

Máster en Bioinformática

Secuenciación Genómica y Análisis De Variantes Para Medicina Personalizada y De Precisión

Curso académico 2024-25

Edición Abril



Universidad
Internacional
de Valencia

Dra. Laura Gutiérrez Macías
laura.gutierrez.m@professor.universidadviu.com

Capítulo 1. La estructura del genoma humano y patrones de transmisión de enfermedades genéticas

1.1.

- Estructura del genoma humano
 - 1.1.1.- Genoma mitocondrial
 - 1.1.2.- Genoma nuclear
 - 1.1.3.- ¿Qué tipos de secuencias encontramos en el genoma humano?
 - 1.1.4.- Secuencias repetidas y secuenciación masiva

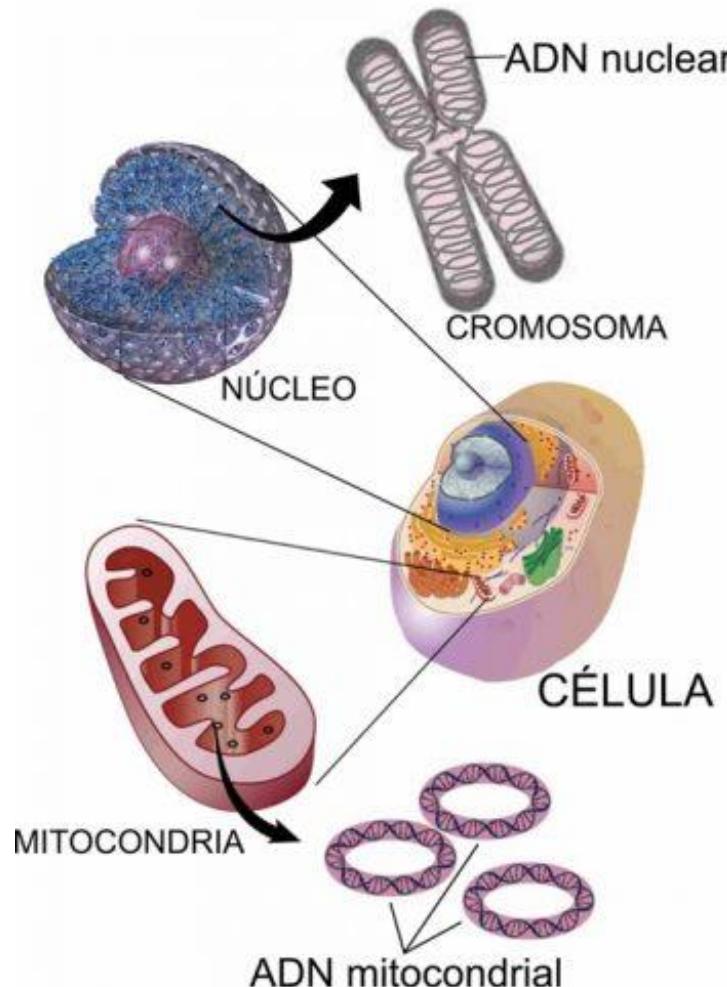
1.2.

- Patrones de transmisión de enfermedades genéticas
 - 1.2.1.- Variabilidad del genoma
 - 1.2.2.- Tipos de enfermedades genéticas.

1.1

Estructura del genoma humano

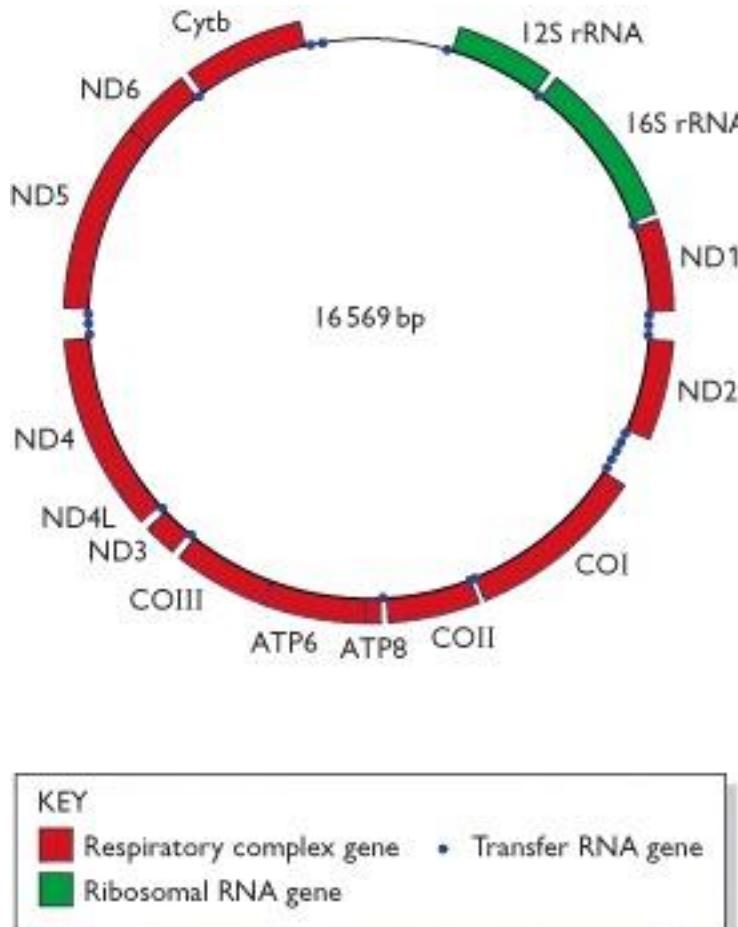
1.1. Estructura del genoma humano



Llamamos genoma al contenido total de ácido desoxirribonucleico (ADN) presente en una célula. Podríamos decir que es el manual de operaciones que contiene todas las instrucciones que ayudan al desarrollo y funcionamiento de un organismo vivo.

El genoma de los organismos eucariotas consta de dos tipos: **genoma nuclear** y **genoma mitocondrial**. En general, para los estudios de genética básicos, nos centramos en el genoma nuclear.

1.1.1. Genoma mitocondrial



El **genoma mitocondrial** fue la primera parte del genoma de la que se conoció la secuencia, en 1981 (Anderson et al., 1981).

El **ADN mitocondrial (ADNmt)** codifica para la mayor parte de los componentes que son esenciales en la cadena de fosforilación oxidativa, crucial para la generación de adenosín trifosfato (ATP). Por lo tanto, salvo en el caso de algunas levaduras que son capaces de sobrevivir en condiciones anaerobias sin la presencia de este ADN, el ADNmt es esencial para la vida (Sharma & Sampath, 2019).

El ADNmt está constituido por una molécula de ADN circular de cadena doble y de 16.6 kb de longitud. Este es un tamaño pequeño, en comparación con el genoma nuclear.

Contiene 37 genes, de los cuales 2 codifican para ARN ribosómico (ARNr), 22 para ARN de transferencia (ARNt) y 13 para proteínas necesarias para el funcionamiento de la mitocondria. El resto de las proteínas son codificadas por el genoma nuclear.

El 97 % del genoma mitocondrial es codificante y tiene una codificación policistrónica, es decir, se transcribe en forma de transcripto continuo (Anderson et al., 1981).

1.1.1. Genoma mitocondrial

 FOSWIKI

You are here: Foswiki > MITOMAP Web > WebHome (28 Dec 2023, UnknownUser)

Jump Search Edit Attach

MITOMAP

MITOMAP
MITOMASTER
POLG Server
MitoScape

Tools
Help
Search
Index

Service provided by the Center for Mitochondrial & Epigenomic Medicine at the Children's Hospital of Philadelphia

MITOMAP
A human mitochondrial genome database

A compendium of polymorphisms and mutations in human mitochondrial DNA

MITOMAP reports published data on human mitochondrial DNA variation. If you would like to add a paper and its data into MITOMAP, please email a pdf to mitomap@email.chop.edu. We appreciate your help.

2023 Update #2: On July 15, 2023 we added 1,779 new full-length (FL) and 1,410 new control region (CR) GenBank sequences to our database. This brings our total number of FL sequences to 61,168 and the number of CR sequences to 80,294. Our SNVs now total 19,747. We update our GenBank sequences every 4-6 months. Hand curation of variants and references continues weekly. See the [GenBank Frequency Info](#) page for details about our current sequence sets.

★ We have starred our user Favorites for seeking information on specific variants and for understanding the contents of our database.

MITOMAP Quick Reference & Tools

- ★ **Allele Search** - get point mutation data based on position
- ★ **MITOMASTER** - analyze any human mito SNV or nucleotide sequence
- ★ **Tool Launchpad**
- The rCRS is GenBank number **NC_012920.1**. [Click here for details.](#)

General References	Illustrations:
The Annotated Human Mitochondrial DNA Sequence	View Figures
The rCRS & other mtDNAs	★ GenBank Frequency Info
Amino Acid Translation Tables	Mitochondrial DNA Map
Mitochondrial References, ALL	-MitoTIP tRNA Scoring
★ Haplogroup Markers	-Eleven pathological mutations in tRNA
High Frequency Haplogroups	-A-L only
mtDNA Polypeptide Assignments	-M-Z only
	-Mitochondrial energetics
	-Diabetes metabolism & the mitochondria
	-World migrations
	-mtDNA Trees

 7,791 Pageviews Nov 29th - Dec 29th



Enfermedades mitocondriales

Tabla 1*Enfermedades mitocondriales y su localización genómica*

Revisión interesante disponible en aula virtual: Feillet et al., 2014

Enfermedad [código OMIM]	Manifestación clínica	Localización genómica	Prevalencia y edad de debut
Neuropatía óptica hereditaria de Leber (LHON) [308905, 535000]	Insuficiencia visual subaguda bilateral e indolora; pérdida súbita	Mutaciones en los genes <i>MT-ND1</i> , <i>MT-ND4</i> o <i>MT-ND6</i>	1-9/100 000 Adolescencia o edad adulta temprana
Epilepsia mioclónica con fibras rojas rasgadas (MERRF) [545000]	Epilepsia mioclónica, ataxia, debilidad muscular y demencia	Mutaciones en <i>MT-LK</i> (codifica para tRNALys, con la mutación A8344G)	Desconocido Infancia o edad adulta
Síndrome de Pearson [557000]	Anemia sideroblástica y disfunción exocrina del páncreas	Deleción en el ADNmt (4,977 pb)	< 1/1 000 000 Infancia, neonatal
Síndrome de Kearns-Sayre (KSS) [530000]	Oftalmoplejia externa progresiva (OEP), retinosis pigmentaria, pérdida auditiva, ataxia cerebelosa, bloqueo cardiaco	Deleción en el ADNmt (4,977 pb)	1-9/100 000 < 20 años
Encefalopatía mitocondrial, acidosis láctica y episodios similares al ictus (MELAS) [540000]	Encefalomielitis, acidosis láctica y episodios similares a un accidente cerebrovascular; en algunas ocasiones endocrinopatías, enfermedad cardíaca, diabetes, pérdida auditiva y manifestaciones neurológicas y psiquiátricas	Mutaciones puntuales en el gen codificante de tRNAleu (A3243G)	1-9/1 000 000 Infancia, adolescencia

Nota. Adaptado de "Mitochondrial DNA Integrity: Role in Health and Disease", por P. Sharma y H. Sampath, 2019, Cells, 8(2); p. 100 y Portal sobre enfermedades raras y medicamentos huérfanos, por Orphanet, 27 de diciembre de 2021, <https://www.orpha.net/consor4.01/www/cgi-bin/?lng=ES>.

1. Estructura del genoma humano

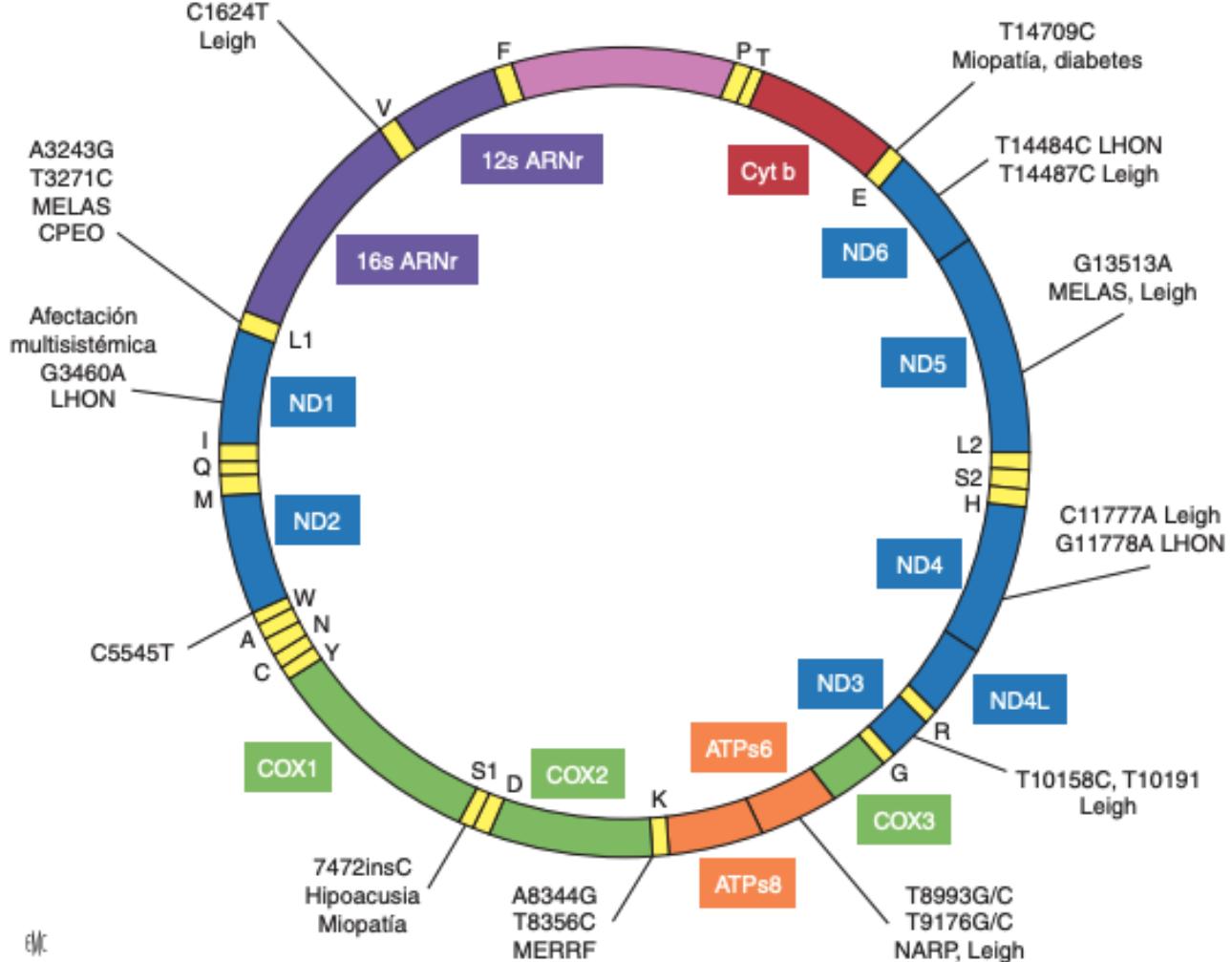


Figura 2. Ácido desoxirribonucleico mitocondrial (ADNm). El genoma mitocondrial es una molécula de ADN circular bicatenaria de 16.569 pares de bases. Esta molécula posee 37 genes que codifican dos ácidos ribonucleicos ribosómicos (ARNr) (violeta), 22 ARN de transferencia (amarillo) y 13 genes que codifican las subunidades proteicas de los complejos de la cadena respiratoria (complejo I: azul; complejo III: rojo; complejo IV: verde; complejo V: naranja). Los ARN de transferencia codifican los aminoácidos siguientes: prolina (P), treonina (T), glutamato (E), leucina (L1 y L2), serina (S1 y S2), histidina (H), arginina (R), glicina (G), lisina (K), aspartato (D), triptófano (W), asparagina (N), tirosina (Y), alanina (A), cisteína (C), isoleucina (I), glutamina (Q), metionina (M), valina (V), fenilalanina (F). Las principales mutaciones del ADNm se describen en el esquema. Una misma mutación puede provocar varios tipos de afecciones y una misma enfermedad puede relacionarse con varias mutaciones. LHON: neuropatía óptica hereditaria de Leber; MELAS: encefalomiopatía, acidosis láctica y episodios de seudoictus; CPEO: oftalmoplejía crónica progresiva; NARP: neuropatía sensitiva, epilepsia y retraso mental.

Nuevo método para secuenciar el genoma mitocondrial humano

A method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome

Ieva Keraite, Philipp Becker, Davide Canevazzi, Cristina Frias-López, Marc Dabad, Raúl Tondá-Hernandez, Ida Paramonov, Matthew John Ingham, Isabelle Brun-Heath, Jordi Leno, Anna Abulí, Elena Garcia-Arumí, Simon Charles Heath, Marta Gut & Ivo Glynne Gut

Nature Communications 13, Article number: 5902 (2022) | [Cite this article](#)

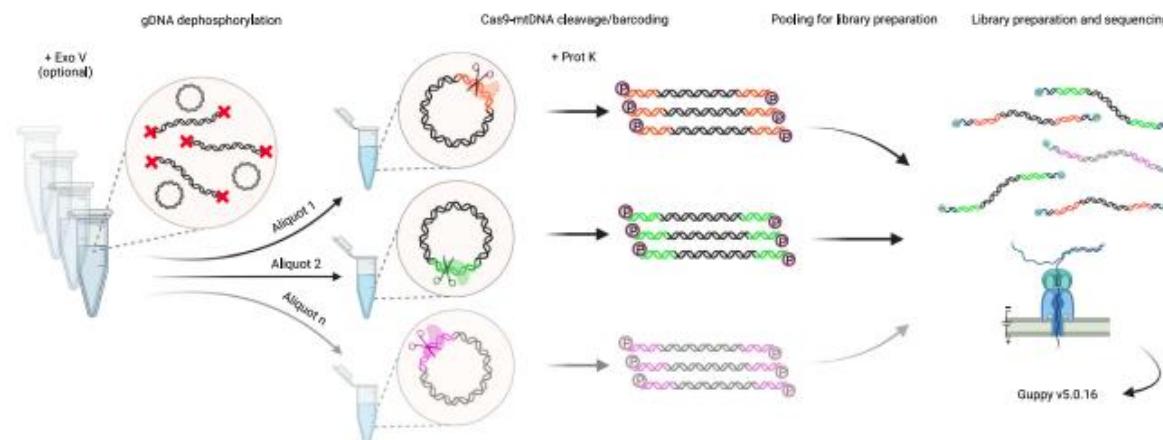


Fig. 1 | Cas9-mtDNA-enrichment, barcoding, pooling and demultiplexing approach for long-read sequencing. A schematic overview of full-length mtDNA targeting with selected dual-guide Cas9 cut-sites. After optional treatment of gDNA with Exonuclease V (Exo V) and dephosphorylation, each sample is split into two or more aliquots for dual-guide targeted cleavage. Each cut-site serves downstream as

a barcode in the analysis pipeline. The circular mtDNA molecules are opened, Cas9 is removed by Proteinase K (Prot K) digestion, followed by mtDNA da-tailing and pooling of all of the aliquots. ONT library is prepared from the pooled samples, sequenced on a nanopore flow cell, followed by basecalling (Guppy v5.0.16). Figure created with BioRender.com.

You are here: Foswiki > MITOMAP Web > MutationsRNACfrm (28 Dec 2023, ShipingZhang)

[Edit](#)[Attach](#)

MITOMAP: Mitochondrial DNA Base Substitution Diseases: rRNA/tRNA Mutations with Cfrm Status

Last Edited: Dec 27, 2023

For Mitomap to assign a status of "Cfrm" to a possibly pathogenic variant, we look for confirming reports which address the criteria outlined in [Mitchell et al 2006](#), [Yarham et al 2011](#), [Wong 2007](#), and [Gonzalez-Viogue et al 2014](#). These criteria include the following: (1) independent reports of two or more unrelated families with evidence of similar disease; (2) evolutionary conservation of the nucleotide (for RNA variants) or amino acid (for coding variants); (3) presence of heteroplasmcy; (4) correlation of variant with phenotype / segregation of the mutation with the disease within a family; (5) biochemical defects in complexes I, III, or IV in affected or multiple tissues; (6) functional studies showing differential defects segregating with the mutation (cybrid or single fiber studies); (7) histochemical evidence of a mitochondrial disorder; and (8) for fatal or severe phenotypes, the absence or extremely rare occurrence of the variant in large mtDNA sequence databases. For the exact scoring systems of Yarham et al 2011 and Mitchell et al 2006, please see the respective papers above. A new scoring system is under development for these criteria, and will be linked here once published.

