Análisis genético mediante secuenciación masiva

Biocomputación
Grado en Bioquímica

Motivación

Existen diferentes situaciones que requieren la determinación del genotipo de un individuo

Asesoramiento genético: análisis genético de una pareja que está pensando en tener un niño y en cuyas familias hay antecedentes.

Diagnóstico pre-natal: es el test genético de un feto. Puede ser realizado en los casos en los que existe riesgo de que el bebé presente genes asociados a un retraso mental o discapacidad física.

Medicina personalizada: Detectar toda la variación de secuencia de un individuo para i) estimar si existen predisposiciones a ciertas enfermedades y ii) predecir mejor el pronóstico y iii) poder diseñar un tratamiento personalizado

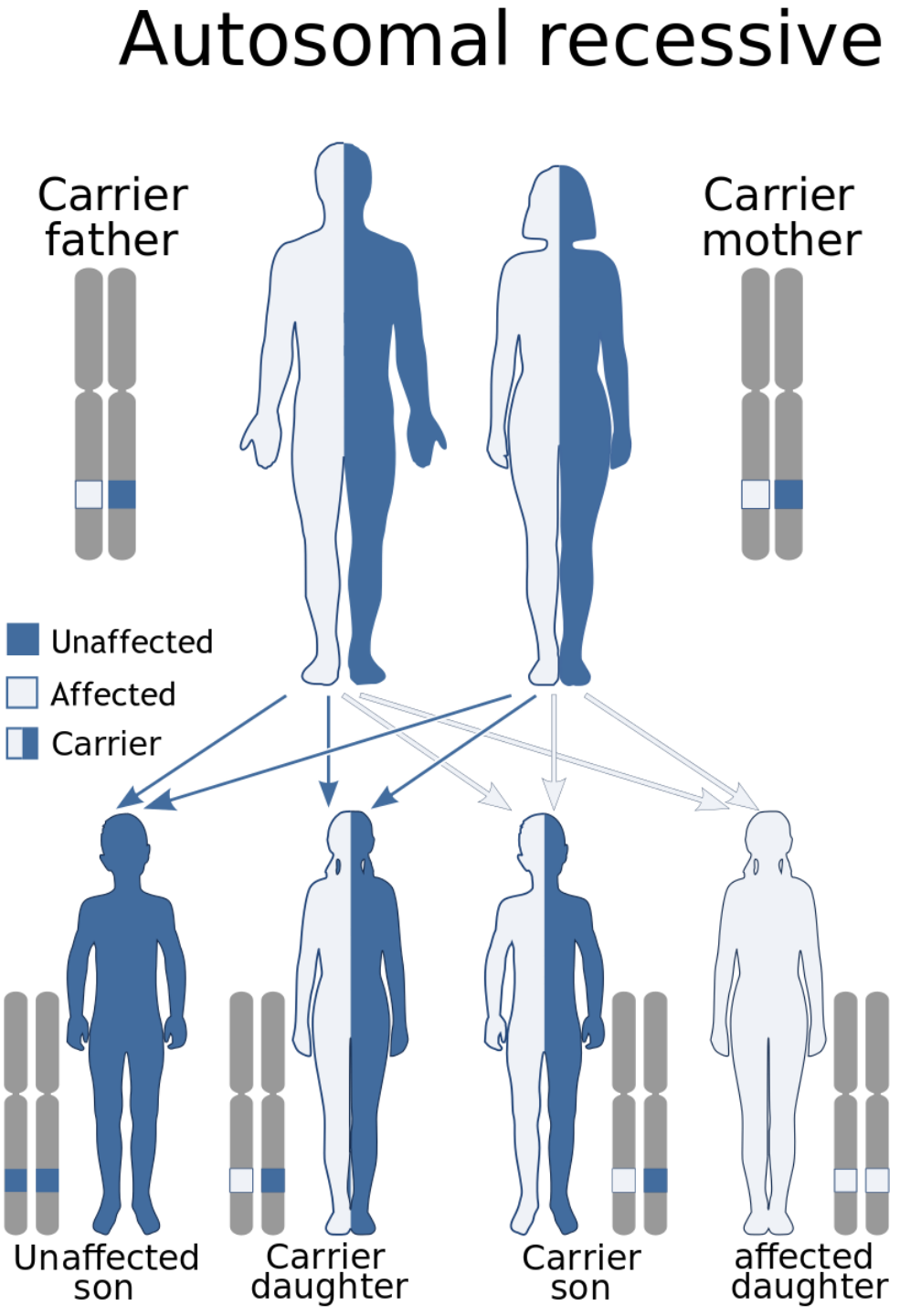
Trastornos de aparición tardía: incluye el análisis de enfermedades en adultos como, por ejemplo, cáncer y enfermedades cardiacas. Estas enfermedades son complejas y implican tanto la genética como el ambiente (GWAS – Genome wide association study).

Genómica de poblaciones: Estudios poblacionales y evolutivos

Herencia mendeliana

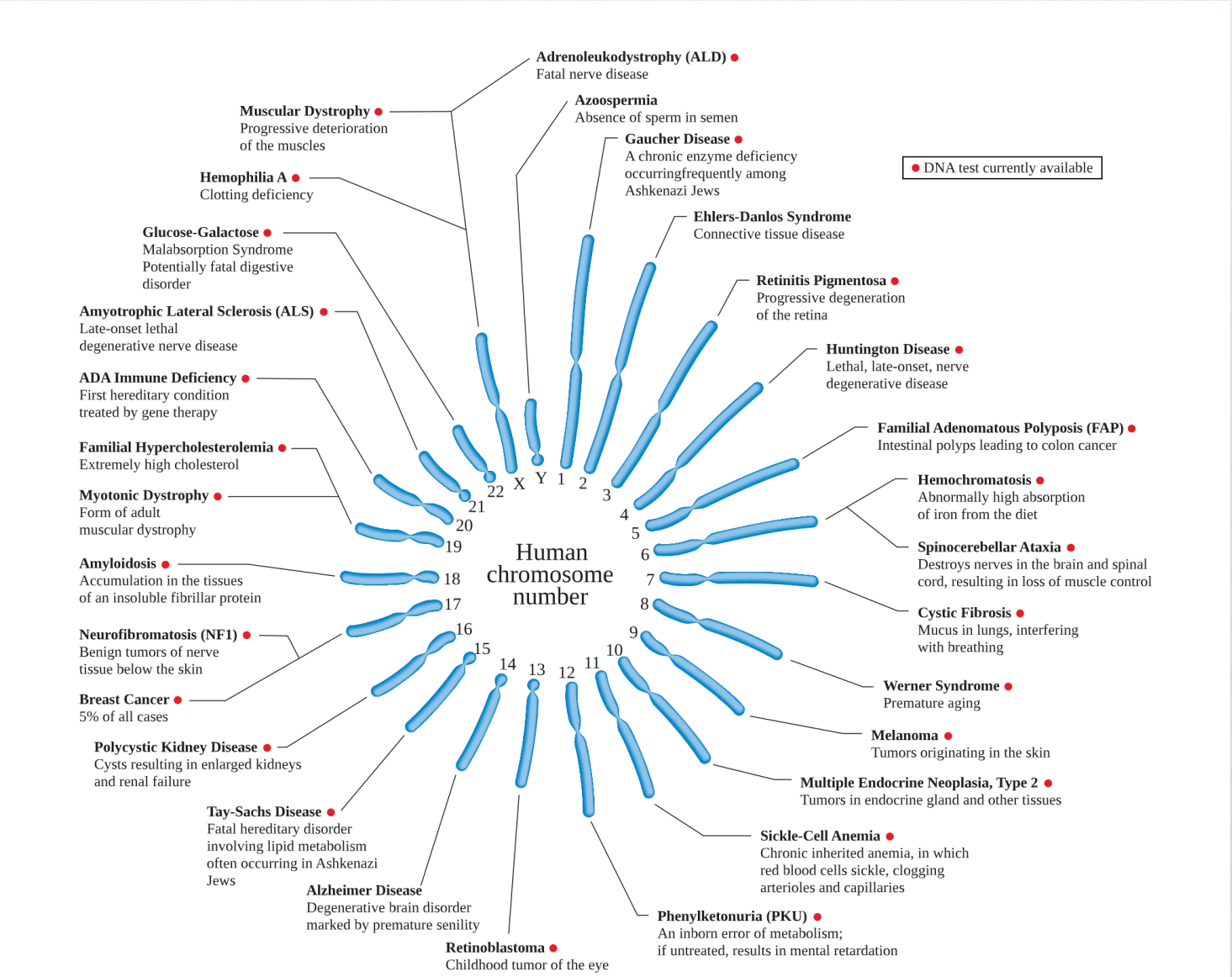
Dos individuos heterocigotos, portadores de un alelo recesivo (cromosoma autosómico) que causa enfermedad tendrán

