Identificación de mutaciones patogénicas.

Prácticas de Genética Médica. 1º Curso. Grado en Medicina.

Departamento de Biología, Genética, Histología y Farmacología.



Proyecto de innovación docente. PDI_003_2021/2022.

Desarrollo de un repositorio de datos genómicos e implantación de prácticas online para mejorar el aprendizaje en el análisis genómico humano.

Universidad de Valladolid

1. Introducción.

1.1 Identificación de mutaciones patogénicas.

La identificación de variantes patogénicas es uno de los principales objetivos en el diagnóstico genético, ya que permite conocer la base genética de una enfermedad y, por lo tanto, la toma de decisiones sobre el tratamiento, la evolución y pronóstico de dicha patología.

Aunque existen diferentes estrategias para la detección de variantes genéticas, la secuenciación bidireccional de exones y zonas de splicing mediante el método Sanger es una de las técnicas más utilizadas en los laboratorios clínicos. Para realizar esta técnica se requiere una muestra de sangre periférica u otro tejido del individuo, a partir de la cual se aísla su ADN genómico. A continuación, mediante la reacción en cadena de la polimerasa (PCR) se amplifican todos los exones, junto con las secuencias intrónicas flanqueantes de cada exón que son examinados nucleótido a nucleótido mediante la reacción de secuenciación. De esta manera se puede identificar cualquier diferencia con respecto a la secuencia de referencia de dicho gen (podéis recordar la técnica visualizando el siguiente video: https://www.youtube.com/watch?v=NEu0mO-2ras).



Figura 1: Electroferograma procedente de una paciente con una mutación puntual en heterocigosis.

1.2 Tipos de mutaciones.

Mediante la secuenciación del ADN se pueden identificar distintos tipos de variantes de secuencia que alteran a un único nucleótido (puntuales) o a unos pocos. Las variantes de secuencia se pueden clasificar como **mutaciones claramente patogénicas**, si trucan la proteína resultante o si alteran las secuencias canónicas de splicing (AG/GT). En este grupo se hallan las **mutaciones denominadas sin sentido** (nonsense) que originan un codón de stop (UAG, UAA o UGA), que termina prematuramente la traducción produciendo una proteína truncada. También las mutaciones denominadas de **cambio de pauta de lectura (frameshift)** que implican la **inserción o la deleción de uno o varios nucleótidos cuyo número no es un múltiplo de tres**, de forma que la pauta de lectura de la secuencia del ARNm (ARN mensajero) queda alterada. El resultado es una proteína truncada, debido a la creación de un codón de stop más adelante. Por último, las mutaciones de splicing que alteran el procesamiento del ARNm en la eliminación de los intrones y unión de los exones. Las mutaciones de splicing pueden producirse en distintas partes de un gen, pero las que se consideran claramente patogénicas son las que modifican las

secuencias canónicas donadoras (GT) o aceptadoras (AG) de splicing situadas en los dos nucleótidos iniciales y finales de cada intrón.

Por otro lado, algunas variantes de secuencia tienen una <u>repercusión funcional incierta en la proteína</u> resultante puesto que **no alteran su pauta de lectura (in frame)**. Estos cambios a priori son considerados **variantes de secuencia no clasificadas (UCVs, unclassified sequence variants)**. En este grupo se incluyen las variantes de cambio de aminoácido denominadas de **cambio de sentido (missense)**, **deleciones o inserciones de un número de nucleótidos múltiplo de tres** y las **variantes de secuencia potencialmente implicadas en el splicing** pero que no alteran los nucleótidos canónicos (AG/GT).



Figura 1: Tipos de variantes genéticas. Fuente: NCBI.

1.3 Evaluación del impacto clínico de las variantes genéticas.

Como ya se ha comentado, las variantes de secuencia que trucan la proteína resultante o que alteran las secuencias canónicas de splicing son claramente patogénicas. El problema surge cuando se detecta una variante de significado incierto. La mejor forma de evaluar la patogenicidad de este tipo de variantes es la realización de un análisis funcional, pero no siempre es posible, dado que se trata de estrategias muy laboriosas no siempre al alcance un de un laboratorio de análisis clínico. Para ello se recurre a otras estrategias que utilizan herramientas "in sílico" para evaluar el impacto de la variante:

- 1- La búsqueda de datos en bases de datos de mutaciones, polimorfismos, y/o datos bibliográficos.
- 2- El grado de conservación del aminoácido en un alineamiento múltiple de secuencias con proteínas ortólogas.
- 3- La diferencia bioquímica y biofísica entre el aminoácido original y el mutado.
- 4- Programas de predicción de patogenicidad (Polyphen y SIFT)
- 5- Análisis de la segregación de la variante con la enfermedad en la familia.
- 6- El análisis de cromosomas control.

Estos análisis permiten clasificar las variantes UCVs como: 1) muy probablemente patogénica, 2) probablemente patogénica, 3) variante de efecto indeterminado y 4) probablemente neutra (polimorfismo).

2. Objetivo de la práctica:

El objetivo de esta práctica es aprender a identificar variaciones de secuencias obtenidas por el método de secuenciación SANGER, así como a interpretar el impacto de la variante identificada en la función de la proteína y en la salud del paciente.

3. Metodología:

Para realizar esta práctica necesitamos:

- Ordenador con conexión a internet.
- Programa de análisis de secuencia: Chromas o Finch TV.
- Electroferogramas Sanger procedentes de casos reales (repositorio online en campus virtual).
- Bases de datos y herramientas de análisis online:
 - Ensembl: https://www.ensembl.org/index.html
 - ClinVar: <u>https://www.ncbi.nlm.nih.gov/clinvar/</u>
 - PolyPhen: <u>http://genetics.bwh.harvard.edu/pph2/</u>

Nota: Instalación y manejo del programa: Finch TV.

