



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**DOCTORADO EN CIENCIAS BIOMÉDICAS**

**LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO**

IDENTIFICACIÓN DE FACTORES GENÉTICOS DE RIESGO PARA EL DESARROLLO DE MELANOMA

EXAMEN DE TITULACIÓN

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS BIOMÉDICAS

PRESENTA:

**M.C. RAUL OSSIO VELA**

DIRECTORA DE TESIS: DRA. CARLA DANIELA ROBLES ESPINOZA

LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO, UNAM.

TUTOR: DR. DANIEL PIÑERO DALMAU

INSTITUTO DE ECOLOGÍA, UNAM.

TUTOR: DR. FRANCISCO XAVIER SOBERÓN MAINERO

INSTITUTO NACIONAL DE MEDICINA GENÓMICA

JURIQUILLA, QUERÉTARO. ABRIL 2020



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Este trabajo se realizó en el Grupo de Genética del Cáncer y Bioinformática, del Laboratorio Internacional de Investigación sobre el Genoma Humano de la Universidad Autónoma de México (UNAM), bajo la dirección de la Dra. Carla Daniela Robles Espinoza.

# Agradecimientos

Al director del LIIGH-UNAM, Dr. Rafael Palacios, al personal administrativo del LIIGH-UNAM, en especial, al, Abigayl Hernández y Eglee Lomelín, por las facilidades y su buena disposición siempre que hubiera algún problema.

Al personal técnico del LIIGH-UNAM, Alejandra Castillo, Carina Uribe y Jair S. García-Sotelo, por ayuda crucial en llevar a cabo partes del trabajo reportado en esta tesis. Asimismo, a Luis A. Aguilar, del LAVIS-UNAM, y Diego Said Anaya Mancilla de la UAQ, por su invaluable ayuda en la creación de la herramienta reportada en este trabajo.

A mi comité tutor, Dr. Daniel Piñero Dalmau y Dr. Xavier Soberón Mainero, por sus buenas ideas y su guía durante la realización de este trabajo. Asimismo, a mis colaboradores internacionales, en especial el Dr. David Adams, por consejos y la financiación de una parte de este proyecto.

Al Programa de Doctorado en Ciencias Biomédicas (PDCB) de la UNAM, en especial a la Dra. Aurea Orozco y al Dr. Pavel Rueda, por su apoyo y su disposición para la resolución de situaciones complicadas.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), de México por la asignación de una beca para estudios de doctorado (número 573128), para poder llevar a cabo el estudio aquí detallado.

Al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la Universidad Nacional Autónoma de México por el apoyo en las diferentes fases del proyecto (IA200318).

A mi tutora, la Dra. Carla Daniela Robles Espinoza, por la dirección del trabajo presentado aquí, y demás miembros del Grupo de Genética del Cáncer y Bioinformática por su ayuda y apoyo siempre, en especial, a Omar Isaac García Salinas y Carolina Castañeda García.

# Resumen

El descubrimiento de nuevas variantes genéticas de línea germinal que incrementan el riesgo a desarrollar melanoma es de vital importancia para entender la biología de la enfermedad y para brindar consejo genético a pacientes portadores. Uno de los genes con variantes patogénicas descritos recientemente en familias con melanoma es *protection of telomeres (POT1)*, el cual es parte del complejo protector de los telómeros y ejerce su función de protección y regulación de la longitud telomérica uniéndose al extremo de ADN de cadena sencilla localizado al final de los cromosomas. Este gen también se ha encontrado mutado en familias con otros tipos de cáncer como glioma, leucemia y linfoma.

En el trabajo reportado en esta tesis, se realizó la secuenciación de todos los exones del gen *POT1* en 6,227 individuos (2,929 casos de melanoma y 3,298 controles poblacionales) provenientes del Reino Unido para detectar nuevas variantes que confieran un riesgo elevado a padecer la enfermedad. Para ayudar en el proceso de identificación, filtrado y priorización de estas variantes, se creó un nuevo algoritmo, llamado VCF/Plotein, el cual está diseñado para que investigadores sin formación en ciencias bioinformáticas sean capaces de llevar a cabo análisis de exomas de manera gratuita, rápida y segura. Las pruebas de este algoritmo en un Caso de Uso y otras publicaciones validaron su utilidad para ser aplicado en el presente trabajo.

De esta manera, se identificaron 43 variantes distintas posiblemente patogénicas en *POT1*. Aplicando VCF/Plotein, se hicieron predicciones de patogenicidad para cada una de ellas, con estas predicciones indicando que 15 alterarían la función de la proteína y por tanto tendrían consecuencias funcionales. Subsecuentemente, y como prueba de principio, se realizaron experimentos de unión a ADN para las proteínas con estas variantes, los cuales confirmaron las predicciones de patogenicidad de VCF/Plotein para 7/10 variantes. Subsecuentemente se realizaron experimentos de medición de longitud telomérica por PCR para todos los individuos portadores.

Como conclusión, se describió que las variantes patogénicas en *POT1* tienen una contribución limitada al riesgo poblacional a desarrollar melanoma, con aproximadamente 1/200 casos portando una variante de este tipo en este gen. Las variantes que interrumpen el complejo POT1-ADN en general conllevan a una longitud telomérica mayor, pero la variabilidad observada en los experimentos de PCR, y otros posibles factores genéticos y/o ambientales, complican la interpretación biológica de la relación entre las variantes de POT1 y la longitud telomérica de sus portadores.

# Tabla de contenido

<b>1</b>	<b>INTRODUCCIÓN</b>	<b>12</b>
<b>1.1</b>	<b>El cáncer es una enfermedad del genoma</b>	<b>12</b>
<b>1.2</b>	<b>Identificación y análisis de variantes genéticas</b>	<b>14</b>
1.2.1	Tipos de variantes genéticas	15
1.2.2	Anotación de variantes genéticas	17
1.2.3	Efectos de las variantes genéticas	19
1.2.4	Priorización de variantes genéticas	21
1.2.5	Interpretación de los efectos de variantes genéticas	22
<b>1.3</b>	<b>El melanoma</b>	<b>24</b>
1.3.1	Factores de riesgo	26
1.3.2	El genoma de los tumores de melanoma	32
<b>1.4</b>	<b>Justificación del proyecto presentado en esta tesis</b>	<b>34</b>
<b>2</b>	<b>VCF/PLOTEIN: CREACIÓN DE UN ALGORITMO PARA EL ANÁLISIS DE VARIANTES GENÉTICAS PROVENIENTES DE ESTUDIOS DE SECUENCIACIÓN</b>	<b>35</b>
<b>2.1</b>	<b>Introducción</b>	<b>35</b>
<b>2.2</b>	<b>Métodos</b>	<b>37</b>
2.2.1	Características del servidor y lenguajes de programación	37
2.2.2	Interfaz de programación de aplicaciones	37
2.2.3	Archivo de entrada	38
2.2.4	Flujo funcional de la aplicación	38
2.2.5	Construcción y partes de la gráfica generada	39
<b>2.3</b>	<b>Resultados</b>	<b>42</b>
2.3.1	Panorama	42
2.3.2	Seguridad de los datos	43
2.3.3	Filtrado de variantes y visualización	44
2.3.4	Desempeño	44



2.3.5	Archivos marcador	46
2.3.6	Comparación con otras herramientas similares	46
<b>2.4</b>	<b>Caso de uso</b>	<b>46</b>
2.4.1	Priorizando variantes genéticas en el gen <i>BAP1</i> que predisponen a melanoma	46
<b>2.5</b>	<b>Discusión</b>	<b>52</b>
<b>2.6</b>	<b>Disponibilidad de la aplicación</b>	<b>53</b>
<b>2.7</b>	<b>Equipo de trabajo</b>	<b>53</b>
<b>2.8</b>	<b>Perspectivas</b>	<b>53</b>
<b>3</b>	<b>APLICACIÓN DE VCF/PLOTEIN: PRIORIZANDO VARIANTES EN <i>POT1</i> EN UN ESTUDIO DE CASOS Y CONTROLES DE MELANOMA</b>	<b>55</b>
<b>3.1</b>	<b>Introducción</b>	<b>55</b>
<b>3.2</b>	<b>Métodos</b>	<b>56</b>
3.2.1	Participantes y recolección de las muestras	56
3.2.2	Secuenciación e identificación de variantes	57
3.2.3	Predicción de variantes deletéreas con VCF/Plotein	60
3.2.4	Traducción <i>in vitro</i> y ensayos de unión a telómeros	63
3.2.5	Alineamiento multi-especies	63
3.2.6	Análisis de longitud telomérica por medio de PCR	64
<b>3.3</b>	<b>Resultados</b>	<b>65</b>
<b>3.4</b>	<b>Discusión</b>	<b>69</b>
<b>3.5</b>	<b>Equipo de trabajo y agradecimientos</b>	<b>72</b>
<b>4</b>	<b>CONCLUSIONES Y ESTUDIOS FUTUROS</b>	<b>74</b>
<b>5</b>	<b>BIBLIOGRAFÍA</b>	<b>78</b>
<b>6</b>	<b>ANEXOS Y MATERIAL SUPLEMENTARIO</b>	<b>89</b>

<b>6.1</b>	<b>Tabla Suplementaria 1</b>	<b>89</b>
<b>6.2</b>	<b>Línea de comando HaplotypeCaller (GATK)</b>	<b>89</b>
<b>6.3</b>	<b>Tabla Suplementaria 2</b>	<b>90</b>
<b>6.4</b>	<b>Tabla Suplementaria 3</b>	<b>90</b>
<b>6.5</b>	<b>Tabla Suplementaria 4</b>	<b>91</b>
<b>7</b>	<b>PUBLICACIONES ORIGINADAS EN ESTE DOCTORADO</b>	<b>92</b>

# Índice de figuras

Figura 1. Acumulación de mutaciones debido a procesos endógenos y exógenos a lo largo de la vida de una célula.....	13
Figura 2. La frecuencia alélica de las variantes genéticas en una población correlaciona inversamente con el tamaño del efecto sobre el fenotipo. ....	14
Figura 3. Estructura de un archivo VCF .....	17
Figura 4. Representación gráfica de los distintos tipos de variantes de acuerdo a la clasificación hecha por Ensembl.....	21
Figura 5. Subtipos histopatológicos más importantes de melanoma. ....	26
Figura 6. Distribución de subtipos comunes de melanoma en diferentes países alrededor del mundo .....	27
Figura 7. Funciones de los genes involucrados en el riesgo a melanoma y vías biológicas alteradas.....	29
Figura 8. Diagrama de flujo con el flujo funcional de VCF/Plotein y la arquitectura del sistema .....	40
Figura 9. Pantalla de trabajo de VCF/Plotein .....	42
Figura 10. Pantalla de carga y filtrado.....	43
Figura 11. Pantalla de carga de VCF/Plotein con los datos del caso de uso.....	48
Figura 12. Pantalla con vista del péptido resultante del transcrito canónico de BAP1.....	49
Figura 13. Menú de filtrado de consecuencias por variante.....	50
Figura 14. Menú de filtrado por predicción de patogenicidad por SIFT y PolyPhen-2.....	51
Figura 15. Vista de las variantes restantes en BAP1 después de los filtros de priorización.....	52
Figura 16. Gráfica de mutaciones presentes en POT1 elaborada con VCF/Plotein.....	60
Figura 17. Modelo lineal ajustando por sexo y edad al momento del diagnóstico para muestras de individuos que no eran portadores de variantes clasificadas en el Grupo 1.....	65
Figura 18. Longitudes teloméricas de portadores de variantes en POT1 clasificadas en el Grupo 1 mostradas sobre una distribución de longitudes teloméricas de casos y controles de melanoma que no portan estas variantes. ....	66
Figura 19. Variabilidad de las réplicas técnicas de las mediciones de las longitudes teloméricas de los individuos portadores de variantes de POT1 clasificadas en el Grupo 1 .....	67
Figura 20. Experimentos EMSA probando la capacidad de unión a ssDNA de proteínas mutantes POT1 traducidas in vitro. ....	68
Figura 21. Conservación en distintas especies de residuos con mutaciones en POT1 en este estudio .....	69

# Índice de tablas

Tabla 1. Clasificación de los efectos de las variantes genéticas.....	20
Tabla 2. Variantes genéticas en el gen POT1 encontradas en familias propensas a desarrollar melanoma. ....	32
Tabla 3. Desempeño de VCF/Plotein después de cargar tres archivos VCF distintos en diferentes sistemas operativos con un rango de especificaciones de hardware. ....	45
Tabla 4. Comparación de las principales características de VCF/Plotein con aquellas de otras herramientas similares. ....	47
Tabla 5. Variantes con consecuencia de alteración de estructura de la proteína identificadas por medio de secuenciación Fluidigm y confirmadas por medio de resecuenciación por capilar o Illumina. ....	59
Tabla 6. Predicción de patogenicidad de variantes encontradas en POT1 de acuerdo a VCF/Plotein.....	61

# 1 Introducción

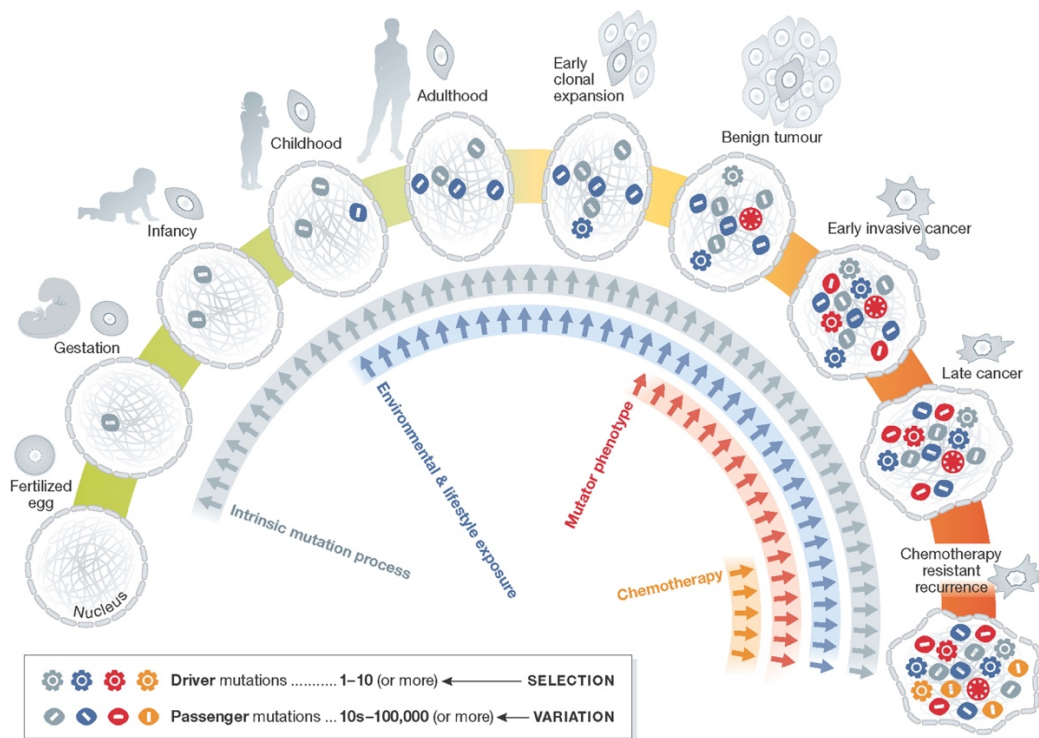
## 1.1 El cáncer es una enfermedad del genoma

El cáncer es un conjunto de enfermedades caracterizado por el crecimiento anormal de células, las cuales se dividen incontroladamente y son capaces de invadir tejidos adyacentes (Weinberg 2013). En México, los tumores malignos representan la tercera causa de muerte en la población en general después de las enfermedades del corazón y la diabetes mellitus (INEGI 2018), mientras que en el mundo ocupan el segundo lugar después de las enfermedades cardiovasculares (Roth et al. 2018). Es por esto que identificar las causas de los tumores y por lo tanto mejorar las herramientas para su prevención se ha vuelto una prioridad de los sistemas de salud a nivel internacional.

El cáncer es una enfermedad del genoma. Es una teoría bien establecida que la acumulación progresiva de mutaciones en una célula puede llevar a la adquisición de características como el crecimiento incontrolado y la colonización de tejidos tanto adyacentes como lejanos (Weinberg 2013). Hay dos mecanismos principales por los cuales puede suceder esta acumulación de mutaciones: Mecanismos causados por agentes exógenos, como lo son la exposición a la radiación ultravioleta y el fumar, y mecanismos endógenos, como lo son errores introducidos por las polimerasas a la hora de replicar el ADN y la presencia de variantes en línea germinal que disminuyen la capacidad de la maquinaria de reparación de ADN (**Figura 1**).

Las variantes en línea germinal pueden conferir un rango de distintos riesgos a padecer cáncer (desde bajo hasta elevado), que generalmente correlaciona de manera inversa con la frecuencia de estas variantes en la población (**Figura 2**). Por ejemplo, las variantes que confieren un riesgo pequeño pueden ser comunes en la población en general y se identifican por medio de estudios de asociación de genoma completo (GWAS, por sus siglas en inglés) (Manolio et al. 2009). La identificación de estas variantes es importante para complementar nuestro conocimiento de las causas de esta enfermedad pero no son al momento clínicamente accionables, al menos en su gran mayoría (Struck, Mannakee, and Gutenkunst 2018). Por otro lado, las variantes genéticas

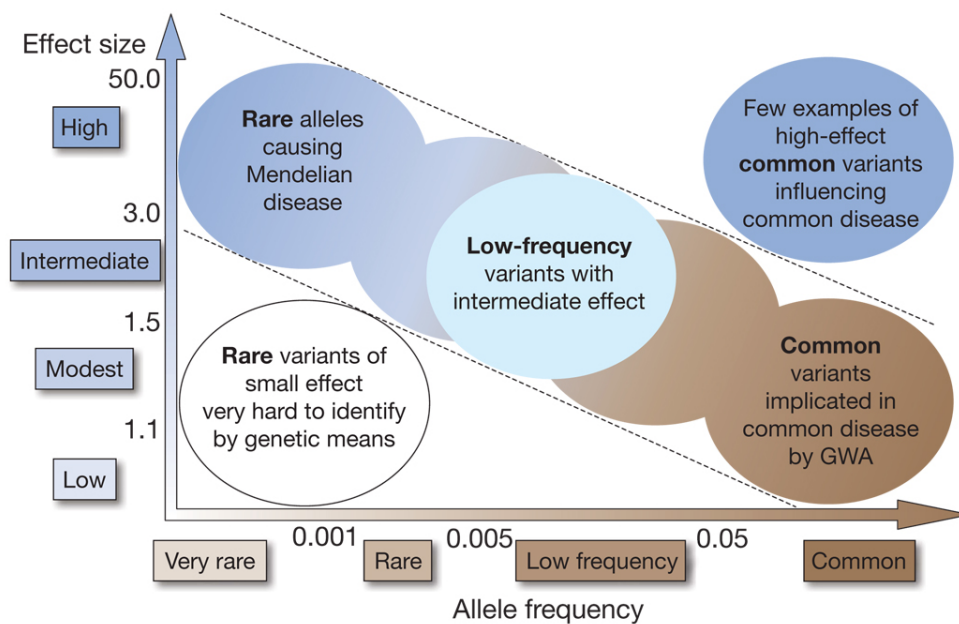
que confieren alto riesgo a desarrollar ciertos tipos de cáncer, como las localizadas en los genes *BRCA1* y *BRCA2* que aumentan el riesgo a desarrollar cáncer de mama, de ovario y de próstata (Petrucelli, Daly, and Pal 1993), son infrecuentes en la población en general, y pueden ser detectadas por métodos de análisis de ligamiento o secuenciación de exomas o genomas (Bamshad et al. 2011). Identificar y asignar un grado de patogenicidad a estas variantes es invaluable desde el punto de vista clínico ya que este conocimiento es utilizado para brindar consejo genético a familias portadoras y en algunos casos para guiar el tratamiento (Elliott and Friedman 2018; Richards et al. 2015; Tung and Garber 2018).



**Figura 1**  
**Acumulación de mutaciones debido a procesos endógenos y exógenos a lo largo de la vida de una célula.** Los procesos endógenos son representados con flechas grises, mientras que los exógenos con flechas azules. La exposición de las células a éstos resulta en una acumulación de mutaciones, ya sea conductoras (que contribuyen al crecimiento descontrolado) o pasajeras, y que puede llevar a la transformación neoplásica. Figura reproducida de (Stratton 2013) por medio de una licencia de atribución Creative Commons CC-BY.

Sin embargo, asignar un grado de patogenicidad a las nuevas variantes detectadas tras haber llevado a cabo un estudio de secuenciación de exomas o genes completos es bastante complicado, ya que a menudo se identifica un número considerable de éstas y nuestro

conocimiento biológico sobre la gran mayoría de los genes es aún insuficiente para predecir sus posibles consecuencias. Tomando lo anterior en cuenta, el objetivo del trabajo de investigación descrito en esta tesis tuvo por objetivo crear una nueva herramienta bioinformática (llamada **VCF/Plotein**) para contribuir en esta área, y aplicarla a un número considerable de variantes encontradas en **POT1**, un gen involucrado en alto riesgo a desarrollar **melanoma**, el tipo de cáncer de piel más agresivo. En este capítulo, iniciamos describiendo información pertinente a las variantes genéticas y los métodos establecidos para analizarlas, seguido de una introducción al tipo de cáncer de nuestro interés, melanoma. Los capítulos siguientes describen el desarrollo de VCF/Plotein (**Capítulo 2**), su aplicación a un conjunto de datos biológico para identificar variantes causales de melanoma (**Capítulo 3**), y finalmente una conclusión y consideración de estudios futuros (**Capítulo 4**).



**Figura 2**  
**La frecuencia alélica de las variantes genéticas en una población correlaciona inversamente con el tamaño del efecto sobre el fenotipo.** Figura reimpressa con permiso de Springer Nature, *Nature* (Manolio et al. 2009), todos los derechos reservados.

## 1.2 Identificación y análisis de variantes genéticas

Para identificar variantes genéticas en individuos de interés, se puede realizar una secuenciación de nucleótidos, la cual puede ser realizada por medio de distintas tecnologías como la secuenciación por capilar, la pirosecuenciación o la secuenciación por síntesis utilizada por las

máquinas Illumina (Mardis 2017). Una vez realizada esta secuenciación, y después de aplicar filtros de calidad extensivos, es necesario realizar una alineación de las lecturas resultantes al genoma de referencia por medio de algoritmos de mapeo como BWA (Li and Durbin 2009) o Bowtie2 (Langmead and Salzberg 2012; Bao et al. 2014). Finalmente se debe correr un algoritmo identificador de variantes, como puede ser Samtools mpileup (Li et al. 2009) o GATK HaplotypeCaller (McKenna et al. 2010).

### 1.2.1 Tipos de variantes genéticas

Las variantes genéticas pueden ser clasificadas en distintos tipos, dependiendo de sus características de secuencia:

*Variantes de un sólo nucleótido (SNVs, por sus siglas en inglés, de “single nucleotide variants”).* También son conocidas como sustituciones de una sola base, y son el tipo más simple de variación genética. Pueden ser sub-categorizadas en transiciones y transversiones. Si un SNV es común en la población, entonces se le llama polimorfismo de nucleótido sencillo (SNP, por sus siglas en inglés, “single nucleotide polymorphism”).

*Variantes de nucleótidos múltiples (MNVs, por sus siglas en inglés, de “multi-nucleotide variants”).* Estas variantes involucran el cambio consecutivo de dos o más bases. Al igual que los SNVs, si estos se encuentran en frecuencias elevadas en la población en general, entonces se les puede llamar polimorfismos de nucleótidos múltiples (MNVs, por sus siglas en inglés, “multi-nucleotide polymorphism”).

*Indeles (conjunción de las palabras en inglés “INsertions” y “DELetions”).* Este tipo de variante involucra la ganancia o pérdida de una o más bases en una secuencia. Usualmente, a lo que se le llama “indel” tiende a contener un número reducido de bases.

*Variantes estructurales.* Estas variantes involucran segmentos más extensos del genoma. Incluyen las inversiones, que es cuando una sección del genoma se invierte con respecto a su



posición original, y cambios de número de copias los cuales pueden ser amplificaciones, cuando una parte del genoma se duplica una o más veces, o deleciones grandes, en los que grandes secciones del genoma se pierden. No existe una regla específica entre un indel y una variante estructural, pero en general, se les llama variantes estructurales a aquellos cambios que involucran a más de una kilobase de secuencia.

La mayoría de los algoritmos identificadores de variantes pueden identificar SNVs, pero sólo algunos reportan indeles o variantes estructurales (Xu 2018). Esto se debe a que los métodos para encontrar estos distintos tipos de variación son conceptualmente muy diferentes: La identificación de SNVs, MNVs e indeles cortos requieren comparar un grupo de lecturas que mapean al mismo lugar contra el genoma de referencia, mientras que la identificación de indeles más largos y variantes estructurales requiere detectar aquellas lecturas pareadas que caen fuera de la distribución de tamaños del inserto y que tienen direcciones de alineamiento no concurrentes (Kosugi et al. 2019).

Los algoritmos de identificación de variantes usualmente generan un archivo de formato de variantes (VCF, por sus siglas en inglés, “variant call format”). Estos son archivos de texto simple que contienen información de los genotipos de todas las muestras secuenciadas en el proyecto. Un archivo VCF está organizado como una matriz, con la información de los cromosomas y las coordenadas de las variantes en renglones y la información de estas variantes y de las muestras en columnas (**Figura 3**) (Danecek et al. 2011). La información a nivel de muestras contiene el genotipo identificado por el algoritmo y otras métricas de interés como son (dependiendo del programa), las verosimilitudes de los genotipos y la cobertura evidencia para cada alelo del genotipo, entre otros. Para cada posición, el archivo también contiene información arrojada por el algoritmo como los alelos de referencia y alternativo y el puntaje de calidad escalado por Phred (Ewing and Green 1998). De manera importante, este archivo también contiene una columna llamada “FILTER”, en donde información sobre otros filtros de calidad aplicados posteriormente puede ser anotada.

(a) VCF example

```

##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29
1 1 2 . C T,CT . PASS H2;AA=T GT 0|1 2/2
1 1 5 rs12 A G 67 PASS . GT:DP 1|0:16 2/2:20
X 100 . T <DEL> . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36

```

**Figura 3**

**Estructura de un archivo VCF.** Al inicio del archivo, encontramos el encabezado, o “header”, que contiene información sobre el procesamiento por el que han pasado los datos. Cada línea del cuerpo, o “body”, contiene información sobre una variante genética encontrada en la población bajo estudio. Reimpreso bajo permiso de Oxford University Press, *Bioinformatics*, (Danecek et al. 2011), todos los derechos reservados.

### 1.2.2 Anotación de variantes genéticas

La anotación de variantes puede ayudar a los investigadores a filtrar y priorizar variantes con potenciales consecuencias biológicas para que puedan ser estudiadas funcionalmente. La primera anotación que se lleva a cabo generalmente es por calidad, y pueden aplicarse distintos filtros como son:

**Filtro por sesgo de cadena.** Ciertos eventos dentro del proceso de secuenciación podrían causar que una base incorrecta sea identificada en lecturas con una dirección específica, mientras que las lecturas en la otra dirección no la muestran. Este error de secuenciación puede ser identificado por medio de un filtro que aplica una prueba de Fisher comparando el número de lecturas en ambas direcciones que muestran la variante, y si el  $p$ -valor resultante es menor al establecido en un umbral entonces la variante se descarta.

**Filtro por distancia y distancia al final de la lectura.** Estos filtros fueron inicialmente desarrollados para lidiar con errores de secuenciación en datos de RNA cuando son alineados al genoma de referencia (Danecek et al. 2012). Si la evidencia de la existencia de una variante está mayormente basada en la última porción de la mayoría de las lecturas entonces podría ser identificada como un falso positivo resultante de la alineación de la última parte de un RNA mensajero (mRNA) a

un intrón aledaño a un sitio de empalme. El filtro de distancia al final de una lectura realiza una prueba  $t$  para determinar si la variante ocurrió a una distancia fija del final de las lecturas, por lo que podría ser marcada como un falso positivo.

*Filtro de distancia a indels u otros SNPs.* Estos filtros fueron diseñados para marcar variantes que están muy cercanas a un indel o a otro SNP, ya que podrían ser resultado de un error de alineamiento. Por ejemplo, un valor de 10 para este filtro querría decir que los SNPs que están a 10 pares de bases o menos de un indel, o los grupos de SNPs que se encuentren a menos de 10 pares de bases de distancia entre ellos, serían eliminados del conjunto de variantes de alta calidad.

Estos y otros filtros post-identificación de variantes pueden ser aplicados a un archivo VCF. Programas como bcftools (Li 2011) o GATK VariantFiltration (McKenna et al. 2010) pueden ser utilizados para este propósito, y sus parámetros pueden ser ajustados para adecuarse a las necesidades de los investigadores. Generalmente, después de que estos algoritmos son aplicados a un VCF, el campo 'FILTER' de este archivo para cada variante será anotado con la etiqueta 'PASS' si ésta pasó todos los filtros de calidad, o especificará el o los filtros que haya fallado. Usualmente los análisis consecuentes se realizan solamente en las variantes que han pasado todos los filtros.

Generalmente, el segundo nivel de anotación que se lleva a cabo en un conjunto de variantes genéticas es a nivel de predicción funcional. Varias herramientas para este propósito han sido desarrolladas, algunas de éstas basadas en datos públicos y por tanto basadas en variantes ya conocidas, mientras que otras han sido desarrolladas para la anotación de nuevos SNPs.

La predicción funcional de variantes puede ser realizada por medio de distintos acercamientos, desde los análisis basados en secuencia hasta el posible impacto estructural de los cambios en las proteínas. La predicción de las consecuencias de las variantes genéticas pueden ser inferidas por medio de herramientas como Ensembl VEP (McLaren et al. 2016) y SnpEff (Cingolani et al. 2012). Además de este nivel de anotación (que indica, por ejemplo, si una variante es de cambio

de aminoácido, de ganancia de codón de paro, de cambio de marco de lectura, etc.), estas herramientas permiten realizar anotaciones al nivel de frecuencia alélica en bases de datos públicas como son GnomAD (Lek et al. 2016) y el Proyecto de los 1000 Genomas (Auton et al. 2015), si la variante se ha visto con anterioridad en otros proyectos (dbSNP, (Sherry et al. 2001)) o, por ejemplo, si se ha encontrado mutada somáticamente en cáncer (COSMIC, (Tate et al. 2019)), el nivel de conservación evolutiva de la región donde se encuentra la variante (GERP (Davydov et al. 2010)), PolyPhen-2 (Adzhubei, Jordan, and Sunyaev 2013)), o si se ha encontrado que tiene relevancia clínica (anotaciones de ClinVar (Landrum et al. 2018)). La predicción funcional de estas variantes también puede ser complementada con anotaciones a nivel de regiones genómicas, por ejemplo, si la variante cae en sitios de pegado a factores de transcripción o sitios que son blancos de microRNAs, entre otros. Todas estas anotaciones ayudan a los investigadores a enfocarse en aquellas variantes que tienen más probabilidad a estar asociadas a su fenotipo de interés.

Sin embargo, todos estos pasos para identificar variantes candidatas son sólo parte del conjunto de consideraciones que debe tener un investigador antes de alcanzar conclusiones acerca de su posible patogenicidad. Aunque las variantes sean reales y parezca que tienen un efecto en la función de los genes y/o las proteínas, esto por sí solo no es evidencia suficiente para ligar a éstas causalmente con un fenotipo (MacArthur et al. 2014). Investigadores y bioinformáticos deben tener cautela al considerar cualquier asociación positiva, y deben tomar en cuenta hipótesis alternativas antes de reportar sus resultados (Minikel and MacArthur 2016; Verhagen et al. 2018).

### 1.2.3 Efectos de las variantes genéticas

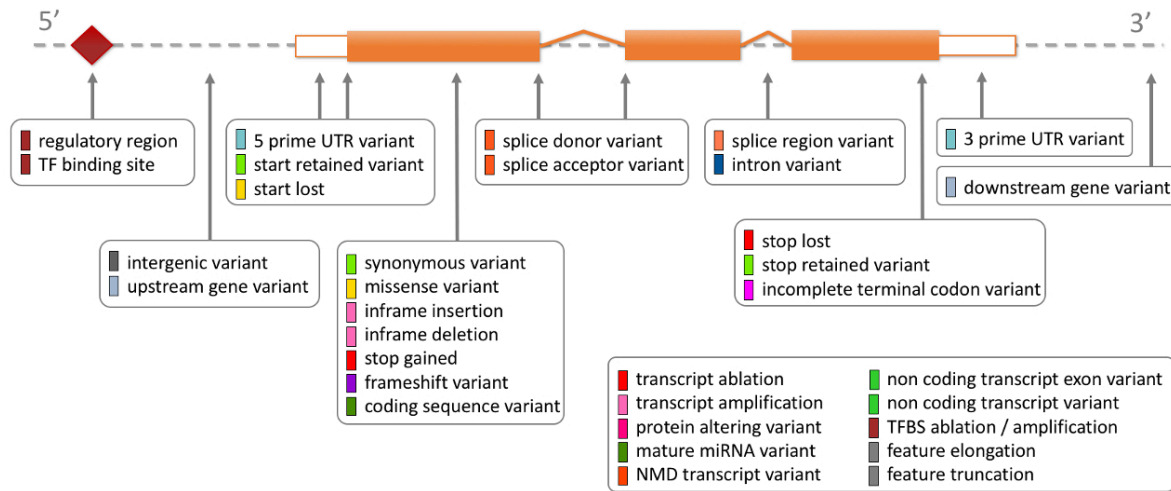
Las variantes genéticas pueden tener distintos efectos funcionales dependiendo de la región en la que se encuentren, la conservación evolutiva de ésta y el efecto de los genes cercanos a ellas. Una de las clasificaciones más utilizadas para estos efectos es la realizada por el proyecto Ensembl a través de su herramienta VEP (McLaren et al. 2016), la cual considera 36 distintas

categorías (Ensembl versión 99, enero de 2020) (**Figura 4**). Las más comunes se encuentran enlistadas en la **Tabla 1**:

**Tabla 1**

**Clasificación de los efectos de las variantes genéticas.** Se muestran la consecuencia en la cadena de aminoácidos, su descripción y el impacto en la función para las variantes más comunes encontradas por estudios masivos de secuenciación.

Consecuencia	Descripción	Impacto
Variante en receptor de sitio de corte y empalme	Una variante en sitio de corte y empalme que cambia la región invariante en el extremo 3' de un intrón ( <i>Splice acceptor variant</i> )	Alto
Variante en donador de sitio de corte y empalme	Una variante en sitio de corte y empalme que cambia la región invariante en el extremo 5' de un intrón ( <i>Splice donor variant</i> )	Alto
Ganancia de codón de paro	Una variante en la que al menos una base de un codón resulta cambiada, resultando en la ganancia de un codón de paro prematuro ( <i>Stop gained</i> )	Alto
Cambio de marco de lectura	Una variante que causa una disrupción del marco de lectura original, ya que el número de nucleótidos insertados o deletados no es múltiplo de tres ( <i>Frameshift variant</i> )	Alto
Cambio de aminoácido	Una variante que cambia una o más bases resultado en una secuencia de aminoácidos distinta pero en la que la longitud del transcrito es preservada ( <i>Missense variant</i> )	Moderado
Inserción sin cambio en el marco de lectura	Una variante no sinónima que inserta bases dentro de secuencia codificante, pero que no resulta en cambio de marco de lectura ya que el número de bases es múltiplo de tres ( <i>Inframe insertion</i> )	Moderado
Delección sin cambio en el marco de lectura	Una variante no sinónima que elimina bases dentro de secuencia codificante, pero que no resulta en cambio de marco de lectura ya que el número de bases es múltiplo de tres ( <i>Inframe deletion</i> )	Moderado
Variante en sitio de corte y empalme	Una variante en la que un cambio ocurre en la región de corte y empalme a 1-3 bases del exón o 3-8 bases del intrón ( <i>Splice region variant</i> )	Bajo
Variante sinónima	Una variante en región codificante en la que no existe un cambio en la secuencia de aminoácidos ( <i>Synonymous variant</i> )	Bajo
Variante en sitio de intrón	Una variante génica que ocurre dentro de un intrón ( <i>Intron variant</i> )	Modificador



**Figura 4**  
**Representación gráfica de los distintos tipos de variantes de acuerdo a la clasificación hecha por Ensembl.** En la parte superior se representa la estructura de un gen. Las cajas sólidas color anaranjado representan exones codificantes, las cajas vacías representan exones no codificantes, y la línea punteada representa el genoma. El diamante rojo representa una región regulativa río arriba del gen, e.g. un promotor. Figura reproducida con permiso del Proyecto Ensembl (Yates et al. 2020)

Como lo hemos discutido anteriormente, estas anotaciones pueden ser utilizadas para priorizar variantes genéticas con el propósito de identificar aquellas con mayor probabilidad de estar asociadas al fenotipo de interés.

### 1.2.4 Priorización de variantes genéticas

Es común que los investigadores asuman que una variante rara (generalmente definida como aquellas con una frecuencia alélica menor al 1% en la población de interés (Jeng et al. 2016)) que cae dentro de un gen con una función biológica relevante debe ser causal de su fenotipo de interés (MacArthur et al. 2014). Sin embargo, esto es casi siempre falso: Dependiendo de la ancestría genética, se estima que los humanos pueden cargar hasta 20,000 variantes ‘singletons’ (definidas como aquellas que se observan en un solo individuo en una población) (Auton et al. 2015), y que pueden cargar más de 50 variantes genéticas clasificadas como causales de enfermedades (Lek et al. 2016). En un claro ejemplo ilustrativo, Goldstein y colaboradores analizaron secuencia de ADN de un individuo control y reportaron haber encontrado variantes genéticas de baja frecuencia alélica dentro de regiones codificantes altamente conservadas evolutivamente, las cuales también estaban asociadas a consecuencias biológicas de alto impacto y en genes que han sido previamente asociados a fenotipos específico (Goldstein et al. 2013).

Aún habiendo llenado todos estos criterios, es evidente que estas variantes no están asociadas a un fenotipo ya que fueron encontradas en un individuo control. De esta manera, en su estudio, Goldstein *et al* llaman a la tendencia a asumir la causalidad de estas variantes el ‘potencial narrativo’, el cual es desafortunadamente bastante común en la literatura (Bell *et al.* 2011).

Dado el enorme número de variantes genéticas identificadas en estudios NGS (Aproximadamente 12,000 en exomas y 5 millones en genomas) (Pabinger *et al.* 2014), el filtrado y el post-procesamiento de éstas para elegir variantes candidatas podría ser una de las tareas más intensivas dentro de todo el flujo de trabajo. Dependiendo de la pregunta biológica original, es posible que el investigador deba refinar estos parámetros para poder brindar una respuesta acertada. Por ejemplo, si están buscando variación de frecuencia alélica baja que podría elevar el riesgo a padecer algún fenotipo en familias afectadas, entonces es posible que sea necesario relajar los filtros de calidad para no perder ninguna variante candidata, bajo el entendimiento de que éstas deberán ser confirmadas por medio de otro método ortogonal como la secuenciación por capilar. Si en cambio el objetivo del proyecto de investigación es identificar variantes en una cohorte grande, entonces se requerirá fijar umbrales de calidad mucho más estrictos.

#### 1.2.5 Interpretación de los efectos de variantes genéticas

Después de que todos los pasos anteriores han sido llevados a cabo (identificación, filtrado, anotación y priorización de variantes genéticas), un investigador necesita después considerar la cantidad de evidencia disponible para asignar causalidad a sus variantes de interés. La primera línea de evidencia debe ser estadística: Asumiendo que existen variantes candidatas, la primera pregunta a examinar debería ser, ¿Qué tan probable sería obtener resultados equivalentes a los obtenidos al azar, si cualquier otro gen fuera considerado? Por ejemplo, un estudio publicado en 2007 por Chiu y colaboradores asumió que dos variantes genéticas que habían identificado en el gen *CARD3* eran causales de cardiomiopatía hipertrófica familiar (Chiu *et al.* 2007). La manera en la que llegaron a esta conclusión se basó en un conjunto de cuatro puntos:

1. Que la variante haya sido detectada en otros pacientes con cardiomiopatía,
2. Que estuviera ausente de una cohorte de 200 individuos control,
3. Que el sitio estuviera altamente conservado evolutivamente a través de las especies y
4. Que co-segregara con el fenotipo en familias afectadas.

Sin embargo, la probabilidad de que todos estos criterios fueran cumplidos de manera azarosa si otros genes hubieran sido considerados es elevada, como fue demostrado por un estudio subsiguiente que estudio un panel génico más extenso (Verhagen et al. 2018). De manera adicional, evidencia tanto a favor como en contra de la hipótesis bajo consideración debe ser evaluada cuidadosamente. Por ejemplo, en el mismo estudio de cardiomiopatía, algunas de las variantes predichas por algoritmo bioinformáticos como potencialmente causales no co-segregaron con el fenotipo en familias afectadas (Verhagen et al. 2018). La disponibilidad de números cada vez mayores de genomas y exomas secuenciados en cohortes grandes como gnomAD sin duda ayudará a establecer la causalidad de estas variantes ya que cada vez se refinan más los estimados de frecuencias alélicas poblacionales (Lek et al. 2016). De la misma manera, es de suma importancia que las frecuencias alélicas sean analizadas en la población específica bajo estudio, ya que es sabido que estas pueden variar de manera dramática entre poblaciones (Auton et al. 2015).

Otro criterio importante, recomendado por MacArthur y colaboradores (MacArthur et al. 2014) argumenta que cuando se analizan enfermedades de origen posiblemente monogénico, se recomienda analizar en primera instancia los genes que hayan sido ligados anteriormente con fenotipos similares. Si se procede a analizar otros genes distintos, entonces se requiere identificar a varios individuos independientes que presenten síntomas clínicos similares. Adicionalmente, es deseable que se analice la distribución de variantes genéticas en el gen de interés en una población relevante. Por ejemplo, si se ha identificado una nueva variante de cambio de marco de lectura en un gen candidato, una buena práctica es analizar la frecuencia de variantes de impactos similares en catálogos de variación genética poblacional.



Finalmente, es posible que la evidencia estadística y múltiples análisis computacionales sugieran que una variante es causante del fenotipo bajo estudio. Si este es el caso, es deseable que se realicen estudios biológicos funcionales que sugieran si éste es el caso, ya sea utilizando tejido derivado de los pacientes afectados, usando líneas celulares u organismos modelo. De esta manera, la visión brindada por el conjunto de estudios funcionales, estadísticos y computacionales podría ser suficiente para reportar una potencial asociación causal entre una o varias variantes genéticas y un fenotipo. Cuando esto se lleve a cabo, se recomienda que se reporte una descripción detallada de toda la evidencia disponible, de manera clara, y que aquellas asociaciones inciertas sean igualmente especificadas (MacArthur et al. 2014). A su vez, es buena práctica depositar los datos genéticos con los que se llegó a esta conclusión en un repositorio como dbGaP (Mailman et al. 2007) o EGA (Lappalainen et al. 2015), de manera que otros investigadores puedan analizarla y contribuir al conocimiento científico sobre las causas del fenotipo de interés.

En la presente tesis, se creó un algoritmo interactivo y fácil de usar para auxiliar a grupos de investigación que estén buscando establecer la causalidad de las variantes genéticas de estudios de secuenciación de exomas. En este algoritmo, se implementaron pasos para facilitar el filtrado, la anotación y la priorización de variantes. Consecuentemente, lo aplicamos a un estudio de secuenciación del gen *POT1*, el cual confiere alto riesgo a desarrollar melanoma, para identificar aquellas variantes candidatas a causar una consecuencia funcional dentro de este gen. Finalmente, y tomando en consideración lo discutido anteriormente, llevamos a cabo experimentos funcionales para determinar la posible causalidad de las variantes descubiertas.

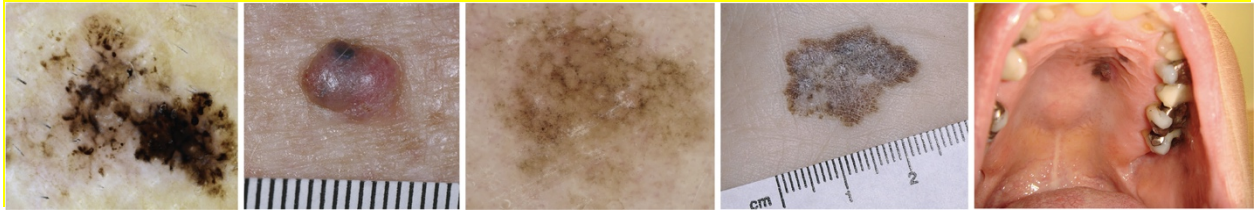
### 1.3 El melanoma

El melanoma es un cáncer de los melanocitos, células de origen neuroectodérmico productoras de pigmento que se localizan principalmente en la lámina basal de la epidermis. Aunque constituye solamente el 2% de los casos de cáncer dermatológico, se estima que es causante de más del 75% de las muertes por cáncer en este tejido (A. J. Miller and Mihm 2006). Además, se conoce que es altamente agresivo y resistente al tratamiento terapéutico cuando se detecta en

etapas avanzadas, y que sólo aproximadamente un 15% de los pacientes presenta una supervivencia mayor a 5 años cuando existe metástasis (A. J. Miller and Mihm 2006) La incidencia de este cáncer ha aumentado drásticamente en las últimas cuatro décadas y se ha convertido en un importante problema de salud pública en algunos países como Reino Unido, Estados Unidos y Australia (Whiteman, Green, and Olsen 2016). Sin embargo, este no es un problema único de estos países, ya que estudios recientes indican que también ha habido un aumento en la incidencia en países como Brasil y Colombia (Erdmann et al. 2013) y que en México, su incidencia en las últimas décadas ha aumentado hasta 500% (Herrera González and Aco Flores 2010). Esto ha incrementado la importancia de conocer los mecanismos epidemiológicos, fisiopatológicos y los factores de riesgo que predisponen al desarrollo de melanoma.

Desde la perspectiva clínica, históricamente el melanoma se ha clasificado de acuerdo con las características anatómicas e histológicas de las lesiones en cinco subtipos: melanoma de extensión superficial, melanoma acral lentiginoso, melanoma nodular, lentigo maligna melanoma, y melanoma mucoso (**Figura 5**). Dicha clasificación tiene una correlación epidemiológica relacionada con el origen étnico y el color de la piel, ya que el melanoma de extensión superficial se presenta principalmente en pacientes de ascendencia europea y está altamente relacionado con la exposición intermitente y prolongada a la luz ultravioleta, mientras que el melanoma acral lentiginoso es uno de los subtipos más comunes en el resto de las poblaciones y no se relaciona con la radiación solar (Scolyer, Long, and Thompson 2011; Rabbie et al. 2019). Asimismo, los casos de melanoma nodular que se caracterizan por ser altamente metastásicos, se distribuyen de manera más homogénea en las diferentes poblaciones, a excepción de en algunos países africanos como Nigeria en donde representan hasta el 78% de los diagnósticos de melanoma (Atanda and Umar 2012). Finalmente, el lentigo maligna melanoma se presenta principalmente en personas de edad avanzada y de manera equilibrada entre las poblaciones, mientras que el melanoma mucoso que es raro en las poblaciones de ascendencia europea, es común en algunos países asiáticos (Xiong, Charifa, and Chen 2020; Fujisawa et al. 2019; Guo et al. 2015). Por lo tanto, la distribución de casos pertenecientes a cada

subtipo del total de diagnósticos de melanoma varía enormemente en distintos países alrededor del mundo (Ossio et al. 2017) (Figura 6).



**Figura 5**

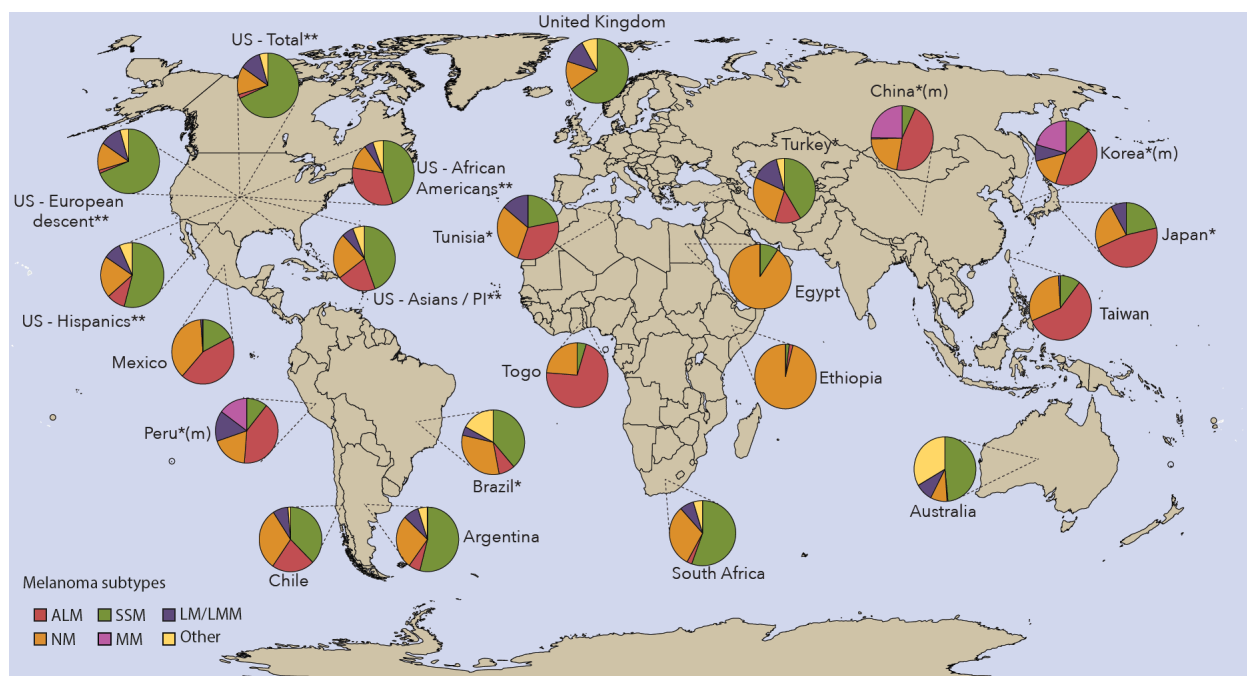
**Subtipos histopatológicos más importantes de melanoma.** De izquierda a derecha, melanoma de extensión superficial, el cual surge principalmente en pacientes más jóvenes y zonas de la piel intermitentemente expuestas al sol; melanoma nodular, el cual surge como un nódulo que rápidamente presenta un crecimiento vertical e invasivo; melanoma lentigo maligna, el cual surge en pacientes de edad avanzada y en sitios crónicamente expuestos al sol; melanoma lentiginoso acral, el cual surge en las palmas de las manos, las plantas de los pies y zonas subungueales y no está asociado a la radiación ultravioleta; y melanoma mucoso, un subtipo de baja frecuencia que se presenta en regiones mucosas y presenta características muy distintas a los melanomas cutáneos. Paneles 1 y 3 reproducidos con permiso de Elsevier, *Dermatol Clin* (Longo and Pellacani 2016), paneles 2 y 4 reproducidos de *Mol Oncol* (Scolyer, Long, and Thompson 2011) a través de una licencia CC-BY, y panel 5 reproducido con permiso de Elsevier, *JAAD* (Tacastacas et al. 2014).

### 1.3.1 Factores de riesgo

Como la mayoría de los tipos de cáncer, el melanoma tiene agentes causales tanto ambientales como genéticos. Se ha establecido un modelo en el que el efecto de las variantes genéticas que confieren riesgos altos o moderados interactúan con los efectos ambientales, modificando el efecto de estos últimos (Cust, Mishra, and Berwick 2018). En esta sección se discuten estos factores y los estudios relevantes que han llevado a su identificación.

#### 1.3.1.1 Factores de riesgo ambientales

El factor de riesgo más importante para el desarrollo de melanoma es sin duda alguna la exposición a la radiación ultravioleta (UV) (Lo and Fisher 2014). De hecho, se estima que el dramático cambio que se ha observado en la incidencia de melanoma de algunas décadas al presente en un gran número de países alrededor del mundo se debe a cambios en el comportamiento de los individuos respecto a los hábitos de exposición al sol (Parkin, Mesher, and Sasieni 2011). Esto último es debido en gran medida a la popularidad y accesibilidad de viajar a regiones soleadas y de utilizar camas solares.



**Figura 6**  
**Distribución de subtipos comunes de melanoma en diferentes países alrededor del mundo.** ALM: melanoma lentiginoso acral; NM: melanoma nodular; SSM: melanoma de extensión superficial; LM/LMM: lentigo maligna/melanoma lentigo maligna; MM: melanoma mucoso. Los países con asteriscos tienen una proporción de casos asignados a un subtipo desconocido: (\*) : menos de 25% de los casos asignados a un subtipo desconocido; (\*\*) : entre 25% y 50% de los casos asignados a un subtipo desconocido. Diferencias metodológicas entre países podrían significar que las proporciones mostradas no son representativas y por tanto que éstas no son comparables. Figura reproducida con autorización de Springer, *Nature Reviews Cancer* (Ossio et al. 2017).

Sin embargo, la cantidad y modalidad de la exposición a la luz UV es importante en la determinación del riesgo. Por ejemplo, un meta-análisis que incluyó 57 estudios y un total de 38,671 pacientes identificó que la exposición intermitente e intensa a la luz UV significativamente aumenta el riesgo a padecer la enfermedad en la edad adulta comparada con la exposición crónica (Gandini et al. 2005). Esto, aunado a otros estudios, llevó en 2009 a la Agencia Internacional sobre la Investigación en Cáncer (IARC), de la Organización Mundial de la Salud (WHO), a declarar a los instrumentos emisores de radiación UV como carcinogénicos para los humanos (Clase 1) (Ghissassi et al. 2009).

### 1.3.1.2 Factores de riesgo genéticos

Algunos estudios han demostrado que aproximadamente el 10% de los casos de melanoma cutáneo pueden atribuirse a factores hereditarios (Skolnick, Cannon-Albright, and Kamb 1994). Definitivamente, el gen que subyace la mayor parte del riesgo genético es *CDKN2A* (*cyclin-*

*dependent kinase inhibitor 2A*), el cual se encuentra mutado en ~25% de las familias con melanoma, dependiendo del lugar de residencia (Harland et al. 2014). De manera similar, *CDK4* (*cyclin-dependent kinase 4*), el cual interactúa con *CDKN2A* y de esta manera contribuye al control del ciclo celular, también se encuentra mutado en algunas familias con este tipo de cáncer (Zuo et al. 1996). Estos genes fueron descubiertos por estudios de ligamiento hace más de 20 años (Cannon-Albright et al. 1992; Zuo et al. 1996), y durante varios años no se realizó ningún avance significativo en la búsqueda de más genes que contribuyen a elevar el riesgo de melanoma en familias afectadas.

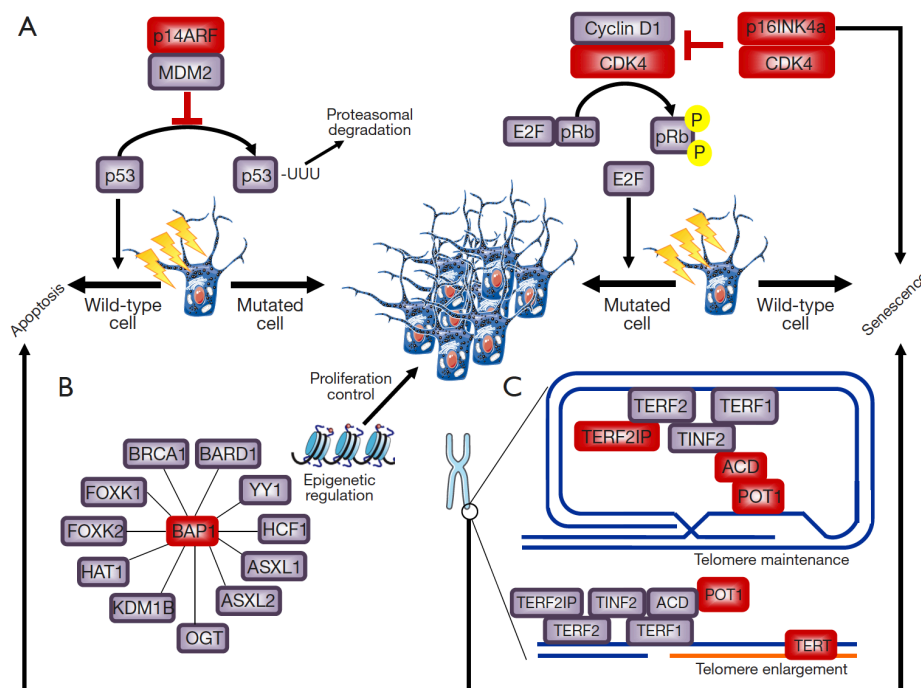
Sin embargo, en la última década, los grandes avances de las técnicas de secuenciación masiva permitieron el análisis de genomas y exomas completos de grandes números de pacientes y familias afectados. Es así que ha habido una nueva ola de genes descritos con una influencia en el riesgo a desarrollar melanoma en los últimos años, como son *BAP1* (*BRCA1-associated protein*) también relacionado con el control del ciclo celular (Wiesner et al. 2011), así como *TERT* (*telomerase reverse transcriptase*) (Horn et al. 2013), *POT1* (*protection of telomeres*) (Robles-Espinoza et al. 2014; Shi et al. 2014), *ACD* (*adrenocortical dysplasia homolog*) y *TERF2IP* (*telomeric repeat binding factor 2 interacting protein*) (Aoude et al. 2015), que codifican proteínas que participan en el mantenimiento adecuado de los extremos teloméricos. Como se ha mencionado anteriormente, el descubrimiento de estos últimos genes se ha realizado mediante la secuenciación de exomas de familias con alta penetrancia de la enfermedad, lo cual ha permitido la identificación de variantes de un sólo nucleótido, variantes estructurales, y deleciones e inserciones en las regiones codificantes del genoma a una resolución sin precedentes. Sin embargo, aún con estos grandes avances, la causa de alrededor del 50% de los casos de melanoma familiar continúa sin ser identificada, lo cual sigue siendo un importante foco de investigación (Rossi et al. 2019).

Los estudios GWAS llevados a cabo en grandes números de casos y controles también ha identificado regiones del genoma moderadamente asociadas a desarrollar melanoma. Uno de los últimos meta-análisis, que incluyó 15,990 casos y 26,409 controles, encontró veinte regiones de

importancia y reafirmó las asociaciones previas encontradas en los genes *melanocortin 1 receptor* (*MC1R*), *tyrosinase* (*TYR*), *TERT* y *CDKN2A* (Law et al. 2015). Existen muchos esfuerzos actuales que tienen como objetivo realizar el mapeo fino de estas regiones, esto es, identificar las variantes genéticas específicas dentro de ellas que subyacen el elevado riesgo a padecer la enfermedad.

### 1.3.1.2.1 Vías biológicas afectadas en el riesgo genético a desarrollar melanoma

Las vías biológicas que se ven alteradas por las variantes y los genes descubiertos hasta el momento como modificadores del riesgo a padecer melanoma afectan principalmente al control del ciclo celular y el mantenimiento de los telómeros (**Figura 7**) y al proceso de pigmentación.



**Figura 7**

**Funciones de los genes involucrados en el riesgo a melanoma y vías biológicas alteradas.** En rojo se señalan todas las proteínas codificadas por genes de alto riesgo. A) *CDKN2A* codifica para dos proteínas, p16INK4a y p14ARF, las cuales tienen funciones en el control del ciclo celular. B) BAP1 está involucrado en el control epigenético y regulación de la transcripción de una gran cantidad de proteínas, y su incorrecto funcionamiento puede llevar a una proliferación celular descontrolada. C) El complejo *shelterin*, formado por las seis proteínas TERF1, TERF2, TERF2IP, TINF2, ACD y POT1 protegen los telómeros contra el daño y controlan su longitud. Cuando estas proteínas están mutadas y no funcionan correctamente, los telómeros sufren una elongación progresiva y daño, lo que puede conducir al desarrollo de cáncer. Figura reproducida con permiso de Nancy International Ltd Subsidiary AME Publishing Company, de (Potrony et al. 2015), permiso obtenido a través de Copyright Clearance Center.

Ciclo celular. Los primeros genes descubiertos en el estudio de familias susceptibles a desarrollar melanoma fueron *CDKN2A* y *CDK4*, que codifican proteínas importantes en el control del ciclo celular (Cannon-Albright et al. 1992; Zuo et al. 1996). En particular, *CDKN2A* codifica para dos proteínas estructuralmente independientes, p14ARF e y p16INK4A (**Figura 7A**). La primera de éstas es capaz de inhibir a mouse double minute 2 homolog (*MDM2*), la cual a su vez inhibe a *tumour protein p53 (TP53)* (Potrony et al. 2015). De esta manera, esta interacción permite la activación de TP53, el cual tiene un papel importante en la respuesta al daño al ADN, el estrés oxidativo y la desregulación oncogénica, entre otros (D. Liu and Xu 2011). La segunda proteína, p16INK4A, es un inhibidor del complejo ciclina D1/CDK4, el cual promueve la progresión del ciclo celular y la proliferación (Potrony et al. 2015). La mayoría de las variantes genéticas encontradas en familias con riesgo elevado a desarrollar melanoma en estos genes inactivan al gen *CDKN2A*, ya sea a uno o ambos supresores de tumores, o activan a *CDK4*, impidiendo su inhibición por p16INK4A. BAP1 también es una proteína involucrada en control de ciclo celular, que se ha encontrado con variantes genéticas deletéreas en individuos con alto riesgo (**Figura 7B**).

Pigmentación. La región más significativa en el último gran meta-análisis de estudios GWAS de melanoma (Law et al. 2015) está asociada al gen *MC1R*. El producto de este gen participa en la producción de melanina, el pigmento que le da color a la piel y que tiene la capacidad de absorber el espectro más dañino de la radiación UV (Nasti and Timares 2015). Un gran número de estudios han caracterizado sus diferentes variantes funcionales, las cuales generalmente afectan de manera grave su función y por ende la generación de este pigmento (Dessinioti et al. 2011). Otro de los genes que se ha encontrado asociado a melanoma tanto por estudios GWAS como por experimentos funcionales es *TYR*, el cual tiene un papel crucial en la ruta biológica de la síntesis de melanina a partir del aminoácido tirosina (Iozumi et al. 1993). Aunque estos genes no conllevan un riesgo elevado, están asociados a un riesgo moderado y se consideran modificadores de los genes con penetrancia elevada (Box et al. 2001).

Mantenimiento de los telómeros. Una de las vías biológicas más recientes en ser asociadas al desarrollo de melanoma es aquella relacionada con la función y el mantenimiento de los

telómeros (Potrony et al. 2015). En la última década, estudios de secuenciación de exomas en familias con melanoma llevó a la identificación de los genes *TERT*, *POT1*, *TERF2IP* y *ACD* como causales de esta enfermedad (Horn et al. 2013; Robles-Espinoza et al. 2014; Shi et al. 2014; Aoude et al. 2015) (**Figura 7C**). Algunas de las variantes identificadas en estas familias están asociadas a una longitud telomérica elevada y a señales de daño a ADN constante en estas regiones (Robles-Espinoza et al. 2014; Shi et al. 2014). Se ha observado que la mayoría de las variantes identificadas desactivan la función de estos genes, aunque sus consecuencias biológicas aún están bajo estudio.

#### *1.3.1.3 POT1: Un gen de alta penetrancia para el melanoma*

Uno de los genes asociados al riesgo a desarrollar melanoma cutáneo y que fue identificado por estudios de secuenciación de exomas es *POT1*. En 2014, dos grupos de investigación independientes encontraron distintas variantes en este gen co-segregando con el fenotipo en familias con melanoma de distintas partes del mundo (Robles-Espinoza et al. 2014; Shi et al. 2014) (

**Tabla 2**). Estas variantes estaban asociadas a una longitud telomérica elevada y perturban el mantenimiento de los telómeros. Asimismo, estudios posteriores identificaron variantes en este gen asociadas a otros tipos de cáncer como el glioma (Bainbridge et al. 2015), linfoma (McMaster et al. 2018), cardioangiosarcoma (Calvete et al. 2015), y leucemia linfocítica crónica (CLL) (Speedy et al. 2016), en donde también se ha encontrado mutado somáticamente (Ramsay et al. 2013). Estudios *in vitro* de estas y otras variantes han demostrado la progresión constante de la longitud telomérica en células portadoras, lo que, junto con evidencia observada en familias, podría indicar que las variantes deletéreas en este gen conllevan a un nuevo síndrome de alargamiento progresivo de telómeros, el cual sería uno de los primeros síndromes de este tipo descritos en humanos (Stanley and Armanios 2015).

Dada la amplia evidencia existente sobre la implicación de *POT1* en un un gran número de tipos de cáncer, incluyendo el síndrome multi-tumores *Li-Fraumeni-like* (Calvete et al. 2017), se ha incluido en los paneles de pruebas genéticas con el propósito de brindar consejo genético a



individuos de familias con predisposición a estas enfermedades. Sin embargo, no todas las variantes encontradas en este y otros genes afectarán la función biológica, y es por eso que uno de los grandes problemas con los que se enfrenta este campo médico es la interpretación de variantes de significancia incierta (VUS, por sus siglas en inglés, “*variants of uncertain significance*” (Richards et al. 2015)) resultantes de experimentos de secuenciación. Es por eso que consideramos de gran importancia definir un catálogo de variantes específicas asociadas a desarrollar cáncer en los genes presentes en estos paneles. En esta tesis, escribimos un algoritmo computacional que facilite la tarea de priorizar las posibles variantes asociadas, y después realizamos una validación experimental para comprobar nuestras hipótesis.

**Tabla 2**

**Variantes genéticas en el gen *POT1* encontradas en familias propensas a desarrollar melanoma.** Estos y otros estudios subsecuentes han demostrado que estas variantes son causales. La secuencia de referencia de la proteína es la codificada por el transcrito con Ensembl ID ENST00000357628.

Posición genómica (GRCh37)	Consecuencia	Efecto en la proteína	Referencia
g.7:124503684 T>C	Cambio de aminoácido	p.Tyr89Cys	(Robles-Espinoza et al. 2014)
g.7:124465412 C>T	Variante en receptor de sitio de corte y empalme	-	(Robles-Espinoza et al. 2014)
g.7:124503670 G>C	Cambio de aminoácido	p.Gln94Glu	(Robles-Espinoza et al. 2014)
g.7:124493077 C>A	Cambio de aminoácido	p.Arg273Leu	(Robles-Espinoza et al. 2014)
g.7:124493086 C>T	Cambio de aminoácido	p.Ser270Asn	(Shi et al. 2014)
g.7:124464052 C>G	Cambio de aminoácido	p.Gln623His	(Shi et al. 2014)
g.7:124503540 C>T	Cambio de aminoácido	p.Arg137His	(Shi et al. 2014)
g.7:124499043 C>T	Cambio de aminoácido	p.Asp224Asn	(Shi et al. 2014)
g.7:124469308 C>G	Cambio de aminoácido	p.Ala532Pro	(Shi et al. 2014)

### 1.3.2 El genoma de los tumores de melanoma

Aunque no es el enfoque del estudio presentado en esta tesis, consideramos pertinente dedicar una breve sección a discutir el perfil somático de los tumores de melanoma, ya que éstos muestran la interacción de los efectos de las variantes genéticas y los factores ambientales de riesgo. Este perfil, el cual está formado por todas las alteraciones somáticas encontradas cuando se secuencia el ADN de tumores y se “restan” las variantes de línea germinal, puede mostrar no solamente aquellas mutaciones que contribuyeron al crecimiento del tumor sino también los

procesos mutacionales que le dieron origen (Alexandrov et al. 2013). De esta manera, el genoma tumoral puede considerarse como un “registro arqueológico”, el cual posee las huellas de todos los procesos que han estado activos durante la vida del tumor (Stratton 2013).

*Genes conductores de melanoma.* Diversos estudios genómicos en las últimas dos décadas han identificado los genes más comúnmente mutados en tumores de melanoma. Generalmente, estos son *BRAF*, los genes *RAS*, y *NF1* (Cancer Genome Atlas Network 2015). Aunque el resultado de la mutación de estos genes es el mismo, *i.e.*, la activación de la cascada de señalización de las cinasas activadas por mitógenos (MAPK por sus siglas en inglés, *mitogen-activated protein kinases*), dando como resultado un descontrolado crecimiento celular, los pacientes muestran características clínicas distintas dependiendo del gen que sus tumores tengan alterado. Por ejemplo, los pacientes con mutaciones activadoras en *BRAF* tienden a ser más jóvenes, aquéllos con mutaciones en los genes *RAS* tienden a tener una activación más pronunciada de la vía de señalización MAPK, y los pacientes con mutaciones en *NF1* tienden a tener edad más avanzada y una carga mutacional más elevada (Cancer Genome Atlas Network 2015). Asimismo, los tumores que no tienen mutaciones en ninguno de estos genes, llamados triple negativos o triple silvestres, son sujeto de investigación activa para identificar nuevos conductores.

*Firmas mutacionales.* Los análisis computacionales de firmas mutacionales, los cuales son recientes al haberse popularizado a partir de 2013 (Alexandrov et al. 2013), son capaces de extraer los patrones de alteraciones causadas por distintos mutágenos, así como caracterizar el nivel de exposición a éstos en el tumor. Este tipo de análisis ha logrado, por ejemplo, identificar los tipos de mutaciones causadas por el ácido aristolóquico y la influencia de las enzimas APOBEC en la evolución tumoral (Rosenquist and Grollman 2016; Petljak et al. 2019). El estudio de tumores de melanoma ha logrado la identificación de la firma mutacional de la radiación UV, dominada por transiciones C>T, en la mayoría de los melanomas cutáneos, así como la firma del agente quimioterapéutico temozolomida en pacientes tratados (Alexandrov et al. 2013). En tipos de melanoma con etiología desconocida, como el melanoma acral lentiginoso, este tipo de análisis podría identificar posibles agentes causales y por tanto es un área de investigación activa.

Potencialmente, el algoritmo descrito en esta tesis podría ser aplicado también al estudio de mutaciones somáticas, proveyendo una manera de visualizar a los genes potencialmente conductores y analizar los patrones mutacionales en genes candidatos.

## 1.4 Justificación del proyecto presentado en esta tesis

El algoritmo descrito en esta tesis busca facilitar la visualización y la interpretación de variantes provenientes de estudios masivos de secuenciación. Rutinariamente, y dependiendo de la escala del estudio, se pueden identificar cientos de miles a millones de mutaciones genéticas, dificultando la identificación de aquéllas verdaderamente asociadas con el fenotipo de interés.

Algunas de las razones que podrían explicar esta dificultad, y una que busca remediar el algoritmo presentado en esta tesis, es que en la mayoría de los estudios realizados no se han utilizado estrategias alternativas de análisis más allá de lo convencional (*e.g.*, identificar mutaciones que introducen codones de paro, que cambian el marco de lectura, que destruyen sitios esenciales de corte y empalme o que alteran aminoácidos altamente conservados). Algunos ejemplos de análisis alternativos son la búsqueda de variantes sinónimas considerando efectos más allá de la estructura protéica, o la búsqueda de variantes que generen sitios críticos de corte y empalme. Estas características han sido incluidas en **VCF/Plotein**, y pueden analizarse visual e interactivamente en conjunto con otros tipos de variantes.

Finalmente, aplicamos este algoritmo a las mutaciones descubiertas en el gen *POT1* en un estudio de secuenciación de 2,929 casos de melanoma y 3,298 controles, y lo utilizamos para identificar posibles mutaciones deletéreas que pudieran aumentar el riesgo a padecer este tipo de cáncer. Después, realizamos ensayos biológicos de medición de longitud telomérica en los pacientes portadores para probar las predicciones del algoritmo, y para aportar conocimiento que podría ser útil en el mejoramiento y la interpretación de resultados de paneles utilizados para brindar consejo genético.

## **2 VCF/Plotein: Creación de un algoritmo para el análisis de variantes genéticas provenientes de estudios de secuenciación**

### **2.1 Introducción**

Los grandes avances en las tecnologías de secuenciación, así como la importante disminución en el costo para realizar este tipo de experimentos a gran escala han permitido que cada vez más biólogos y médicos puedan utilizar estas herramientas para responder preguntas de investigación relacionadas con el genoma, tales como:

1. ¿Qué genes están relacionados con este fenotipo de interés?
2. ¿Qué variantes genéticas dentro de esos genes son causales, y cuáles carecen de efectos?
3. ¿Cómo contribuyen los genes involucrados para fomentar el desarrollo de cáncer?

De esta manera, la secuenciación de exomas ha tenido gran éxito en la identificación de variación genética que tiene influencia en el desarrollo de un gran número de fenotipos humanos, desde cambios en el ADN de línea germinal que subyacen riesgo elevado a padecer enfermedades mendelianas y rasgos complejos, hasta mutaciones somáticas que contribuyen al desarrollo de cáncer (Gilissen et al. 2011; Do, Kathiresan, and Abecasis 2012). Sin embargo, el proceso de identificar variantes y mutaciones causantes de enfermedades continúa siendo una tarea altamente compleja, y a menudo una que requiere al menos un mínimo de conocimientos bioinformáticos. Esto se debe en gran parte a la enorme cantidad de variantes que son identificadas rutinariamente en proyectos de secuenciación de exomas, la gran diversidad de mecanismos biológicos por el que estas variantes podrían estar actuando, y la necesidad de integrar grandes cantidades de información, como aquella proveniente de algoritmos que asignan patogenicidad a las variantes como SIFT y PolyPhen-2 (Ng and Henikoff 2003; Adzhubei,

Jordan, and Sunyaev 2013), y bases de datos clínicas y poblacionales (Landrum et al. 2018; Sherry et al. 2001; Lek et al. 2016).

En este contexto, se han desarrollado herramientas de software interactivo y fácil de utilizar que son capaces de filtrar, mostrar y contextualizar datos de secuenciación de exomas con el objetivo de acelerar el descubrimiento de variantes genéticas causantes de enfermedades. Estas herramientas varían en la cantidad de información de bases de datos externas que integran, su interactividad, y el nivel de experiencia en bioinformática que un usuario requiere para ejecutarlas. Por ejemplo, *Genome Mining* (GEMINI) (Paila et al. 2013) permite que el usuario explore interactivamente sus propios archivos de variación genética y sobrelapa información de dbSNP (Sherry et al. 2001), el proyecto ENCODE (Dunham et al. 2012), ClinVar (Landrum et al. 2018) y *The Kyoto Encyclopedia of Genes and Genomes* (KEGG, (Ogata et al. 1999)), pero requiere que los usuarios tengan un buen entendimiento de la línea de comando y que cuenten con el conocimiento para realizar consultas con el sistema MySQL. De la misma manera, los algoritmos VCF-Miner (Hart et al. 2016) y BrowseVCF (Salatino and Ramraj 2017) permiten que el usuario utilice filtros interactivos en su propio archivo VCF a través de una interface web, pero no consideran información externa de fuentes como dbSNP o COSMIC (Tate et al. 2019). Otras herramientas como BiERapp (Alemán et al. 2014) y exomeSuite (Maranhao et al. 2014) permiten aplicar una gran cantidad de filtros personalizados pero no cuentan con soporte para visualización de datos genómicos. Asimismo, se han desarrollado otras herramientas que se enfocan en visualización a nivel de secuencia de proteína en lugar de en filtrado de variantes, tales como ProteinPaint (Zhou et al. 2016), VizGVar (Solano-Román et al. 2018) y vcf.iobio (C. A. Miller et al. 2014). Estos recursos son altamente interactivos, pero tienen algunas deficiencias: Por ejemplo, VizGVar sólo puede mostrar información existente en la base de datos Ensembl sin permitir al usuario analizar sus propias variantes, vcf.iobio no permite visualizar información a nivel de gen, y ProteinPaint requiere que el usuario realice varios pasos bioinformáticos antes de poder visualizar sus datos.

Dado este panorama, decidimos que si nuestro objetivo es identificar variantes causales en un estudio de secuenciación de melanoma, lo primero a realizar debería ser una manera de filtrar de manera fácil y visual los resultados de estos grandes conjuntos de datos, que a menudo contienen miles o millones de variantes. Con esto en mente, decidimos desarrollar el algoritmo **VCF/Plotein**, una aplicación que permitiera no sólo a los investigadores, sino también a profesionales médicos e incluso a pacientes, visualizar, filtrar e interactuar con sus datos genómicos de forma fácil y segura. Como requisitos, este programa no debía requerir tener conocimientos de bioinformática para poder ser utilizado, debía preservar la seguridad de los datos, y debía ser capaz de analizar una gran cantidad de variantes genéticas rápidamente.

## 2.2 Métodos

### 2.2.1 Características del servidor y lenguajes de programación

Implementamos VCF/Plotein completamente como una aplicación de una sola página hospedada en un servidor con un procesador Intel Xeon E5-4627 v4 con 2 núcleos y 2.60 Ghz, el cual ejecuta una máquina virtual VMware versión 6.5.0 en un sistema operativo Linux Centos 7.5. El servidor se encuentra en el Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH), parte de la Universidad Nacional Autónoma de México (UNAM). Este servidor también cuenta con 4GB de memoria RAM y un disco duro de estado sólido con 1TB de espacio de almacenamiento. Escribimos la aplicación principalmente en el lenguaje de programación JavaScript utilizando la plataforma Nuxt.js, basada en Vue.js, para controlar el almacenamiento de la información, así como su flujo y presentación en el navegador.

### 2.2.2 Interfaz de programación de aplicaciones

Desarrollamos una interfaz de programación de aplicaciones (*application programming interface*, API, por sus siglas en inglés) para obtener información de bases de datos externas instaladas localmente en el servidor del LIIGH-UNAM. Las bases de datos incorporadas en VCF/Plotein son aquellas que consideramos más importantes para el filtrado y la priorización de variantes genéticas causales de fenotipos:

1. **gnomAD**, versión 2.1, tamaño: 59.23 GB (Lek et al. 2016)

2. **dbSNP**, versión 151, tamaño: 14.6 GB (Sherry et al. 2001)
3. **COSMIC**, versión 86, tamaño: 421.8 MB (Tate et al. 2019)
4. **ClinVar**, versión 86, tamaño: 170.7 MB (Landrum et al. 2018)
5. Relaciones de fenotipo de la base de datos **Human Phenotype Ontology**, versión de febrero de 2019, tamaño: 5.9 MB (Köhler et al. 2019)
6. Información de términos de **Gene Ontology (GO)**, versión de septiembre de 2018, tamaño: 7 MB, para cada gen anotado (Ashburner et al. 2000)

### 2.2.3 Archivo de entrada

VCF/Plotein trabaja con archivos de tipo VCF, el formato estándar para el almacenamiento de información de variantes genéticas (Danecek et al. 2011). Dicho archivo es un archivo de texto plano con 9 columnas obligatorias, aunque puede contener más. En éstas, se encuentra información sobre la posición genómica, los alelos de referencia y alternativos, el ID de la variante si ya ha sido observada en otros estudios, puntaje de calidad, filtros aplicados, y otras métricas como anotaciones y genotipos de cada una de las muestras en el estudio de secuenciación.

### 2.2.4 Flujo funcional de la aplicación

Al cargarse, lo primero que la aplicación realiza es validar que el archivo que introduce el usuario sea en efecto un archivo de tipo VCF. En caso de no serlo, la aplicación genera un mensaje de error invitando al usuario a revisar el archivo. Posteriormente, la aplicación realiza un ejercicio de análisis línea por línea y almacena en memoria la información relacionada al cromosoma, la posición y la información del alelo de referencia y el alternativo.

Después de identificar la versión del ensamblaje a partir de la línea que contiene esta información en el VCF, los genes representados por las variantes en el VCF son identificados utilizando un algoritmo de árbol de intervalos en donde se realiza la comparación de cada una de las variantes con una librería que contiene las posiciones genómicas de cada gen. En este punto, es importante mencionar que distintos genes pueden compartir varias posiciones genómicas, lo cual aumenta el tiempo de procesamiento ya que cada una de las variantes debe de ser evaluada contra las coordenadas de cada uno de los genes.

Al obtener una lista de cada uno de los genes representados en el VCF, la aplicación genera tres índices para que el usuario pueda filtrar los genes en su VCF por proceso biológico, función molecular o componente celular (términos de tipo GO) (Ashburner et al. 2000). Una vez que el usuario selecciona un gen, la aplicación solicita información y los transcritos asociados al gen a la base de datos Ensembl por medio de servicios web a través de su REST API (Zerbino et al. 2018), independientemente de la existencia de otras anotaciones en el VCF. El acotar el proceso de selección a un solo gen es un aspecto clave para mejorar el tiempo de procesamiento de la aplicación, ya que reduce considerablemente el esfuerzo computacional así como la transferencia de información a través del internet.

Consecuentemente, información sobre las consecuencias de las variantes genéticas, así como los puntajes asignados por los algoritmos SIFT (Ng and Henikoff 2003) y PolyPhen-2 (Adzhubei, Jordan, and Sunyaev 2013) para cada una de las variantes se obtienen a partir del Ensembl Variant Effect Predictor (McLaren et al. 2016). Posteriormente, se revisa la información adicional que pueda existir en las bases de datos instaladas en el servidor sobre cada una de las variantes. Esto se realiza a través de una búsqueda a la base de datos interna a través del motor de búsqueda *Elasticsearch* (“Elasticsearch Reference [7.5] | Elastic” n.d.) (**Figura 8**).

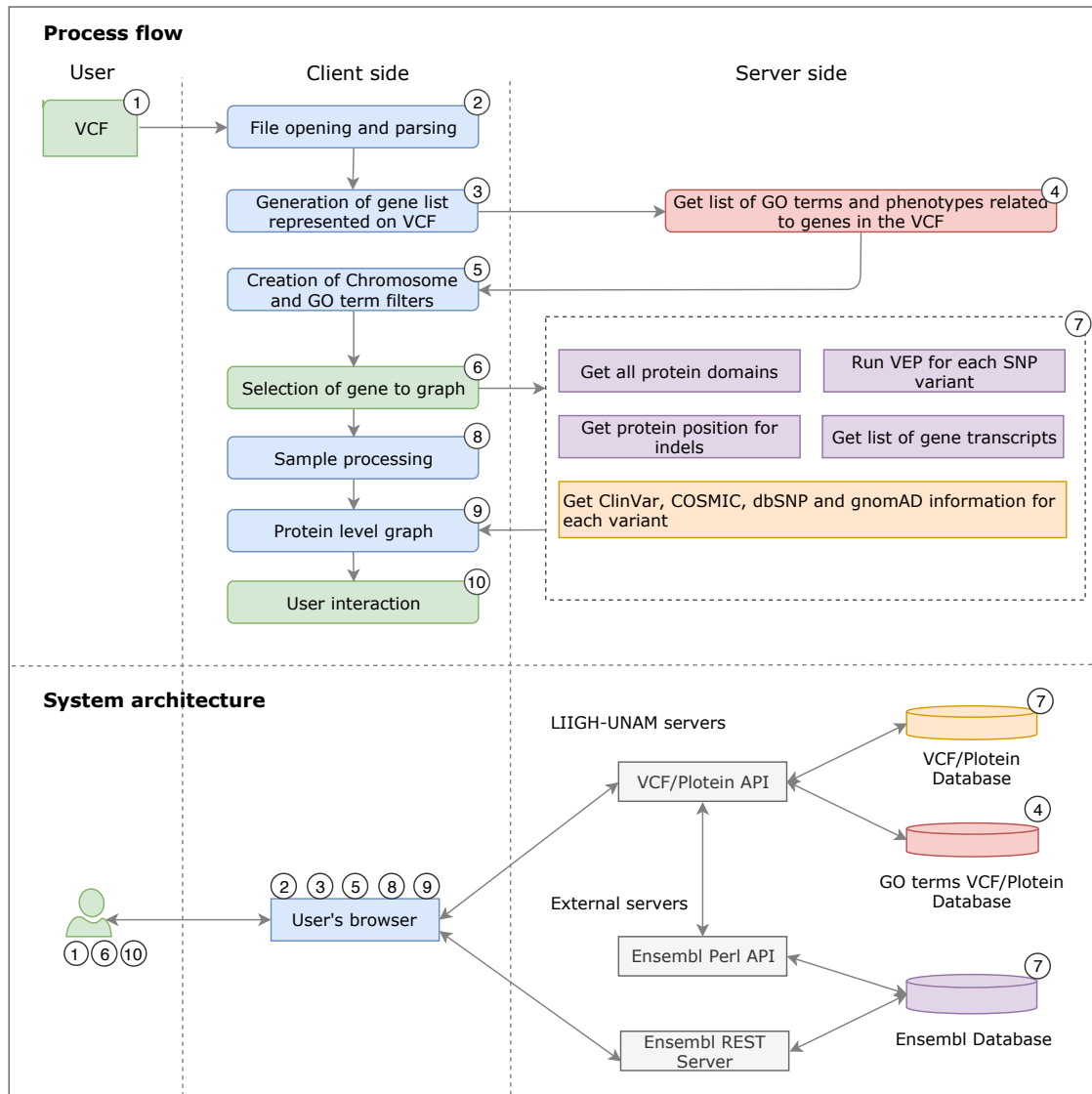
#### 2.2.5 Construcción y partes de la gráfica generada

Toda la información recolectada se almacena como un objeto en un archivo tipo JSON y se envía al navegador para generar una gráfica interactiva que represente de forma resumida toda la información recolectada. La gráfica representa la estructura primaria de la proteína del transcrito canónico, y fue creada utilizando la librería D3.js, ya que permite mostrar la información de manera personalizada y en forma de vectores en tipo SVG. Todo el procesamiento, a excepción de la solicitud de información a las fuentes de información externa, se realiza de forma local utilizando el CPU del usuario (**Figura 9**).



### 2.2.5.1 Parte central de la gráfica.

Muestra la estructura primaria de la proteína del transcrito canónico al lado de los dominios y otras características presentes en la proteína, con las variantes genéticas detectadas en ella como “paletas”. Los colores de estas paletas indican la consecuencia de la variante en la estructura de la proteína, y los cuadrados debajo indican la presencia o ausencia de la variante en distintas bases de datos externas (**Figura 9**, inciso 1).



**Figura 8**

**Diagrama de flujo con el flujo funcional de VCF/Plotein y la arquitectura del sistema.** Los colores en cada cuadro en el panel superior corresponden al color del elemento en la arquitectura del sistema en el cual es ejecutado (en el panel inferior), e.g., las acciones en los cuadros verdes son ejecutados por el usuario, los que están en recuadros azules son ejecutados por el navegador del usuario, y las acciones en rojo, morado y anaranjado son procesos utilizados para obtener información de distintas bases de

datos (cada color representa una base de datos distinta). Los números en la figura indican el flujo funcional, y representan las mismas acciones en el diagrama de la arquitectura del sistema: 1) El usuario selecciona y carga el archivo VCF (comprimido o sin comprimir) en el panel “Open file” de la aplicación web, 2) El programa abre el archivo, lo descomprime si estaba comprimido, y procede a su análisis línea por línea, 3) Los genes representados en el archivo VCF son encontrados al introducir las variantes en un árbol de intervalos, 4) La lista de genes es enviada a un servidor para generar una lista de términos GO y fenotipos relacionados con los genes representados en el archivo VCF, 5) Los filtros de términos GO y fenotipos son creados en la aplicación web, 6) El usuario selecciona el gen de interés para generar su gráfica, 7) El programa extrae todas las variantes del gen seleccionado y clasifica la información asociada en delicada y no delicada. La información delicada se define como nombres de muestras o IDs, información a nivel de genotipo de cada muestra, cualquier anotación previamente agregada al archivo VCF por el usuario, o información en el encabezado del archivo VCF. La información no delicada son el cromosoma, la posición genómica, y el cambio de base. Esta última es enviada al Ensembl REST API para obtener información del Ensembl VEP. El programa también envía esta información no delicada a los servidores de la UNAM para extraer información de bases de datos públicas (gnomAD, dbSNP, ClinVar y COSMIC) utilizando el motor de búsqueda *Elasticsearch*, 8) Al mismo tiempo que la información no delicada es enviada a los servidores, el programa comienza a procesar la información sobre las muestras instaladas de forma local en la computadora del usuario, 9) Después de que el programa tiene toda la información que necesita de Ensembl y las distintas bases de datos, la procesa y la presenta de forma gráfica en un diagrama de la estructura primaria de la proteína.

#### *2.2.5.2 Parte superior de la gráfica*

Este panel muestra información sobre el gen y el transcrito siendo visualizado al momento, así como ya sea información general acerca de todas las variantes en el gen, o si alguna variante específica es seleccionada, información relacionada con ella (la posición genómica, las consecuencias en la proteína, y las muestras que la contienen) (**Figura 9**, inciso 2).

#### *2.2.5.3 Parte inferior de la gráfica*

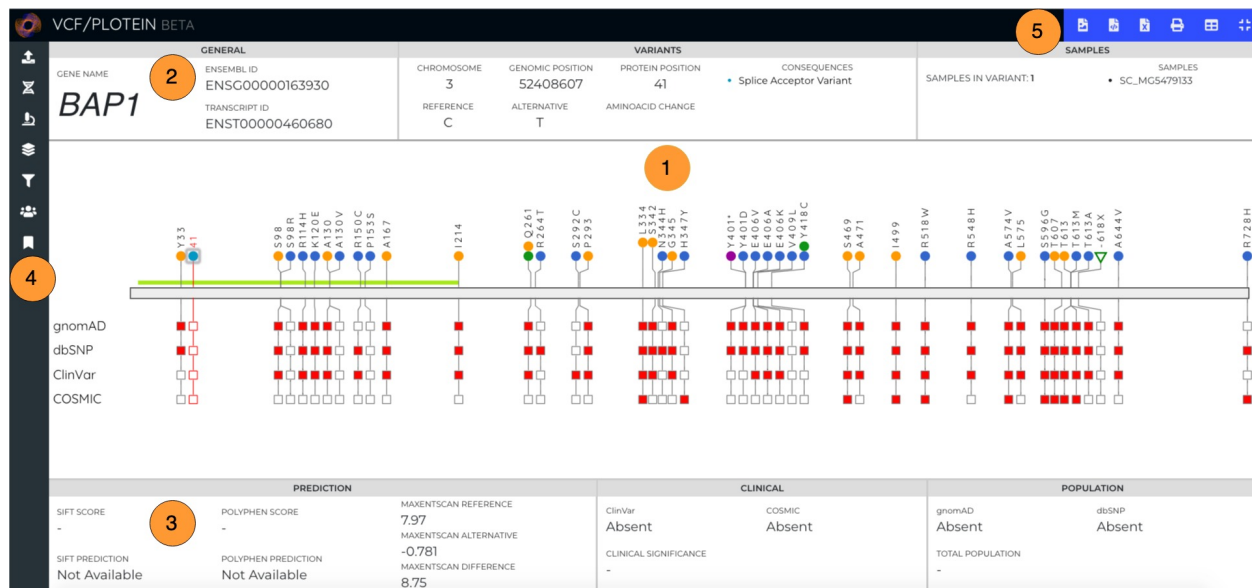
Si una variante en particular es seleccionada, el panel inferior muestra los puntajes asignados por los algoritmos SIFT, PolyPhen-2 y MaxEntScan (Yeo and Burge 2004), dependiendo del tipo de variante, e información acerca de su presencia en bases de datos externas como dbSNP, ClinVar, COSMIC y gnomAD (**Figura 9**, inciso 3).

#### *2.2.5.4 Menú lateral*

Con el objetivo de poder colocar filtros y controlar varios aspectos de la gráfica, colocamos un menú retraíble en el eje vertical izquierdo de la pantalla. Este menú permite al usuario personalizar e interactuar con la gráfica, y tiene funciones para subir un nuevo archivo para analizar, seleccionar un gen distinto en el archivo siendo examinado al momento, mostrar distintos dominios y características de la proteína, y filtrar las variantes de formas distintas (**Figura 9**, inciso 4).

### 2.2.5.5 Menú superior

Este menú permite al usuario visualizar y exportar información en distintos formatos. Se incluyó la opción de exportar toda la información en un archivo tipo JSON para almacenarla o compartirla. Este archivo se puede introducir de nuevo al programa y permite al usuario ver e interactuar su información de manera inmediata sin tener que realizar todo el procesamiento y solicitud de información (**Figura 9**, inciso 5).



**Figura 9**  
**Pantalla de trabajo de VCF/Plotein.** Vista del navegador cuando se ha seleccionado un gen para analizar. 1) Panel central, 2) Panel superior, 3) Panel inferior, 4) Menú lateral, 5) Menú superior.

## 2.3 Resultados

### 2.3.1 Panorama

Los únicos requisitos para utilizar VCF/Plotein son una computadora con un navegador y conexión a internet, así como un archivo VCF. Una vez que el usuario ha cargado el VCF en el programa, el ensamblaje del genoma es identificado y los genes representados por las variantes localizadas en el archivo se agrupan en una lista que el usuario puede filtrar por medio de categorías de términos GO (**Figura 10**). Una vez que un gen es seleccionado, se carga una nueva página que muestra la estructura primaria de la proteína codificada por el transcrito canónico con sus dominios y otras características, así como todas las variantes que se encuentran en éste.

Las variantes se muestran con una indicación de su frecuencia entre las muestras presentes en la cohorte del VCF, sus consecuencias en el transcrito seleccionado, y su presencia o ausencia en las bases de datos gnomAD, ClinVar y COSMIC (**Figura 9**). El usuario puede hacer click en cualquier variante para acceder a más información sobre ella, como es su posición genómica, una predicción de su patogenicidad de acuerdo a los algoritmos de SIFT y PolyPhen-2, y la lista de muestras que contienen. El menú a la izquierda permite al usuario realizar varias tareas relacionadas con cargar nuevos archivos o seleccionar distintas variantes y filtros en el archivo actual, así como marcar las características seleccionadas. El menú superior, la información sobre las variantes puede ser descargada en distintos formatos, o impresa en los formatos SVG o PNG.

The screenshot shows the VCF/PROTEIN BETA interface with three numbered callouts:

- 1**: A file upload area with a dashed border containing the text "all\_genes\_bap1\_example\_sorted.vcf.gz". Below it, the selected file name "all\_genes\_bap1\_example\_sorted.vcf.gz" and reference genome information "Reference genome: (37) Human assembly GRCh37 (hg 19)" are displayed.
- 2**: A "Filter genes" section with tabs for "Chromosome", "Biological process", "Molecular function", "Cellular component", and "Phenotype". The "Phenotype" tab is selected, showing a list of melanoma-related terms: "melanoma", "Choroidal melanoma", "Ciliary body melanoma", "Cutaneous melanoma", "Iris melanoma", and "Uveal melanoma".
- 3**: A "Genes (16)" table with a search bar and a list of genes. The table has columns for "NAME" and "ID".

NAME	ID
AC135541	ENSG00000264813
ACE	ENSG00000159640
ACE3P	ENSG00000224353
BAP1	ENSG00000163930
CDX2	ENSG00000165556
ERCC2	ENSG00000104884
FOXP1	ENSG00000150907

**Figura 10**

**Pantalla de carga y filtrado.** 1) El usuario puede hacer click en la caja gris para cargar un archivo VCF o un archivo marcador (creado con VCF/Plotein a partir de un archivo cargado previamente). El ensamblaje de referencia es identificado automáticamente de la línea con esta información en el VCF. 2) Se muestra una lista de distintos criterios para auxiliar con la priorización de genes: Los usuarios pueden seleccionar genes basados en un cromosoma de preferencia, o un proceso biológico, función molecular, componente celular o fenotipo asociado. 3) Se muestra una lista de genes que cumplen con los criterios seleccionados, de los cuales uno puede ser seleccionado para ser mostrado en la pantalla de trabajo.

### 2.3.2 Seguridad de los datos

Esta aplicación está especialmente diseñada para ser utilizada con facilidad por la comunidad médica y biológica, que por lo general tiene como prioridad la privacidad del paciente. El API y las bases de datos internas se encuentran instaladas detrás de un cortafuegos Fortinet, y la transferencia de información se realiza a través de un puerto HTTPS con un certificado SSL para lograr una transferencia segura de la información. Es muy importante mencionar que ninguna información relacionada a las muestras se evalúa fuera del navegador de usuario. De hecho, este

fue uno de los requerimientos que determinamos como obligatorio al inicio del proyecto, y gran parte de la arquitectura de la solución se realizó sobre este aspecto. Esto es de gran relevancia debido a las políticas de seguridad que la mayoría de los comités de investigación solicita a los estudios que involucran la secuenciación de ADN de individuos humanos. La única información que sale del navegador del usuario es el conjunto de las primeras cuatro columnas del VCF, las cuales contienen información no delicada. Asimismo, todo el procesamiento de datos relacionado a las muestras e información del genotipo se realiza de manera local. Es por esto que el servidor no almacena información sobre los nombres, IDs, genotipos o anotaciones de las muestras.

### 2.3.3 Filtrado de variantes y visualización

Las variantes encontradas en cualquier transcrito codificante de proteína de cualquier gen puede ser filtrado y graficado. Los usuarios pueden filtrar variantes por consecuencia de proteína, por predicción clínica, por puntaje de patogenicidad o por frecuencia alélica en la base de datos gnomAD, o pueden seleccionar un subconjunto personalizado para mostrar. Los usuarios también pueden seleccionar cuáles dominios de proteína y características quieren ver. La gráfica personalizada puede ser después exportada como archivo SVG o PNG.

### 2.3.4 Desempeño

VCF/Plotein es capaz de procesar archivos VCF de estudios de secuenciación de exoma completo en un tiempo reducido. Uno de los aspectos principales que tienen que ver con el desempeño tiene que ver con el proceso de abrir y cargar el archivo VCF, el cual requiere tanta RAM como el tamaño del archivo. Entonces, no existe un límite duro en este paso: Las computadoras que posean más memoria RAM serán capaces de realizar mejor esta tarea y serán capaces de abrir archivos más extensos. Una relación similar existe entre el tipo de procesador y el tiempo de procesamiento: Los procesadores con tiempo de reloj de pared más rápido podrán leer la información del archivo VCF con mayor rapidez. Ya que la aplicación se ejecuta desde el navegador, el sistema operativo no juega un papel importante en la calidad de su desempeño. Otros pasos que consumen tiempo son aquellos que requieren enviar y recibir datos a través del internet, los cuales son afectados por la velocidad de transferencia de datos y el número de

variantes que se envía a los servidores para la extracción de información de las bases de datos externas. Para ilustrar el desempeño de VCF/Plotein bajo distintas arquitecturas de sistema, probamos nuestra aplicación en varias máquinas distintas (**Tabla 3**). Mientras que VCF/Plotein debería correr sin problemas en la mayoría de los navegadores web, lo hemos probado más extensamente en el navegador Chrome corriendo en los sistemas operativos MacOS y Linux, así como en el navegador Edge del sistema operativo Windows 10 (**Tabla 3**).

**Tabla 3**

**Desempeño de VCF/Plotein después de cargar tres archivos VCF distintos en diferentes sistemas operativos con un rango de especificaciones de hardware.** Las pruebas fueron realizadas en el navegador Google Chrome en los sistemas operativos MacOS y Linux, así como el navegador Edge en Windows 10. Todos los tiempos están en milisegundos (ms).

	<b>Especificaciones del sistema</b>	<b>Procesos</b>	<b>VCF de un solo gen Tamaño: 2.9 mb Número de genes: 1 Gen seleccionado: 490 variantes Muestras: 201</b>	<b>VCF ClinVar Tamaño: 170.7 mb Número de genes: 7357 Gen seleccionado: 340 variantes Sin muestras</b>	<b>VCF de COSMIC Tamaño: 421.8 mb Número de genes: 23141 Gen seleccionado: 1979 variantes Sin muestras</b>
1	<b>Sistema operativo: macOS Mojave Procesador: 2.2 GHz intel core i7 RAM: 8 GB 1600MHz DDR3 Disco duro: 500 GB</b>	Apertura de archivo y procesamiento	62	1665	4390
		Generación de listas de genes	80	2821	32300
		Creación de filtros términos GO	5	1486	11277
		Consulta de información de la variante	5724	6499	22192
		Total	<b>5871</b>	<b>12471</b>	<b>70159</b>
2	<b>Sistema operativo: Manjaro Linux Procesador: 1.7GHz AMD A8 x 4 RAM: 8 GB Disco duro: 750 GB</b>	Apertura de archivo y procesamiento	222	11528	60492
		Generación de listas de genes	16	4954	75800
		Creación de filtros términos GO	255	12566	32418
		Consulta de información de la variante	6159	31141	106972
		Total	<b>6652</b>	<b>60189</b>	<b>275682</b>
3	<b>Sistema operativo: Windows 10 Procesador: intel core i7-6500U RAM: 8 GB 1600MHz DDR3 Disco duro: 1 TB</b>	Apertura de archivo y procesamiento	113	9753	70859
		Generación de listas de genes	94	7897	6348
		Creación de filtros términos GO	323	9130	11784
		Consulta de información de la variante	7277	8558	63012
		Total	<b>7807</b>	<b>35338</b>	<b>152003</b>
4	<b>Sistema operativo: macOS Mojave Procesador: 2.2 GHz intel core i7 RAM: 64 GB 1600MHz DDR3 Disco duro: 1 TB Solid state drive</b>	Apertura de archivo y procesamiento	35	2666	33034
		Generación de listas de genes	15	10648	110945
		Creación de filtros términos GO	112	10488	29218
		Consulta de información de la variante	4974	7042	19251
		Total	<b>5136</b>	<b>30844</b>	<b>192448</b>

### 2.3.5 Archivos marcador

Los archivos marcador permiten al usuario guardar de manera sencilla las características seleccionadas de cualquier número de transcritos en un archivo de texto, en formato JSON, que pueden subsecuentemente ser cargados en VCF/Plotein y compartidos con otros colegas e investigadores.

### 2.3.6 Comparación con otras herramientas similares

Como mencionamos en la introducción, existen otras herramientas similares que permiten realizar algunas de las funciones de VCF/Plotein, pero requieren ya sea al menos un mínimo de experiencia en técnicas bioinformáticas, no importan información de bases de datos externas, o no están disponibles de manera libre (**Tabla 4**).

## 2.4 Caso de uso

Para ilustrar como utilizar VCF/Plotein, hemos creado un caso de uso basado en un VCF real obtenido de la publicación de (O'Shea et al. 2017), quienes llevaron a cabo estudios funcionales para identificar aquellas variantes en el gen *BRCA1-associated protein 1 (BAP1)* con alta probabilidad de conferir un riesgo elevado a desarrollar melanoma. Hemos también suplementado este VCF con datos simulados de mutaciones para agregar información de más genes.

### 2.4.1 Priorizando variantes genéticas en el gen *BAP1* que predisponen a melanoma

Para este proyecto, se secuenciaron todos los exones de un número reducido de genes en 1,977 casos de melanoma y 754 controles, que fueron reclutados en el hospital de St. James's, Universidad de Leeds, quienes son todos de ascendencia genética europea. Se sabe que algunas variantes de frecuencia alélica baja en el gen *BAP1* han sido asociadas a lesiones melanocíticas, así como a una variedad de otros tipos de cáncer (Affar and Carbone 2018), por lo que sería beneficioso para la comunidad médica y científica si pudiéramos priorizar aquellas variantes, de todas las encontradas en un estudio grande de secuenciación, para que fueran analizadas de manera funcional y clínica. Esto nos permitiría reunir suficiente información para que después

éstas sean incluidas en paneles de pruebas genéticas y consideradas en casos de consejo genético a pacientes.

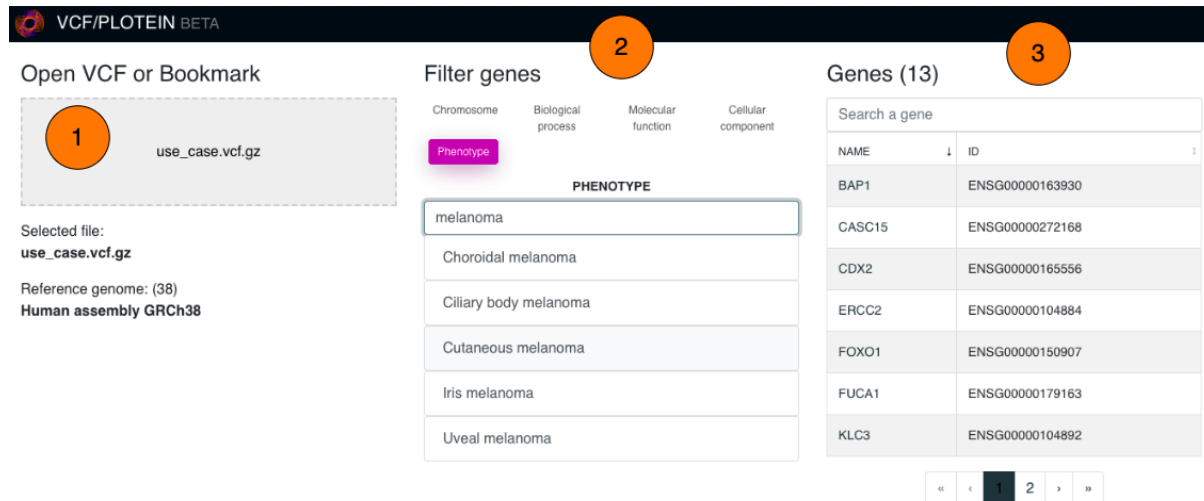
**Tabla 4**

**Comparación de las principales características de VCF/Plotein con aquellas de otras herramientas similares.** Las herramientas marcadas con (\*) no están disponibles de manera gratuita, por lo que se obtuvieron sus características de la documentación disponible. Los programas marcados con (\*\*\*) permiten seleccionar variantes por tipo de consecuencia pero no por frecuencia alélica o predicción de patogenicidad. NA: No disponible.

	VCF/Plotein	GEMINI	VCF-minor	BrowseVCF	BIERapp	exomesuite	PeCan	VCF Iobio	VizGVar	VariantStudio*	Alamut*	Ingenity*	VarSeq*	IGV
<b>No requiere instalación</b> El usuario no requiere llevar a cabo ningún proceso de instalación, ya sea desde la línea de comando o por estrategia de apunte y clickeo.	✓	x	x	x	x	x	✓	✓	✓	x	x	✓	x	✓
<b>No requiere conocimiento de línea de comando.</b> No requiere conocimiento de línea de comando por el usuario, ya sea para instalar u operar la aplicación	✓	x	x	x	✓	x	✓	✓	✓	✓	✓	✓	✓	x
<b>Interfaz gráfica (GUI)</b> El programa cuenta con una interfaz gráfica de usuario para hacer más sencilla la interacción, análisis y visualización de información	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<b>Análisis de VCF propio</b> El programa es capaz de analizar un VCF provisto por el usuario	✓	✓	✓	✓	✓	✓	x	✓	x	✓	✓	✓	✓	✓
<b>Consulta de bases de datos externas</b> El programa es capaz de obtener información de distintas bases de datos públicas para ayudar al usuario a priorizar variantes	✓	✓	x	x	✓	✓	✓	x	✓	✓	✓	✓	✓	x
<b>Sin pasos de pre-procesamiento requeridos</b> El programa no requiere que el VCF esté pre-procesado o que sea convertido en un formato de base de datos	✓	x	x	x	✓	x	x	x	NA	✓	x	✓	✓	x
<b>Priorización de variantes</b> El programa permite al usuario priorizar variantes utilizando filtros o especificando un set de criterios personalizados.	✓	✓	✓	✓	✓	✓	✓	x	***	✓	✓	✓	✓	x
<b>Gráfica de información a nivel de proteína</b> El programa genera una gráfica para analizar el impacto de las variantes en la estructura primaria de la proteína.	✓	x	x	x	x	x	✓	x	✓	x	x	x	x	x
<b>Procesamiento local de muestras (seguridad)</b> Datos delicados como el ID de la muestra u otras anotaciones no son enviados al servidor, sino que son procesados localmente en la computadora del usuario.	✓	✓	✓	✓	✓	✓	NA	x	NA	✓	✓	x	✓	✓
<b>Permite exportación de la imagen como SVG</b> La gráfica de los resultados puede ser exportada a un formato gráfico de vectores escalables para que puedan ser utilizados en cualquier publicación sin pérdida de calidad.	✓	NA	NA	NA	x	NA	✓	x	✓	NA	x	NA	x	x
<b>Disponible de manera gratuita</b> El programa no requiere una suscripción ni una cuota	✓	✓	✓	✓	✓	✓	✓	✓	✓	x	x	x	x	✓




La información de la secuencia de los 1,977 casos y los 754 controles está hospedado en un VCF, que puede ser descargado de la página de **Online Materials** de nuestra publicación (Ossio et al. 2019). Lo primero que hacemos es cargar el VCF en VCF/Plotein (**Figura 11-1**).



**Figura 11**

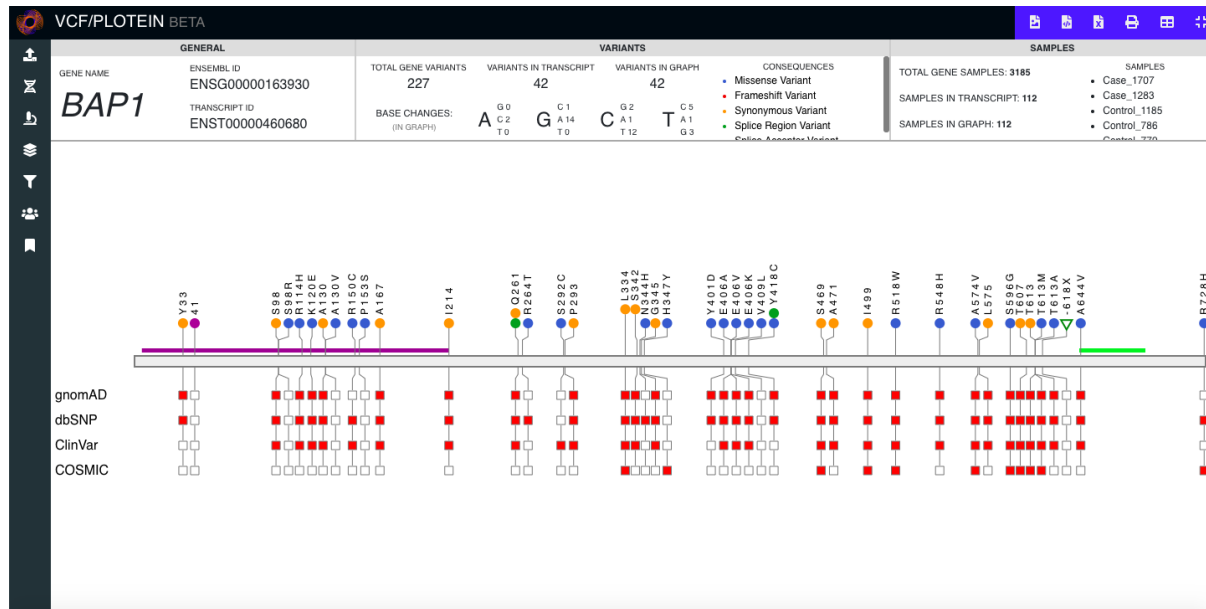
**Pantalla de carga de VCF/Plotein con los datos del caso de uso.** 1) Ventana utilizada para cargar un archivo VCF, en este caso el VCF basado en la publicación de (O’Shea et al. 2017). 2) Pantalla de priorización de genes por diversas categorías. 3) Lista de genes que cumplen con los criterios seleccionados, entre los que vemos a *BAP1*.

De aquí, podemos ver que tenemos información sobre varios genes que se encuentran con variaciones en el VCF. Podemos encontrar genes que se encuentran asociados al fenotipo de interés (en este caso, melanoma) seleccionando éste en el filtro de fenotipos (**Figura 11-2**) o directamente escribiendo el nombre del gen en la caja de búsqueda (**Figura 11-3**).

Podemos seleccionar el transcrito de interés utilizando el menú con el símbolo , en este caso el transcrito de interés es el *ENST00000460680* (el cual es el más largo). VCF/Plotein muestra el transcrito canónico, anotado en la base de datos Ensembl (Yates et al. 2020), por defecto.


En la imagen de este transcrito podemos ver una gráfica de todas las variantes encontradas en todos los individuos presentes en el VCF de entrada (**Figura 12**). Si el usuario hace click en las diferentes variantes, puede visualizar más información acerca de las muestras portadoras,

predicciones de patogenicidad, puntajes de MaxEntScan (Yeo and Burge 2004) cuando estos estén disponibles, y la presencia o ausencia de estas variantes en distintas bases de datos.




**Figura 12**

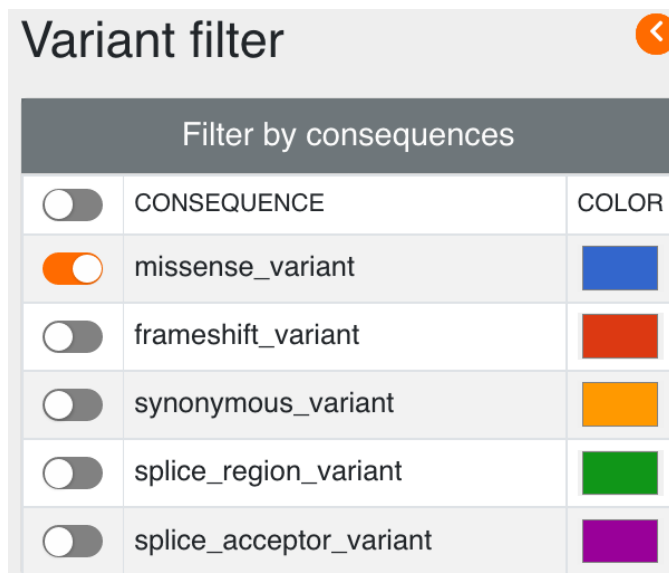
**Pantalla con vista del péptido resultante del transcrito canónico de *BAP1*.** El programa muestra la estructura primaria con los dominios como líneas de colores, las variantes como “paletas” de colores, donde cada color representa un tipo de variante distinto (siguiendo las categorías mostradas en la **Tabla 1**) y la presencia (rojo) o ausencia (blanco) de las variantes en distintas bases de datos se muestra en recuadros debajo de cada variante.

Para realizar un primer filtro de estas variantes, el usuario podría enfocarse en aquellas cuyos puntajes de patogenicidad predican que son las más deletéreas para la proteína (típicamente, variantes que afectan sitios de corte y empalme, o que causan una ganancia de codón de paro o de cambio de marco de lectura, **Tabla 1**). Una búsqueda rápida por medio del menú marcado con el símbolo  muestra que existe una variante de cambio de marco de lectura (en el aminoácido 618) y una en sitio de corte y empalme (cerca al aminoácido 4). Entonces, estas variantes podrían ser interesantes para realizar un seguimiento funcional.

Pero, ¿Qué hay acerca de otras variantes? Éstas podrían cambiar la estructura de la proteína pero sus efectos podrían ser un poco más complicados de predecir. Es por eso que es muy útil integrar distintas líneas de evidencia para realizar esta priorización, y es aquí donde VCF/Plotein puede aportar de manera importante a este proceso.

Lo primero que el usuario puede hacer es enfocarse en variantes que caigan sobre dominios importantes de la proteína, ya que éstas podrían afectar de manera más severa la función de la proteína dado el sitio donde se encuentran. En este ejemplo, existe un dominio, anotado en la base de datos PFM (El-Gebali et al. 2019), cerca del extremo N-terminal de la proteína marcado como *'Peptidase C12, ubiquitin carboxyl-terminal hydrolase'*. BAP1 funciona primordialmente como una deubiquitinasa (Affar and Carbone 2018), por lo cual esta región parece un sitio plausible para enfocarse. VCF/Plotein muestra todos los dominios anotados en la base de datos PFM (El-Gebali et al. 2019), a través de la base de datos Ensembl (Yates et al. 2020), por defecto.

Ahora, si el usuario muestra todas las consecuencias de las variantes, puede observar que existen 11 variantes posibles (excluyendo el sitio de corte y empalme que ha sido mencionado anteriormente) que caen sobre este dominio funcional. Entonces, el usuario puede mostrar sólo aquellas variantes que cambian la estructura de la proteína. Para realizar este paso, puede ir al menú  y filtrar por consecuencia, seleccionado sólo aquellas variantes que son de cambio de aminoácido (**Figura 13**).



**Figura 13**  
Menú de filtrado de consecuencias por variante.

Al aplicar este filtro, el usuario puede comprobar que solamente quedan seis variantes, y que todas menos P153S caen en casos de melanoma (lo cual puede ser comprobado en el menú superior al hacer click en cada variante o al filtrar para mostrar solamente los casos en el menú de la izquierda). Un paso común que sigue en los ejercicios de priorización de variantes es enfocarse solamente en

aquellas que tienen puntajes de “deletéreas” o “probablemente dañinas” por los algoritmos SIFT

(Ng and Henikoff 2003) y PolyPhen-2 (Adzhubei, Jordan, and Sunyaev 2013) respectivamente (**Figura 14**), y en aquéllas que tiene una frecuencia alélica muy baja en la población en general (Lek et al. 2016). Al aplicar todos estos filtros, solamente dos variantes quedan: S98R y A130V, las cuales, para resumir, caen en los dominios funcionales de la proteína, tienen efectos predichos como deletéreos por dos algoritmos bioinformáticos, y no se encuentran en la población en general. Entonces, estos serían buenos candidatos para realizar un seguimiento funcional.

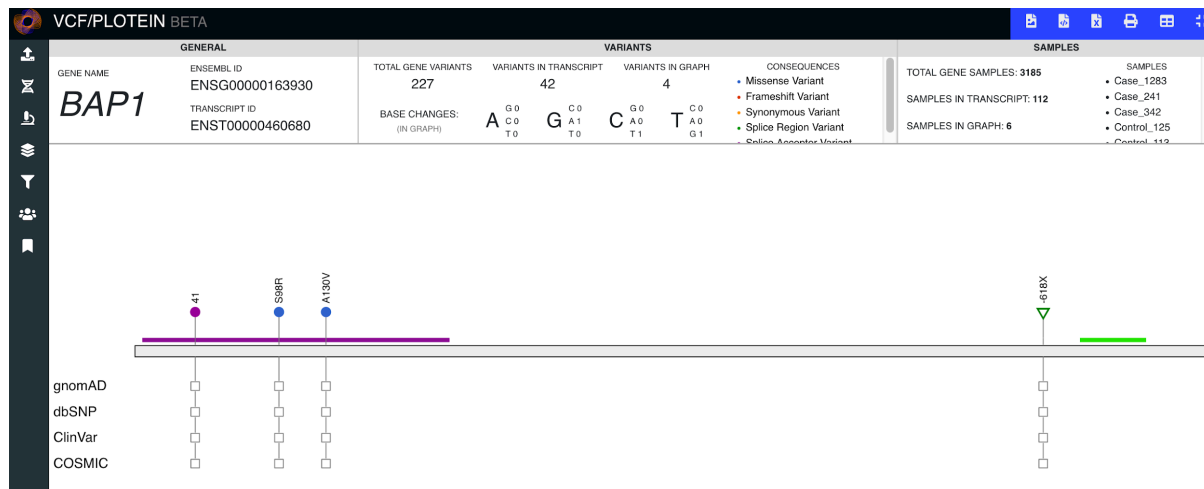
Filter by SIFT prediction	
<input type="checkbox"/>	ALL SIFT CATEGORIES
<input checked="" type="checkbox"/>	deleterious
<input type="checkbox"/>	tolerated
<input type="checkbox"/>	tolerated_low_confidence
<input type="checkbox"/>	deleterious_low_confidence
<input type="checkbox"/>	undefined
<input type="checkbox"/>	Not Available
Filter by PolyPhen prediction	
<input type="checkbox"/>	ALL POLYPHEN CATEGORIES
<input checked="" type="checkbox"/>	probably_damaging
<input type="checkbox"/>	possibly_damaging
<input type="checkbox"/>	benign
<input type="checkbox"/>	undefined
<input type="checkbox"/>	Not Available

**Figura 14**  
Menú de filtrado por predicción de patogenicidad por SIFT y PolyPhen-2.

En la publicación original (O'Shea et al. 2017), las variantes con consecuencias de cambio de aminoácido fueron probadas funcionalmente por medio de ensayos de deubiquitinación, y, tal cual es predicho por VCF/Plotein y estos pasos de priorización, la variante S98R causa que BAP1 pierda su función y por lo tanto se vuelve un candidato posible a aumentar el riesgo a desarrollar melanoma, sumada a las variantes que afectan sitios de corte y empalme y cambio de marco de lectura (**Figura 15**).

Sin embargo, esta no es la única estrategia que un usuario podría seguir para priorizar variantes: Existen muchos otros criterios que se podrían aplicar a estas variantes, dependiendo en el objetivo del proyecto. Por ejemplo, también se podría enfocar en

aquellas variantes que alteran la estructura de la proteína reportadas en la base de datos ClinVar (Landrum et al. 2018) (R114H, K120E, y R150C) aunque éstas han sido previamente descritas como presentes en la población o tienen puntajes de patogenicidad conflictivos.



**Figura 15**  
**Vista de las variantes restantes en BAP1 después de los filtros de priorización.** Estas variantes son candidatos para aumentar el riesgo a desarrollar melanoma.

## 2.5 Discusión

La priorización de variantes genéticas por medio del análisis interactivo del impacto que tienen éstas en la estructura y función de las proteínas es un ejercicio fundamental para poder identificar factores que contribuyan al desarrollo de una gran variedad de fenotipos. Con este trabajo, esperamos que VCF/Plotein permita a los investigadores, especialmente en laboratorios pequeños y con especialidades distintas a la bioinformática, a enfocarse en preguntas biológicas relevantes relacionadas a su proyecto de investigación en lugar de tener que aprender a instalar dependencias de software, o tener que utilizar herramientas más complicadas de anotación y búsqueda en otras bases de datos, o familiarizarse con UNIX o la línea de comando de MySQL. Las ventajas más importantes que ofrece la herramienta que desarrollamos por encima de otras similares son la facilidad de uso, la capacidad de mostrar información de un archivo VCF propio, que está disponible de manera gratuita, y que puede procesar archivos de manera local. También, ilustramos con un caso de uso que, al aplicar un número razonable de filtros, un investigador puede identificar un subconjunto pequeño de variantes dentro de un gen que

contiene aquéllas con predicciones deletéreas para la función de la proteína codificada. Al combinar el filtrado de variantes y la anotación en una sola herramienta gráfica e interactiva, hemos demostrado que la priorización y la visualización de variantes genéticas puede realizarse de una manera fácil, rápida e intuitiva.

## 2.6 Disponibilidad de la aplicación

VCF/Plotein se encuentra disponible de manera libre y gratuita en la dirección <https://vcfplotein.liigh.unam.mx>. El sitio web fue implementado en JavaScript utilizando el marco de trabajo Vue.js, y fue probado y es soportado por los principales navegadores. El código fuente está disponible de manera libre y gratuita en <https://github.com/raulossio/VCF-plotein>.

## 2.7 Equipo de trabajo

Esta aplicación fue desarrollada con el apoyo de Diego Said Anaya-Mancilla, estudiante de licenciatura que se encontraba realizando su servicio social en el grupo de la Dra. Robles Espinoza, Omar Isaac García Salinas, tesista de licenciatura en el mismo grupo, y Jair S. García Sotelo y Luis Alberto Aguilar, del Laboratorio de Investigación sobre el Genoma Humano (LIIGH) y el Laboratorio Nacional de Visualización Científica Avanzada (LAVIS), de la Universidad Nacional Autónoma de México, respectivamente. Este proyecto fue liderado en todo momento por mí. Recibí consejo del Dr. David Adams, del Instituto Wellcome Sanger, y de mi asesora, la Dra. Daniela Robles Espinoza.

## 2.8 Perspectivas

La realización de este programa es el primer paso para su aplicación a proyectos de secuenciación más grandes, como es el foco de la presente tesis, para identificar variantes que contribuyen o aumentan el riesgo a desarrollar un fenotipo. La utilidad del presente trabajo ha sido evidenciada por dos publicaciones en las que participé como co-autor con el código y el programa generado, las cuales se encuentran en la sección de **Publicaciones originadas de este doctorado**. En la primera, el código escrito para VCF/Plotein fue adaptado en la herramienta *feature-map2*, parte de la suite *Regulatory Sequence Analysis Toolkit (RSAT)*, para representar visualmente y de

manera moderna distintas anotaciones en una o más secuencias, así como permitir el análisis de varios mapas en paralelo (Nguyen et al. 2018). En la otra, se utilizó un prototipo de VCF/Plotein para analizar las variantes encontradas por un estudio con 181 familias de todo el mundo portadoras de variantes en el gen *BAP1* (Walpole et al. 2018). VCF/Plotein reveló el espectro de variantes de cambio de aminoácido presentes, así como su presencia en distintos dominios de la proteína, su evidencia de patogenicidad y su presencia o ausencia en distintas bases de datos (**Figura 1** de la citada publicación).

Entonces, el desarrollo de esta herramienta no sólo permitirá a otros investigadores aplicarla a sus proyectos, ya sea completa o por modificación del código provisto, sino también es el primer paso en nuestro proyecto para identificar variantes que predisponen a melanoma, tema que será abordado en el siguiente capítulo.

### 3 Aplicación de VCF/Plotein: Priorizando variantes en *POT1* en un estudio de casos y controles de melanoma

Habiendo escrito y probado un programa que puede ser utilizado para priorizar e identificar variantes genéticas que incrementan el riesgo a desarrollar cáncer (Ossio et al. 2019), el interés del presente capítulo es su aplicación a un estudio de secuenciación de casos y controles con melanoma.

#### 3.1 Introducción

Desde el descubrimiento de alelos patogénicos en el gen *CDKN2A* hace más de veinticinco años (Kamb et al. 1994), se han descrito otras tantas variantes que incrementan el riesgo a desarrollar melanoma por medio de estudios GWAS (Law et al. 2015) y el análisis de familias con predisposición a este tipo de cáncer. Estas variantes patogénicas afectan genes que regulan la pigmentación, como *MC1R*, el conteo de lunares, como *PLA2G6*, el control del ciclo celular y senescencia celular, como *CDKN2A* y *CDK4*, y la regulación de los telómeros (Ribero, Glass, and Bataille 2016). De manera notable, las mutaciones encontradas en el promotor del gen de la telomerasa, *TERT*, reportadas por vez primera en una familia con historia de melanoma (Horn et al. 2013), sugirieron que los telómeros podrían jugar un papel importante en el desarrollo de este tipo de cáncer. Estudios siguientes revelaron variantes patogénicas en el gen *POT1* en familias con melanoma (Robles-Espinoza et al. 2014; Shi et al. 2014), así como en familias con otros tipos de cáncer como glioma (Bainbridge et al. 2015), leucemia (Speedy et al. 2016) y linfoma (McMaster et al. 2018).

El gen *POT1* codifica para una proteína de unión a ADN de cadena sencilla (ssDNA, por su abreviatura en inglés) que forma parte del complejo protector de telómeros (*shelterin complex*, en inglés). Este complejo está formado por seis proteínas que tienen una función importante en el mantenimiento de la estructura telomérica y permiten a las células distinguir entre el final de los cromosomas y sitios de daño a ADN, así como la regulación de la actividad de la telomerasa para controlar la longitud telomérica (de Lange 2005). Como mencionábamos anteriormente, los



estudios de secuenciación de individuos propensos a desarrollar melanoma han revelado un número importante de alelos patogénicos de POT1 que afectan su capacidad de reconocer al ssDNA, teniendo como resultado una elongación anormal de los telómeros o aberraciones en estas secuencias (Robles-Espinoza et al. 2014; Shi et al. 2014; Wong et al. 2019). Actualmente no se conoce el mecanismo por el cuál estas variantes llevan al desarrollo de tumores, pero podría incluir una perturbación del programa de senescencia celular, defectos en la maquinaria de reparación de ADN, y/o alteraciones cromosómicas o teloméricas.

Ya que se han encontrado variantes patogénicas en el gen *POT1* en familias con melanoma y otros tipos de cáncer, éste ha sido incluido en paneles de pruebas genéticas que se utilizan para diagnosticar y aconsejar a estos pacientes. Es por esto que existe una necesidad de identificar aquellas variantes genéticas que alteran la función de POT1, así como de medir la frecuencia de éstas en casos de melanoma y controles poblacionales. En este estudio, se secuenciaron todos los exones codificantes del gen *POT1* en 2,929 casos de melanoma y 3,298 controles, para un total de 6,227 individuos británicos de tres cohortes distintas. Identificamos todas las variantes que alteran la estructura de la proteína tanto en casos como controles, realizamos una priorización con VCF/Plotein, y evaluamos la patogenicidad de cada una de las variantes candidatas por medio de experimentos de cambio de movilidad electroforética (EMSAs), los cuales miden la capacidad de las proteínas mutadas de unirse a ssDNA telomérico. Asimismo, también realizamos mediciones de longitudes teloméricas en todos los individuos portadores para identificar aquéllos con cambios notables en este aspecto de la función de POT1.

## 3.2 Métodos

### 3.2.1 Participantes y recolección de las muestras

Las muestras utilizadas en este estudio provinieron de tres cohortes distintas. El estudio de casos y controles de melanoma de la Universidad de Leeds ha reclutado casos y controles pareados por sexo y grupo de edad (+/- 5 años) de la región de Yorkshire, Reino Unido, desde el año 2000 (Newton-Bishop et al. 2010). Asimismo, se incluyeron muestras de la serie de estudios poblacionales *Epidemiology and Risk Factors in Cancer Heredity* (SEARCH), basado en el este de

Inglaterra (Pooley et al. 2010). Finalmente, se incluyeron controles provenientes del Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium 2007) (**Tabla Suplementaria 1**). Todos los pacientes firmaron un consentimiento informado. Se extrajo ADN de sangre periférica para llevar a cabo la amplificación de exones del gen *POT1* por medio de PCR, y su posterior secuenciación. Esta parte del trabajo fue realizada por nuestros colaboradores M. Harland, D. T. Bishop, J. A. Newton Bishop, de la Universidad de Leeds, y K. A. Pooley, de la Universidad de Cambridge, en Reino Unido.

### 3.2.2 Secuenciación e identificación de variantes

Se utilizó la tecnología de secuenciación de amplicones basada en PCR Fluidigm para amplificar todos los exones codificantes y los sitios de corte y empalme del gen *POT1* en 7,024 muestras, los cuales fueron subsecuentemente secuenciados en la plataforma Illumina MiSeq. Después, se realizó un alineamiento con el algoritmo BWA (Li and Durbin 2009), y un filtrado por calidad para mantener solamente aquellas muestras con alta cobertura (Aquellas con >94% de las bases exónicas de *POT1* cubiertas con una profundidad  $\geq 10$  lecturas de alta calidad, definidas como aquellas con calidad de mapeo  $\geq 50$  y calidad de base  $\geq 20$ ). Subsecuentemente, se filtró por muestra: se mantuvo en el estudio solamente una muestra de cada par de individuos emparentados, se eliminaron muestras que no eran de ancestría genética europea, controles emparentados con casos y las de aquellos individuos que retiraron su consentimiento durante la realización de este estudio. Al aplicar todos estos filtros, se mantuvieron en el estudio un total de 6,227 muestras. Estas muestras incluyen 2,929 casos de melanoma (1,574 del estudio de la Universidad de Leeds y 1,355 del consorcio SEARCH) y 3,298 controles (1,431 de la colección de WTCCC, 459 del estudio de la Universidad de Leeds y 1,408 del consorcio SEARCH) (**Tabla Suplementaria 1**). Para la identificación de variantes, para reducir la tasa de falsos positivos, incluimos en el estudio la unión de los conjuntos de variantes resultantes de correr el algoritmo HaplotypeCaller de GATK (McKenna et al. 2010) (ver **Línea de comando HaplotypeCaller (GATK)**, sección de **Anexos**) y Samtools mpileup (Li et al. 2009) (parámetros -t DP,SP -C50 -m2 -F0.0005 -d 10000 -ug). Posteriormente realizamos el filtrado por calidad de variante (mpileup: QUAL $\geq$ 20 && (DP4[2]>30 || DP4[3]>30); GATK: QUAL $\geq$ 20 && AD[0:1]>30). Asimismo, también

eliminamos los indeles identificados entre las coordenadas genómicas GRCh37 7:124475296-124475328, ya que esta región es altamente repetitiva y podría ser propensa a identificar un gran número de falsos positivos. En total, identificamos 141 variantes distintas con predicción de consecuencias que alteran la estructura protéica (cambio de aminoácido, ganancia de codón de paro, alteración de sitios de corte y empalme o de cambio de marco de lectura) utilizando Ensembl VEP versión 96 (McLaren et al. 2016), filtrando de manera laxa para capturar todas las posibles variantes existentes para su posterior validación (transcrito de referencia: *POT1 ENST00000357628*, **Tabla Suplementaria 2**). Esta validación fue realizada por medio de resecuenciación por capilar o por tecnología Illumina en al menos una muestra para todas las variantes detectadas (12 y 180 muestras, respectivamente, **Tabla Suplementaria 2**). La resecuenciación por Illumina fue llevada a cabo en el Wellcome Sanger Institute previa captura exónica por medio de sondas de la compañía Agilent Technologies utilizando el panel WTSI v4 Solid Tumour que incluye otros genes de predisposición a desarrollar melanoma como *CDKN2A*, *CDK4* y *BAP1*. Esta secuenciación obtuvo datos para todos los exones y sitios de corte y empalme de *POT1*. Esta secuenciación fue llevada a cabo por medio de nuestro colaborador D. J. Adams (Wellcome Sanger Institute), mientras que la secuenciación por capilar fue hecha por M. Harland (Universidad de Leeds). Los análisis aquí descritos los llevé a cabo en colaboración con mi tutora, C. D. Robles Espinoza (LIIGH-UNAM) y J. Taylor (Universidad de Leeds).

En total, 174 de 180 muestras obtuvieron cobertura promedio  $\geq 10$  en todos los exones codificantes de *POT1*. Se identificaron variantes por medio de Samtools mpileup en modo *pooled*, y de esta manera, logramos validar 42 de las 141 variantes detectadas inicialmente por el método de Fluidigm, mientras que no identificamos ninguna variante nueva. La resecuenciación por capilar en las 12 muestras que se perdieron durante la preparación de librerías identificó una variante adicional, para un total de 43/141 variantes confirmadas (**Tabla 5, Tabla Suplementaria 3**). No se identificaron variantes nuevas en las muestras resecuenciadas. En resumen, de 247 variantes con consecuencias de alteración de estructura de proteína originalmente identificadas en las 180 muestras resecuenciadas, pudimos validar 76, lo cual se traduce en una tasa de confirmación del 30.8%, sin falsos negativos detectados. Ya que, como mencionamos

anteriormente, nuestro método de captura nos permitió secuenciar genes adicionales de predisposición a melanoma, analizamos todas las muestras con mutaciones en *POT1* para descartar la posibilidad de que también fueran portadoras de variantes patogénicas en los genes conductores *CDKN2A*, *CDK4* y *BAP1*. Tres variantes con predicción de alteración de estructura protéica fueron identificadas en *CDKN2A* en portadores de variantes en *POT1*, aunque estas variantes en *POT1* fueron predichas como benignas de acuerdo a los experimentos EMSA (**Tabla Suplementaria 4**, transcritos de referencia *CDKN2A*: *ENST00000304494*, *CDK4*: *ENST00000257904*, *BAP1*: *ENST00000460680*).

**Tabla 5**

**Variantes con consecuencia de alteración de estructura de la proteína identificadas por medio de secuenciación Fluidigm y confirmadas por medio de resecuenciación por capilar o Illumina.** La primera columna muestra la posición de la variante en el genoma de referencia GRCh37, la segunda la consecuencia en la estructura de *POT1*, y la tercera el tipo de consecuencia. Para una versión completa de esta Tabla, ver la **Tabla Suplementaria 3** en la sección de **Anexos**.

Variante	Cambio en la proteína	Consecuencia
7:124464036,C/G	V629L	Cambio de aminoácido
7:124464049,A/C	I624M	Cambio de aminoácido
7:124464068,TTA/T	DN617-618EX	Cambio de marco de lectura
7:124464080,T/C	N614S	Cambio de aminoácido
7:124464089,T/C	N611S	Cambio de aminoácido
7:124465356,T/C	K581R	Cambio de aminoácido
7:124465412,C/T	Sp. aa. 563	Sitio de receptor de corte y empalme
7:124469308,C/G	A532P	Cambio de aminoácido y sitio de corte y empalme
7:124469337,G/A	T522I	Cambio de aminoácido
7:124469346,A/G	V519A	Cambio de aminoácido
7:124469347,C/T	V519I	Cambio de aminoácido
7:124475396,T/C	E481G	Cambio de aminoácido
7:124481086,T/C	H437R	Cambio de aminoácido
7:124481116,T/C	K427R	Cambio de aminoácido
7:124481116,T/G	K427T	Cambio de aminoácido
7:124481140,T/C	Y419C	Cambio de aminoácido
7:124481185,C/A	G404V	Cambio de aminoácido
7:124481210,C/G	D396H	Cambio de aminoácido
7:124481210,C/T	D396N	Cambio de aminoácido
7:124481218,T/C	H393R	Cambio de aminoácido
7:124481224,A/G	V391A	Cambio de aminoácido
7:124481233,C/T	Sp. aa. 388	Sitio de receptor de corte y empalme

7:124482894,G/A	S377F	Cambio de aminoácido
7:124482897,T/C	Q376R	Cambio de aminoácido
7:124482912,G/A	P371L	Cambio de aminoácido
7:124482936,C/T	R363Q	Cambio de aminoácido
7:124482952,G/GA	-357-358X	Cambio de marco de lectura
7:124491972,C/A	Q301H	Cambio de aminoácido
7:124493077,C/A	R273L	Cambio de aminoácido
7:124493119,A/C	L259*	Ganancia de codón de paro
7:124499043,C/T	D224N	Cambio de aminoácido
7:124503499,G/C	L151V	Cambio de aminoácido
7:124503540,C/T	R137H	Cambio de aminoácido
7:124503600,C/T	R117H	Cambio de aminoácido
7:124503608,G/C	I114M	Cambio de aminoácido
7:124503655,A/G	S99P	Cambio de aminoácido
7:124510966,T/G	K85T	Cambio de aminoácido y sitio de corte y empalme
7:124510988,T/C	I78V	Cambio de aminoácido
7:124510996,T/C	N75S	Cambio de aminoácido
7:124511044,C/T	C59Y	Cambio de aminoácido
7:124532327,T/A	K39N	Cambio de aminoácido
7:124532380,T/C	I22V	Cambio de aminoácido
7:124532435,C/T	Sp. aa. 4	Sitio de receptor de corte y empalme

### 3.2.3 Predicción de variantes deletéreas con VCF/Plottein

Introduciendo el VCF de este proyecto en VCF/Plottein, podemos visualizar la distribución de estas variantes y hacer hipótesis sobre aquéllas que tendrán un efecto deletéreo en los pacientes (Figura 16).

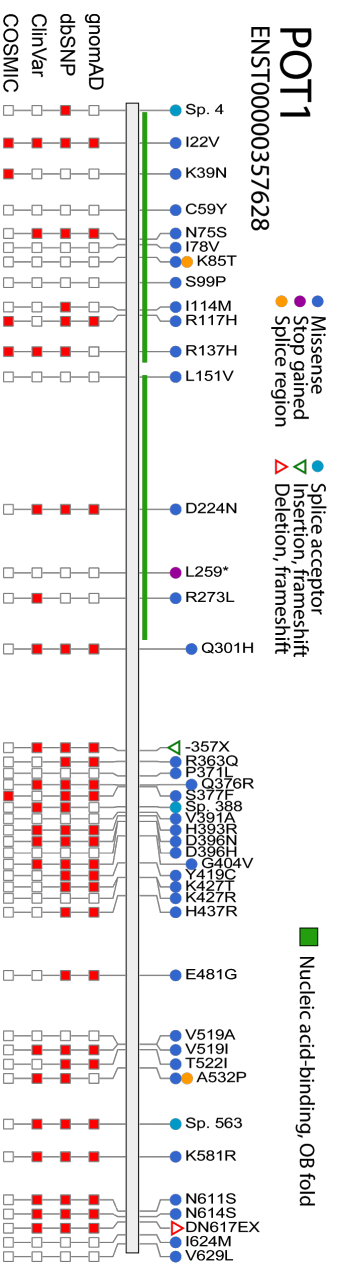


Figura 16

Gráfica de mutaciones presentes en POT1 elaborada con VCF/Plottein. En la gráfica se muestran las 43 variantes con consecuencia de cambio de en la proteína, con sus consecuencias. Y presencia en dominios y en bases de datos externas.

En total, identificamos 43 variantes que alteran la estructura de la proteína POT1 a través de secuenciación Fluidigm seguida de validación por medio de captura con sondas de Agilent Technologies y secuenciación Illumina o directamente por medio de secuenciación por capilar. En esta gráfica y en la **Tabla 5**, podemos observar que existen tres variantes con predicción de afectar sitios de corte y empalme, una variante que introduce un codón de paro, y dos variantes que alteran el marco de lectura. Asimismo, esta gráfica nos muestra que sólo 15 de estas variantes se encuentran en los dominios OB (de unión a oligonucleótidos u oligosacáridos (*oligonucleotide/oligosaccharide-binding*)), los cuales median la función de unión de POT1 al ssDNA. Veintitrés del total de variantes se encuentran presentes en la base de datos poblacional gnomAD, indicando la probabilidad de que no sean patogénicas, o representen alelos de bajo riesgo. Diecinueve de éstas se encuentran reportadas en ClinVar, indicando un posible grado de patogenicidad, y cinco se encuentran en la base de datos COSMIC. Poniendo todas estas observaciones en conjunto, podríamos hacer hipótesis sobre el conjunto de variantes que afectarán la función de unión a ssDNA, siguiendo los pasos que indicamos en el Caso de Uso discutido en el **Capítulo 3 (Tabla 6)**.

**Tabla 6**  
Predicción de patogenicidad de variantes encontradas en POT1 de acuerdo a VCF/Plotein

Cambio en la proteína	Predicción de patogenicidad	Razones
V629L	No patogénica	Cambio de aminoácido fuera de dominios OB
I624M	No patogénica	Cambio de aminoácido fuera de dominios OB
DN617-618EX	Patogénica	Cambio de marco de lectura
N614S	No patogénica	Cambio de aminoácido fuera de dominios OB
N611S	No patogénica	Cambio de aminoácido fuera de dominios OB
K581R	No patogénica	Cambio de aminoácido fuera de dominios OB
Sp. aa. 563	Patogénica	Sitio de receptor de corte y empalme
A532P	No patogénica	Cambio de aminoácido fuera de dominios OB
T522I	No patogénica	Cambio de aminoácido fuera de dominios OB
V519A	No patogénica	Cambio de aminoácido fuera de dominios OB
V519I	No patogénica	Cambio de aminoácido fuera de dominios OB
E481G	No patogénica	Cambio de aminoácido fuera de dominios OB
H437R	No patogénica	Cambio de aminoácido fuera de dominios OB

K427R	No patogénica	Cambio de aminoácido fuera de dominios OB
K427T	No patogénica	Cambio de aminoácido fuera de dominios OB
Y419C	No patogénica	Cambio de aminoácido fuera de dominios OB
G404V	No patogénica	Cambio de aminoácido fuera de dominios OB
D396H	No patogénica	Cambio de aminoácido fuera de dominios OB
D396N	No patogénica	Cambio de aminoácido fuera de dominios OB
H393R	No patogénica	Cambio de aminoácido fuera de dominios OB
V391A	No patogénica	Cambio de aminoácido fuera de dominios OB
Sp. aa. 388	Patogénica	Sitio de receptor de corte y empalme
S377F	No patogénica	Cambio de aminoácido fuera de dominios OB
Q376R	No patogénica	Cambio de aminoácido fuera de dominios OB
P371L	No patogénica	Cambio de aminoácido fuera de dominios OB
R363Q	No patogénica	Cambio de aminoácido fuera de dominios OB
-357-358X	Patogénica	Cambio de marco de lectura
Q301H	No patogénica	Cambio de aminoácido fuera de dominios OB
R273L	Patogénica	Presente en dominio OB, ausencia en bases de datos poblacionales, presencia en ClinVar
L259*	Patogénica	Ganancia de codón de paro
D224N	Información conflictiva	Presente en dominio OB y ClinVar, sin embargo presente en base de datos poblacional
L151V	Patogénica	Presente en dominio OB, sin presencia en bases de datos poblacionales
R137H	Patogénica	Presente en dominio OB, ausente en bases de datos poblacionales, presente en COSMIC y ClinVar
R117H	Información conflictiva	Presente en dominio OB y COSMIC, sin embargo presente en base de datos poblacional
I114M	Patogénica	Presente en dominio OB, no presente en bases de datos poblacionales
S99P	Patogénica	Presente en dominio OB, sin presencia en bases de datos poblacionales
K85T	Patogénica	Presente en dominio OB, sin presencia en bases de datos poblacionales
I78V	Patogénica	Presente en dominio OB, sin presencia en bases de datos poblacionales
N75S	Información conflictiva	Presente en dominio OB y ClinVar, sin embargo presente en base de datos poblacional
C59Y	Patogénica	Presente en dominio OB, sin presencia en bases de datos poblacionales
K39N	Patogénica	Presente en dominio OB y COSMIC, sin presencia en bases de datos poblacionales
I22V	Información conflictiva	Presente en dominio OB y ClinVar, sin embargo presente en base de datos poblacional
Sp. aa. 4	Patogénica	Sitio de receptor de corte y empalme

### 3.2.4 Traducción *in vitro* y ensayos de unión a telómeros

Se realizaron experimentos EMSA siguiendo el protocolo publicado anteriormente por miembros de nuestro grupo de investigación (Robles-Espinoza et al. 2014), y el cual explicamos aquí. Se enviaron a sintetizar ADN complementarios del gen completo POT1 con cada una de las variantes identificadas a la compañía Invitrogen GeneArt y se tradujeron a proteína en lisados de reticulocitos de conejo (ProMega). La expresión de proteínas fue confirmada por medio de Western Blot. La proteína traducida fue combinada con una sonda con secuencia telomérica marcada radiactivamente (P32) y los complejos proteína-ADN fueron visualizados en geles de poliacrilamida para separar aquellas sondas unidas a proteínas de las sondas libres. Esta parte del proyecto fue realizada por nuestro colaborador Chi Wong, en el Wellcome Sanger Institute. Después, las variantes se clasificaron en tres grupos: El Grupo 1 consistió en aquellas variantes que mostraron capacidad reducida o nula de unión a ADN por este experimento, así como las de ganancia de codón de paro, afectación de sitios de corte y empalme o cambio de marco de lectura. Las variantes pertenecientes al Grupo 2 fueron aquellas con predicción de consecuencias deletéreas por medio de los algoritmos SIFT y PolyPhen-2, pero que no mostraron cambios en su capacidad de unión a ADN de acuerdo a este experimento. El resto de las variantes fueron clasificadas en el Grupo 3.

### 3.2.5 Alineamiento multi-especies

Se descargaron secuencias de la proteína POT1 de la base de datos NCBI Protein en octubre de 2018 (humano: NP\_056265.2, ratón: NP\_598692.1, vaca: DAA30462.1, armadillo: XP\_004478310.1, elefante: XP\_003407293.1, zarigüeya: XP\_007504312.1, ornitorrinco: XP\_001508179.2, pollo: NP\_996875.1, rana: AAI71328.1, pez zebra: ADY16707.1) y las secuencias completas fueron alineadas con Clustal O 1.2.4 (Sievers et al. 2011) con los parámetros por defecto. El alineamiento de secuencias múltiples fue visualizado con el programa Jalview (Waterhouse et al. 2009).

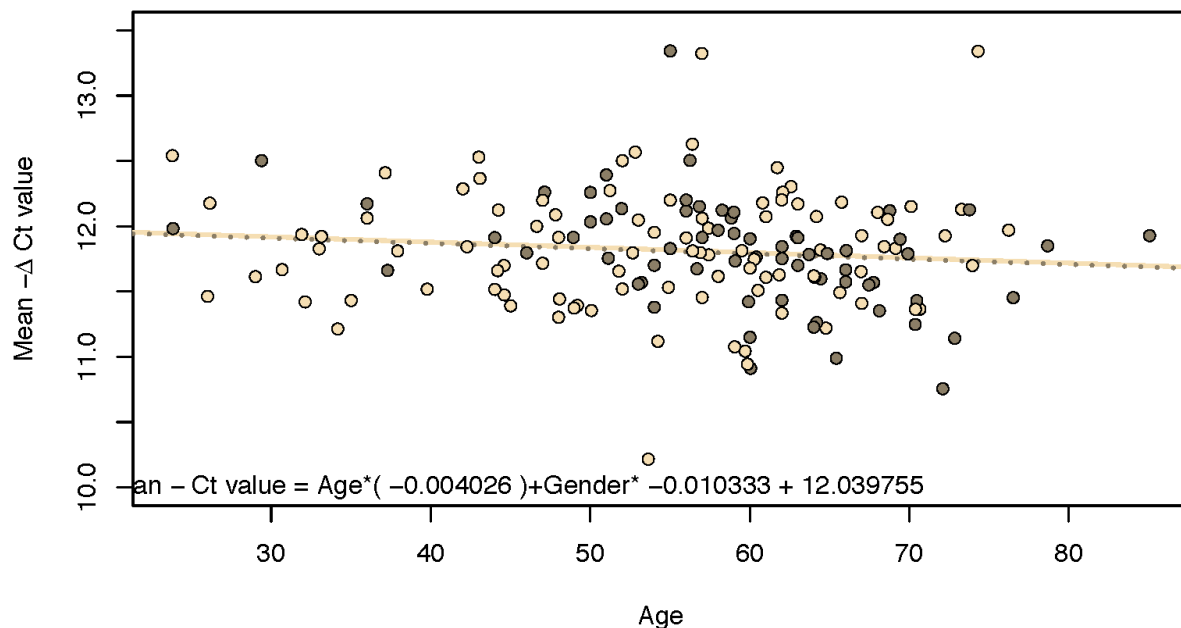


### 3.2.6 Análisis de longitud telomérica por medio de PCR

La longitud de los telómeros fue medida en todos los casos y controles resecuenciados en la plataforma Illumina provenientes de las cohortes de la Universidad de Leeds, SEARCH y WTCCC que portaban variantes en *POT1* de acuerdo al análisis inicial de Fluidigm. Se midió la longitud media telomérica relativa por medio de SYBR Green RT-PCR utilizando una versión de protocolos de Q-PCR publicados anteriormente (Cawthon 2002; McGrath et al. 2007) y que han sido modificados y adaptados por nuestros colaboradores (Pooley et al. 2010). Se extrajo ADN genómico de sangre periférica y se calculó la longitud telomérica por medio del cociente del valor de la fluorescencia detectada de la amplificación de unidades de repetición telomérica (TEL) y del valor de la fluorescencia de la amplificación de una secuencia referencia de una sola copia del gen HBB ( $\beta$  globina) (CON). Las reacciones para el control y para los telómeros se realizaron por separado, con cuatro réplicas técnicas por muestra. Para cada experimento, se registró el ciclo de PCR en el cual la reacción cruzó un umbral de fluorescencia predeterminado (valor Ct). La longitud telomérica se determinó por medio del promedio de las diferencias de valores Ct,  $\Delta Ct = Ct\ TEL - Ct\ CON$ , de las cuatro réplicas técnicas por muestra, como se ha publicado en otros trabajos (Robles-Espinoza et al. 2014; Pooley et al. 2010).

Para el análisis, las muestras con  $Ct\ TEL > 18$ ,  $Ct\ CON > 29$  o con  $Ct\ CON > 2$  desviaciones estándar lejos de la media fueron eliminadas y consideradas como reacciones fallidas. La información de edad y sexo para las muestras de la cohorte WTCCC no se encontraba disponible. Después de eliminar aquellas muestras control para las cuales no se encontraba información de edad y sexo, o aquellas para las cuales fallaron las réplicas técnicas del experimento, quedaron para análisis 162 muestras, las cuales utilizamos para realizar un modelo lineal. Este modelo lineal se realizó ajustando por sexo y edad en el momento del diagnóstico para todos los individuos que no portaban variantes clasificadas en el Grupo 1, ya sea casos de melanoma o controles (**Figura 17**). Los residuales de este modelo lineal fueron utilizados para crear una distribución de longitud telomérica para esta cohorte. El ajuste por sexo y edad en el momento del diagnóstico para los individuos portadores de variantes pertenecientes al Grupo 1 se hizo de manera separada con los mismos parámetros calculados para la distribución poblacional, y las medias de estas

mediciones fueron comparadas con la distribución poblacional (**Figura 18**). Cuando realizamos este análisis, notamos que existió gran variabilidad en las réplicas técnicas para las mediciones de longitud telomérica para los individuos con variantes en el Grupo 1 (**Figura 19**).



**Figura 17**

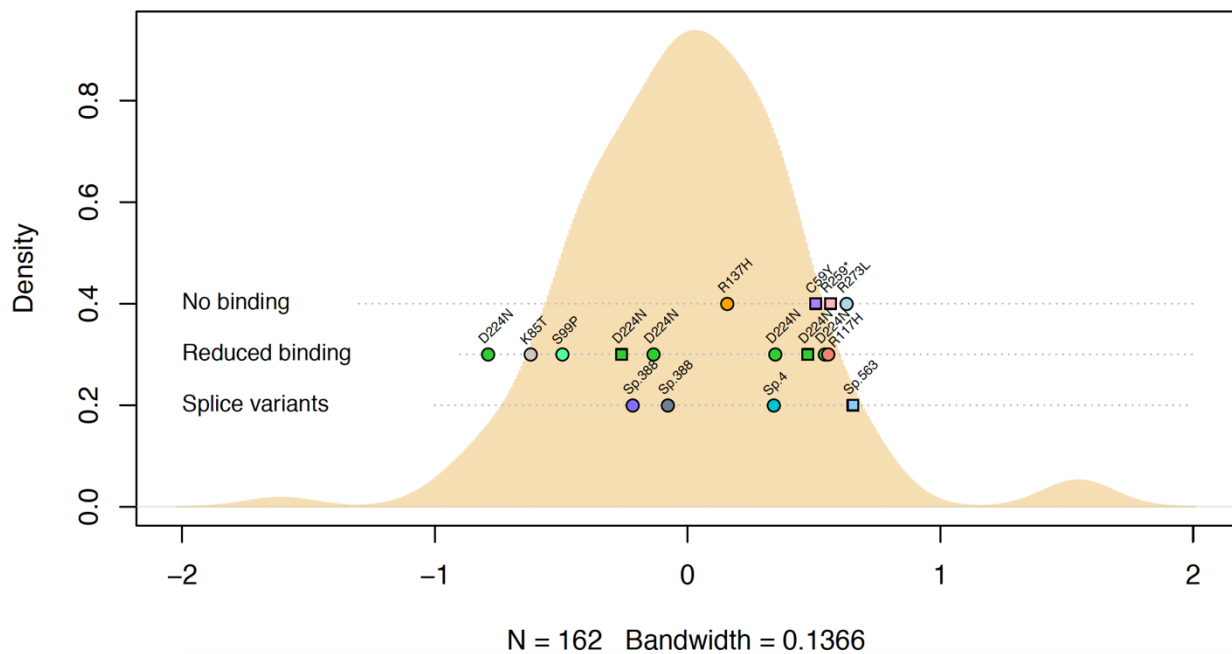
**Modelo lineal ajustando por sexo y edad al momento del diagnóstico para muestras de individuos que no eran portadores de variantes clasificadas en el Grupo 1.** Cada punto representa la medida media de cada individuo. Color beige: mujer, café: hombre. Se muestran las líneas de regresión para cada sexo, con el mismo esquema de color. Esta figura fue realizada con ayuda de Irving Simonin Wilmer, estudiante de doctorado de la Dra. Robles-Espinoza.

### 3.3 Resultados

En total, y como fue mencionado en los Métodos, fueron identificadas 43 variantes con predicción de alteración de la estructura de la proteína en POT1 por medio de secuenciación Fluidigm, seguido de validación por medio de captura exónica y secuenciación por medio de tecnologías Illumina o por medio de secuenciación por capilar (**Figura 16, Tabla 5**).

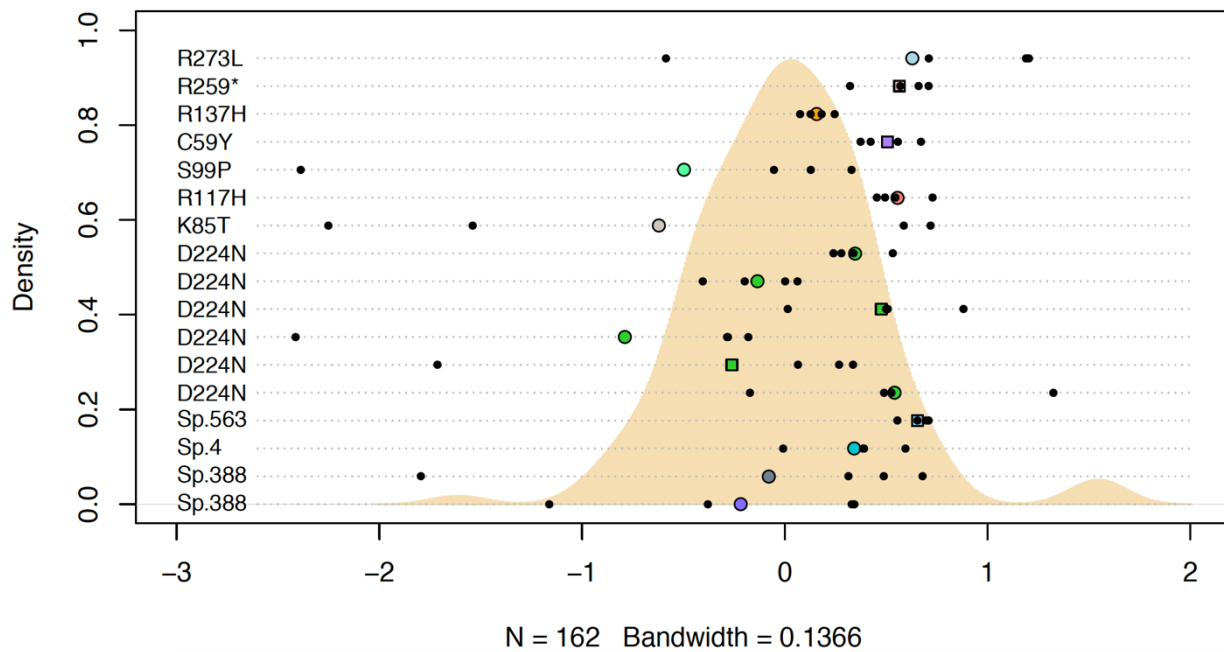
Para determinar si las variantes detectadas podrían potencialmente afectar la regulación telomérica, analizamos la capacidad de unión a ssDNA de proteínas POT1, cada una conteniendo una de las variantes de cambio de aminoácido o de codón de paro, traducidas *in vitro* por medio

de experimentos EMSA (Un total de 38 variantes, **Tabla 5**). Nuestros resultados indican que cuatro variantes completamente destruyen la capacidad de la proteína para formar el complejo POT1-ssDNA (C59Y, R137H, L259\* y R273L) mientras que otras cinco podrían resultar en una reducción de la fuerza de esta interacción (K39N, K85T, S99P, R117H, y D224N) (**Figura 20**). Es importante mencionar que, como se esperaba, todas las variantes que alteran la capacidad de POT1 de unirse a ssDNA se encuentran en los dominios funcionales OB. Cabe destacar que de éstas nueve variantes, siete fueron predichas como patogénicas por medio de los análisis realizados con VCF/Plotein mientras que dos fueron clasificadas como aquéllas con información conflictiva (**Tabla 6**). Asimismo, de todas las variantes clasificadas como patogénicas por VCF/Plotein, tres de éstas no mostraron diferencias en su capacidad de unión a ssDNA, indicando que la integración de más información, o quizá la evaluación de otras de las interacciones de POT1 podrían ayudar a refinar las predicciones de nuestra herramienta.



**Figura 18**

**Longitudes teloméricas de portadores de variantes en POT1 clasificadas en el Grupo 1 mostradas sobre una distribución de longitudes teloméricas de casos y controles de melanoma que no portan estas variantes.** En color beige se muestra la distribución de los residuales del modelo lineal de longitudes medias teloméricas para individuos sin variantes patogénicas en POT1 o que sólo portaban variantes clasificadas en los Grupos 2 y 3, con o sin melanoma. Las longitudes teloméricas ajustadas por sexo y edad al momento del diagnóstico se muestran encima de acuerdo al tipo de variante (*No binding*, sin unión a ssDNA, *reduced binding*, con unión reducida a ssDNA, de acuerdo a los experimentos EMSA, y variantes en sitios de corte y empalme). Los casos se muestran en círculos y los controles en cuadrados. Los números negativos en el eje X indican longitudes más cortas, mientras que números positivos longitudes más largas.

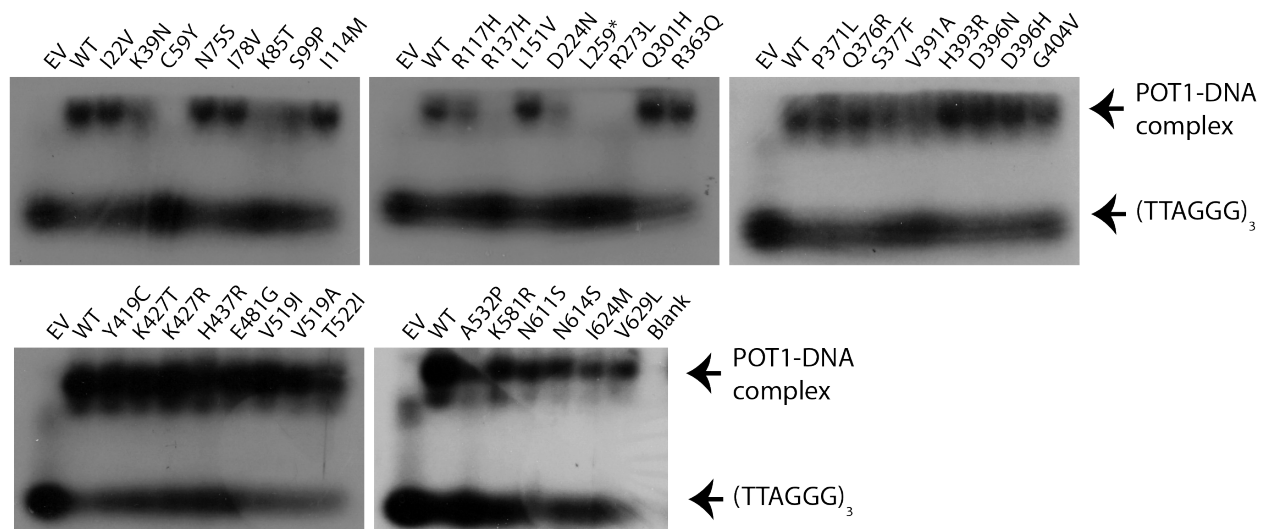


**Figura 19**

**Variabilidad de las réplicas técnicas de las mediciones de las longitudes teloméricas de los individuos portadores de variantes de POT1 clasificadas en el Grupo 1.** Los residuales de cada medición réplica para los portadores de variantes clasificadas en el Grupo 1 se muestran como puntos negros, con la media de las mediciones en colores. La distribución de las longitudes teloméricas es la misma que aquella mostrada en la Figura 18. Los números negativos en el eje X indican longitudes más cortas, mientras que números positivos longitudes más largas. Esta figura fue hecha por Irving Simonin Wilmer, estudiante de doctorado en el grupo de la Dra. Robles-Espinoza.

Clasificamos a las variantes en tres grupos de acuerdo con su patogenicidad por el experimento EMSA: Las variantes del Grupo 1 fueron aquellas que abolieron por completo la formación del complejo POT1-ssDNA, o aquellas con consecuencias drásticas (cambio de marco de lectura o en sitios de corte y empalme). Decidimos incluir las variantes con capacidad de unión reducida en esta categoría dada su conservación en distintas especies (**Figura 21**) y la evidencia existente de que éstas podrían ser patogénicas (R117H (Calvete et al. 2015), D224N (Shi et al. 2014)). En total, 14/43 variantes fueron clasificadas en este grupo, con 10 de éstas dentro de los dominios OB (**Figura 16**). Las variantes clasificadas en el Grupo 2 fueron aquellas predichas como deletéreas y probablemente dañinas por los algoritmos SIFT (Ng and Henikoff 2003) y PolyPhen-2 (Adzhubei, Jordan, and Sunyaev 2013), pero que no alteraron la formación del complejo POT1-ssDNA de acuerdo a los experimentos EMSA (4/43 variantes). Estas variantes podrían afectar la función de la proteína de otras maneras que no hemos probado en este estudio, por ejemplo, su capacidad

de interactuar con otras proteínas. El resto de las variantes (25/43) fueron clasificadas en el Grupo 3. Para el Grupo 1, 15 casos fueron portadores de una de estas variantes (0.51%), mientras que 8 controles fueron portadores (0.24%) ( $P$  valor=0.095, prueba exacta de Fisher con dos colas). Para los Grupos 1+2, 22 casos fueron portadores de variantes (0.75%), así como 14 controles ( $P$  valor=0.096, prueba exacta de Fisher con dos colas). Finalmente, para el Grupo 3, 127 casos (4.3%) fueron portadores, lo mismo que 151 controles (4.6%) ( $P$  valor=0.66, prueba exacta de Fisher con dos colas). Entonces, mientras que aproximadamente el doble de casos fueron portadores de variantes patogénicas en *POT1* en comparación con los controles, esta diferencia no fue estadísticamente significativa (siguiendo las convenciones de la comunidad científica) ya que estas variantes tienen frecuencias alélicas muy bajas en la población.

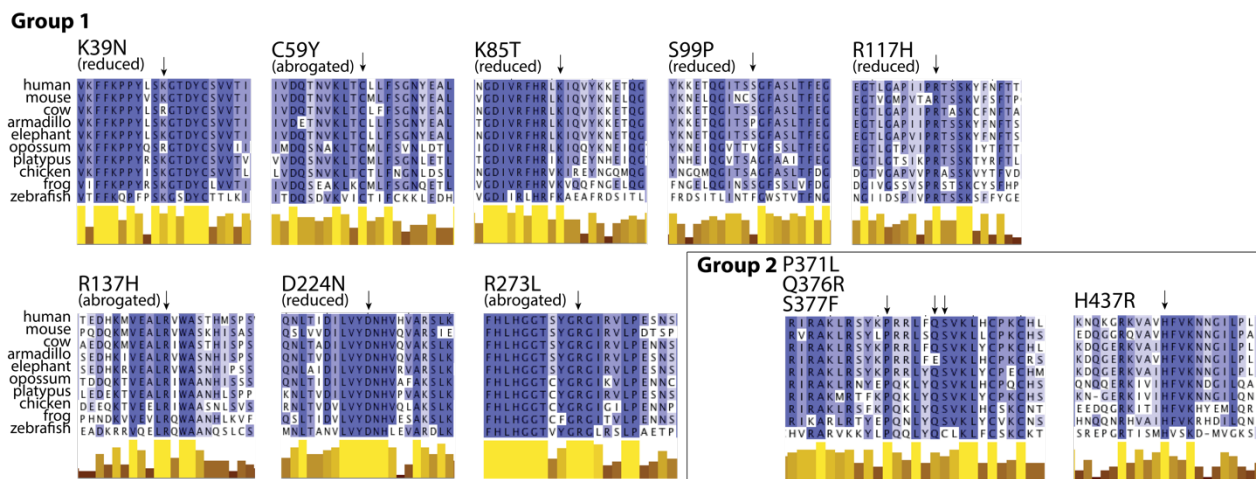


**Figura 20**

**Experimentos EMSA probando la capacidad de unión a ssDNA de proteínas mutantes *POT1* traducidas *in vitro*.** EV: Vector vacío (control negativo). WT: Proteína tipo silvestre (*wild type*). La secuencia de las proteínas utilizada como base para introducir las variantes detectadas se basaron en el transcrito de *POT1* ENST00000357628 de la base de datos Ensembl. (TTAGGG)<sub>3</sub> indica la secuencia del ssDNA (TTAGGGTTAGGGTTAGGG, imitando la secuencia telomérica). Este experimento fue realizado por nuestro colaborador Chi Wong, en el Wellcome Sanger Institute.

Subsecuentemente, buscamos determinar si las variantes detectadas tenían algún efecto en la regulación telomérica. Para esto, se realizaron mediciones de longitud telomérica de sangre periférica por medio de PCR de casos de melanoma y controles poblacionales que fueron portadores de las variantes detectadas. Después de ajustar estas longitudes por sexo y edad al momento del diagnóstico (**Figura 17**), pudimos observar que las longitudes teloméricas de los

individuos con variantes que interrumpían la formación del complejo POT1-ssDNA se encontraban en la región de la distribución que contenía las longitudes más elevadas (**Figura 18**, variantes marcadas como ‘no binding’). También observamos que algunos individuos con variantes en sitios de corte y empalme o variantes con capacidad reducida de unión a ssDNA también mostraban longitudes más largas (e.g. R117H, Sp. 563), pero no todos mostraron este comportamiento (e.g. K85T, S99P, Sp. 388). Los individuos portadores de variantes en D224N mostraron longitudes teloméricas repartidas en toda la distribución (**Figura 18**). Sin embargo, observamos una alta variabilidad en las réplicas técnicas de algunos de estos estimados, lo cual no fue el caso con aquellas variantes en la categoría de ‘no binding’ (**Figura 19**).



**Figura 21**  
**Conservación en distintas especies de residuos con mutaciones en POT1 en este estudio.** Los aminoácidos en los cuales se encontraron variantes se indican con una flecha. Se muestran las variantes clasificadas en los Grupos 1 y 2. Las posiciones de los alineamientos se muestran coloreadas de acuerdo al porcentaje de identidad: entre más oscuro se encuentre un aminoácido, mayor es la conservación a través del tiempo evolutivo. Las columnas amarillas representan la conservación en distintas especies del aminoácido en esa posición: Mientras más amarilla y más alta sea una columna, mayor la conservación).

### 3.4 Discusión

Aunque *POT1* parece ser uno de los genes que confiere mayor riesgo a padecer melanoma, con aproximadamente el 2-4% de familias sin mutaciones en *CDKN2A/CDK4* siendo portadoras de variantes patogénicas en este gen, su contribución al riesgo de desarrollar este tipo de cáncer en la población en general parecer ser menor, ya que solamente aproximadamente 0.5% de los casos en el contexto de la población en general parece portar variantes patogénicas en este gen.

De manera notable, cuando se estudian casos en el contexto de la población en general, no se espera que los individuos porten variantes de línea germinal de alta penetrancia (ya que si no posiblemente serían detectados en estudios familiares). En estos individuos, la influencia genética en el riesgo a padecer melanoma es muy probablemente dominada por factores poligénicos, incluyendo alelos de *MC1R*, los cuales están asociados al cabello pelirrojo, la presencia de pecas y una capacidad reducida de bronceado. De manera similar a lo que reportamos en este estudio, mientras que la frecuencia de alelos patogénicos de *CDKN2A* en melanoma familiar ronda el 20-40%, la frecuencia de estas variantes en cohortes poblacionales es aproximadamente 2% (Harland et al. 2014). Asimismo, cuando reunimos nuestra cohorte control, compuesta de grupos incluyendo la colección de ADN de WTCCC, no contamos con información para excluir a aquellos individuos con un diagnóstico previo de melanoma, cuya frecuencia se estima cerca del 1/50.

Los cálculos de longitud telomérica realizados en este estudio confirmaron las asociaciones ya conocidas de variantes a telómeros más largos (por ejemplo, R273L (Robles-Espinoza et al. 2014) y R117H (Shi et al. 2014; Calvete et al. 2015) ,y encontraron nuevas asociaciones con otras variantes patogénicas (C59Y y un portador de Sp. 563, reportada originalmente en (Robles-Espinoza et al. 2014)), pero para otras variantes esta asociación no es tan clara. Las razones de esta observación pueden ser varias: Podría reflejar la naturaleza no cuantitativa de los experimentos EMSA (por ejemplo, podría ser el caso que las variantes K85T y S99P en realidad tienen afinidades al ssDNA comparables a la proteína tipo silvestre, pero el experimento muestra variabilidad técnica), o podría reflejar la complejidad de medir longitudes teloméricas en muestras de ADN que han sido archivadas por un periodo de tiempo largo. De acuerdo con esta hipótesis, mientras que las réplicas técnicas para las mediciones de los portadores de muchas de estas variantes (*e.g.* R259\*, R137H, C59Y, R117H, y Sp. 563) fueron muy similares entre ellas, hubo algunas otras (R273L, S99P, K85T y Sp. 388) para las cuales las réplicas mostraron alta variabilidad entre ellas y por lo tanto la longitud calculada podría no ser confiable (**Figura 19**). Particularmente, para R273L, nuestro grupo de investigación y colaboradores mostraron en un experimento independiente que esta variante se asocia con telómeros largos (Robles-Espinoza

et al. 2014), y de esta manera tres de las cuatro réplicas técnicas muestran mediciones largas para esta variante. En este estudio, identificamos seis portadores de la variante D224N, la cual ha sido reportada como patogénica anteriormente (Shi et al. 2014), pero cuatro de estos portadores mostraron alta variabilidad en las mediciones teloméricas (**Figura 19**), lo cual vuelve complicada la interpretación en el presente experimento. También observamos algunas mediciones que podrían diferir de aquéllas reportadas previamente, por ejemplo, la longitud estimada para un portador de la variante R117H ha sido reportada como ligeramente más larga que la de controles por otro grupo (Shi et al. 2014) mientras que en nuestro estudio esta longitud cae dentro de la distribución poblacional. Esto también podría ser indicativo de los muchos otros mecanismos, incluyendo otras variantes genéticas presentes en estos individuos y otros factores de estilo de vida, por los cuales podría verse afectada la longitud telomérica. También observamos un incremento en la longitud telomérica para aquellos individuos control (sin indicios de melanoma) portadores de variantes patogénicas, lo cual podría indicar que otros factores adicionales son necesarios para el desarrollo de este tipo de cáncer.

Aunque en este estudio intentamos identificar alelos patogénicos de *POT1* por medio de ensayos EMSA de unión proteína-ssDNA, la función de las proteínas con variantes fuera de los dominios OB también podría estar afectada por otros mecanismos. Por ejemplo, un estudio publicado previamente reportó que la variante A532P muestra una capacidad reducida de unión a la proteína ACD, la cual es también parte del complejo protector de telómeros, y la que podría llevar a una desregulación en los mecanismos de mantenimiento de telómeros (J. Liu et al. 2015).

En conclusión, llevamos a cabo el estudio más largo reportado hasta el momento de casos y controles de melanoma enfocados en variantes del gen *POT1*. Hemos demostrado que algunas de estas variantes interrumpen la formación del complejo POT1-ssDNA y que, de manera importante, también se localizan en individuos que no han desarrollado melanoma, lo cual indica que se requieren factores adicionales, ya sea genéticos o de estilo de vida, para el desarrollo de este tipo de cáncer. También hemos demostrado que aquellas variantes que destruyen por completo la formación del complejo POT1-ssDNA se asocian de manera consistente con



mediciones más elevadas de longitud telomérica, pero que la relación entre las variantes asociadas a una reducción moderada en la afinidad de la unión de POT1 y ssDNA y la longitud de telómeros es más compleja. También mencionamos las complicaciones de realizar mediciones de longitud telomérica en muestras de ADN que han sido archivadas por grandes periodos de tiempo, mostrando la necesidad de realizar réplicas técnicas para obtener un estimado más exacto.

Los datos presentados en este estudio contribuyen al conocimiento sobre la influencia de las variantes en *POT1* presentes en casos de la población en general, la cual es muy limitada dada nuestra estimación de que solamente aproximadamente 1 en 200 casos en el Reino Unido porta una variante patogénica en este gen. También esperamos que el catálogo de variantes presentado aquí y los experimentos biológicos asociados contribuyan a complementar el conocimiento necesario para brindar consejo genético informado a pacientes con historia familiar de melanoma.

### 3.5 Equipo de trabajo y agradecimientos

Este proyecto fue una colaboración multinacional y multidisciplinaria dada la gran cantidad de individuos analizados y los diferentes tipos de experimentos realizados. Particularmente, quiero agradecer al Dr. David Adams por el financiamiento de una parte de los experimentos aquí mostrados, así como a Mark Harland (Universidad de Leeds) por su ayuda con el manejo de la cohorte de casos y controles, a Chi Wong (Instituto Wellcome Sanger) por realizar los experimentos EMSA y a Karen Pooley (Universidad de Cambridge) por realizar los experimentos de mediciones teloméricas. Los análisis bioinformáticos de análisis y filtrado de variantes, así como de comparaciones de secuencias a través de la historia evolutiva, y análisis de datos de longitudes teloméricas fueron realizados por mí, con consejo de Mark Iles (Universidad de Leeds), Irving Simonin Wilmer (LIIGH-UNAM), D. Timothy Bishop (Universidad de Leeds) y mi tutora, la Dra. Robles Espinoza (LIIGH-UNAM). Los resultados de este estudio están siendo preparados para su envío a una revista científica indizada y revisada por pares.

Asimismo, quiero agradecer la ayuda de Abigayl Hernández, Eglee Lomelín, Alejandra Castillo, Carina Uribe y Jair S. García-Sotelo, del LIIGH-UNAM, y de Luis A. Aguilar, Carlos S. Flores y Alejandro de León del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS-UNAM). Los experimentos aquí descritos fueron financiados por los donativos Wellcome Trust Ref. 204562/Z/16/Z de la Dra. Robles Espinoza, y Medical Research Council Ref. MR/S01473X/1 a la Dra. Robles Espinoza y el Dr. D. J. Adams. Asimismo, fue financiado por un donativo de Cancer Research UK al Dr. D. J. Adams.

## 4 Conclusiones y estudios futuros

En esta tesis, partimos de la necesidad que tienen muchos investigadores que se encuentran analizando datos de secuenciación masiva y un fenotipo de interés: Encontrar aquellas variantes genéticas que se asocian al fenotipo bajo estudio. En el caso particular de este trabajo, esto consistía en datos de secuenciación de un gen en particular, *POT1*, en más de 6,000 individuos de cohortes de casos y controles de melanoma originarios del Reino Unido.

De esta manera, y al comenzar los pasos bioinformáticos de análisis y filtrado de variantes, me di cuenta de que este proceso era complicado y requería una formación particular en informática y biología, así como un equipo multidisciplinario. Esto significa que este tipo de análisis no puede ser realizado por cualquier biólogo, investigador o médico, que muchas veces son quienes más necesitan hacer estos estudios. Es por esto por lo que decidí primero crear una herramienta de software que fuera fácil de usar, rápida, segura, y que pudiera integrar una gran cantidad de información de bases de datos externas para facilitar este proceso. Este esfuerzo culminó en la publicación de la herramienta basada en web **VCF/Plotein** (Ossio et al. 2019), la cual es capaz de analizar archivos en formato VCF (el archivo de salida estándar de cualquier algoritmo alineador de lecturas moderno) y mostrar, de manera interactiva, los genes que se encuentran mutados y varios de sus atributos para facilitar su priorización, así como las características de las variantes encontradas como consecuencia en la estructura primaria de la proteína, su frecuencia alélica en bases de datos poblacionales, su presencia en bases de datos clínicas o de cáncer, así como predicciones de patogenicidad de los algoritmos más utilizados. Esta herramienta también permite al investigador realizar el filtrado de variantes de acuerdo con los criterios que le parezcan adecuados de acuerdo con el tipo de estudio, e imprimir las imágenes con alta calidad para compartir con colaboradores o facilitar la preparación de figuras para presentaciones o publicaciones. De manera importante, esta herramienta también es segura, ya que no guarda ni envía información sobre el proyecto, los nombres de las muestras o las anotaciones previamente encontradas en el archivo.

Al probar nuestra herramienta primero con un Caso de Uso, basado en el gen *BAP1* y con un archivo VCF previamente publicado (O'Shea et al. 2017), nos percatamos de que contenía información suficiente para poder predecir, con una estrategia razonable basada en la función del gen, cuáles variantes resultarían en una reducción de su función por medio de alteraciones en el dominio funcional de la proteína. Subsecuentemente, aplicamos este tipo de estrategia a nuestro estudio de melanoma con 6,227 casos y controles, prediciendo que 15/43 variantes en el gen *POT1* eran patogénicas dada su localización en los dominios funcionales, su consecuencia o su presencia o ausencia en bases de datos poblacionales, clínicas o de cáncer (**Tabla 6**). De éstas, y como parte de una prueba de principio, realizamos junto con ayuda de nuestros colaboradores experimentos para comprobar si la función principal de POT1, que es la unión de la proteína al ADN telomérico para auxiliar en su protección y la regulación de su longitud, se veía afectada por estas variantes. Estos experimentos demostraron que 7/10 variantes que la estrategia de VCF/Plotein predijo como patogénica y que fueron evaluadas en estos experimentos sí interrumpieron o redujeron la capacidad del POT1 de unirse al ssDNA. Dos de las cuatro variantes con información conflictiva redujeron la afinidad del complejo POT1-ssDNA. Asimismo, ninguna de las variantes predichas como no patogénicas afectó la formación de este complejo. Esta herramienta, o módulos de ésta, también fue utilizada en otras dos publicaciones a las cuales contribuí (Walpole et al. 2018; Nguyen et al. 2018).

Estos resultados nos indican que, aunque VCF/Plotein es una buena herramienta que puede ayudar a priorizar variantes para estudios funcionales subsecuentes, para lograr una mejor priorización y predicción otras funciones son necesarias. Proyectos futuros en el laboratorio o quizá en otros grupos de investigación podrían centrarse en agregar una pantalla que muestre una visión integral del archivo VCF, con aquellos genes que se encuentran más frecuentemente mutados, las muestras con más mutaciones (en el caso de estudios de tumores, por ejemplo), y quizá un análisis de firmas mutacionales previo a la elección de un gen para examinar sus variantes en detalle. No cabe duda de que un enfoque integral en el desarrollo de este tipo de herramientas beneficiará a la comunidad científica y a la aplicación de las tecnologías de secuenciación masiva a la práctica clínica.

Respecto a la conclusión biológica de esta tesis, por medio del algoritmo descrito anteriormente, y una colaboración internacional e interdisciplinaria, logramos identificar nuevas variantes del gen *POT1* presentes en la población británica que impiden su unión al ADN telomérico y que, dado el conocimiento que hemos acumulado en la última década (Robles-Espinoza et al. 2014; Shi et al. 2014; Calvete et al. 2015; Ramsay et al. 2013; Calvete et al. 2017; Wong et al. 2019), conllevan alto riesgo a desarrollar melanoma.

La manera en la que estas variantes ejercen su función es aún desconocida. Se ha propuesto un modelo en el que estas variantes podrían actuar de un modo dominante negativo (Loayza and De Lange 2003; Calvete et al. 2015), en el cual la presencia de una proteína mutada interrumpe la función del complejo protector aunque se encuentre presente la proteína tipo silvestre. Algunas mutaciones particulares de *POT1* podrían mostrar este comportamiento por medio de tener una afinidad mayor al ssDNA o al complejo protector, secuestrándolo, mientras que otras variantes podrían afectar la interacción de *POT1* con otros miembros de este complejo como ACD o TEF2IP. Sin embargo, aunque los detalles de este mecanismo aún están siendo explorados, se ha demostrado que la presencia de una proteína mutada, aún sin pérdida del alelo silvestre, conlleva a daño telomérico en experimentos *in vitro* (Loayza and De Lange 2003). Este comportamiento refleja lo que se observa en los pacientes con melanoma familiar portadores de estas variantes, los cuales hasta el momento son todos heterocigotos (Robles-Espinoza et al. 2014; Shi et al. 2014). Esta hipótesis también tiene soporte de análisis de secuencias tumorales, en las cuales se ha encontrado que por lo general *POT1* no muestra pérdida de heterocigidad (Ramsay et al. 2013; Calvete et al. 2015), con alelos silvestres presentes aún cuando existe desregulación telomérica.

No obstante que no sepamos los detalles de su función biológica, se ha observado una frecuencia mayor de varios tipos de tumores en personas portadoras de alelos patogénicos en *POT1* (Calvete et al. 2017; Shen et al. 2020), por lo que ha sido incluido en varios paneles de genes utilizados para brindar consejo genético a individuos portadores. Es por esto por lo que es de vital

importancia la identificación de aquellas variantes que afecten la función del gen, lo cual requiere analizar un número elevado de muestras y llevar a cabo diversos experimentos biológicos. Con este trabajo, esperamos haber contribuido de manera importante a asignar una clasificación patogénica a aquellas variantes presentes en la población británica, y por ende, a mejorar la calidad de la información disponible utilizada por consejeros y asesores genéticos.

Finalmente, aunque este estudio se llevó a cabo en una población del Reino Unido debido principalmente a los recursos disponibles y las colaboraciones ya establecidas, considero que es de vital importancia realizar este tipo de trabajos en la población mexicana para aumentar su relevancia y aplicabilidad en México. Ya que en nuestro país el tipo de melanoma más común, llamado melanoma lentiginoso acral, es de muy baja frecuencia en poblaciones de ascendencia genética europea (Ossio et al. 2017), estudios futuros muy probablemente se enfoquen en estudiar no solamente la arquitectura genética del riesgo al melanoma y otros tipos de cáncer en nuestro país, sino a descifrar los elementos genómicos y ambientales que contribuyen a su desarrollo. Esfuerzos actuales como los llevados a cabo en el laboratorio de la Dra. Robles Espinoza y otros en la UNAM y otras instituciones en México como el Instituto Nacional de Medicina Genómica (INMEGEN) y el Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO) sin duda contribuirán a este esfuerzo. El establecimiento del Biobanco Mexicano (MX Biobank, <https://mxbiobankproject.org/>), liderado por científicos del LANGEBIO, es un paso importante en el establecimiento de la arquitectura genómica de la población mexicana y podrá ser utilizado para poner en contexto nuevas variantes genéticas que se encuentren en estudios de riesgo genético a fenotipos, como es el caso del abordado en la presente tesis.

## 5 Bibliografía

- Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. 2013. "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2." *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 07 (January): Unit7.20. <https://doi.org/10.1002/0471142905.hg0720s76>.
- Affar, El Bachir, and Michele Carbone. 2018. "BAP1 Regulates Different Mechanisms of Cell Death." *Cell Death & Disease* 9 (12): 1–3. <https://doi.org/10.1038/s41419-018-1206-5>.
- Alemán, Alejandro, Francisco Garcia-Garcia, Francisco Salavert, Ignacio Medina, and Joaquín Dopazo. 2014. "A Web-Based Interactive Framework to Assist in the Prioritization of Disease Candidate Genes in Whole-Exome Sequencing Studies." *Nucleic Acids Research* 42 (Web Server issue): W88–93. <https://doi.org/10.1093/nar/gku407>.
- Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500 (7463): 415–21. <https://doi.org/10.1038/nature12477>.
- Aoude, Lauren G., Antonia L. Pritchard, Carla Daniela Robles-Espinoza, Karin Wadt, Mark Harland, Jiyeon Choi, Michael Gartside, et al. 2015. "Nonsense Mutations in the Shelterin Complex Genes ACD and TERF2IP in Familial Melanoma." *Journal of the National Cancer Institute* 107 (2). <https://doi.org/10.1093/jnci/dju408>.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25 (1): 25–29. <https://doi.org/10.1038/75556>.
- Atanda, A. T., and A. B. Umar. 2012. "Clinical and Histological Features of Melanoma in Adult Nigerians." *West African Journal of Medicine* 31 (3): 149–53.
- Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. <https://doi.org/10.1038/nature15393>.
- Bainbridge, Matthew N., Georgina N. Armstrong, M. Monica Gramatges, Alison A. Bertuch, Shalini N. Jhangiani, Harsha Doddapaneni, Lora Lewis, et al. 2015. "Germline Mutations in Shelterin Complex Genes Are Associated with Familial Glioma." *Journal of the National Cancer Institute* 107 (1): 384. <https://doi.org/10.1093/jnci/dju384>.
- Bamshad, Michael J., Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson, and Jay Shendure. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews. Genetics* 12 (11): 745–55. <https://doi.org/10.1038/nrg3031>.
- Bao, Riyue, Lei Huang, Jorge Andrade, Wei Tan, Warren A. Kibbe, Hongmei Jiang, and Gang Feng. 2014. "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing." *Cancer Informatics* 13s2 (January): CIN.S13779. <https://doi.org/10.4137/CIN.S13779>.
- Bell, Callum J., Darrell L. Dinwiddie, Neil A. Miller, Shannon L. Hateley, Elena E. Ganusova, Joann Mudge, Ray J. Langley, et al. 2011. "Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing." *Science Translational Medicine* 3 (65): 65ra4. <https://doi.org/10.1126/scitranslmed.3001756>.

- Box, N. F., D. L. Duffy, W. Chen, M. Stark, N. G. Martin, R. A. Sturm, and N. K. Hayward. 2001. "MC1R Genotype Modifies Risk of Melanoma in Families Segregating CDKN2A Mutations." *American Journal of Human Genetics* 69 (4): 765–73. <https://doi.org/10.1086/323412>.
- Calvete, Oriol, Pablo Garcia-Pavia, Fernando Domínguez, Gaele Bougeard, Kristin Kunze, Andreas Braeuninger, Alex Teule, et al. 2017. "The Wide Spectrum of POT1 Gene Variants Correlates with Multiple Cancer Types." *European Journal of Human Genetics: EJHG* 25 (11): 1278–81. <https://doi.org/10.1038/ejhg.2017.134>.
- Calvete, Oriol, Paula Martinez, Pablo Garcia-Pavia, Carlos Benitez-Buelga, Beatriz Paumard-Hernández, Victoria Fernandez, Fernando Dominguez, et al. 2015. "A Mutation in the POT1 Gene Is Responsible for Cardiac Angiosarcoma in TP53-Negative Li-Fraumeni-like Families." *Nature Communications* 6 (September): 8383. <https://doi.org/10.1038/ncomms9383>.
- Cancer Genome Atlas Network. 2015. "Genomic Classification of Cutaneous Melanoma." *Cell* 161 (7): 1681–96. <https://doi.org/10.1016/j.cell.2015.05.044>.
- Cannon-Albright, L. A., D. E. Goldgar, L. J. Meyer, C. M. Lewis, D. E. Anderson, J. W. Fountain, M. E. Hegi, R. W. Wiseman, E. M. Petty, and A. E. Bale. 1992. "Assignment of a Locus for Familial Melanoma, MLM, to Chromosome 9p13-P22." *Science (New York, N.Y.)* 258 (5085): 1148–52. <https://doi.org/10.1126/science.1439824>.
- Cawthon, Richard M. 2002. "Telomere Measurement by Quantitative PCR." *Nucleic Acids Research* 30 (10): e47. <https://doi.org/10.1093/nar/30.10.e47>.
- Chiu, Christine, Molly Tebo, Jodie Ingles, Laura Yeates, Jonathan W. Arthur, Joanne M. Lind, and Christopher Semsarian. 2007. "Genetic Screening of Calcium Regulation Genes in Familial Hypertrophic Cardiomyopathy." *Journal of Molecular and Cellular Cardiology* 43 (3): 337–43. <https://doi.org/10.1016/j.yjmcc.2007.06.009>.
- Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6 (2): 80–92. <https://doi.org/10.4161/fly.19695>.
- Cust, Anne E., Kriti Mishra, and Marianne Berwick. 2018. "Melanoma - Role of the Environment and Genetics." *Photochemical & Photobiological Sciences: Official Journal of the European Photochemistry Association and the European Society for Photobiology* 17 (12): 1853–60. <https://doi.org/10.1039/c7pp00411g>.
- Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58. <https://doi.org/10.1093/bioinformatics/btr330>.
- Danecek, Petr, Christoffer Nellåker, Rebecca E McIntyre, Jorge E Buendia-Buendia, Suzannah Bumpstead, Chris P Ponting, Jonathan Flint, Richard Durbin, Thomas M Keane, and David J Adams. 2012. "High Levels of RNA-Editing Site Conservation amongst 15 Laboratory Mouse Strains." *Genome Biology* 13 (4): r26. <https://doi.org/10.1186/gb-2012-13-4-r26>.
- Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLOS Computational Biology* 6 (12): e1001025. <https://doi.org/10.1371/journal.pcbi.1001025>.



- Dessinioti, Clio, Christina Antoniou, Andreas Katsambas, and Alexander J. Stratigos. 2011. "Melanocortin 1 Receptor Variants: Functional Role and Pigmentary Associations." *Photochemistry and Photobiology* 87 (5): 978–87. <https://doi.org/10.1111/j.1751-1097.2011.00970.x>.
- Do, Ron, Sekar Kathiresan, and Gonçalo R. Abecasis. 2012. "Exome Sequencing and Complex Disease: Practical Aspects of Rare Variant Association Studies." *Human Molecular Genetics* 21 (R1): R1-9. <https://doi.org/10.1093/hmg/dds387>.
- Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74. <https://doi.org/10.1038/nature11247>.
- "Elasticsearch Reference [7.5] | Elastic." n.d. Learn/Docs/Elasticsearch/Reference/7.5. Accessed January 27, 2020. <https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html>.
- El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. "The Pfam Protein Families Database in 2019." *Nucleic Acids Research* 47 (D1): D427–32. <https://doi.org/10.1093/nar/gky995>.
- Elliott, Alison M., and Jan M. Friedman. 2018. "The Importance of Genetic Counselling in Genome-Wide Sequencing." *Nature Reviews Genetics* 19 (12): 735–36. <https://doi.org/10.1038/s41576-018-0057-3>.
- Erdmann, Friederike, Joannie Lortet-Tieulent, Joachim Schüz, Hajo Zeeb, Rüdiger Greinert, Eckhard W. Breitbart, and Freddie Bray. 2013. "International Trends in the Incidence of Malignant Melanoma 1953–2008—Are Recent Generations at Higher or Lower Risk?" *International Journal of Cancer* 132 (2): 385–400. <https://doi.org/10.1002/ijc.27616>.
- Ewing, Brent, and Phil Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." *Genome Research* 8 (3): 186–94. <https://doi.org/10.1101/gr.8.3.186>.
- Fujisawa, Yasuhiro, Shusuke Yoshikawa, Akane Minagawa, Tatsuya Takenouchi, Kenji Yokota, Hiroshi Uchi, Naoki Noma, et al. 2019. "Clinical and Histopathological Characteristics and Survival Analysis of 4594 Japanese Patients with Melanoma." *Cancer Medicine* 8 (5): 2146–56. <https://doi.org/10.1002/cam4.2110>.
- Gandini, Sara, Francesco Sera, Maria Sofia Cattaruzza, Paolo Pasquini, Orietta Picconi, Peter Boyle, and Carmelo Francesco Melchi. 2005. "Meta-Analysis of Risk Factors for Cutaneous Melanoma: II. Sun Exposure." *European Journal of Cancer* 41 (1): 45–60. <https://doi.org/10.1016/j.ejca.2004.10.016>.
- Ghissassi, Fatiha El, Robert Baan, Kurt Straif, Yann Grosse, Béatrice Secretan, Véronique Bouvard, Lamia Benbrahim-Tallaa, et al. 2009. "A Review of Human Carcinogens—Part D: Radiation." *The Lancet Oncology* 10 (8): 751–52. [https://doi.org/10.1016/S1470-2045\(09\)70213-X](https://doi.org/10.1016/S1470-2045(09)70213-X).
- Gilissen, Christian, Alexander Hoischen, Han G. Brunner, and Joris A. Veltman. 2011. "Unlocking Mendelian Disease Using Exome Sequencing." *Genome Biology* 12 (9): 228. <https://doi.org/10.1186/gb-2011-12-9-228>.
- Goldstein, David B., Andrew Allen, Jonathan Keebler, Elliott H. Margulies, Steven Petrou, Slavé Petrovski, and Shamil Sunyaev. 2013. "Sequencing Studies in Human Genetics: Design and

- Interpretation." *Nature Reviews. Genetics* 14 (7): 460–70. <https://doi.org/10.1038/nrg3455>.
- Guo, Jun, Shukui Qin, Jun Liang, Tongyu Lin, Lu Si, Xiaohong Chen, Zhihong Chi, et al. 2015. "Chinese Guidelines on the Diagnosis and Treatment of Melanoma (2015 Edition)." *Annals of Translational Medicine* 3 (21). <https://doi.org/10.3978/j.issn.2305-5839.2015.12.23>.
- Harland, Mark, Anne E. Cust, Celia Badenas, Yu-Mei Chang, Elizabeth A. Holland, Paula Aguilera, Joanne F. Aitken, et al. 2014. "Prevalence and Predictors of Germline CDKN2A Mutations for Melanoma Cases from Australia, Spain and the United Kingdom." *Hereditary Cancer in Clinical Practice* 12 (1): 20. <https://doi.org/10.1186/1897-4287-12-20>.
- Hart, Steven N., Patrick Duffy, Daniel J. Quest, Asif Hossain, Mike A. Meiners, and Jean-Pierre Kocher. 2016. "VCF-Miner: GUI-Based Application for Mining Variants and Annotations Stored in VCF Files." *Briefings in Bioinformatics* 17 (2): 346–51. <https://doi.org/10.1093/bib/bbv051>.
- Herrera González, Norma Estela, and Aramara Yasmín Aco Flores. 2010. "El melanoma en México." *Revista de Especialidades Médico-Quirúrgicas* 15 (3): 161–64.
- Horn, Susanne, Adina Figl, P. Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, et al. 2013. "TERT Promoter Mutations in Familial and Sporadic Melanoma." *Science (New York, N.Y.)* 339 (6122): 959–61. <https://doi.org/10.1126/science.1230062>.
- INEGI. 2018. "Características de Las Defunciones Registradas En México Durante 2017." <https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2018/EstSociodemo/DEFUNCIONES2017.pdf>.
- Iozumi, K., G. E. Hoganson, R. Pennella, M. A. Everett, and B. B. Fuller. 1993. "Role of Tyrosinase as the Determinant of Pigmentation in Cultured Human Melanocytes." *The Journal of Investigative Dermatology* 100 (6): 806–11. <https://doi.org/10.1111/1523-1747.ep12476630>.
- Jeng, Xinge Jessie, Zhongyin John Daye, Wenbin Lu, and Jung-Ying Tzeng. 2016. "Rare Variants Association Analysis in Large-Scale Sequencing Studies at the Single Locus Level." *PLOS Computational Biology* 12 (6): e1004993. <https://doi.org/10.1371/journal.pcbi.1004993>.
- Kamb, A., N. A. Gruis, J. Weaver-Feldhaus, Q. Liu, K. Harshman, S. V. Tavtigian, E. Stockert, R. S. Day, B. E. Johnson, and M. H. Skolnick. 1994. "A Cell Cycle Regulator Potentially Involved in Genesis of Many Tumor Types." *Science (New York, N.Y.)* 264 (5157): 436–40. <https://doi.org/10.1126/science.8153634>.
- Köhler, Sebastian, Leigh Carmody, Nicole Vasilevsky, Julius O. B. Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, et al. 2019. "Expansion of the Human Phenotype Ontology (HPO) Knowledge Base and Resources." *Nucleic Acids Research* 47 (D1): D1018–27. <https://doi.org/10.1093/nar/gky1105>.
- Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. "Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing." *Genome Biology* 20 (1): 117. <https://doi.org/10.1186/s13059-019-1720-5>.
- Landrum, Melissa J, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. "ClinVar: Improving Access to Variant Interpretations

- and Supporting Evidence.” *Nucleic Acids Research* 46 (Database issue): D1062–67. <https://doi.org/10.1093/nar/gkx1153>.
- Lange, Titia de. 2005. “Shelterin: The Protein Complex That Shapes and Safeguards Human Telomeres.” *Genes & Development* 19 (18): 2100–2110. <https://doi.org/10.1101/gad.1346005>.
- Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. <https://doi.org/10.1038/nmeth.1923>.
- Lappalainen, Ilkka, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif ur-Rehman, Gary Saunders, et al. 2015. “The European Genome-Phenome Archive of Human Data Consented for Biomedical Research.” *Nature Genetics* 47 (7): 692–95. <https://doi.org/10.1038/ng.3312>.
- Law, Matthew H., D. Timothy Bishop, Jeffrey E. Lee, Myriam Brossard, Nicholas G. Martin, Eric K. Moses, Fengju Song, et al. 2015. “Genome-Wide Meta-Analysis Identifies Five New Susceptibility Loci for Cutaneous Malignant Melanoma.” *Nature Genetics* 47 (9): 987–95. <https://doi.org/10.1038/ng.3373>.
- Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91. <https://doi.org/10.1038/nature19057>.
- Li, Heng. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93. <https://doi.org/10.1093/bioinformatics/btr509>.
- Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
- Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. <https://doi.org/10.1093/bioinformatics/btp352>.
- Liu, Dongping, and Yang Xu. 2011. “P53, Oxidative Stress, and Aging.” *Antioxidants & Redox Signaling* 15 (6): 1669–78. <https://doi.org/10.1089/ars.2010.3644>.
- Liu, Jinqiang, Clinton Yu, Xichan Hu, Jin-Kwang Kim, Jan C. Bierma, Hyun-Ik Jun, Scott D. Rychnovsky, Lan Huang, and Feng Qiao. 2015. “Dissecting Fission Yeast Shelterin Interactions via MICro-MS Links Disruption of Shelterin Bridge to Tumorigenesis.” *Cell Reports* 12 (12): 2169–80. <https://doi.org/10.1016/j.celrep.2015.08.043>.
- Lo, Jennifer A., and David E. Fisher. 2014. “The Melanoma Revolution: From UV Carcinogenesis to a New Era in Therapeutics.” *Science (New York, N.Y.)* 346 (6212): 945–49. <https://doi.org/10.1126/science.1253735>.
- Loayza, Diego, and Titia De Lange. 2003. “POT1 as a Terminal Transducer of TRF1 Telomere Length Control.” *Nature* 423 (6943): 1013–18. <https://doi.org/10.1038/nature01688>.
- Longo, Caterina, and Giovanni Pellacani. 2016. “Melanomas.” *Dermatologic Clinics* 34 (4): 411–19. <https://doi.org/10.1016/j.det.2016.05.004>.

- MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, et al. 2014. "Guidelines for Investigating Causality of Sequence Variants in Human Disease." *Nature* 508 (7497): 469–76. <https://doi.org/10.1038/nature13127>.
- Mailman, Matthew D., Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, et al. 2007. "The NCBI DbGaP Database of Genotypes and Phenotypes." *Nature Genetics* 39 (10): 1181–86. <https://doi.org/10.1038/ng1007-1181>.
- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53. <https://doi.org/10.1038/nature08494>.
- Maranhao, B., P. Biswas, J. L. Duncan, K. E. Branham, G. A. Silva, M. A. Naeem, S. N. Khan, et al. 2014. "ExomeSuite: Whole Exome Sequence Variant Filtering Tool for Rapid Identification of Putative Disease Causing SNVs/Indels." *Genomics* 103 (2–3): 169–76. <https://doi.org/10.1016/j.ygeno.2014.02.006>.
- Mardis, Elaine R. 2017. "DNA Sequencing Technologies: 2006–2016." *Nature Protocols* 12 (2): 213–18. <https://doi.org/10.1038/nprot.2016.182>.
- McGrath, Monica, Jason Y. Y. Wong, Dominique Michaud, David J. Hunter, and Immaculata De Vivo. 2007. "Telomere Length, Cigarette Smoking, and Bladder Cancer Risk in Men and Women." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 16 (4): 815–19. <https://doi.org/10.1158/1055-9965.EPI-06-0961>.
- McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. <https://doi.org/10.1101/gr.107524.110>.
- McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122. <https://doi.org/10.1186/s13059-016-0974-4>.
- McMaster, Mary L., Chongkui Sun, Maria T. Landi, Sharon A. Savage, Melissa Rotunno, Xiaohong R. Yang, Kristine Jones, et al. 2018. "Germline Mutations in Protection of Telomeres 1 in Two Families with Hodgkin Lymphoma." *British Journal of Haematology* 181 (3): 372–77. <https://doi.org/10.1111/bjh.15203>.
- Miller, Arlo J., and Martin C. Mihm. 2006. "Melanoma." *The New England Journal of Medicine* 355 (1): 51–65. <https://doi.org/10.1056/NEJMra052166>.
- Miller, Chase A., Yi Qiao, Tonya DiSera, Brian D'Astous, and Gabor T. Marth. 2014. "Bam.lobio: A Web-Based, Real-Time, Sequence Alignment File Inspector." *Nature Methods* 11 (12): 1189. <https://doi.org/10.1038/nmeth.3174>.
- Minikel, Eric Vallabh, and Daniel G. MacArthur. 2016. "Publicly Available Data Provide Evidence against NR1H3 R415Q Causing Multiple Sclerosis." *Neuron* 92 (2): 336–38. <https://doi.org/10.1016/j.neuron.2016.09.054>.
- Nasti, Tahseen H., and Laura Timares. 2015. "MC1R, Eumelanin and Pheomelanin: Their Role in Determining the Susceptibility to Skin Cancer." *Photochemistry and Photobiology* 91 (1): 188–200. <https://doi.org/10.1111/php.12335>.
- Newton-Bishop, Julia A., Yu-Mei Chang, Mark M. Iles, John C. Taylor, Bert Bakker, May Chan, Susan Leake, et al. 2010. "Melanocytic Nevi, Nevus Genes, and Melanoma Risk in a Large

- Case-Control Study in the United Kingdom.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 19 (8): 2043–54. <https://doi.org/10.1158/1055-9965.EPI-10-0233>.
- Ng, Pauline C., and Steven Henikoff. 2003. “SIFT: Predicting Amino Acid Changes That Affect Protein Function.” *Nucleic Acids Research* 31 (13): 3812–14.
- Nguyen, Nga Thi Thuy, Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Walter Santana-Garcia, Raul Ossio, Carla Daniela Robles-Espinoza, Mathieu Bahin, et al. 2018. “RSAT 2018: Regulatory Sequence Analysis Tools 20th Anniversary.” *Nucleic Acids Research* 46 (W1): W209–14. <https://doi.org/10.1093/nar/gky317>.
- Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. 1999. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 27 (1): 29–34. <https://doi.org/10.1093/nar/27.1.29>.
- O’Shea, Sally J., Carla Daniela Robles-Espinoza, Lauren McLellan, Jeanine Harrigan, Xavier Jacq, James Hewinson, Vivek Iyer, et al. 2017. “A Population-Based Analysis of Germline BAP1 Mutations in Melanoma.” *Human Molecular Genetics* 26 (4): 717–28. <https://doi.org/10.1093/hmg/ddw403>.
- Ossio, Raul, O. Isaac Garcia-Salinas, Diego Said Anaya-Mancilla, Jair S. Garcia-Sotelo, Luis A. Aguilar, David J. Adams, and Carla Daniela Robles-Espinoza. 2019. “VCF/Plotein: Visualization and Prioritization of Genomic Variants from Human Exome Sequencing Projects.” *Bioinformatics* 35 (22): 4803–5. <https://doi.org/10.1093/bioinformatics/btz458>.
- Ossio, Raul, Rodrigo Roldán-Marín, Héctor Martínez-Said, David J. Adams, and Carla Daniela Robles-Espinoza. 2017. “Melanoma: A Global Perspective.” *Nature Reviews. Cancer* 17 (7): 393–94. <https://doi.org/10.1038/nrc.2017.43>.
- Pabinger, Stephan, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, and Zlatko Trajanoski. 2014. “A Survey of Tools for Variant Analysis of Next-Generation Genome Sequencing Data.” *Briefings in Bioinformatics* 15 (2): 256–78. <https://doi.org/10.1093/bib/bbs086>.
- Paila, Umadevi, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. 2013. “GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations.” *PLoS Computational Biology* 9 (7): e1003153. <https://doi.org/10.1371/journal.pcbi.1003153>.
- Parkin, D M, D Mesher, and P Sasieni. 2011. “13. Cancers Attributable to Solar (Ultraviolet) Radiation Exposure in the UK in 2010.” *British Journal of Cancer* 105 (Suppl 2): S66–69. <https://doi.org/10.1038/bjc.2011.486>.
- Petljak, Mia, Ludmil B. Alexandrov, Jonathan S. Brummel, Stacey Price, David C. Wedge, Sebastian Grossmann, Kevin J. Dawson, et al. 2019. “Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis.” *Cell* 176 (6): 1282–1294.e20. <https://doi.org/10.1016/j.cell.2019.02.012>.
- Petrucci, Nancie, Mary B. Daly, and Tuya Pal. 1993. “BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer.” In *GeneReviews*<sup>®</sup>, edited by Margaret P. Adam, Holly H. Ardinger, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen Stephens, and

- Anne Amemiya. Seattle (WA): University of Washington, Seattle. <http://www.ncbi.nlm.nih.gov/books/NBK1247/>.
- Pooley, Karen A., Jonathan Tyrer, Mitul Shah, Kristy E. Driver, Jean Leyland, Judith Brown, Tina Audley, et al. 2010. "No Association between TERT-CLPTM1L Single Nucleotide Polymorphism Rs401681 and Mean Telomere Length or Cancer Risk." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 19 (7): 1862–65. <https://doi.org/10.1158/1055-9965.EPI-10-0281>.
- Potrony, Miriam, Celia Badenas, Paula Aguilera, Joan Anton Puig-Butille, Cristina Carrera, Josep Malvehy, and Susana Puig. 2015. "Update in Genetic Susceptibility in Melanoma." *Annals of Translational Medicine* 3 (15). <https://doi.org/10.3978/j.issn.2305-5839.2015.08.11>.
- Rabbie, Roy, Peter Ferguson, Christian Molina-Aguilar, David J Adams, and Carla D Robles-Espinoza. 2019. "Melanoma Subtypes: Genomic Profiles, Prognostic Molecular Markers and Therapeutic Possibilities." *The Journal of Pathology* 247 (5): 539–51. <https://doi.org/10.1002/path.5213>.
- Ramsay, Andrew J., Víctor Quesada, Miguel Foronda, Laura Conde, Alejandra Martínez-Trillos, Neus Villamor, David Rodríguez, et al. 2013. "POT1 Mutations Cause Telomere Dysfunction in Chronic Lymphocytic Leukemia." *Nature Genetics* 45 (5): 526–30. <https://doi.org/10.1038/ng.2584>.
- Ribero, Simone, Dan Glass, and Veronique Bataille. 2016. "Genetic Epidemiology of Melanoma." *European Journal of Dermatology: EJD* 26 (4): 335–39. <https://doi.org/10.1684/ejd.2016.2787>.
- Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24. <https://doi.org/10.1038/gim.2015.30>.
- Robles-Espinoza, Carla Daniela, Mark Harland, Andrew J. Ramsay, Lauren G. Aoude, Víctor Quesada, Zhihao Ding, Karen A. Pooley, et al. 2014. "POT1 Loss-of-Function Variants Predispose to Familial Melanoma." *Nature Genetics* 46 (5): 478–81. <https://doi.org/10.1038/ng.2947>.
- Rosenquist, Thomas A., and Arthur P. Grollman. 2016. "Mutational Signature of Aristolochic Acid: Clue to the Recognition of a Global Disease." *DNA Repair* 44: 205–11. <https://doi.org/10.1016/j.dnarep.2016.05.027>.
- Rossi, Mariarita, Cristina Pellegrini, Ludovica Cardelli, Valeria Ciciarelli, Lucia Di Nardo, and Maria Concetta Fargnoli. 2019. "Familial Melanoma: Diagnostic and Management Implications." *Dermatology Practical & Conceptual* 9 (1): 10–16. <https://doi.org/10.5826/dpc.0901a03>.
- Roth, Gregory A., Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, et al. 2018. "Global, Regional, and National Age-Sex-Specific Mortality for 282 Causes of Death in 195 Countries and Territories, 1980–2017: A Systematic Analysis for the Global Burden of Disease Study 2017." *The Lancet* 392 (10159): 1736–88. [https://doi.org/10.1016/S0140-6736\(18\)32203-7](https://doi.org/10.1016/S0140-6736(18)32203-7).

- Salatino, Silvia, and Varun Ramraj. 2017. "BrowseVCF: A Web-Based Application and Workflow to Quickly Prioritize Disease-Causative Variants in VCF Files." *Briefings in Bioinformatics* 18 (5): 774–79. <https://doi.org/10.1093/bib/bbw054>.
- Scolyer, Richard A., Georgina V. Long, and John F. Thompson. 2011. "Evolving Concepts in Melanoma Classification and Their Relevance to Multidisciplinary Melanoma Patient Care." *Molecular Oncology* 5 (2): 124–36. <https://doi.org/10.1016/j.molonc.2011.03.002>.
- Shen, Erica, Joanne Xiu, Giselle Y. Lopez, Rex Bentley, Ali Jalali, Amy B. Heimberger, Matthew N. Bainbridge, Melissa L. Bondy, and Kyle M. Walsh. 2020. "POT1 Mutation Spectrum in Tumour Types Commonly Diagnosed among POT1-Associated Hereditary Cancer Syndrome Families." *Journal of Medical Genetics*, January. <https://doi.org/10.1136/jmedgenet-2019-106657>.
- Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. "DbSNP: The NCBI Database of Genetic Variation." *Nucleic Acids Research* 29 (1): 308–11. <https://doi.org/10.1093/nar/29.1.308>.
- Shi, Jianxin, Xiaohong R. Yang, Bari Ballew, Melissa Rotunno, Donato Calista, Maria Concetta Fagnoli, Paola Ghiorzo, et al. 2014. "Rare Missense Variants in POT1 Predispose to Familial Cutaneous Malignant Melanoma." *Nature Genetics* 46 (5): 482–86. <https://doi.org/10.1038/ng.2941>.
- Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (October): 539. <https://doi.org/10.1038/msb.2011.75>.
- Skolnick, M. H., L. A. Cannon-Albright, and A. Kamb. 1994. "Genetic Predisposition to Melanoma." *European Journal of Cancer* 30 (13): 1991–95. [https://doi.org/10.1016/0959-8049\(94\)00392-I](https://doi.org/10.1016/0959-8049(94)00392-I).
- Solano-Román, Antonio, Verónica Alfaro-Arias, Carlos Cruz-Castillo, and Allan Orozco-Solano. 2018. "Visualization Portal for Genetic Variation (VizGVar): A Tool for Interactive Visualization of SNPs and Somatic Mutations in Exons, Genes and Protein Domains." *Bioinformatics (Oxford, England)* 34 (6): 1048–49. <https://doi.org/10.1093/bioinformatics/btx694>.
- Speedy, Helen E., Ben Kinnersley, Daniel Chubb, Peter Broderick, Philip J. Law, Kevin Litchfield, Sandrine Jayne, et al. 2016. "Germ Line Mutations in Shelterin Complex Genes Are Associated with Familial Chronic Lymphocytic Leukemia." *Blood* 128 (19): 2319–26. <https://doi.org/10.1182/blood-2016-01-695692>.
- Stanley, Susan E., and Mary Armanios. 2015. "The Short and Long Telomere Syndromes: Paired Paradigms for Molecular Medicine." *Current Opinion in Genetics & Development* 33 (August): 1–9. <https://doi.org/10.1016/j.gde.2015.06.004>.
- Stratton, Michael R. 2013. "Journeys into the Genome of Cancer Cells." *EMBO Molecular Medicine* 5 (2): 169–72. <https://doi.org/10.1002/emmm.201202388>.
- Struck, Travis J., Brian K. Mannakee, and Ryan N. Gutenkunst. 2018. "The Impact of Genome-Wide Association Studies on Biomedical Research Publications." *Human Genomics* 12 (August). <https://doi.org/10.1186/s40246-018-0172-4>.
- Tacastacas, Joselin D., Julie Bray, Yoon K. Cohen, Joshua Arbesman, Julian Kim, Henry B. Koon, Kord Honda, Kevin D. Cooper, and Meg R. Gerstenblith. 2014. "Update on Primary

- Mucosal Melanoma." *Journal of the American Academy of Dermatology* 71 (2): 366–75. <https://doi.org/10.1016/j.jaad.2014.03.031>.
- Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47. <https://doi.org/10.1093/nar/gky1015>.
- Tung, Nadine M., and Judy E. Garber. 2018. "BRCA 1/2 Testing: Therapeutic Implications for Breast Cancer Management." *British Journal of Cancer* 119 (2): 141–52. <https://doi.org/10.1038/s41416-018-0127-5>.
- Verhagen, Judith M. A., Job H. Veldman, Paul A. van der Zwaag, Jan H. von der Thüsen, Erwin Brosens, Imke Christiaans, Dennis Dooijes, et al. 2018. "Lack of Evidence for a Causal Role of CALR3 in Monogenic Cardiomyopathy." *European Journal of Human Genetics* 26 (11): 1603–10. <https://doi.org/10.1038/s41431-018-0208-1>.
- Walpole, Sebastian, Antonia L. Pritchard, Colleen M. Cebulla, Robert Pilarski, Meredith Stautberg, Frederick H. Davidorf, Arnaud de la Fouchardière, et al. 2018. "Comprehensive Study of the Clinical Phenotype of Germline BAP1 Variant-Carrying Families Worldwide." *JNCI: Journal of the National Cancer Institute* 110 (12): 1328–41. <https://doi.org/10.1093/jnci/djy171>.
- Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. "Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench." *Bioinformatics (Oxford, England)* 25 (9): 1189–91. <https://doi.org/10.1093/bioinformatics/btp033>.
- Weinberg, Robert A. 2013. *The Biology of Cancer*. Garland Science.
- Wellcome Trust Case Control Consortium. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447 (7145): 661–78. <https://doi.org/10.1038/nature05911>.
- Whiteman, David C., Adele C. Green, and Catherine M. Olsen. 2016. "The Growing Burden of Invasive Melanoma: Projections of Incidence Rates and Numbers of New Cases in Six Susceptible Populations through 2031." *Journal of Investigative Dermatology* 136 (6): 1161–71. <https://doi.org/10.1016/j.jid.2016.01.035>.
- Wiesner, Thomas, Anna C. Obenaus, Rajmohan Murali, Isabella Fried, Klaus G. Griewank, Peter Ulz, Christian Windpassinger, et al. 2011. "Germline Mutations in BAP1 Predispose to Melanocytic Tumors." *Nature Genetics* 43 (10): 1018–21. <https://doi.org/10.1038/ng.910>.
- Wong, Kim, Carla Daniela Robles-Espinoza, David Rodriguez, Saskia S. Rudat, Susana Puig, Miriam Potrony, Chi C. Wong, et al. 2019. "Association of the POT1 Germline Missense Variant p.I78T With Familial Melanoma." *JAMA Dermatology* 155 (5): 604–9. <https://doi.org/10.1001/jamadermatol.2018.3662>.
- Xiong, Michael, Ahmad Charifa, and Chih Shan J. Chen. 2020. "Cancer, Lentigo Maligna Melanoma." In *StatPearls*. Treasure Island (FL): StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/books/NBK482163/>.
- Xu, Chang. 2018. "A Review of Somatic Single Nucleotide Variant Calling Algorithms for Next-Generation Sequencing Data." *Computational and Structural Biotechnology Journal* 16 (February): 15–24. <https://doi.org/10.1016/j.csbj.2018.01.003>.



- Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, et al. 2020. "Ensembl 2020." *Nucleic Acids Research* 48 (D1): D682–88. <https://doi.org/10.1093/nar/gkz966>.
- Yeo, Gene, and Christopher B. Burge. 2004. "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11 (2–3): 377–94. <https://doi.org/10.1089/1066527041410418>.
- Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. "Ensembl 2018." *Nucleic Acids Research* 46 (D1): D754–61. <https://doi.org/10.1093/nar/gkx1098>.
- Zhou, Xin, Michael N. Edmonson, Mark R. Wilkinson, Aman Patel, Gang Wu, Yu Liu, Yongjin Li, et al. 2016. "Exploring Genomic Alteration in Pediatric Cancer Using ProteinPaint." *Nature Genetics* 48 (1): 4–6. <https://doi.org/10.1038/ng.3466>.
- Zuo, Lin, John Weger, Qingbei Yang, Alisa M. Goldstein, Margaret A. Tucker, Graeme J. Walker, Nicholas Hayward, and Nicholas C. Dracopoli. 1996. "Germline Mutations in the P16INK4a Binding Domain of CDK4 in Familial Melanoma." *Nature Genetics* 12 (1): 97–99. <https://doi.org/10.1038/ng0196-97>.

## 6 Anexos y material suplementario

### 6.1 Tabla Suplementaria 1

**Muestras de línea germinal secuenciadas en este estudio.** La columna A indica el ID de la muestra, la columna B la proporción de bases de POT1 con cobertura de alta calidad, la columna C el estatus de caso/control, y la columna D, la cohorte de origen (Leeds, SEARCH o WTCCC). Debido a su longitud, no es posible imprimir esta tabla, pero puede ser descargada del siguiente enlace: [https://www.dropbox.com/s/k30egw8s5ueqcsi/Tabla\\_Suplementaria\\_1.xlsx?dl=0](https://www.dropbox.com/s/k30egw8s5ueqcsi/Tabla_Suplementaria_1.xlsx?dl=0)

### 6.2 Línea de comando HaplotypeCaller (GATK)

```
analysis_type=HaplotypeCaller input_file=[example.bam] showFullBamList=false read_buffer_size=null phone_home=NO_ET gatk_key=gatk.key
tag=NA read_filter=[] disable_read_filter=[] intervals=[pot1.bed, 7:1-159138663] excludeIntervals=null interval_set_rule=INTERSECTION
interval_merging=ALL interval_padding=0 reference_
sequence=hs37d5.fa nonDeterministicRandomSeed=false disableDithering=false maxRunTi
me=-1 maxRuntimeUnits=MINUTES downsampling_type=BY_SAMPLE downsample_to_fraction=null downsample_to_coverage=500 baq=OFF
baqGapOpenPena
lty=40.0 refactor_NDN_cigar_string=false fix_misencoded_quality_scores=false allow_potentially_misencoded_quality_scores=false useOri
ginalQualities=false defaultBaseQualities=-1 performanceLog=null BQSR=null quantize_qual=0 disable_indel_qual=false
emit_original_qual=false preserve_qscores_less_than=6 globalQScorePrior=-1.0 validation_strictness=SILENT remove_program_records=false
keep_program_records=false sample_rename_mapping_file=null unsafe=null
disable_auto_index_creation_and_locking_when_reading_rods=false no_cmdline_in_header=false sites_only=false
never_trim_vcf_format_field=false bcf=false bam_compression=null simplifyBAM=false disable_bam_indexing=false generate_md5=false
num_threads=1 num_cpu_threads_per_data_thread=1 num_io_threads=0 monitorThreadEfficiency=false num_bam_file_handles=null
read_group_black_list=null pedigree=[] pedigreeString=[] pedigreeValidationType=STRICT allow_intervals_with_unindexed_bam=false
generateShadowBCF=false variant_index_type=DYNAMIC_SEEK variant_index_parameter=-1 logging_level=ERROR log_to_file=null help=false
version=false out=/example_path/1_gatk_haplotype_caller_with_genome_chunking/7_1-159138663.gatk_haplotype.vcf.gz
likelihoodCalculationEngine=PairHMM heterogeneousKmerSizeResolution=COMBO_MIN dbsnp=(RodBinding name= source=UNBOUND)
dontTrimActiveRegions=false maxDiscARExtension=25 maxGGAARExtension=300 paddingAroundIndels=150 paddingAroundSNPs=20 comp=[]
annotation=[ClippingRankSumTest, DepthPerSampleHC] excludeAnnotation=[] debug=false useFilteredReadsForAnnotations=false
emitRefConfidence=NONE bamOutput=null bamWriterType=CALLED_HAPLOTYPES disableOptimizations=false annotateNDA=false
heterozygosity=0.001 indel_heterozygosity=1.25E-4 standard_min_confidence_threshold_for_calling=4.0
standard_min_confidence_threshold_for_emitting=4.0 max_alternate_alleles=6 input_prior=[] sample_ploidy=2 genotyping_mode=DISCOVERY
alleles=(RodBinding name= source=UNBOUND) contamination_fraction_to_filter=0.0 contamination_fraction_per_sample_file=null
p_nonref_model=null exactcallslog=null output_mode=EMIT_VARIANTS_ONLY allSitePLs=false gcpHMM=10
pair_hmm_implementation=VECTOR_LOGLESS_CACHING pair_hmm_sub_implementation=ENABLE_ALL
always_load_vector_logless_PairHMM_lib=false phredScaledGlobalReadMismatchingRate=45 noFpga=false sample_name=null kmerSize=[10,
```

```
25] dontIncreaseKmerSizesForCycles=false allowNonUniqueKmersInRef=false numPruningSamples=1 recoverDanglingHeads=false
doNotRecoverDanglingBranches=false minDanglingBranchLength=4 consensus=false maxNumHaplotypesInPopulation=128
errorCorrectKmers=false minPruning=2 debugGraphTransformations=false allowCyclesInKmerGraphToGeneratePaths=false graphOutput=null
kmerLengthForReadErrorCorrection=25 minObservationsForKmerToBeSolid=20 GVCFGQBands=[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55,
56, 57, 58, 59, 60, 70, 80, 90, 99] indelSizeToEliminateInRefModel=10 min_base_quality_score=10 includeUmappedReads=false
useAllelesTrigger=false doNotRunPhysicalPhasing=true keepRG=null justDetermineActiveRegions=false dontGenotype=false
dontUseSoftClippedBases=false captureAssemblyFailureBAM=false errorCorrectReads=false pcr_indel_model=CONSERVATIVE
maxReadsInRegionPerSample=10000 minReadsPerAlignmentStart=10 mergeVariantsViaLD=false activityProfileOut=null activeRegionOut=null
activeRegionIn=null activeRegionExtension=null forceActive=false activeRegionMaxSize=null bandPassSigma=null
maxProbPropagationDistance=50 activeProbabilityThreshold=0.002 min_mapping_quality_score=20 filter_reads_with_N_cigar=false
filter_mismatching_base_and_qual=false filter_bases_not_stored=false
```

## 6.3 Tabla Suplementaria 2

**Todas las variantes con predicción de alteración de estructura de la proteína identificadas por medio de secuenciación de Fluidigm.** Columna A, posición de la variante, columna B, consecuencia patogénica predicha, columnas C-E, número total de muestras, número de casos y número de controles, respectivamente, columnas F-G, IDs de los casos y los controles en donde la variante fue encontrada, respectivamente, columna H, número de muestras resecuenciadas via Illumina o por capilar para validación de la variante, columna I, IDs de las muestras resecuenciadas, columna J, número de muestras en los que la variante fue confirmada, columna K, proporción de muestras en los que la variante fue confirmada (confirmadas / evaluadas), columna L, detalles de las variantes confirmadas. Debido a su longitud, no es posible imprimir esta tabla, pero puede ser descargada del siguiente enlace:

[https://www.dropbox.com/s/ate6cnm2cbiu6ev/Tabla\\_Suplementaria\\_2.xlsx?dl=0](https://www.dropbox.com/s/ate6cnm2cbiu6ev/Tabla_Suplementaria_2.xlsx?dl=0)

## 6.4 Tabla Suplementaria 3

**Todas las variantes con predicción de alteración de estructura de la proteína identificadas por medio de secuenciación de Fluidigm y confirmadas por medio de secuenciación Illumina o por capilar.** Columna A, posición de la variante, columna B, consecuencia patogénica predicha, columna C, consecuencia en la proteína, columna D, variación co-existente en esa posición en

otras bases de datos (de acuerdo a lo reportado por Ensembl Variant Effect Predictor v96), columnas E-F, predicciones de patogenicidad de acuerdo a los algoritmos SIFT y PolyPhen-2 respectivamente, columnas G-H, frecuencia alélica en la población global y europeos no finlandeses de la base de datos gnomAD, versión 2.1 respectivamente, columnas I-K, número total de muestras, número de casos y número de controles, respectivamente, columnas L-M, IDs de los casos y los controles en donde la variante fue identificada, respectivamente, columna N, número de muestras resecuenciadas por medio de Illumina o capilar para validación de la variante, columna O, IDs de las muestras resecuenciadas, columna P, número de muestras en las que la variante fue confirmada, columna Q, proporción de muestras en las que la variante fue confirmada (confirmadas / evaluadas), columna R, detalles de las variantes confirmadas, columna S, grupo de patogenicidad asignado. Debido a su longitud, no es posible imprimir esta tabla, pero puede ser descargada del siguiente enlace:

[https://www.dropbox.com/s/ft3kdf1zgx8jib1/Tabla\\_Suplementaria\\_3.xlsx?dl=0](https://www.dropbox.com/s/ft3kdf1zgx8jib1/Tabla_Suplementaria_3.xlsx?dl=0)

## 6.5 Tabla Suplementaria 4

**Co-ocurrencia de variantes en *POT1* con variantes en otros genes que confieren riesgo elevado a padecer melanoma en las muestras resecuenciadas por Illumina.** Columna A, ID de la muestra, columna B, variante detectada en *POT1*, columna C, variante detectada en un gen de riesgo a desarrollar melanoma.

<b>Muestra</b>	<b>Variante en <i>POT1</i></b>	<b>Variante en gen patogénico</b>
PD30744a	Q376R	<i>CDKN2A</i> A148T
PD30549a	G404V	<i>CDKN2A</i> A148T
PD30730a	G404V & Y419C	<i>CDKN2A</i> A60T

## 7 Publicaciones originadas en este doctorado

Ossio, Raul, O. Isaac Garcia-Salinas, Diego Said Anaya-Mancilla, Jair S. Garcia-Sotelo, Luis A. Aguilar, David J. Adams, and Carla Daniela Robles-Espinoza. 2019. "**VCF/Plotein: Visualization and Prioritization of Genomic Variants from Human Exome Sequencing Projects.**" *Bioinformatics* 35 (22): 4803–5. <https://doi.org/10.1093/bioinformatics/btz458>.

Ossio, Raul, Rodrigo Roldán-Marín, Héctor Martínez-Said, David J. Adams, and Carla Daniela Robles-Espinoza. 2017. "**Melanoma: A Global Perspective.**" *Nature Reviews. Cancer* 17 (7): 393–94. <https://doi.org/10.1038/nrc.2017.43>.

Nguyen, Nga Thi Thuy, Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Walter Santana-Garcia, Raul Ossio, Carla Daniela Robles-Espinoza, Mathieu Bahin, et al. 2018. "**RSAT 2018: Regulatory Sequence Analysis Tools 20th Anniversary.**" *Nucleic Acids Research* 46 (W1): W209–14. <https://doi.org/10.1093/nar/gky317>.

Walpole, Sebastian, Antonia L. Pritchard, Colleen M. Cebulla, Robert Pilarski, Meredith Stautberg, Frederick H. Davidorf, Arnaud de la Fouchardière, et al. 2018. "**Comprehensive Study of the Clinical Phenotype of Germline BAP1 Variant-Carrying Families Worldwide.**" *JNCI: Journal of the National Cancer Institute* 110 (12): 1328–41. <https://doi.org/10.1093/jnci/djy171>.

**Nota sobre la publicación Walpole et al (2018).** Dado el elevado número de autores, no todos se encuentran enlistados en la primera página del documento. Mi nombre se encuentra en la sección de Notas: Autores, en la página 1339.