

# Máster en Bioinformática

## Generación y mantenimiento de datos ómicos

Curso académico 2024-25



**Universidad**  
Internacional  
de Valencia

Dr. Jordi Tronchoni León  
[Jordi.tronchoni@campusviu.es](mailto:Jordi.tronchoni@campusviu.es)

17/04/2024

De:  
 Planeta Formación y Universidades

## Tema 2

# Principales flujos de trabajo en bioinformática

17/04/2024

### **Tema 1. Introducción a la bioinformática**

- 1.1 Historia de la bioinformática
- 1.2 Bioética aplicada al análisis de datos

### **Tema 2. Principales flujos de trabajo en bioinformática**

- 2.1 Genómica
- 2.2 Metagenómica y metataxonómica
- 2.3 Transcriptómica
- 2.4 Proteómica

### **Tema 3. Gestión de entornos y paquetes**

- 3.1 Conda

### **Tema 4. Bases de datos y herramientas bioinformáticas**

- 4.1 Principales bases de datos
- 4.2 Otros recursos online

### **Tema 5. Alineamiento de secuencias**

- 5.1 Introducción al alineamiento de secuencias
- 5.2 Alineamientos Pairwise
- 5.3 Alineamientos Múltiples

### **Tema 6. Métodos de secuenciación**

- 6.1 Primera generación de secuenciadores
- 6.2 Segunda generación de secuenciadores
- 6.3 Tercera generación de secuenciadores
- 6.4 Comparación de plataformas de secuenciación

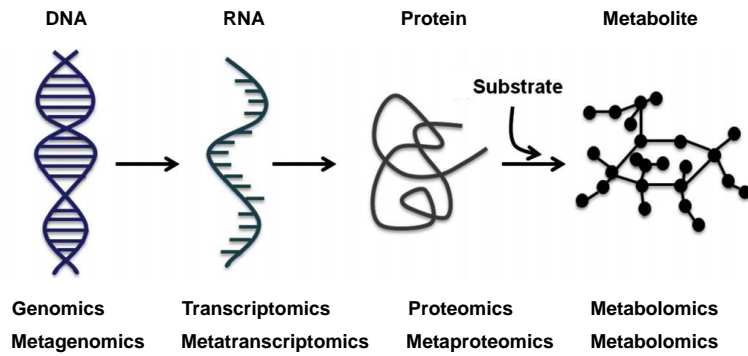
### **Tema 7. Pre-procesado y calidad de secuencias**

- 7.1 Calidad de secuencias
- 7.2 Pre-procesado de secuencias

17/04/2024

Vamos a introducir algunos de los flujos de trabajo más importantes en bioinformática, también conceptos y términos, tipos de formato de archivos y empezar a familiarizarnos con los términos.

## Principales flujos de trabajo



17/04/2024

## Principales flujos de trabajo

### DNaseq

Conjunto de técnicas para conocer la secuencia de nucleótidos de ADN y los análisis derivados, como alineamientos y filogenia.

### RNAseq

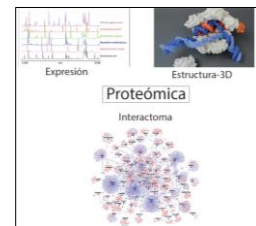
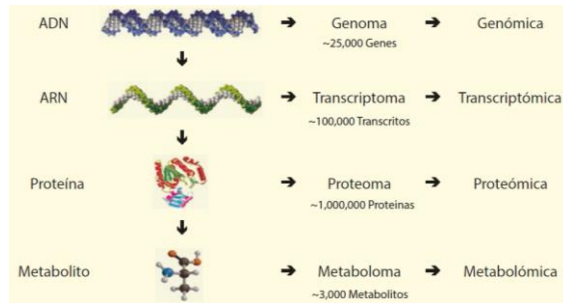
Análisis del transcriptoma (presencia y cantidad de ARN) en la muestra en un momento dado

### Proteómica

Análisis del proteoma (presencia y cantidad de proteínas) en la muestra en un momento dado. Estudio de la estructura proteica, análisis de función proteica.

### Metabolómica

Análisis de los metabolitos, también diferencialmente, (presencia y cantidad de metabolitos) en la muestra en un momento dado.



Existen distintas disciplinas dentro de la bioinformática que estudian los distintos aspectos de la biología molecular. El DNaseq o secuenciación de ADN estudia los genomas de especies o poblaciones complejas del mundo microbiano o eucariótico. A partir de la secuencia de ADN podemos realizar multitud de análisis como la caracterización de los genes por comparación respecto a una base de datos o estudios evolutivos mediante estudios filogenéticos. Mediante RNAseq podemos estudiar las secuencias de ADN codificante derivado de ARN mensajero, con el objetivo de saber qué genes se están expresando y con qué proporción en una muestra de interés y en un momento concreto. A partir de ARN produciremos proteínas, las cuales se estudian dentro de la proteómica. Aquí hay que tener en cuenta que existen dos ramas dentro de la proteómica: **la proteómica diferencial y la estructural**. La proteómica diferencial o de expresión diferencial, busca al igual que la transcriptómica, comparar muestras y evaluar su perfil, en este caso proteico, incidiendo en las diferencias. La estructural, conocer la estructura de las proteínas (terciaria y cuaternaria especialmente) para arrojar luz sobre su posible función. La metabolómica (estudio de metabolitos derivados de las reacciones que se tienen en los sistemas biológicos).

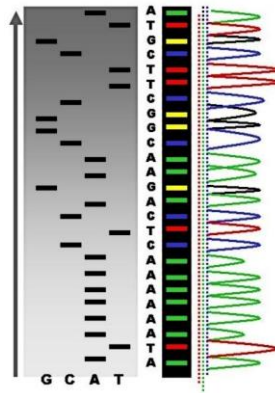


# DNAseq: Secuenciación Genómica

## DNaseq: Secuenciación genómica

Se conoce como DNaseq a la secuenciación de ADN, al conjunto de técnicas y métodos de laboratorio para determinar la secuencia exacta de bases (A, C, G y T) en una molécula de ADN.

- Genómica
- Metagenómica
- Metataxonómica



17/04/2024

La secuenciación de ADN se basa en la lectura de los genomas de los organismos de un sistema, a nivel de organismo único (genómica) o comunidad de organismos (metagenómica y metataxonómica), dando lugar a un archivo con la lectura de la secuencia en la forma de su secuencia de bases nitrogenadas. Dentro del DNaseq podríamos hablar de distinto tipo de especializaciones, una posible diferenciación sería la genómica, metagenómica y metataxonómica.

## Historia

- 1975: Frederick Sanger desarrolla el método de terminación de cadena para la secuenciación del ADN
- 1977: Walter Gilbert y Allan Maxam desarrollan el método químico para la secuenciación del ADN
- 1986: Se establece el primer laboratorio comercial de secuenciación de ADN
- 1990: Se lanza el Proyecto Genoma Humano
- 1995: Se publica la secuencia completa del primer genoma bacteriano, *Haemophilus influenzae*
- 2000: Se publica la secuencia "completa" del genoma humano

17/04/2024

La genómica engloba el conjunto de técnicas para la obtención de la secuencia de genomas individuales. Los comienzos de la secuenciación de ADN se dieron con la metodología de Sanger en 1975, lo que dio lugar a la secuencia del fago PhiX174 y la instauración de la primera tecnología de secuenciación, la cual aún se usa a día de hoy (aunque mejorada y actualizada). Desde este momento el campo empezó a evolucionar y se dieron otras técnicas de secuenciación mediante PCR multiplex o el uso de transposones en vectores, aunque su gran potencial fue durante el desarrollo del proyecto más ambicioso de finales del siglo XX, la secuenciación del genoma humano. En este proyecto se aunaron varios grupos de investigación para paralelizar el coste y el esfuerzo para lograrlo. Se evolucionó en la técnica, la tecnología de secuenciación del ADN se hizo más rápida y menos costosa como parte del proyecto genoma humano.

## Ejemplos de aplicaciones

Detección de Enfermedades Genéticas: evaluar el riesgo de enfermedades genéticas de un individuo.

Farmacogenómica: la composición genética de un individuo afecta su respuesta a los medicamentos.

Biología Evolutiva: comprender la historia evolutiva y las adaptaciones genéticas.

Ciencia Forense: identificar sospechosos, determinar la paternidad y analizar pruebas.

Metagenómica: estudiar comunidades microbianas y sus funciones en un ambiente concreto.

Agricultura y ganadería: mejora de cultivos, resistencia a enfermedades, programas de cría.

17/04/2024

**Detección de Enfermedades Genéticas:** La secuenciación genómica ayuda a evaluar el riesgo de enfermedades genéticas al analizar el genoma de un individuo. Facilita la detección temprana y el tratamiento personalizado.

**Farmacogenómica:** Este campo estudia cómo la composición genética de un individuo afecta su respuesta a los medicamentos, permitiendo la medicina personalizada.

**Biología Evolutiva:** La secuenciación genómica proporciona información sobre las relaciones evolutivas entre las especies. Comparar genomas ayuda a rastrear la historia evolutiva y comprender las adaptaciones genéticas.

**Ciencia Forense:** En investigaciones criminales, la secuenciación genómica puede identificar sospechosos, determinar la paternidad y analizar pruebas biológicas (como muestras de sangre o cabello).

**Metagenómica:** La metagenómica implica secuenciar ADN de muestras ambientales (por ejemplo, suelo, agua o microbiota intestinal). Ayuda a estudiar comunidades microbianas y sus funciones.

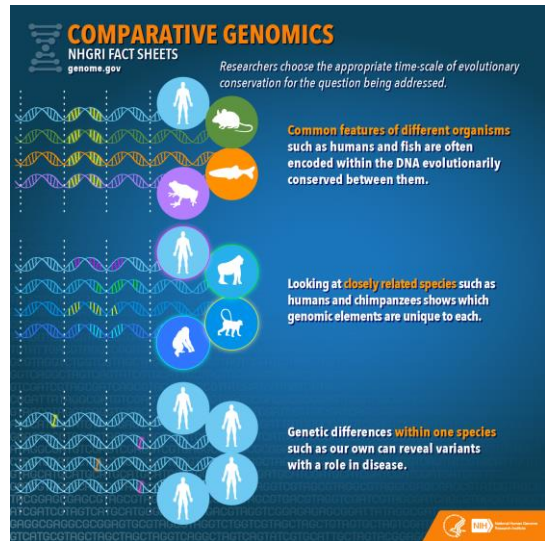
Agricultura y ganadería: La secuenciación genómica ayuda en la mejora de cultivos, la resistencia a enfermedades y los programas de cría. Permite identificar rasgos beneficiosos y optimizar prácticas agrícolas.

## Genómica comparada

Muchos de los ejemplos anteriores se realizaron mediante la comparación de las características genómicas de diferentes organismos. Estas características pueden incluir:

- la secuencia de ADN,
- los genes,
- el orden de los genes,
- las secuencias reguladoras
- y otros puntos de referencia estructurales del genoma.

Según la pregunta que queramos contestar, compararemos a nivel de individuo (adaptación evolutiva), de especie, especies altamente relacionadas o especies alejadas filogenéticamente.



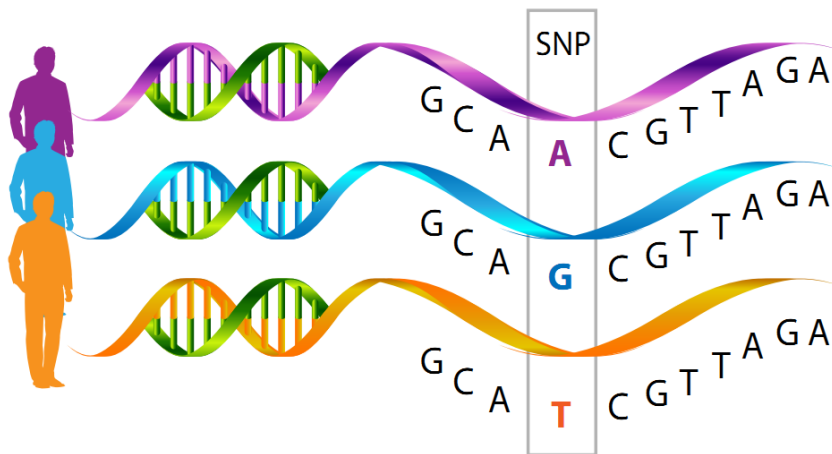
## Genómica comparada



De gran aplicabilidad, la genómica comparada busca polimorfismos, variantes génicas (SNPs e InDels), reordenamientos cromosómicos, variación en el número de cromosomas, etc en un determinado individuo.

