



UNIVERSIDAD DE CHILE

AATAGGCGTACTTTCATCCATGCGCGTATACTAGCCGACACG
AATAGGCGTACTTC
ATAGGCGTACTTCA
AGGCCTACTACATC
GCCTACTTTCATCCA
CTACTTCA GCCATG

Biomedicina I

Elementos de Biología Computacional en Biomedicina

I

TGAACTGATTACCCCTTTTTCGCGTAGCCGAGATGCCAC
CCCCTATTGCGTAG
CCCTATTGCGTAGC
CCTATTGCTTAGCC
CTATTGCGTA*CCG
TATTGCGTAGCCGA
ATTGCGTA*CCGAC



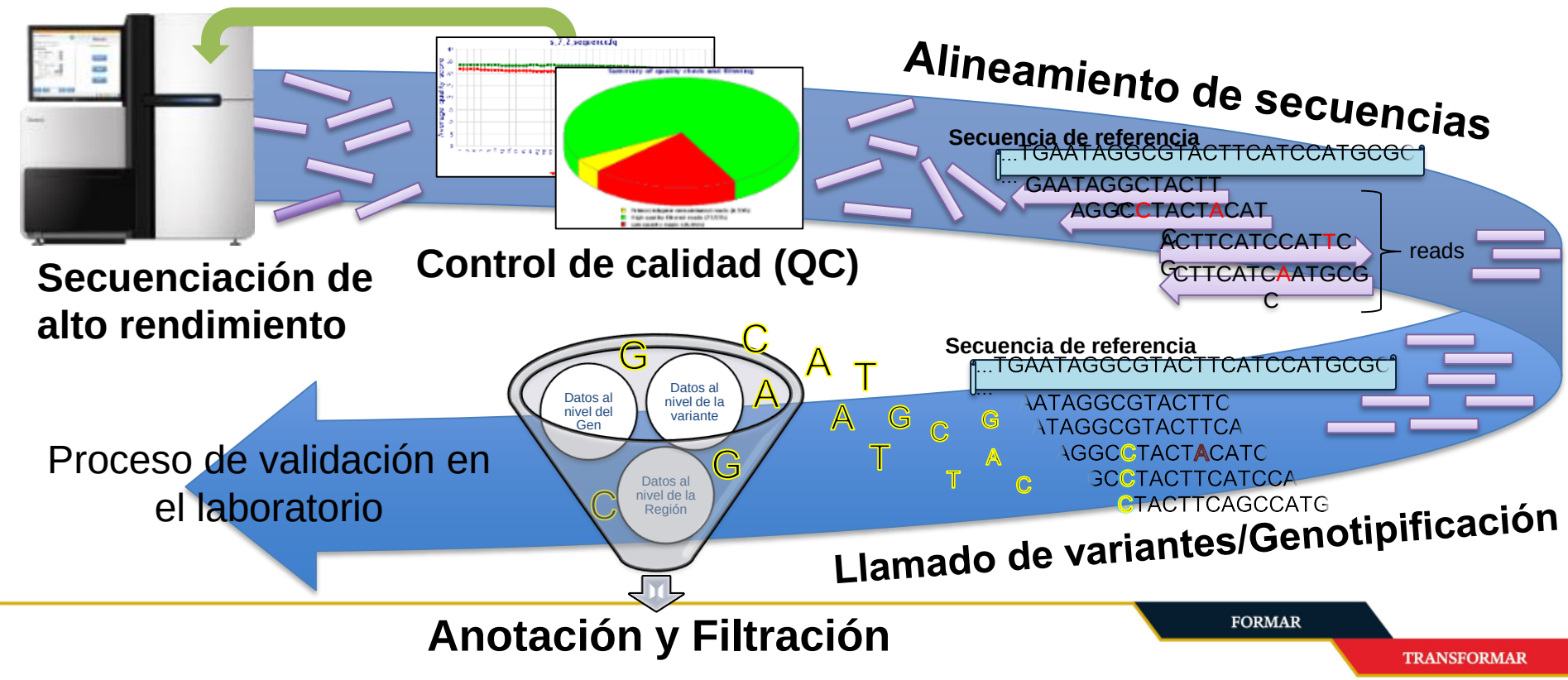
Universidad
Andrés Bello

FORMAR

TRANSFORMAR

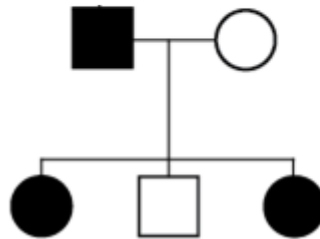
Estructura de la Clase

- En esta clase vamos a explorar las etapas indispensables en el descubrimiento de variantes genómicas que contribuyen a enfermedades.



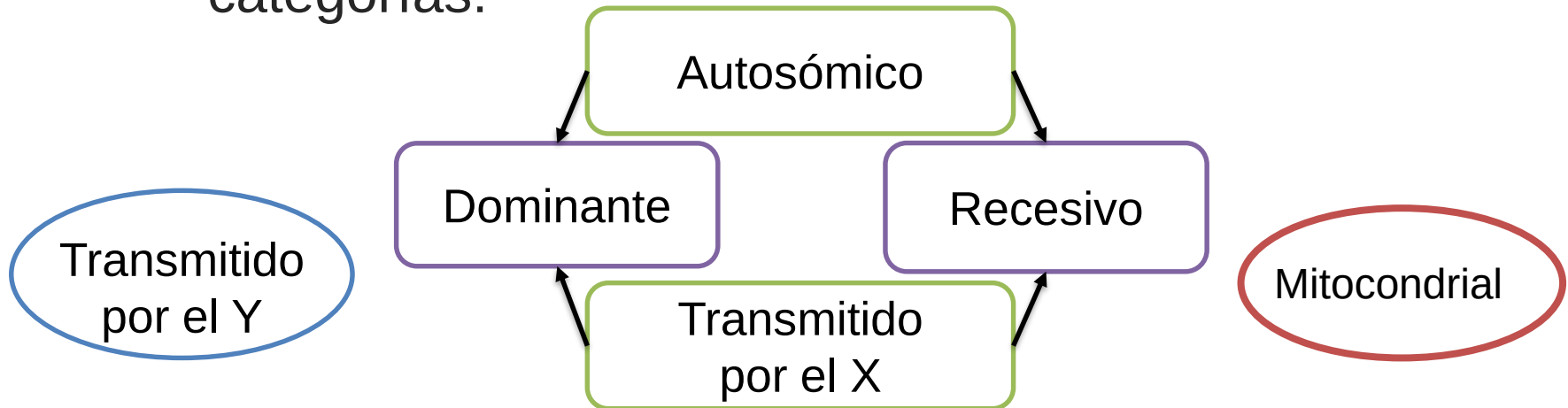
Un Escenario

- Tenemos una familia en la que el padre tiene una condición rara que se manifestó también en dos de tres de sus hijos.
 - Al revisar el historial de la familia con **consejero genético**, el fenotipo de la enfermedad parece seguir un patrón que sugiere una **enfermedad monogénica**.



Enfermedades Mendelianas

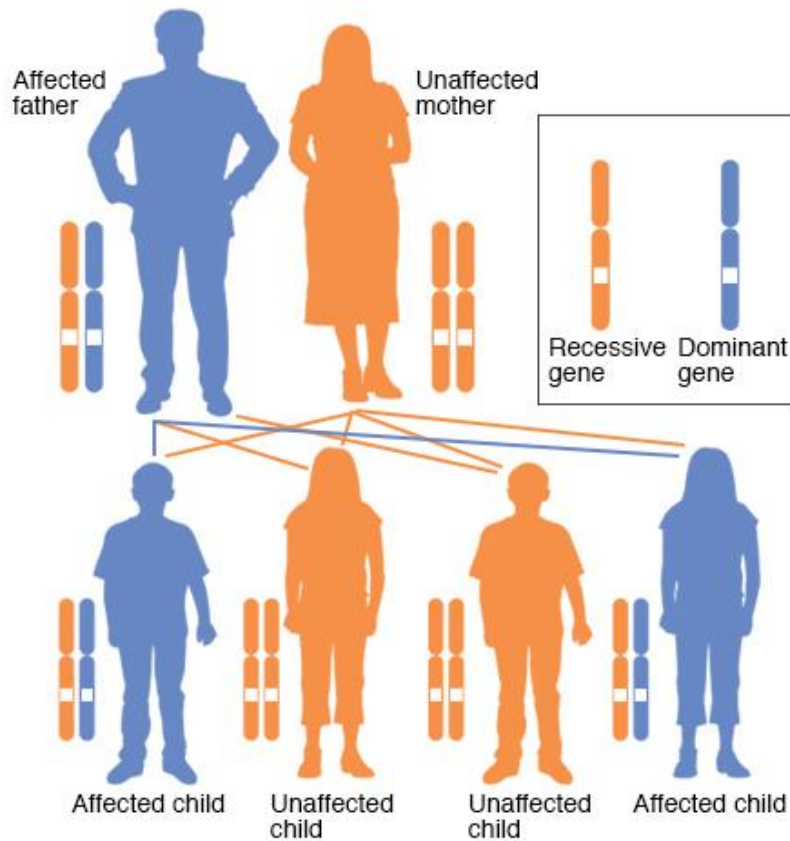
- Las **enfermedades Mendelianas** (controlado por un solo gen defectuoso) son el producto de una mutación ancestral recién que afecta solo los miembros de una familia específica.
 - Podemos categorizar esas enfermedades en 6 categorías:



Autosómico

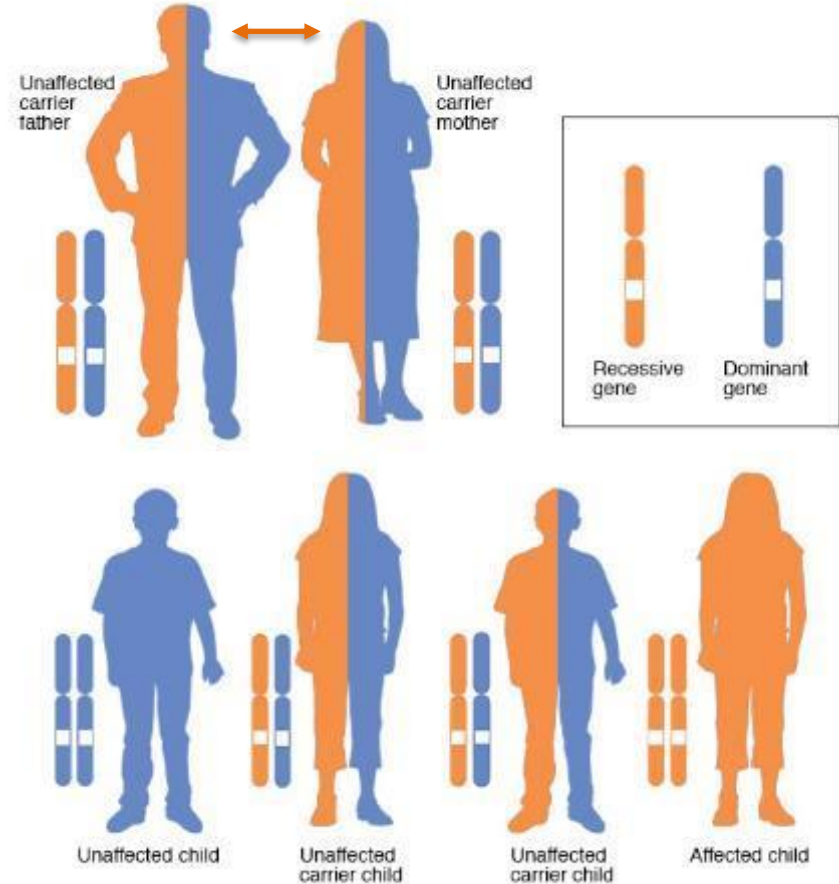
Dominante

Recesivo



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Primos?