When investigating a novel variant or a variant of uncertain significance (VUS) in cases of suspected mitochondrial disease, further studies which address any or all of the above criteria are strongly recommended. We encourage publication of such case reports as they are extremely helpful to the mitochondrial research community.

The GB frequency data in Mitomap is derived from **61168** GenBank sequences with size greater than 15.4kbp and **80294** Control Region sequences with size 0.4-1.6kbp. These sequences have been pre-loaded into Mitomaster and represent almost all haplogroups known to date. We will be updating and refining this set of sequences on a regular basis. As a caveat, please note that GenBank sequences may not be of equal quality ([Yao, et al, 2009](#)), that some of these sequences are from individuals with past, current or future disease, and that this portion of our data set has not been hand-curated by Mitomap.

For more details about the current GenBank sequence set, please see [GB Frequency Info](#).

For information about the predictive MitoTIP scoring for tRNA variants, see [MitoTIP Info](#).

[https://www.mitomap.org
/foswiki/bin/view/MITOMAP/MutationsRNACfrm](https://www.mitomap.org/foswiki/bin/view/MITOMAP/MutationsRNACfrm)

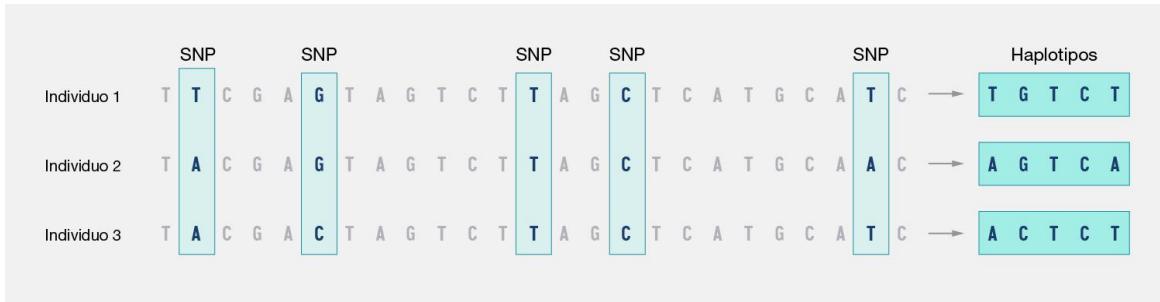
1. Estructura del genoma humano

Position	Locus	Disease	Allele	RNA	Homo plasmcy	Hetero plasmcy	Status (Mitomap [ClinGen])	MitoTIP ↑	GB Freq FL (CR) *‡	GB Seqs FL (CR) *	References
583	MT-TF	MELAS / MM & EXIT	G583A	tRNA Phe	-	+	Cfrm [VUS*]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	5
616	MT-TF	Maternally inherited epilepsy / mito tubulointerstitial kidney disease (MITKD) / Gitelman-like syndrome	T616C	tRNA Phe	+	+	Cfrm [LP]	Pathogenic ↑↑↑	0.002% (0.000%)	1 (0)	6
1494	MT-RNR1	DEAF	C1494T	12S rRNA	+	-	Cfrm [LP]	N/A	0.008% (0.000%)	5 (0)	33
1555	MT-RNR1	DEAF; autism spectrum intellectual disability; possibly antiatherosclerotic	A1555G	12S rRNA	+	+	Cfrm [P]	N/A	0.141% (0.000%)	86 (0)	158
1606	MT-TV	AMDF	G1606A	tRNA Val	-	+	Cfrm [VUS*]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	5
1630	MT-TV	MNGIE-like disease / MELAS	A1630G	tRNA Val	-	+	Cfrm [VUS*]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	9
1644	MT-TV	Leigh Syndrome / HCM / MELAS	G1644A	tRNA Val	-	+	Cfrm [LP]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	6
3243	MT-TL1	MELAS / Leigh Syndrome / DMDF / MIDD / SNHL / CPEO / MM / FSGS / ASD / Cardiac+multi-organ dysfunction	A3243G	tRNA Leu (UUR)	-	+	Cfrm [P]	Pathogenic ↑↑↑	0.016% (0.000%)	10 (0)	468
3243	MT-TL1	MM / MELAS / SNHL / CPEO	A3243T	tRNA Leu (UUR)	-	+	Cfrm [LP]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	12
3256	MT-TL1	MELAS; possible atherosclerosis risk	C3256T	tRNA Leu (UUR)	-	+	Cfrm [LP]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	21
3258	MT-TL1	MELAS / Myopathy	T3258C	tRNA Leu (UUR)	-	+	Cfrm [LP]	Pathogenic ↑↑↑	0.002% (0.000%)	1 (0)	5
3260	MT-TL1	MMC / MELAS	A3260G	tRNA Leu (UUR)	-	+	Cfrm [LP]	Pathogenic ↑↑↑	0.000% (0.000%)	0 (0)	18

<https://www.mitomap.org/foswiki/bin/view/MITOMAP/MutationsRNACfrm>

Haplótipos & Haplogrupos

Haplótipos: grupo de alelos de diferentes lugares (loci) de un cromosoma que se heredan juntos. --> Un haplotipo es una **agrupación física de variantes genómicas (o polimorfismos) que tienden a heredarse juntas**. Un haplotipo específico por lo general refleja una combinación única de variantes que residen una cerca de la otra en un cromosoma.



Un haplotipo es, en su sentido más general, un conjunto de variaciones de ADN a lo largo de un cromosoma que tienden a ser heredados juntos porque están muy próximos. El motivo por el cual se heredan juntos es porque no suele haber cruzamientos o recombinaciones entre estos marcadores, o diferentes polimorfismos, al estar tan cerca. Así que un haplotipo se puede referir a una combinación de alelos en un solo gen, o alelos en múltiples genes. Podrían ser polimorfismos de nucleótido sencillo que no están en un gen, sino en la región intergénica. Básicamente, sólo significa que se trata de variaciones en el ADN que están tan juntas que no tienen tendencia a recombinarse, y por lo tanto tienden a ser transmitidas juntas a través de generaciones. El **Proyecto Internacional HapMap** nos ha dado una herramienta excelente para la detección de estas regiones de haplotipos que se transmiten juntas y utilizarlas en estudios genéticos.

Haplotipos & Haplogrupos

Haplogrupo : grupo grande de haplotipos que se heredan juntos y que comparten un ancestro común. Son una única línea hereditaria. Se estudian para trazar la evolución molecular de una especie y trazar el árbol genealógico de nuestra especie.

Haplogrupos en genética humana:

- ADNmt: haplogrupos de ADN mitocondrial transmitidos a través de la madre (matrilineal) [Eva-mitocondrial]
- ADN-Y : haplogrupos del cromosoma Y humano, linaje patrilineal. [Adán cromosomal-Y]

Tradicionalmente se ha hecho el estudio de polimorfismos por restricción enzimática (RFLP), donde las dianas de corte de las enzimas de restricción nos darán unos patrones de corte específicos. Hoy en día, se utiliza la secuenciación masiva para la secuenciación de los lugares polimórficos.

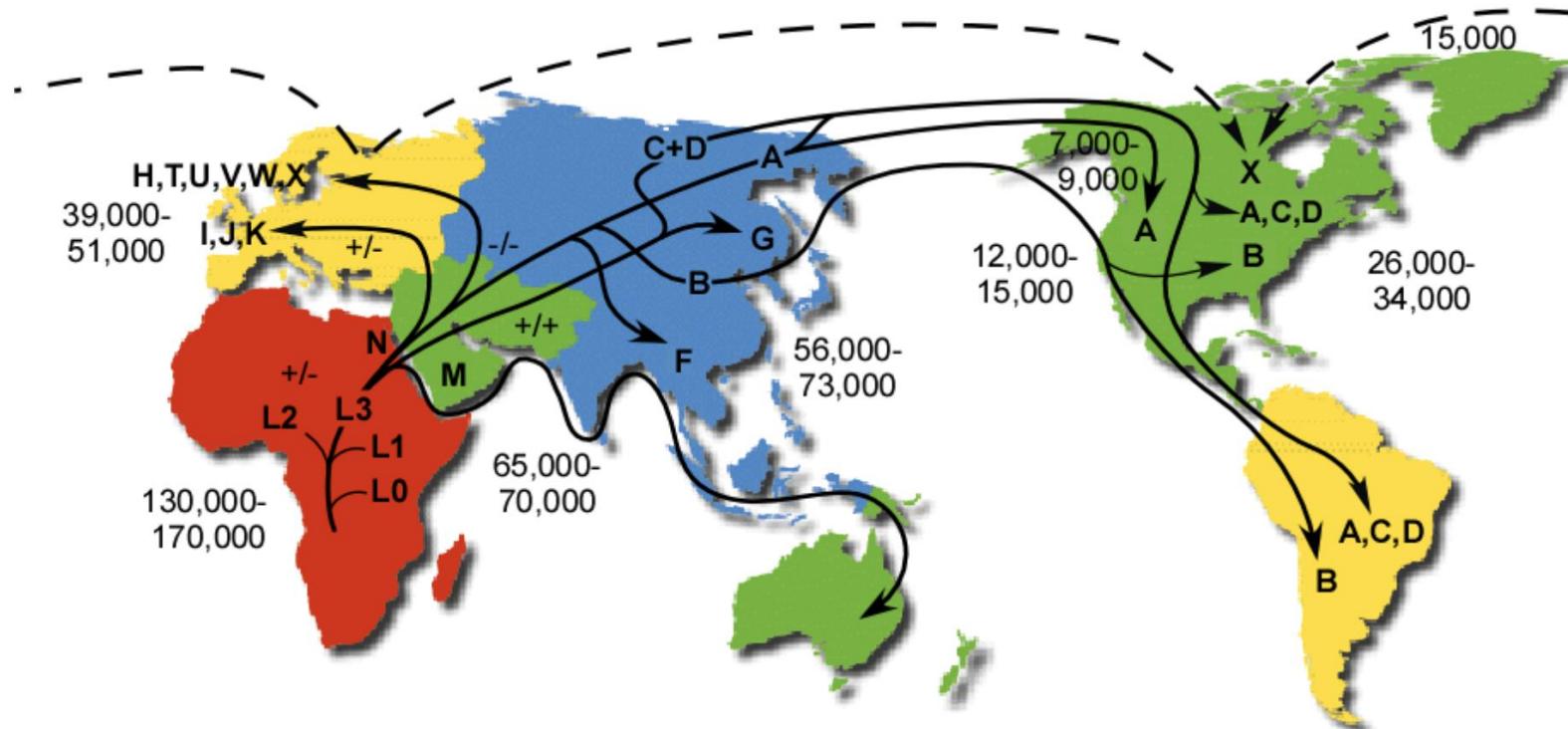
¿Por qué se eligen Polimorfismos SNPs? : baja probabilidad de desaparición; acumulación a lo largo del tiempo en la población.

<https://www.mitomap.org/foswiki/bin/view/MITOMAP/HaplogroupMarkers>

Lineage	Top Level Haplogroup	# Seqs	Ancestral Marker Motif † (the "RSRS50")	HG Markers	Other Selected Markers
L	L0	1674	247A, 750G, 769A, 825A, 1018A, 2758A, 2885C, 3594T, 4104G, 4312T, 4769G, 7028T, 7146G, 7256T, 7521A, 8468T, 8655T, 8701G, 8860G, 9540C, 10398G, 10664T, 10688A, 10810C, 10873C, 10915C, 11719A, 11914A, 12705T, 13105G, 13276G, 13506T, 13650T, 14766T, 15326G, 16187T, 16189C, 16223T, 16230G, 16311C	1048T, 3516A, 5442C, 6185C, 9042T, 9347G, 9755A, (263A=) (522d, 4586C, 8251A, 9818T, 16172C)	(73G, 148C, 152C, 195C, 1438G, 2706G, 16129A, 16519C)
L	L1	851	73G, 152C, 247A, 263G, 750G, 769A, 825A, 1018A, 2706G, 2758A, 2885C, 3594T, 4104G, 4769G, 7028T, 7146G, 7256T, 7521A, 8468T, 8655T, 8701G, 8860G, 9540C, 10398G, 10668A, 10810C, 10873C, 11719A, 12705T, 13105G, 13506T, 13650T, 14766T, 151T, 186A, 189C, 316A, 522d, 2395d, 5951G, 6071C, 15326G, 16187T, 16189C, 16223T, 16278T, 16311C, 16519C	182T, 3666A, 7055G, 7389C, 13789C, 14178C, 14560A 8027A, 9072G, 10321C, 10586A, 12810G, 13485G, 14000A, 14911T, 16294T, 16360T)	(195C, 1438G, 16129A)
L	L2	1325	73G, 146C, 152C, 195C, 263G, 750G, 769A, 1018A, 1438G, 2706G, 3594T, 4104G, 4769G, 7028T, 7256T, 7521A, 8701G, 8860G, 9540C, 10398G, 10873C, 11719A, 12705T, 13650T, 14766T, 15326G, 16223T, 16278T	2416C, 8206A, 9221G, 10115C, 11944C, 13590A, 15301A, 15 16390A (2789T, 7175C, 7274T, 7771G, 12693G, 13803G, 14566G, 15784C, 16294T, 16309G)	N P 413 73G, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G 15607G N R 1185 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 14766T, 15326G (73G, 11719A, 16519C)
L	L3	2140	73G, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8701G, 8860G, 9540C, 10398G, 10873C, 11719A, 12705T, 14766T, 15326G, 16223T	15301A (150T)	35 16 (16519C)
L	L4	111	73G, 263G, 750G, 769A, 1018A, 1438G, 2706G, 4769G, 7028T, 8701G, 8860G, 9540C, 10398G, 10873C, 11719A, 12705T, 14766T, 15326G, 16223T, 16311C	3918A, 15301A, 16362C (244G, 315CC, 1413C, 8104C, 9855G, 12609C, 13470G, 16293T, 16355T, 16399G)	35 49 73G, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 8404C (152C, 16223T)
L	L5	37	73G, 247A, 263G, 750G, 769A, 825A, 1018A, 1438G, 2706G, 3594T, 4104G, 4769G, 7028T, 7256T, 7521A, 8655T, 8701G, 8860G, 9540C, 10398G, 10873C, 11719A, 12705T, 13506T, 13650T, 14766T, 15326G, 16129A, 16223T, 16278T, 16311C	182T, 3423C, 5147A, 7972G, 12432T, 12950G, 14581C, 16148T, 16166G (522d, 709A, 851G, 1822C, 511T, 5656G, 6182A, 6297C, 7424G, 8155A, 8188G, 8582T, 9305A, 9329A, 11025C, 11881T, 12236A, 13722G, 14212C, 14239T, 14905A, 14971C, 15217A, 15884A, 16355T, 16362C)	Le ref 45 (11812G, 14233G)
L	L6	12	73G, 146C, 152C, 263G, 750G, 769A, 1018A, 1438G, 2706G, 3594T, 4769G, 7028T, 7256T, 8701G, 8860G, 9540C, 10398G, 10873C, 11719A, 12705T, 13650T, 14766T	182T, 185C, 709A, 770T, 961C, 1461G, 4964T, 5267C, 6002G, 6284G, 9332T, 10978G, 1116C, 11743T, 12771A,	189G, 204C, 207A, 709A, 1243C, 3505G, 5046A, 5460A, 8251A, 8994A, 11674T, 11947G, 12414C, 15884C, 16292T (194T)
					N U 5145 73G, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G 11467G, 12308G, 12372A N V 827 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 15326G 72C, 4580A, 15904T, 16298C (310C) N W 620 73G, 195C, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 15326G 189G, 204C, 207A, 709A, 1243C, 3505G, 5046A, 5460A, 8251A, 8994A, 11674T, 11947G, 12414C, 15884C, 16292T N X 643 73G, 195C, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 15326G 153G, 1719A, 6221C, 6371T, 13966G, 14470C N Y 194 73G, 263G, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 10398G, 11719A, 12705T, 14766T, 15326G 5417A, 8392A, 14178C, 14693G, 16126C, 16231C (225A) (310C, 3834A, 7933G, 16266T) (146C, 16519C)

Human mtDNA Migrations

From <http://www.mitomap.org>



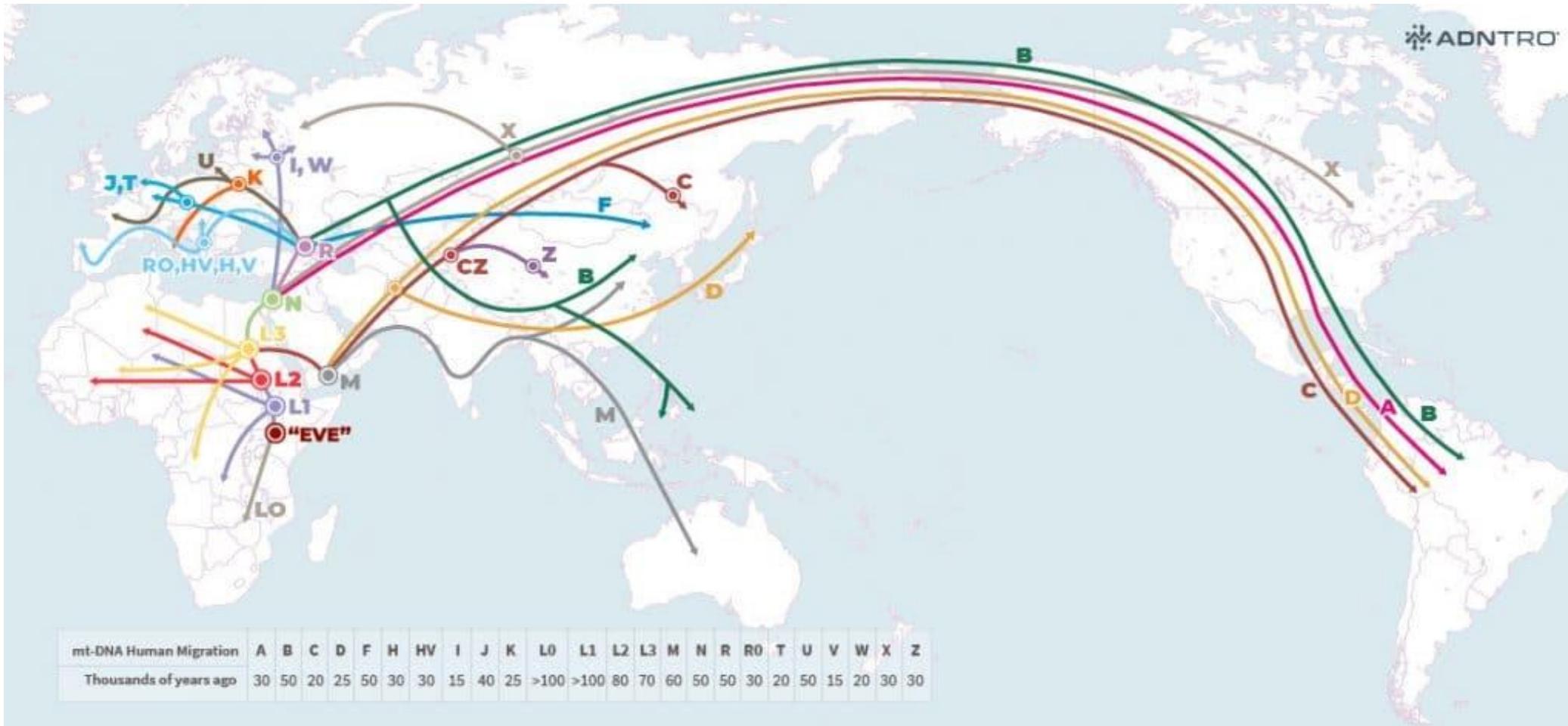
Symbols +/-, +/+, and -/- represent
RFLP status for Dde I 10394 / Alu I 10397