17/04/2024

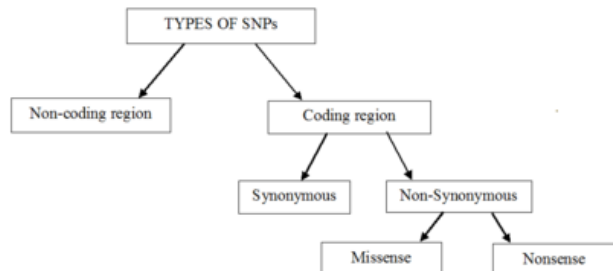
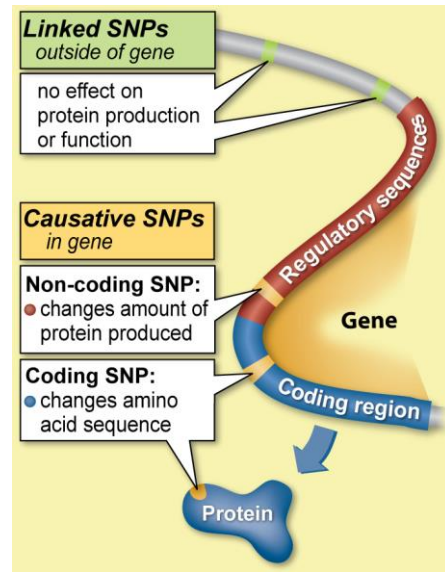
## Recordemos: SNPs, InDels & CNVs



17/04/2024

SNP significa Polimorfismo de Nucleótido Único. Es un tipo común de variación genética que ocurre cuando un solo nucleótido (A, C, G o T) en la secuencia de ADN es diferente entre individuos dentro de una población. Los SNP pueden ocurrir tanto en regiones codificantes como no codificantes del genoma y pueden afectar a una amplia gama de procesos biológicos, incluyendo la expresión génica, la estructura de proteínas y la susceptibilidad a enfermedades. Los SNP se utilizan ampliamente en la investigación genética, como en los estudios de asociación del genoma completo (GWAS), para identificar factores genéticos asociados con diversas enfermedades y rasgos.

## SNPs



17/04/2024

Existen varios tipos de SNPs, pero los principales son:

**Codificantes:** ocurren en una región que codifica para un gen:

SNPs sinónimos: son aquellos que no cambian el aminoácido codificado por el genoma, por lo que no afectan la función de la proteína producida.

SNPs no sinónimos: son aquellos que cambian el aminoácido codificado por el genoma, lo que puede afectar la función de la proteína producida:

Los SNPs sin sentido (missense) son un tipo de SNP no sinónimo que produce un cambio en la secuencia de aminoácidos de una proteína.

Los SNPs sin sentido (nonsense) son un tipo de SNP no sinónimo que produce un codón de parada prematuro en la secuencia de ADN de un gen, lo que resulta en una proteína truncada o no funcional.

**No codificantes:**

SNPs intrónicos: son aquellos que ocurren en las regiones no codificantes del genoma, llamadas intrones, que pueden afectar la regulación de la expresión génica.

SNPs intergénicos: son aquellos que ocurren en las regiones entre los genes, que pueden afectar la regulación de la expresión génica de los genes cercanos.

**Indel examples**

wild-type sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

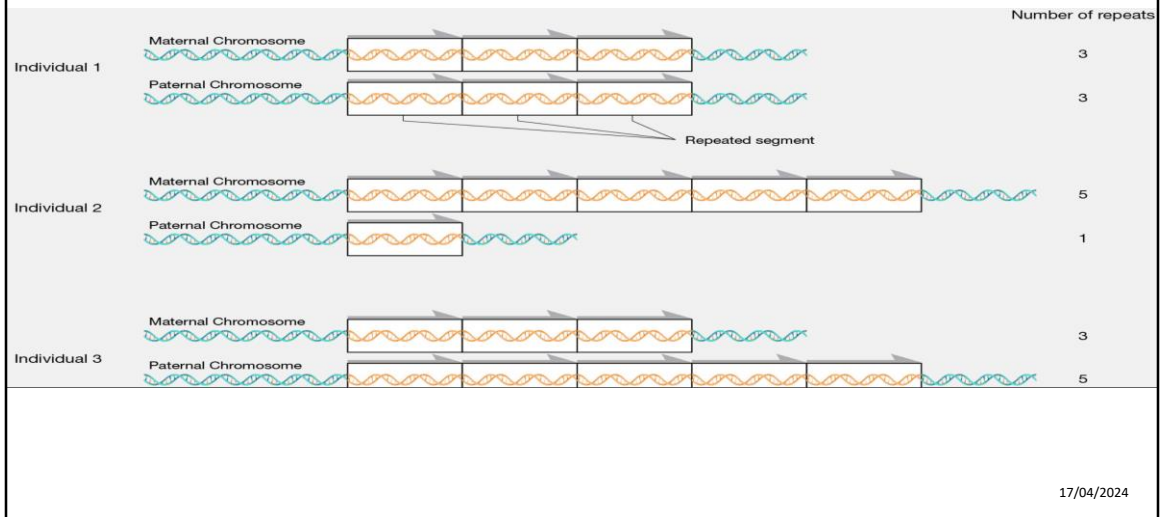
ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

17/04/2024

Un indel es una mutación genética que se produce cuando se insertan o eliminan uno o más nucleótidos en una secuencia de ADN. El término “indel” es una contracción de “inserción o delección”. Los indels pueden compararse con una mutación puntual.

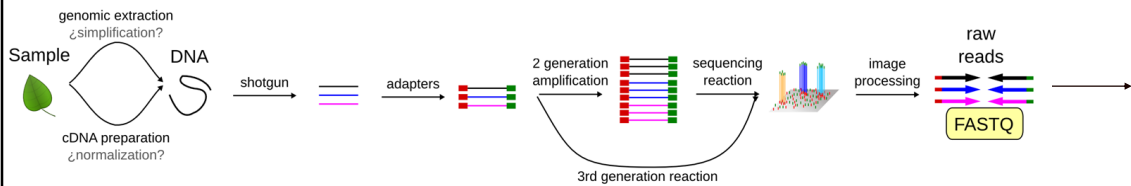


17/04/2024

Un CNV (del inglés, Copy Number Variation) se refiere a una circunstancia en la que el número de copias de un segmento específico de ADN varía entre los genomas de distintos individuos. Las variantes individuales pueden ser cortas o incluir miles de bases. Estas diferencias estructurales pueden haberse producido por duplicaciones, deleciones u otros cambios y pueden afectar a largos tramos de ADN. Tales regiones pueden contener o no un gen o genes. Los cambios en el número de copias de pequeñas secciones de nuestros genomas pueden tener grandes consecuencias. Un ejemplo interesante es un gen llamado amilasa. Este gen es importante para digerir alimentos ricos en almidón como las patatas o los cereales. Los científicos descubrieron variaciones en el número de copias del gen de la amilasa en distintos grupos de personas. Las personas de lugares en los que los alimentos ricos en almidón eran históricamente muy importantes suelen tener más copias del gen de la amilasa que las personas de entornos en los que los alimentos ricos en almidón eran menos comunes. Lugares donde la carne y el pescado eran partes más importantes de la dieta en relación con los almidones. Esto sugiere que los pueblos antiguos que tenían más copias del gen de la amilasa eran capaces de obtener más energía de los alimentos ricos en almidón, y por lo tanto eran capaces de prosperar en regiones donde las fuentes de alimentos ricos en almidón eran importantes. Y tal vez superar a los habitantes de esas regiones que tenían menos copias del gen de la amilasa. Esta

es una de las muchas formas en que la variación del número de copias puede tener efectos importantes. (desde el **Talking Glossary of Genomic and Genetic Terms:**[Copy Number Variation \(CNV\) \(genome.gov\)](#))

## ¿Como se realiza la secuenciación genómica?



Preparación de la muestra - Fragmentación del ADN - Ligación de adaptadores y preparación de la biblioteca - Enriquecimiento mediante PCR - Secuenciación

17/04/2024

Ejemplo de posible flujo de trabajo, en amarillo los distintos formatos de archivo que podemos encontrar.

**Preparación de la muestra:** Comienza extrayendo ADN puro y de alto rendimiento de la muestra. La calidad de la muestra es crucial para obtener resultados precisos.

**Fragmentación del ADN:** Para secuenciar un genoma completo o una secuencia más grande, es importante fragmentar el ADN. Los fragmentos más pequeños son más manejables y se pueden secuenciar con mayor precisión.

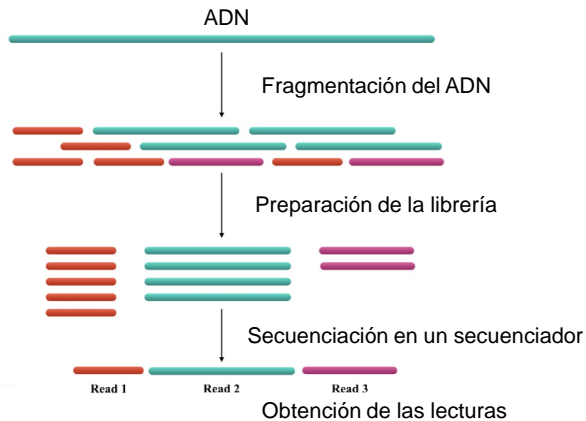
**Ligación de adaptadores y preparación de la biblioteca:** Se agregan adaptadores a los extremos de los fragmentos de ADN. Estos adaptadores permiten la unión a una superficie sólida para la secuenciación.

**Enriquecimiento de la biblioteca mediante amplificación (2ª generación):** La biblioteca de fragmentos de ADN se amplifica mediante PCR (reacción en cadena de la polimerasa). Esto crea múltiples copias de los fragmentos, lo que facilita la detección.

**Secuenciación del ADN:** Se lleva a cabo con distintos tipos de secuenciadores. Obtención de las **lecturas (reads)** del secuenciador, generalmente en un archivo **fastq**, conocido como datos crudos o RAW data.

## Read o lectura

Una lectura se refiere a la secuencia inferida de pares de bases (o probabilidades de pares de bases) correspondientes a la totalidad o a parte de un único fragmento de ADN. Esta secuencia es inferida a través de los distintos métodos de secuenciación, generalmente mediante el uso de secuenciadores.



17/04/2024

Una lectura se refiere a la secuencia inferida de pares de bases (o probabilidades de pares de bases) correspondientes a la totalidad o a parte de un único fragmento de ADN. Esta secuencia es inferida a través de los distintos métodos de secuenciación, generalmente mediante el uso de secuenciadores.

## Archivo fastq

1. La primera línea siempre comienza con @ y contiene información sobre la descripción de la secuencia y una identificación única.
2. La segunda línea contiene la secuencia sin procesar.
3. Por lo general, la tercera línea solo contiene el signo «+» o el signo «+» y una repetición de la identificación de la secuencia.
4. La cuarta línea contiene los valores de calidad y debe contener el mismo número de caracteres que la línea de secuencia. Se codifican código ASCII dependiendo de la tecnología de secuenciación