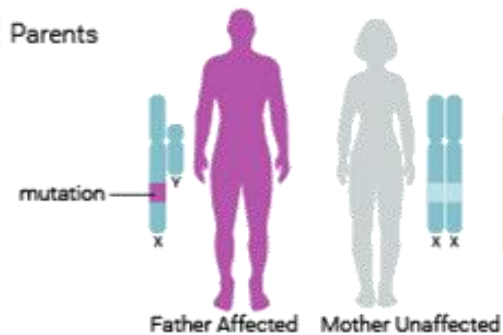


© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

FORMAR

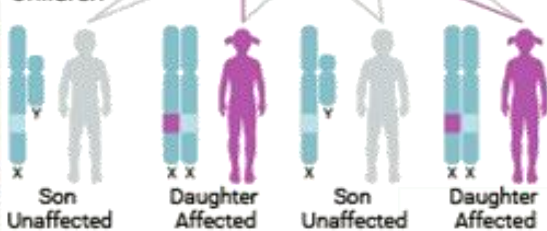
TRANSFORMAR

Parents



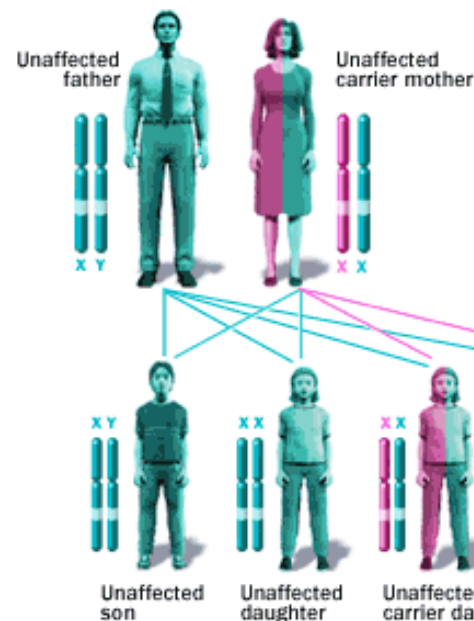
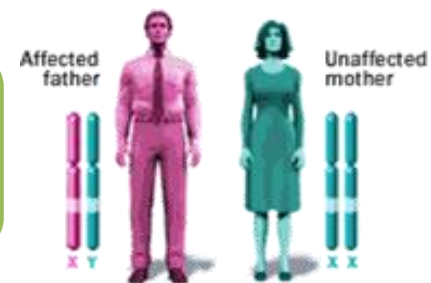
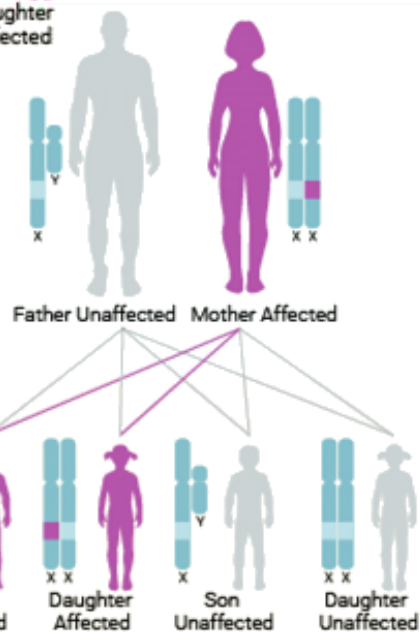
Transmitido por el X

Children



NIH U.S. National Library of Medicine

Dominante



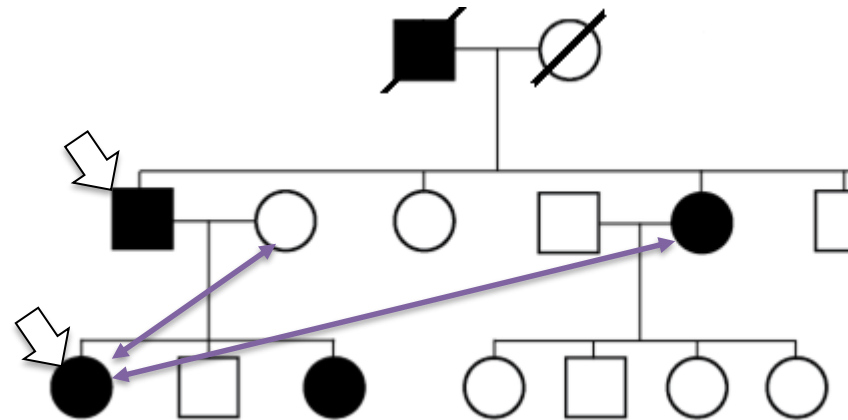
Recesivo

FORMAR

TRANSFORMAR

Un Escenario Simple

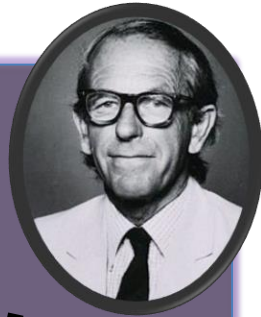
- Tenemos una familia en que hemos identificado una condición rara que se manifestó en algunos miembros.
- El fenotipo de la enfermedad parece seguir un patrón que sugiere una **modo de herencia autosómico dominante**.
 - ¿A quién podemos extraer muestras y secuenciar para buscar esta mutación hipotética?
 - ¿Tiene sentido secuenciar la madre? ¿Por qué?



Secuenciación

Un poco de contexto histórico

Durante aproximadamente 40 años, el **método Sanger** fue la técnica principal para secuenciar el DNA. Originalmente, se necesita más trabajo en laboratorio, pero progresivamente fue automatizado. Este método permitió secuenciar por primera vez los genomas de varios organismos, incluso el del ser humano!



Frederick Sanger

**Muestras
de DNA de
la familia**



**Secuenciación de
alto rendimiento**



1990

**Human Genome
Project (HGP)**
lanzado

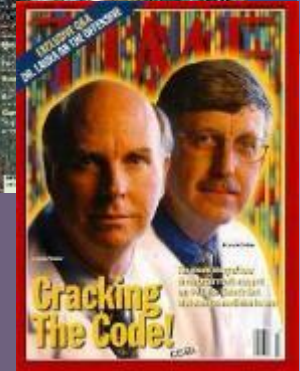
2000

Primer
borrador

2003

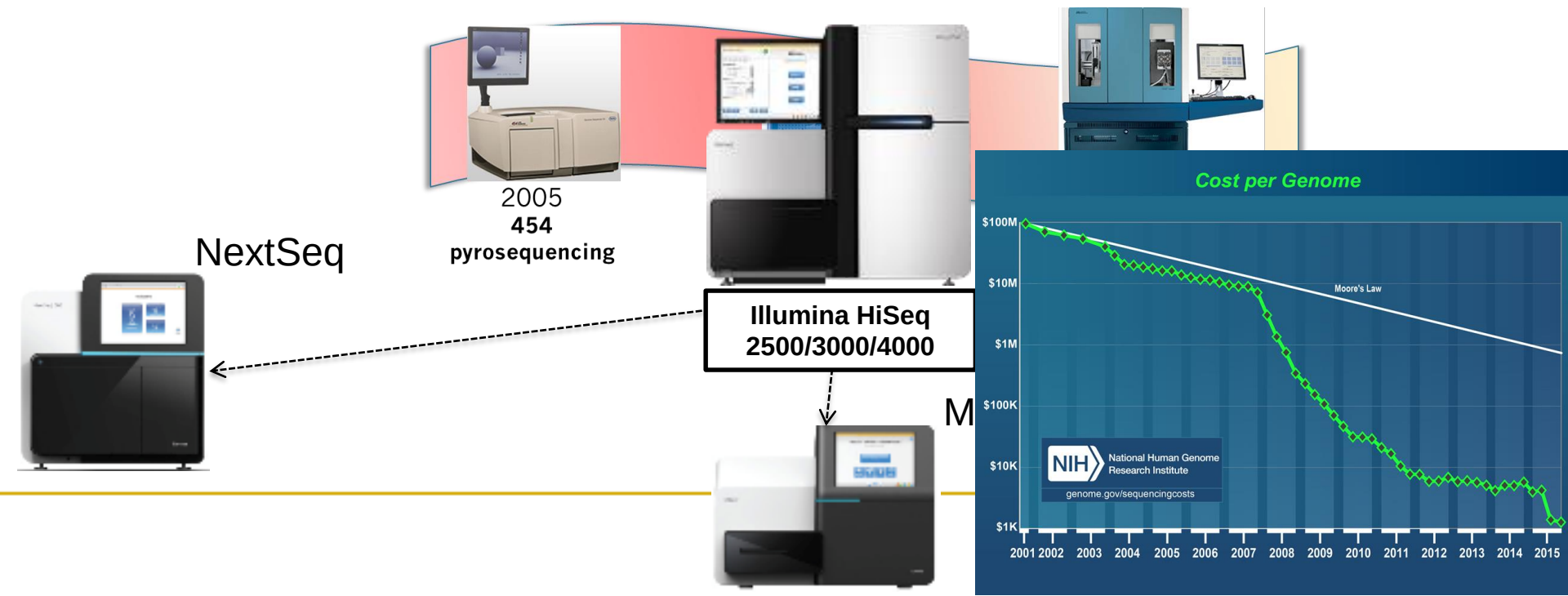
Genoma
completo

13 años para secuenciar un genoma humano.
¿Y ahora?



Secuenciación de Alto Rendimiento

- En 2005 aparecieron nuevas plataformas para secuenciar **masivamente en paralelo** y con **alto rendimiento**. (ing. High-throughput sequencing). Usan **secuenciación-por-síntesis**.



Innovaciones en Secuenciación

- Una nueva ola de plataformas ha traído muchas innovaciones.



En este momento, la tasa de error por bases es **demasiada alta** para algunos proyectos de genómica. Por ejemplo, si necesitas identificar variaciones específicas para relacionarlas con enfermedades, Estas plataformas no van a servir.

En este momento, es mejor usar la plataforma Illumina.



2010 Ion-Torrent

Un sensor eléctrico detecta directamente los **iones de hidrogeno** liberado durante la secuenciación-por-síntesis.

2012 PacBio SMRT system

Secuenciación de **una molécula única** de DNA en tiempo real!

2012 Oxford Nanopore MinION

No secuenciación-por-síntesis! Reemplazado por una estructura Nanopore!

Secuenciación Escopeta (1/2)

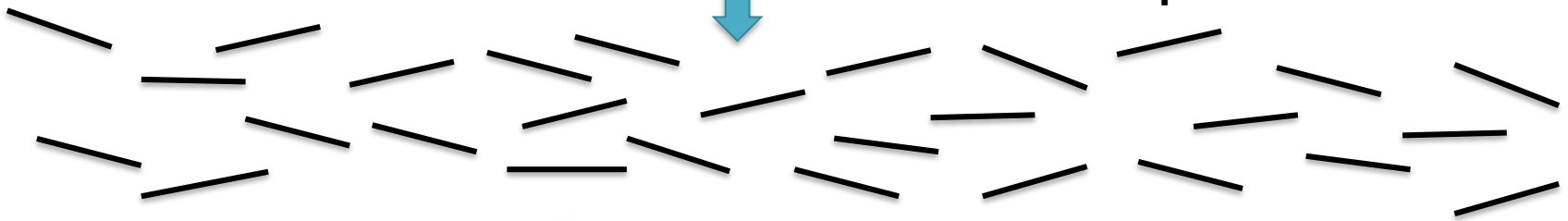
- Limitación de las plataformas nombradas:
Solo puede procesar un fragmento de DNA con un numero limitado de **pares de bases** al vez.
 - ¿Como fue posible secuenciar todo el genoma del ser humano (con su **3 billones** de bases) con el método Sanger? (limitado a 100-1000 pares de bases)
 - ¿Ahora como es posible secuenciar genomas con Illumina con una limita de 75-300 pares de bases?
- **Secuenciación escopeta** (ing. **shotgun**).
 - Pequeños fragmentos de ADN de mismo tamaño se obtiene con el **cizallamiento aleatorio** del ADN genómico. ¡Se pierde el orden de los fragmentos!



Secuencia Escopeta Ilustrado



Cizallamiento '**escopeta**' aleatorio de ADN



Para cada fragmento...



...**amplificación clonal**



Secuenciación

fragmentos están interpretados como
secuencia de bases → **reads**

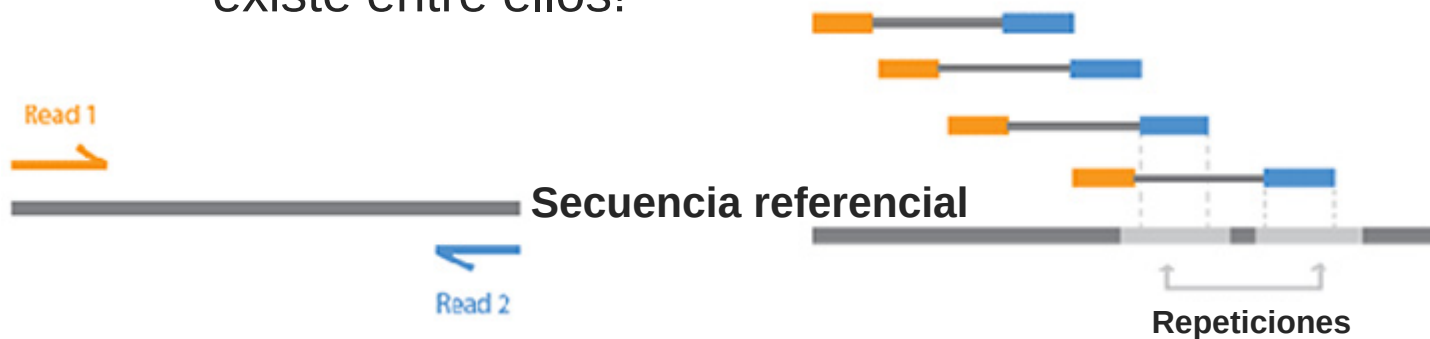
...ATCGGCTGAATGGACACT...
...TTTGCCTAGCTACAGAAG
...
...ATTGATGTATTTTGGCT
...CCCTGATGTCTCAGAGCA
...TAAATGGACACTATCCGC
...
...

Secuenciación Escopeta (2/2)

- La gran mayoría de las técnicas de secuenciación que se usan hoy usan la **secuenciación escopeta**.
 - El tamaño de los fragmentos depende de la plataforma de secuenciación.
- Resolvió un gran problema de rendimiento...
 - ...para crear otro problema, uno de los grandes retos computacionales de la genómica.
 - Dado que el orden de fragmentos **no se conserva**, un gran tema de la genómica es el **re-ensamblaje de secuencias**.