Mutation rate = 2.2 - 2.9 % / MYR
Time estimates are YBP

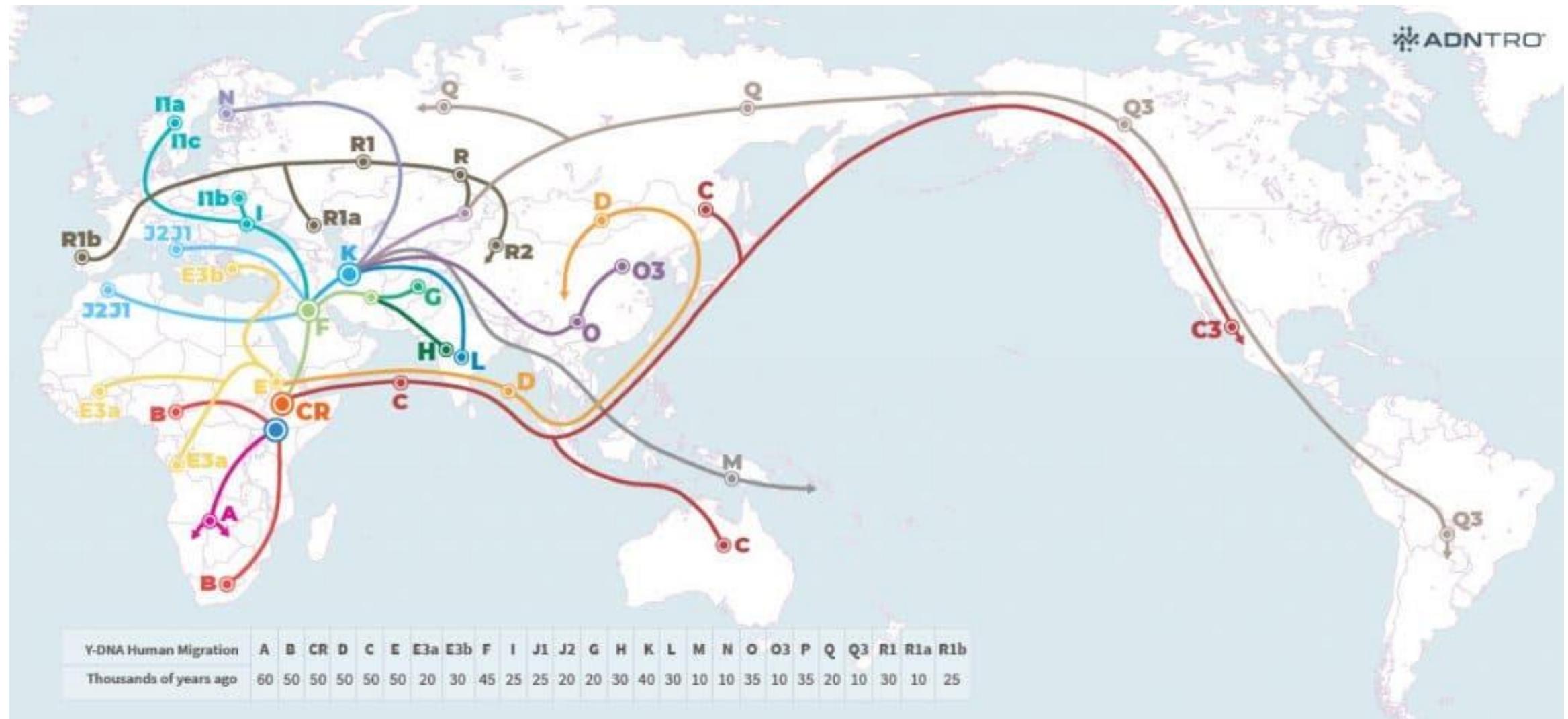


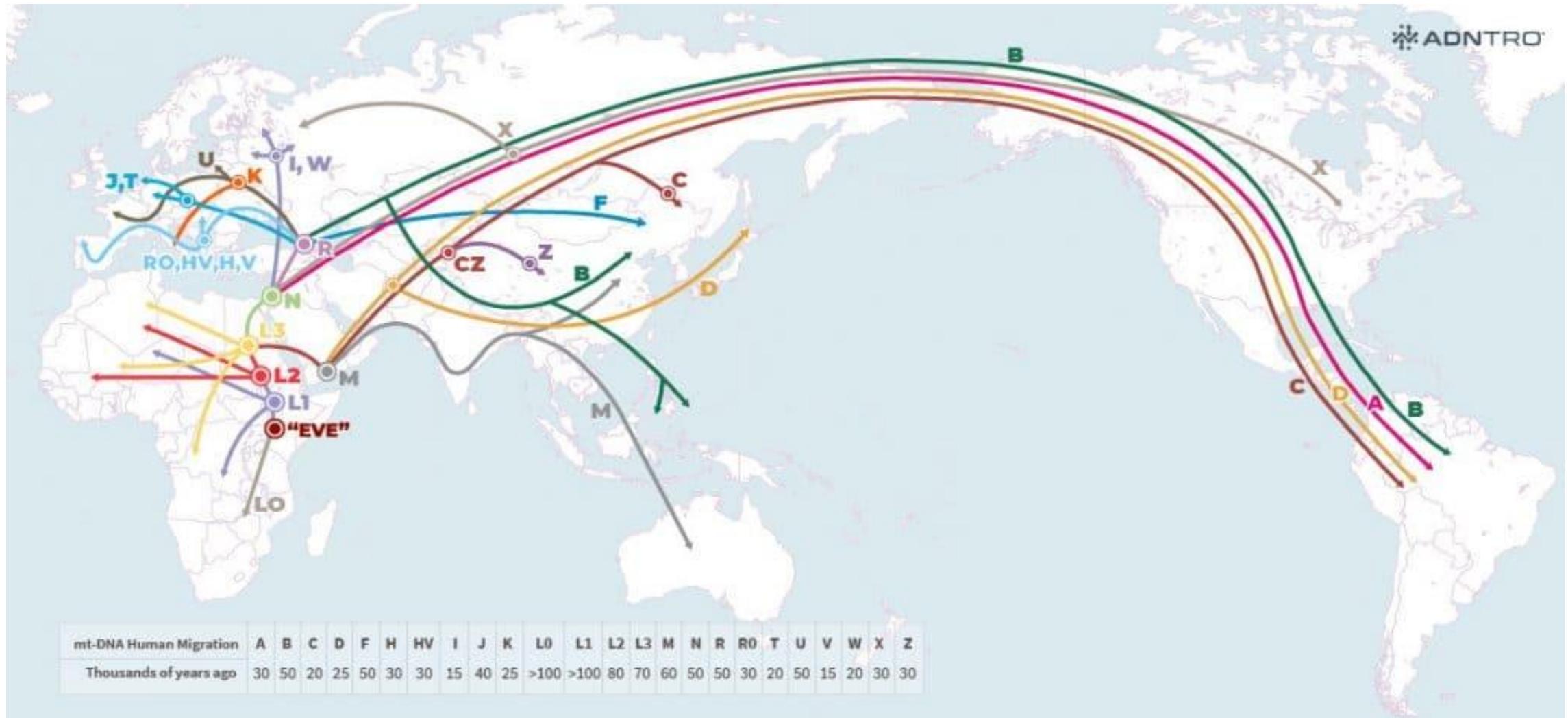
This is licensed by a Creative Commons Attribution 3.0 license.

<https://www.mitomap.org/foswiki/pub/MITOMAP/MitomapFigures/mitomap-phylogeny.pdf>
02/07/2024



<https://adntro.com/es/blog/ancestria/haplogrupos-la-huella-de-un-gran-viaje/>





Lo que nos cuenta de ti el ADN

La genética y bioinformática nos permiten preguntar cosas distintas a nuestro ADN. Una de ellas consiste en escudriñar en nuestro pasado como especie.

Por Ana M. Rojas Mendoza a.rojas.m@csic.es

Científica Titular, CSIC. Computational Biology and Bioinformatics Group. Centro Andaluz de Biología del Desarrollo. CABD-CSIC.

Recurso interesante:

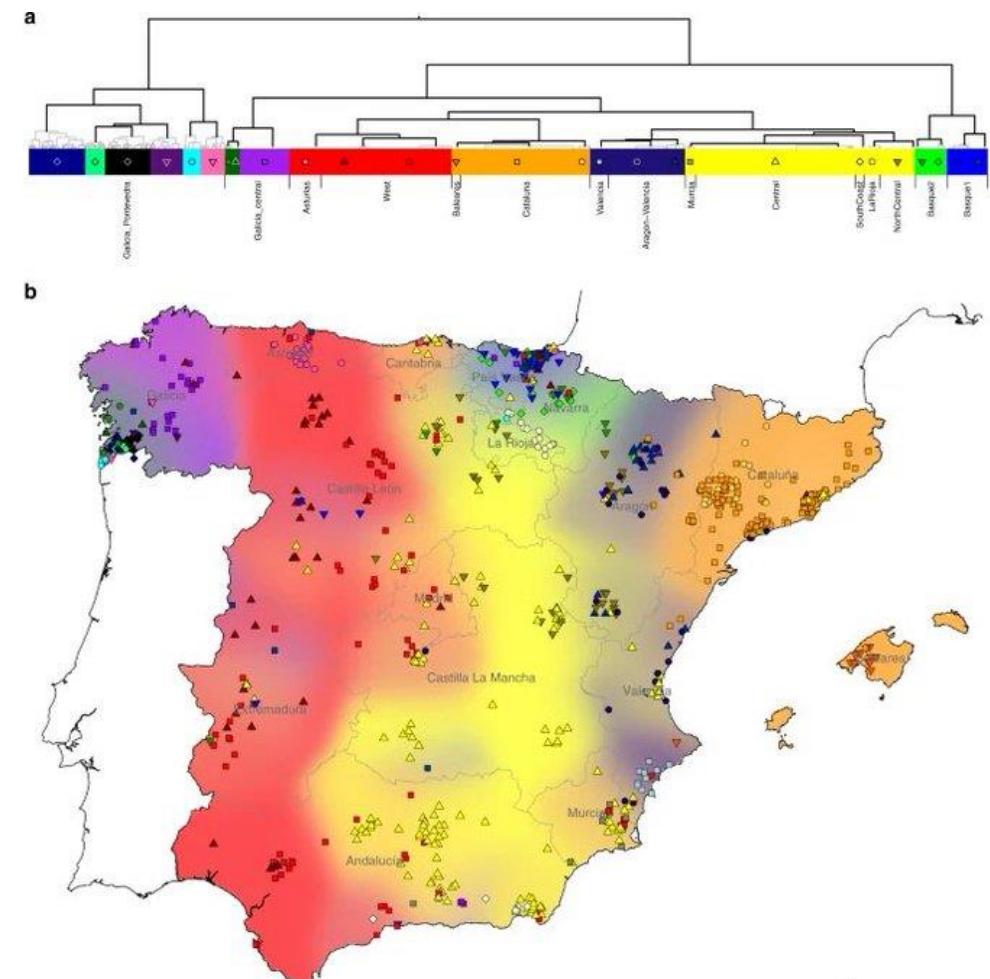
<https://sebbm.es/rincon-del-aula/lo-que-nos-cuenta-de-ti-el-adn/>

Con las referencias:

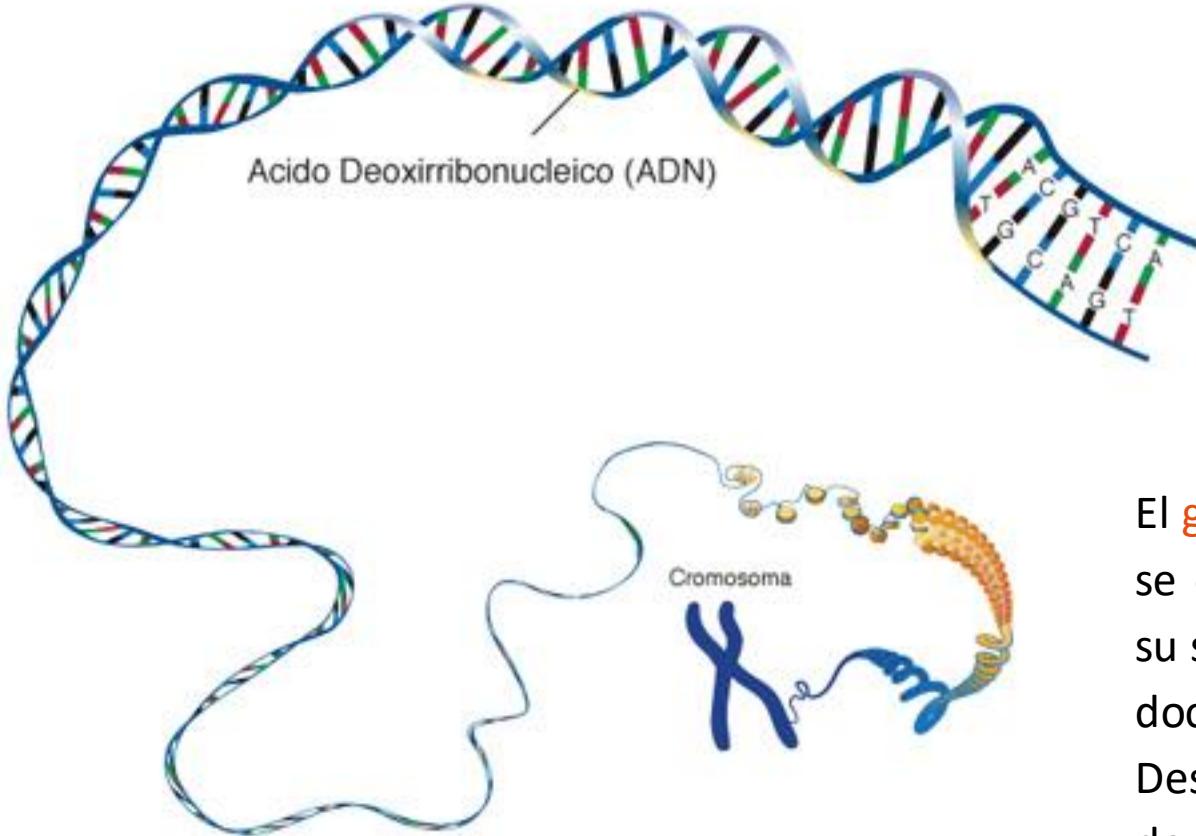
Olalde I, et al., *The genomic history of the Iberian Peninsula over the past 8000 years*. *Science*. 2019 Mar 15;363(6432):1230-1234. doi: 10.1126/science.aav4040. PMID: 30872528; PMCID: PMC6436108.

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C. et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* **10**, 551 (2019).

<https://doi.org/10.1038/s41467-018-08272-w>



1.1.2. Genoma nuclear



El **genoma nuclear**, tal y como su nombre indica, es el ADN que se encuentra dentro del núcleo celular. El primer borrador de su secuencia y estructura se obtuvo en el año 2001, tras más de doce años de intenso trabajo en el Proyecto Genoma Humano. Desde ese momento, esta información se ha ido actualizando de manera constante.

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Login/Register

 Search Human... 

Search Human (Homo sapiens)

Search all categories

▼ Search...

Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

 More information and statistics

 Download DNA sequence (FASTA)

 Convert your data to GRCh38 coordinates

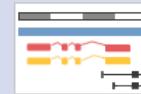
 Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go



View karyotype



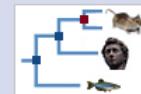
Example region

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

 More about comparative analysis

 Download alignments (EMF)



Example gene tree

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

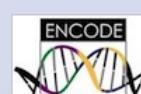
 More about the Ensembl regulatory build and microarray annotation

 Experimental data sources

 Download all regulatory features (GFF)



Example regulatory feature



Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

 More about this genebuild

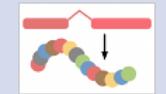
 Download FASTA files for genes, cDNAs, ncRNA, proteins

 Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

 Update your old Ensembl IDs

Pax6 INS
FOXP2
BRCA2
DMD
ssh

Example gene



Example transcript

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

 More about variation in Ensembl

 Download all variants (GVF)

 Variant Effect Predictor

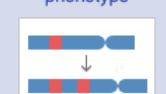


ATCGAGCT
ATCCAGCT
ATCGAGAT

Example variant



Example phenotype



Example structural variant

1.1.2. Genoma nuclear



El genoma nuclear está dividido en un conjunto de moléculas de **ADN lineares**, cada una de ellas contenida en un cromosoma. No existen excepciones a este patrón en los genomas eucariotas, todos los genomas eucariotas conocidos tienen al menos dos cromosomas y la molécula de ADN es siempre linear.

La única variación a este respecto está en el número de cromosomas. Por ejemplo, una levadura tiene 16 cromosomas, mientras que el genoma humano está compuesto por 23 pares de cromosomas.

Por otra parte, **este número tampoco está relacionado con el tamaño del genoma**. Algunas salamandras tienen un genoma 30 veces superior al genoma humano, pero dividido en la mitad de los cromosomas. Aunque estas comparaciones son muy curiosas, realmente no nos dicen nada sobre los genomas, sino que simplemente nos dan idea de la falta de uniformidad de los eventos evolutivos que han moldeado la arquitectura genómica en los distintos organismos eucariotas.

1.1.2. Genoma nuclear

Tabla 2

Tamaño de distintos genomas eucariotas.

Actualización 30/12/2023

Especie	Tamaño genoma	Num. Chr	Número de genes
Hongos			
<i>Saccharomyces cerevisiae</i> (taxonomy ID 4932)	12.16 Mb	16	6 463
<i>Aspergillus nidulans</i>	29.83 Mb	8*	9 586
Protozoos			
<i>Tetrahymena pyriformis</i> ***	116 Mb	**	26 866
Invertebrados			
<i>Caenorhabditis elegans</i>	100.3 Mb	6 + Mt	46 926
<i>Drosophila melanogaster</i>	143.7 Mb	7 + Mt	17 894
<i>Bombyx mori</i> (gusano de seda)	460.3 Mb	28 + Mt	17 047
<i>Strongylocentrotus purpuratus</i> (erizo de mar morado)	921.8 Mb	Un	33 503
<i>Locusta migratoria</i> (langosta migratoria)	6.3 Gb	12 + Mt	**
Vertebrados			
<i>Takifugu rubripes</i> (pez puffer /torafugu)	384.1 Mb	22 + Mt	27 412
<i>Homo sapiens</i>	3.1 Gb	23 + Mt	59 265
<i>Mus musculus</i>	2.7 Gb	20 + Mt	50 561
Plantas			
<i>Arabidopsis thaliana</i>	119.1 Mb	5 + Mt +Pltd	38 312
<i>Oryza sativa</i> (arroz)	373.8 Mb	12 + Mt +Pltd + B1	35 223
<i>Zea mays</i> (maíz)	2.2 Gb	10 + Mt + Pltd	49 897
<i>Pisum sativum</i> (guisante)	3.8 Gb	7 + Mt	65 672
<i>Triticum aestivum</i> (trigo)	14.6 Gb	21 + Mt	155 463
<i>Fritillaria assyriaca</i> (lila)	120 Gb	**	**

Nota. Información adaptada de Genomes 4, por T. A. Brown, 2017, y Home - Genome – NCBI, 2021 (<https://www.ncbi.nlm.nih.gov/genome/>).* Aún no hay ningún genoma depositado en <https://www.ncbi.nlm.nih.gov/genome/> que se encuentre cerrado completamente.** No está contenido ningún genoma ni completo ni parcial en <https://www.ncbi.nlm.nih.gov/genome/>.

Un : no se encuentra ensamblado su genoma.

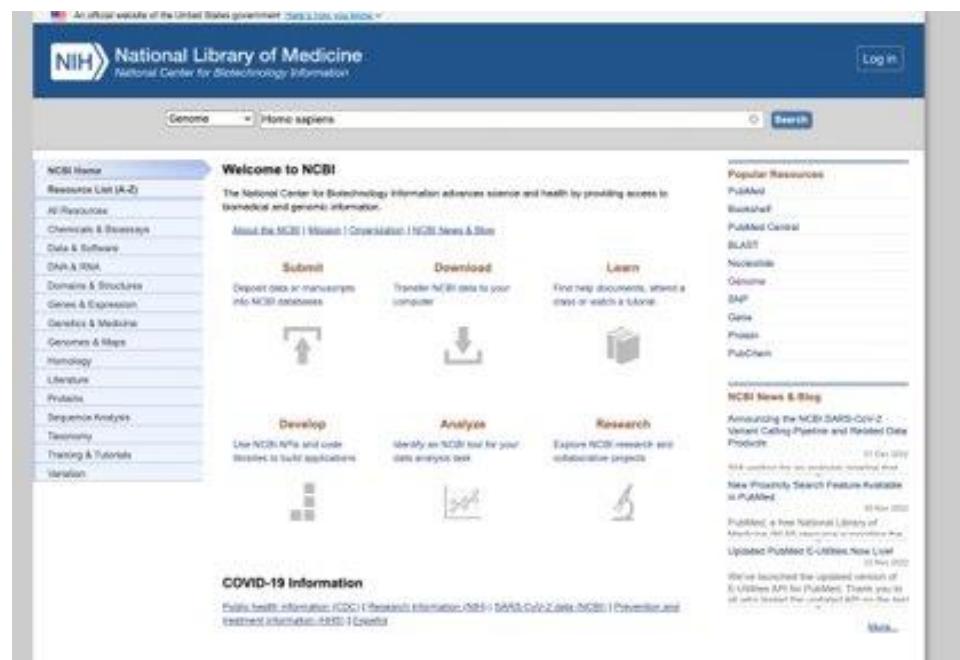
***<http://ciliate.ihb.ac.cn/tcgd/database/home/>

Tabla 2023

Ejercicio!!!!

