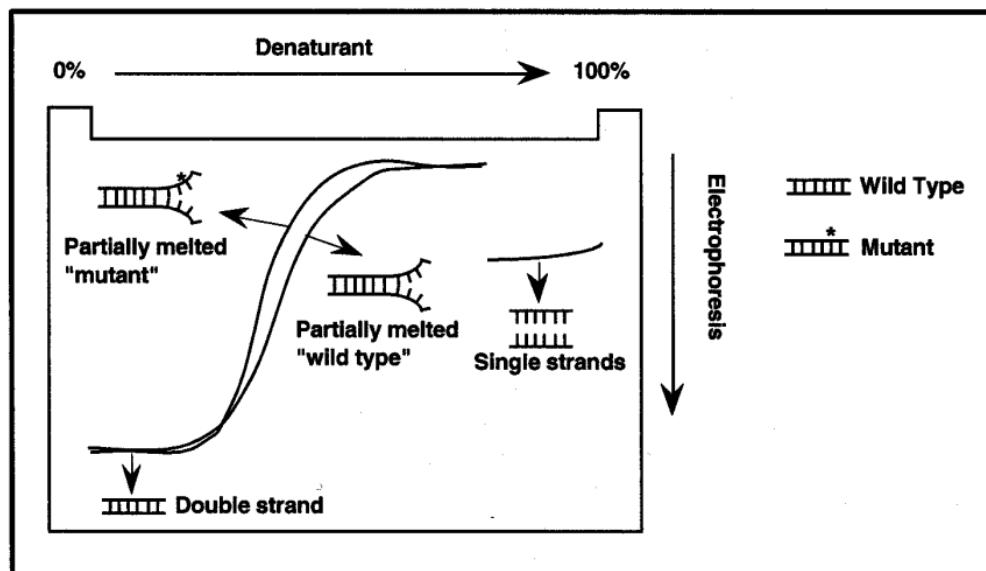
Ecology genomics

- 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 rRNA genes, 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:

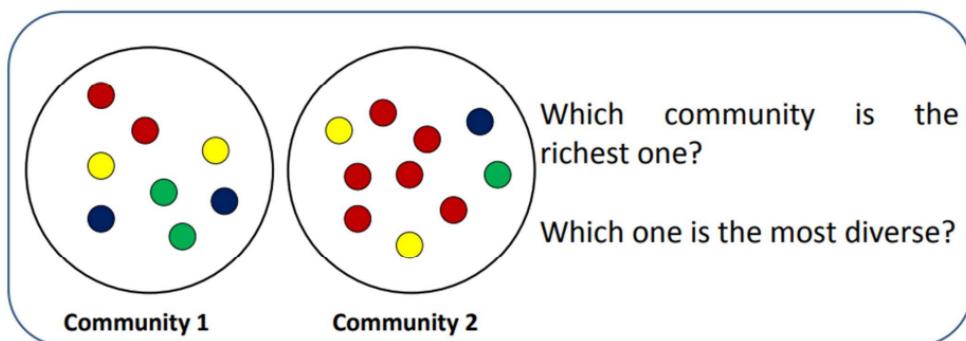
- **Absolute abundance:** Numbers represent real abundance of things being measured.
Example: The actual quantity of a particular gene or organism
- **Relative abundance:** Numbers represent the proportion of things being measured within a sample. In almost all cases microbiome studies are measuring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

**ONE
WAY**

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)
Calle Pelayo, 5. 08001 Barcelona

¡Únete y recibe una bebida de regalo!



**NEW
YORK
BURGER**
A fuego, but lento

**NEW
YORK
BURGER**
A fuego, but lento

Calle Pelayo, 5.
08001 Barcelona

¡Únete y recibe una bebida de regalo!