Accede al campus virtual y en la carpeta "práctica 13: Identificación de variantes" encontrarás una carpeta comprimida con el nombre FINCHTV. Accede e instala el programa FINCHTV. Visualiza el video "Manejo FINCHTV" disponible en la misma carpeta para saber como se maneja el programa.



Figura 2: Programa FINCHTV

4. Desarrollo de la práctica:

En la carpeta: "práctica 13: Identificación de variantes" podrás encontrar una serie de archivos con formato "abi" que corresponden a los electroferogramas que vamos explorar en esta práctica. Como se ha explicado en la introducción, con este tipo de tecnológia secuenciamos tanto la hebra de ADN fordward (+) como la hebra de ADN reverse (-), por eso cada paciente tendrá mímino dos electroferogramas. Hay que comprobar ambas secuencias de cada paciente que vayamos a analizar.

En esta práctica vamos a analizar dos secuencias ejemplo que corresponden a los pacientes 1 y 2. Despues dispondrás de un respositorio de secuencias de diferentes pacientes para practicar de forma autónoma desde casa.

Paciente 1

4.1 Identificación de variantes.

Paso I. Abre el archivo ENG_48_fw con el programa FINCHTV e identifica la alteración:



Figura 3: Variante identificada.

En el ejemplo puedes observar una posición donde vemos dos picos que solapan: uno negro (que corresponde a una G) y uno verde (que corresponde a una A). Se trata de una sustición de una base por otra en heterocigosis.

Paso II. Una vez identificada la variante, desde el mismo programa, podemos comparar la secuencia que estamos analizando con una secuencia de referencia, mediante el análisis de alineamiento de secuencias "BLAST". Esto nos perimitirá determinar el gen que estamos analizando (en base a la similitud de secuencias) y por otro lado la variante que hemos encontrado (es decir, el cambio de nucleotidos/nucleotidos presentes en nuestra secuencia). Para ello, en el propio FINCHTV seleccionamos con el ratón un fragmento de la secuenica de interés que incluya la variante (unos 50-100 nucleótidos) o si la variación es muy grande, la zona

inmediatamente anterior o posterior a la variante. Pinchamos con el botón derecho para que aparezca un desplegable en el que seleccionaremos la opción: "BLAST Sequence"- "Nucleotide, BLASTn"; tal y como se muestra en la figura.



Figura 4: Análisis de alineamiento BLAST desde FINCHTV.

Automáticamente el programa nos redirecciona a la página web del NIH (National Library of Medicine) donde podremos hacer el BLAST de nuestro fragmento de secuencia. Dejamos la configuración que viene por defecto y pinchamos sobre el botón azul (BLAST). Esperamos a que el programa haga su búsqueda.

	National Library of Me Viational Center for Botechnology inf	ormation	Log in
	BLAST [®] » blastn suite	Home Recent Results Saved Strategi	ies Help
blastn b	lastp blastx tblastn tblastx	Standard Nucleotide BLAST	
		BLASTN programs search nucleotide databases using a nucleotide query. more	Reset page Bookmark
Enter Query	Sequence		
Seng 8f (7) GACACTTGTAGC CGACGACGCCAT	number(s), g(s), of MS1A sequence(s) Clair Cuery sublange Council (s) Clair Cuery sublange Council (s) Clair Council (s)		
Or, upload file	Examinar No se ha seleccionado ningún archivo. 😯		
Job Title	ENG 81 (7)		
	Enter a descriptive title for your BLAST search 😧		
Align two or m	ore sequences 😯		
Choose Sear	rch Set		
Database	Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus Nucleotide collection (nn/ht)		
Organism Optional	Enter organism name or id—completions will be suggested exclude (Add organism)		
	Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown 🚱		
Exclude Optional	Models (XMXP) Uncutured/environmental sample sequences		
Limit to	Sequences from type material		
Entrez Query	Tauitte Create custom database		
Optional	Enter an Entrez query to limit search 🕜		
Program Sel	ection		
Optimize for	Tiphy siniar acquetose (negatiast) Ilor of dosimile requencies (discritipiosus negatiast) Somerna III una requencias (disato) Sones a BLAST algoritme 🕑		
BLAST	Search database Nucleotide collection (nrint) using Blastn (Optimize for somewhat similar sequences)		
+ Algorithm	parameters		

Figura 5: National Library of Medicine: BLASTn.

Paso III: Una vez finaliza la búsqueda, el programa nos mostrará todas las secuencias que muestran mayor similitud con la secuencia que hemos introducido. Seleccionamos la que tenga un mayor grado de similitud (cobertura: Query cover). Ten en cuenta que las muestras que estamos analizando son muestras de pacientes, por lo que el organismo de referencia debe ser: "Homo sapiens".

Ł Downle	oad 🗸	Ge	enBank (<u>Graphics</u>				
Homo s	apier	ns end	loglin (E	NG), transcrip	t variant 1, mR	NA		
Sequence	e ID: <u>NI</u>	<u>M_001</u>	114/53.3	Length: 2946 Nu	mber of Matches:	1		
Range 1:	1307 t	to 1407	GenBank	Graphics		▼ <u>N</u> e	ext Match	Previous Match
Score			Expect	Identities	Gaps	Stran	d	
179 bits(198)		4e-41	100/101(99%)	0/101(0	%) Plus/	Plus	
Query 1	L	AGACC		ACCGATCCAGACC				60
Sbjct 1	307	AGACC	tcacccdc	ACCGATCCAGA	TCCTCCCAAGGACA	CTTGTAGCCCGGA	GCTGC	1366
Query 6	51	TCATG	TCCTTGAT	CCAGACAAAGTNTG	CCGACGACGCCATG	101		
Sbjct 1	367	TCATG	TCCTTGAT	CCAGACAAAGTGTG	CCGACGACGCCATG	1407		

Figura 6: Resultados del BLASTn. En nuestro ejemplo: pinchariamos sobre la primera secuencia que corresponde con el ARNm del gen ENG de humano.