Vamos a consultar la web de NCBI y vamos a buscar las características de los siguientes genomas:

- *Homo sapiens*
- *Saccharomyces cerevisiae*
- *Mus musculus*
- *Caenorhabditis elegans*



Indica:

- Tamaño medio de genoma
- Número de cromosomas
- Número de genes
- Número de genomas disponibles secuenciados

<https://www.ncbi.nlm.nih.gov/>



Genome assembly GRCh38.p14

reference

Download

datasets

curl

Reference sequence	RefSeq GCF_000001405.40	⋮
Submitted sequence	GenBank GCA_000001405.29	⋮
Taxon	<i>Homo sapiens</i> (human)	
Synonym	hg38	
Assembly type	haploid-with-alt-loci	
Submitter	Genome Reference Consortium	
Date	Feb 3, 2022	

[View the legacy Assembly page](#)

Assembly statistics

These statistics describe the RefSeq genome sequence GCF_000001405.40

Genome size	3.1 Gb
Total ungapped length	2.9 Gb
Gaps between scaffolds	349
Number of chromosomes	24
Number of scaffolds	470
Scaffolds	67.0 Mb

Additional Genomes

[Browse all *Homo sapiens* genomes \(1084\)](#)

BioProject

[PRJNA31257](#)

The Human Genome Project, currently maintained by the Genome Reference Consortium (GRC)

Publications

Showing 5 of 183

Genome Biol 2008

[Finishing the finished human chromosome 22 sequence](#)

CG Cole, et al.

Nature 2006

[The DNA sequence and biological annotation of human chromosome 1](#)

SG Gregory, et al.

Nature 2006

[The DNA sequence, annotation and analysis of human chromosome 3](#)

DM Muzny, et al.

Nature 2006

[DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage](#)

Assembly statistics

These statistics describe the RefSeq genome sequence GCF_000001405.40

Genome size	3.1 Gb
Total ungapped length	2.9 Gb
Gaps between scaffolds	349
Number of chromosomes	24
Number of scaffolds	470
Scaffold N50	67.8 Mb
Scaffold L50	16
Number of contigs	996
Contig N50	57.9 Mb
Contig L50	18
GC percent	40.5
Assembly level	Chromosome

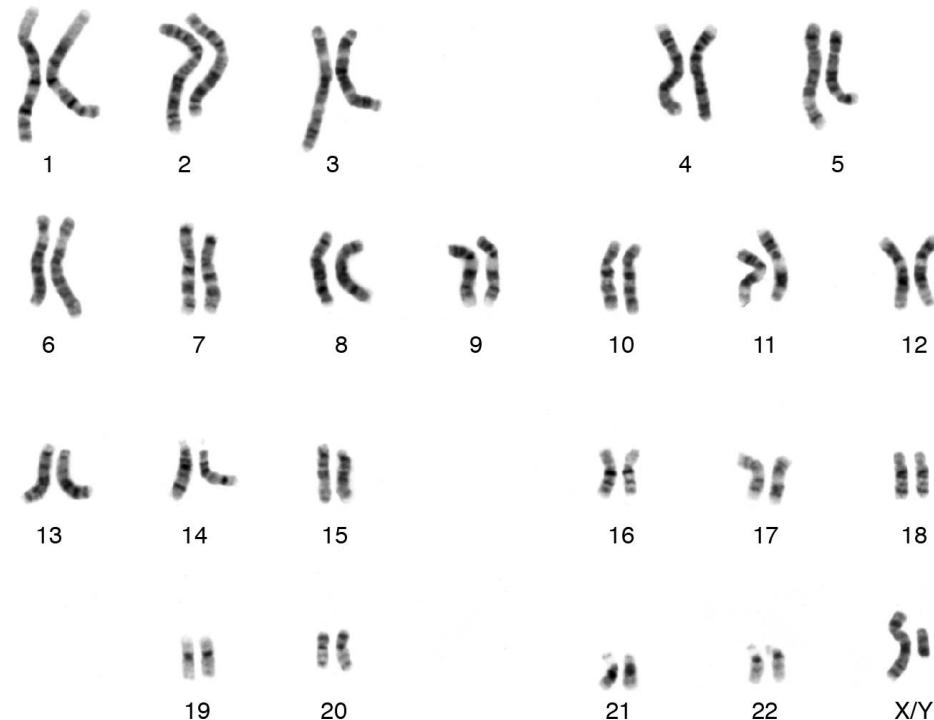
Assembly methods

Comment	The DNA sequence is composed of genomic sequence, primarily finished clones that were sequenced as part of the Human Genome Project. PCR products and WGS shotgun sequence have been added where necessary to fill gaps or correct errors. All such additions are manually curated by GRC staff. For more information see: http://genomereference.org
---------	--

Revision history

GenBank	RefSeq	Name	Level	Date	Action
GCA_000001405.29	GCF_000001405.40	GRCh38.p14	Chromosome	Feb 3, 2022	⋮
GCA_000001405.28	GCF_000001405.39	GRCh38.p13	Chromosome	Feb 28, 2019	⋮
GCA_000001405.27	GCF_000001405.38	GRCh38.p12	Chromosome	Dec 21, 2017	⋮
GCA_000001405.26	GCF_000001405.37	GRCh38.p11	Chromosome	Jun 14, 2017	⋮
GCA_000001405.25	GCF_000001405.36	GRCh38.p10	Chromosome	Jan 6, 2017	⋮
GCA_000001405.24	GCF_000001405.35	GRCh38.p9	Chromosome	Sep 26, 2016	⋮
GCA_000001405.23	GCF_000001405.34	GRCh38.p8	Chromosome	Jun 30, 2016	⋮
GCA_000001405.22	GCF_000001405.33	GRCh38.p7	Chromosome	Mar 21, 2016	⋮
GCA_000001405.21	GCF_000001405.32	GRCh38.p6	Chromosome	Dec 21, 2015	⋮
GCA_000001405.20	GCF_000001405.31	GRCh38.p5	Chromosome	Sep 22, 2015	⋮
GCA_000001405.19	GCF_000001405.30	GRCh38.p4	Chromosome	Jun 25, 2015	⋮
GCA_000001405.18	GCF_000001405.29	GRCh38.p3	Chromosome	Apr 3, 2015	⋮
GCA_000001405.17	GCF_000001405.28	GRCh38.p2	Chromosome	Dec 5, 2014	⋮
GCA_000001405.16	GCF_000001405.27	GRCh38.p1	Chromosome	Oct 3, 2014	⋮
GCA_000001405.15	GCF_000001405.26	GRCh38	Chromosome	Dec 17, 2013	⋮
GCA_000001405.14	GCF_000001405.25	GRCh37.p13	Chromosome	Jun 28, 2013	⋮

1.1.3. ¿Qué tipos de secuencias encontramos en el genoma humano?



En el caso del genoma humano ya hemos dicho que la secuencia de ADN está contenida en 23 cromosomas en el núcleo de cada célula humana. De los 23 pares, 22 son cromosomas autosómicos y 2 son los determinantes del sexo (dos cromosomas X en mujeres y un cromosoma X y un Y en hombres).

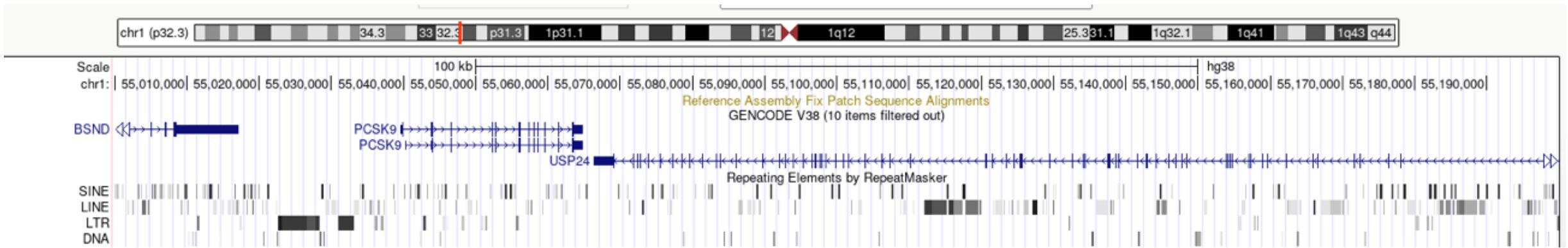
El genoma humano disponible actualmente en la base de datos Ensembl corresponde a la versión **GRCh38.p14**, cuya última actualización se realizó **en febrero de 2022**. Este genoma haploide (una sola representación por cada par) tiene un tamaño de 3.1 Gb (2.9 Gb si eliminamos gaps) y consta de las características siguientes.

Tabla 3

Características del genoma humano recogidas en la base de datos Ensembl.

Genoma: GRCh38.p14		
Pares de bases (pb)	3 088 286 401 (3,1 Gb)	
Ensamblaje primario	Genes codificantes	19 813
	Genes no-codificantes	25 972
	- Cortos	4 864
	- Largos	18 887
	- Miscelánea	2 221
	Pseudogenes	15 241
	Transcritos	252 477
Ensamblaje alternativo	Genes codificantes	3 028
	Genes no-codificantes	1 682
	- Cortos	297
	- Largos	1 198
	- Miscelánea	187
	Pseudogenes	1 796
	Transcritos	21 630

fuente: https://www.ensembl.org/Homo_sapiens/Info/Annotation



Análisis de un segmento del genoma humano, que comprende la región 55 000 000 a 55 200 000. Los genes se muestran en azul, con los exones representados como cajas azules y los intrones como líneas azules. En la parte inferior se representan los elementos de repetición más comunes (LINE, SINE, LTR y transposones ADN).

Nota. Figura tomada de UCSC Genome Browser, Hg38.

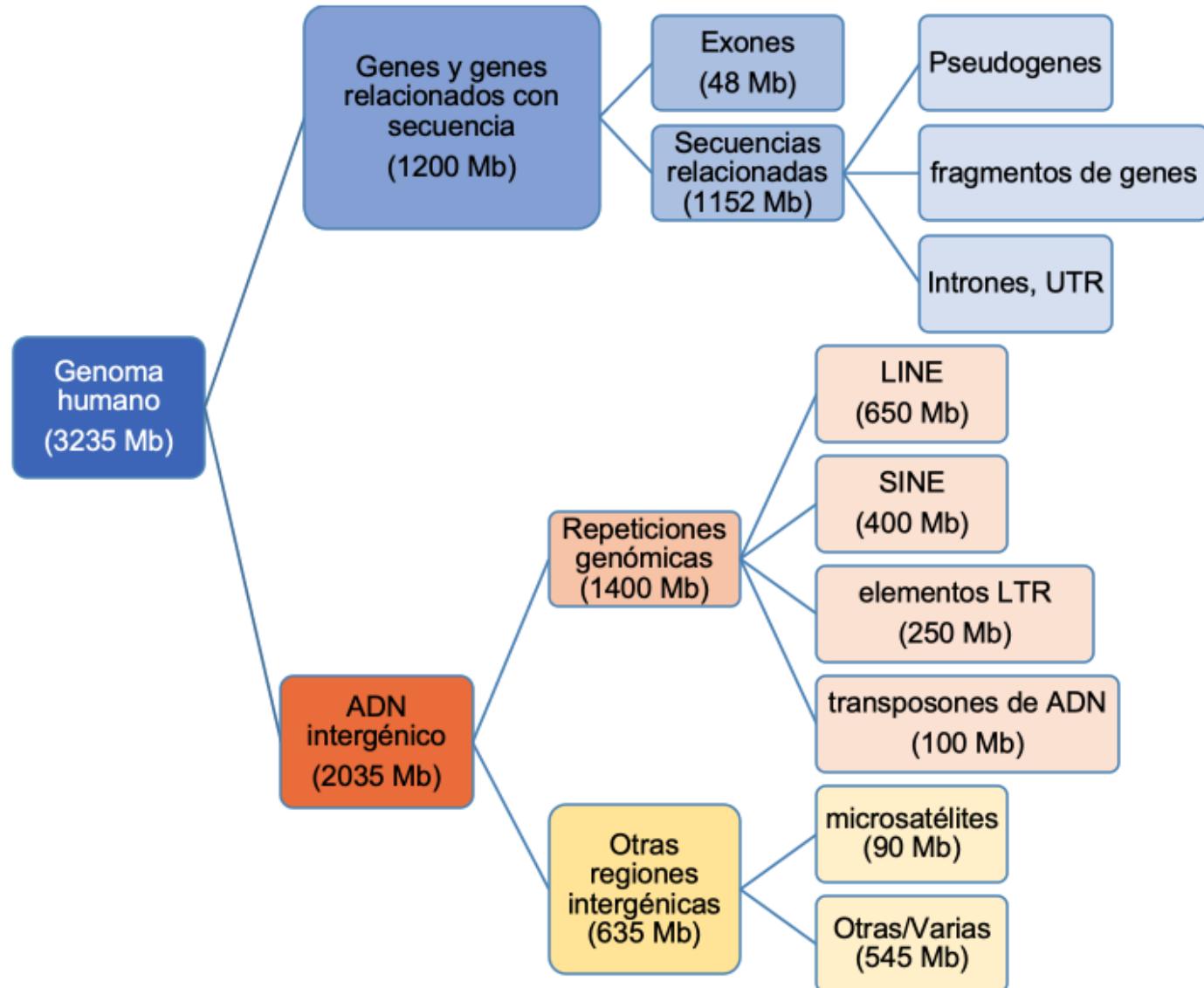
1.1.3. ¿Qué tipos de secuencias encontramos en el genoma humano?

Actualmente conocemos que genoma humano presenta una densidad de genes muy inferior a lo que se pensaba, solo un 1,5 % de su longitud está compuesto por exones que codifican proteínas. Hay un 70 % compuesto por ADN extragénico y un 30 % por secuencias relacionadas con genes. Del total de ADN extragénico, aproximadamente el 70 % corresponde a repeticiones dispersas, con lo que aproximadamente el 50 % del genoma humano corresponde a secuencias repetidas.

Por otra parte, del ADN relacionado con genes, el 95 % corresponde a ADN no codificante (pseudogenes, fragmentos de genes, intrones o secuencias no traducidas (del inglés untranslated región, UTR). La siguiente figura muestra la composición del genoma humano, donde los genes y las secuencias relacionadas con genes corresponden a 1200 Mb del total del genoma (aprox. 3235 Mb).

Composición del genoma humano.

Nota. Adaptado de (Brown, 2017b)



1. Estructura del genoma humano

Tabla 2
Tamaño de distintos genomas eucariotas.

Actualización 30/12/2023

Especie	Tamaño genoma	Num. Chr	Número de genes
Hongos			
<i>Saccharomyces cerevisiae</i> (taxonomy ID 4932)	12.16 Mb	16	6 463
<i>Aspergillus nidulans</i>	29.83 Mb	8*	9 586
Protozoos			
<i>Tetrahymena pyriformis</i> ***	116 Mb	**	26 866
Invertebrados			
<i>Caenorhabditis elegans</i>	100.3 Mb	6 + Mt	46 926
<i>Drosophila melanogaster</i>	143.7 Mb	7 + Mt	17 894
<i>Bombyx mori</i> (gusano de seda)	460.3 Mb	28 + Mt	17 047
<i>Strongylocentrotus purpuratus</i> (erizo de mar morado)	921.8 Mb	Un	33 503
<i>Locusta migratoria</i> (langosta migratoria)	6.3 Gb	12 + Mt	**
Vertebrados			
<i>Takifugu rubripes</i> (pez puffer / torafugu)	384.1 Mb	22 + Mt	27 412
<i>Homo sapiens</i>	3.1 Gb	23 + Mt	59 265
<i>Mus musculus</i>	2.7 Gb	20 + Mt	50 561
Plantas			
<i>Arabidopsis thaliana</i>	119.1 Mb	5 + Mt + Pltd	38 312
<i>Oryza sativa</i> (arroz)	373.8 Mb	12 + Mt + Pltd + B1	35 223
<i>Zea mays</i> (maíz)	2.2 Gb	10 + Mt + Pltd	49 897
<i>Pisum sativum</i> (guisante)	3.8 Gb	7 + Mt	65 672
<i>Triticum aestivum</i> (trigo)	14.6 Gb	21 + Mt	155 463
<i>Fritillaria assyriaca</i> (lila)	120 Gb	**	**

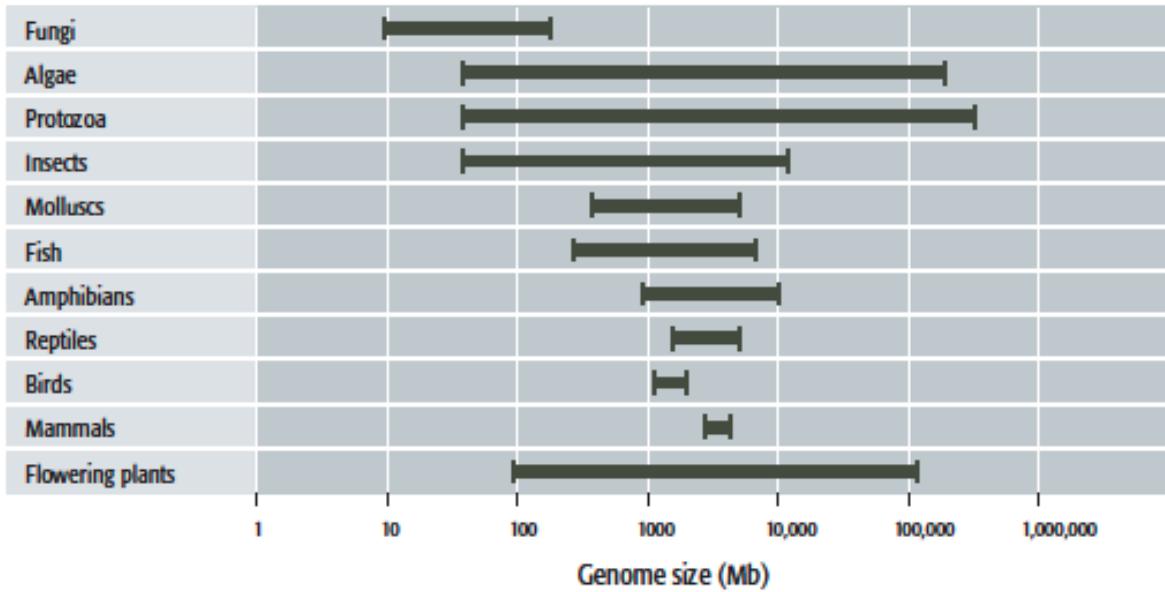
Nota. Información adaptada de Genomes 4, por T. A. Brown, 2017, y Home - Genome – NCBI, 2021 (<https://www.ncbi.nlm.nih.gov/genome/>).

* Aún no hay ningún genoma depositado en <https://www.ncbi.nlm.nih.gov/genome/> que se encuentre cerrado completamente.

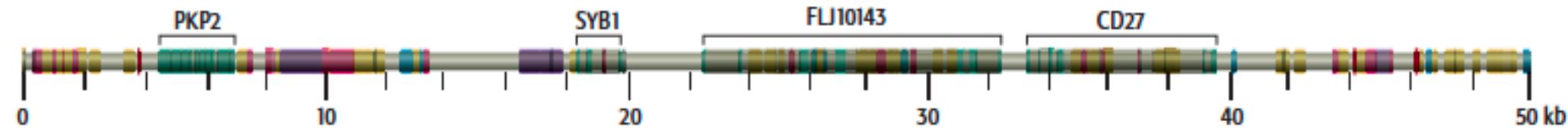
** No está contenido ningún genoma ni completo ni parcial en <https://www.ncbi.nlm.nih.gov/genome/>.