- 0.25 de probabilidad de tener un niño sano, no-portador
- 0.5 de probabilidad de tener un niño sano, pero portador
- 0.25 de probabilidad de tener un niño afectado



https://commons.wikimedia.org/wiki/File:Autosomal_recessive_EN.svg

Enfermedades genéticas (monogénicas)



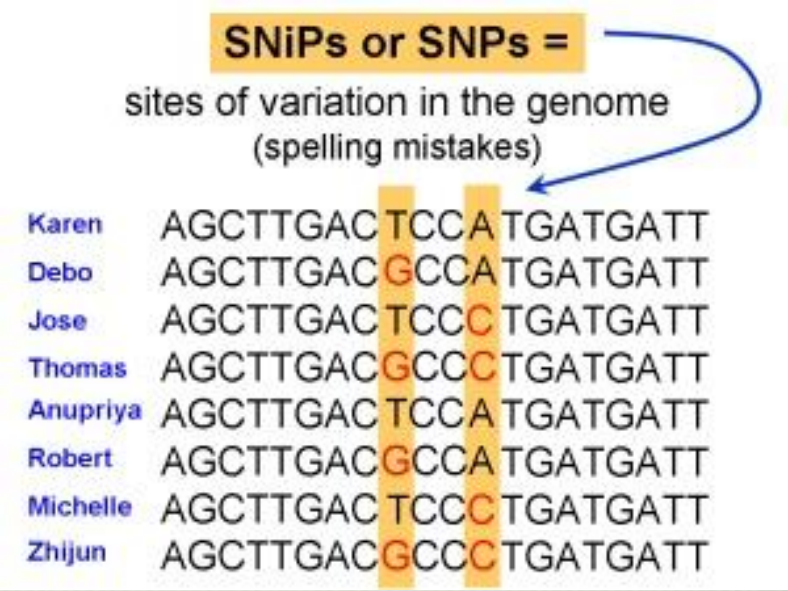
https://en.wikipedia.org/wiki/Genetic_disorder

Tipos de variación: SNPs

En un principio podríamos usar diferentes tipos de variación para detectar las diferencias entre poblaciones o grupos (sanos/enfermos) pero el análisis genético se basa frecuentemente en SNPs

SNPs: Polimorfismo de nucleótido simple

- Suelen tener solo dos alelos, por ejemplo C/T
- Mas de 2/3 de todos los SNPs derivan de un dinucleótido CpG (mutación por metilación)
- MAF: Minor Allele Frequency (frecuencia del alelo menor)
- A partir de una MAF del 5% (algunas veces del 1%) se define como SNP común
- Los SNPs comunes se suelen emplear en los estudios de asociación



Tipos de variación: STR

STR: Short Tandem Repeats

Alelo 1: 2 repeticiones ATTACATCATAATACCTAAATCA

Alelo 2: 3 repeticiones ATTACATCATCATAATACCTAAA

Alelo 3: 4 repeticiones ATTACATCATCATCATAATACCT

Se detectan fácilmente mediante PCR y se usan extensivamente en tests de paternidad y la medicina forense

Electroforesis en gel

Tipos de variación: indels

Dos alelos: ACT/-

```
Individuo 1     AATACGACTTATCGGCTACTACTCCTACTACTA
Individuo 2     AATACGACTTATCGGCTACTACTCCTACTACTA
Individuo 3     AATACGACTTATCGGCT---ACTCCTACTACTA
Individuo 4     AATACGACTTATCGGCT---ACTCCTACTACTA
Individuo 5     AATACGACTTATCGGCT---ACTCCTACTACTA
```

Si el alelo ancestral es ACT → deleción

Dos alelos: GAC/-

```
Individuo 1     AATACGACTTATCGGCT---ACTCCTACTACTA
Individuo 2     AATACGACTTATCGGCT---ACTCCTACTACTA
Individuo 3     AATACGACTTATCGGCTGACACTCCTACTACTA
Individuo 4     AATACGACTTATCGGCTGACACTCCTACTACTA
Individuo 5     AATACGACTTATCGGCTGACACTCCTACTACTA
```

Si el alelo ancestral es - → GAC es una inserción

Genotipo / Haplotipo

Genotipo: la combinación alélica (un alelo de la madre y el otro del padre) en un locus

Heredado del padre ATTACATACTACCTAACGTACGGATCA

Heredado de la madre ATTACATAATACCTAACGAACGGATCA

Genotipos C/A T/A

Haplotipo: la sucesión de los alelos a lo largo de un cromosoma

Haplotipo C/T

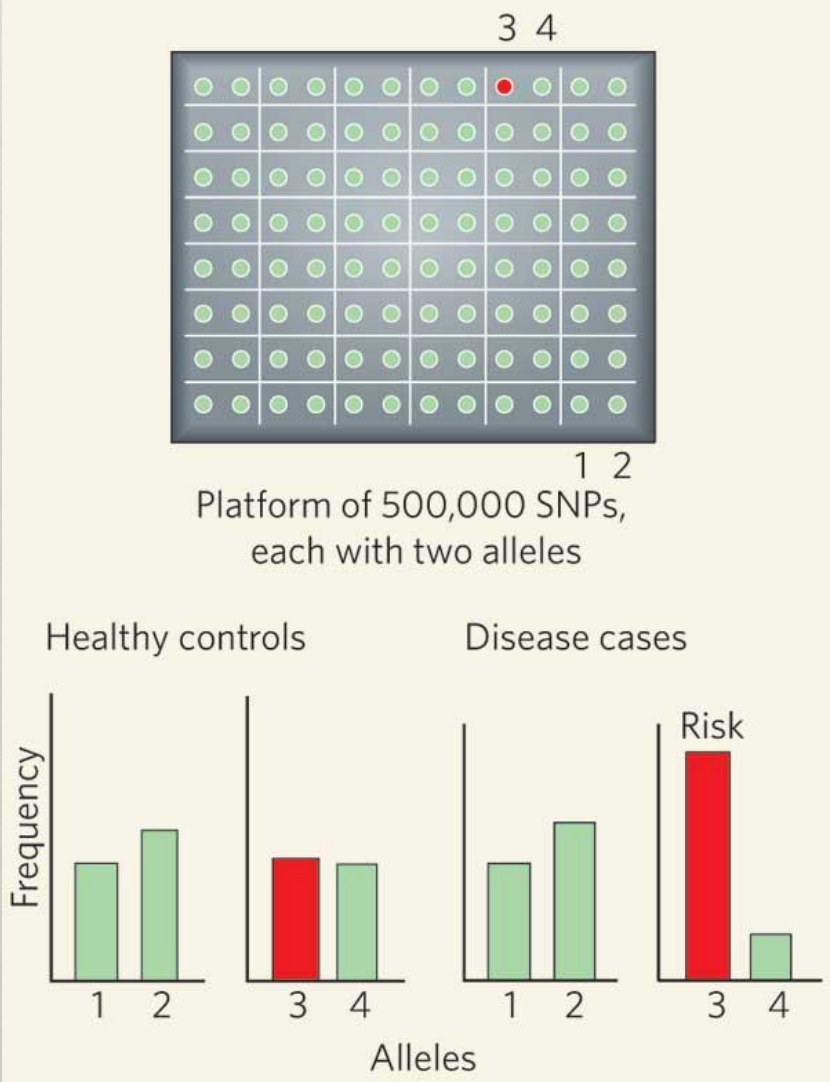
Heredado del padre ATTACATACTACCTAACGTACGGATCA