**ONE
WAY**

Gana premios y descuentos únicos por unirte a nuestro programa One Way ;)

Community 1: More even
Community 2: More rich

Rarefaction

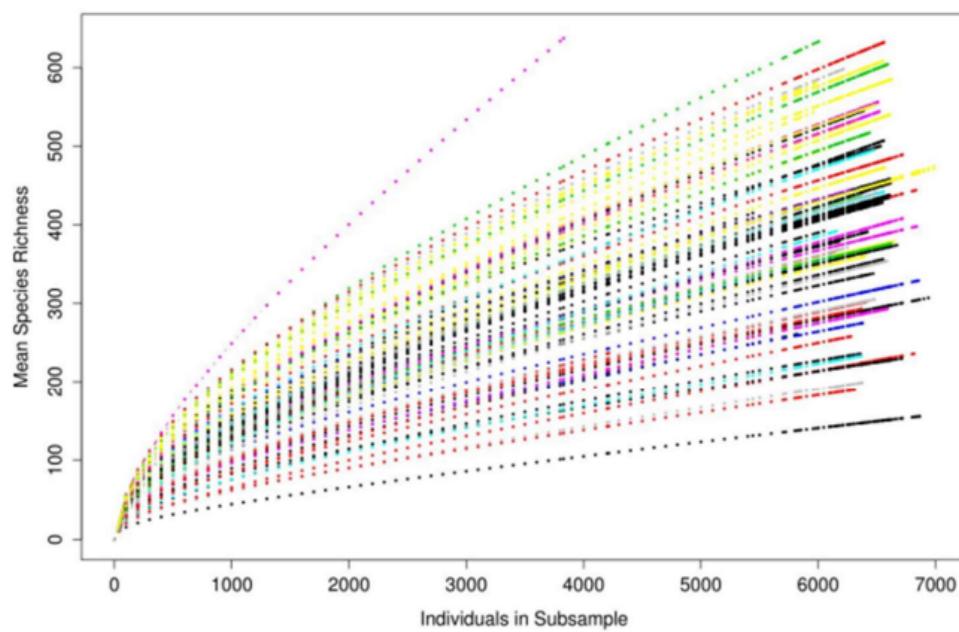
How do we know if we need more sequencing depth or if it is enough? Using the rarefaction curve.

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.

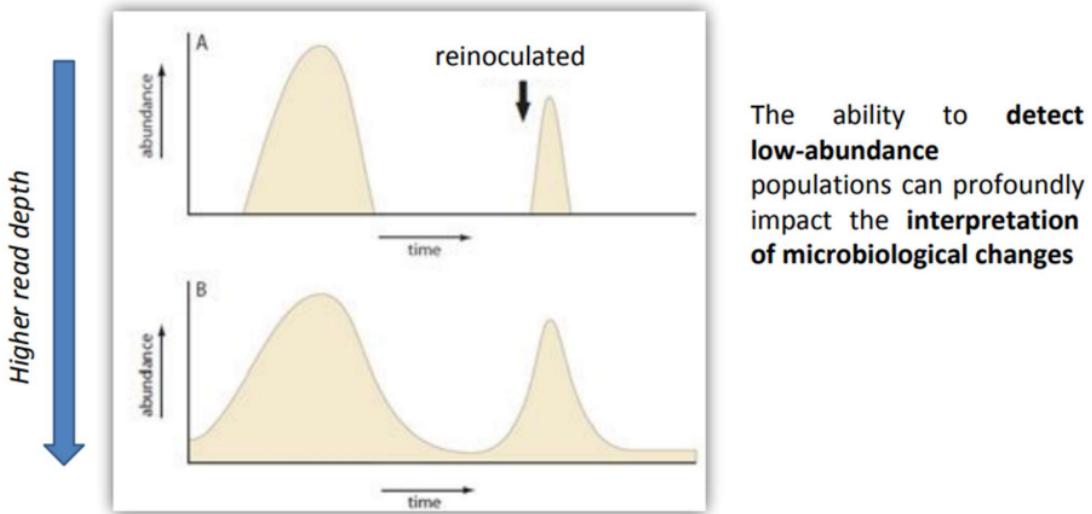
WUOLAH

Technical challenges

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

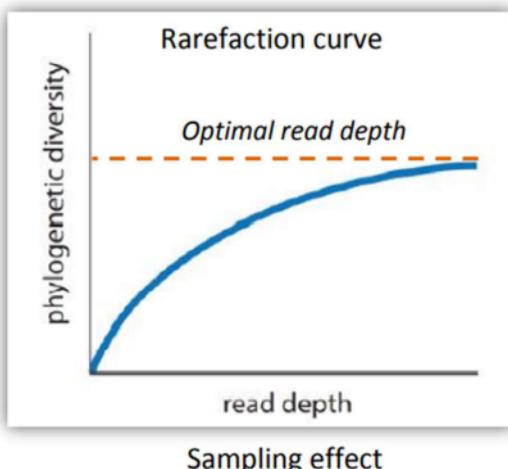
Read depth

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.