Como puedes ver en el ejemplo, la secuencia que estamos analizando (query) corresponde al transcrito del gen de la endoglina, que presenta una similitud del 99% con la secuencia de referencia (sbjct). Esto se debe a que de 101 nucleotidos que hemos introducido en el programa, 1 es diferente a la secuencia de referencia. En nuestra secuencia este cambio de base esta marcado con una N, mientras que en la secuencia de referencia la base nitrogenada que conforma el nucleótido sería una guanina (G). Si vuelves al programa FINCHTV puedes ver que en este caso el cambio que se ha producido es de una G por una A (esto se representa de la siguiente manera: G>A).

- **Paso IV:** Comprueba que la variante identificada se encuentra tanto en la hebra fordware como en la hebra reverese (para ello no olvides darle la vuelta a la secuencia reverse).



Figura 7: Comparación de la secuencia fordware y reverse. Fijate como en la secuencia reverse los picos se superponen tanto que es dificil identificar la variante.

4.2 Evaluación del impacto de la variante en la estructura y función de la proteína.

Una vez identificado el gen de estudio y la variante de interés, debemos averiguar que impacto tendrá la variante en la estructura y función de la proteína y por lo tanto en la salud del paciente. Para ello:

 Paso I. Abrimos la página web: ENSEMBL: <u>https://www.ensembl.org/index.html.</u> Introducimos el nombre del gen y la esepcie que estamos analizando.

LIISE	BLAST/BLAT VEP 100	IS BIOMART Downloads Help &	Docs Blog
Tools	BioMart >	BLAST/BLAT >	Variant Effect Predictor >
<u>All tools</u>	Export custom datasets from Ensembl with this data-mining tool	Search our genomes for your DNA or protein sequence	Analyse your own variants and predict the functional consequences of known and unknown variants
	Search		
	Human	for	

Figura 8: Consulta en la base de datos ensembl.

- **Paso II.** De todas opciones que nos proporciona ensembl seleccionamos el gen de interes que exploraremos.

De nuevo tomando como ejemplo el gen de la endoglina (ENG), obtenemos un listado genes. En este caso pinchamos sobre la primera entrada de la lista.



Figura 9: Resultados de una búsqueda realizada con ensembl.

Paso III. Al pinchar sobre el gen aparece la descripción del mismo, el número de transcritos y mucha información sobre la secuencia. En este punto nos centraremos en la tabla de transcritos para seleccionar el transcrito con más información disponible para explorarlo. Es importante que te fijes en los códigos que usa ensembl para referirse a un gen (ENS<u>G</u>), un transcrito (ENS<u>T</u>) o una proteína (ENS<u>P</u>).

Pincha sobre el ENST que disponga de más información (por ejemplo, un transcrito que contenga una proteina codificante, de mayor tamaño, con código UniProt y RefSeq Match y que presente muchos "flags", que identifican transcritos de mayor calidad o más relevantes).

🥰 Ensembl 💵	AST/BLAT VEP Tools	BioMart	Dow	vnloads 1	Help & Docs Blog				🛃 🗸 Se
Human (GRCh38 p13) 🔻								
calion: 9:127 815 013-127 854 65	8 Gene: ENG Jobs V								
ne-based displays									
Summary Splice variants	Gene: ENG ENSG	60000010	6991						
- Transcript comparison	Description			endoglin [S	ource:HGNC Symbol;A	vcc:HGNC:3349	67		
Gene alleles	Gene Synonyms			CD105, EN	ID, HHT1, ORW, ORW	1			
Secondary Structure	Location			Chromoso	me 9: 127,815,013-127	854,658 reverse	e strand.		
comparative Genomics				GRCh38:C	M000671.2				
· Gene tree	About this gene			This gene	has 5 transcripts (splice	variants), 159 o	orthologues, 2 paralo	gues and is associated	d with <u>5 phenotypes</u> .
 Gene gain/loss tree Orthologues 	Transcripts			Hide tra	nscript table				
- Paralogues				_					
 Ensembl protein families Ontologies 	Show/hide columns (1	I hidden)							Filter
GO: Biological process	Transcript ID 🖕 N	Name 🖕	bp 🖕	Protein 🖕	Biotype 🖕	CCDS 🖕	UniProt Match	RefSeq Match 🔺	Flags
GO: Molec	ENST00000373203.9 E	NG-202	2946	<u>658aa</u>	Protein coding	CCDS48029@	P P17813-1@	NM_001114753.3	MANE Select v0.95 Ensembl Canonical GENCODE basic APPRIS P4 TSI
enotypes netic Variation	ENST0000344849.4 E	NG-201	3059	<u>625aa</u>	Protein coding	<u>CCDS6880</u> ഗ്ര	P17813-2@		GENCODE basic APPRIS ALT2 TSL:1
Variant table	ENST00000480266.5 E	ENG-204	2804	<u>476aa</u>	Protein coding	CCDS75906	P F5GX88	-	GENCODE basic TSL:2
Variant image Structural variants	ENST00000486329.1 E	ENG-205	663	No protein	Processed transcrip		-	-	TSL:2
ne expression	ENST00000462196.1 E	ENG-203	495	No protein	Processed transcrip		-	-	TSL:3
Pathway Regulation	Summary @								
External references Supporting evidence	•••••••••••••••••••••••••••••••••••••••								
) History	Name			ENG 🖻 (HO	GNC Symbol)				
Gene history	CCDS			This gene	is a member of the Hun	nan CCDS set: C	CDS48029.1@, CC	DS6880.1 P. CCDS75	5906.1 @
	UniProtKB			This gene	has proteins that corres	pond to the follo	wing UniProtKB ider	ntifiers: P17813	
Custom tracks	RefSeq			This Enser	nbl/Gencode gene cont	ains transcript(s)) for which we have	selected identical Ref9	Seq transcript(s). If there are other RefSeq transcripts available they will be in the
Custom tracks	LRG			LRG_589	provides a stable genor	nic reference fran	mework for describin	ng sequence variants f	for this gene
	Ensembl version			ENSG0000	00106991.14				
Share this name	Other assemblies			This gene	maps to 130,577,292-1	<u>30,616,937</u> in GF	RCh37 coordinates.		
onare uns page				View this k	ocus in the GRCh37 ard	hive: ENSG0000	<u>00106991</u> ജ		
	Gene type			Protein coo	ling				
	Annotation method			Annotation	for this gene includes I	ooth automatic a	nnotation from Ense	embl and Havana manu	ual curation, see article.