Las Reads 'Paired-End'

- Una **read** se crea a partir del borde de un fragmento.
 - Cuando un par de reads se crea a partir de los dos bordes de un fragmento, hablamos de **paired-end reads**.
 - A diferencia de otras reads, conocemos la distancia que existe entre ellos!

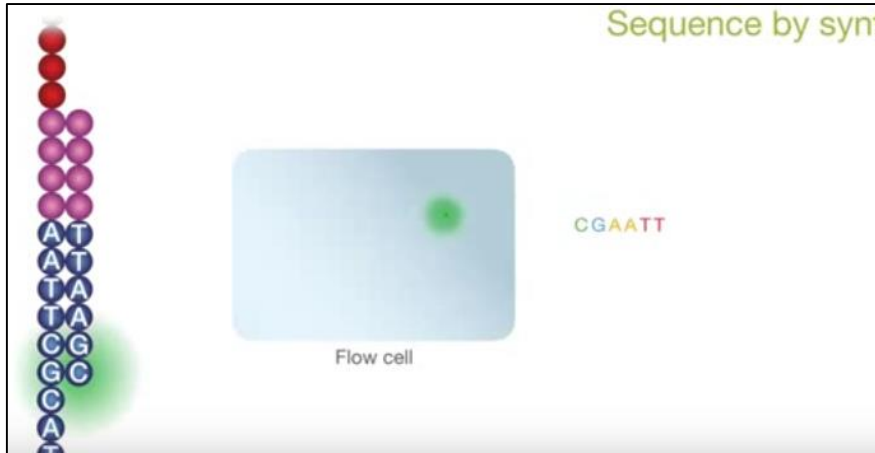


<https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html>

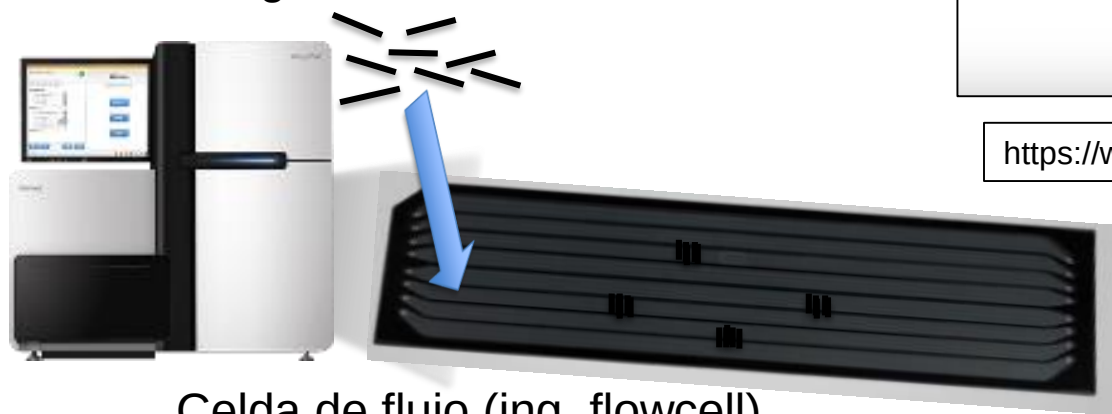
Secuenciación-por-Síntesis (Illumina)

Sequence by synthesis

“masivamente en paralelo y con alto rendimiento”



Fragmentos de DNA



<https://www.youtube.com/watch?v=9YxExTSwgPM>

Celda de flujo (ing. flowcell)

FORMAR

TRANSFORMAR

Genoma o Exoma

- ¿Qué tipo de secuenciación necesitamos para encontrar la mutación en esta familia?
¿Secuenciación entera del genoma? (ing. Whole-Genome Sequencing, WGS)
 - En un mundo casi perfecto, WGS es el elegido obvio. Por lo tanto...
 - Aunque el costo de secuenciación sigue bajando, ya esta **demasiado alto**.
 - Mucho segmentos del genoma **no están anotados**.
 - Si el objetivo es encontrar una única mutación. Buscar el genoma entero es como encontrar una aguja en una pila de agujas.



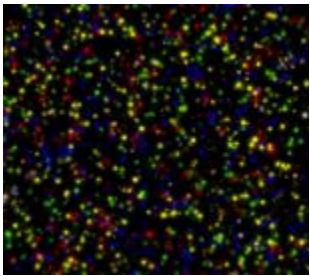
Genoma o Exoma

- ¿Secuenciación entera del exoma? (ing. **Whole-Exome Sequencing, WES**)
 - **WES** proporciona una alternativa barata que focaliza la secuenciación solo sobre las regiones **exómicas** del genoma.
 - El exoma solo cubre ~**1%** del genoma humano, pero contiene todas las **secuencias que codifican por proteínas**.
 - Según estimaciones, estas regiones serían responsables por ~**85% de las mutaciones deletéreas de todo el genoma**.
- Ahora es solo como encontrar una aguja en un pajar !



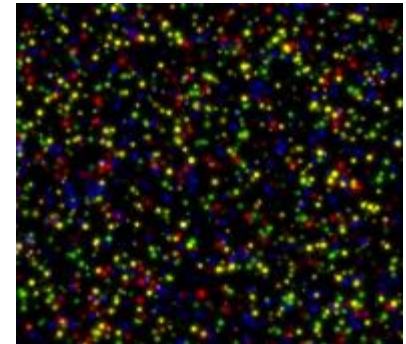
Errores en determinación de base

- El proceso de **secuenciación-por-síntesis** es lejos de ser perfecto, por lo que puede producir señales ambiguas.
 - Hacer una serie de determinación de las bases observadas no es tan simple!!
 - ❖ Las señales que provienen de racimos diferentes pueden mezclarse si están demasiado cerca.
 - ❖ En un racimo, las reacciones bioquímicas pierden su sincronización poco a poco
 - ❖ La intensidad de las señales pueden variar.
- La incertidumbre así creada conduce a la asignación incorrecta de bases.



Phred (para bases)

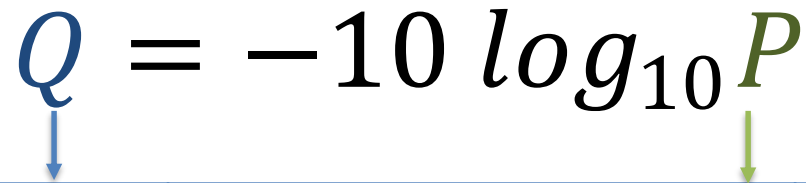
- Cada determinación de base en un 'read' se ve asignada una probabilidad P de que **la determinación es incorrecta**.
 - Esta probabilidad depende de:
 - ✓ La intensidad del señal.
 - ✓ La proximidad de otros racimos.
 - ✓ El ciclo de secuenciación.
- Podemos representar eso en forma logarítmica, un índice de calidad Phred (ing. Phred-quality score):



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

Índice Phred $\Leftrightarrow$ Probabilidad

- La conversión de índice a probabilidad es bastante intuitivo...

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


Índice de calidad Phred	Probabilidad de mala identificación de una base	Exactitud de designación de base
10	1/10	90%
20	1/100	99%
30	1/1 000	99.9%
40	1/10 000	99.99%
50	1/100 000	99.999%

Phred en Formato ASCII

- A menudo, los índices en archivos de secuenciación se presentan en el formato ASCII:

Qual 0.....10.....20.....30.....40
ASCII !"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

- Este tipo de codificación se llama **Phred-33** (por qué '!' corresponde al símbolo ASCII numero 33).

El Formato FASTQ

- Tenemos datos de secuenciación para cada individuo en la forma de **pares de reads**.
 - Típicamente, estos datos están almacenados en un archivo en el formato **FASTQ**.

```
@SN431_1138:5:1101:1176:10390#33/2
CGGGTGTGTGACCTGGGAGAGCTGGTGCATCACGGGCGTATCTTTGCACTGCTGGCTGGTGCGCCCATGGGGAAGCATCCGCGTTGGGTCCCACATTCTC
+
=ED>>=<<ECCD0/GDF;D.IGGDCEA;<H.H=;)H:<;B-HF--;GAEHC:<BEI.;<;8.;99F@<.DD'8@,, 'D6E9;@7GB.H-(7C6',A880
@SN431_1138:5:1101:1177:20812#33/2
AGTCAGTGAGGCCTTGTCGGTAAGGACCTGGTCCTTGAGGCCTTGCCAGTGAGGCCTTGTCAGTAAGGTCTGCTCACTGAGGCCTTGTCAGTAAGGACT
+
=EDCE=FGFGFC=FGFGE.D<GGGGBIFIHGHGD?EHHHF9@EIEBGDGHFHHD<<EIH<GEEGA(;C/DGD;EEFD'EGEH-HAHHHDG,D>EFE?G0
@SN431_1138:09354:5:1101:1178:83518#33/2
TTTATTCTATGTATGAATAGATGCATATTATGTCAATGACTTTCTTGATGAAATAACTATTTTTCTTGTAATCTCATAAAACAGTTTGAGATTATCA
+
>ADD>EBGE>FFFEGBEEADECG:CEHDGHGHGD=EHGHCBEHFIDIGFCF;CF<HF;F<FHHBGC@HBADFEGDG=DGEEEDHGHGHGGDEEAHEGHE
```

Línea 1: ID Read (ID plataforma de secuenciación : numero de carril de la celda de flujo: numero de azulejo en carril : ubicación (x, y) del racimo : miembro (1 o 2) del pare de reads).

Línea 2: Bases del read en el orden en que se secuencian.

Línea 4: Los índices de calidad correspondientes.

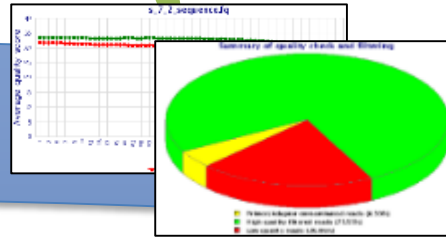
Control de Calidad (QC)

Muestras
de DNA de
la familia



Secuenciación de
alto rendimiento

Control de calidad (QC)



El control de calidad es una etapa **crucial** de cualquier proyecto de secuenciación.

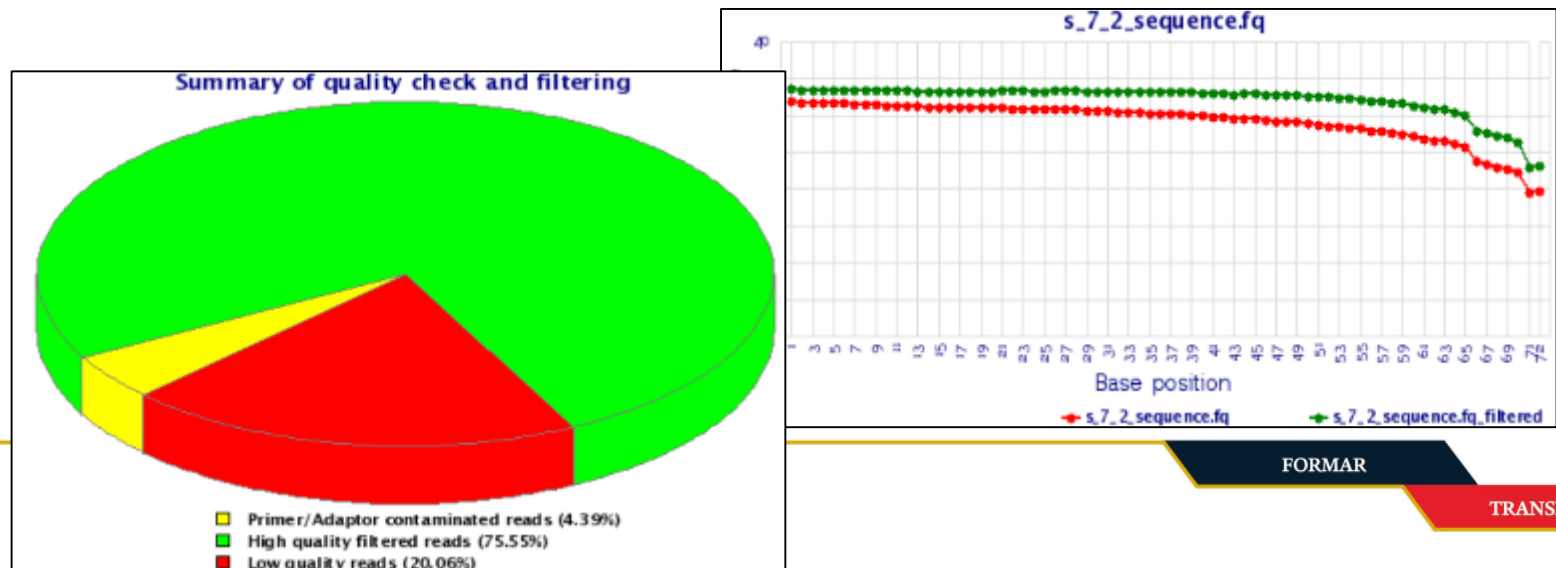
Un poco de control de calidad es importante en cada etapa **pero** el control de datos de **secuencias en bruto** es esencial para identificar problemas que resultan durante el **procesamiento de muestras** y de **secuenciación**.



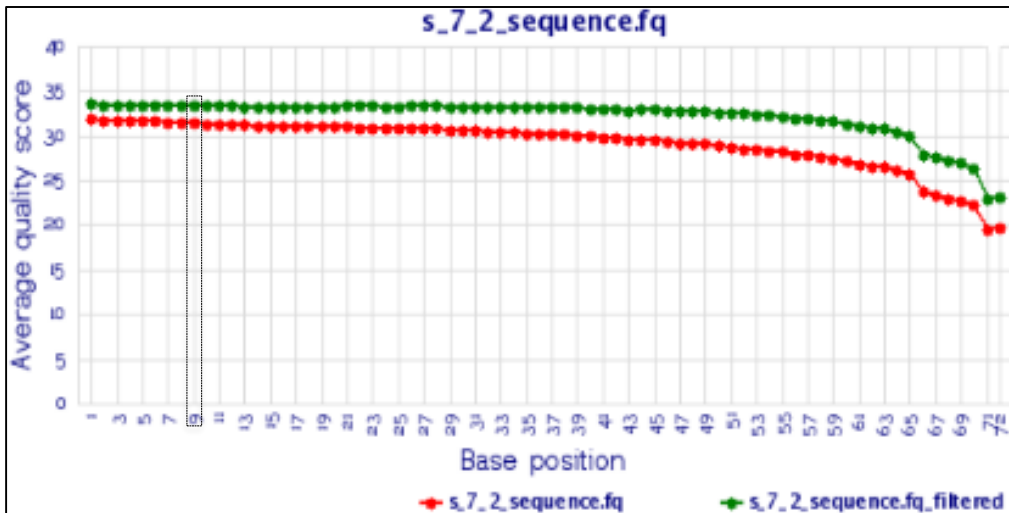
Intentar extraer información significativa de datos de mala calidad nunca es una buena idea. **Basura dentro → Basura fuera.**

Métricas del Control de Calidad

- Con el **control de calidad** (abr. ing. **QC**) Podemos **evaluar** el nivel de integridad de los datos de secuenciación usando varias métricas.
 - Existen herramientas como **NGS QC toolkit** o **FASTQC** (entre otros) que permiten la colección y visualización de métricas.



Ejemplo de métricas: Calidad Por Posición de Base



Como leer:

El índice de calidad (Phred) en una posición de base específica (e.g. 9 de 72) en cada read esta agregado. A partir de ellos **un promedio** se obtiene por esta posición.

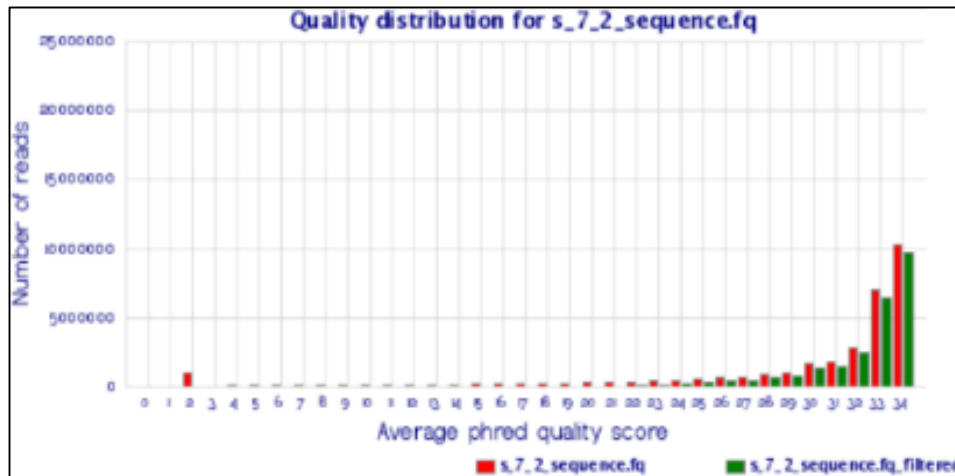


Una caída en calidad debajo del índice Phred promedio **20** señala datos de mala calidad.

Por lo tanto, una caída en las **últimas posiciones** de base no es dramático, es el resultado del perdido progresivo de sincroniza dentro de los ramitos.

Ejemplo de métricas:

Distribución de Índices de Calidad



Como leer:

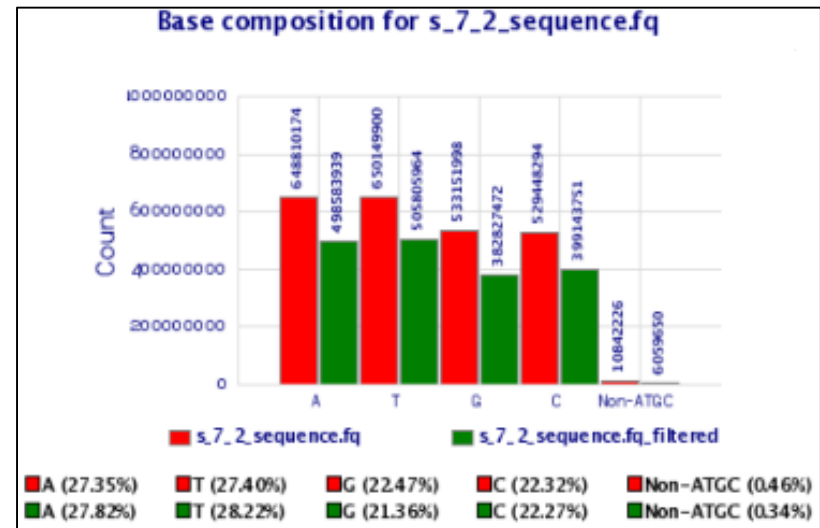
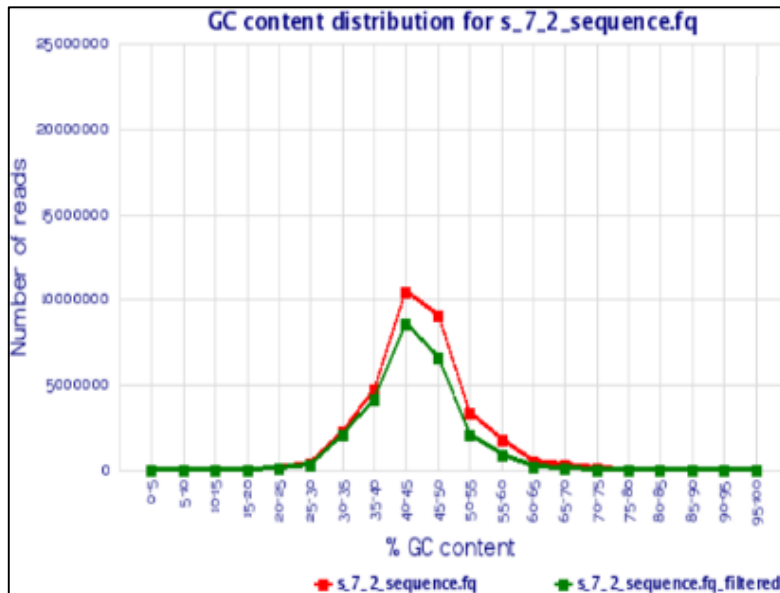
Para cada read, los índices Phred en cada posición de un read están agregados entre ellos para obtener un promedio para el read. Luego, las reads están agrupado por su promedio índice de calidad.



Solo deberíamos ver un solo una punta en la cola de distribución. Si la punta aparece antes o si existen más de uno, eso es problemática. Es importante también prestar atención al Phred promedio al fin de la distribución. No es mismo si se terminara con 34 o 20...

Ejemplo de métricas:

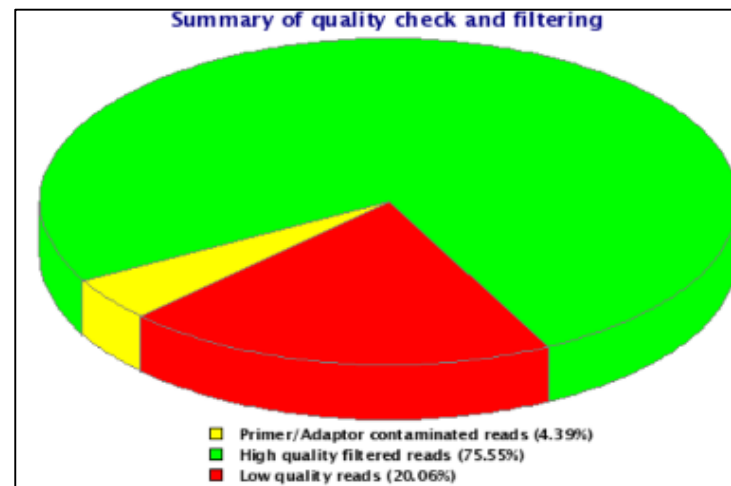
Distribución del contenido o de bases



Un poquito de variación es típico. La punta debe ser cerca del rango 45-55%.

Resumen del Control de Calidad

- Herramientas de control de calidad (e.g. **NGS QC toolkit**) proveen un resumen de QC de los datos para rápidamente identificar problemas.



Esto no

métricas



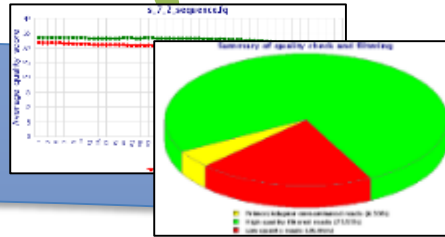
Alineamiento de Secuencias

Muestras de DNA de la familia



Secuenciación de alto rendimiento

Control de calidad (QC)



Alineamiento de secuencias

Secuencia de referencia

...TGAATAGGCGTACTTCATCCATGCGC...
GAATAGGCTACTT
AGGCCTACTACAT
ACTTCATCCATTC
CCTTCATCAATGCG
C

reads

Rec^ordatorio

La secuenciación escopeta **no preserva el orden** de las reads.

Entonces necesitamos un mecanismo para ensamblar las diferente reads en su orden original.

Una solución : Alineamiento de secuencias

FORMAR

TRANSFORMAR

Ensamblar Secuencias

- Sabemos el orden de las bases en un read, pero no el orden de las reads en si mismos!
- Existe un nivel de **redundancia** entre las reads que podemos usar para el ensamblaje

```
          CCATCGGATG...  
...ATTGAGAGCACACCATCGGATG...  
...ATTGAGAGCATCACC  
          ATCACCATCGGATG...
```

- Pero, debido a las limites de las plataformas actuales, **las reads son cortitos** (ej. para Illumina ~150-300 pares de bases). Además, comparar cada read con cada otro seria **muy costoso computacionalmente**, en termino de **tiempo** y de **memoria**.
- ¿Qué alternativa tenemos para ensamblar una secuencia entonces?

Resecuenciación (1/2)

- A diferencia de las primeras secuencias del genoma humano (que fue ensamblado *de novo*), cada nueva secuencia puede ser alineado contra una **referencia genómica** existente. Cada read con la referencia.
 - Este procedimiento originalmente se llamaba **resecuenciación**
 - i.e. el genoma humano se secuencia otra vez.
 - Ahora es más común hablar simplemente de **alineamiento de secuenciación**.

El Genoma Humano de Referencia

- El primer ensamblaje del **genoma humano** ha sido actualizado **20 veces** (!).
 - Las últimas **5 versiones** del genoma humano están disponibles a través del portal **UCSC** o **GRC** (ing. Genome Reference Consortium).
 - Cada genoma es un **compuesto** de varios **individuos anónimos**.
 - El genoma resultante es **una secuencia de consenso**.
 - E.j. el ultimo genoma, **GRCh38/hg38** fue construido a partir de la secuencias de 10 contribuidores anónimos.
 - Esta secuencia de consenso sirve de **referencia** para todos los seres humanos.

Resecuenciación (2/2)

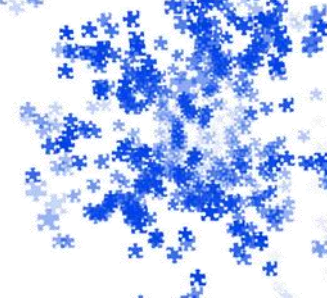
- La resecuenciación es posible debido a las **grande similitudes** que existen entre cada ser humano.
 - La secuencia genómica de dos individuos (non-gemelos) es idéntica al 99.5%.
 - Ciertos fragmentos de una secuencia humana pueden mapear perfectamente contra el genoma humano referencial.
- Pero...
- ★ **Errores de secuenciación** van a complicar el alineamiento.
 - ★ También existen regiones problemáticas debido a la presencia de **repeticiones** o la presencia de **secuencias homologas** en otra parte del genoma.

Desafíos del Alineamiento

- El alineamiento no es tan simple como parece a primera vista.
 - Una **gran cantidad** de reads (todavía cortitos) deben ser alineados contra la referencia **entera** del genoma humano.
 - De vez y cuando, existe una disparidad de una o dos bases en una read contra la referencia. Esto se debe a una mezcla de **errores** que surgieron durante el llamado de bases y **variación real**, la que existe en cada individuo (SNVs y indels).

Alineamiento es como un puzle con demasiadas piezas pequeñas, con muchas piezas que son idénticas y con piezas que no encajan perfectamente juntas!

Tipos de disparidades



Disparidad en un solo nucleótido:

(Fuente: **error** o **variación** en individuo)

Ref ATGATGCCATGACTGACCCTGAT

Read ATGATGCCATGACTGAC**A**CTGAT

Inserción (fuente: error o inserción real)

Ref TCCATGTGTGACTA*****CACC

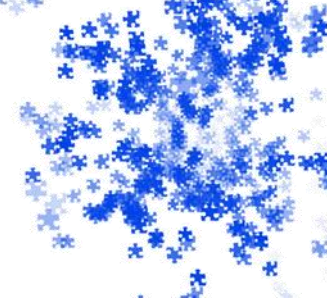
Read TCCATGTGTGACTATTTGTCACC

Delección (fuente: error o delección real)

Ref AAAC**T**TAGTGCAACAGTGACGAG

Read AAAC**AGTGCAACAGTGACGAG

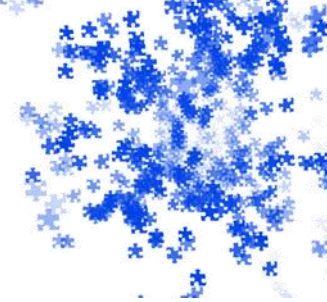
¿Como hacer el alineamiento?



- Podemos imaginar comparar **cada parte del genoma** con **cada fragmento** uno a uno hasta que obtenemos el mejor alineamiento posible...
 - Debido a los varios desafíos del alineamiento, este enfoque será **costoso computacionalmente**.
 - Irrealizable en un tiempo adecuado.
 - En lugar de obtener el alineamiento perfecto, podemos obtener un alineamiento bueno con menos costo computacional usando **heurísticas**.

No dejes que la perfección sea el enemigo del bien !

Heurísticas



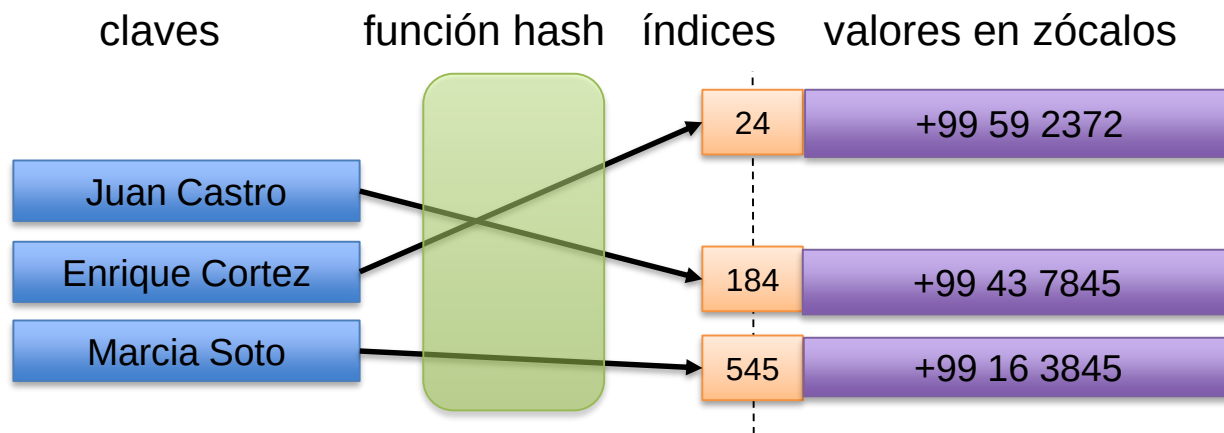
- La **heurística** es muy común en la bioinformática (y en la resolución de problemas en informática en general).
- **Heurísticas** corresponde a **algoritmos** para resolver un problema que **no siempre va a encontrar la solución optima**, sino una solución **bastante precisa** para cumplir con su objetivo.
 - Para el alineamiento de secuencias, significa simplificar nuestros datos (las reads, la referencia o ambos) en nuevos datos que **pierden en especificidad para ganar en eficiencia**. Esos datos podemos manejarlos en **estructuras de datos** que facilitan el alineamiento.

Estructuras Alineamiento

- Existen dos categorías de herramientas para **alineamiento rápido** (ing. **fast-aligners**) definido por el tipo de **estructura de datos** que usan para acceder a secuencias.
 - **Tabla hash**
 - Lo mismo que se encuentra en clases de programación.
 - **Trie**
 - de la palabra 'retrieval' (esp. recuperación) para parecer como la palabra 'tree'. Es una estructura de tipo árbol.
 - Vamos a ver en breve estos dos tipos de estructuras.

Rec😊rdatorio: Tabla Hash

- Una **tabla hash** asocia **claves** con **valores en un array de zócalos** (ing. **buckets**) mediante una función hash.
- La función hash **transforma la clave en la ubicación** en la tabla hash donde se encuentra la valor.

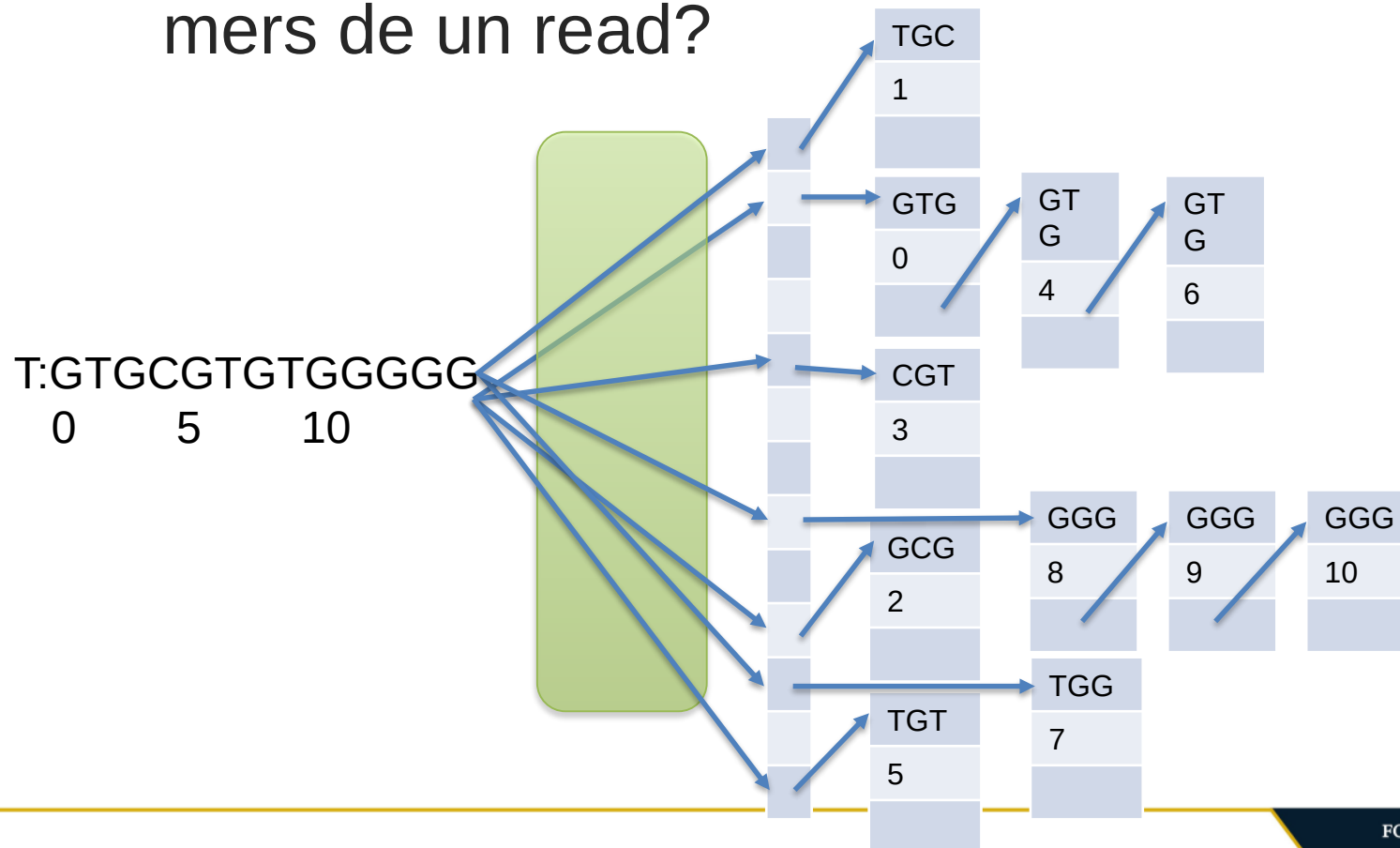


Simplificar las Reads con los *K*-mers

- Podemos quebrar nuestras reads en subconjuntos, lo que se llaman ***k*-mer**.
 - El *k* corresponde al tamaño de cada subconjunto.
 - Ej. Los 3-mers de ATGAGAGCAA son...
ATG TGA **GAG** AGA AGC GCA CAA
- ¿Cuál es la ventaja alinear los ***k*-mers** de las reads y no las reads en sí?
 - Existen menos *k*-mer posible que reads debido al alfabeto limitado de secuencias (i.e. A,T,G,C).
 - Ej. GAG aparece dos veces en ATGAGAGCAA.

Tabla Hash Para K -mers

- ¿Como se rellena una tabla hash con los k -mers de un read?

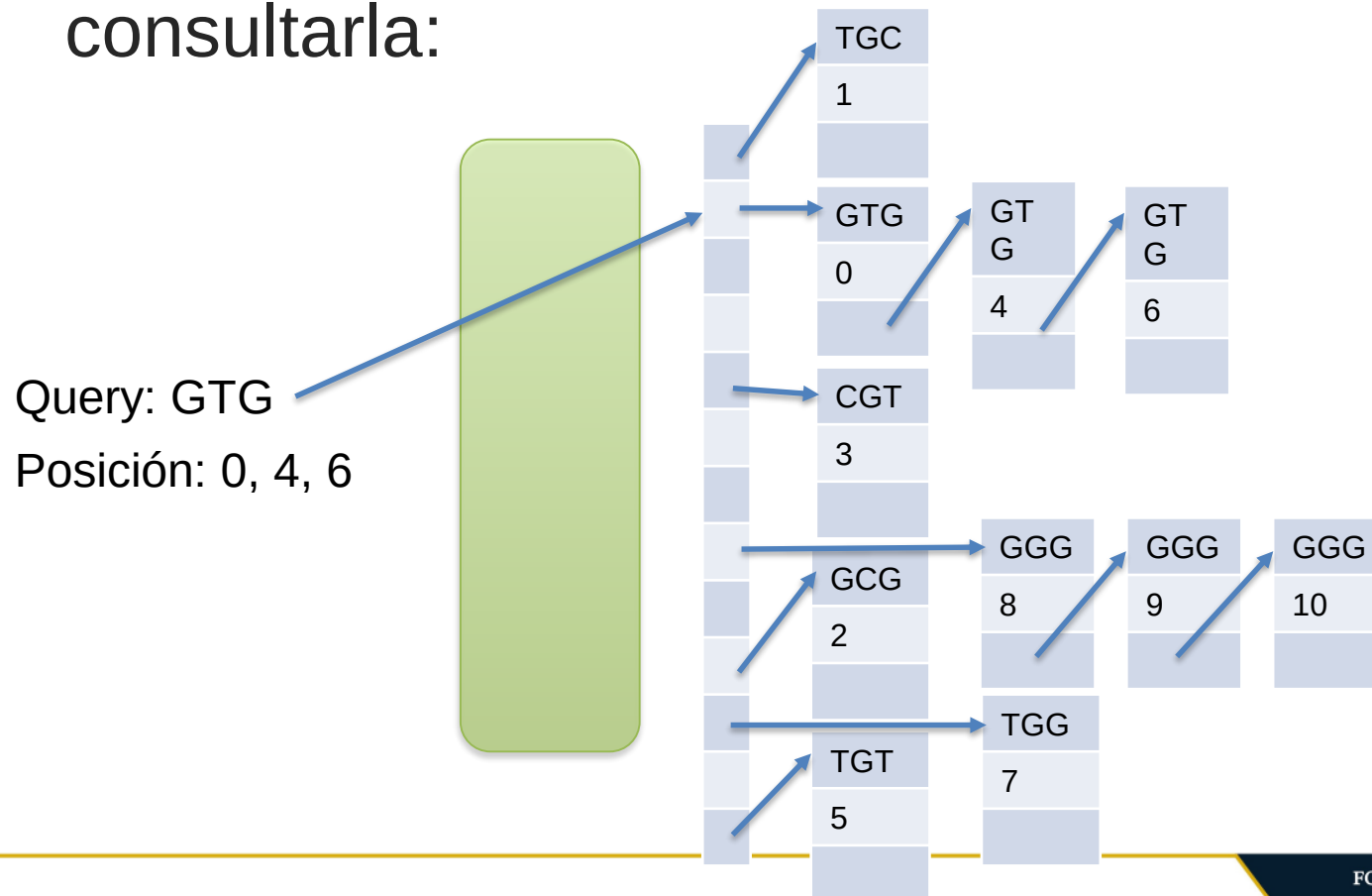


FORMAR

TRANSFORMAR

Tabla Hash Para *K*-mers

- La tabla hash esta rellena, ahora podemos consultarla:



FORMAR

TRANSFORMAR

Trie

- Es una **estructura tipo árbol**.
 - Corresponde al **árbol más pequeño** que cumple con las condiciones siguientes:
 - A cada arista se le designa un carácter de un **alfabeto** definido Σ .
 - Para secuencias: $\Sigma = A, T, G, C$ y $\$$ (=representa el fin de la secuencia)
 - **Cada nodo** tiene a los menos una **vértice saliente** marcado con x , con $x \in \Sigma$.
 - Cada **clave** es deletreada a través de un camino a partir de la raíz.

Trie (de sufijos)

- Otra vez, se trata de obtener **valores** a partir de **claves**.
 - En este caso, las claves corresponden a **todos los pre/sufijos posible** a partir de una secuencia.
 - Las valores almacenadas en los **nodos hojas** corresponden a la **posición de cada pre/sufijo** en la secuencia original.
- Vamos a ilustrar esto con un trie de **sufijos** por la siguiente secuencia:
GTCCGAAGCTCCGG\$

Sufijos de una Clave

Secuencia original

GTCCGAAGCTCCGG\$

TCCGAAGCTCCGG\$

CCGAAGCTCCGG\$

CGAAGCTCCGG\$

GAAGCTCCGG\$

AAGCTCCGG\$

AGCTCCGG\$

GCTCCGG\$

CTCCGG\$

TCCGG\$

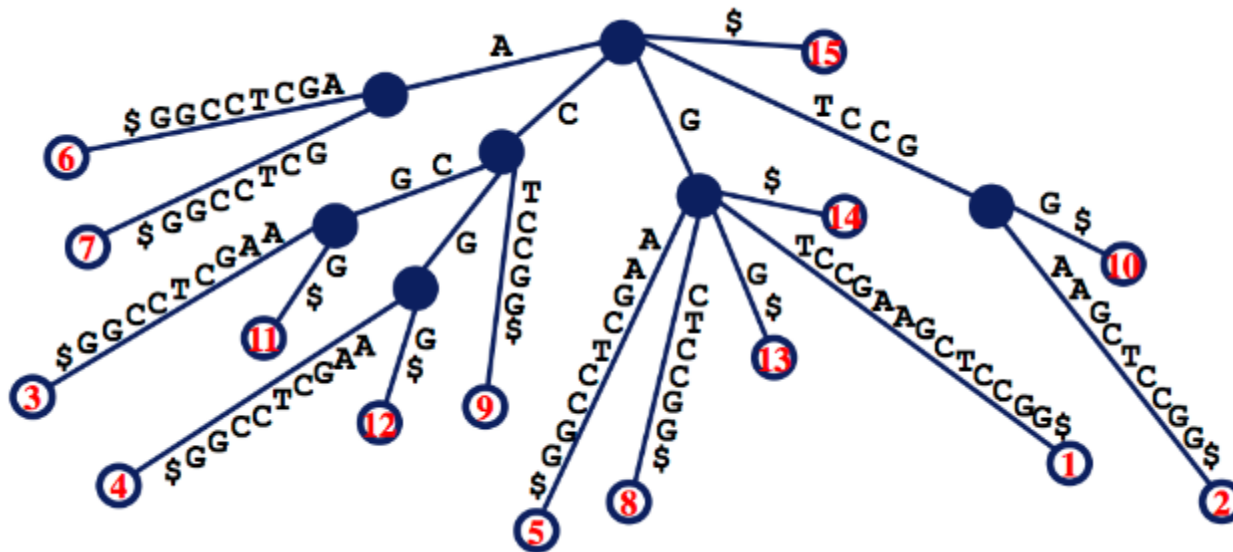
CCGG\$

CGG\$

GG\$

G\$

\$



Los últimos nodos tienen el índice a partir de que comienzo el sufijo representado.

Phred (para alineamiento)

- No todos los alineamientos son de igual calidad debido a la existencia de
 - **Disparidades** en ciertas bases.
 - **incertitud en el llamado** de bases.
- Otro vez, usamos el índice Phred para poner un valor sobre la incertidumbre de un alineamiento.
 - La calidad se aplica a **todo el read** y no a bases individuales.
 - El **llamado de variantes** (que vamos a ver siguiente) se hace con el apoyo de los dos índices Phred.

El Formato SAM

- El formato privilegiado para los alineamientos de secuencias es el formato SAM (ing. Sequence alignment/Mapping format).
 - Los reads reciben coordenadas relativas a la secuencia de referencia utilizada en el alineamiento.
 - Es un archivo legible para ojos humanos.
 - Se puede abrir con el programa SAMtools, usando el comando VIEW (con parámetros).

Ejemplos de SAM

- Ejemplo de la documentación.

```
@HD VN:1.5 SO:coordinate } Encabezamiento (otro posible @RG, @PG, @CO)
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

- En realidad es más como...

```
HWI-ST508_0109:8:2103:19403:137111#ATCACG 83 chr1 16234 255 100M = 16155 -179 T
TGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTG #####
#A:AABFGB;GGGGGGEDBACCDE5>?<@>DE<?D?FCBFEEBDBFDFFFC>@>CDDADD>FDFFCECEEDGGFGE GEGGGGGGGGEGGF NM:i:0 NH:i:1
HWI-ST508_0109:7:1204:3497:194785#ATCACG 163 chr1 16237 255 100M = 16357 220 C
ACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAG DD@D=DEEE
E@GGEEGGFDF<GD@CEEEEEG=FFGFBFBFHHGHDEGGF@EEEBD>=>B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC##### NM:i:2 NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG 355 chr1 16241 3 100M = 16337 196 C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG HHHHFHHHH
HCHHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHHHHFHHHHHHHHHHHAFE?FCFFFFHEHDFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC NM:i:2 NH:i:2 C
C:Z:chr15 CP:i:102514823 HI:i:0
```

SAM format

- Cada entrada contiene lo siguiente:

Primera línea

- ID del read
- Indicador bitwise
- Cromosoma (Nombre referencia)
- Position la más a la izquierda (1-base)
- Calidad del alineamiento
- El string CIGAR
- Nombre referencia siguiente (símbolos)
- Position del otro read en el pare
- Tamaño de la plantilla observado

Otras líneas

- Bases del read ordenados.
- Índices de calidad de los bases ordenados.

```
HWI-ST508_0109:8:2103:19403:137111#ATCACG      83      chr1    16234   255      100M    =      16155   -179    T
TGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTG      #####
#A:AABFGB;GGGGGGEDBACCCDE5>?<@>DE<?D?FCBFEEBDBFDFFFC>@>CDDADD>FDFFCECEEDGGFEGEGEGGGGGGGEGGF      NM:i:0  NH:i:1
HWI-ST508_0109:7:1204:3497:194785#ATCACG      163      chr1    16237   255      100M    =      16357    220     C
ACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAG      DD@D=DEEE
E@GGEEGGFDF<GD@CEEEEEEG=FFGFBBFBFHGHGDEGGF@EEEBD>>=B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC#####      NM:i:2  NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG      355      chr1    16241    3      100M    =      16337    196     C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG      HHHHFHHHH
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFEFEFEEGEGFGHHH?GDCFGGHHHF?FCGGC      NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823      HI:i:0
```

Indicador BITWISE

- El indicador almacena información sobre atributos de un read en la forma de un bit.

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

- El indicador final es una combinación de todos los atributos que corresponden.
 - E.g. 1033 es la combinación de 1, 8 and 1024.

String CIGAR

- Necesitamos una nueva estructura para representar **inserciones** y **deleciones** relativo a la referencia.
- Un **string CIGAR** describe como diferente bases se alinean a la secuencia de referencia.
 - ¿Como interpretar **100M**?
 - significa que cada de los 100 bases en el read se alinean a la referencia*.

➤ ¿Como interpretar **40M5I30M2D25M**?

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTA*****CACCACATGCTGGATAGGTGCCCGTGAAACTTAGTGCAACAGTG
CACGAGATGAGGAGTG

ATGATGCCATGACTGACCCTGATGGTCCATGTGTGACTATTTGTACCACATGCTGTATAGGTGCCCGTGAAAC**AGTGCAACAGTG
CACGAGATGAGGAGTG

M – alineamiento , I – inserción, D – deleción...entre otros.

*No significa necesariamente que corresponden! Podemos tener un A alineado con un C!

Comprimir SAM en un BAM o CRAM

- Gran proyectos involucrando muchos datos pueden rápidamente ocupar espacio en esta etapa.
 - **BAM** ofrece una compresión binaria del SAM.
 - La mayoría de los herramientas que leen SAM también pueden leer **BAM**. Viene con **alta recomendación**!
 - **CRAM** es una innovación mas recién. Usa la secuencia de referencia como clave para comprimir, guardando en el archivo solo la información que se desvía de la referencia.



No debe perder el archivo que contiene la secuencia de referencia!
No puede descomprimir con otro referencia non-idéntica.

Antes del Llamado de Variantes

- Usando herramientas como **Picard** o **SAMtools**, podemos tomar medidas para simplificar el llamado de variantes.
 - **Comprimir** archivos (SAM>BAM>CRAM).
 - **Ordenar** y **indexar** archivos para un acceso más rápido.
 - **Eliminar** secuencias problemáticas.
 - Eliminar el “soft clipping”.
 - Reads con calidad de alineamiento 0.
 - Duplicados ópticos o debido al PCR.

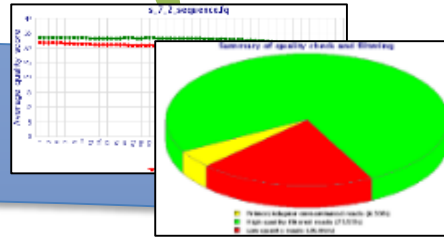
Llamado de Variantes y Genotipificación

Muestras de DNA de la familia



Secuenciación de alto rendimiento

Control de calidad (QC)



Aunque hemos alineado las reads a una referencia basado en los similitudes que existen en entre ambos conjuntos, lo que nos interesa particular son las disparidad a la referencia, es decir las **variantes** propias a los individuos

Alineamiento de secuencias

Secuencia de referencia

...TGAATAGGCGTACTTCATCCATGCGC...

GAATAGGCTACTT

AGGCCTACTACAT

ACTTCATCCATTC

GCTTCATCAATGCG

C

reads

Secuencia de referencia

...TGAATAGGCGTACTTCATCCATGCGC...

ATAGGCGTACTTC

ATAGGCGTACTTCA

AGGCCTACTACATC

GCCTACTTCATCCA

CTACTTCAGCCATG

Llamado de variantes/Genotipificación

FORMAR

TRANSFORMAR

Llamado de Variantes

- Existen reads alineados con **disparidad** relacionada al genoma humano referencial.
 - Disparidades pueden corresponder a **errores** ✖ o a **variantes reales** ✔
 - El **llamado de variantes** (ing. **variant calling**) es el arte de distinguir errores y variantes reales.
 - Una de estas variantes puede ser la causa de una enfermedad Mendeliana (la que buscamos)!
 - Pero no queremos demasiadas candidatas sino las variantes reales se pierden en el mezcla!



FORMAR

TRANSFORMAR

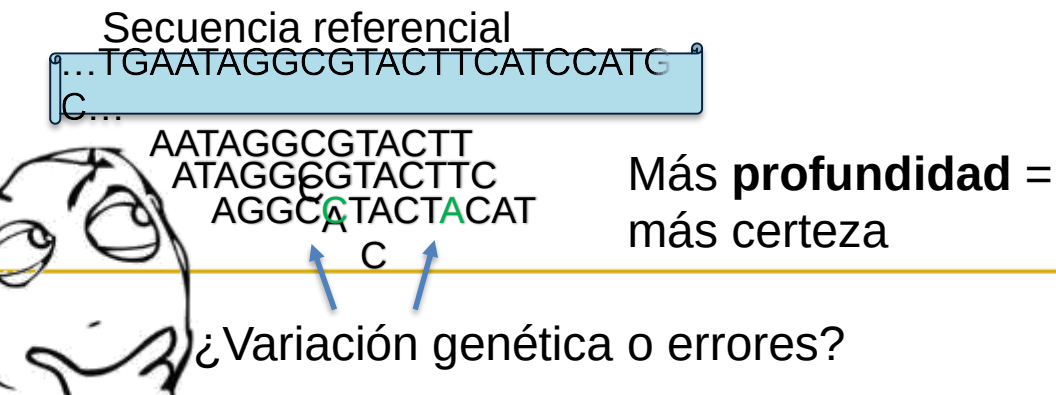
...y Genotipificación

- El **llamado de variantes** es también el proceso con el cual podemos determinar el **genotipo** de cada miembro de una familia secuenciada para cada locus en cual existe una variante.
 - Eso se llama genotipificación (ing. **genotyping**).
 - Típicamente, para un individuo se representa con:
 - 0/0: homocigoto seguido la referencia (ing. wildtype).
 - 0/1: heterocigoto con un variante.
 - 1/1: homocigoto para la variante.

(Con la genotipificación de más gente, podemos ver 1/2, 2/3, 3/3...)

Determinar la Existencia de SNVs

- No existe solo **superposición** entre referencia y reads sino entre las reads en si.
- La cantidad de reads que cubre una posición de base específica se llama la **cobertura** o **profundidad** (ing. **coverage** o **read depth**).
 - Determinar si una disparidad en una base es un error o una variante es más fácil con **cobertura alta**.
 - El **consenso entre reads** proporciona la primera gran prueba para la **existencia o no** de una variante.



Llamado de Variante Básico

- Para presentar mejor el **llamado de variantes**, podemos imaginar una manera de llamar variantes y genotipos que se base solo en las variantes llamadas y la cobertura.
 - En este escenario la frecuencia con cual aparece una disparidad en una posición específica va a determinar si:
 - Se considera verdadero o falso
 - Forma parte de un genotipo heterocigoto o homocigoto.

Un Llamado de Variantes Básico (Het)

En un locus, tenemos dos **alelos**:
2xG (ref) y 3xC (var)

En este escenario, **C** sería detectado como **una variante**. El genotipo sería detectado como heterocigoto (**0/1**) porque la **frecuencia alélica** en este locus es casi 0.5. (A diferencia de **A** que parece ser un error).

Secuencia referencial
...TGAATAGGCGTACTTCATCCATG
C...
AATAGGCGTACTTC
ATAGGCGTACTTCA
AGGCCTACTACATC
GCCTACTTTCATCCA
CTACTTCAGCCATG

En teoría, si un individuo es **heterocigoto** en un **locus**, deberíamos ver (más o menos) el mismo número de reads con base de referencia y con variante.

Un Llamado de Variantes Básico (Hom)

Secuencia referencial
...TGAATAGGCGTACTTTCATCCATG
C...
AATAGGCCTACTTC
ATAGGCCTACTTCA
AGGCCTACTACATC
GCCTACTTTCATCCA
CTACTTCAGCCATG

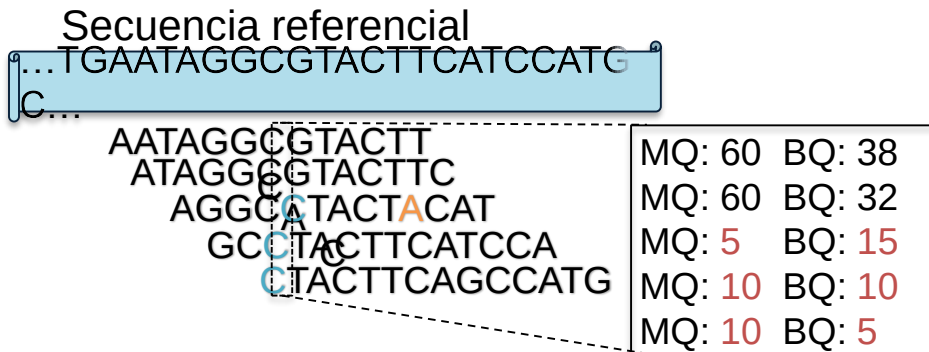
En un locus, vemos:
5xC (var)

En este escenario, **C** sería detectado como **una variante**. El genotipo sería detectado como **homocigoto (1/1)** porque la **frecuencia alélica** en este locus para la variante es 1.

En teoría, si un individuo es **homocigoto** en un **locus**, deberíamos ver solo esta variante en todos los reads. La presencia de unos pocos reads con otro alelo, se considera como errores.

Con los Indices Phred

- ¿Y si la detección de la base **C** fue debido a **errores sistemáticos** en...
 - El alineado de ciertos reads?
 - El llamado de base?
- Los **índices Phred** pueden contar una historia diferente que la cobertura...



Ahora **C** no parece realmente ser un variante. Es un error.

Llamado de Variantes con Probabilidades

- Con el tiempo, los llamadores de variantes han aumentado constantemente en **sofisticación**.
 - Los últimos llamadores usan **modelos probabilísticos** para asignar un **índice de confianza** (ing. confidence score) a las variantes y los genotipos llamados.
 - En particular, llamadores de variantes como **SAMtools**, **FreeBayes** or **GATK HaplotypeCaller** usan de manera extensiva el **teorema de Bayes**.
 - El teorema forma el núcleo de lo que se llama **estadísticas Bayesianas**.

El Teorema de Bayes

- La ecuación que describe el **teorema de Bayes**:


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$: **probabilidad posterior**.

- Probabilidad de un **evento A** dado que **B** es verdadero.

$P(B|A)$: **Función de verosimilitud** (ing. **likelihood function**) de A (¿los datos se concuerdan con el evento A?).

$P(A)$: **probabilidad anterior** del evento A.

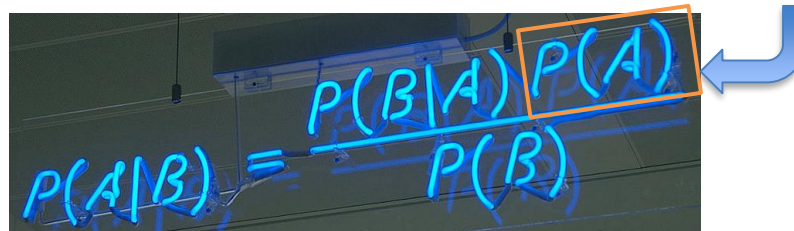
$P(B)$: el factor de **reequilibrio**.

- para obtener un resultado entre 0 y 1
- Se derive así: $P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$

La Idea Básica en Llamado de Variantes

- Supongamos que estamos intentando determinar si un nucleótido encontrado en un locus representa realmente un variante.
 - Sabemos que la probabilidad de un variante en un locus es $P(A)=0.001$. (eso será nuestro **probabilidad anterior**)

Nuestra información originalmente determinada de manera **empírica**.

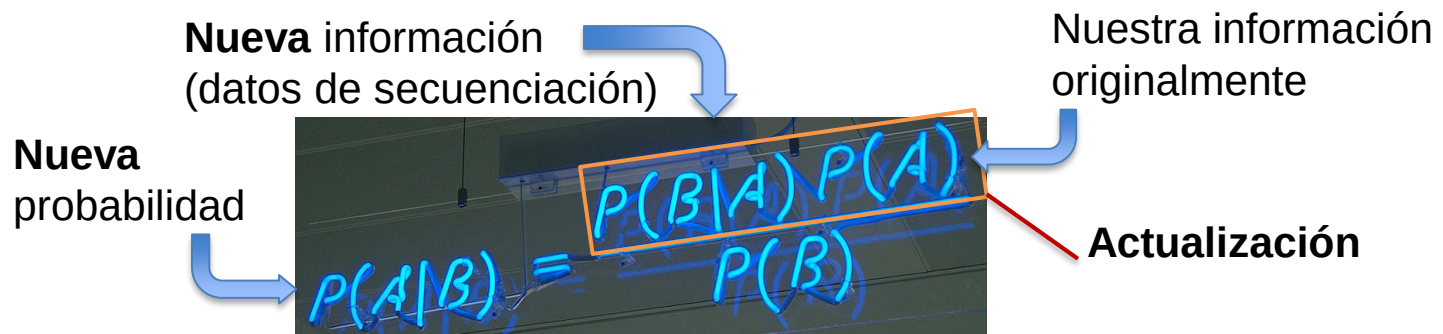


A photograph of a chalkboard with the formula $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ written in blue chalk. An orange rectangular box highlights the term $P(A)$ in the numerator. A blue arrow points from the text above to this box.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

La Idea Básica en Llamado de Variantes

- Supongamos que estamos intentando determinar si un nucleótido encontrado en un locus representa realmente un variante.
 - Con el teorema de Bayes podemos actualizar nuestra probabilidad original con **datos de secuenciación** representados por $P(B|A)$.



El Teorema de Bayes Aplicado a la Genotipacion

- Hay muchas maneras de aplicar el teorema al llamado de variantes y a la genotipacion.
 - E.g., para determinar la probabilidad que un individuo es heterocigoto (A/B) en un locus específico.

$$P(G = AB|datos) = \frac{P(datos | G=AB)P(G=AB)}{P(datos)} = \frac{P(datos | G=AB)P(G=AB)}{\sum_{G \in (AB, AA, BB)} P(datos|G)P(G)}$$

- La probabilidad posterior podría ser comparada a las probabilidades de los otros genotipos.
 - El genotipo que **maximiza la probabilidad** puede ser llamada.

Ejemplo del Variant Call Format File

- La salida de la mayoría de los llamadores de variantes es un archivo **VCF**.
 - El encabezamiento proporciona referencias para cada métrica usada por el llamador de variante.

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G

Cada línea representa

un dón una cali prop Las ultimas columnas proporcionan predicciones para el genotipo de cada individuo así que información específica a cada individuo.

FORMAT	NA000001	NA000002	NA000003
GT:GQ:DP:HQ	0/0:48:1:51,51	1/0:48:8:51,51	1/1:43:5:.,.
GT:GQ:DP:HQ	0/0:49:3:58,50	0/1:3:5:65,3	0/0:41:3
GT:GQ:DP:HQ	1/2:21:6:23,27	2/1:2:0:18,2	2/2:35:4
GT:GQ:DP:HQ	0/0:54:7:56,60	0/0:48:4:51,51	0/0:61:2
GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

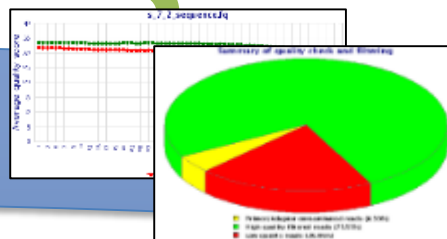
Llamado de Variantes y Genotipificación

Muestras de DNA de la familia



Secuenciación de alto rendimiento

Control de calidad (QC)



Alineamiento de secuencias

Secuencia de referencia

...TGAATAGGCGTACTTCATCCATGCGC
...GAATAGGCTACTT
AGGCCTACTACAT
ACTTCATCCATTC
CCTTCATCAATGCG
C

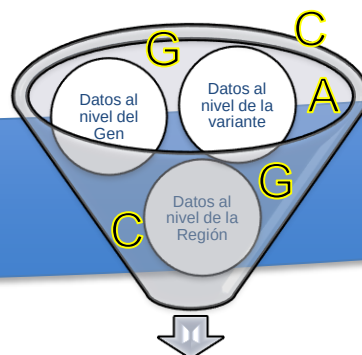
reads

Secuencia de referencia

...TGAATAGGCGTACTTCATCCATGCGC
...TAGGCGTACTTC
TAGGCGTACTTCA
AGGCCTACTACATC
GCCTACTTCATCCA
CTACTTCAGCCATG

Llamado de variantes/Genotipificación

Anotación y Filtración

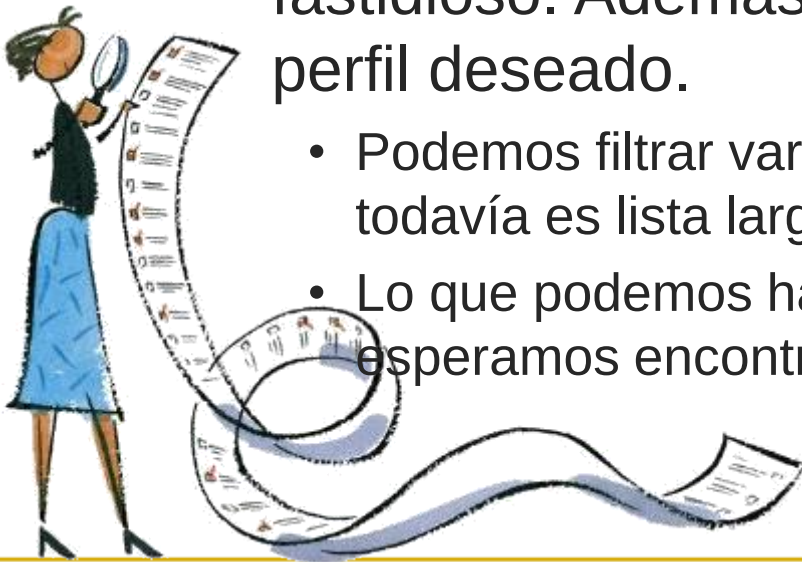


FORMAR

TRANSFORMAR

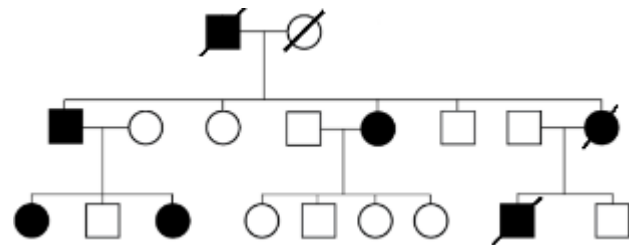
Identificar las Variantes con Caracterizas de interés

- Después del **alineamiento de secuencias** y del **llamado de variantes** tenemos una lista de variantes demasiada larga.
 - Hace poco sentido verificar cada una, seria un trabaja fastidioso. Además, no todas las variantes tienen el perfil deseado.
 - Podemos filtrar variantes por calidad, pero lo que queda todavía es lista larga.
 - Lo que podemos hacer es **filtrar** la lista según **atributos** que esperamos encontrar en nuestra variante de interés.



Pistas para el Filtrado

- Volvemos a nuestro escenario.



- Recordatorio: Estamos intentando identificar una variante que puede causar enfermedad en una familia.
- Entonces, la variante debe tener la siguiente características...
 - Ser **deletérea**.
 - Ser **rara** (i.e. no se ve en la población general).
 - Debe estar **presente** en miembros de la familia afectados y **ausente** en otros.

Presencia y Ausencia en Individuos

- Para filtrar variantes según la presencia y ausencia de variantes en diferentes miembros de la familia, simplemente podemos referirnos al VCF correspondiente.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA000001	NA000002	NA000003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:...
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

- Si queremos ser estricto, vamos a guardar solo variantes detectadas en individuos afectados con **1/0 (heterocigoto)** y no detectadas en otros **0/0**.
 - Otra posibilidad es ignorar genotipo e incluir **1/1** también como criterio para individuos afectados.

Variantes Conocidas en dbSNP

Short Genetic Variations

- Existe una forma de anotación que se puede incluir en el VCF.
 - La línea 'ID' contiene un **identificador** (número rs) que corresponde a la base de datos **dbSNP**.
 - Esta base de datos contiene millones de variantes que podemos usar para filtrar nuestros resultados.

```
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002 NA000003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:,,
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

- Si queremos ser estricto, podemos filtrar todas las variantes con un numero rs para dejar solo variantes “desconocidas”...pero existe otras opciones.

Variantes Conocidas en Otras Base de Datos

- Podemos agregar a nuestro VCF datos sobre variantes ya identificadas en otras personas en el mundo de otros estudios



Exome Aggregation Consortium

ExAC

¡60 706 individuos!

Genome Aggregation Database

¡ ¡ 138 632 individuos ! !

(123,136 exomas y 15,496 genomas)

Potencial Deletéreo (1/4)

- Podemos seleccionar variantes con potencial de ser deletéreas en 3 etapas:
 1. Eliminando variantes con muy poca probabilidad de ser deletéreas.
 2. Intentar predecir el efecto de una variante sobre la estructura proteica subyacente.
 3. Medir la conservación de una variante o regiones a través el tiempo evolutivo basado en su presencia en otras especies.

Potencial Deletéreo (2/4)

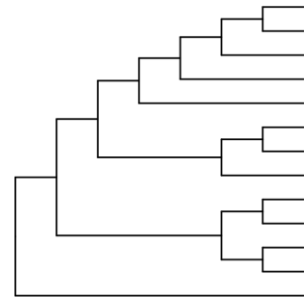
- Para esta primera etapa, podemos agregar a nuestro archivo VCF datos del **RefSeq**, **KnownGene** o **Ensembl Gene**.
 - Además de nombrar los genes en que se ubican variantes (que servirá más tarde...) estas anotaciones permiten determinar si la variante resulta en un **cambio en aminoácidos**.
 - Recordatorio: No todas las mutaciones cambian la estructura de la proteína de un gen debido a la redundancia que existe entre **codones** y **aminoácidos**. Hablamos de **variantes sinónimos** que podemos eliminar de nuestra lista.

Potencial Deletéreo (3/4)



- Basado en la información que existe sobre la estructura de proteínas diferentes herramientas van a intentar predecir el efecto de una variante sobre esa estructura.
 - Se va a basar en diferentes criterios como la ubicación en la proteína (a menudo menos daño al fin de una ristra de aminoácidos que en el medio) o la subestructura afectada (ciertas serán más cambiadas que otras).
 - Programas: MutationTaster, SIFT, PolyPhen2...

Potencial Deletéreo (4/4)



- Basado en la información que existe sobre la conservación de una región o de una base a través especies (y entonces a través tiempo evolutivo), diferentes herramientas van a determinar si una variante está particularmente conservada.
 - Una variante más conservada sugiere que tiene un rol muy importante y que entonces su cambio va a tener un efecto muy deletéreo.
 - Programas: GERP++, PhyloP, SiPhy...



- Existen muchos recursos para agregar anotación a variantes dispersado a través diferentes servicios.
 - Juntarlos puede ser un trabajo fastidioso.
- Agradecimiento, ya existen agregadores de herramientas. Un que es mas útil se llama **Annovar**.
 - Para lo que faltan espacio para almacenar varios base de datos agregado por Annovar, tienen la posibilidad de usar el servicio en línea **wAnnovar**.

¿Que Sigue?

- Ahora tenemos una lista mucho más manejable (todavía larga, pero manejable).
 - En esa etapa típicamente vamos a consultar la literatura científica para ver si los genes asociados a variantes ya están asociados con la enfermedad o si se expresa más o menos en un órgano de interés.
 - **El Online Mendelian Inheritance in Man (OMIM)**
 - **Genecards**
 - **Publicaciones** (Pubmed, Google Scholar...)



Muestras de DNA de la familia

Control de calidad (QC)

Diagrama de embudo de datos genéticos. Muestra tres niveles de datos: 'Datos al nivel del Gen' (top left), 'Datos al nivel de la variante' (top right), and 'Datos al nivel de la Región' (bottom center). Las etiquetas G, C, and A están asociadas con los niveles de datos.

Alineamiento de secuencias

Secuencia de referencia
...TGAATAGGCGTACTTCATCCATGCGC

reads
GAATAGGCTACTT
AGGCCTACTACAT
ACTTCATCCATTC
GCTTCATCAATGCG
C

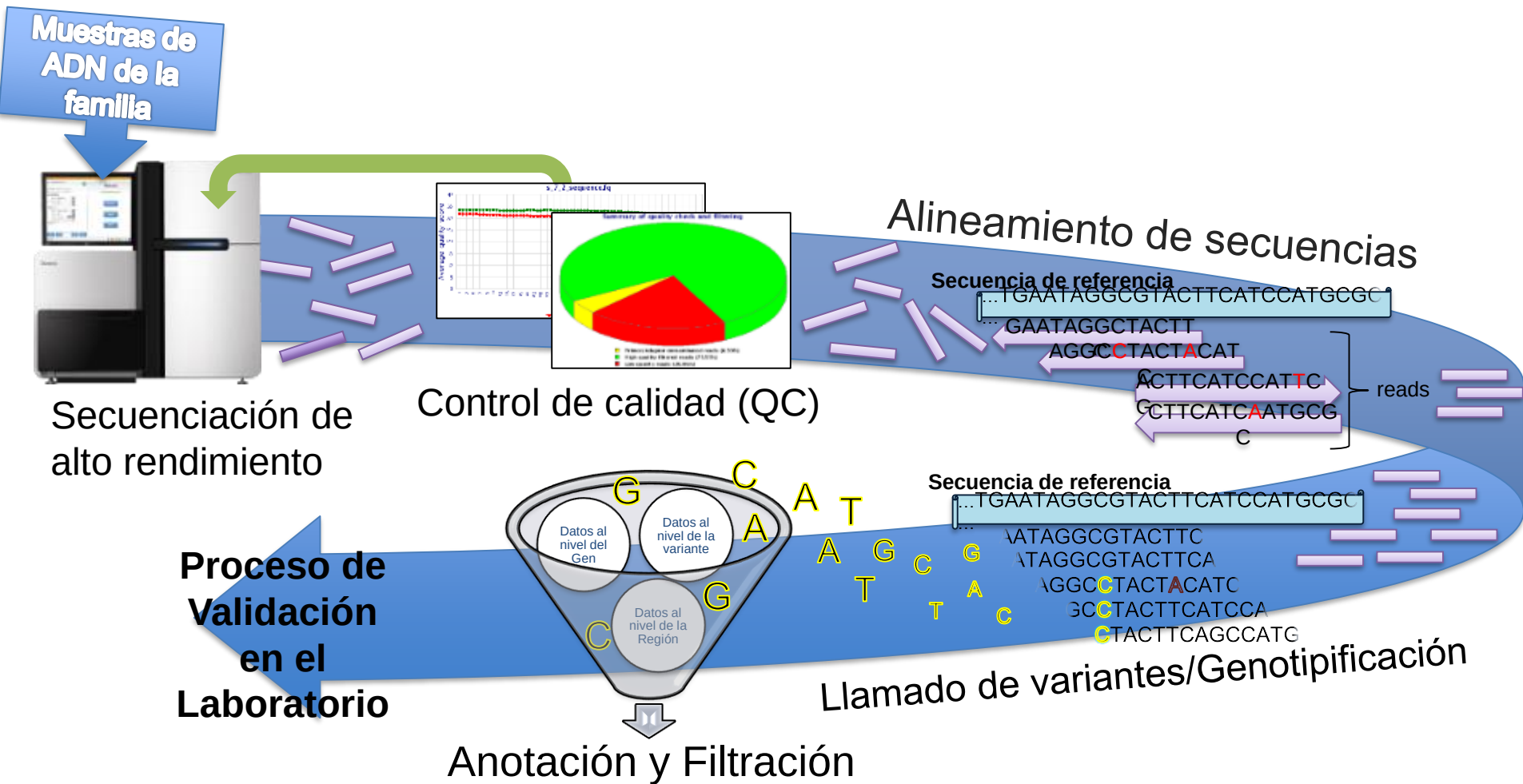
Secuencia de referencia
...TGAATAGGCGTACTTCATCCATGCGC

A T
A G C G
T
T A C
CTACTTCATCCA
CTACTTCAGCCATG

Llamado de variantes/Genotipificación

TRANSFORMAR

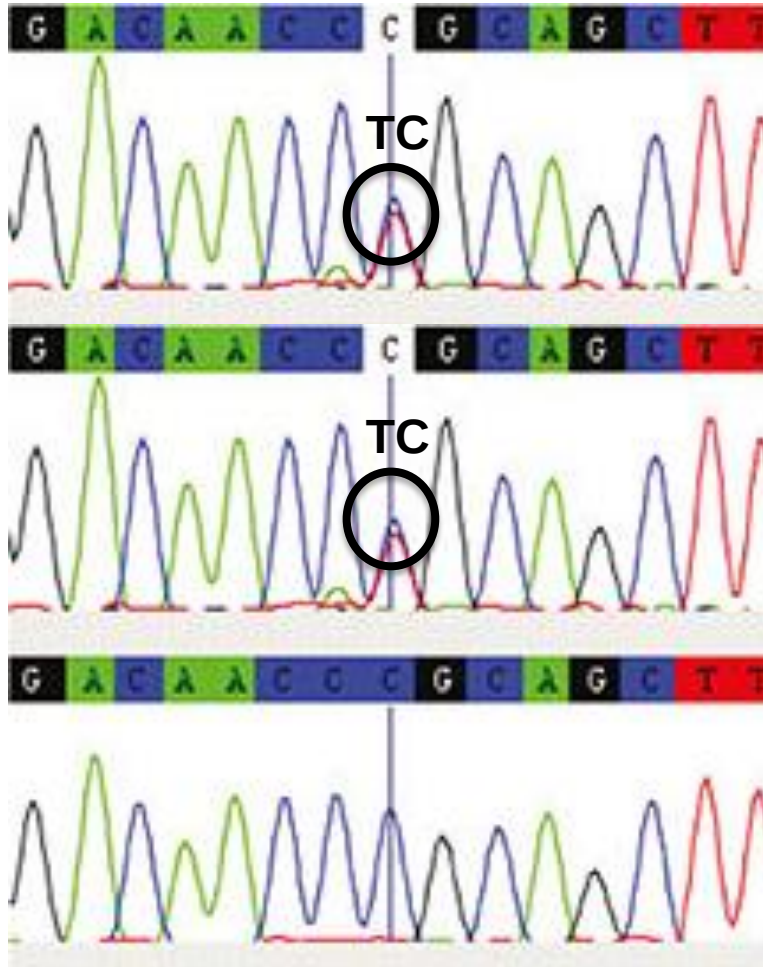
De Secuenciación a Filtración



Validación con El Método Sanger

- Al final de un estudio (y a menudo, con el ayuda de la literatura y otras herramientas), vamos a elegir una lista **cortita** de **variantes candidatas**.
 - Para cada variante, vamos a seleccionar un fragmento que contiene la variante (si existe) y secuenciarlo con Sanger.
 - Sirve no solo para **validar la existencia** de una variante, sino su **ausencia** (en la madre en nuestro ejemplo).
 - Podemos hacer **validación** en varios miembros de una familia, **incluyendo algunos no secuenciados antes!** (menos costoso que secuenciar todo el genoma/exoma)

Ejemplo de Validación



Niña
presente

Padre
presente

Madre
ausente

Mas allá del Trabajo *In Silico*

- Siguiendo una validación con éxito, termina el proceso *in silico* del descubrimiento de variantes.
- Ahora necesitamos volver a trabajar *in vitro* o *in vivo*, con modelos experimentales que contienen la mutación o mutaciones de interés (i.e. células, pesca, ratón...).
- Si sigue el éxito, se puede publicar los resultados y ayudar a los pacientes a tomar decisión sobre su salud y el de su familia presente y futura.