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.



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):

$$\text{Redundancy} = (N \cdot R) / G = 132130 \text{ Mb} / 695 \text{ Mb} = 190.12$$

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:

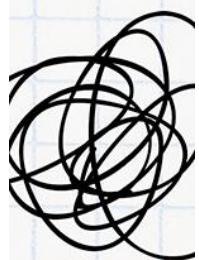
- sequence paired-end reads.**
- eliminate repetitive regions.
- sequence the transcriptome by RNA-seq.
- compare contigs with a database of proteins of a nearby species.

Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

pierdo
espacio



Necesito
concentración

ali ali ooooh
esto con 1 coin me
lo quito yo...

WUOLAH

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 leafy 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.

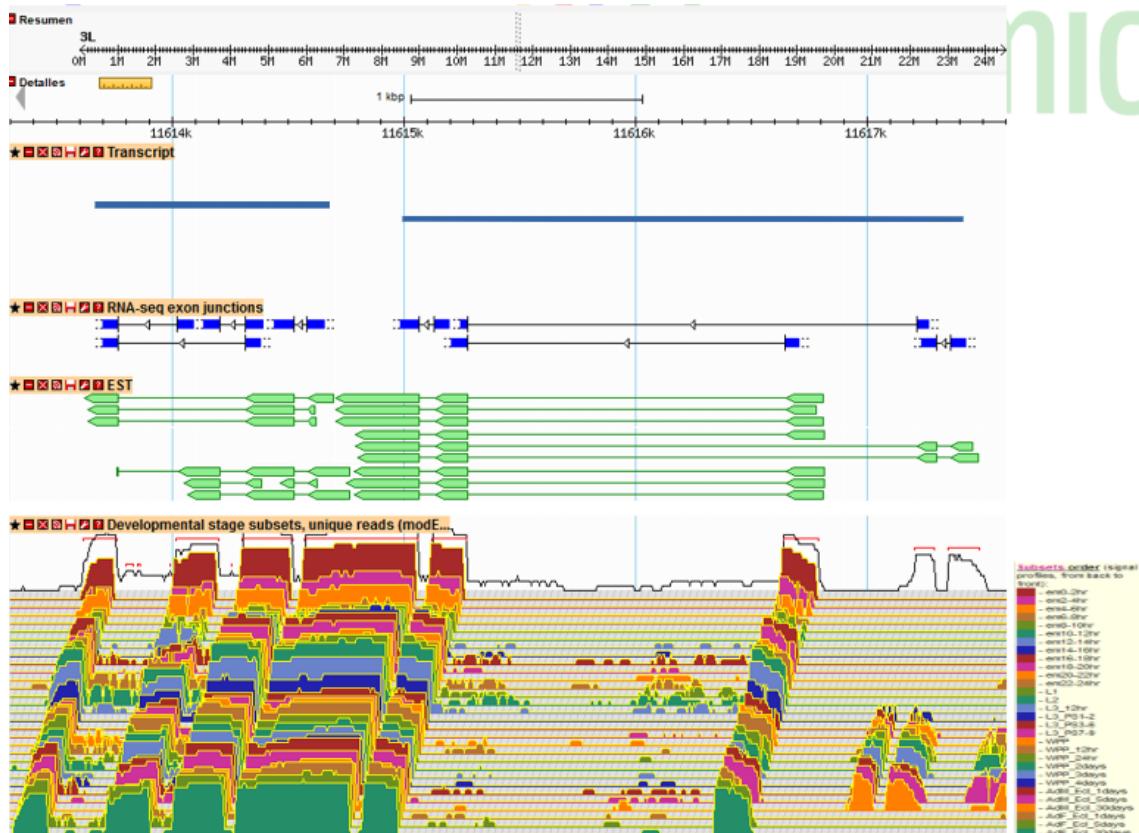
The study of RNAs gives information about:

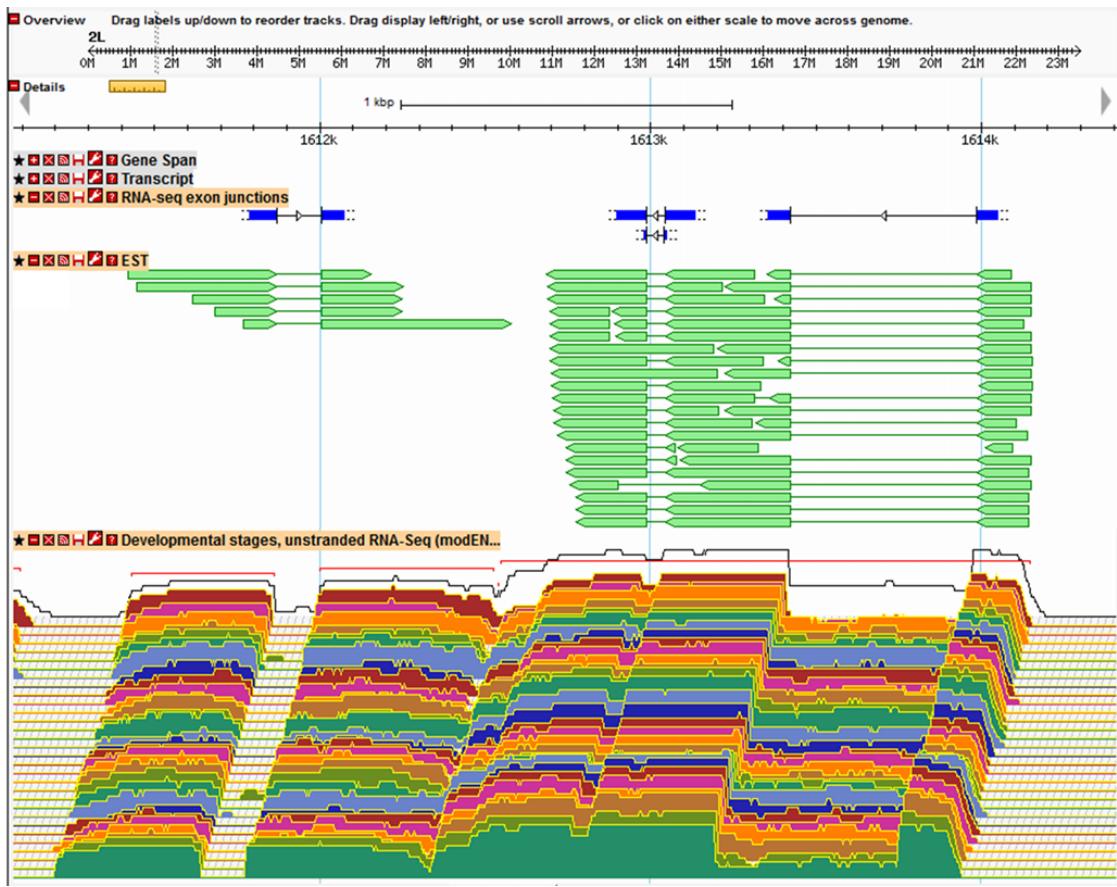
- Genes and other expressed sequences of a genome
- Gene regulation and regulatory sequences
- Function of the genes and their interaction
- Functional differences between tissues and cell types
- Identification of candidate genes for any given process or disease
- Gene expression for a given condition

WUOLAH

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	+	117133 465	117134 193	Possible insertion ~1,2 kb
F1 rev	8	99,4%	7	-	117172 022	117173 660	
F2 fwd	87	96,3%	19	+	218377 76	218385 34	Insertion
F2 rev	182	98,2%	19	-	218683 65	218700 73	
F3 fwd	7	100,0%	X	+	153560 230	153560 952	Inversion
F3 rev	7	98,0%	X	+	153586 670	153587 167	