Figura 10: Información del gen. En este ejemplo pinchamos sobre el primer transcrito de la lista.

- **Paso IV.** Una vez seleccionamos el transcrito que nos interesa aparecerá en el menu de la parte izquierda un desplegable que nos permite explorar los exones, el cDNA y la proteina codificante.



Figura 11. Menu para explorar el transcrito del gen de interes selecionado.

Si seleccionamos la opción "exones" podemos hacer una búsqueda del fragmento de la secuencia donde se encontraba nuestra variante. De esta manera podremos ver en qué exón se localiza nuestra variante y si ya se ha descrito con anterioridad. Para ello, copiamos un fragmento pequeño de unos 10-15 nucleótidos de la secuencia inmediatamente anterior o posterior a la variante que hemos identificado y la buscamos (Ctrl + F) en la secuencia de nucleótidos de los exones que muestra Ensembl.

Ten en cuenta que la secuencia de nucleótidos tiene varias líneas por lo que la variación puede estar al inicio o final de la secuencia y puede resultar difícil identificar el fragmento.

F Transcript history Protein history	Exons 🚱								
Configure this page									
A Custom tracks	Download seq	uence							
Export data	Exons/ Introns	Translated sequence Flanking sequence UTR							
< Share this page	Variants	3 prime UTR 5 prime UTR Coding sequence	Frameshift Inframe deletion Infr	ame insertion Missense					
R Bookmark this page		Protein altering variant Splice acceptor Splice d	onor Splice region Start lost S	Stop gained Stop lost					
		Stop retained Synonymous							
	Markup	loaded							
	Show All v entrie			Showhide columns					Filter
	No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence	
	1	ENSE00001181101	127.854.658	127.854.289		1	370		TECTOTICAL TOLATOLATOLATO ALTICIASCOLLECCIONAL GRANGETERANTA AND ALTICAL INCOLLECTOR SOLDES TOL COLLECTOR SOLDES TOL COLLECTOR SOLDES TOL TOLET SOLDES TOLATOL TOLET SOLDES TOLATOL TOLET SOLDES TOLATOL TOLET SOLDES TOLATOL
	2	ENSE00003346729	127.843.245	127.843.094	1	0	152	GTOTICCAGAALACASTOCATUCH SACCTICACGUMAGIG CATATACCACTAGCAACGTOTCCAALCCCCTCCACCCTCCAACGCCTCCCTCCACCACCTCCCTCC	CCCCASCOCCASCASCA MCCCCCCATCCCATCCTT
	3	ENSE00003297799	<u>127.829.827</u>	127,829,687	0	0	141	CORCELLA ACTIVACE TOAT TO TOACE A CORAC A REFERENCE AND A CORCELLA ACTIVACE A CORAC A CORRECT A CORCELLA ACTIVACE A CORCELLA ACTIVACE A CORAC A TO CORRECT A CORCELLA ACTIVACE A CORE A CORAC	<mark>GAARTIGCALCTIGCCCCAR</mark> DTICCAT <mark>NICA</mark> RG <mark>CCCCTGCS</mark> A
	4	ENSE00003486427	<u>127.826.672</u>	<u>127.826.510</u>	0	1	163	HATTOLS CONTONING CHICKAGA COMPONING TO THE COCALENCE CHICKAGE CONTONING COMPONING CONTONING GARCING MICH OF CONTON CATOOR CO	<mark>BAGACCATONCTORC</mark> CCC <mark>ATO</mark> C SCENECCATONCCTORCCTORT CCAA <mark>D</mark>
	5	ENSE00003438349	127.825.860	127.825.695	1	2	166	COLOSE TOMORIS CONCERNING CATES ISSAACS CASE A STOCESSOOD STATE ARE TITE STOCESSOOTS A STOCESSOOD STATE ARE TO TOOLS SOOTS	MAGA CA TERCO COMONINA Cact <mark>teg</mark> aag <mark>oog 1<mark>000000000000000000000000000000000000</mark></mark>
	6	ENSE00000806866	<u>127.825.357</u>	127.825.231	2	0	127	C 00 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	10010056051_01067160051 10010056051_01067160051
	7	ENSE00003738385	<u>127.824.974</u>	127.824.800	0	1	175	2 CONTREMENTATION CONTRACTOR TO THE CONTROL OF A CONTRACT	AM <mark>ONTICE TOGOTICE NOCICE CELENTICE CACENTICE CON CALLES OF GOMENTICE CON</mark>
	8	ENSE00003730037	127.824.446	127.824.304	1	0	143	CTCC ACCTS AND ATTACCED TO TO TO CONTRACT AND A ADDRESS AN	CENCER <mark>CO 21/C</mark> ENCACE (CETA 2 <mark>0 CACCAL CO</mark> CATICA <u>CO 100</u>
	9	ENSE00003605233	127.820.037	127.819.900	0	0	138	CATETICANONCE ACCETICA CONCECTION CONCERNING CONACCETICAN CONCERNING CONCERNI	COLASCE TORONORIO STORES TOTACE ATCONTINUES



Figura 12. Búsqueda de la variante en los exones del transcrito selecionado. *(ejemplo: TTGATCCAGACAAAGT). En nuestro ejemplo la variante identificada se localiza en el exón 7.*

Paso V. Una vez identificado el nucleótido afectado, podemos explorar las variantes descritas. Como habrás podido comprobar, cada nucleotido tiene un color/colores distinto que hace referencia al tipo de mutación o variante que se ha descrito en esa posición.

Exons/ Introns	Translated sec	quence Flar	nking sequence	UTR	1					
Variants	3 prime UTR	5 prime UTF	R Coding sec	luence	Frames	shift	Inframe	deletion	Inframe insertior	Missense
	Protein altering	g variant Sp	olice acceptor	Splice	donor	Splice	e region	Start los	t Stop gained	Stop lost
	Stop retained	Synonymou	s							

Figura 13. Código de colores que identifican el tipo de variante encontrada.

En nuestro ejemplo el nucleótido G donde se localiza la variante (G>A) aparece en amarillo donde se ha descrito una mutación "missense".

Paso VI. Podemos pinchar sobre el nucleótido y explorar qué variantes se han descrito y si alguna corresponde a la variante que hemos identificado.

0	138	CATT	GA <mark>AGTGC</mark> ACC	ATCA <mark>CG</mark> GGCC <mark>T</mark> GA	.c <mark>chweile</mark> ggacc <mark>cca</mark> gct <mark>e</mark> tgagecagag <mark>ea</mark> c Ntactccag <mark>cte</mark> tg <mark>c</mark> atg <mark>cag</mark> igt <mark>cacc</mark> a
		2 featu	res		
0	Variation: <u>CM01</u>	<u>2090</u>	Variation: <u>rs158</u>	<u>8580782</u>	
Ŭ	Class	SNP	Class	SNP	
0	Source	HGMD-PUBLIC	Source	dbSNP	CCTCTCTTTCCAGCTGGGCCTCTACCTCAGC
	Location	9:127824350	Location	9:127824350	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
0	Alleles	HGMD_MUTAT (Forward strand)	Alleles	C/G (Forward strand)	ICCTGC <mark>T</mark> C <mark>CA</mark> GT <mark>TAGAC</mark> AGCTGCCACCTGGAC
	cDNA position	1391	cDNA position	1391	CATCCAGGCCCGGCCGCCAAGGCCAACTGT
	Protein position	363	Protein position	363	reaccceeccificacontcerectecacific co <mark>tc</mark> acc <mark>c</mark> caccccccccccccac
	Consequences	coding sequence variant	Amino acids	C/S	
1	Explore this ver	ient	Codons	tGt/tCt	GAACATCATCAGCCCTGACCTGTCTG
1	Gene/Transcript	Locations	Consequences	missense variant	CCCCTCCTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
	Phenotype Data	1	Explore this var	iant	TGGTACATCTACCTTCCCCCCCCCCCCCCCCCCCCCCCC
-	101				<u>JGCG</u> GT <mark>GC</mark> CC <mark>CG</mark> GCCTCCT <mark>CG</mark> GAGAG <mark>C</mark> A

Figura 14. Variantes descritas en el nucleótido de interés.

En nuestro ejemplo, se han descrito dos variantes, aunque ninguna de ellas corresponde con la que hemos identificado (G>A).

Puesto que hay una parecida (C/G) aunque no sea la misma la vamos a explorar.

Para ello le damos a "explore this variant" y de nuevo aparece otra ventana sobre con información sobre la variante. Fíjate que las variantes descritas tienen un código que siempre se identifica como: rs seguido de un número (ej: rs1588580782).

La variante que hemos elegido es una variante de cambio de sentido (missense) y si exploramos las consecuencias fenotípicas (Phenotype Data), la variante en cuestión se relaciona con la TELANGIECTASIA HEMORRÁGICA HEREDITARIA. A)



B)

Location: 9:127,815,013-127,854,658	Gene: ENG Transcript: ENG-202	Variant: rs1588580782 Jobs 🔻
Variant displays Explore this variant Group Genomic context	rs1588580782 SNP	
- Genes and regulation	Most severe consequence	missense variant See all predicted consequences
Population genetics	Alleles	C/G Ancestrai: C
 Phenotype data 	Change tolerance	CADD: G:23.5 GERP: 2.38
Sample genotypes Linkage diseguilibrium	Location	Chromosome 9:127824350 (forward strand) VCF. 9 127824350 rs1588580782 c g
- Phylogenetic context	Co-located variant	HGMD-PUBLIC CM012090
Citations JD Protein model	Evidence status 🕖	A a
Configure this page	Clinical significance	
K Coninguie and page	HGVS names	This variant has 20 HGVS names - Show 🐵
Custom tracks	Synonyms	This variant has 3 synonyms - Show 🕑
r⊌n Export data	Original source	Variants (including SNPs and indels) imported from dbSNP (release 154) <u>View in dbSNP @</u>
	About this variant	This variant overlaps 6 transcripts, is associated with 1 phenotype and is mentioned in 5 citations.
< Share this page	Bhanatuna Data O	
🗱 Bookmark this page	Phenotype Data	
	Significant association(s)	
	Show/hide columns	Filer
	Phenotype, disease and trait	Source(s) Mapped Ontology Supporting External Clinical Reported Associated Terms Accessions evidence reference significance gene(s) allele
	HEREDITARY HEMORRHAGIC TELANGIECTASIA	Clinvlar@ - Orphanet:774@ Evidence @ - A **** ENG G

Figura 15. A) Información sobre una variante descrita previamente. B) Consecuencias fenotípicas de la variante identificada.

Paso VII. Podemos explorar que consecuencia tendrá el cambio detectado en la secuencia de aminoácidos que codifica el fragmento de ADN que contiene la variante y anotar la variante. Para ello, en la pestaña de transcritos, en vez de seleccionar exones **exploraremos el cDNA**.

Copiamos de nuevo el fragmento de secuencia donde se localiza la variante e identificamos el codón que se verá afectado y el cambio que supone dicha variante.

A)		
1381 1078 360	*R***M*M *SB* BRVNVRY** * YSS*RR*** ***********************	1440 1137 379
В)	Códigos de Aminoácido	DS

-												Código de	Códig		Código de	Códig
			SEGUN	IDA LET	RA DEL C	CODÓN					Aminoácido	tres letras	o una	Aminoácido	tres letras	o una
	U		C	2	A	4	(G					letra			letra
	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U		Alanina	Ala	A	Leucina	Leu	L
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	С		A		_	T false	· · · ·	
0	UUA	Leu	UCA	Ser	UAA	STOP	UGA	STOP	Α		Arginina	Arg	R	Lisina	Lys	K
	UUG	Leu	UCG	Ser	UAG	STOP	UGG	Trp	G	긆	Asparagina	Asn	Ν	Metionina	Met	М
	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U	RC	Aspártico	Asn	D	Fenilalanina	Pho	-
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	С	ER,			D	i cinananna	T IIC	F
C	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	Α	ALI	Cisteina	Cys	С	Prolina	Pro	Р
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G	井	Glutámico	Chu	-	Corino	Con	
	AUU	lle	ACU	Thr	AAU	Asn	AGU	Ser	U	A	Gratanico	Giù	E	Senna	Ser	S
•	AUC	lle	ACC	Thr	AAC	Asn	AGC	Ser	С	Ĕ	Glutamina	Gln	Q	Treonina	Thr	т
~	AUA	lle	ACA	Thr	AAA	Lys	AGA	Arg	Α	8	Glicina	Cly	0	Triptófano	Trp	14/
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G	DČ	Gnema	uly	G			vv
	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U	z	Histidina	His	н	Tirosina	Tyr	Y
C	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	С		Isoleucina	Пе		Valina	Val	V
G	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	Α				1			v
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G		Selenocisteina	Sec	U	Pirrolisina	Pyl	0
	U C G	UUU UUU UUU UUG UUG UUG CUU CUU CUU CUU CUU CUU CUU CUG CUA AUU AUU AUA AUG GUU GUG GUG GUG GUG	UUU Phe UUU Phe UUC Leu CUC Leu CUG Leu CUG Leu CUG Leu AUC Ile AUC Ile AUG Met GUU Val GUU Val GUG Val GUG Val	SEGUN U Phe UCC UUU Phe UCC UUC Phe UCC UUC Phe UCC UUC Phe UCC UUG Leu UCG CUU Leu CCU CUC Leu CCC CUG Leu CCG CUG Leu CCG AUU Ile ACU AUC Ile ACC AUG Met ACG GUU Val GCU GUU Val GCU GUG Val GCA GUG Val GCA GUG Val GCA	U SEGUNDA LET U C UUU Phe UCU Ser UUC Phe UCU Ser UUC Phe UCC Ser UUG Leu UCA Ser UUG Leu UCA Ser UUG Leu UCG Ser CUU Leu CCU Pro CUC Leu CCA Pro CUG Leu CCA Pro CUG Leu CCA Pro CUG Leu CCA Pro AUU Ile ACU Thr AUC Ile ACC Thr AUG Met ACG Thr AUG Met ACG Thr GUU Val GCU Ala GUZ Val GCC Ala GUG Val GCA Ala GUG Val GCG	UUU Phe UCU Ser UAU UUU Phe UCU Ser UAU UUC Phe UCU Ser UAU UUC Phe UCA Ser UAA UUG Leu UCA Ser UAA UUG Leu UCA Ser UAA UUG Leu UCA Ser UAA CUU Leu UCA Ser UAG CUU Leu CCC Pro CAA CUG Leu CCG Pro CAA CUG Leu CCG Pro CAA CUG Leu CCG Pro CAA AUU Ile ACU Thr AAA AUG Ile ACA Thr AAA AUG Met ACG Thr AAG GUU Val GCU Ala GAA GUC Val <	SEGUNDA LETRA DEL CODÓN U C A UUU Phe UCU Ser UAU Tyr UUC Phe UCU Ser UAU Tyr UUC Phe UCA Ser UAA STOP UUG Leu UCA Ser UAA STOP UUG Leu UCG Ser UAG STOP UUG Leu UCA Ser UAG STOP CUU Leu CCC Pro CAU His CUC Leu CCA Pro CAA Gin CUG Leu CCG Pro CAA Gin CUG Leu CCG Pro CAA Gin CUG Leu CCG Pro CAA Asn AUU Ile ACA Thr AAA Lys AUG Met ACG Thr AAG Lys <tr< td=""><td>$\begin{tabular}{ c c c c } \hline &$</td><td>SEGUNDA LETRA DEL CODÓN U C A G UUU Phe UCU Ser UAU Tyr UGU Cys UUC Phe UCC Ser UAU Tyr UGC Cys UUC Phe UCA Ser UAA STOP UGA STOP UUG Leu UCA Ser UAG STOP UGA STOP UUG Leu UCG Ser UAG STOP UGA STOP CUU Leu CCC Pro CAU His CGC Arg CUG Leu CCC Pro CAG His CGC Arg CUG Leu CCG Pro CAG Gin CGA Arg AUU Ile ACU Thr AAU Asn AGG Ser AUG Ile ACG Thr AAA Asp AGG Arg</td><td>$\begin{tabular}{ c c c c } \hline \$ U \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$</td><td>$\begin{tabular}{ c c c c } \hline \hline \$ U \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$</td><td>SEGUNDA LETRA DEL CODÓN Aminoácido U C A G Aminoácido UUU Phe UCU Ser UAU Tyr UGU Cys U Alanina UUU Phe UCA Ser UAA STOP UGC Cys C Alanina UUU Phe UCA Ser UAA STOP UGA STOP A Arginina UUG Leu UCG Ser UAA STOP UGG Arg Asparagina Asparagina CUU Leu CCA Pro CAA His CGU Arg A Giutámico CUG Leu CCG Pro CAA Gin CGA Arg A Giutámico CUG Leu CCG Pro CAA Asn AGC Ser U Giutámico CUG Leu CCG Pro CAA Asn AGC Ser C Giutámina AUG Ile ACG Thr AAG<!--</td--><td>$\begin{tabular}{ c c c c c c c c c c c c c c c c c c c$</td><td>SEGUNDA LETRA DEL CODÓN U C A G UUU Phe UCU Ser UAU Tyr UGU Cys U UUC Phe UCC Ser UAU Tyr UGC Cys U UUC Phe UCC Ser UAU Tyr UGC Cys C UUC Phe UCC Ser UAA STOP UGC Cys C UUG Leu UCG Ser UAA STOP UGA STOP A UUG Leu UCG Ser UAG STOP UGG Trp G CUU Leu CCG Pro CAU His CGC Arg U CUC Leu CCG Pro CAU His CGG Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUU Leu CCG Pro CAG Gin CGA Arg A AUU Ile ACG Thr</td><td>$\begin{array}{ c c c c c c c c c c c c c c c c c c c$</td><td>$\begin{array}{c c c c c c c c c c c c c c c c c c c$</td></td></tr<>	$\begin{tabular}{ c c c c } \hline & & & & & & & & & & & & & & & & & & $	SEGUNDA LETRA DEL CODÓN U C A G UUU Phe UCU Ser UAU Tyr UGU Cys UUC Phe UCC Ser UAU Tyr UGC Cys UUC Phe UCA Ser UAA STOP UGA STOP UUG Leu UCA Ser UAG STOP UGA STOP UUG Leu UCG Ser UAG STOP UGA STOP CUU Leu CCC Pro CAU His CGC Arg CUG Leu CCC Pro CAG His CGC Arg CUG Leu CCG Pro CAG Gin CGA Arg AUU Ile ACU Thr AAU Asn AGG Ser AUG Ile ACG Thr AAA Asp AGG Arg	$\begin{tabular}{ c c c c } \hline $ U $ $ $ $ $ $ $ $ $ $ $ $ $ $ $ $ $$	$\begin{tabular}{ c c c c } \hline \hline $ U $ $ $ $ $ $ $ $ $ $ $ $ $ $ $ $$	SEGUNDA LETRA DEL CODÓN Aminoácido U C A G Aminoácido UUU Phe UCU Ser UAU Tyr UGU Cys U Alanina UUU Phe UCA Ser UAA STOP UGC Cys C Alanina UUU Phe UCA Ser UAA STOP UGA STOP A Arginina UUG Leu UCG Ser UAA STOP UGG Arg Asparagina Asparagina CUU Leu CCA Pro CAA His CGU Arg A Giutámico CUG Leu CCG Pro CAA Gin CGA Arg A Giutámico CUG Leu CCG Pro CAA Asn AGC Ser U Giutámico CUG Leu CCG Pro CAA Asn AGC Ser C Giutámina AUG Ile ACG Thr AAG </td <td>$\begin{tabular}{ c c c c c c c c c c c c c c c c c c c$</td> <td>SEGUNDA LETRA DEL CODÓN U C A G UUU Phe UCU Ser UAU Tyr UGU Cys U UUC Phe UCC Ser UAU Tyr UGC Cys U UUC Phe UCC Ser UAU Tyr UGC Cys C UUC Phe UCC Ser UAA STOP UGC Cys C UUG Leu UCG Ser UAA STOP UGA STOP A UUG Leu UCG Ser UAG STOP UGG Trp G CUU Leu CCG Pro CAU His CGC Arg U CUC Leu CCG Pro CAU His CGG Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUU Leu CCG Pro CAG Gin CGA Arg A AUU Ile ACG Thr</td> <td>$\begin{array}{ c c c c c c c c c c c c c c c c c c c$</td> <td>$\begin{array}{c c c c c c c c c c c c c c c c c c c$</td>	$\begin{tabular}{ c c c c c c c c c c c c c c c c c c c$	SEGUNDA LETRA DEL CODÓN U C A G UUU Phe UCU Ser UAU Tyr UGU Cys U UUC Phe UCC Ser UAU Tyr UGC Cys U UUC Phe UCC Ser UAU Tyr UGC Cys C UUC Phe UCC Ser UAA STOP UGC Cys C UUG Leu UCG Ser UAA STOP UGA STOP A UUG Leu UCG Ser UAG STOP UGG Trp G CUU Leu CCG Pro CAU His CGC Arg U CUC Leu CCG Pro CAU His CGG Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUG Leu CCG Pro CAG Gin CGA Arg A CUU Leu CCG Pro CAG Gin CGA Arg A AUU Ile ACG Thr	$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	$ \begin{array}{c c c c c c c c c c c c c c c c c c c $

Figura 16. A) Posición del nucleótido y el aminoácido alterado. B) Código genético.

La variante que hemos identificado (G>A) supondría un cambio en el aminoácido 363, que cambia de cisteína (C, Cys, codón: U<u>G</u>U) a tirosina (Y, Tyr, codón: U<u>A</u>U). Por lo tanto, la variante que nosotros hemos identificado es también una missense. La mutación se anota de la siguiente manera: **c.1088 G>A p.Cys363Tyr**. (donde la c.1088 hace referencia a la posición del nucleótido del cDNA alterado, y p.363 la posición del aminoácido alterado).

Paso VIII. Otra opción para saber si la variante es patogénica o no es consultar en diferentes bases de datos. Para ello entramos en ClinVar (<u>https://www.ncbi.nlm.nih.gov/clinvar/</u>):

- Si disponemos del código rs de la variante podemos buscarlo directamente.
 - *Ej.: rs1588580782: descrita como probablemente patogénica.*
- Si no disponemos del código rs de la variante podemos buscarlo usando la anotación de la variante identificada junto con el código del transcrito NM.
 - Ej: NM_001114753.3: c.1088 G>A p.Cys363Tyr. En nuestro caso no esta descrita.

4.3 <u>Predicción bioinformática de los efectos potenciales de las variantes genéticas</u> <u>en la función y estructura de proteínas.</u>

Si la variante identificada no se ha descrito previamente deberíamos predecir si ésta será patogénica o no. Para ello podemos acceder a simuladores "in sílico" de los efectos potenciales de las variantes genética, por ejemplo, PolyPhen: <u>http://genetics.bwh.harvard.edu/pph2/</u>.

Para ello necesitamos:

- el código de la proteína: *ENSP0000362299*, o un archivo que contenga la secuencia completa de aminoácidos (en un formato especial denominado FASTA que podemos descargar de Ensembl).
- la posición donde se produce el cambio de aminoácido.
- o los aminoácidos que cambian.

En nuestro caso introducimos el código de la proteína del Ensembl: ENSP00000362299 la posición del aminoácido: 363 y la sustitución: C>Y.

Le damos a "Submit Query" y pasado un tiempo aparecerá nuestro resultado.

A)

PolyPhen-2	prediction	of fun	ctional	l effect	sofh	umai	n nsS	NPs										
Home Abo	at F	leip	Do	wnioads	Ba	itch qi	uery	WI	IESS.	db	_	-	-	-	-	-	-	
Query Data																		
Protein or SNP identifie	r	ENS	SP00	00036	6229	9												
Protein sequence in FASTA forma	•																	
Position	1															363		
Substitution	AA ₁ AA ₂	A R A R	N N	D C	E	Q Q	G G	H H	 	L K	C N	1 F 1 F	P P	S S	T T	W W	Y Y	V V
Query description	1																	
								1	Sub	mit	Qu	ery	Cle	ar	Che	eck (Stat	us
								Dis	spla	ay a	adv	and	ed	qu	ery	ор	tio	ns

B)

	Grid Gate	way Interface v2.2.8 bleshooting/FAQ	se caret Sunya	evLab
Service Name:	PolyPhen-2			
Session ID:	b4081256b9f69b50	35d22bd5044ade732c5	8501f	Overwrite default
Grid Status:				
Load Heal Light 889	th Jobs: Pend % 5	ing Running 21		
Jobs total):				
	Complet	ed (1)		
ID R	esults Errors	Date/Time	Delete	Description
8351878	View -	2022-05-16 13:22:11		
Refresh All ite	ems with Delete bo	kes checked <u>will be re</u>	moved!	

Figura 17. A) Búsqueda en PolyPhen. B) Pantalla para acceder al resultado.

Pinchamos en "view results", y nos aparece el resultado: en este caso Polyphen determina que la variante identificada es probablemente patogénica.





Paciente 2:

Repetimos los pasos realizados para el paciente 1.

5.1. Identificación de variantes.

- **Paso I.** Abre el archivo TP53 268 4f_E01 con el programa FINCHTV e identifica la alteración:



Figura 19: Variante identificada.

En el ejemplo puedes observar que apartir de una determinada posición toda la secuenica de nucleótidos cambia y por lo tanto tambien lo hara la secuencia de aminoácidos. Esto es una mutación que afecta al marco de lectura (frameshift variant).

- Paso II y III. Identificamos el gen y la secuencia de referencia.

Gen: TP53.

Secuencia de referencia: ATGCTGTCCC<mark>C</mark>GGACGATATTGAA

Secuencia problema: ATGCTGTCCCGGACGATATTGAA

En este caso vemos claramente que se ha producido una delecion de un C que ha hecho que la secuencia se desplace a la izquierda. Esto se reproduce tanto en la hebra fordware como en la reverse. Se trata de una delecion de un nucleotido que cambia el marco de lectura y que por definición tendrá consecuencias patogénicas.

5.2. Impacto de la variante en la estructura y función de la proteína.

- **Pasos I-VI:** Buscamos en ensembl la variante para ver si la deleción ya ha sido descrito y a que enfermedad se asocia.

La variante se localiza en el exon 4 del gen TP53. Se ha identificado un variante idéntica: rs1567557016 que se relaciona con el Síndrome de Li-Fraumeni.

- **Paso VII y VIII:** Buscamos la posción del nucleótido afectado y del aminoácido en esembl.

241 99 33	*VB***BYHR*YB*****************************	300 158 53
301 159 53	*Y***********************************	360 218 73
361 219 73	**YY****M*H**RSR*S*Y*****************************	420 278 93
421 279 93	***D**********************************	480 338 113

Figura 20: Variante identificada.

- o posción del nucléotido en el cDNA: 140
- tipo de mutación: : delecion de una C (del).
- Posición del primer aminoácido que cambia y tipo: 47 P (Pro)
- Consecuencia: cambio en el marco de lectura: frameship (fs)

Anotación: c.140del (p.Pro47fs).

En clinVar esta variante está clasificada como patogénica. NM_000546.6(TP53):c.140del (p.Pro47fs).

Podemos seguir explorando la variante en ClinVar y vemos que se trata de una mutación que crea un codón de stop prematuro en la posición del aminoácido 76 y que se ha relacionado con el síndrome de Li-Fraumeni. Si nos metemos en Orphanet podemos ver en que consiste la enfermedad: <u>https://www.orpha.net/consor/cqi-bin/index.php?lng=ES</u>.

En este caso ya no es necesario seguir investigando puesto que la variante es claramente patogénica.

Repite el proceso con el resto de secuencias disponibles en el repositorio para practicar de forma autónoma y si tienes alguna duda utiliza el FORO.