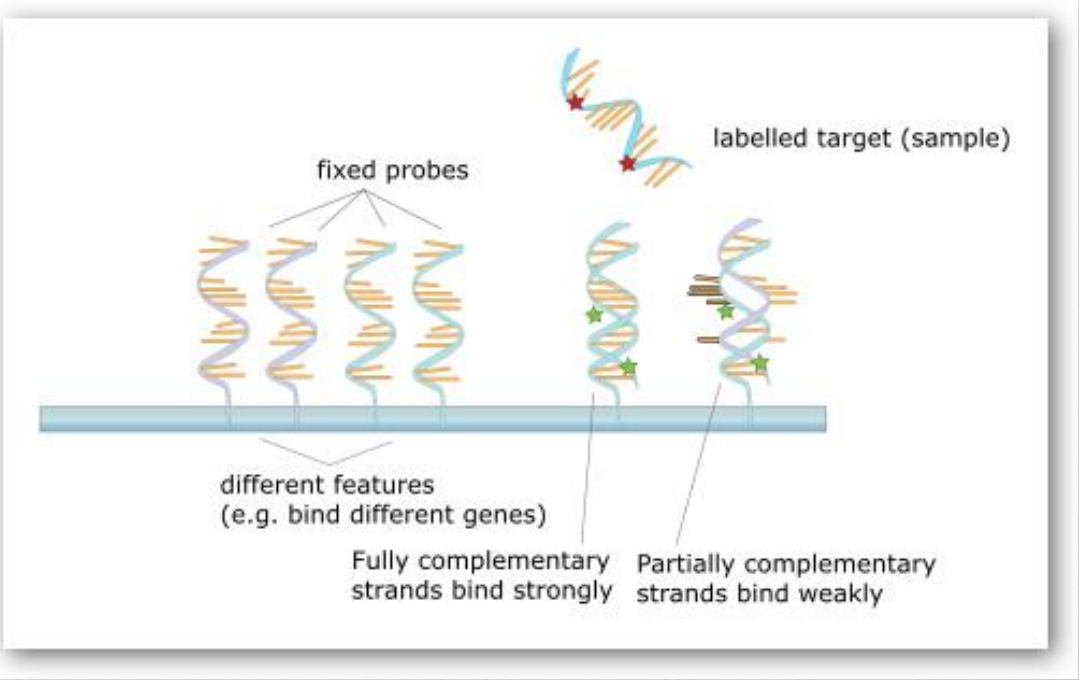
Heredado de la madre ATTACATAATACCTAACGAACGGATCA

Haplotipo A/A

Detectar la variación de secuencia

El principio de detectar variación mediante **chips de DNA** es el mismo que determinar los perfiles de expresión:



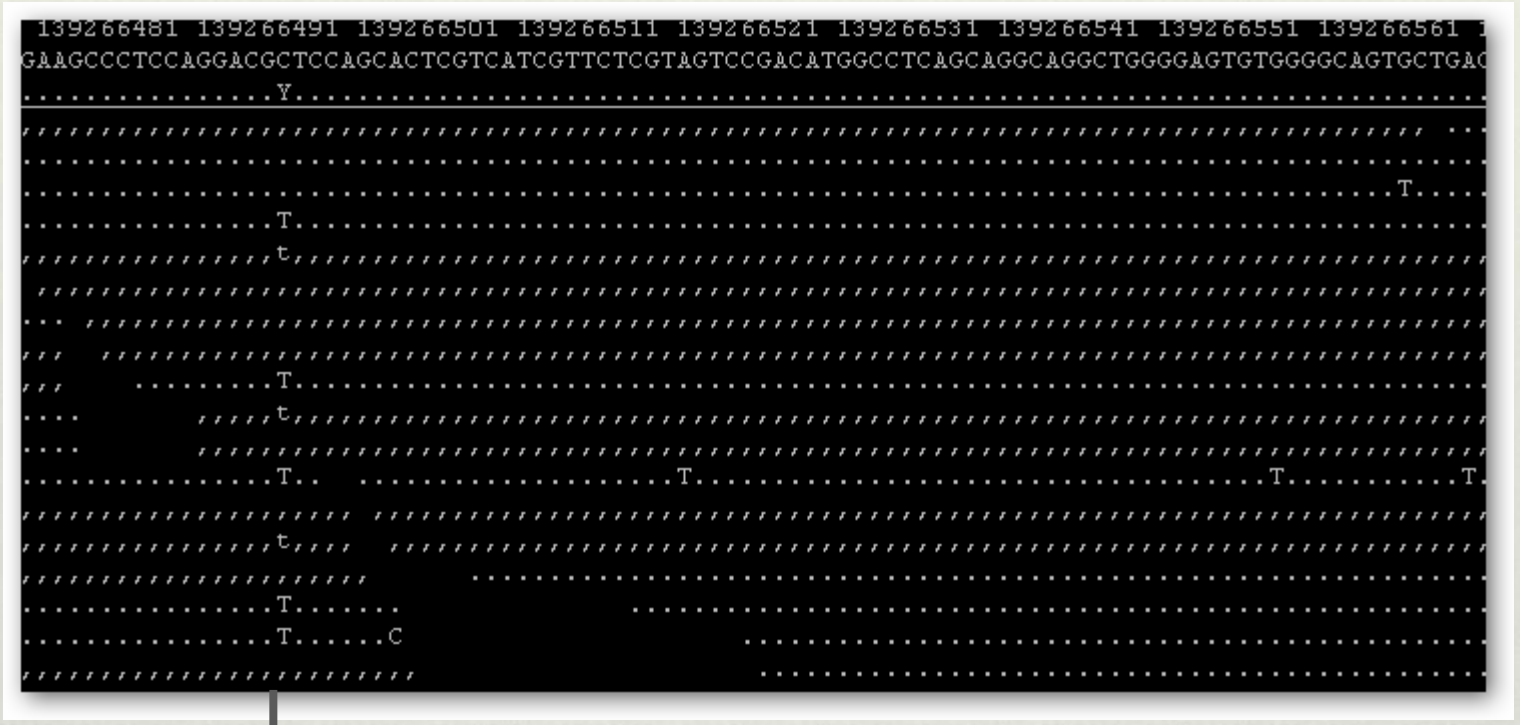


Detectar la variación de secuencia

Mediante secuenciación masiva

- Secuenciar una región (o el genoma completo) del individuo
- Alinear las lecturas frente a una secuencia de referencia
- Obtener el genotipo de los alineamientos