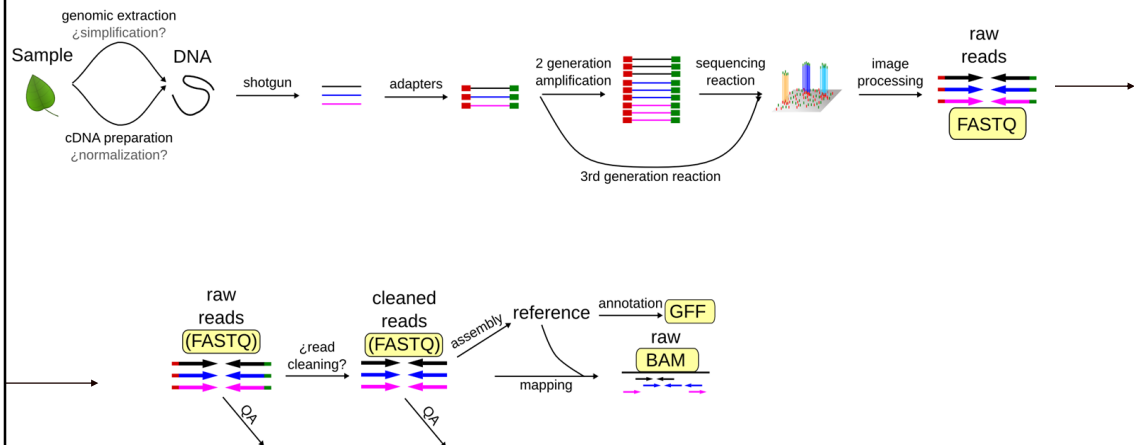
```
@SRR6043417.1 1 length=125
CCGGGGGGCGGAGACGGGGAGGAGGACGGACGGACGGACGGGGCCCCCGAGCCACCTCCCCCGGGCCTCCAGCCGTCCTCCGGAGCCGGTCGGGCGCACCCGCCGCGTGGA
+SRR6043417.1 1 length=125
/ <BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SRR6043417.2 2 length=125
GGACCTTGAGAGCTTGTGAGGTTCTAGCAGGGGAGCGCAGCTACTCGTATACCTTGACCAAGACCGTCTCTCTATCGGGATGGTCGTCTTCGACCGAGCGCGCAGCTTCGGGA
+SRR6043417.2 2 length=125
/ <BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

17/04/2024

La estructura que observamos en los archivos .fastq es la que se observa en la imagen. Para cada una de las lecturas tenemos:

- una cabecera o ID única
- seguida de la lectura de la secuencia
- en la tercera línea tenemos una línea que contiene el símbolo "+" o el símbolo y una repetición del identificador
- por último la cuarta línea contiene la calidad, en formato de caracteres ASCII de forma individual para cada una de las bases leídas. Es importante tener en cuenta que esta codificación de la calidad es dependiente del método de secuenciación.

## ¿Cómo se realiza la secuenciación genómica?



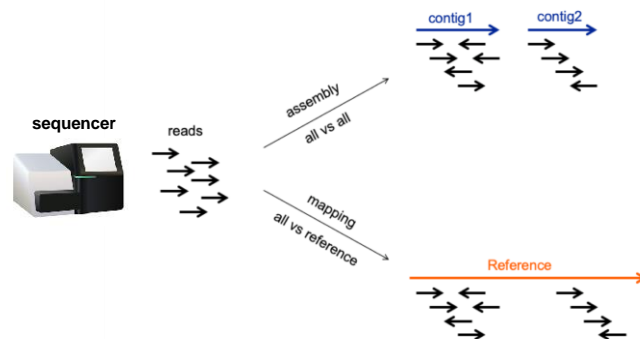
17/04/2024

**Análisis de datos y secuencia:** Los datos de secuenciación se recopilan y se analizan para obtener la secuencia completa del ADN mediante una serie de programas bioinformáticos:

**Limpieza y calidad de lecturas:** Las secuencias de ADN generadas por la secuenciación son fragmentos cortos llamados “lecturas” o “reads”. El primer paso será comprobar su calidad y “limpiarlas”.

**Ensamblaje de lecturas o mapeado de lecturas en función de la estrategia de secuenciación:** El ensamblaje combina estas lecturas para reconstruir la secuencia completa del genoma. Se utilizan algoritmos para superponer y ensamblar las lecturas en una secuencia coherente. El mapeado de lecturas utiliza un genoma de referencia conocido para situar las lecturas sobre este genoma.

## Dos estrategias de secuenciación: mapeado de novo vs contra referencia o re-secuenciación



17/04/2024

Existen dos estrategias de secuenciación, el mapeado de *novo* y el mapeado utilizando un genoma de referencia, llamado comúnmente, re-secuenciación.

El **mapeado utilizando un genoma de referencia** es un enfoque en el que las secuencias de ADN de una muestra se comparan con un genoma de referencia previamente conocido. En este enfoque, se utilizan algoritmos para alinear las secuencias de ADN de la muestra con las secuencias correspondientes en el genoma de referencia. Este enfoque es comúnmente utilizado para identificar SNPs y otras variantes genéticas comunes en la población.

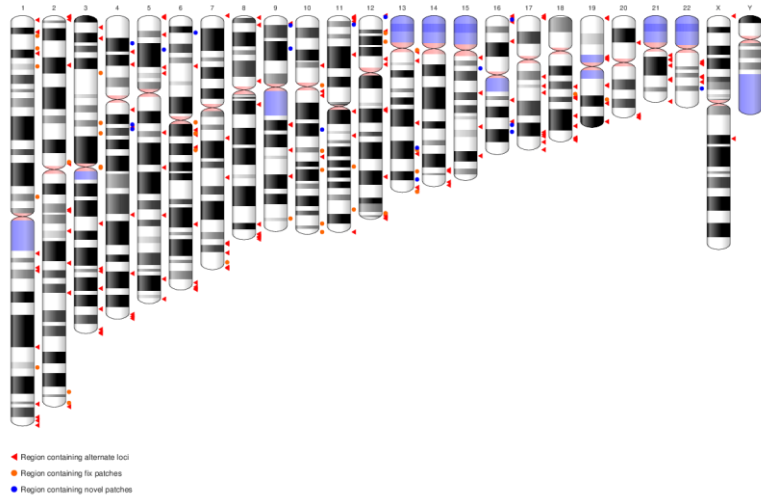
El éxito de la técnica dependerá en gran medida de la calidad del **genoma de referencia**. En humanos y organismos modelo tenemos genomas de referencia de gran calidad.

Por otro lado, el **mapeado de novo** es una técnica que se utiliza para identificar y mapear secuencias de ADN previamente desconocidas en un genoma sin utilizar un genoma de referencia previo. En este enfoque, se utilizan algoritmos para ensamblar secuencias de ADN a partir de lecturas individuales de ADN sin un genoma de referencia previo. Este enfoque es útil para identificar variantes genéticas que no

están presentes en el genoma de referencia utilizado en el mapeado de lecturas cortas, como variantes estructurales complejas y mutaciones raras asociadas con enfermedades genéticas. Además, el mapeado de novo también puede utilizarse para ensamblar el genoma completo de una especie desconocida.

Ambas técnicas tienen sus ventajas y desventajas y se utilizan dependiendo del objetivo específico de la secuenciación de ADN. El mapeado utilizando un genoma de referencia es más rápido y menos computacionalmente intensivo, pero puede perder información sobre variantes genéticas que no están presentes en el genoma de referencia. Por otro lado, el mapeado de *novo* es más preciso para la identificación de variantes genéticas no comunes y desconocidas, pero es más computacionalmente intensivo y requiere mayores cantidades de datos de secuenciación de alta calidad.

## Genoma de referencia



17/04/2024

Un **genoma de referencia** es una secuencia de ADN que se utiliza como punto de comparación para analizar y comprender los genomas de individuos de la misma especie. Un genoma de referencia se crea mediante la secuenciación del ADN de un número limitado de individuos de una población y la posterior combinación de esas secuencias para crear una representación del genoma de la especie en cuestión. Los genomas de referencia se utilizan ampliamente en la investigación genómica para identificar variantes genéticas en los genomas de individuos de la misma especie. Esto es posible gracias a que los genomas de referencia son secuencias completas y detalladas de una especie, que permiten comparar las secuencias de ADN de otros individuos con ellas. Además, los genomas de referencia también se utilizan para la identificación y anotación de genes y otras regiones funcionales del ADN. Es importante tener en cuenta que un genoma de referencia no representa necesariamente a toda la población de una especie, ya que se basa en una muestra limitada de individuos. Por lo tanto, pueden existir variaciones en los genomas de individuos que no están representados en el genoma de referencia utilizado. Además, los genomas de referencia pueden contener errores, especialmente en regiones complejas o repetitivas del ADN, lo que puede afectar la precisión de los análisis genómicos.

Inicialmente los genomas de referencia correspondían a un único individuo, por ejemplo, el genoma de referencia de la levadura *Saccharomyces cerevisiae* ha sido durante muchos años (y aun hoy es “uno de los genomas de referencia”) la cepa de laboratorio S288c. Esta cepa, llamada de laboratorio porque había sido seleccionada artificialmente en busca de ciertos fenotipos útiles en el laboratorio resultó estar genómicamente muy alejada del resto de *S. cerevisiae* y a medida que se han ido secuenciando más y más cepas, ha quedado patente el gran sesgo con el que estábamos observando esta especie. De ahí la importancia de captar la diversidad genómica de una especie.

Un **buen genoma de referencia** es aquel que es lo más completo y preciso posible, y que representa adecuadamente la variabilidad genética de la especie en cuestión. Un genoma de referencia de alta calidad debe tener una alta cobertura, lo que significa que se han secuenciado y ensamblado la mayor parte del genoma. También debe tener una baja tasa de errores, lo que significa que la secuencia del genoma es precisa y está libre de artefactos de secuenciación.

Además, un buen genoma de referencia debe ser representativo de la variabilidad genética de la especie. Esto significa que se deben incluir secuencias de múltiples individuos de diferentes poblaciones para abarcar la variabilidad de la especie. Un genoma de referencia que representa adecuadamente la variabilidad genética puede ser utilizado para estudiar la diversidad genética, las adaptaciones evolutivas y la estructura poblacional de la especie.

Por último, es importante que el genoma de referencia esté correctamente anotado con información funcional, como la ubicación y función de los genes y otras regiones del ADN. La anotación adecuada del genoma de referencia permite la identificación y el estudio de genes específicos y otras regiones funcionales del ADN.

En resumen, un buen genoma de referencia es aquel que es lo más completo, preciso y representativo posible de la especie en cuestión, y que está correctamente anotado con información funcional.

Los genomas de referencia están continuamente mejorándose. Si tomamos como referencia el organismo modelo *Saccharomyces cerevisiae*, vemos que la última *assembly* liberada es la R64. Es decir que el genoma de referencia corresponde a la versión 64. Recientemente el genoma de referencia humano dio un gran paso con la liberación del genoma de referencia del consorcio *telomer to telomer* que proporcionó la versión más completa de este genoma (y finalmente cerrada) tras 20 años.

## Genoma de referencia

Amazon WorkSpaces

Amazon WorkSpaces View Settings Support

Terminal

Archivo Editar Ver Buscar Terminal Ayuda

```
AATAATACGGTAGTGGCTCAAACCTCATGCGGGTCTATGATACAATTATATCTTATTTCC
ATTCCCATATGCTAACCGCAATATCCTAAAAGCATAAGTGATGATCTTTAATCTTGAT
GTGACACTACTCATACGAAGGGACTATATCTAGTCAAGACGATACTGTGATAGGTACGTT
(base) [UNIVERSIDADVIU\jtronchoni2@a-2ucwa72qat2w1 S288C_reference_genome_R64-4-1_20230830]$ head -n 100 S288C_reference_sequence_R64-4-1_20230823.fsa
>ref[NC_001133] [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=I]
CAGACCAACAGCCACACCCGACACCAACCAACCAACCAACCAACCAACCAACCAACCAACCA
CATCTTAACACTACCTTAACACAGGCCAATCTAACCTGGGCAACCTGTCTCTCAACTT
ACCCCTCATTACCTGGCTCTCACTCGTTACCTGTCTCCATTCAACGATACCACTCCGAAC
CAACCATCCATCCCTCTACTTACTACCACTCACCCAGCTTACCCCTCAATTACCCATATC
CAACCCACTGCGCACTTACCCTACCACTTACCCTACCACTGACCTGACCTACTCACCATAC
TGTTCTTCTACCCACCATATTGAACGCTAACAAATGATCGTAAATACACACACGCTGCT
TACCTTACCACTTTATACCAACCAACATGACCACTACCCCTCACTTGTATCTGATTT
TACGTACGACACGAGGTACAGTATATACCATCTCAAACTTACCTACTCTCAGATTCT
CACTTCACTCCATGGCCATCTCTCACTGAATCAGTACCAATGCACTCACATCATTATG
CACGGCACTTGCTCAGCGGTCTATACCTGTGCTTATACCCATAAGCCCATCATTAT
CCACATTTTGATATCTATATCTCTTGGCGGTCCCAATATTGTATAACTGCCCTTAAT
ACATACGTTATACCACTTTTGACCATATATCTACCACTCCATTATATACACTTATGTC
ATATTACAGAAATCGCCACAAATCAGCTAAACATAAATATTCTACTTTTCAG
AATAATACATAAACATATTGGCTTGTGGTAGCAACATATCATGGTATCACTAACGTAA
AGTCTCTCAATATTGCAATTGCTTGAACGATGCTATTTCAGAAATATTCTGACTTACA
CAGGCCATACATTAGAATAATATGTCACATCACTGTCTGAACACTCTTTATCACCGAGC
AATAATACGGTAGTGGCTCAAACCTCATGCGGGTCTATGATACAATTATCTTATTTCC
ATTCCCATATGCTAACCGCAATATCCTAAAAGCATAAGTGATGATCTTTAATCTTGAT
GTGACACTACTCATACGAAGGGACTATATCTAGTCAAGACGATAGTGATAGGTACGTT
ATTTAATAGGATCTATAACGAAATGTCAAATATTTACGGTAATAATACTATACGCGG
CATACTAATAAGCGGTTCGATTTGCTGATGATGATGATGATGATGATGATGATGATGATGAT
```

17/04/2024

Los genomas de referencia los encontraremos en texto plano, en un archivo de texto generalmente denominado .fasta, pero también, .fsa; .faa.

### Archivo FASTA del genoma de referencia:

El archivo FASTA contiene la secuencia de bases de nucleótidos del genoma de referencia.

Cada línea del archivo representa una región del genoma, con una etiqueta ">" seguida de la descripción y la secuencia de bases.

Por ejemplo, el archivo FASTA para el cromosoma 6 humano se vería así:

```
>chr6
ATCG...
```

Ejemplos de la secuencia de S288c de *Saccharomyces cerevisiae*:



viu  
Universidad  
Internacional  
de Valencia

17/04/2024

Como este .gff. Ejemplos de la secuencia de S288c de *Saccharomyces cerevisiae*: [sgd-archive.yeastgenome.org](http://sgd-archive.yeastgenome.org)

## Genoma de referencia

Amazon WorkSpaces

Amazon WorkSpaces View Settings Support

Archivo Editar Ver Buscar Terminal Ayuda