Un : no se encuentra ensamblado su genoma.

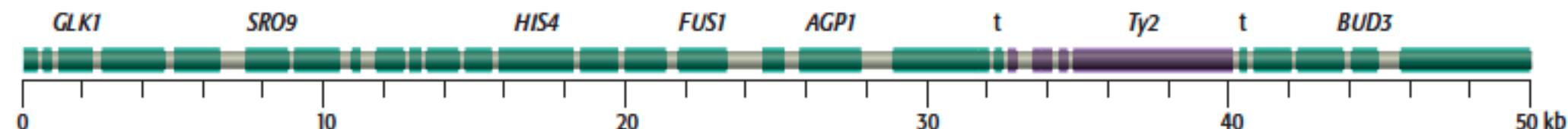
***<http://ciliate.ihb.ac.cn/tcgdb/database/home/>



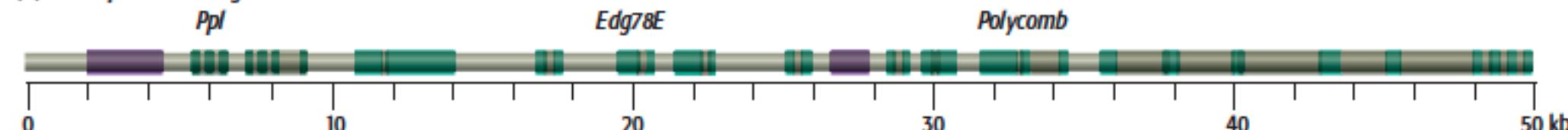
(A) Human



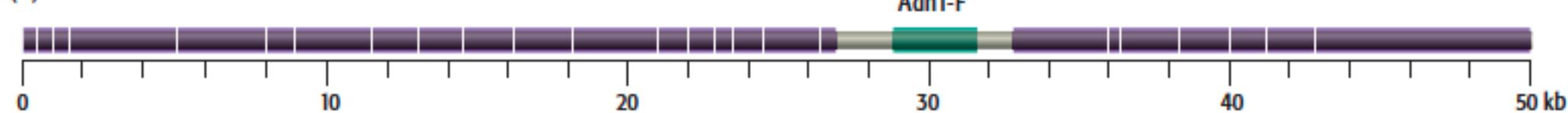
(B) *Saccharomyces cerevisiae*



(C) *Drosophila melanogaster*



(D) Maize



KEY

exon intron

LINE

SINE

LTR element

DNA transposon

Other genome-wide repeat

Microsatellite

tRNA gene

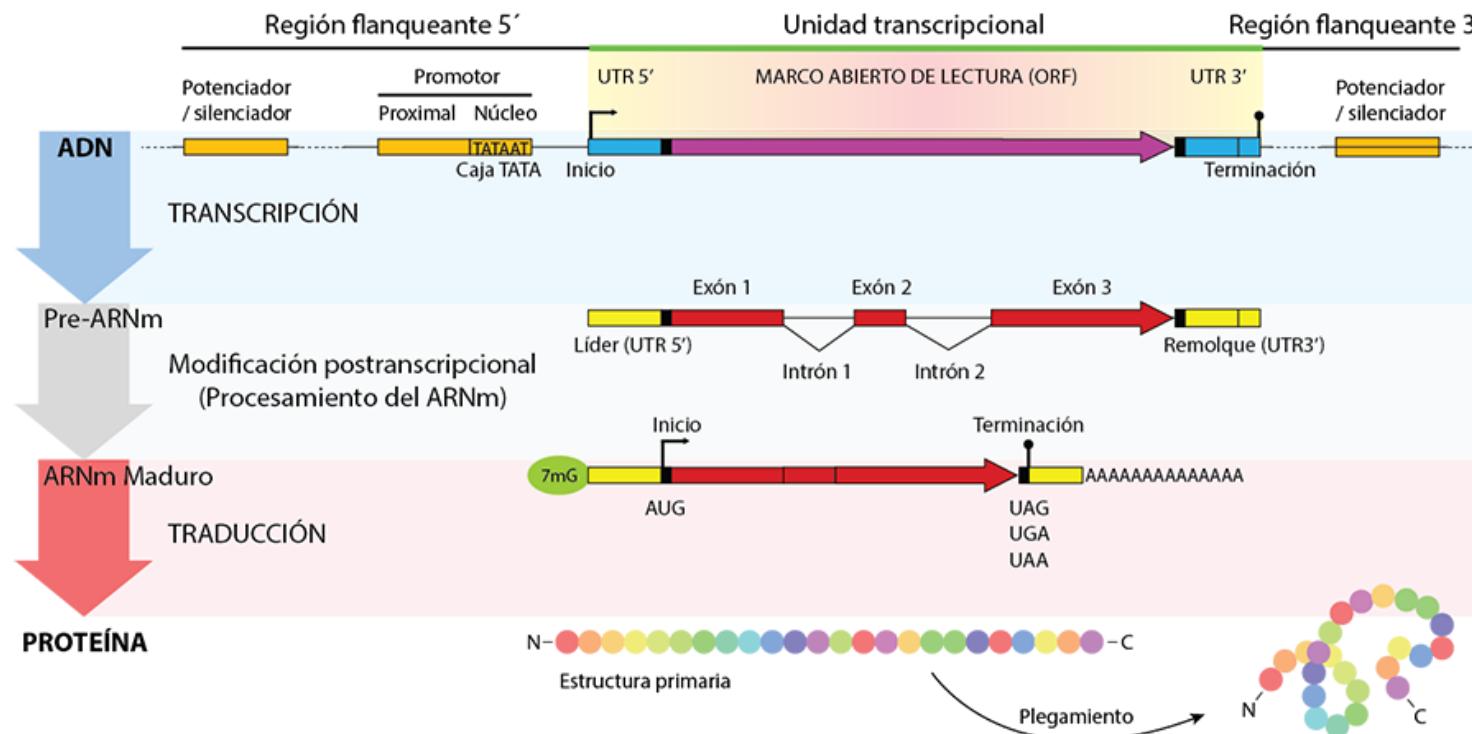
t

Table 7.3 Compactness of the yeast, fruit fly, and human genomes

Feature	Yeast	Fruit fly	Human
Gene density (average number per Mb)	496	76	11
Introns per gene (average)	0.04	3	9
Amount of the genome that is taken up by genome-wide repeats	3.4%	12%	44%

Secuencias en el genoma humano:

- Genes que codifican proteínas
- Intrones
- Genes con información ARN no-codificante



Todos conocemos que el gen es la unidad básica de herencia que porta la información genética necesaria para sintetizar una proteína (en el caso de genes codificantes), o de un ARN no codificante (genes de ARN).

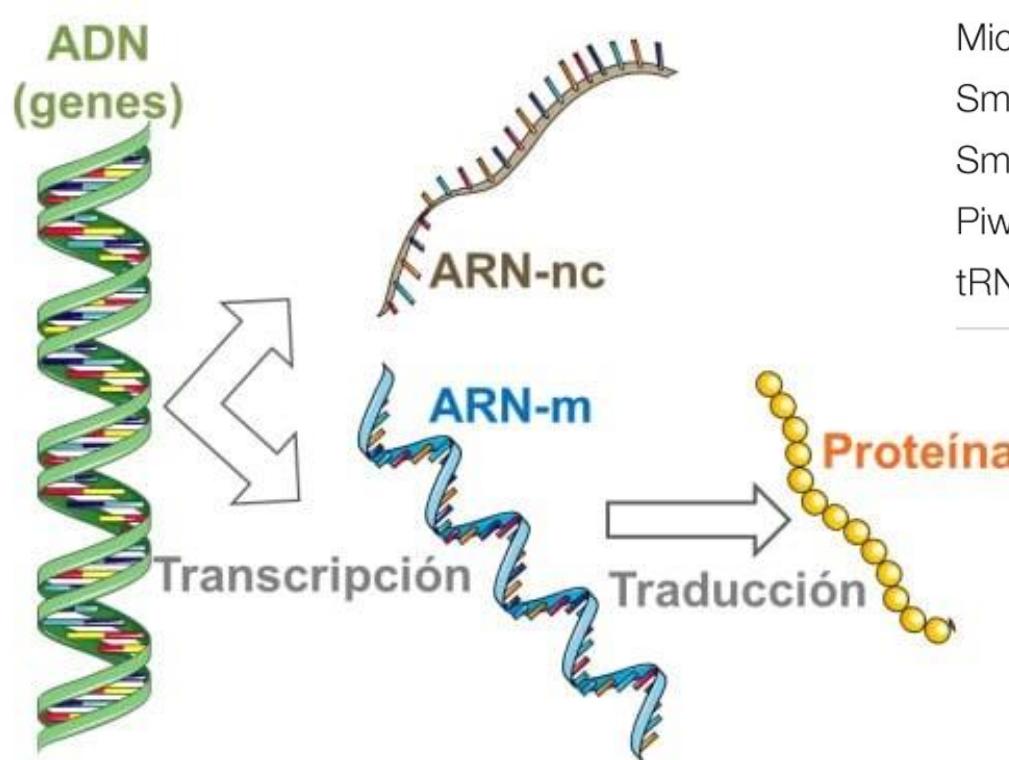
Los genes están formados por una secuencia promotora que regula su expresión y una secuencia que se transcribe. Esta secuencia que se transcribe a su vez se compone por secuencias UTR (regiones flanqueantes no traducidas, pero necesarias para la traducción y la estabilidad del ARN mensajero), exones (codificantes) e intrones (secuencias de ADN no traducidas situadas entre dos exones, y que serán eliminadas en el procesamiento del ARN mensajero).

Actualmente, y en base a los resultados arrojados por el proyecto **ENCODE** (Encyclopedia Of DNA Elements), algunos autores consideran que debemos revisitar el concepto clásico de gen, formado por UTR, exones e intrones.

Algunos estudios han mostrado que las secuencias UTR 5' se encuentran muy distantes de la secuencia traducida, abarcando secuencias muy largas y dificultando delimitar lo que es el gen.

Además, un mismo transcripto puede dar lugar a ARN maduros totalmente diferentes, debido al proceso de *splicing* alternativo. De este modo, un mismo transcripto primario puede dar lugar a proteínas de secuencia y funcionalidad completamente distintas. En este proceso ya no se consideran ni los UTR ni los intrones. De acuerdo a esta observación, en este caso deberían considerarse genes diferentes, según algunos autores (Gerstein et al., 2007).

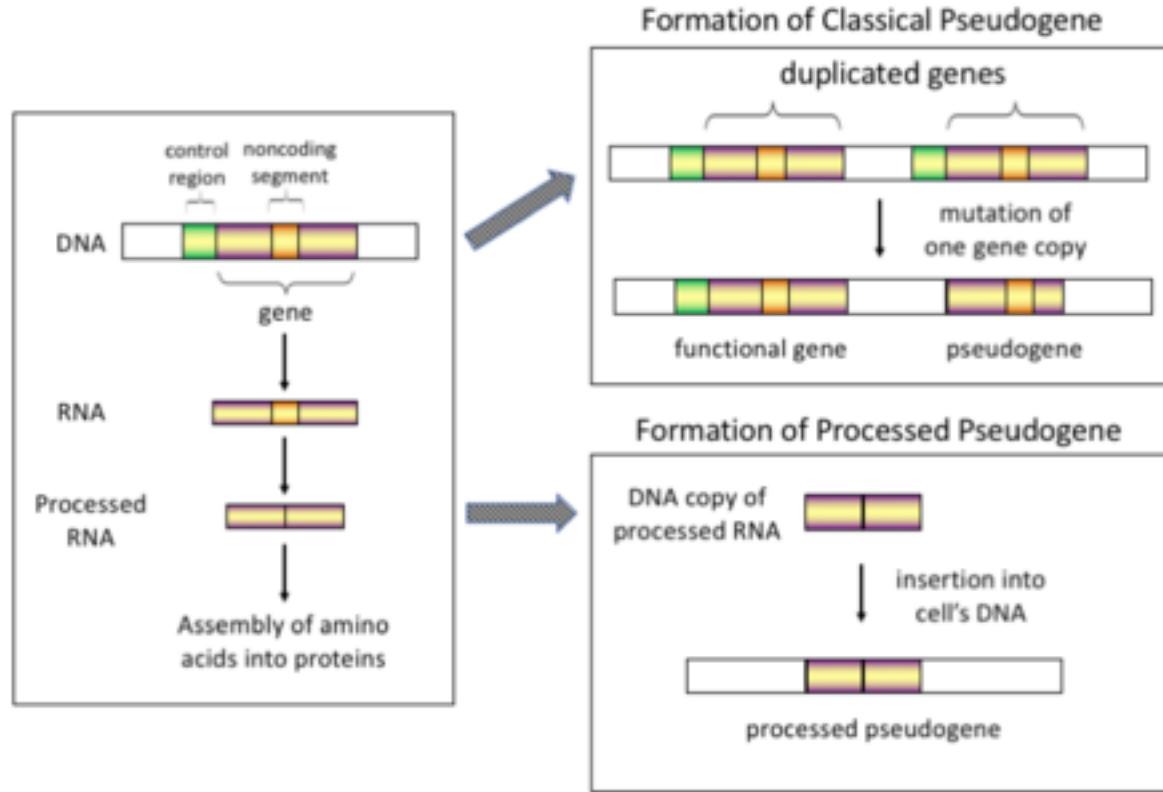
Finalmente, encontramos los **genes que codifican ARN no-codificantes**. Estos genes se expresan para dar lugar a un ARN funcional, como son los microARN y los piwiARN. Este tipo de ARN se involucran en la regulación de la expresión génica.



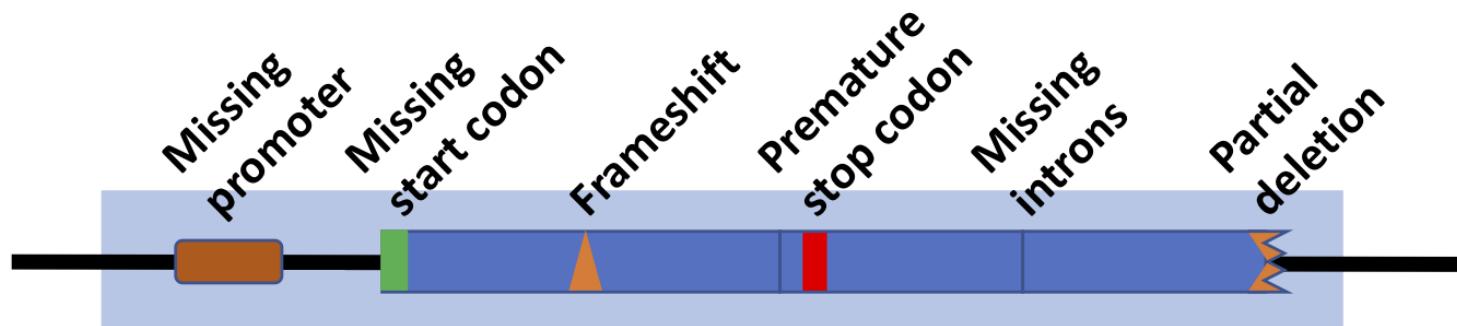
Type of small non-coding RNA	Size (nts)	Function
MicroRNA (miRNA)	~22	Ago – RNAi
Small nuclear RNA (snRNA)	~150	Spliceosome components
Small nucleolar RNA (snoRNA)	60–140	RNA modification
Piwi-interacting RNA (piRNA)	26–31	PIWI – RNAi
tRNA derived small RNA (tsRNA)	15–50	Diverse

Por otro lado, los **pseudogenes** son versiones completas o parciales de genes que han ido acumulando mutaciones a lo largo de la evolución y que generalmente no se transcriben. De esta forma, tenemos el gen original activo y una copia inactiva. Aunque tradicionalmente se han pensado que estos pseudogenes son “genoma basura”, recientemente se ha visto que son capaces de regular la expresión de los genes de los que proceden mediante fenómenos de regulación antisentido, competencia de ARN o ARN de silenciamiento. Se estima que el genoma humano contiene unos 19 000 pseudogenes, clasificados entre no-procesados (aprox. 30 %) y procesados (aprox. 70 %). En este grupo encontramos:

- **Pseudogenes no-procesados** son copias de genes originadas por duplicación que no se transcriben por carecer de una secuencia promotora y tras haber acumulado múltiples mutaciones, algunas de las cuales no tienen sentido originando codones de parada prematuros. Se caracterizan por poseer exones e intrones.
- **Pseudogenes procesados**: son copias de ARN mensajero que se han retrotranscrito (ADN complementario, ADNc) que se introducen en otra posición del genoma. Este pseudogen no tiene promotor, pero pueden activarse por acción del promotor de un gen cercano (retrogenes). Estos pseudogenes al ser copias insertadas en el genoma carecen de intrones y de secuencia promotora.



Common defects of pseudogenes:



ADN intergénico. Las regiones intergénicas o también llamadas extragénicas comprenden la mayor parte de la secuencia del genoma humano y su función es generalmente desconocida.

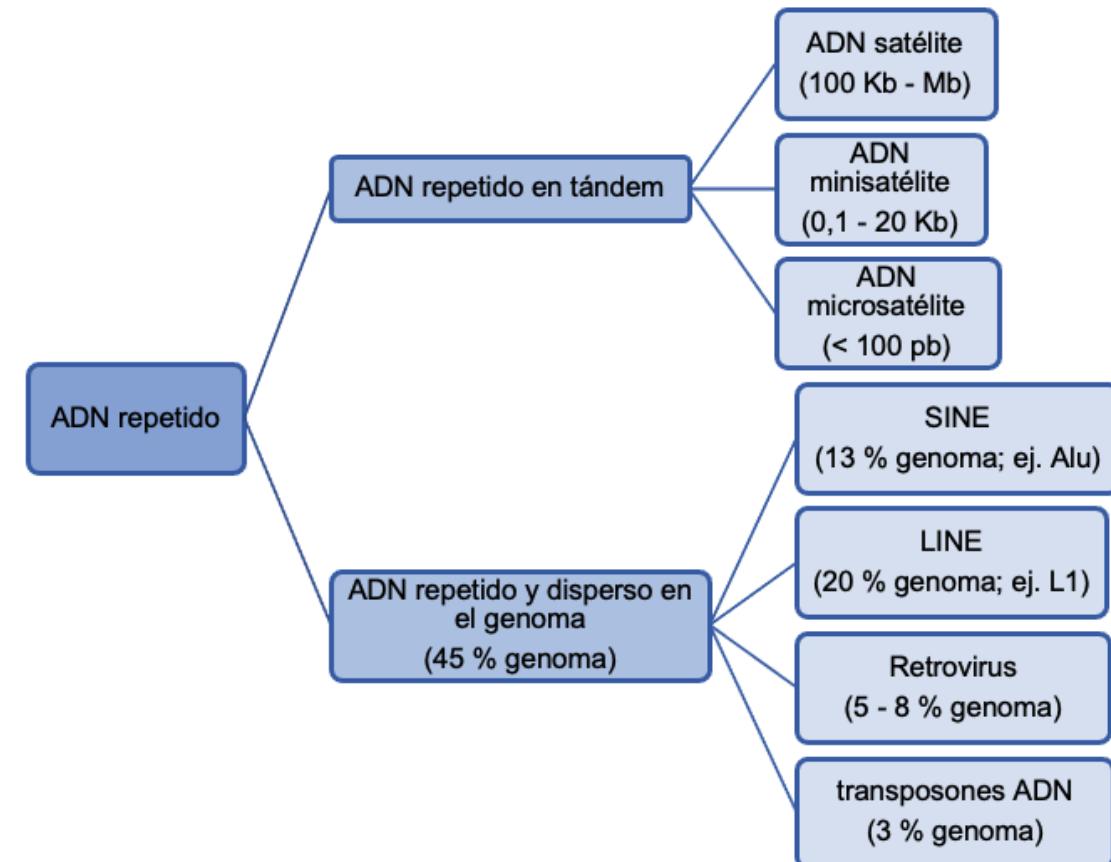
La mayor parte de estas regiones están compuestas por **elementos repetitivos**, como veremos a continuación, y el resto de la secuencia no responde a un patrón definido o clasificable.

Se piensa que gran parte de este ADN intergénico puede ser un artefacto evolutivo sin función determinada en el genoma actual, sin funcionalidad, por lo que ha sido denominado “ADN basura”. Esta denominación incluye el ADN intergénico, secuencias intrónicas y pseudogenes.