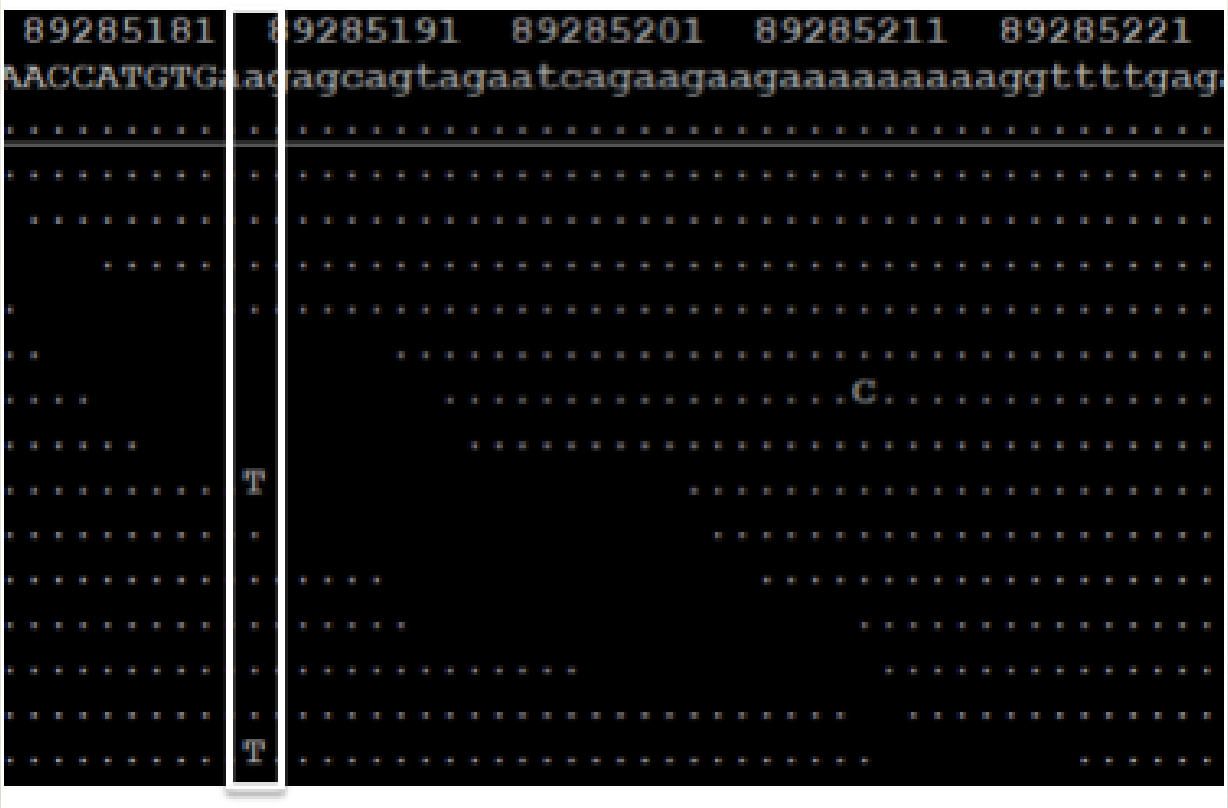
Como detectamos la variación



Posición: chr9:139266466-139266536
Alelo de la referencia: C
8 reads indican timina en la muestra
10 reads indican citosina en la muestra

Heterocigoto con genotipo: C/T

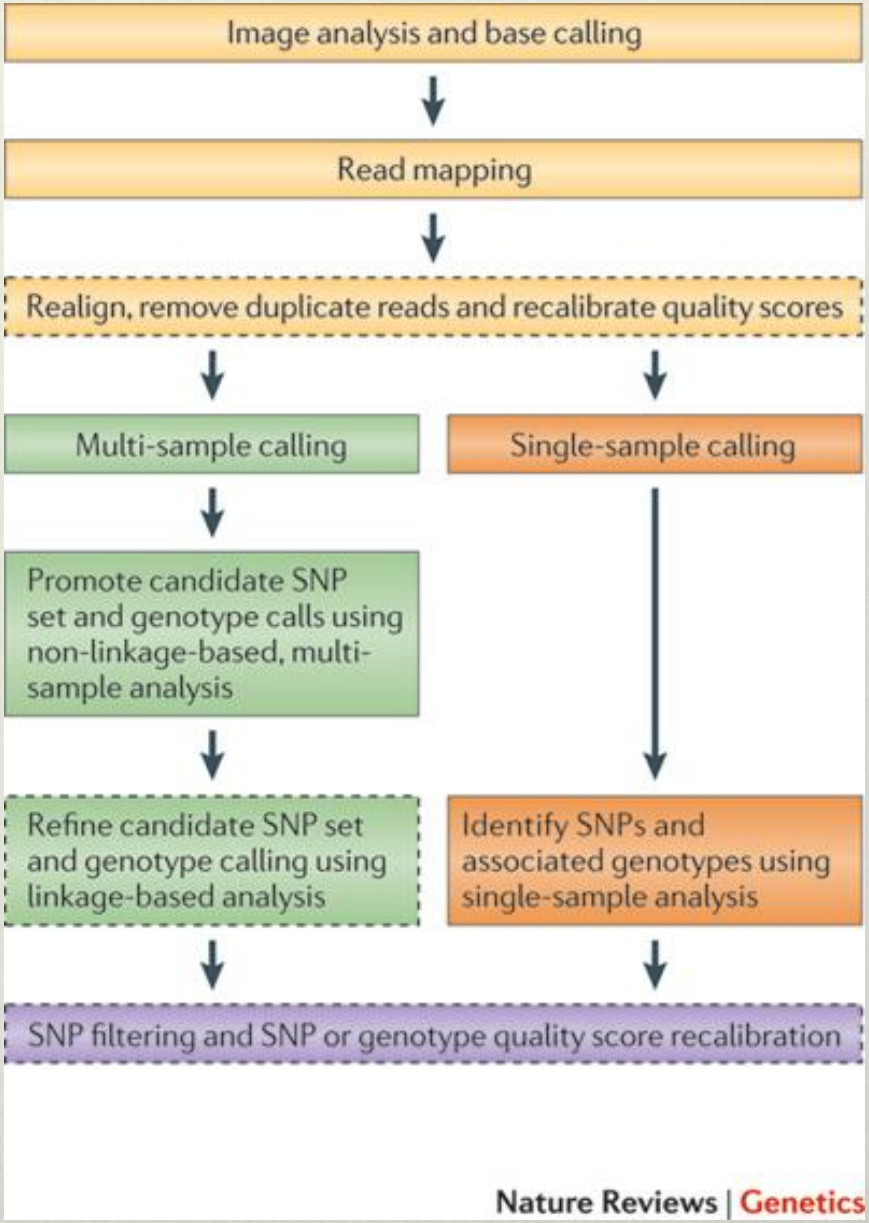
Como detectamos la variación



La posición muestra variación A (9)/T (2), lo que puede deberse a diferentes razones:

- errores de secuenciación
- variantes de secuencia
- mutaciones somáticas

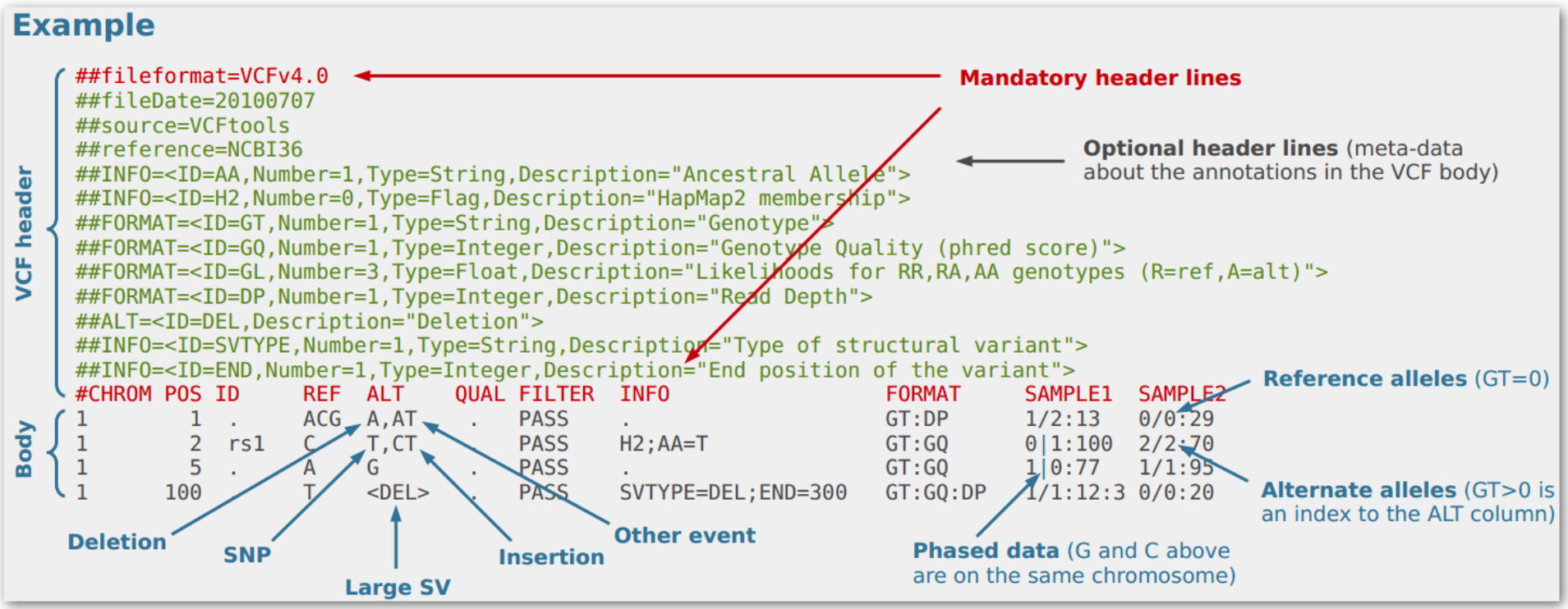
Detectar la variación de secuencia



FROM THE FOLLOWING ARTICLE:
Genotype and SNP calling from next-generation sequencing data
Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen & Yun S. Song
Nature Reviews Genetics **12**, 443-451 (June 2011)
doi:10.1038/nrg2986

El formato VCF

http://vcftools.sourceforge.net/VCF-poster.pdf



El formato VCF

Types of variants

SNPs

Alignment
```
ACGT
ATGT
```

VCF representation

POS	REF	ALT
2	C	T

Insertions

Alignment
```
AC-GT
ACTGT
```

VCF representation

POS	REF	ALT
2	C	CT

Deletions

Alignment
```
ACGT
A--T
```

VCF representation

POS	REF	ALT
1	ACG	A

Complex events

Alignment
```
ACGT
A-TT
```

VCF representation

POS	REF	ALT
1	ACG	AT

Large structural variants

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

Bases de datos

dbSNP es la base de datos de referencia
http://www.ncbi.nlm.nih.gov/SNP/index.html

dbSNP almacena tanto SNPs como indels

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	137	37.3	187,852,828	53,558,214 (38,077,993)	22,508,883	74,323,968	35,994,337
Mus musculus	137	38.1	117,483,135	70,036,850 (687,521)	28,547,196		77
Bos taurus	137	6.1	26,542,614	13,704,221 (3,003)	4,803,553		953
Total: 3 Organisms		genomes	331,878,577	137,299,285 (38,768,517)	55,859,632	74,323,968	35,995,367

Bases de datos:

OMIM Online Mendelian Inheritance in Man
http://omim.org/about

Toda la información acerca de enfermedades mendelianas conocidas incluyendo casi 15000 genes y la relación entre el genotipo y fenotipo

Distribution of Phenotypes across Genes (Updated November 7th, 2017) :

Number of genes with 1 phenotype	2,628
Number of genes with 2 phenotypes	717
Number of genes with 3 phenotypes	260
Number of genes with 4+ phenotypes	232

Number of Entries in OMIM (Updated November 7th, 2017) :

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	14,955	725	49	35	15,764
Gene and phenotype, combined +	75	0	0	2	77
Phenotype description, molecular basis known #	4,769	322	4	31	5,126
Phenotype description or locus, molecular basis unknown %	1,467	124	5	0	1,596
Other, mainly phenotypes with suspected mendelian basis	1,665	106	2	0	1,773
Totals	22,931	1,277	60	68	24,336

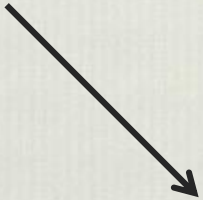
Efecto de la variación

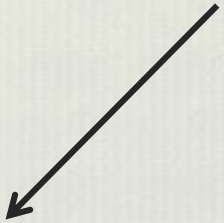
Mutación en la línea germinal

- Tienen relevancia evolutiva si llegan a transmitirse a la descendencia
- Tumores Germinales de Testículo - Testicular germ cell tumour (TGCT)

Mutación en una célula somática

- No tienen relevancia evolutiva pero si para el individuo que lo porta: son la base de enfermedades como el cáncer





¿Mutaciones en que regiones tendrán mayor probabilidad de impactar en el fenotipo?

Efecto de la variación

Mutación en la línea germinal

- Tienen relevancia evolutiva si llegan a transmitirse a la descendencia
- Tumores Germinales de Testículo - Testicular germ cell tumour (TGCT)

Mutación en una célula somática

- No tienen relevancia evolutiva pero si para el individuo que lo porta: son la base de enfermedades como el cáncer

¿Mutaciones en que regiones tendrán mayor probabilidad de impactar en el fenotipo?

Regiones funcionales:

- Genes codificantes y no-codificantes
- Regiones reguladoras: Promotor, potenciador, dianas de microRNAs, etc. etc.

SNP/SNV el la región codificante

Mutación sinónima

↓

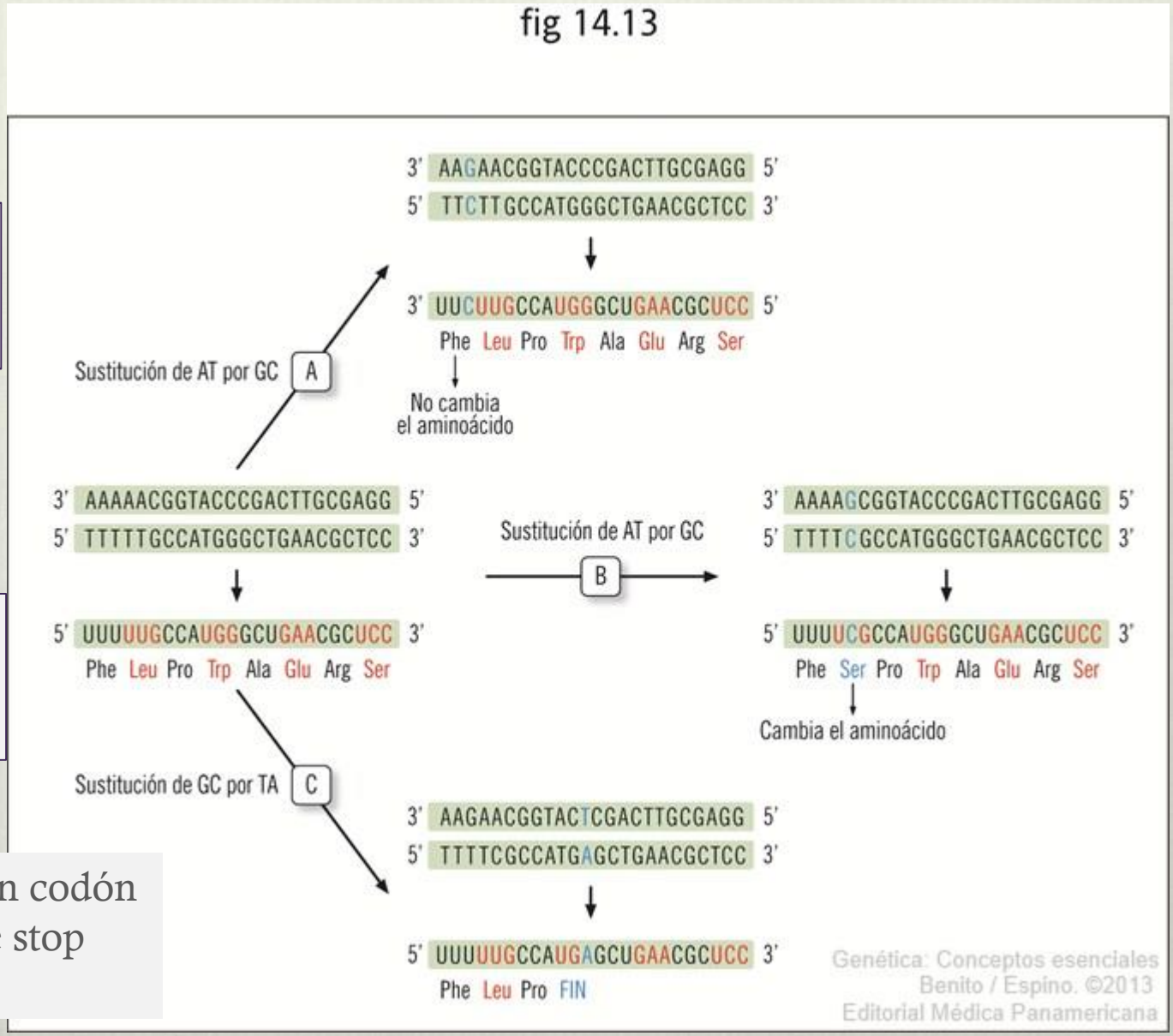
El código genético define si una mutación es sinónima o no