```
rw-r--r-- 1 UNIVERSIDADVIU\tranchoni2 UNIVERSIDADVIU\domain users 5328889 ago 30 21:00 saccharomyces_cerevisiae_R64-4-1_20230830.gff.gz
(base) [UNIVERSIDADVIU\tranchoni2@a-2wcwa72qat2wi 5288C_reference_genome_R64-4-1_20230830]$ head -n 100 5288C_reference_sequence_R64-4-1_20230830.gff
head: no se puede abrir «5288C_reference_sequence_R64-4-1_20230830.gff» para lectura: No such file or directory
(base) [UNIVERSIDADVIU\tranchoni2@a-2wcwa72qat2wi 5288C_reference_genome_R64-4-1_20230830]$ head -n 100 saccharomyces_cerevisiae_R64-4-1_20230830.gff
#gff-version 3
#date-produced 2023-08-30 11:45:31
#data-source SGD
#assembly R64-3-1
#refseq-version GCF_000146045.2

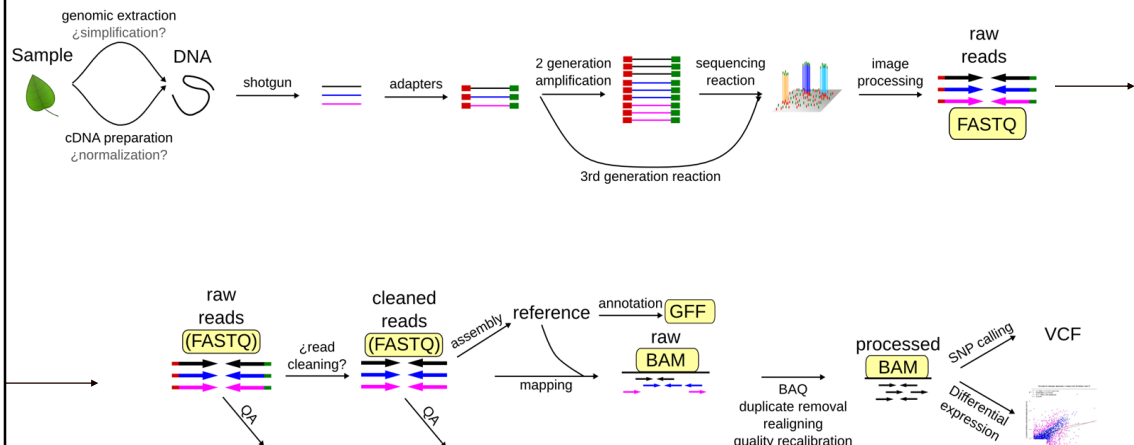
# Saccharomyces cerevisiae S288C genome (version=R64-3-1)
#
# Features from the 16 nuclear chromosomes labeled chrI to chrXVI,
# plus the mitochondrial genome labeled chrmt.
#
# Created by Saccharomyces Genome Database (http://www.yeastgenome.org/)
#
# Weekly updates of this file are available for download from:
# https://downloads.yeastgenome.org/latest/saccharomyces_cerevisiae.gff.gz
#
# Please send comments and suggestions to sgdb-helpdesk@lists.stanford.edu
#
# SGD is funded as a National Human Genome Research Institute Biomedical Informatics Resource from
# the U. S. National Institutes of Health to Stanford University.
#
chrI    SGD    chromosome    1    230218    .    .    .    ID=chrI;dbxref=NCBI:BK006935.2;Name=chrI
chrI    SGD    telomere    1    801    .    .    .    ID=TEL01L;Name=TEL01L;Note=Telomeric%20on%20the%20left%20arm%20of%20Chro
rie=SGD:5000028062
chrI    SGD    X_element    337    801    .    .    .    Parent=TEL01L;Name=TEL01L_X_element
chrI    SGD    X_element_combinatorial_repeat    63    336    .    .    .    Parent=TEL01L;Name=TEL01L_X_element_combinatorial_repeat
chrI    SGD    telomeric_repeat    1    62    .    .    .    Parent=TEL01L;Name=TEL01L_telomeric_repeat_1
chrI    SGD    telomere    1    801    .    .    .    ID=TEL01L_telomere;Name=TEL01L_telomere;Parent=TEL01L
chrI    SGD    gene    335    649    .    +    .    ID=YAL069W;Name=YAL069W;Ontology_term=GO:0003674,GO:0005575,GO:0008150,SO:0000704;Note=Du
13
chrI    SGD    CDS    335    649    .    +    0    Parent=YAL069W_mRNA;Name=YAL069W_CDS;orf_classification=Dubious;protein_id=UniProtKB:Q135
chrI    SGD    mRNA    335    649    .    +    .    ID=YAL069W_mRNA;Name=YAL069W_mRNA;Parent=YAL069W
```

17/04/2024

Ampliación de la anterior imagen:

[sgd-archive.yeastgenome.org](https://archive.yeastgenome.org)

## ¿Cómo se realiza la secuenciación genómica?



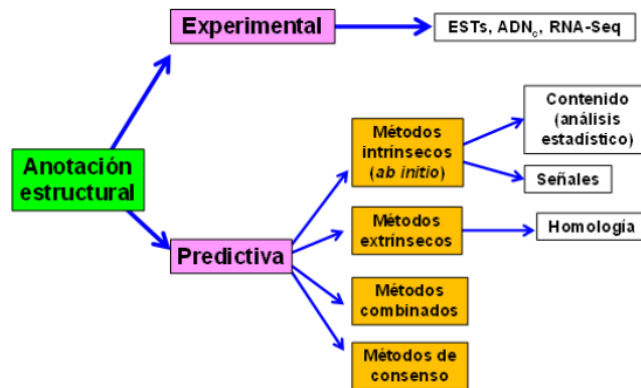
17/04/2024

**Anotación de genes y regiones:** Identificar los genes y otras regiones funcionales en el genoma es esencial. Los programas de anotación buscan características como exones (partes codificantes), intrones, promotores y regiones reguladoras.

**Identificación de variantes:** Se buscan diferencias entre el genoma secuenciado y una referencia (como el genoma humano de referencia). Las variantes pueden ser SNPs (polimorfismos de un solo nucleótido), inserciones, deleciones o duplicaciones. Herramientas de detección de variantes identifican estas diferencias.

## Anotación de secuencias y genomas

Una secuencia sin analizar no nos aporta ninguna información, es necesario extraerla a partir de la secuencia.



17/04/2024

Finalmente, la anotación del genoma es un proceso crítico en la investigación genómica, ya que permite identificar y comprender las regiones funcionales del ADN y los genes que están presentes en el genoma. La anotación del genoma implica asignar funciones biológicas y características a las diferentes regiones del genoma, y proporcionar información sobre las secuencias de ADN que son importantes para la regulación de la expresión génica y otros procesos celulares.

La anotación del genoma es importante por varias razones:

1. **Identificación de genes:** La anotación del genoma permite identificar los genes que están presentes en el genoma, y proporciona información sobre su ubicación y función. Esto es fundamental para comprender cómo los genes interactúan y regulan los procesos celulares y fisiológicos.
2. **Interpretación de variaciones genéticas:** La anotación del genoma permite identificar las variaciones genéticas que pueden estar asociadas con enfermedades o rasgos específicos. Esto es importante para la investigación médica y para la identificación de objetivos terapéuticos.
3. **Estudio de la evolución:** La anotación del genoma permite comparar el genoma de diferentes especies y comprender cómo ha evolucionado el genoma a lo largo del tiempo. Esto puede proporcionar información sobre la historia evolutiva de una

especie y sobre cómo se han desarrollado características biológicas específicas.

4. Mejora de la calidad del genoma: La anotación del genoma puede ayudar a identificar y corregir errores en la secuencia del genoma, lo que puede mejorar la calidad del genoma de referencia utilizado en la investigación genómica.

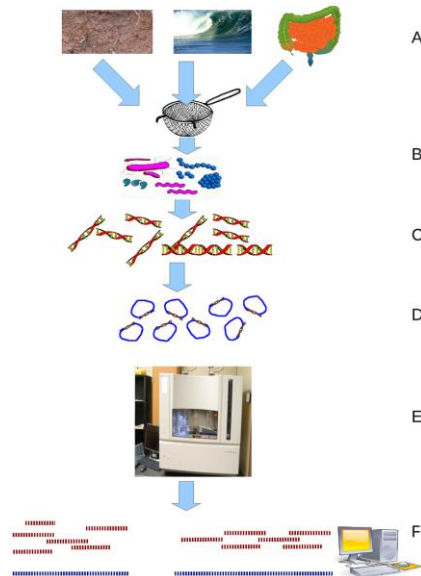
En resumen, la anotación del genoma es fundamental para la identificación de genes y la interpretación de variaciones genéticas, y proporciona información crítica para el estudio de la evolución y la mejora de la calidad del genoma.



# Metagenómica y metataxonómica

## Metagenómica

- Estudio del conjunto de genomas de una muestra compleja
- La diversidad varía según el tipo de muestra
- Permite conocer el conjunto de genes y sus funciones asociadas
- Permite realizar estudio entre comunidades, como pacientes enfermos y sanos
- Es necesaria una profundidad de secuenciación mínima

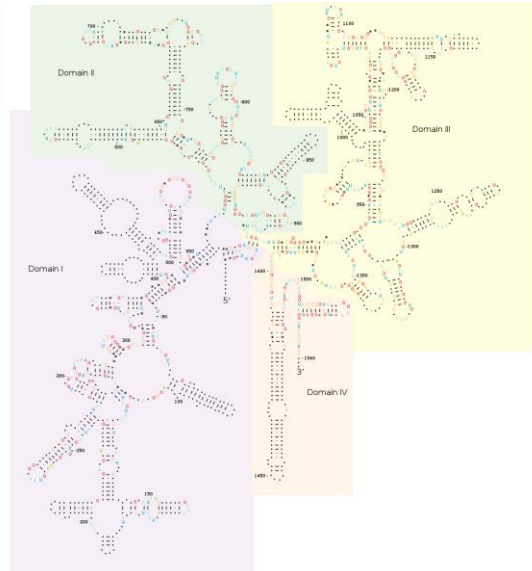


La metagenómica es la parte de la genómica que estudia la comunidad de genomas de microorganismos desde muestras complejas, de las cuales en la mayoría de los casos desconocemos su composición. A partir de estas muestras purificamos los organismos de interés y obtenemos sus genomas. Estos se secuencian siguiendo diferentes estrategias y se ensamblan y analizan posteriormente. Este tipo de estudio nos permite evaluar la diversidad de los organismos que encontramos en estas muestras y podemos realizar estudios comparativos entre, por ejemplo, muestras de pacientes enfermos y sanos. La particularidad de esta técnica es que obtenemos un conjunto amplio de genes, por lo que se necesita una mayor capacidad de secuenciación y de análisis.