La denominación de estas zonas como “ADN no codificante” se debe a que se ha observado que algunas de estas regiones poseen papel regulador y otras están altamente conservadas a lo largo de la evolución, y podrían poseer otras funciones esenciales aún desconocidas o poco conocidas.

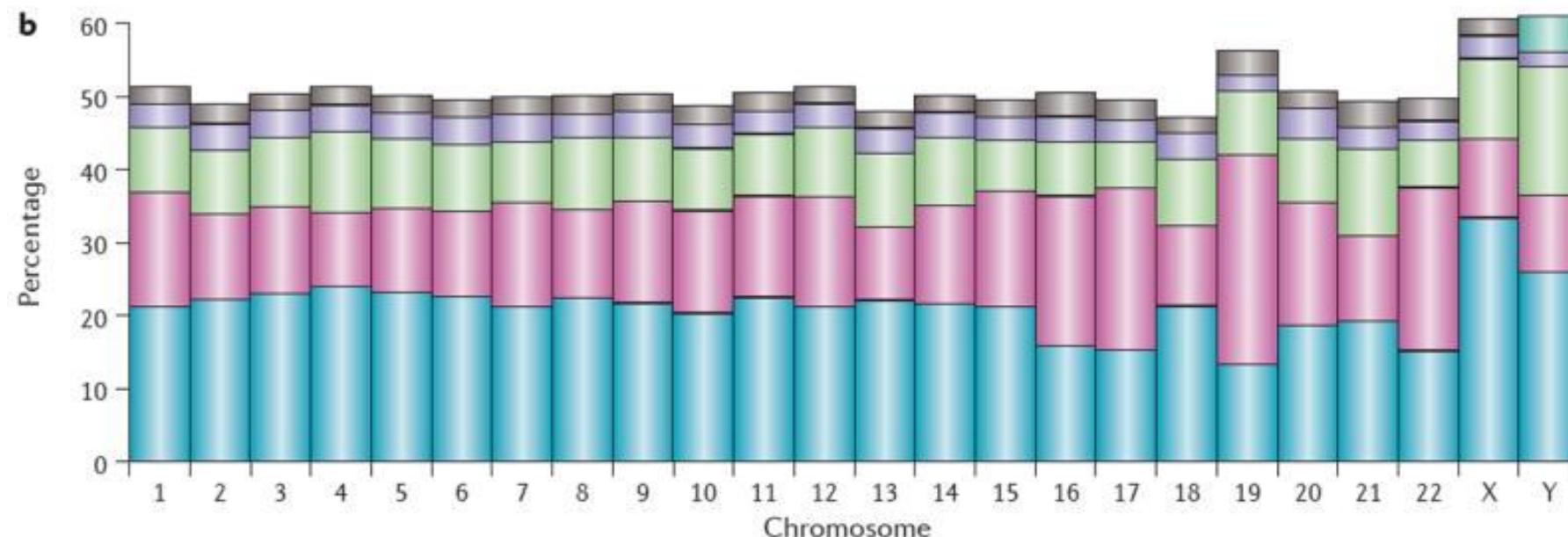
Las secuencias de **ADN repetido** forman parte de este ADN intergénico, constituyendo más del 50% del genoma humano. En este grupo de secuencias están aquellas derivadas de procesos de transposición (elementos transponibles), que suponen aproximadamente un 45% del total del genoma. El otro 5% corresponde a secuencias repetidas, sencillas y cortas, que se encuentran en miles de copias. Este grupo de secuencias no tienen una función muy conocida, pero son responsables de la variabilidad que da lugar a la evolución.

Tipo de secuencias de DNA repetidas en el genoma humano.
Nota. Figura de elaboración propia.



a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

b

1.1.4. Secuencias repetidas y secuenciación masiva

Tal y como hemos visto, las secuencias de DNA repetido son abundantes a lo largo del genoma, tanto genomas eucariotas como el humano, como también veremos en temas posteriores, en genomas bacterianos. Este tipo de repeticiones son un reto técnico para el alineamiento y ensamblaje de los genomas a nivel bioinformático.

Desde un punto de vista computacional, estas repeticiones crean ambigüedades, que pueden producir errores en la interpretación de resultados. Debemos tener en cuenta que este problema se ve agravado en la secuenciación con lecturas cortas (tipo Illumina), mientras que la secuenciación de lecturas largas (Oxford Nanopore, PacBio) pueden paliar este efecto si la repetición es de menor tamaño que la zona secuenciada.

Tal y como se expone en el artículo de (Treangen & Salzberg, 2011), las repeticiones que son suficientemente divergentes no presentan problemas. Sin embargo, consideran que repeticiones de al menos 100 pb de longitud que ocurran dos o más veces en el genoma, que muestren más de un 97% de identidad con al menos otra copia de sí misma, son las que muestran un desafío computacional. Éstas nos suponen diversos problemas:

Mapeo de multilecturas: al alinear las lecturas frente al genoma de referencia, el mayor problema que encontramos son aquellas lecturas que pueden “encajar” en varias localizaciones. Aunque estas lecturas no son un problema para el alineador, sí que los son para análisis posteriores, como puede ser la identificación de mutaciones (SNP). Este problema puede paliarse utilizando lecturas más largas. A longitud mayor, mejor anclaje de la secuencia leída frente al genoma de referencia. Dado que la mayor parte de las secuencias repetidas no son exactamente iguales, muchas de ellas tendrán un único macheo (*best match*).

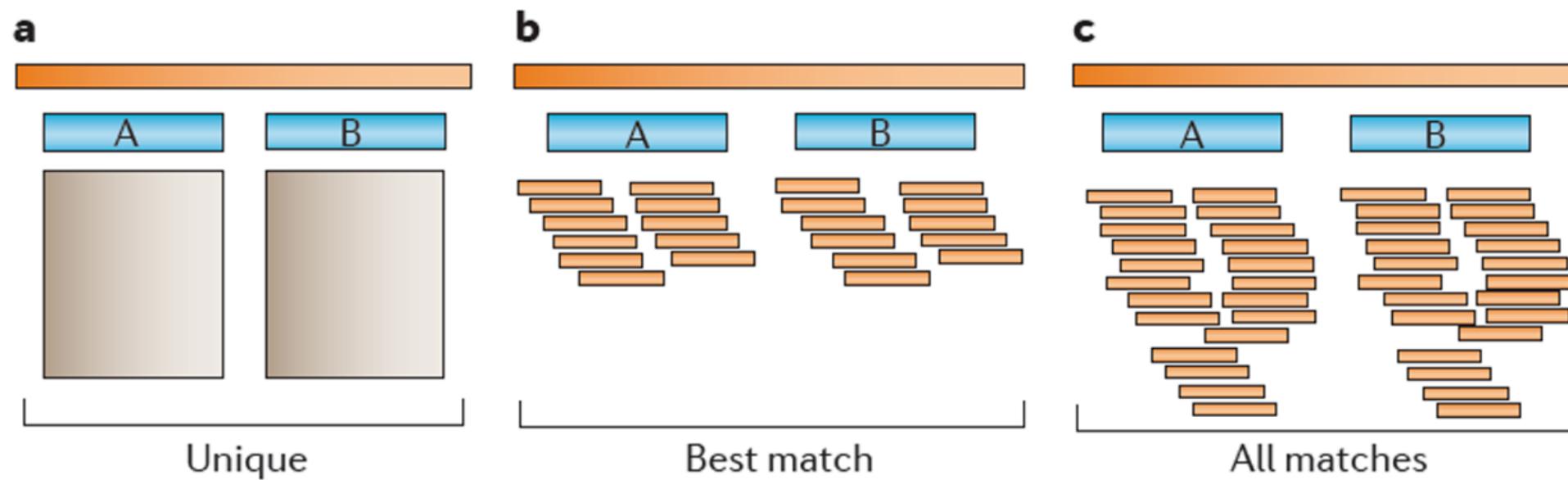
Existen varias estrategias para aminorar este problema:

- Ignorarlo, de manera que eliminemos todas las lecturas multiposición. Esta aproximación es aparentemente la más sencilla, pero limita el análisis no solo de genes de repetición, sino de familias multigen, con alta similitud. Esto puede conllevar que nos perdamos variantes de importancia biológica.
- **Aproximación del mejor alineamiento (*best match*)**. El alineamiento con menor número de desacuerdos (*mismatches*) es el reportado. Si hay múltiples alineamientos igualmente válidos, el alineador elegirá uno al azar, o los informará todos. Esta es quizás la estrategia más aceptada comúnmente y que produce una estimación de cobertura más real.
- Tomar todos los alineamientos hasta un número máximo (X) o bien, ignorar aquellas lecturas alineadas más de X veces. Esta aproximación es la más flexible y hace posible encontrar errores de emplazamiento erróneo de las lecturas.

Estrategias para el manejo de lecturas multi-mapeo. Los rectángulos anaranjados en la parte superior de la figura representan una región del genoma. Los dos rectángulos azules, etiquetados como A y B representan dos copias idénticas de un gen. Los pequeños rectángulos anaranjados en la parte inferior representan las lecturas de secuenciación.

- (a) Representa la estrategia en la que sólo se reportan las lecturas que mapean únicamente. Dado que A y B son idénticos, ninguna lectura será reportada.
- (b) En la aproximación del mejor alineamiento se reporta solo el mejor alineamiento para cada lectura, de acuerdo con el algoritmo que clasifica estas lecturas.
- (c) Estrategia de reporte de todos los alineamientos para cada lectura multi-mapeo, incluyendo aquellas que sean de peor calidad de mapeo.

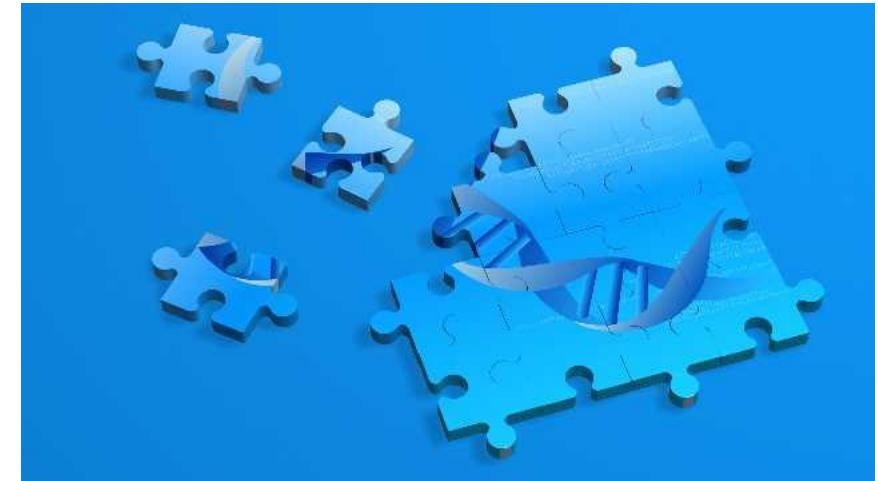
Nota. Figura tomada de (Treangen & Salzberg, 2011).



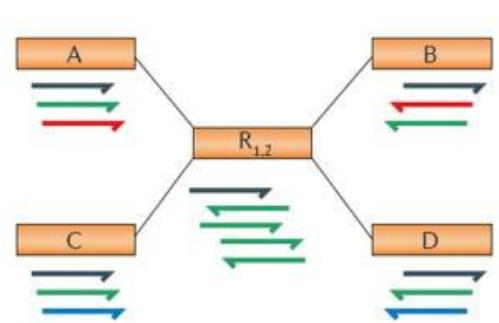
Ensamblaje de genomas *de novo*: En este proceso lo que queremos es reconstruir el genoma sin tener una referencia (montar el puzzle sin tener la fotografía final de cómo debe ser).

Las lecturas cortas son en sí mismas un problema a la hora de reconstruir un genoma desde cero. Tenemos millones de fragmentos (alta cobertura) pero de un tamaño muy limitado. En este caso, esta alta cobertura puede solventar algunos problemas de las lecturas cortas, sin embargo, las zonas repetidas no se solucionan de esta forma. **Nos encontramos dos problemas principales:**

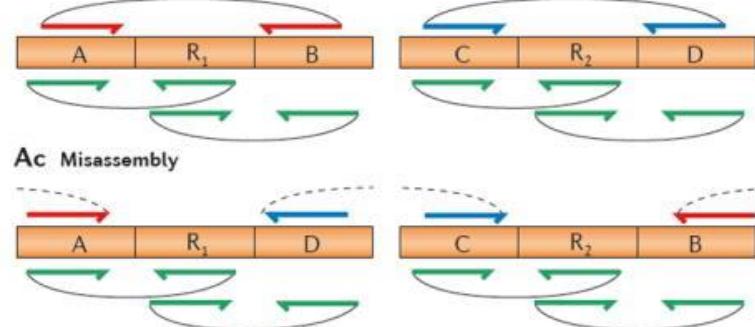
- Cuando la repetición es mayor que el tamaño de la lectura se crean huecos (*gaps*) en el ensamblaje. Esto desemboca en genomas altamente fragmentados.
- Las repeticiones pueden ser colapsadas en una única secuencia, produciendo reorganizaciones erróneas y secuencias quimera.



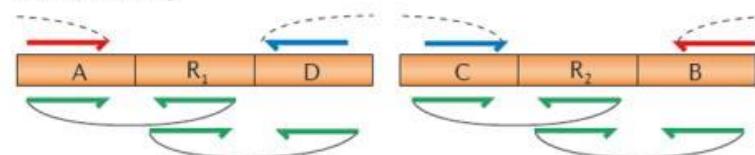
Aa Assembly graph



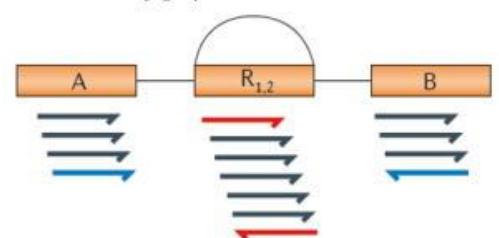
Ab Correct assembly



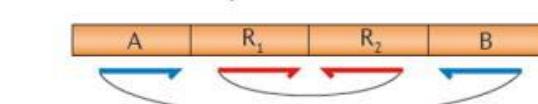
Ac Misassembly



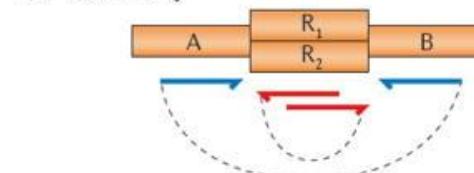
Ba Assembly graph



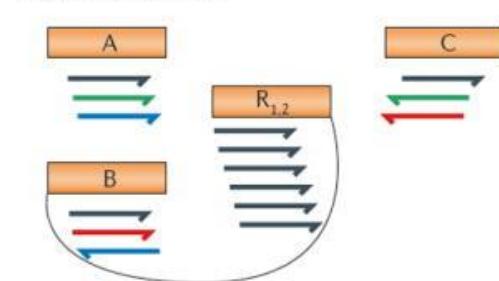
Bb Correct assembly



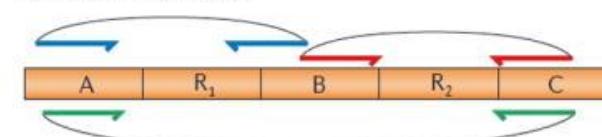
Bc Misassembly



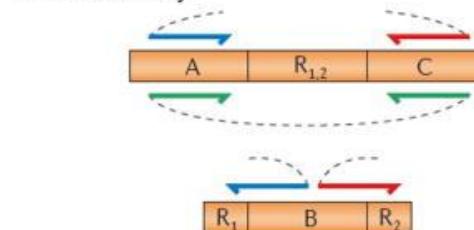
Ca Assembly graph



Cb Correct assembly



Cc Misassembly



Los ensambladores, tal y como veremos en un tema posterior, pueden reconstruir el genoma bien por solapamiento de las lecturas completas o por fragmentos de estas. Desde el punto de vista técnico, las secuencias repetidas son un “bucle” o “burbuja”, una bifurcación en el camino de la resolución del genoma.

Si este bucle se resuelve de manera errónea se generan uniones falsas ([secuencias quimeras](#)). Si el ensamblador es más conservador, romperá el ensamblaje en ese punto de bifurcación, dejando un fragmento corto pero exacto.

Cuando manejamos lecturas pareadas, la información de distancia media entre ellas es utilizada por algunos ensambladores para colocarlas y determinar la mejor posición, e incluso, para llenar con nucleótidos indefinidos (Ns) esos huecos indeterminados.

Alineamiento y ensamblaje de secuencias de ARN: aunque no es tema de esta asignatura, el análisis de expresión génica y ensamblaje de transcritos también se ve influenciado por las secuencias repetidas.

El análisis de expresión génica se basa en el mapeo de las lecturas frente a un genoma de referencia, con el posterior recuento de éstas en cada unidad genómica; asimismo el **ensamblaje de transcritos es una situación similar al ensamblaje *de novo*.**

Por tanto, los problemas que hemos visto en los dos puntos anteriores son similares a los que encontraremos en el análisis de secuencias repetidas de un transcriptoma.

1.2



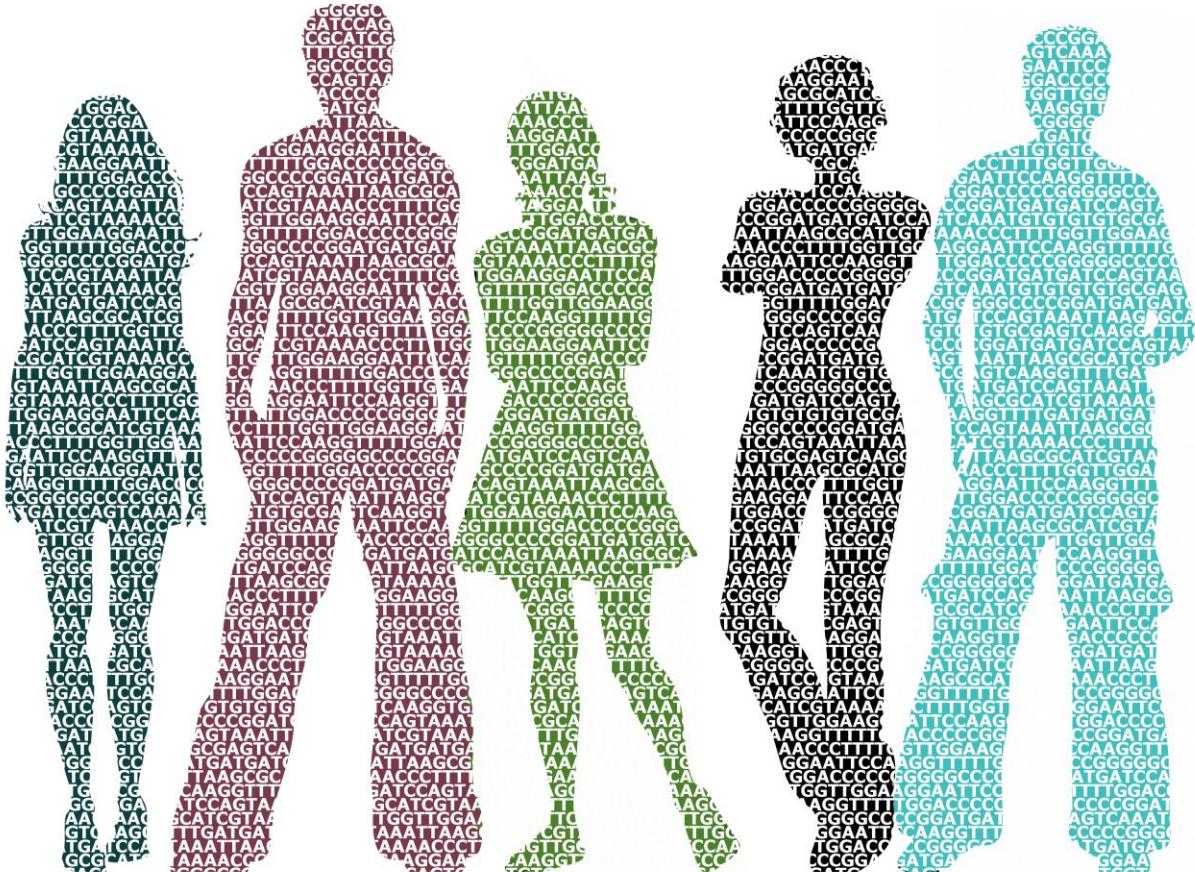
Patrones de transmisión de las
enfermedades genéticas

1.2.1. Variabilidad del genoma

La variabilidad dentro de un genoma puede deberse a una **sustitución, delección o inserción**. De esta manera, encontramos polimorfismos o alelos genéticos.