Mutación de sentido erróneo

↓

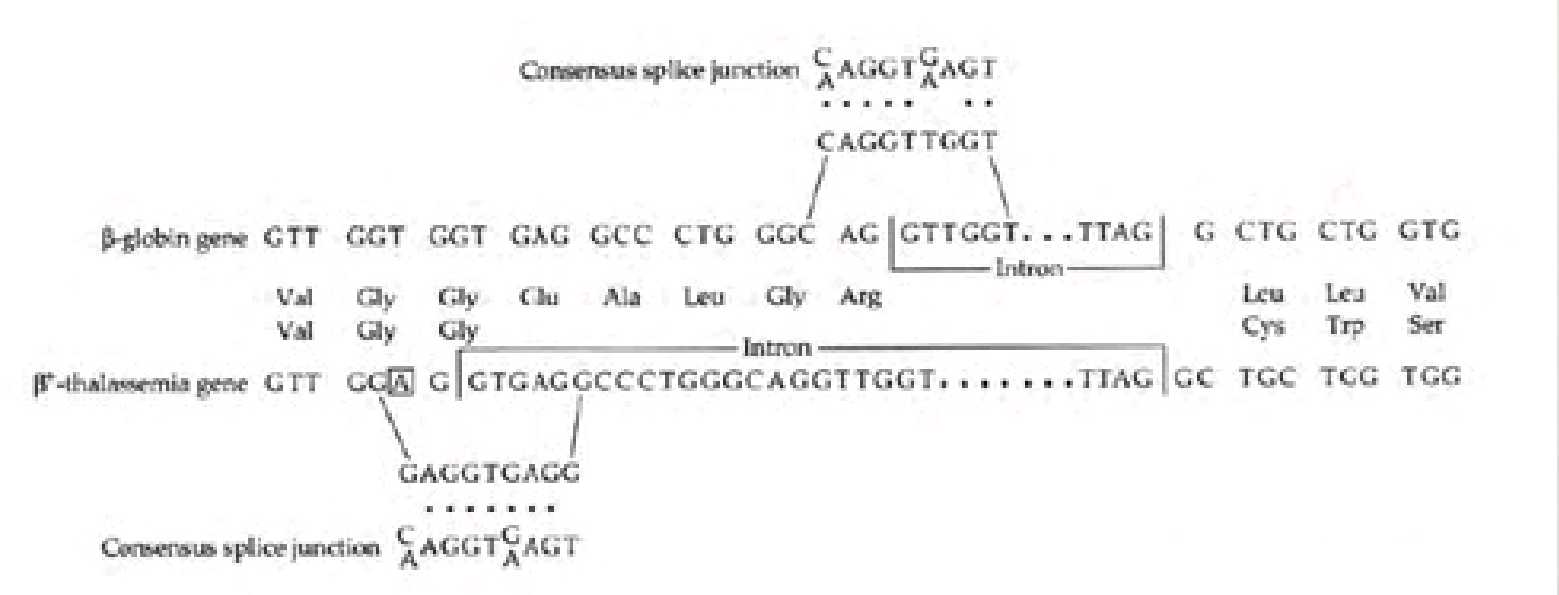
Sustitución (reemplazamiento) del aminoácido

Mutación sin sentido (un codón muta hacia un codón de stop



SNP/SNV el la región codificante

Una mutación sinónima del codón GGT hacia GGA (glicina) en el primer exon del gen β-globina causa un splicing diferente resultando en un marco de lectura erróneo



Mutación sinónima ≠ Mutación silenciosa

http://judgestarling.tumblr.com/post/57494092909/mutations-synonymous-nonsynonymous-silent

SNP/SNV el la región codificante

Mutación sinónima

↓

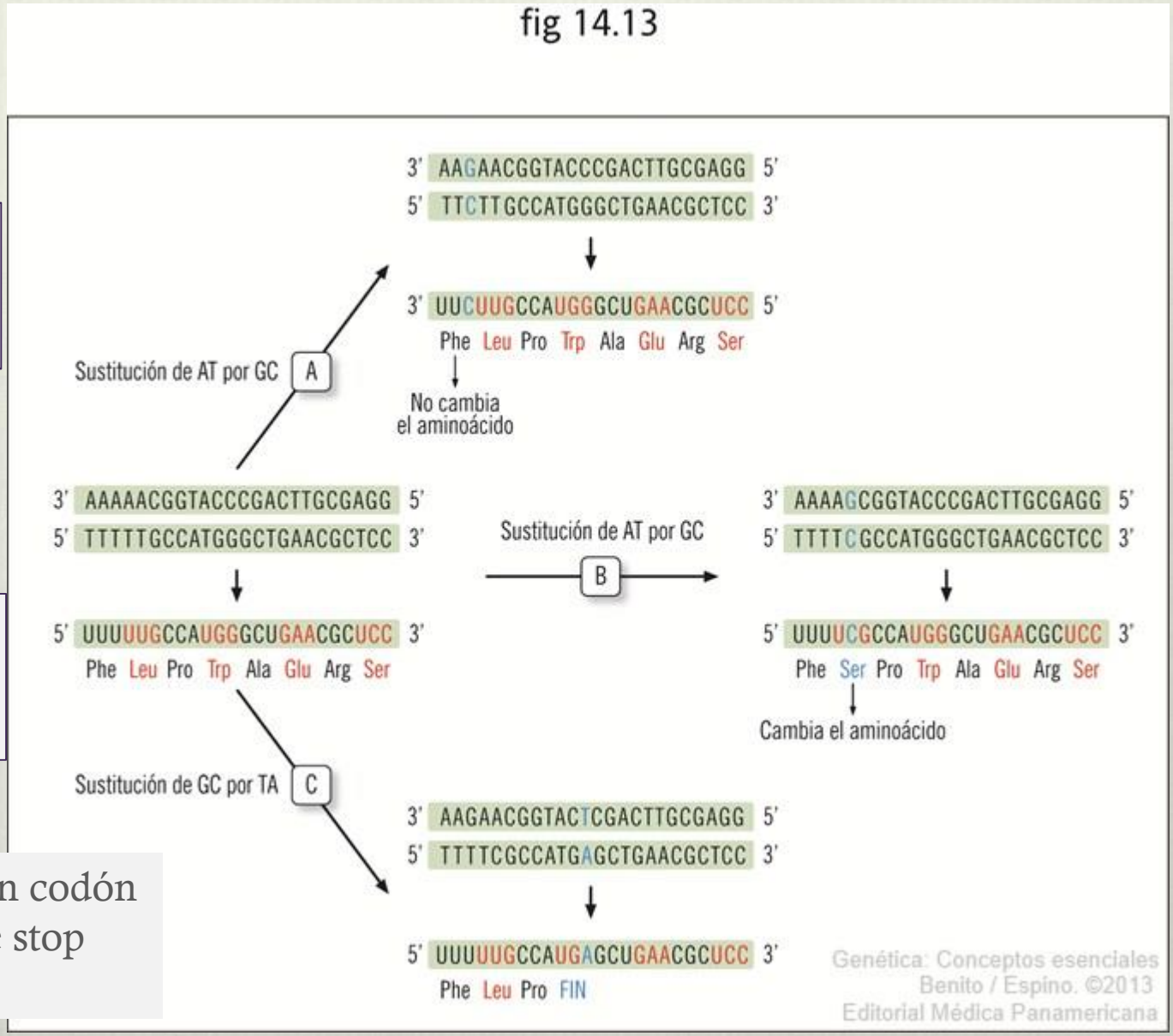
El código genético define si una mutación es sinónima o no

Mutación de sentido erróneo

↓

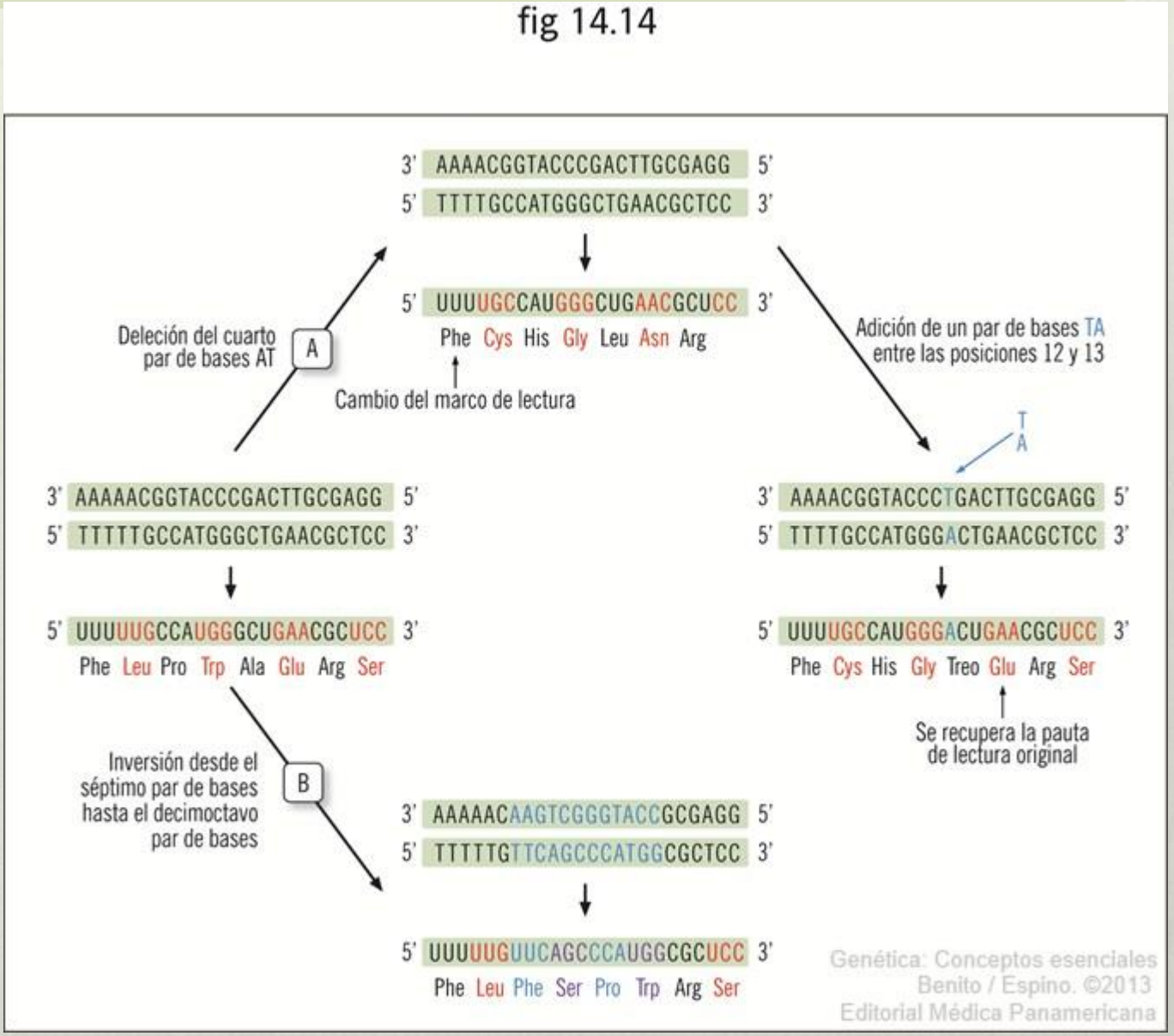
Sustitución (reemplazamiento) del aminoácido

Mutación sin sentido (un codón muta hacia un codón de stop



SNP/SNV el la región codificante

Inserciones y deleciones que no son múltiples de tres cambiaran la pauta de lectura



SNP/SNV en aceptores y donores

FREE
Lens | September 2010

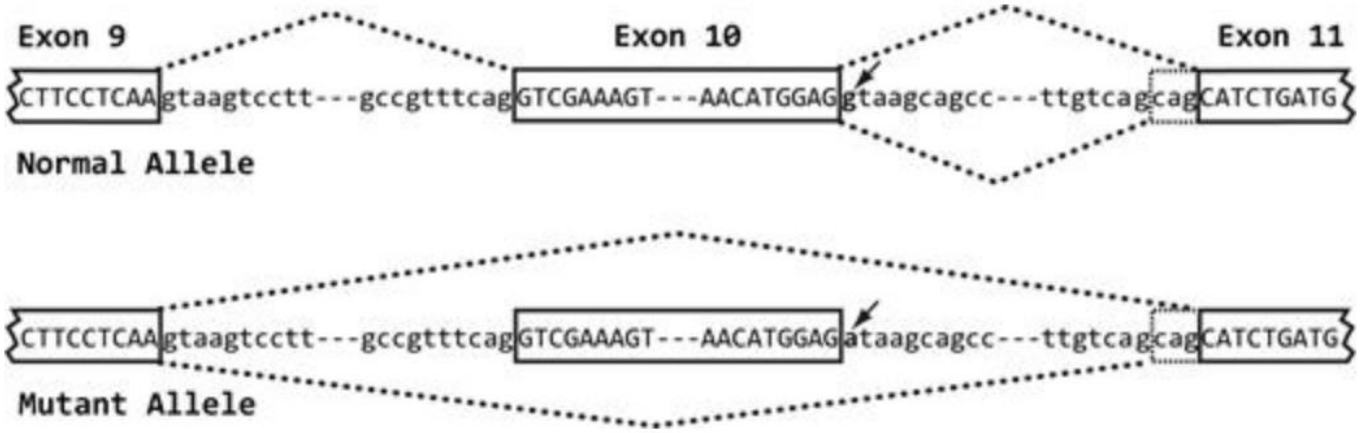
An *ADAMTS17* Splice Donor Site Mutation in Dogs with Primary Lens Luxation

Fabiana H. G. Farias; Gary S. Johnson; Jeremy F. Taylor; Elizabeth Giuliano; Martin L. Katz; Douglas N. Sanders; Robert D. Schnabel; Stephanie D. McKay; Shahnawaz Khan; Puya Gharahkhani; Caroline A. O'Leary; Louise Pettitt; Oliver P. Forman; Mike Boursnell; Bryan McLaughlin; Saija Ahonen; Hannes Lohi; Elena Hernandez-Merino; David J. Gould; David R. Sargan; Cathryn Mellersh

+ Author Affiliations & Notes

Investigative Ophthalmology & Visual Science September 2010, Vol.51, 4716-4721. doi:10.1167/iovs.09-5142

Figure 3.

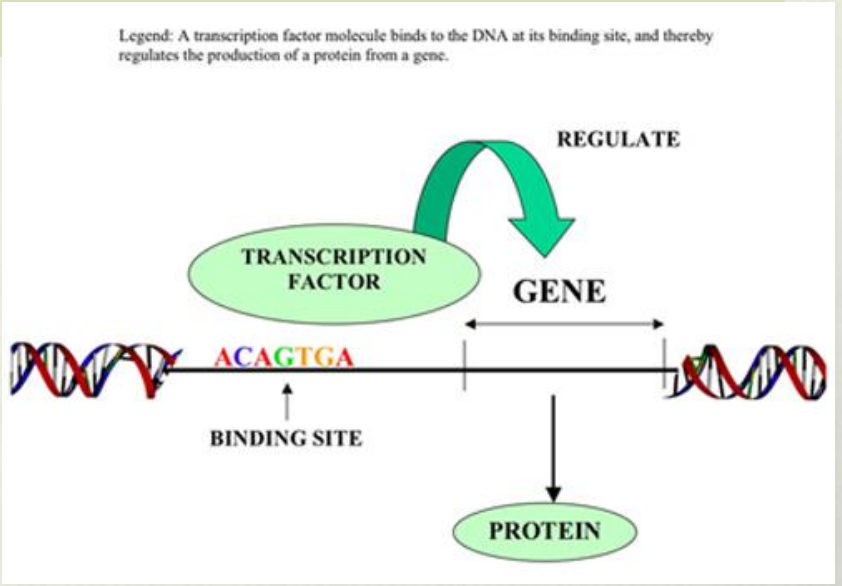


View Original | Download Slide

Exon splicing patterns of transcripts from normal and mutant *ADAMTS17* alleles. In mRNA from the normal allele, exon 9 is spliced to exon 10, which is spliced to each of the two alternative splice acceptor sites at the 5′ end of exon 11. In RNA from the mutant allele, exon 10 is skipped and exon 9 is spliced to the alternative exon 11 splice acceptor sites. *Arrows*: position of the transition at *ADAMTS17:c.1143+1*.

SNP/SNV el los TFBS

Mutaciones en los sitios de unión a factores de transcripción pueden provocar que este no se pueda unir
→ el gen no se expresa



Ejemplo: mutación en el TFBS del gen F9 (factor IX – factor de coagulación) que impide la unión del factor de transcripción HNF4α. En consecuencia, el gen no se transcribe y el fenotipo del portador es el de hemofilia (ausencia del factor IX).

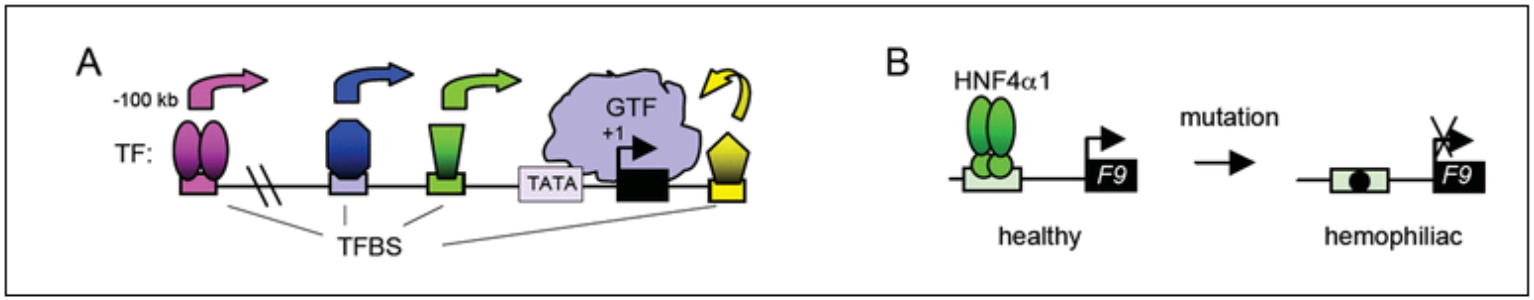


Fig. 1. Role of transcription factor binding sites (TFBS) in the regulation of eukaryotic gene expression.
A. Schematic diagram of a eukaryotic promoter showing Transcription Factor (TF) Binding Sites (TFBS), the TATA box and the start site of transcription (+1). GTF, General Transcription Factors. Not shown are histones, co-regulators, mediator or chromatin remodeling complexes, etc. **B**. Effect of a mutation in the HNF4α binding site on expression of the blood coagulation gene *F9*.

http://fmslab.ucr.edu/binplone/about

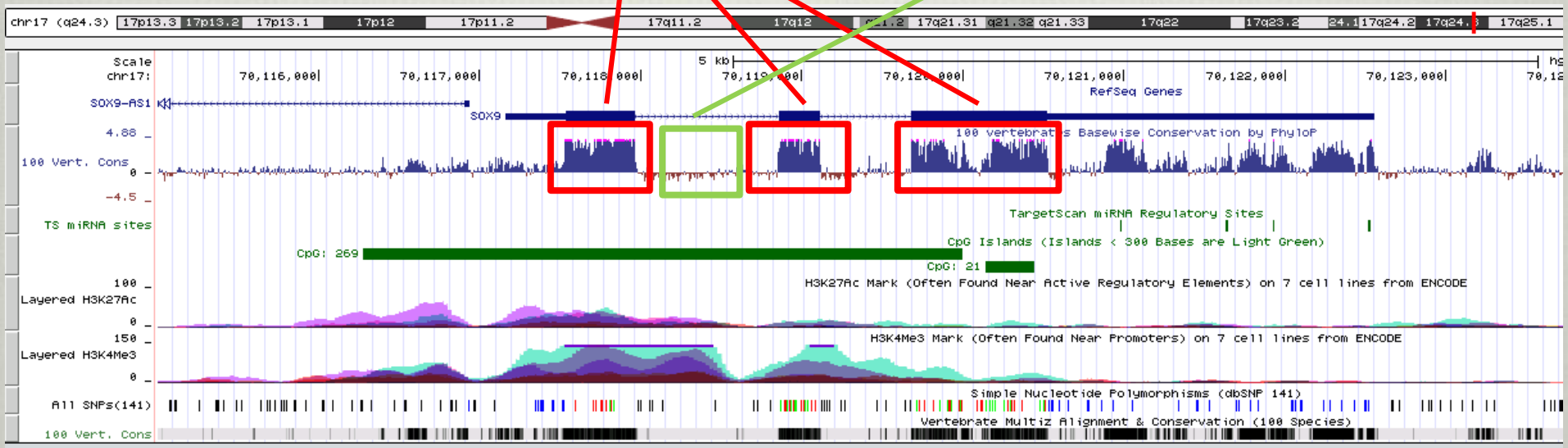
Efecto de la variación y conservación

Mediante un **alineamiento** múltiple se puede determinar el grado de conservación para cada base

Un alto grado de conservación indica función a nivel de secuencia (la selección negativa actúa sobre mutaciones deletéreas)

La región codificante se conserva a nivel de secuencia

La secuencia de los intrones no se conserva generalmente (excepto subsecuencias reguladoras)



Ejemplo

- La fibrosis quística (FQ) es una enfermedad que causa secreciones anómalas y espesas de las glándulas exocrinas (mucosidad muy espesa)
- La FQ muestra herencia autosómica recesiva
- Se produce por la mutación en el gen CFTR (Homo sapiens cystic fibrosis transmembrane conductance regulator)

- El SNP rs12190947 se ubica en el cromosoma 7, posición 117230409 y tiene dos alelos C/A
- Coincide con la región codificadora del gen CFTR (exon 13)

Codón con alelo mayor: GCA → Alanina (Individuos sanos)
Codón con alelo menor: GAA → Glutamina (Individuos enfermos en homocigosis)