## Metataxonómica

- Estudio de amplicones 16S – ITS (Metataxonómica)
- Útil para la clasificación taxonómica
- Regiones conservadas hipervariables
- Se valora una alta profundidad por encima de longitud de lectura
- Bases de datos SILVA y Greengenes (otras)

NO es metagenómica



La metataxonómica se centra en el análisis y la clasificación de las comunidades microbianas presentes en muestras ambientales. En la metataxonómica no secuenciamos el ADN de una muestra compleja de organismos en su totalidad. Se secuenciamos una zona concreta del ADN, mediante la amplificación de un fragmento de ADN, generalmente, genes específicos altamente conservados.

El gen del ARN ribosómico 16S (ADNr) está altamente conservado en procariotas y arqueas, pero no está presente en eucariotas como los seres humanos. Junto a las regiones altamente conservadas, el rRNA 16S tiene nueve regiones hipervariables que permiten la clasificación por género e incluso a nivel de especie. La metataxonómica 16S se basa en la amplificación por PCR de regiones variables en el rDNA 16S utilizando cebadores específicos, seguida de la secuenciación de los amplicones. En este tipo de secuenciación se valora más una alta profundidad de secuenciación que una mayor longitud de secuencia.

Para recuperar identidades bacterianas de los datos de secuenciación, se utiliza un enfoque bioinformático en el que cada secuencia de amplicón de rDNA 16S se compara con secuencias de rDNA 16S en bases de datos de referencia. La metataxonómica 16S se ha utilizado con frecuencia para estudiar la diversidad de comunidades bacterianas en distintos entornos naturales, como el intestino humano.

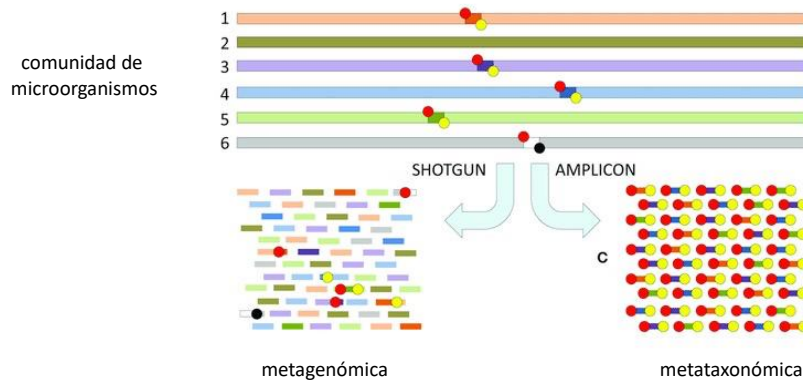
o muestras de agua de océano.

Hay que dejar un concepto claro en este punto, la diferencia entre metagenómica y metataxonómica. A partir de la metagenómica estudiamos el conjunto de todos los genomas provenientes de una muestra compleja, mientras que mediante la metataxonómica solo vamos a amplificar el gen de ARN ribosómico en cuestión, teniendo solo la información de la taxonomía de la comunidad analizada. También es importante destacar que conseguir recomponer todos los genomas de la muestra es a priori posible pero en la práctica extremadamente difícil debido a la distinta proporción en la que se encuentran los microorganismos en la muestra.

Para eucariotas como hongos y levaduras, la región de interés son los ITS (Internal Transcribed Spacer), también del ADNr.

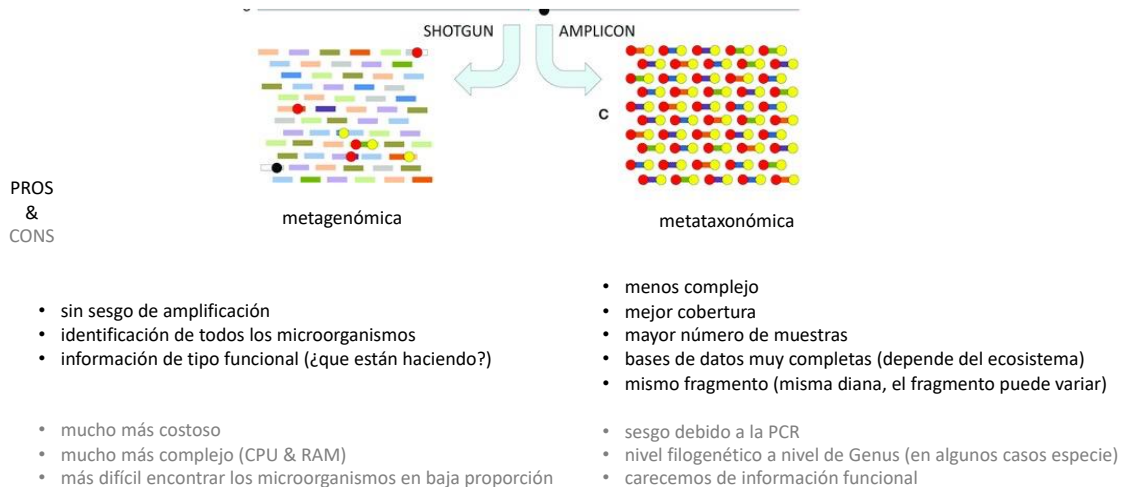
## shotgun seq vs amplicon seq (metataxonomics)

- Tenemos que distinguir entre metagenómica y metataxonómica.
  - Secuenciación de todo el ADN.
  - Secuenciación de una zona concreta del ADN, mediante la amplificación de un fragmento de ADN.



17/04/2024

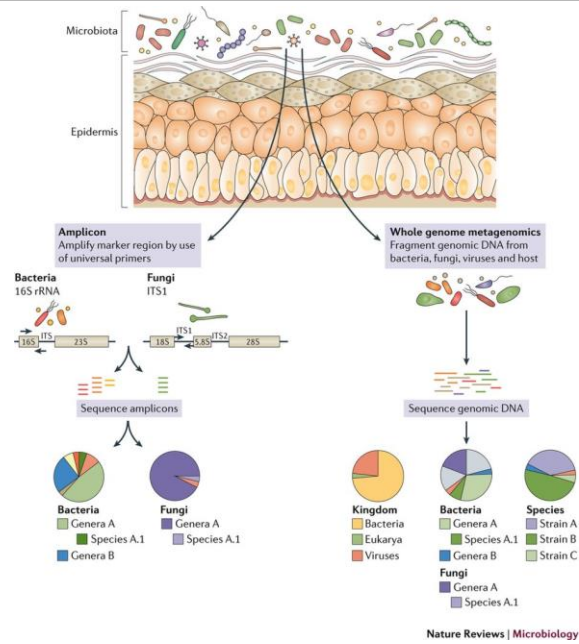
## shotgun seq vs amplicon seq (metataxonomics)



17/04/2024

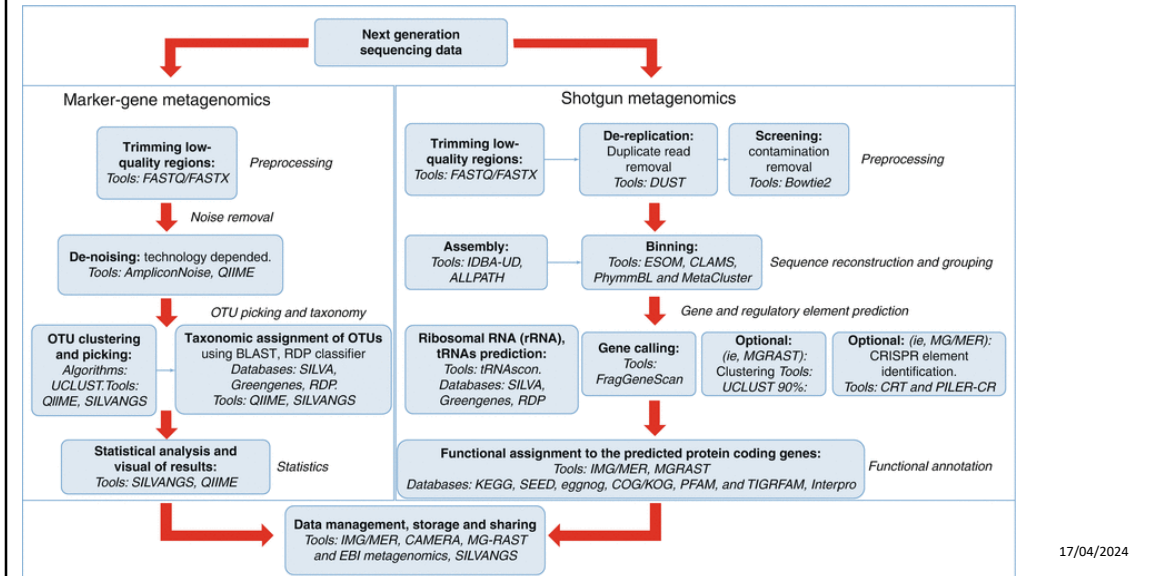
La secuenciación de amplicones tiene una serie de ventajas y desventajas. Entre las principales ventajas destaca el ser una metodología sencilla y muy barata, además de que el análisis de los resultados es “sencillo” ya que está muy bien estudiado y las bases de datos hace años que tiene un gran porcentaje de la información sobre especies más frecuentes en todos los medios (se actualizan con frecuencia pero añadiendo muy poca información nueva entre versiones). Como contraparte, solo estamos analizando un único gen del conjunto de genes que tenemos en nuestra comunidad bacteriana. La resolución a niveles taxonómicos bajos (género o especie) es cuestionable, ya que la variabilidad entre distintas especies en la mayoría de ocasiones es tan baja que podemos obtener resultados poco significativos.

## shotgun seq vs amplicon seq (metataxonomics)



17/04/2024

## Estrategia experimental y de análisis diferente:



17/04/2024

Atención al uso de metagenomics indistinto en algunas fuentes...

## Se requiere una cantidad/calidad mínima de masa

Dependiendo del protocolo de secuenciación

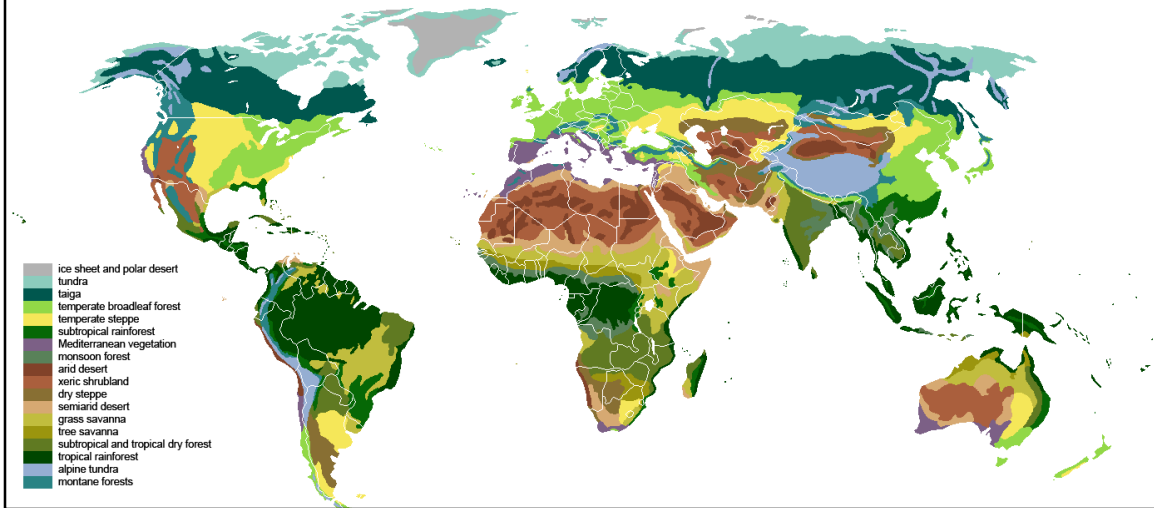
100ng < ADN < 1 microgramo

17/04/2024

Antes de acabar con el bloque de DNaseq dejar claro que la cantidad de masa necesaria para realizar cualquier proceso de secuenciación es crítica. Dependiendo del protocolo que vayamos a seguir y la tecnología de secuenciación esta puede variar, pero actualmente es probable tener problemas si la cantidad de masa mínima está por debajo de 100ng. Este problema puede aparecer si analizamos muestras recogidos con hisopos desde mucosas, como la boca, nariz o vagina, ya que la cantidad de masa está en el orden de picogramos. Este problema se puede solventar solo mediante soluciones parciales, como muestrear del mismo punto varias veces o realizar un proceso de amplificación inespecífica de la muestra antes del paso de secuenciación, lo cual introduce sesgos, en mayor o menor medida, en los resultados de posteriores análisis.

## Algunos conceptos...

Un **bioma** es una gran colección de flora y fauna que ocupa un hábitat principal.

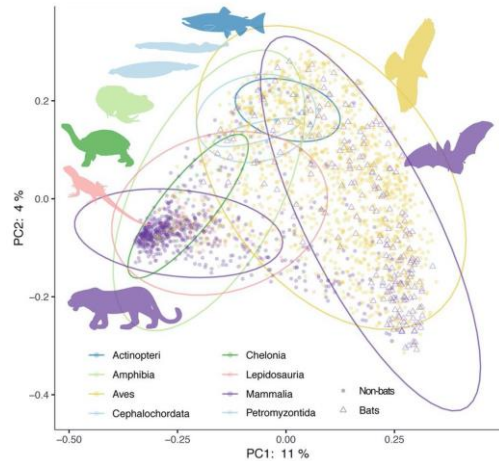
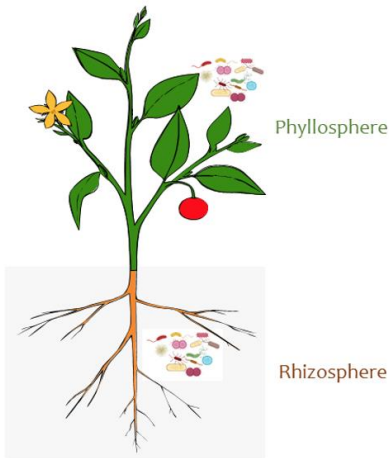


Vamos a definir una serie de términos que se utilizan comúnmente en metagenómica y metataxonómica.

Un **bioma** es una comunidad de plantas y animales que comparten las mismas condiciones climáticas y ambientales en una región geográfica determinada. Los biomas se caracterizan por tener una vegetación y fauna específicas, así como por factores abióticos, como el clima, la geología y el suelo. Los biomas son una forma de clasificar y comprender la diversidad biológica en todo el mundo y se dividen en varios tipos principales, como los bosques tropicales, los desiertos, las praderas, los tundras y los océanos. Cada bioma tiene características únicas que influyen en la diversidad de especies y ecosistemas que se encuentran en él.

## Algunos conceptos...

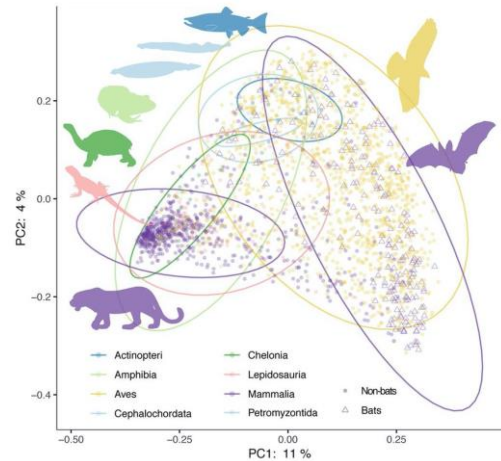
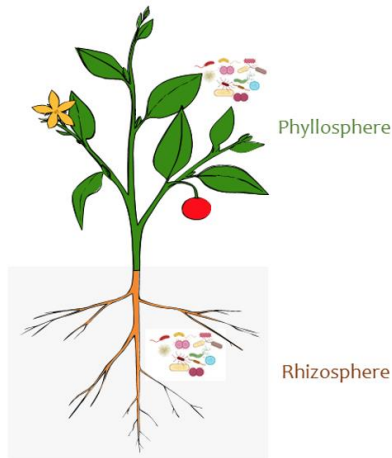
Un **microbioma** es la comunidad de microorganismos que generalmente se pueden encontrar viviendo juntos en determinado hábitat (o *bioma*) sumado a metabolitos y condiciones ambientales que lo forman.



Por lo general, se refiere a los microorganismos que habitan en el cuerpo humano o animal, aunque también puede aplicarse a otros entornos, como el suelo o el agua. El microbioma humano es particularmente importante porque estos microorganismos tienen un papel clave en la digestión, el sistema inmunológico, la regulación hormonal y muchas otras funciones importantes del cuerpo. Un microbioma desequilibrado puede estar relacionado con una serie de problemas de salud, como enfermedades autoinmunitarias, trastornos metabólicos y obesidad.

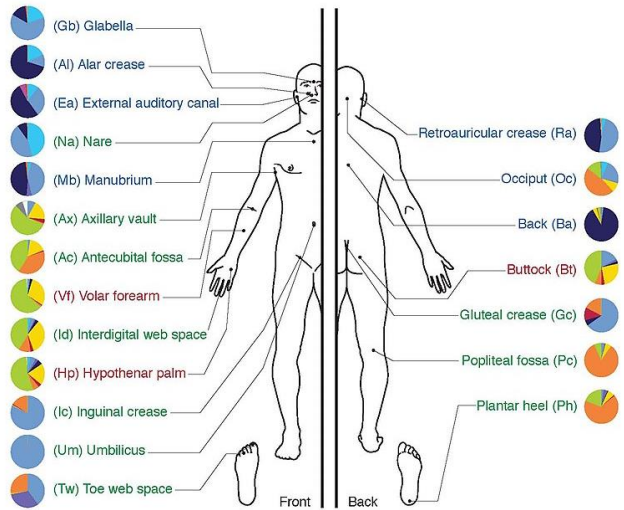
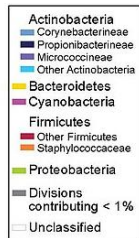
## Algunos conceptos...

La **microbiota** está formada por todos los miembros vivos que forman el **microbioma**. Las bacterias, las arqueas, los hongos, las algas y los pequeños protistas se consideran miembros del microbioma (La integración de fagos, virus, plásmidos y elementos genéticos móviles es más controvertida).



## Algunos conceptos...

El **microbioma humano** es el agregado de toda la **microbiota** que reside sobre o dentro de los tejidos y biofluidos humanos junto con los lugares anatómicos correspondientes en los que residen, incluida la piel, las glándulas mamarias, el líquido seminal, el útero, los folículos ováricos, los pulmones y la saliva, mucosa oral, conjuntiva, tracto biliar y tracto gastrointestinal. Los **tipos de microbiota humana** incluyen bacterias, arqueas, hongos, protistas y virus.

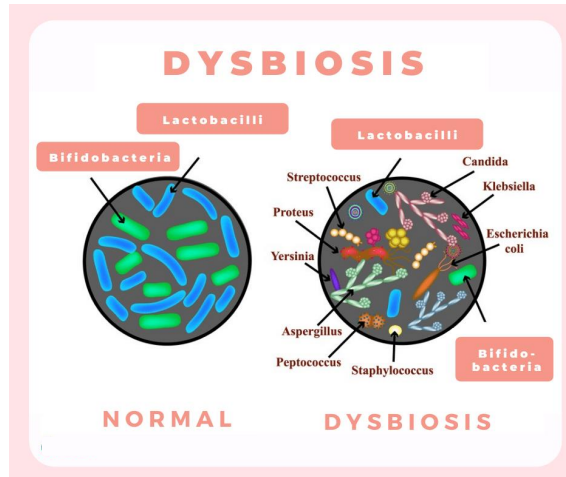


17/04/2024

## Algunos conceptos...

La **disbiosis** es la alteración de la homeostasis de la microbiota causada por un desequilibrio en la microflora que llevan a cambios en su composición funcional y actividades metabólicas, o un cambio en su distribución local.

Se trata de un desequilibrio microbiano.



17/04/2024

A large orange semi-circle is positioned at the top of the slide, with its flat edge aligned with the top border of the content area. The semi-circle is centered horizontally and its diameter spans most of the width of the slide.

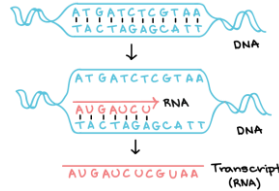
Transcriptómica

## Transcriptómica

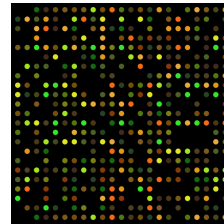
Análisis del transcriptoma, presencia y cantidad de ARN (también llamados transcritos) en una muestra en un momento dado

Análisis de todos los tipos de ARN

- Codificante
- microARN
- Ribosómico
- Transferente



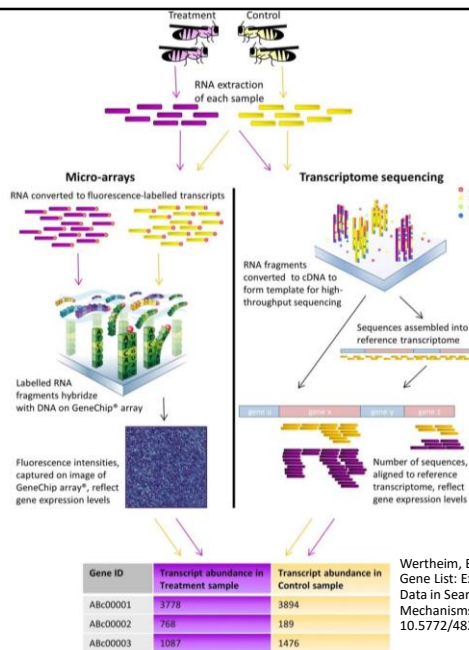
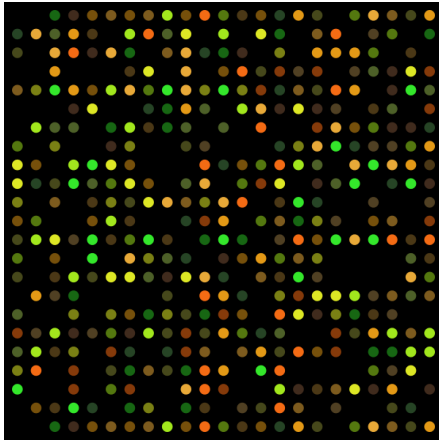
Comúnmente el análisis de transcritos se ha realizado con microarrays:



17/04/2024

La transcriptómica se encarga de estudiar el transcriptoma. El transcriptoma se refiere al conjunto de todos los transcritos o moléculas de ARN producidos a partir del ADN en una célula o tejido en un momento dado. El ADN se transcribe en ARN a través del proceso de transcripción, y el transcriptoma representa el conjunto completo de todas las moléculas de ARN producidas en una célula, incluyendo los ARNm, los ARN no codificantes, los ARN ribosómicos y los ARN de transferencia.

## Dos estrategias



Wertheim, Bregie. (2012). Beyond the Gene List: Exploring Transcriptomics Data in Search for Gene Function, Trait Mechanisms and Genetic Architecture. 10.5772/48239.

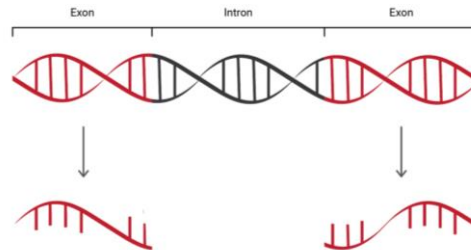
Existen dos técnicas principalmente, microarrays o RNAseq (secuenciación de ARN). Importante destacar que independientemente de la técnica que usemos, partimos de dos muestras, control y problema. Es decir, de una u otra forma, siempre comparamos como ha variado nuestro transcriptoma comparado con el control.

El análisis de estos tipos de ARN se lleva a cabo sobre placas de microarrays, donde mediante una señal fluorescente se puede determinar la sobre o sub expresión de determinados tipos de ARN de forma simultánea. Esta técnica clásica ha quedado relegada al análisis de estos ARN no codificantes mientras que el ARN mensajero se analiza a día de hoy siguiendo técnicas y métodos de RNAseq. Esto no significa que puedan existir justificaciones para el uso de los microarrays. Los microarrays utilizan sondas prediseñadas para detectar la expresión de genes específicos, la hibridación entre las sondas y las muestras de ARN genera señales que se cuantifican.

## Secuenciación de ARN

### Ventajas sobre microarray:

- **Permite análisis de todo el transcriptoma:** Proporciona una visión global de la expresión génica
- **No requiere conocimiento previo de las secuencias:** A diferencia de los microarrays, no depende de sondas diseñadas previamente.
- **Mayor sensibilidad:** Puede detectar transcripciones de baja abundancia y diferenciar isoformas.
- **Identificación de variantes:** Puede identificar variantes genéticas y mutaciones.
- **No sufre problemas de hibridación cruzada:** La secuenciación de ARN no está limitada por problemas de hibridación no específica.

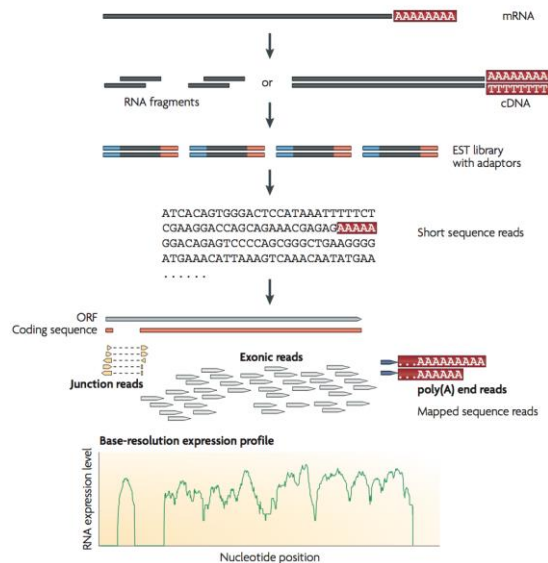


17/04/2024

El RNAseq o secuenciación de ARN mensajero es una técnica actual desarrollada en bioinformática para evaluar la presencia y cantidad de determinados ARN mensajeros en distintas muestras. Presenta una serie de ventajas sobre el análisis de microarray, como una mayor escala o identificación de polimorfismos de nucleótidos (tenemos acceso a la secuencia de ARN).

Es una técnica útil para la identificación de expresión diferencial de genes, es decir, ver qué genes están sobre o sub representados al compararlo con una muestra control, de la misma forma que en un microarray. A partir de la secuenciación de esta información podemos realizar agrupación de secuencia según determinados criterios e identificar genes marcadores relacionados por ejemplo con el desarrollo de alguna enfermedad como cáncer. Los genes que se agrupen basados en su perfil de expresión, por ejemplo siguiendo un método basado en *hierarchical clustering*, pueden no tener exactamente la misma función pero sí estar involucrados en la misma ruta metabólica.

## ¿Como se realiza la secuenciación de ARN?



17/04/2024

**Extracción de ARN:** Se extrae el ARN total de la muestra biológica de interés (por ejemplo, tejido, células o fluidos corporales). La calidad y cantidad del ARN extraído son críticas para el éxito del análisis.

**Preparación de la librería de ARN:** Esto implica la eliminación de ribosomas o selección por poli-A, si se estudia el ARN mensajero, la fragmentación y la adición de adaptadores.

**Secuenciación:** generalmente utilizando secuencias cortas (short read sequencing) de plataformas de segunda generación de secuenciadores

**Análisis de datos y secuencia:** Los datos de secuenciación se recopilan y se analizan para obtener la secuencia completa del ADN mediante una serie de programas bioinformáticos:

**Limpieza y calidad de lecturas:** Las secuencias de ADN generadas por la secuenciación son fragmentos cortos llamados "lecturas" o "reads". El primer paso será comprobar su calidad y "limpiarlas".

**Alineación de las lecturas:** Las lecturas se asignan a regiones específicas del genoma.

**Análisis diferencial de expresión:** Se comparan las muestras (por ejemplo, condiciones de tratamiento vs. control) para identificar genes cuya expresión difiere significativamente. Se utilizan herramientas estadísticas para determinar qué genes están regulados de manera diferente. Existe un proceso de normalización de las

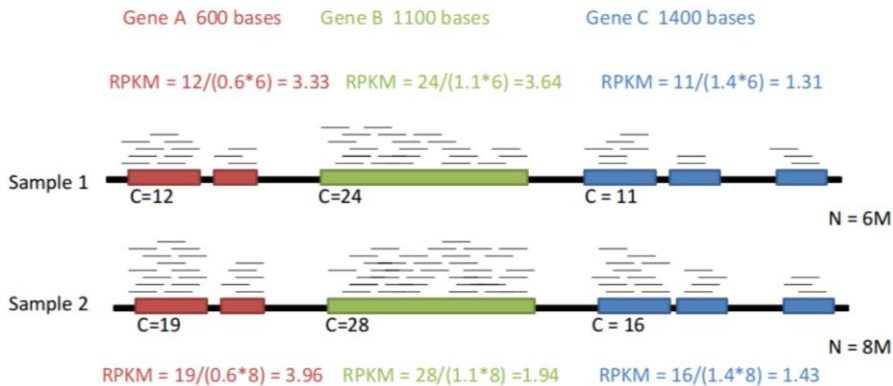
muestras.

**Exploración de vías y redes génicas:** Se analizan las interacciones entre genes y proteínas. Se identifican vías metabólicas y redes de regulación.

**Interpretación biológica:** Se relacionan los hallazgos con procesos biológicos, enfermedades o respuestas a tratamientos.

## Análisis diferencial de expresión: normalización: RPKMs

Abundancia medida en lecturas **mapeadas** por kilobase (RPKM).



17/04/2024

El alineamiento de las secuencias sobre el genoma de referencia se analiza, midiendo la cantidad de secuencias que caen sobre cada posición del genoma. Obteniéndose un valor denominado **RPKM (Reads Per Kilobase Million)** que es la medida utilizada para cuantificar la expresión génica relativa. Los RPKMs representan la **abundancia relativa** de un gen específico en una muestra de ARN. Se calculan **normalizando** la cantidad de lecturas (reads) mapeadas a un gen por su longitud y el número total de lecturas en la muestra. Permiten comparar la expresión de diferentes genes dentro de una muestra.

**Cálculo de RPKMs:** El cálculo se realiza en tres pasos:

Se cuentan las lecturas mapeadas a un gen específico.

Se divide el conteo de lecturas por la longitud del gen en kilobases (KB).

Finalmente, se normaliza dividiendo por el número total de lecturas en millones.



# Exploración de vías y redes génicas

HITS

HITS per million

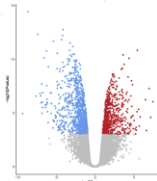
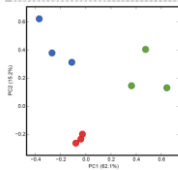
RPKMs

HITS per Kb

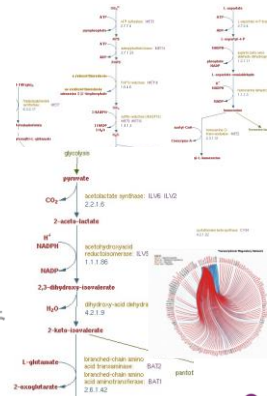
TPMs

Log Fold Changes

id	Associated	T	FC	FC	FC	FC	FC	FC	FC
FV0233C	MT122	0.28	0.22	0.28	0.44	0.40			
FV0233C	MT118	0.09	0.01	0.01	0.11	0.11			
FV0233C	MT111	0.04	-0.02	-0.02	-0.01	-0.07			
FV0233C	MT117	-0.03	0.01	0.01	0.02	0.07			
FV0233C	MT120	0.32	-0.02	-0.02	-0.03	-0.19			
FV0233C	MT114	0.04	-0.01	-0.01	-0.05	-0.12			
FV0233C	MT118	0.34							
FV0233C	MT122	-0.17	-0.01	-0.01	-0.05	-0.14			
FV0233C	MT122	0.02	-0.09	-0.09	-0.32	-0.13			
FV0233C	MT112	0.28	0.01	0.01	0.11	0.11			
FV0233C	MT123	0.12	-0.07	-0.07	-0.19	-0.19			
FV0233C	MT118	0.23	-0.01	-0.01	-0.14	-0.19			
FV0233C	MT118	0.17	-0.01	-0.01	-0.14	-0.19			
FV0233C	MT118	0.14	-0.01	-0.01	-0.14	-0.19			



Gene Ontology  
Pathway enrichment



17/04/2024

A tener en cuenta:

- Variabilidad en el resultado:
  - Diseño experimental: número de réplicas!
  - Tasa de error de la plataforma de secuenciación, los niveles de expresión puede variar entre rondas secuenciación
  - Es importante un número mínimo de muestras biológicas y tener en cuenta la variabilidad biológica de la muestra
- La señal de expresión está limitada por la profundidad de secuenciación
- Un mayor número de lecturas se asignan a secuencias largas → sesgo

¿Cuántas lecturas se requiere por muestra?

5 – 20 millones de lecturas por muestra  
Levaduras simples 30 millones de lecturas cortas (35pb)

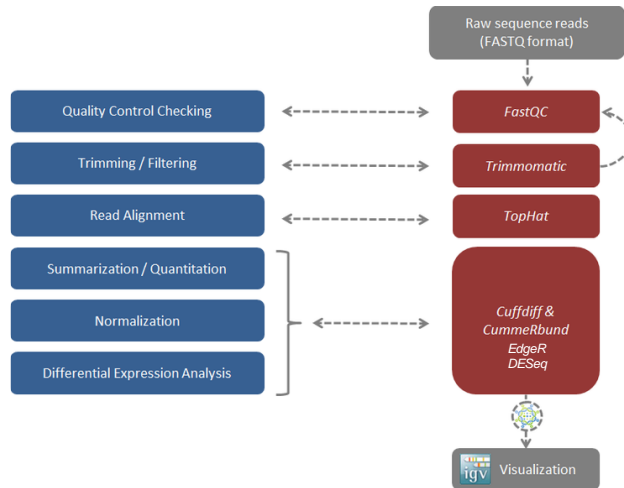
17/04/2024

El análisis de estos datos no es perfecto, y existe una variabilidad que pueden influir en nuestros resultados. Por ejemplo, pueden existir errores derivados de la plataforma de secuenciación en función del momento en que se realice la secuenciación, o muestras biológicas analizadas en distintos momentos. Si bien es cierto que esto influye también a la secuenciación de ADN para otros análisis relacionados, los valores recogidos por RNAseq deben ser lo más precisos posibles, ya que la mínima variación en el cálculo del fold-change en función del gen analizado puede dar resultados significativamente muy diferentes entre análisis del mismo tipo.

Por otra, parte la señal de expresión está limitada por la profundidad de secuenciación y la longitud del mensajero. Por probabilidad, a mayor profundidad conseguiremos representación de genes poco expresados, mientras que a mayor longitud de la secuencia analizada tendremos más problemas en la aparición de sesgos respecto a la representación de estas secuencias. De esta forma llegamos a la pregunta, ¿cuál es la profundidad de secuenciación mínima requerida para RNASeq? Esta pregunta depende de qué queremos estudiar y la cantidad de mensajeros que queramos analizar. Un rango oscila entre los 5 a 20 millones de secuencias por muestra, o en el ejemplo de *S. cerevisiae* donde necesitaremos unas 10 millones de lecturas de 100 pares de bases. Se puede obtener resultados con menor número de

lecturas pero tendremos que ser cuidados con el grado de significatividad.

## RNAseq pipeline



17/04/2024

A large orange semi-circular shape is positioned at the top of the slide, partially cut off by the top edge. It is centered horizontally and its flat edge is at the top.

Proteómica

## Principales flujos de trabajo

### DNaseq

Conjunto de técnicas para conocer la secuencia de nucleótidos de DNA y los análisis derivados, como alineamientos y filogenia.

### RNAseq

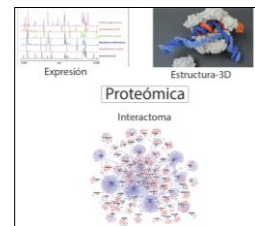
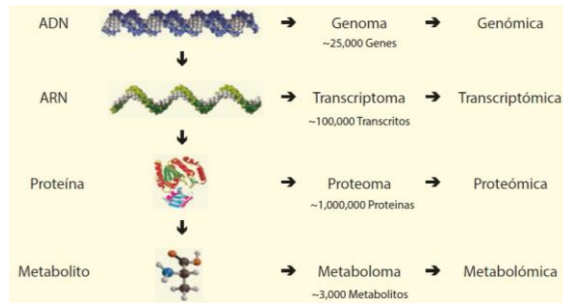
Análisis del transcriptoma (presencia y cantidad de ARN) en la muestra en un momento dado

### Proteómica

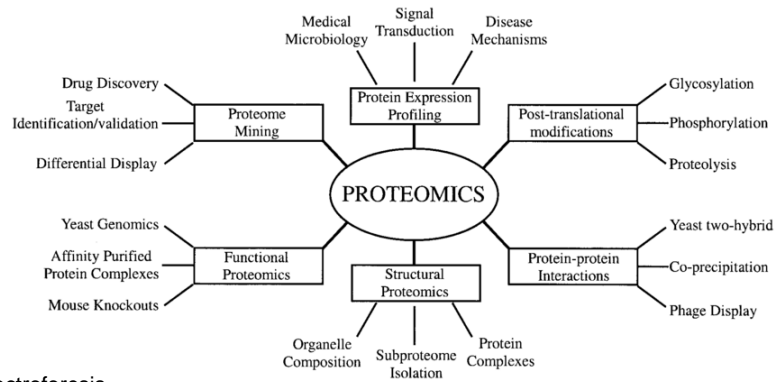
Análisis del proteoma (presencia y cantidad de proteínas) en la muestra en un momento dado. Estudio de la estructura proteica, análisis de función proteica.

### Metabolómica

Análisis de los metabolitos, también diferencialmente, (presencia y cantidad de metabolitos) en la muestra en un momento dado.



Tras el análisis del ADN y ARN pasamos al análisis de otros metabolitos, como las proteínas. Las proteínas se estudian dentro de lo que conocemos como proteómica que, como dijimos anteriormente, no solo engloba el análisis de la secuencia de aminoácidos o su cantidad por muestra, si no también del análisis de su estructura.



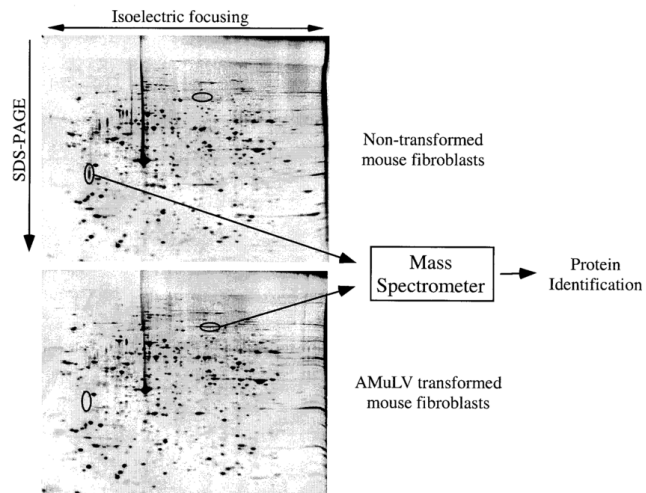
Separación mediante electroforesis  
 Cuantificación basada en espectrometría de masas  
 Proteómica estructural  
 Resolución mediante inteligencia artificial  
 Secuenciación de cadenas polipeptídicas

17/04/2024

El **proteoma** celular es la totalidad de proteínas expresadas en una célula particular bajo condiciones de medioambiente y etapa de desarrollo (o ciclo celular) específicas, como lo puede ser la exposición a estimulación hormonal.

La proteómica tradicional consiste en la extracción proteica de la muestra y separación en un gel de electroforesis en 2D, donde se separan por tamaño y por carga eléctrica, creando un mapa. De esta forma se comparaban muestras y se identificaban cambios en el patrón de expresión proteica. Un paso adelante en esta cuantificación son las técnicas de separación y cuantificación de proteínas basadas en espectrometría de masas. Muy potentes actualmente son capaces de reconocer miles de proteínas de una muestra y comparar su cantidad relativa entre experimentos, similar a un ensayo de RNAseq en el resultado. La proteómica estructural es otra rama de la proteómica encargada del estudio y análisis de la estructura tridimensional de las proteínas. Recientemente en la proteómica han ocurrido dos grandes hitos, la resolución de estructuras tridimensionales mediante modelos de inteligencia artificial y más reciente aun la secuenciación de cadenas polipeptídicas mediante secuenciadores.

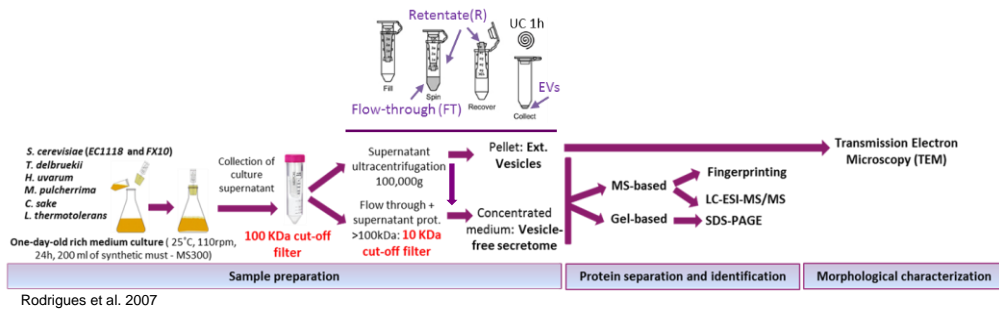
## Proteómica de expresión tradicional



17/04/2024

Separación de proteínas en un gel de electroforesis en 2D.

## Proteómica de expresión: espectrometría de masas



Reverse Phase Liquid Chromatography  
Electro Spray Ionization  
Tandem Mass Spectrometry

RP-LC  
ESI  
MS/MS

Separation  
Ionization  
Identification

17/04/2024

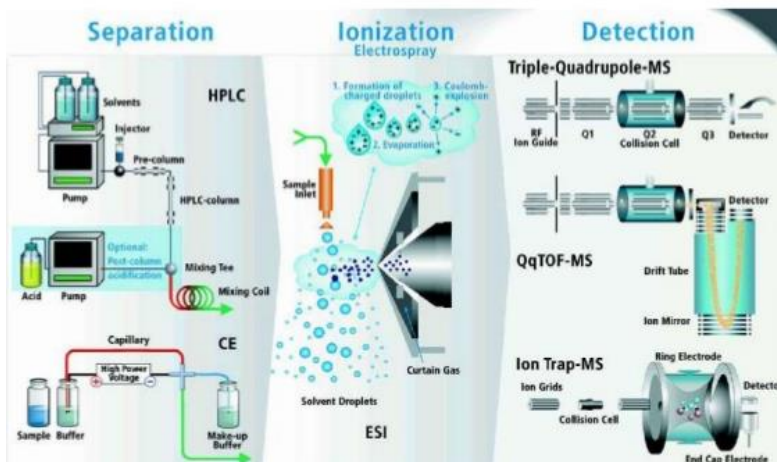
Ejemplo de flujo de trabajo de separación y cuantificación de proteínas basadas en espectrometría de masas.

## Proteómica de expresión: espectrometría de masas

Reverse Phase Liquid Chromatography  
Electro Spray Ionization  
Tandem Mass Spectrometry

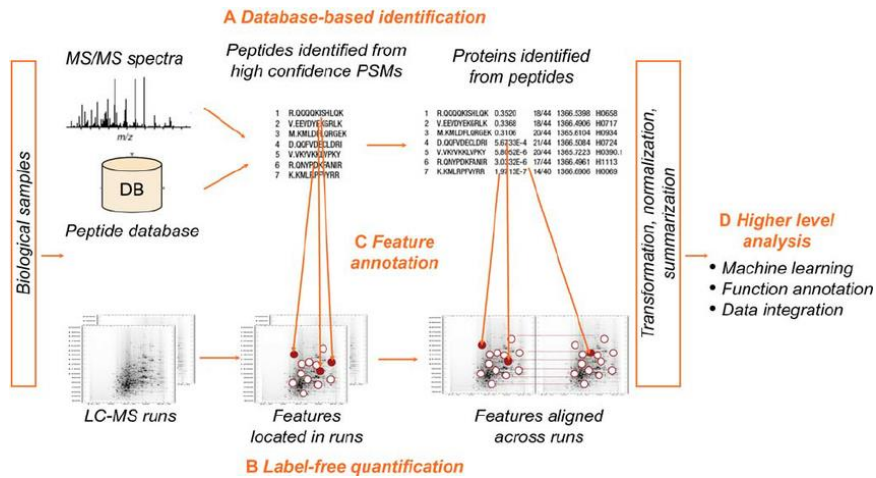
RP-LC  
ESI  
MS/MS

Separation  
Ionization  
Identification

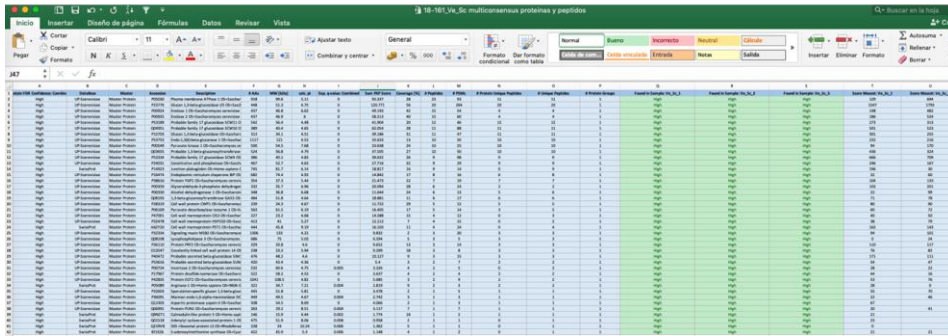


17/04/2024

# Proteómica de expresión: espectrometría de masas



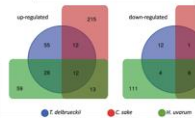
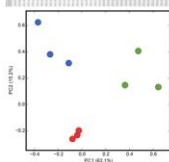
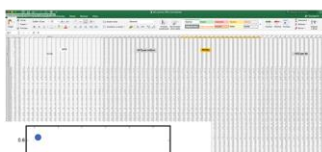
17/04/2024



57

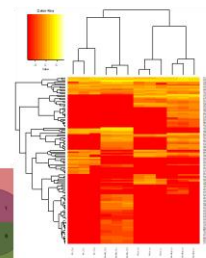
# Proteómica de expresión: espectrometría de masas

## Péptidos

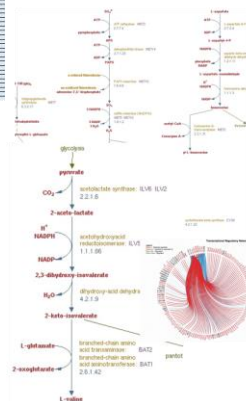


## Relative Abundance Spectral Counting

Protein	Peptide	Abundance
Protein 1	Peptide 1	100
Protein 1	Peptide 2	80
Protein 2	Peptide 3	60
Protein 2	Peptide 4	40
Protein 3	Peptide 5	20
Protein 3	Peptide 6	10



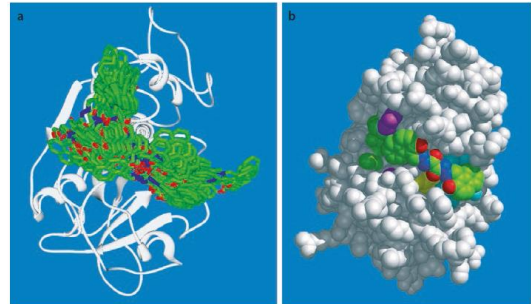
## Gene Ontology Pathway enrichment



17/04/2024

## Proteómica estructural

- Análisis de estructuras proteicas
- No existen métodos analíticos que permitan predecir la estructura tridimensional
  - Cristalografía por rayos X
  - Resonancia magnética nuclear
  - Microscopía electrónica
- Protein Data Bank (PDB)
- Diseño de fármacos

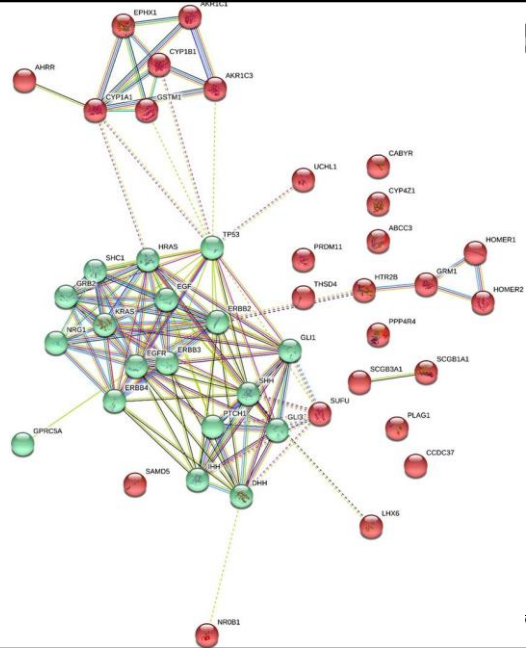


17/04/2024

La proteómica estructural se encarga del análisis de la estructura tridimensional de la proteína. Tradicionalmente la resolución estructural de proteínas se realiza mediante técnicas de biología molecular como cristalografía por rayos X, resonancia magnética nuclear o microscopía electrónica para determinar las distintas isoformas de una misma proteína. La información de estas estructuras se almacenan dentro de la base de datos “Protein Data bank” (PDB) fundada en 1971.

## Interactómica

Interactómica: interacciones proteína-proteína → STRING



1/04/2024

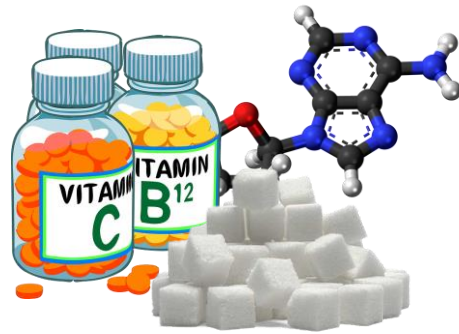
Por último, dentro de la proteómica tenemos la interactómica como disciplina que estudia la interacción entre distintas proteínas en función de la información experimental o bibliográfica de la que se disponga. Aquí por ejemplo vemos una red de diferentes proteínas creada con la aplicación STRING, que se basa en la información bibliográfica para mostrar la interacción de las proteínas de interés en nuestro set de datos. Esto también es extrapolable a los resultados que podemos obtener desde RNASeq, creando redes de genes que tiene una conexión en diferentes rutas metabólicas.

A large orange semi-circle is positioned at the top of the slide, with its flat edge aligned with the top border of the main content area.

Metabolómica

## Metabolómica

- Conjunto de metabolitos de un organismo
- Péptidos, nucleótidos, vitaminas, toxinas...
- Identificación por resonancia magnética nuclear (RMN)
- Espectroscopía de masas
- Farmacogenómica: Variaciones genéticas responsables a cómo una paciente reacciona a distintos medicamentos



17/04/2024

La siguiente disciplina es la metabolómica, la cual estudia otros metabolitos celulares implicados en los distintos procesos metabólicos. Al igual que en proteómica funcional no existen métodos directos para el análisis de estos metabolitos y se depende de resonancia magnética nuclear o espectroscopía de masas para conocer su composición. Este conocimiento es muy útil en el campo de la farmacogenómica, ya que se pueden estudiar las variaciones genéticas (de expresión) al administrar diferentes medicamentos a un paciente.

# Contenidos

## Tema 1. Introducción a la bioinformática

- 1.1 Historia de la bioinformática
- 1.2 Bioética aplicada al análisis de datos

## Tema 2. Principales flujos de trabajo en bioinformática

- 2.1 Genómica
- 2.2 Metagenómica y metataxonómica
- 2.3 Transcriptómica
- 2.4 Proteómica

## Tema 3. Gestión de entornos y paquetes

- 3.1 Conda

## Tema 4. Bases de datos y herramientas bioinformáticas

- 4.1 Principales bases de datos
- 4.2 Otros recursos online

## Tema 5. Alineamiento de secuencias

- 5.1 Introducción al alineamiento de secuencias
- 5.2 Alineamientos Pairwise
- 5.3 Alineamientos Múltiples

## Tema 6. Métodos de secuenciación

- 6.1 Primera generación de secuenciadores
- 6.2 Segunda generación de secuenciadores
- 6.3 Tercera generación de secuenciadores
- 6.4 Comparación de plataformas de secuenciación

## Tema 7. Pre-procesado y calidad de secuencias

- 7.1 Calidad de secuencias
- 7.2 Pre-procesado de secuencias

# ¡Gracias!



**Universidad**  
Internacional  
de Valencia

[universidadviu.com](http://universidadviu.com)

De:  
🌐 Planeta Formación y Universidades