Estas variaciones pueden darse tanto en regiones codificantes como no codificantes, y obviamente, debido a la degeneración del código genético, no toda alteración supone una alteración en la proteína o su nivel de expresión. Muchos de estos cambios son silenciosos y no tienen repercusión (Trent, 2012).

Entre los cambios que pueden darse se encuentran:



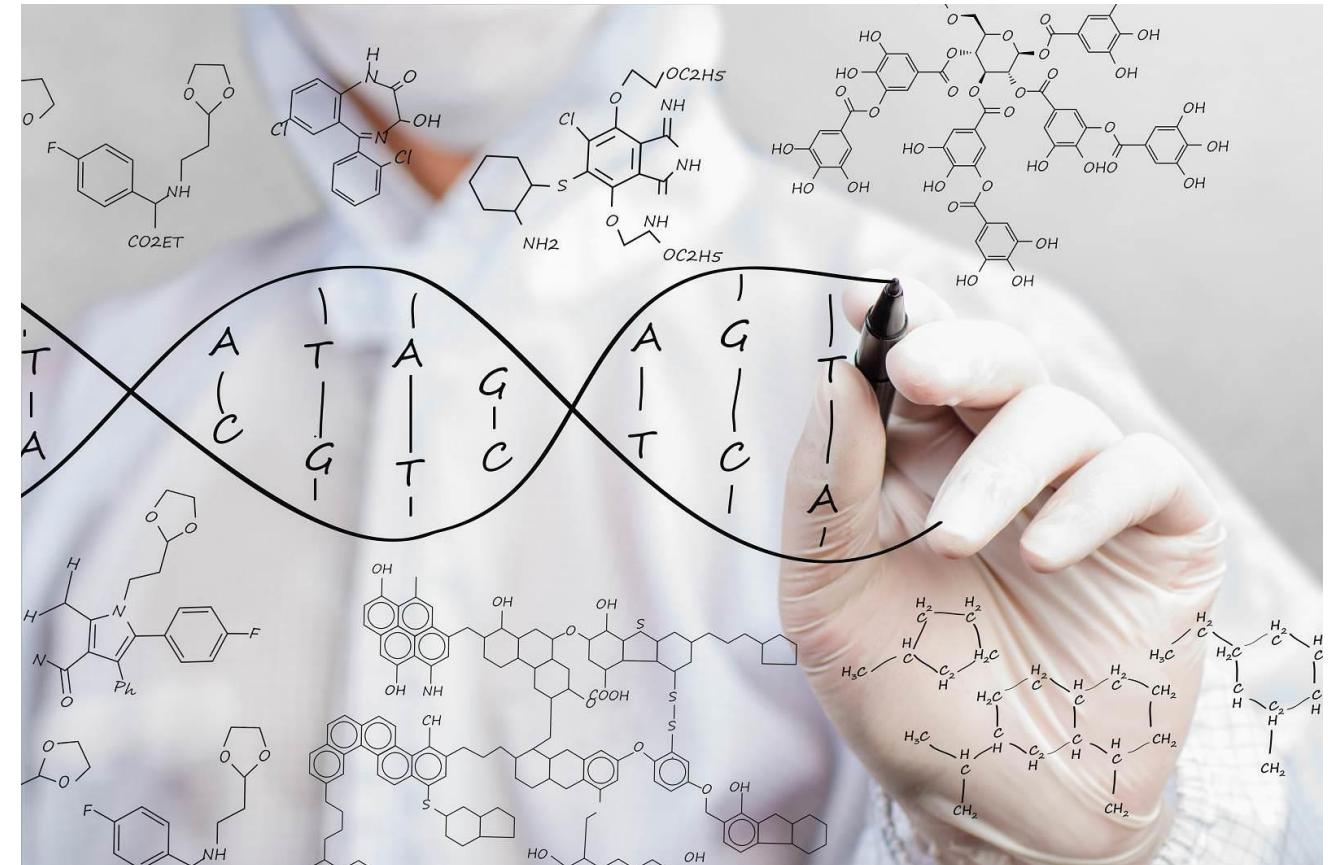
- **Single nucleotide polymorphisms (SNP)**: son la mayor fuente de variabilidad en los genomas de seres humanos, suponiendo la variación de un solo nucleótido. Se estima que su frecuencia es de un SNP por cada 500-1000 pb. Son utilizados como marcadores genéticos en estudios a gran escala mediante el empleo de chips de AND (*microarrays*). Actualmente, **son la base de nuestros estudios de secuenciación masiva**. La identificación de estas variantes de nucleótido único por este método son las SNV (*single nucleotide variants*). Se observan regiones del genoma con mayor grado de conservación, teóricamente por ser regiones con una función altamente esencial para el organismo. **Las zonas que codifican para proteínas presentan menor número de SNV que las zonas intergénicas.**
- **Variación estructural (duplicaciones, inversiones, inserciones o variantes en número de copias)**. Todos estos eventos afectan a una gran proporción del genoma. Este término abarca un gran número de eventos genéticos que **implican segmentos de más de 1 kb de longitud**.

1.2.2. Tipos de enfermedades genéticas

Se considera enfermedad genética aquella en la que existe un componente genético o hereditario implicado.

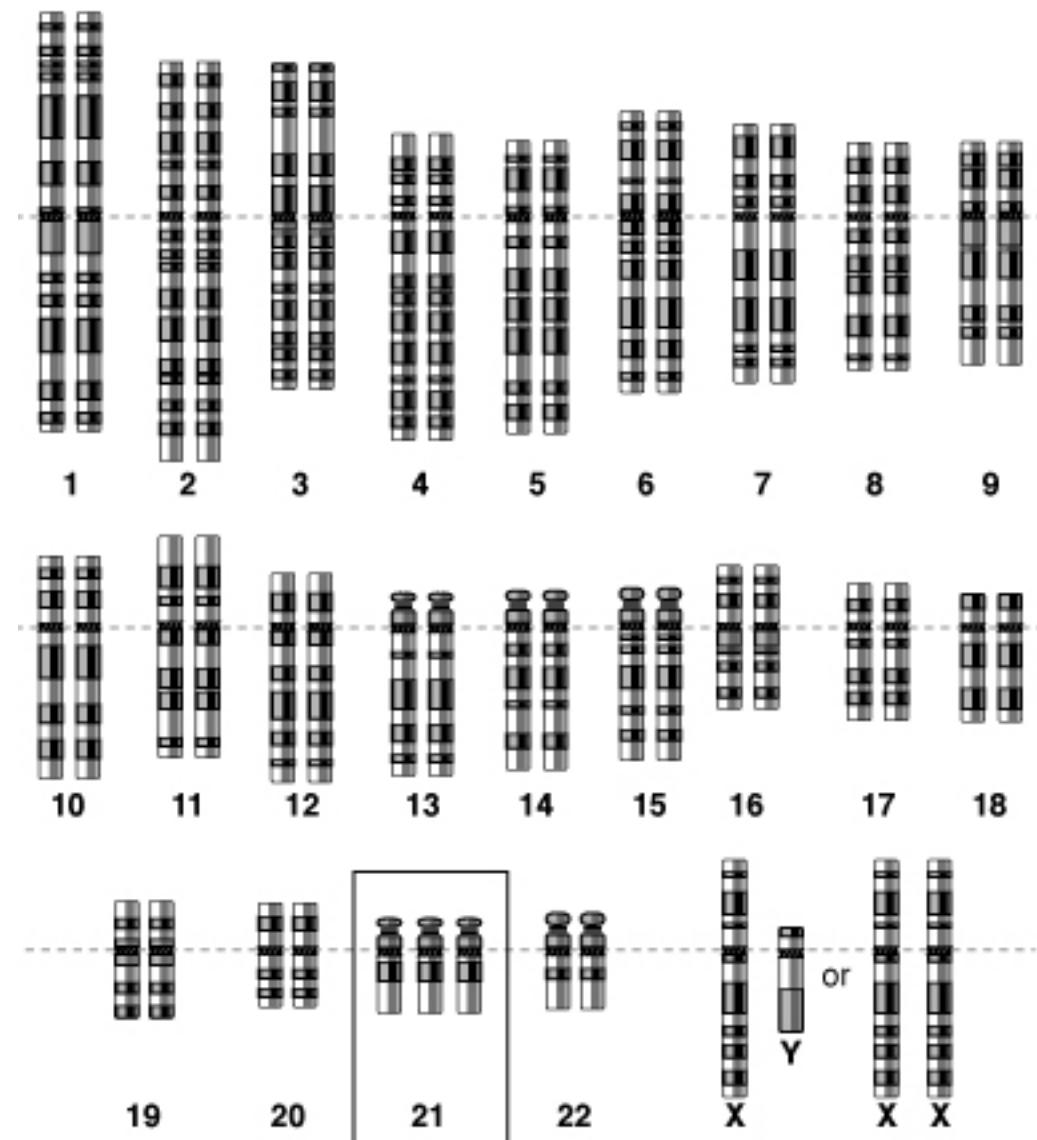
Se clasifican en:

- cromosómicas,
- monogénicas y,
- multifactoriales o de herencia compleja



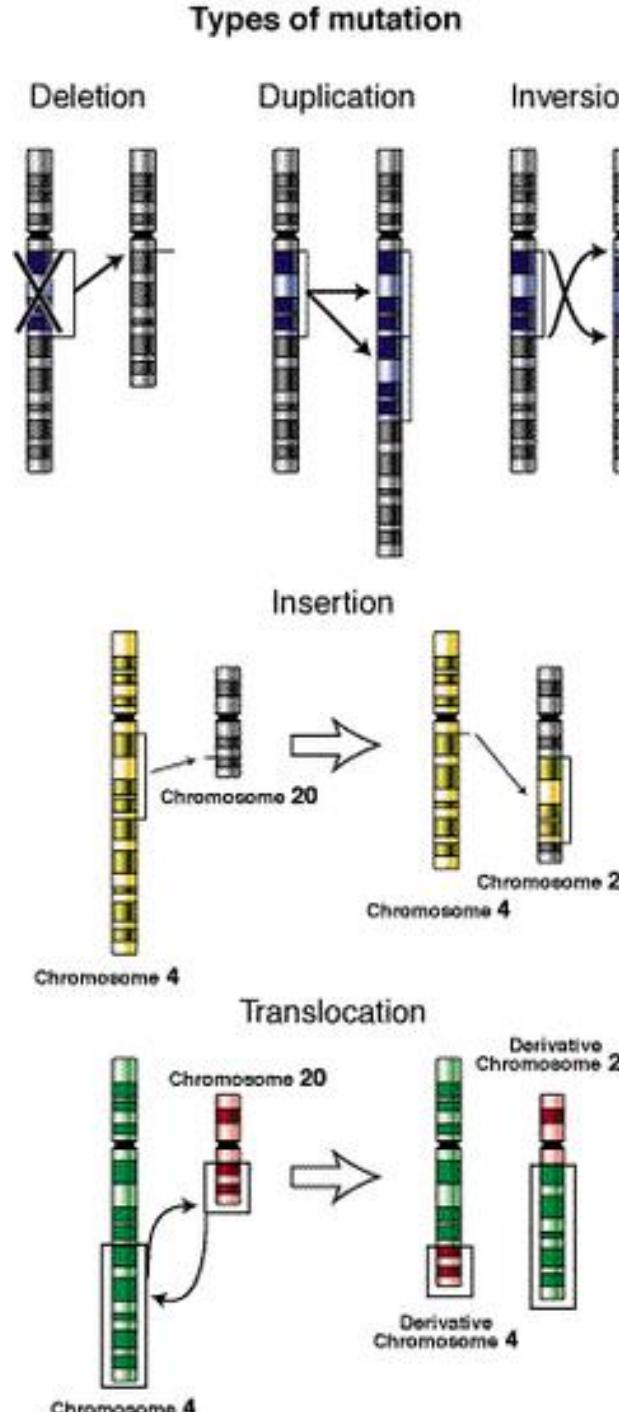
Cromosómicas: Afectan 7/1000 nacimientos. Son debidas a alteraciones en el número o estructura de los cromosomas. Son responsables de aproximadamente el 50% de los abortos espontáneos del primer trimestre. Se pueden clasificar en dos grandes grupos:

- **Numéricas:** alteración en el número normal de cromosomas. Habitualmente encontramos 23 pares. Puede afectar a un solo par de cromosomas (aneuploidía), habiendo un solo cromosoma (monosomía) o más de dos (trisomía, tetrasomía). Ejemplo es el síndrome de Down (trisomía del cromosoma 21), síndrome de Turner (monosomía del cromosoma X), Edwards (trisomía 18), Patau (trisomía 13), Klinefelter (XXY) o XYY. Si la alteración afecta a todos los cromosomas se habla de euploidías, de manera que el individuo puede tener una sola dotación cromosómica (haploidía, 23 cromosomas totales) o más de dos dotaciones completas (triploidía -69-, tetraploidía -92-). Éstas suelen ser letales a nivel embrionario, por lo que habitualmente no llega a término el embarazo.



- **Estructurales:** alteraciones en la estructura de los cromosomas, como pueden ser grandes inserciones o delecciones, reorganizaciones...

- **Delecciones:** eliminación de una porción del cromosoma. Ejemplos: síndrome Wolf-Hirschhorn (deleción parcial del brazo corto del cromosoma 4), síndrome de Jacobsen (deleción 11q terminal).
- **Duplicaciones:** de una región considerable del cromosoma. Enfermedad Charcot-Marie-Tooth con la duplicación del gen PMP22 en el cromosoma 17.
- **Traslocaciones:** transferencia parcial de un cromosoma a otro.
- **Inversiones:** una parte del genoma se rompe y se reorienta en dirección opuesta.



Enfermedades monogénicas o mendelianas: Causadas por mutaciones en genes individuales. Tienen patrones específicos de transmisión y su prevalencia es de 2/100 habitantes. Ejemplos son la fibrosis quística o la enfermedad de Huntington. Son el único tipo de enfermedades que presentan un patrón de herencia claro, al estar causadas por mutaciones en un único gen. Esta mutación tiene un fuerte impacto en el fenotipo, es decir, el riesgo de desarrollar la enfermedad será alto, y además será igual para todas las familias que presentan la mutación. El patrón de herencia que puede mostrar una enfermedad de tipo monogénica o mendeliana se clasifica en:

- Autosómico dominante
- Autosómico recesivo
- Dominante ligado al X
- Recesivo ligado al X
- Ligado a Y
- mitocondrial

Autosómico dominante:

se manifiestan en individuos heterocigotos.

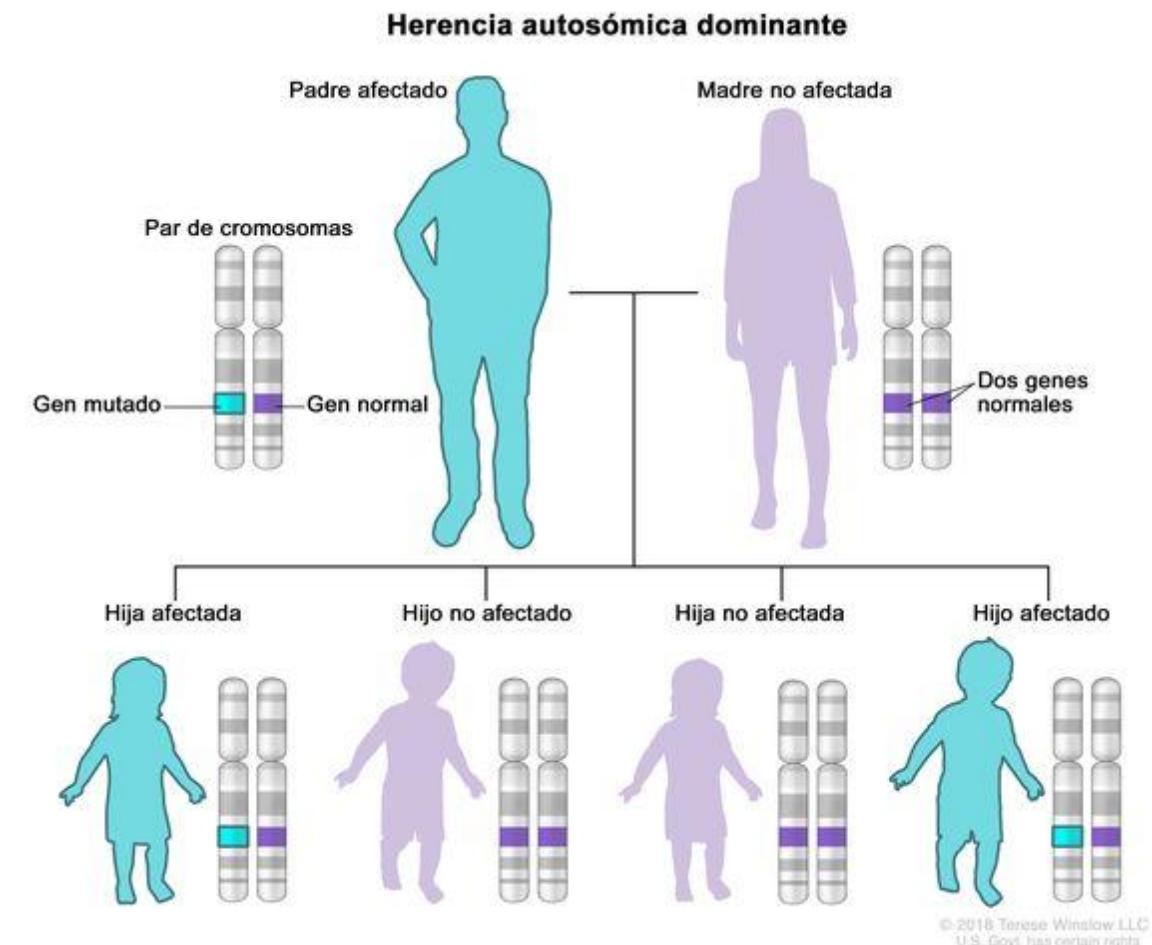
Es suficiente que mute una de las dos copias del cromosoma de un gen para que se manifieste la enfermedad.

Los individuos enfermos suelen tener a uno de sus progenitores enfermos, y asimismo, la probabilidad de tener descendencia afectada es de un 50%.

Frecuentemente son mutaciones de ganancia de función (el alelo mutado tiene una nueva función que provoca el desarrollo de la enfermedad) o bien, por pérdida de función.

Son enfermedades de baja penetrancia (solo una parte de los individuos que portan la mutación desarrollan la enfermedad).

Ejemplos: Huntington, Marfan, algunos tipos de cáncer colorrectal hereditario.



© 2016 Terese Winslow LLC
U.S. Govt. has certain rights

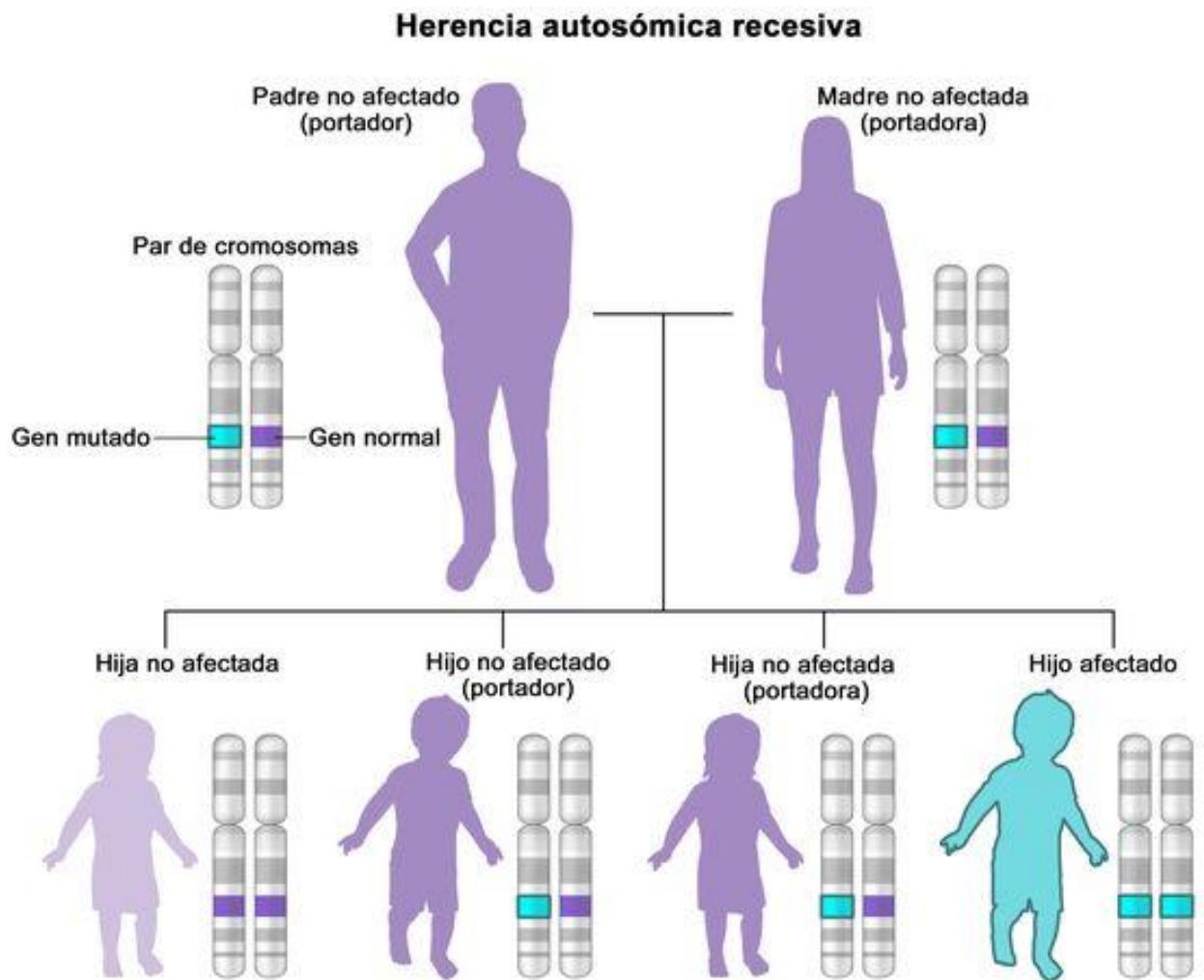
Autosómico recesivo:

la enfermedad solo se manifiesta en individuos homocigóticos recesivos (ambas copias del gen están mutadas).

Suelen ser mutaciones que causan pérdida de función, de modo que la causa de la enfermedad es la ausencia de acción de un gen.

Habitualmente el individuo enfermo tiene ambos progenitores sanos, pero son portadores de la mutación. El 25% de la descendencia es afectada.

Ejemplos: Fibrosis quística, anemia falciforme, Tay-Sachs, atrofia muscular espinal.



Dominante ligado al X:

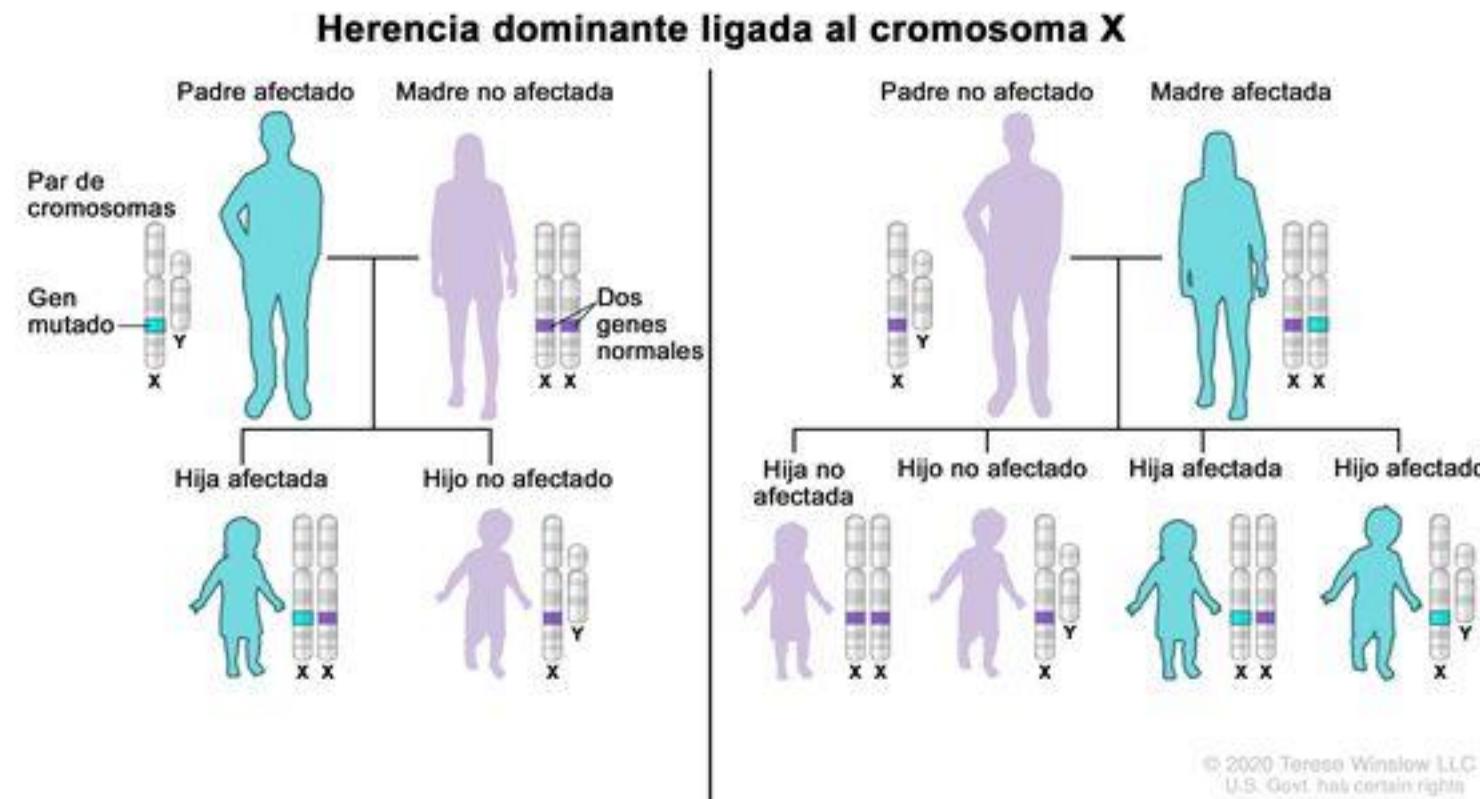
mutaciones en el cromosoma X.

Patrón de herencia especial, siendo poco frecuente.

Las mujeres tienen mayor prevalencia de la enfermedad que los hombres.

Un varón enfermo tendrá a todos sus hijos varones sanos, mientras que a todas sus hijas mujeres enfermas. Por otra parte, una mujer enferma tendrá un 50% de su descendencia enferma, independientemente del sexo.

Ejemplo: Aicardi.



Recesivo ligado al X:

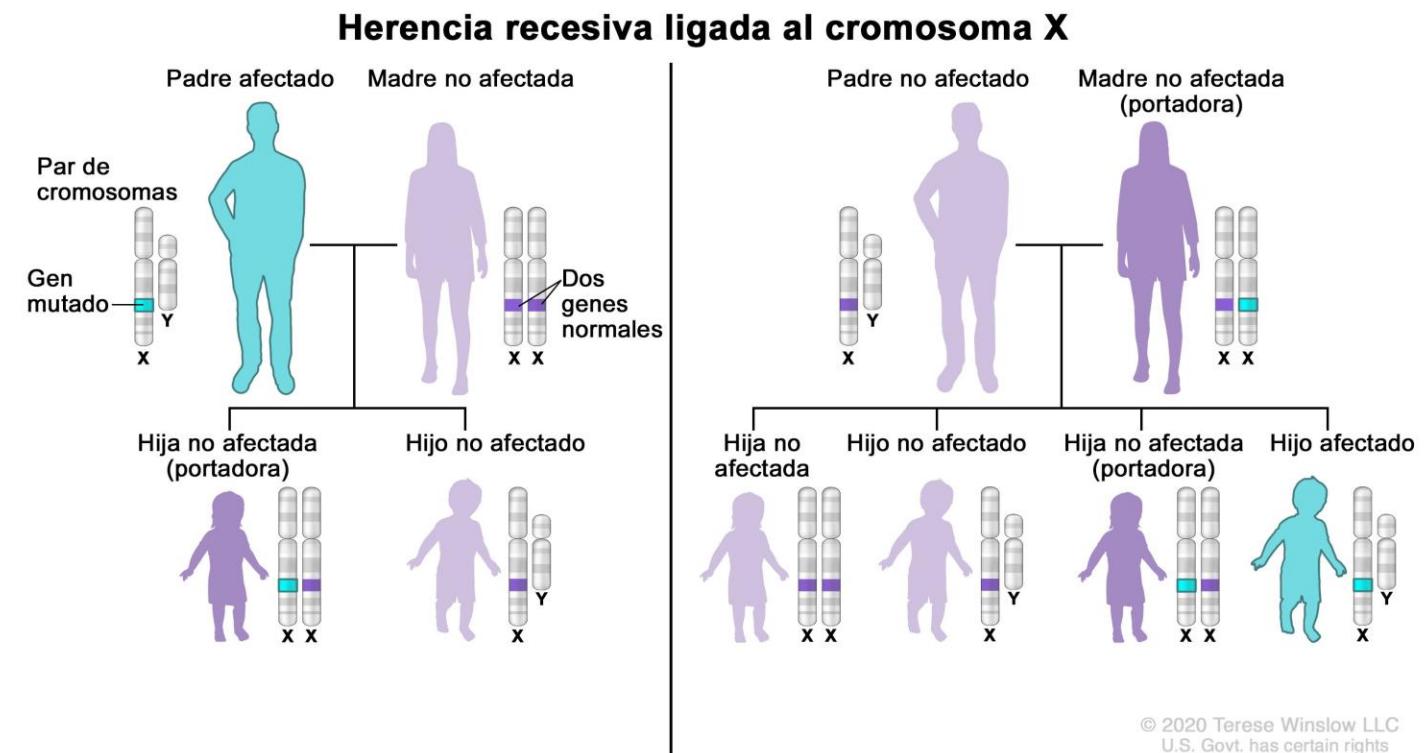
mutaciones en el cromosoma X, en este caso los varones están más frecuentemente afectados.

Un varón portador siempre será enfermo (solo posee un cromosoma X, y éste está afectado).

Su descendencia serán varones sanos (ya que solo les transmite el cromosoma Y) e hijas portadoras.

Una mujer portadora tendrá una descendencia compuesta por un 50% de hijas portadoras y un 50% de hijos enfermos.

Ejemplo: hemofilia A, Duchenne

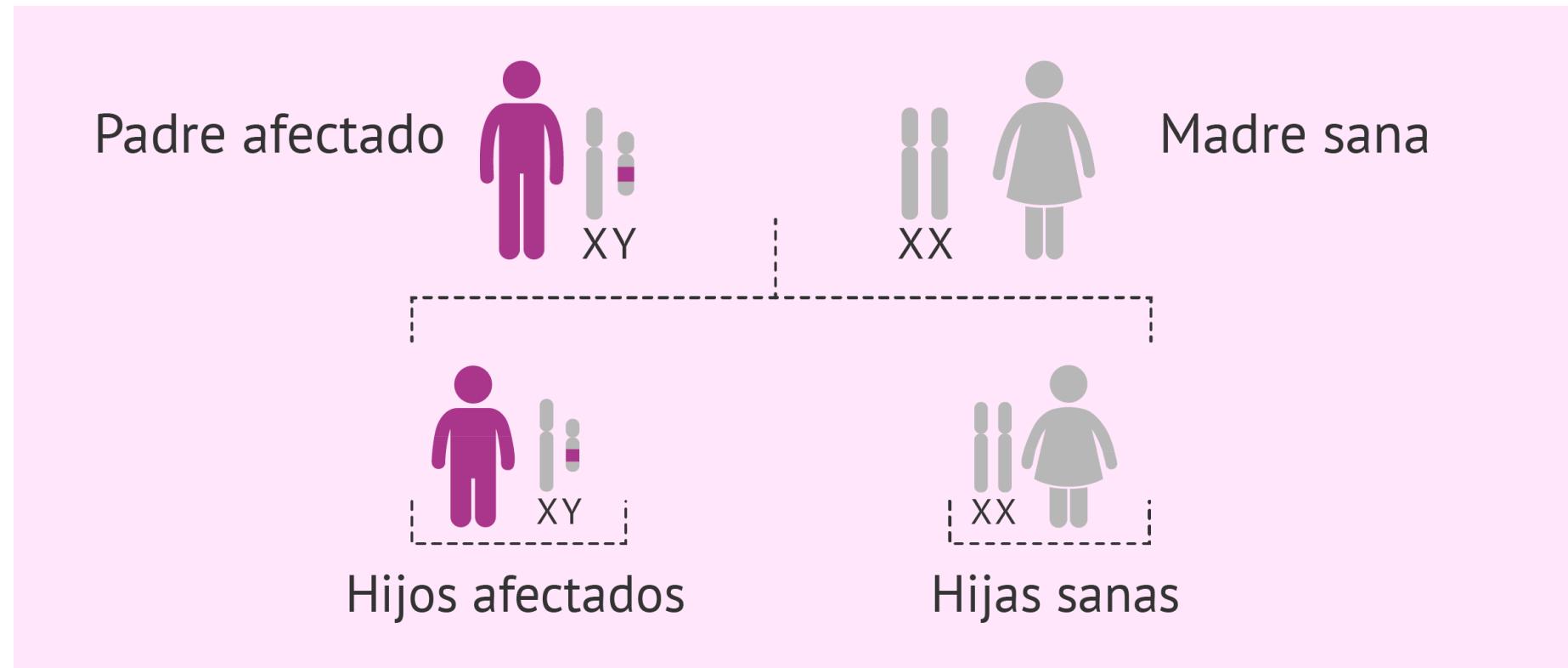


Ligado a Y:

Solo pueden manifestarse en varones, cuya descendencia son 100% de hijas sanas y 100% de varones enfermos.

Habitualmente estas mutaciones causan infertilidad.

Ejemplo: infertilidad masculina hereditaria o síndrome de Jacobs (XYY)

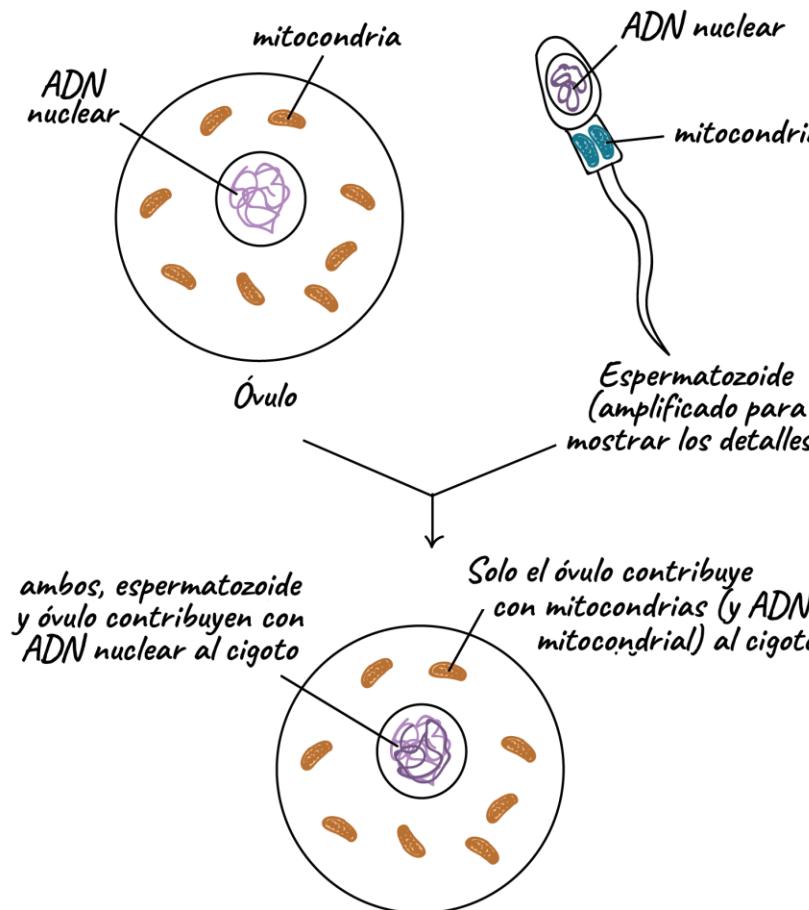


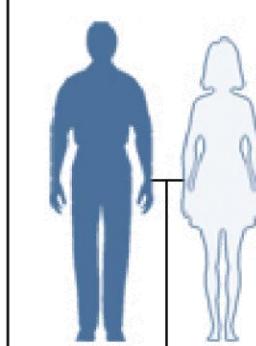
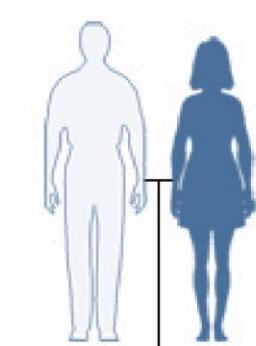
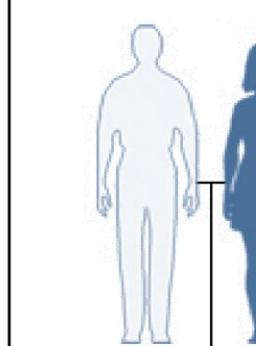
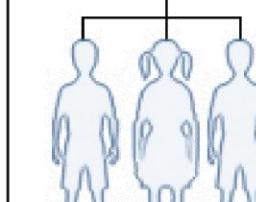
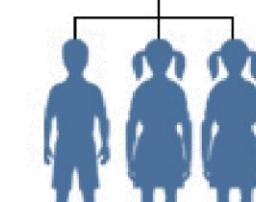
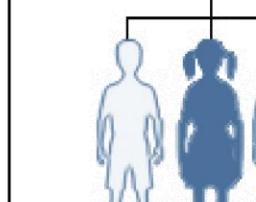
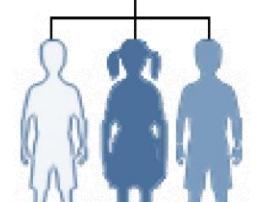
Mitocondrial:

mutaciones en el genoma mitocondrial, que solo se transmite por herencia materna.

La gravedad de la mutación depende del porcentaje de genomas afectados en la población mitocondrial (heteroplasmia).

Ejemplos: neuropatía óptica hereditaria de Leber.



Padre con enfermedad	Madre con enfermedad (todas mitocondrias mutantes)	Madre con enfermedad (algunas mitocondrias normales, otras mutantes)
   <p>Sus hijos no tienen la enfermedad</p>	  <p>Sus hijos tienen la enfermedad (asumiendo penetrancia completa)</p>	  <p>Sus hijos pueden o no tener la enfermedad (y la gravedad puede variar)</p>

Enfermedades multifactoriales o de herencia compleja: En este tipo de enfermedades contribuyen tanto factores genéticos como factores de tipo ambiental, o la interacción entre ambos. Son las enfermedades hereditarias más numerosas, responsables de malformaciones congénitas como labio leporino, alteraciones en el tubo neural, diabetes de tipo 2, algunos tipos de hipertensión, cardiopatías o enfermedades psiquiátricas.

La mutación de un único gen no es suficiente para que se manifieste la enfermedad. Se utiliza el concepto de variación génica o polimorfismo. Para que el individuo desarrolle la enfermedad debe tener una combinación concreta de los genes implicados, de esta forma, las mutaciones en genes individuales pueden tener una frecuencia alta en la población sin causar un efecto en el fenotipo. Es, por tanto, complicado rastrear este tipo de enfermedades, dado que además los factores ambientales tienen un factor contribuyente. Cada mutación tiene una contribución fenotípica distinta y además el riesgo para cada familia puede ser distinto en función del conjunto de factores de riesgo que presente. En estas enfermedades se llevan a cabo estudios de asociación, revisando una batería de genes distribuidos por todo el genoma, localizando aquellos polimorfismos más frecuentes en individuos afectados que en controles. Esto nos arroja un resultado de probabilidad estadística y se establece la hipótesis de que ciertos polimorfismos se relacionan con una enfermedad.

¡Gracias!

The logo consists of the lowercase letters "viu" in white, centered within a dark orange, rounded rectangular shape.

viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades