UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DISEÑO DE UN SISTEMA DE INFORMACIÓN GENÓMICA PARA EL DIAGNÓSTICO DEL NEUROBLASTOMA

AUTOR: CLARA SOLER PELLICER

TUTOR: ÓSCAR PASTOR LÓPEZ

Seleccion

Curso Académico: 2016-17

AGRADECIMIENTOS

En primer lugar, quiero dar las gracias a mi tutor, Dr. Óscar Pastor López, por haberme dado la oportunidad de adentrarme en este proyecto y hacerme descubrir el mundo apasionante de la bioinformática. También quiero dar las gracias a los miembros del grupo PROS, en especial a José F. Reyes R. por todas las correcciones y tiempo invertido en mí, y a Ana León por siempre estar dispuesta a ayudarme en todo lo que necesitase.

Por otro lado, agradecer a mis compañeras y amigas, Belén, Julia, Fuen, Sandra y Paula, todo el apoyo que me han dado durante estos años en la universidad, porque sin ellas no hubiese sido lo mismo.

Por último, a mis padres y hermano, por siempre darme todo lo que he necesitado y porque sin ellos no hubiese llegado a donde estoy ahora.

RESUMEN

La aplicación de técnicas de *Modelado Conceptual* (MC) es un punto clave a la hora de obtener diagnósticos genómicos que permitan potenciar la *Medicina Personalizada* a través de la prevención y diagnóstico con mayor precisión. Actualmente, los expertos en genética y bioinformática no cuentan con herramientas eficientes que les proporcione un diagnóstico precoz para sus pacientes (en este caso específico, enfermos de Neuroblastoma). La información sobre las variaciones asociadas con la enfermedad se encuentra en múltiples repositorios, dando lugar a problemas de dispersión, el cual es un ejemplo de lo que se conoce como "*caos genómico*". Este trabajo, propone el diseño de un *Sistema de Información Genómico* (GeIS) que permita reducir el problema existente en el dominio mediante la creación de un repositorio con datos curados y de calidad. El proceso de obtención de dicho sistema se basa en una revisión y filtrado del contenido que existe en los repositorios genómicos actuales con el fin de obtener sólo la información más relevante sobre la enfermedad en estudio. Dicho proceso culmina con la obtención de un total de 373 variaciones –*curadas*- relacionadas con el "*Neuroblastoma*". Dichas variaciones son cargadas a la *Base de Datos del Genoma Humano* (HGDB) y la información es explotada mediante la herramienta "*VarSearch"*, la cual compara las variaciones encontradas en la muestra de un paciente con las cargadas en la HGDB. De esta manera, se obtiene un diagnóstico mucho más preciso que permite al especialista tomar las decisiones correctas y permitiría la asignación de tratamientos guiados y específicos (personalizados) según las variaciones que posea el paciente.

Palabras clave: Neuroblastoma, medicina de precisión, modelado conceptual, diagnóstico genómico, GeIS.

RESUM

L'aplicació de tècniques de *Modelatge Concepual* (MC) és un punt clau a l'hora d'obtindre diagnòstics genòmics que permeten potenciar la *Medicina Personalitzada* a través de la prevenció i diagnòstic amb més precisió. Actualment, els experts en genética i bioinformática no compten amb ferramentes eficients que els proporcionen un diagnòstic precoç per als seus pacients (en aquest cas específic, malalts de Neuroblastoma). La información sobre les variacions associades amb la malaltia es trova en múltiples repositoris, donant lloc a problemas de dispersió, el qual és un exemple del que es coneix com a "*caos genòmic*". Aquest treball, proposa el disseny d'un *Sistema d'Informació Genòmic* (GeIS) que permeta aminorar el problema existent en el domini per mitjà de la creació d'un repositori amb dades curades i de qualitat. El procés d'obtenció d'aquest sistema es basa en una revisió i filtrat del contingut que existix en els repositoris genòmics actuals a fi d'obtindre només la información més rellevant sobre la malaltia en estudi. Aquest procés dóna lloc a l'obtenció d'un total de 373 variacions –*curades*- relacionades amb el "Neuroblastoma". Dites variacions són carregades a la *Base de Dades del Genoma Humà* (HGDB) i la información és explotada mitjançant la ferramenta "*VarSearch*", la qual compara les variacions trobades en la mostra d'un pacient amb les carregades en la HGDB. D'aquesta manera, s'obté un diagnòstic molt més precís que permet a l'especialista prendre les decisions correctes i permetria l'assignació de tractaments guiats i específics (personalitzats) segons les variacions que posseïsca el pacient.

Paraules clau: Neuroblastoma, medicina de precisió, modelatge conceptual, diagnòstic genòmic, GeIS.

ABSTRACT

The application of techniques of *Conceptual Modeling* (CM) is a key point in obtaining genomic diagnosis that allow to enhance the *personalized medicine* through prevention and diagnosis with better precision. Currently, genetics and bioinformatics experts do not have efficient tools that provide them an early diagnosis for their patients (in this specific case, Neuroblastoma patients). The information on the variations associated with the disease is located in multiple repositories, giving rise to problems of dispersion, which is an example of what is known as "*genomic chaos*". This work proposes a design of a *Genomic Information System* (GeIS) that allows to reduce the problem existing in the domain by creating a repository with curated and quality data. The obtainment process of that system is based on a review and filtering of content that exists in the current genomic repositories in order to get only the most relevant information about the disease in study. This process culminates with the obtaining of a total of 373 variations -*curated*- related to "Neuroblastoma". These variations are loaded into the *Database of the Human Genome* (HGDB) and the information is exploited by the "*VarSearch*" tool, which compares the variations found in the sample of a patient with the loaded into the HGDB. In this way, you get a much more precise diagnosis that allows the specialist to make the right decisions and would allow the allocation of guided and specific (personalized) treatments according to the patient's variations.

Keywords: Neuroblastoma, precision medicine, conceptual modeling, genomic diagnosis, GeIS.

DOCUMENTOS CONTENIDOS EN EL TFG

DOCUMENTO I: MEMORIA.

DOCUMENTO II: PRESUPUESTO.

DOCUMENTO III: ANEXOS.

ÍNDICE DE LA MEMORIA

ÍNDICE DEL PRESUPUESTO

ÍNDICE DE LOS ANEXOS

LISTA DE FIGURAS

LISTA DE TABLAS

MEMORIA

DOCUMENTO I

"Diseño de un Sistema de Información Genómica para el Diagnóstico del Neuroblastoma"

Clara Soler Pellicer

Curso 2016/2017

CAPÍTULO 1. INTRODUCCIÓN

1.1. MOTIVACIÓN

En los últimos años ha habido grandes avances en el campo de la genómica, descubrimientos que permiten detectar la predisposición genética y la patogenia molecular de una persona a padecer enfermedades de origen genético.

Aunque la tecnología convencional de secuenciación ideada por Sanger et al. (1977) permite la detección de variantes genéticas de pequeño tamaño, ésta resulta ser una técnica muy lenta ya que solo se pueden realizar pocas reacciones en paralelo. Esta limitación produce una duración de tiempos largos en los experimentos y que el precio por base secuenciada sea bastante elevado. Las tecnologías de nueva generación (NGS de sus siglas en inglés "*Next-Generation Sequencing*"), conocidas también como secuenciación masiva paralela, permiten secuenciar millones de fragmentos de ADN de forma paralela a un precio por base mucho más barato que las tecnologías de secuenciación convencionales. Estas tecnologías son capaces de detectar todos los tipos de variación genómica en un único experimento, como, por ejemplo, variantes de nucleótido único o mutaciones puntuales, pequeñas inserciones y deleciones, o también tanto variantes estructurales equilibradas (inversiones y translocaciones) como desequilibradas (deleciones o duplicaciones) (Benjamín Rodríguez-Santiago, 2012).

Estas técnicas son útiles para entender las alteraciones genéticas de manera integral. Como parte de estos avances, las alteraciones genómicas de algunas enfermedades, como, por ejemplo, en este trabajo que se aborda el Neuroblastoma, pueden analizarse y entenderse con mayor precisión (Theruvath et al., 2016).

Actualmente, es posible obtener la secuencia completa de un genoma humano por menos de 1000 dólares (Van Dijk et al., 2014), por lo que el reto ya no es la extracción de datos a partir de tecnologías de secuenciación masiva sino el *almacenar*, *procesar* e *interpretar* la inmensidad de datos obtenidos. Este reto precisa de la participación de un amplio espectro de especialistas: biólogos computacionales, asesores genéticos, doctores o ingenieros biomédicos.

Una vez obtenida la secuencia, se alinea con la secuencia de referencia de la que va a diferir entre 4,1-5 millones de posiciones (Auton et al., 2015). La combinación de estas variaciones, lo que se llama "*genotipo*", junto con los factores ambientales determinan las características físicas de cada individuo, es decir, el "*fenotipo*". Concretamente el entendimiento de esta relación genotipo-fenotipo es la principal tarea de la medicina genómica.

Si se realiza una búsqueda de información genética relacionada con cierta enfermedad se aprecia la gran heterogeneidad y dispersión en los repositorios genómicos, lo que da lugar al

famoso "*caos genómico*". Para la correcta gestión de los datos se precisa de mecanismos que recojan toda la información de las diferentes fuentes y la unifiquen creando un único repositorio evitando así problemas de ubicuidad o usabilidad, entre otros. Este es uno de los grandes retos que se plantea la comunidad médico-científica quienes buscan la obtención de sistemas que les permitan obtener diagnósticos eficientes y fiables con el fin de mejorar la calidad de vida de los pacientes.

El constante crecimiento en el entorno bioinformático ha dado lugar a la metodología SILE (*Search-Identification-Load-Exploitation*), que sirve como herramienta para mejorar los procesos de carga de las bases de datos del genoma humano. Con ella se pretende crear bases de datos filtradas en las que se tiene una información validada y relevante que ayude a crear diagnósticos más fiables y precisos. Esta metodología se pone en práctica con la enfermedad del Neuroblastoma en el presente proyecto.

Con la creación de estos *Sistemas de Información Genómicos* (GeIS) para el diagnóstico, nos adentramos en el desarrollo de una medicina mucho más personalizada y precisa.

1.2. OBJETIVOS DEL TRABAJO

El **objetivo general** del trabajo consiste en crear un Sistema de Información Genómico que recoja la información más relevante que se puede extraer de las distintas bases de datos/artículos científicos disponibles sobre una enfermedad en concreto, el Neuroblastoma, de manera que se sitúe toda ella en un marco común evitando problemas de redundancias o calidad de los datos. Podrá ser utilizado por médicos o genetistas los cuales podrán explotar la información disponible y proporcionar diagnósticos más fiables y corroborados por la comunidad científica.

Como **objetivos específicos**:

- Investigar las fuentes de datos heterogéneas existentes con el objetivo de poner solución al "*caos genómico*".
- Seleccionar e identificar aquella información más relevante de las bases de datos convenientes a investigar para la enfermedad del Neuroblastoma.
- Manipular y/o transformar de manera correcta los datos genómicos complejos incluyendo la determinación de criterios de calidad para su correcta gestión.
- Obtener una base de datos homogénea que contenga toda la información correctamente filtrada con el fin de explotar el conocimiento disponible en las diferentes fuentes heterogéneas.

1.3. ESTRUCTURA DEL DOCUMENTO

DOCUMENTO I: MEMORIA. A continuación, se presenta la estructura del siguiente Trabajo Final de Grado, el cual consta de 5 capítulos. Un primer capítulo introductorio donde se presenta el problema, un segundo capítulo donde se sitúa al lector en el contexto, después se describen los materiales utilizados en un tercer capítulo, el cuarto capítulo consta del desarrollo del trabajo utilizando la metodología SILE y, por último, un capítulo a modo de conclusión. A continuación, se muestran de manera detallada cada uno de ellos:

- CAPÍTULO 1. INTRODUCCIÓN.
  Además de contener la estructura del documento, en este primer capítulo se plantea el problema actual con los Sistemas de Información Genómicos, como ha surgido la motivación por ponerle solución y cuáles son los objetivos entorno a los cuales va a girar el proyecto, tanto de forma general como las tareas específicas.

- CAPÍTULO 2. ESTADO DEL ARTE.
  Se pretende situar al lector en el marco actual tanto de la enfermedad del Neuroblastoma como en la obtención del diagnóstico genómico. Se expone toda la actualidad acerca de la enfermedad en cuestión, desde cuándo se detectó por primera vez hasta las soluciones que se proponen actualmente. También se hace hincapié en las bases de datos que utilizan los genetistas para obtener la información, además de una revisión de las empresas genómicas que ya proporcionan servicio de diagnóstico genómico personalizado y cuáles son las ventajas del presente proyecto

- CAPÍTULO 3. MATERIALES Y MÉTODOS.
  Se describen de manera detallada las herramientas y metodologías seguidas para el desarrollo del trabajo con la meta de alcanzar los objetivos planteados.

- CAPÍTULO 4. APLICACIÓN DEL MÉTODO SILE.
  Se describe paso a paso como se ha desarrollado la metodología SILE, la cual proporciona los resultados óptimos, dividida en cuatro partes: *Search*, *Identification*, *Load* y *Exploitation.* En cada parte se detallan los pasos seguidos para la evolución de la información con el fin de poder transformarla en conocimiento y poder explotarla. Es decir, incluye todo el desarrollo e implementación del sistema de información genómico para el diagnóstico del Neuroblastoma.

- CAPÍTULO 5. CONCLUSIONES Y LÍNEAS FUTURAS.
  Se hace un breve repaso del trabajo, remarcando los retos a los que se ha hecho frente, así como los objetivos conseguidos al realizar este trabajo de investigación, y se proponen posibles mejoras para trabajos futuros.

CAPÍTULO 2. ESTADO DEL ARTE

2.1 NEUROBLASTOMA

El **Neuroblastoma** se trata de un tipo de tumor canceroso que se desarrolla a partir de tejido nervioso, en las células neurales inmaduras en desarrollo llamadas "*neuroblastos"*. Normalmente se presenta en bebés y niños, aproximadamente el 37% de los casos se diagnostican en lactantes y el 90% corresponden a niños menores de 5 años (London et al., 2005). Esta condición fue descrita por primera vez en 1864 por el médico alemán Rudolf Virchow quién llamó gliomas a los tumores encontrados en los abdómenes de los niños ("Historia del Neuroblastoma", 2017).

En 1891, el patólogo alemán Felix Marchand describió por primera vez las características de los tumores que se desarrollan en el sistema nervioso simpático y la médula suprarrenal que se encuentra sobre los riñones (Pryse-Phillips, 2009).

Más tarde, en 1910, James Homer Wright señaló que estos tumores se originaron a partir de una forma inmadura de células neurales, y, por tanto, el nombre de "*blastoma*" se refiere a una colección de células inmaduras indiferenciadas (Rothenberg, Berdon, D'Angio, Yamashiro and Cowles, 2008).

Cushing y Wolbach (1927) descubrieron que no todos los neuroblastomas eran cancerosos. Mientras que unos eran malignos y se propagaban rápidamente a varios órganos del cuerpo como el *hígado*, *piel*, *hueso* y *médula ósea*, otros desaparecían sin tratamiento. En algunos casos raros, encontraron que los tumores se convirtieron en masas no malignas llamadas "*ganglioneuromas*", que también pueden resolverse por sí solos. Evenson y Cole (1966) añadieron que la transformación de formas cancerosas a no cancerosas era rara en bebés de más de 6 meses de edad.

Un hecho relevante para la detección de la enfermedad ocurrió cuando Mason et al. (1957) descubrieron que la presencia de catecolaminas[1] podía ser detectada en la orina de niños con Neuroblastoma. Las catecolaminas son las hormonas que se producen en grandes cantidades por los tumores, proporcionando un marcador para la presencia de la enfermedad.

Después de todo este proceso de descubrimiento del Neuroblastoma se conoce que se desarrolla a partir de tejidos que forman el sistema nervioso simpático. Esta es la parte del sistema nervioso que controla funciones del organismo como la frecuencia cardíaca, la presión

[1] Las catecolaminas (CA) o aminohormonas son todas aquellas sustancias que contienen en su estructura un grupo catecol y una cadena lateral con un grupo amino. Pueden funcionar en nuestro organismo como hormonas o como neurotransmisores (Silván, 2017).

arterial, la digestión y los niveles de ciertas hormonas ("Sistema nervioso simpático - Definición", 2017). Por lo general, suele comenzar con mayor frecuencia en las glándulas suprarrenales que se ubican en la parte superior de los riñones, aunque también puede desarrollarse en los tejidos nerviosos del cuello, tórax, abdomen o pelvis como se puede ver en la Figura 1.

El mayor problema es que el Neuroblastoma puede diseminarse a los huesos principalmente, pero también a la médula ósea, el hígado, los ganglios linfáticos, la piel y alrededor de los ojos. Estas formas metastásicas suelen representar el 50% de los casos ("Neuroblastoma", 2017).

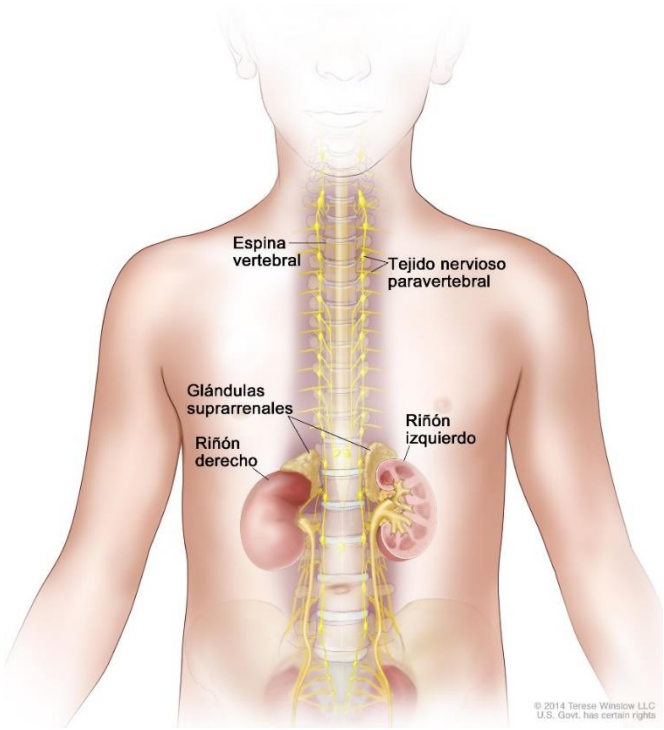


Figura 1. Esquema sobre los principales puntos donde se suele encontrar el Neuroblastoma ("Neuroblastoma", 2017).

El Neuroblastoma representa el 10% de los tumores sólidos de los niños menores de 15 años, con una incidencia anual de alrededor de 1/70.000 niños en esa franja de edad. Los primeros síntomas suelen ser fiebre, sensación de malestar general y dolor. También puede verse reflejada en la pérdida de peso y diarrea.

Como se ha mencionado antes, en niños el Neuroblastoma a veces desaparece de manera espontánea. En el resto de casos, puede que la cirugía sola sea suficiente, pero con frecuencia también se requieren otras terapias como quimioterapia, radioterapia o trasplante de células madre hematopoyéticas ("Neuroblastoma: MedlinePlus enciclopedia médica", 2017). Este

comportamiento clínico tan dispar ha sido relacionado con factores biológicos tales como la edad al diagnóstico, histología del tumor o aberraciones genéticas. Por ello, estos datos se incluyen en los sistemas de clasificación de pacientes.

Durante años, los distintos grupos de trabajo han utilizado factores y protocolos distintos a la hora de **clasificar pacientes**. Esto hacía que los resultados obtenidos de los estudios no fuesen comparables. Por ello, con el objetivo de posibilitar la comparación de trabajos de investigación, se crearon una serie de estándares internacionales: *International Neuroblastoma Staging System* (INSS), *International Neuroblastoma Risk Group Staging System* (INRGSS) and *International Neuroblastoma Risk Group* (INRG) *Classification System* (Monclair et al., 2009).

El INSS es un sistema de estatificación que depende del grado de resección quirúrgica. La clasificación original del INSS se muestra en la Tabla 1.

ESTADIO DEL TUMOR	DESCRIPCIÓN
1	Tumor localizado con extirpación macroscópica completa, con o sin enfermedad residual microscópica; ganglios linfáticos ipsilaterales negativos para el tumor microscópicamente. Los ganglios unidos y eliminados con el tumor primario pueden ser positivos.
2A	Tumor localizado con escisión macroscópica incompleta; ganglios linfáticos ipsilaterales no adherentes negativos para el tumor microscópicamente.
2B	Tumor localizado con o sin escisión macroscópica completa; con ganglios linfáticos ipsilaterales no adherentes positivos para el tumor; ganglios linfáticos contralaterales alargados negativos microscópicamente.
3	Tumor unilateral no operable que se infiltra a través de la línea media (más allá del lado opuesto de la columna vertebral) con o sin afectación de ganglios linfáticos regionales, o tumor de línea media con extensión bilateral con infiltración (inoperable) o afectación ganglionar.
4	Cualquier tumor primario con diseminación a ganglios linfáticos distantes, hueso, piel hígado, médula ósea y/u otros órganos (excepto aquello definido en el estadio 4S)
4S	Tumor primario localizado (definido como enfermedad en el estadio 1, 2A o 2B) con diseminación limitada a la piel, hígado y/o médula ósea (limitado a niños menores de un año, con afectación de la médula de al menos el 10% de las células nucleadas y resultados de la exploración MIBG negativos en la médula ósea).

Tabla 1. Descripciones de los estadios originales de tumores de la INSS (Brisse et al., 2011).

No obstante, dado que según este sistema el estadio es asignado tras una intervención quirúrgica, no es adecuado para agrupar pacientes en grupos de riesgo ya que se busca una clasificación pretratamiento. Por ello se creó un nuevo sistema de estatificación INRG (INRGSS) basado en criterios clínicos y factores de riesgo obtenido a partir de pruebas médicas (*image-*

defined risk factors, IDRFs). Dichos factores de riesgo basados en pruebas médicas quedan definidos en la Tabla 2. La clasificación que realiza el nuevo sistema de estatificación INRGSS se muestra en la Tabla 3, dividiendo la enfermedad en 4 subgrupos.

REGIÓN ANATÓMICA	DESCRIPCIÓN
Compartimientos múltiples del cuerpo	Extensión del tumor ipsilateral entre dos compartimientos del cuerpo (por ejemplo, cuello y pecho, pecho y abdomen, o abdomen y pelvis).
Cuello	Tumor que envuelve la arteria carótida, la arteria vertebral, y/o la vena yugular interna. Tumor que se extiende a la base del cráneo. Tumor que comprime la tráquea.
Unión cervicotorácica	Tumor que cubre las raíces del plexo braquial. Tumor que envuelve vasos subclavios, arteria vertebral y/o arteria carótica. Tumor que comprime la tráquea.
Tórax	Tumor que envuelve la aorta y/o ramas principales. Tumores que comprimen la tráquea y/o los bronquios principales. Tumor mediastínico inferior que se infiltra en la unión costovertebral entre los niveles vertebrales T9 y T12.
Unión toracoabdominal	Tumor que envuelve la aorta y/o la vena cava.
Abdomen y pelvis	Tumor infiltrante porta hepatis y/o ligamento hepatoduodenal. Tumor envolviendo ramas de la arteria mesentérica superior en la raíz mesentérica. Tumor que contiene el origen del eje celiaco y/o el origen de la arteria mesentérica superior. Tumor que invade uno o ambos pedículos renales. Tumor que contiene aorta y/o vena cava. Tumor que envuelve vasos ilíacos. Tumor pélvico que cruza la hendidura ciática.
Extensión de tumor intraespinal	Extensión del tumor intraespinal (sea cual sea la localización) siempre que se invada más de un tercio del conducto espinal en el plano axial, los espacios leptomeníngeos perimedulares no sean visibles o la intensidad de la señal de la médula espinal sea anormal.
Infiltración de órganos y estructuras adyacentes.	Pericardio, diafragma, riñón, hígado, bloqueo duodenopancreático y mesentrio.

Tabla 2. Descripción de los IDRFs (Brisse et al., 2011).

ESTADIO DEL TUMOR	DESCRIPCIÓN
L1	Tumor localizado que no envuelve estructuras vitales, como las definidas por la lista de IDRFs, y confinado a un compartimiento del cuerpo.
L2	Tumor local- regional con la presencia de 1 o más IDRFs.
M	Enfermedad metastásica distante (excepto aquello definido en el estadio MS).
MS	Enfermedad metastásica en niños menores de 18 meses de edad, con metástasis confinada a la piel, hígado y/o médula ósea.

Tabla 3. Descripciones de los nuevos estadios de tumores INRG (Brisse et al., 2011).

Con el uso de esta nueva estadificación (INRGSS) y clasificación de riesgo (INRG) del Neuroblastoma se facilita la comparación basada en hechos clínicos entre diferentes regiones del mundo.

La **causa del tumor** se desconoce, pero los expertos creen que parte del problema puede deberse a un defecto en los genes. Esta enfermedad se ha asociado con numerosas anomalías genéticas que condicionan el pronóstico: la amplificación del oncogén MYCN (2p24.3) (copias extras del oncogén MYCN) es un factor de pronóstico negativo; la triploidía y las anomalías numéricas cromosómicas están asociadas a un buen pronóstico, mientras que la di- o tetraploidía y las anomalías segmentarias cromosómicas (incluyendo pérdidas de 1p, de 11q, ganancia de 17q, etc.) están asociadas a un mal pronóstico. Recientemente, una mutación del gen ALK se ha descrito en el 12% de los casos (Reservados, 2017).

Recientemente, en abril de 2017 se realizó un estudio con 72 niños con Neuroblastoma (Hyun Lee et al., 2017) en el que las alteraciones genéticas más frecuentes se detectaron en los genes que se muestran en la Tabla 4.

GENES	PORCENTAJE DE NIÑOS EN LOS QUE SE DETECTÓ
ALK	16.7%
BRCA1	13.9%
ATM	12.5%
PTCH1	11.1%

Tabla 4. Genes involucrados en la aparición del neuroblastoma (Hyun Lee et al., 2017).

Pero la novedad vino con la detección de alteraciones en la secuencia en **ARID1B** en 5 de los 72 niños sobre los que se hacía el estudio (6.9%), y todos ellos se relacionaban con tumores agresivos de Neuroblastoma.

Existen gran cantidad de **bases de datos biomédicas** en las que se puede extraer conocimiento para la predisposición a la enfermedad, basándose en qué genes se relacionan con el Neuroblastoma o cualquier otra enfermedad. En ellas se recogen los datos disponibles y proporcionados por investigaciones y análisis científicos. Entre las más importantes se encuentra, por ejemplo, *ClinVar*, *dbGaP* o *dbSNP*, bases de datos de mutaciones de acceso libre donde se puede obtener gran cantidad de información. Estas brindan cantidad de beneficios ya que a partir de estos datos es posible extrapolar conocimiento para aplicarlo a los pacientes con predisposición a una cierta enfermedad.

Aunque son una fuente de información muy potente, también poseen problemas. El mayor problema es la heterogeneidad de los datos ya que la información se encuentra de manera muy dispersa y, al realizar una búsqueda, pueden aparecer problemas de redundancias o falta de calidad en los datos.

Los genetistas utilizan estas bases de datos, pero hay que tener en cuenta que la manipulación de estos datos no es trivial. Según sobre qué base de datos se realice la búsqueda, los resultados obtenidos pueden ser diferentes. Además, gran parte del trabajo es manual, por esto, en los resultados presentados pueden aparecer errores debidos al factor humano. Es por esto que los genetistas o médicos precisan de sistemas software que realicen esta tarea de gestión de datos de manera que se facilite una base de datos común en la que se pueda concentrar toda la información de las diferentes fuentes con el fin de explotar el conocimiento genómico existente. Además, los datos son cambiantes debido a los descubrimientos que se realizan día a día, ya que es un campo en constante evolución. Por este motivo, la base de datos debería actualizarse con frecuencia.

Por otro lado, existen diferentes **grupos de investigación y asociaciones** que tratan de buscar solución a la enfermedad en cuestión. Un ejemplo sería la asociación NEN[2] formada por un grupo de padres unidos por la firme determinación de luchar contra el Neuroblastoma. La asociación apoya proyectos de investigación relacionados con el tratamiento de la enfermedad, además de realizar campañas que apoyen a los afectados y reúnan fondos los cuales se destinan a la investigación. Un ejemplo de los proyectos que NEN apoya económicamente es el estudio de cambios epigenéticos, es decir, conjunto de reacciones químicas y demás procesos que modifican la actividad del DNA sin alterar la secuencia, que pueden contribuir al desarrollo y al comportamiento clínico de la enfermedad con el fin de identificar nuevas dianas terapéuticas.

También en la Fundación de investigación Sant Joan de Déu hay un grupo de investigación coordinado por Jaume Mora que estudian el origen del Neuroblastoma con la caracterización fenotípica y genotípica de las diferentes poblaciones celulares que conforman el Neuroblastoma.

[2] Asociación NEN (Niños Enfermos de Neuroblastoma), http://asociacion-nen.org/.

En Valencia, está el *Instituto de investigación Sanitaria La Fe*[3] que coordina a nivel nacional los tratamientos y diagnóstico del Neuroblastoma según protocolo europeo SIOPEN (*Sociedad Internacional de Oncología Pediátrica-Europa*), además de contar con varios ensayos clínicos. También el *Instituto de Investigación Sanitaria INCLIVA*[4] de Valencia que realiza la mayoría de análisis patológicos de las muestras tumorales de los pacientes españoles y lleva a cabo varias investigaciones sobre el Neuroblastoma.

Junto el Centro de Investigación en Métodos de Producción de Software (PROS[5]) de la Universidad Politécnica de Valencia (UPV) se ha trabajado con la Unidad de Oncología Pediátrica de La Fe, con las doctoras Adela Cañete y Victoria Castel, para diseñar un sistema de información para gestionar de forma conjunta los datos clínicos y genómicos sobre el Neuroblastoma. Se trata de una aplicación web que permite gestionar la información almacenada en una base de datos sobre el diagnóstico y tratamiento de pacientes que pasan o han pasado por esta unidad hospitalaria, ya que también se introdujeron datos de carácter retrospectivo que las doctoras habían ido almacenando durante años. El objetivo final es, además de gestionar la información de los pacientes, poder analizarla de forma conjunta para obtener información relevante sobre el estado actual de la enfermedad, así como poder obtener datos estadísticos.

Mediante la gestión de bases de datos curadas se buscar mejorar la atención al paciente, así como actuar en la prevención de la enfermedad, ya que los factores genéticos juegan un papel fundamental en la salud de las personas. Todos los esfuerzos se basan en integrar la información disponible con el objetivo de generar diagnósticos genéticos que permitan mejorar la calidad de vida de los pacientes proporcionando a la población una medicina mucho más personalizada.

2.2 EMPRESAS- DIAGNÓSTICO GENÓMICO

Como resultado de estos avances en NGS, en la actualidad existen numerosas empresas que facilitan *tests genéticos* a la población con el fin de ofrecer un diagnóstico precoz acerca de alguna enfermedad de origen genético.

Los *tests genéticos* identifican cambios en los cromosomas, genes o proteínas. Los resultados pueden confirmar o descartar una condición genética que había sido sospechada o ayudar a determinar la posibilidad de una persona a desarrollar o transmitir un trastorno genético.

Un ejemplo es **23andme** (*https://www.23andme.com/*). Se trata de una empresa fundada en 2006 por Linda Avey, Paul Cusenza y Anne Wojcicki. 23andme ofrece tests genéticos utilizando un paquete personal de prueba genética para determinar las variaciones de un individuo con respecto a una enfermedad determinada y diversos rasgos relacionados con su salud, para ello analizan solamente una muestra de saliva del cliente. El paquete salió a la venta en 2007 por el

[3] www.iislafe.es/
[4] www.incliva.es/
[5] www.pros.webs.upv.es/

precio de 999 dólares, pero las financiaciones han conseguido reducir este precio hasta 99 dólares (Zettler, Sherkow & Greely, 2014).

23andme permite generar listas de variaciones y sus estudios son en base a probabilidades y aspectos ancestrales. El tipo de información que ofrece el kit es el siguiente:

- *Historia genética*: Buscar parientes, personas que comparten la misma información genética y han hecho uso del paquete, así como conocer los orígenes globales a nivel de población mundial.
- *Salud personal*: Conocer si el consumidor es portador de alguna enfermedad y si sus descendientes directos podrán sufrirla. Proporcionar información sobre el riego a padecer ciertas enfermedades con el fin de actuar en la prevención y por último proporcionar un perfil de respuesta personal a ciertos fármacos en cuanto a consecuencias, dosis y efectos secundarios.

De igual forma, en España, se han desarrollado empresas de este tipo como, por ejemplo: Genotest[6], Imegen[7] o TellMeGen[8], con el mismo objetivo de proporcionar a los usuarios tests genéticos de manera sencilla y aportándoles un diagnóstico que les sirva para actuar sobre su salud.

[6] www.trkgenetics.com/genotest
[7] www.imegen.es/
[8] www.tellmegen.com/

CAPÍTULO 3. MATERIALES Y METODOLOGÍA

3.1. BASES DE DATOS GENÓMICAS

Las bases de datos[9] utilizadas en este trabajo para la búsqueda de las variaciones relacionadas con la enfermedad de Neuroblastoma han sido: dbGaP (*https://www.ncbi.nlm.nih.gov/gap/?term*=) y ClinVar (*https://www.ncbi.nlm.nih.gov/clinvar/*).

Por un lado, **ClinVar** es una base de datos de mutaciones, de acceso gratuito donde se encuentran los reportes de relaciones entre variaciones genéticas humanas y fenotipos, con evidencias que soportan esta relación indicando cuantos organismos de investigación han detectado esa variación y la información que han aportado sobre ello (Landrum et al., 2013).

ClinVar procesa las notificaciones sobre variantes encontradas en muestras de pacientes, las afirmaciones hechas con respecto a su significado clínico, información sobre la persona u organismo que proporciona la información y otros datos de apoyo. Los alelos descritos en las notificaciones son mapeados con las secuencias de referencia y se presentan de acuerdo al estándar HGVS (*Human Genome Variation Society*). La nomenclatura HGVS se usa para presentar e intercambiar información de variaciones encontradas en secuencias de ADN, ARN y proteínas y sirve como un estándar internacional (por ejemplo, si se quiere representar una variación G/T en la posición chr19:11087877, según dicho estándar, se representaría como NC_000019.8:g.11087877G>T, donde NC_000019.8 es el número único de acceso a la secuencia usada para posicionar la variación (en NCBI *RefSeq*), la letra g significa que la secuencia es genómica, 11087877 corresponde a la posición en la secuencia referida y G>T describe el cambio de Guanina por Timina). De esta manera, ClinVar presenta los datos tanto para aquellos que estén interesados como para los que necesiten este tipo de información en su trabajo diario. ClinVar trabaja en colaboración con organizaciones interesadas en satisfacer las necesidades de la comunidad médica (genética) de la manera más eficiente y efectiva posible.

ClinVar pertenece al NCBI, que es el Centro Nacional para la Información Biotecnológica (https://www.ncbi.nlm.nih.gov/) y que proporciona acceso a 41 bases de datos (como, por ejemplo, dbVar, dbGaP, Pubmed, Gene, OMIM, entre ellas ClinVar). Se trata de una base de datos con gran cantidad de información. Las estadísticas que se obtuvieron para Clinvar con fecha 23 de mayo de 2017 son las que se muestran en la Tabla 5.

[9] Una base de datos o un banco de datos (designado en ocasiones con la sigla BD o la abreviatura b.d.) es un conjunto de datos ordenados según ciertas reglas y criterios pertenecientes a un mismo contexto, y almacenados sistemáticamente para su posterior uso.

CATEGORY	TOTAL (MAY 23, 2017)
Records submitted	485987
Unique variation records	318582
Total genes represented	27742
Total submiitters	712

Tabla 5. Estadísticas de ClinVar en mayo de 2017.

En segundo lugar, se ha hecho uso de **dbGaP**, una base de datos también dependiente de NCBI que contiene resultados de estudios, los cuales consisten en la investigación de la interacción entre genotipo y fenotipo, incluyendo estudios de asociaciones de genoma, secuenciación médica, ensayos de diagnóstico molecular, así como asociaciones entre genotipo y características no clínicas. Fue desarrollada para archivar y distribuir los datos y los resultados de investigaciones ("Home - dbGaP - NCBI", 2017).

Según las estadísticas de 2017, dbGaP dispone de 852 estudios registrados como se muestra en la Figura 2. Se ve el enorme aumento en el número de estudios que han habido en los últimos 10 años, en donde para el año 2007 se partía solo de 9 estudios.

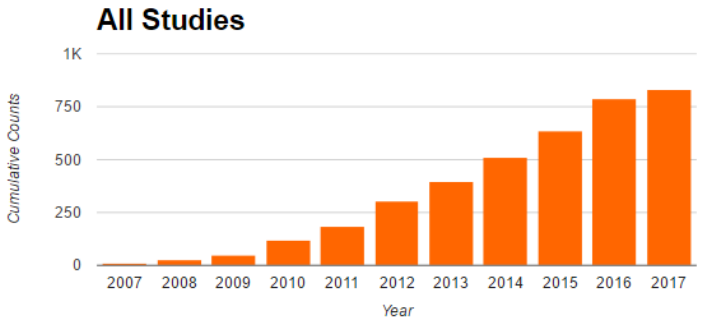


Figura 2. Número de estudios acumulados en dbGaP hasta el 2017 ("Home - dbGaP - NCBI", 2017).

Por otro lado, también se muestra información acerca de la procedencia de los investigadores que proporcionan los estudios a la base de datos. El país que más investigadores proporcionan estudios es EEUU con un total de 3800, seguido de Reino Unido y Canadá como se puede ver en la Figura 3. DbGaP dispone de un total de 5367 investigadores cuyos datos proporcionados han sido aprobados.

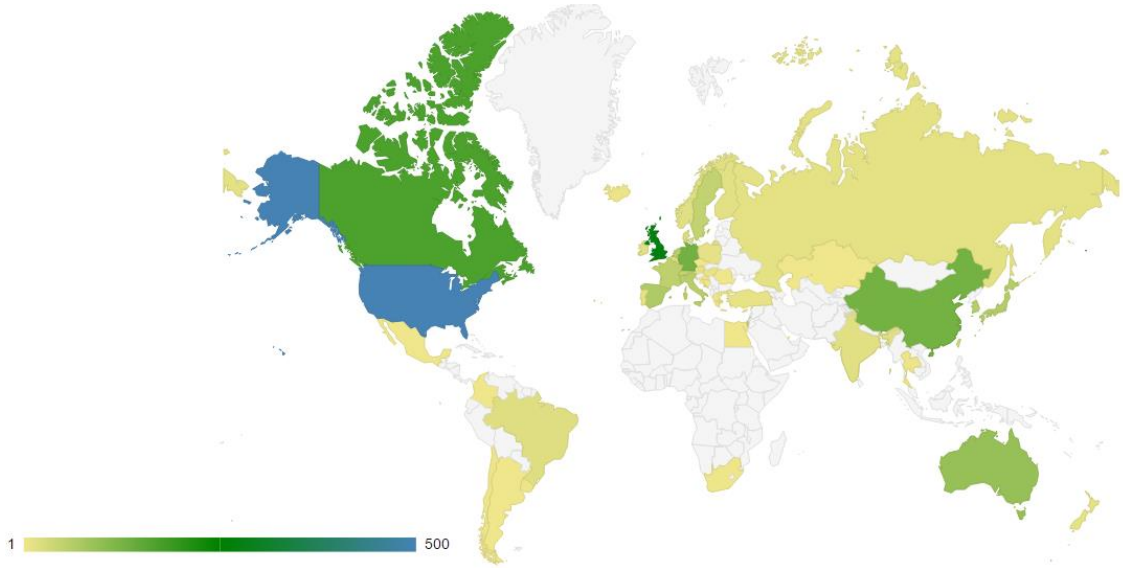


Figura 3. Número de investigadores por país en 2017 ("Home - dbGaP - NCBI", 2017).

3.2. METODOLOGÍA SILE

Para abordar el problema a resolver en el presente trabajo, se hará uso de la **metodología SILE**, la cual ha sido desarrollada por el Centro PROS de la UPV. Esta metodología consiste en una serie de pasos que permiten la obtención de toda la información necesaria y relevante sobre una enfermedad genética (Reyes Román, J. F. and Pastor López, Ó., 2016). Los pasos son los siguientes:

- *Search* (Búsqueda)
  En este primer paso se busca toda la información que tiene que ver con la enfermedad en cuestión y se analizan detenidamente las fuentes científicas disponibles como artículos, bases de datos, etc… con el fin de determinar cuáles son las más óptimas para obtener la información deseada y relevante.
- *Identification* (Identificación)
  Se realiza un filtrado de la información obtenida en el paso anterior (*Search*), extrayendo así la más relevante y evitando repeticiones o carencias en la calidad de los datos. Al final de este paso, se obtendrán aquellas variaciones que están profundamente relacionadas con la enfermedad.
- *Load* (Carga)
  En esta tarea se realiza la carga de todas las variaciones identificadas a una base de datos, en este caso a la HGDB desarrollada también por el grupo de investigación PROS basada en el Modelo Conceptual del Genoma Humano (MCGH) (Reyes, J.F et al., 2016). Para ello

se hace un mapeado entre la información obtenida en el paso de *Identification* y la arquitectura de la base de datos destino.

- *Exploitation* (Explotación)
  En este último paso se convierten los datos ya cargados en conocimiento para poder ser explotados. En este caso se utilizará la herramienta llamada "*VarSearch*".

Esta es la metodología que se va a utilizar para el desarrollo del trabajo ya que permite obtener toda la información necesaria y relevante sobre una enfermedad genética, Figura 4.

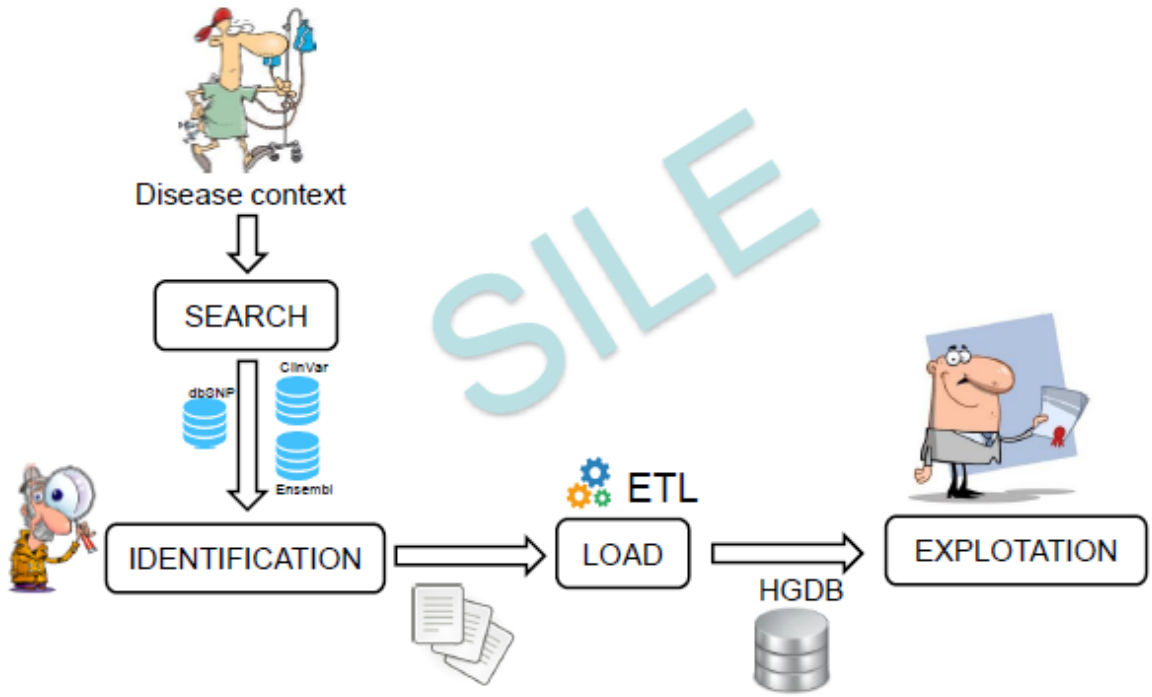


Figura 4. Pasos metodología SILE (León, A., 2017).

3.3. *HUMAN GENOME DATABASE* (HGDB)

La carga de las variaciones detectadas y validadas en los pasos de búsqueda e identificación se va a realizar a la **Base de Datos del Genoma Humano** (HGDB, de sus siglas en inglés "*Human Genome Database*"). La HGDB fue desarrollada siguiendo el MCGH descrito previamente con el fin de manejar eficientemente los datos genómicos (Reyes, J.F et al., 2016). Es por esto por lo que en la HGDB se muestran las vistas extraídas del **Modelo Conceptual del Genoma Humano (MCGH)** (Reyes Román, León Palacio & Pastor López, 2017). La creación del modelo conceptual y la caracterización ontológica asegura consistencia, corrección y una explotación eficiente de estas bases de datos especializadas. Tanto la HGDB como el MCGH se desarrollaron con la finalidad de

definir todas las características que definen el genoma humano y crear una estructura sólida sobre la cual construir un **GeIS** que ayude a explotar el conocimiento genómico por medio de la Medicina de Precisión (Burriel, V. et al., 2017).

La base de datos del genoma humano era de dos vertientes. En un primer momento se creó la base de datos con fines de investigación, donde se realizaba una carga masiva de todos los datos disponibles en la web (por ejemplo, *dbSNP*, *Ensembl*, etc.). Estos datos después ya podían ser utilizados y filtrados por los genetistas e investigadores según las necesidades que tuviesen.

Más tarde, se quiso utilizar la base de datos con el objetivo de generar diagnósticos genéticos (impulsar la *medicina personalizada*), por lo que se pasó a la "*Carga Selectiva"* de los datos ("*Curated*"). Este tipo de carga se realizaba siguiendo la metodología SILE. Esta última vertiente de la HGDB es la que vamos a usar en el presente trabajo.

Para la carga de los datos se crearán **ficheros CSV** con las variaciones detectadas. Estos ficheros serán cargados utilizando un gestor de base de datos con su herramienta de importación de datos, en este caso, se utilizará **HeidiSQL** (https://www.heidisql.com/).

3.4. *VARSEARCH*

Una vez realizada la carga, este repositorio podrá ser explotado mediante la herramienta ***VarSearch***, desarrollada por el grupo PROS y utilizada como mecanismo de validación de la tesis doctoral de José F. Reyes R.

Se trata de una aplicación web capaz de analizar ficheros generados en tecnologías NGS (*Next-Generation Sequencing*) (Rodríguez-Santiago and Armengol, 2012) como, por ejemplo, SANGER (*.fasta) o VCF. A partir de la muestra de un paciente se pueden obtener sus variaciones y presentar los resultados con este tipo de archivos que son analizados por *VarSearch*. Estos ficheros contienen las variaciones identificadas en el paciente, *VarSearch* realiza una búsqueda en el repositorio de información genómica creado y divide las variaciones contenidas en dos listas de resultados:

- Por un lado, devolverá una lista con las **variaciones encontradas**, es decir, aquellas que se encuentran en el fichero y también en la base de datos.
- Por otro lado, se creará una lista con las **variaciones no encontradas**, estas serán aquellas que se encuentren en el fichero, pero no en la base de datos.

Con la obtención de estos resultados y la creación de esta herramienta de diagnóstico la medicina pasa a ser mucho más personalizada, adaptándose a las necesidades de cada paciente.

CAPÍTULO 4. APLICACIÓN MÉTODO SILE

4.1 *SEARCH*

En primer lugar, se aplica el paso 1 de la metodología SILE.

Se ha realizado un amplio análisis de algunas de las bases de datos disponibles con el fin de seleccionar aquellas que cubran las necesidades del proyecto. Para elegir aquellas sobre las que trabajar se han seleccionado aquellas de mayor importancia entre la comunidad científica.

Las diferentes bases de datos que se encuentran en la web se pueden dividir según su contenido en los tipos que se muestran en la Tabla 6.

BASES DE DATOS			
SECUENCIAS	**PROTEÍNAS**	**MUTACIONES**	**PUBLICACIONES**
GenBank	BioGrid	dbGaP	Pubmed
Ensembl	UniProtKB	dbSNP	
RefSeq		ClinVar	

Tabla 6. Ejemplos de diferentes tipos de bases de datos.

Un ejemplo de base de datos de secuencia es **GenBank**[10] .Es una base de datos de secuencias del Instituto Nacional de Salud (*National Institutes of Health*, NIH), este repositorio consiste en una colección de secuencias de ADN disponibles públicamente. GenBank forma parte de la Colaboración en la base de datos internacional de secuencias de nucleótidos INSDC[11], que comprende DDBJ[12], el ENA[13] y GenBank del NCBI[14]. Estas tres instituciones comparten sus datos diariamente con la coordinación de INSDC. Además, **PubMed** complementa la información contenida en GenBank mediante una colección anotada de artículos científicos.

[10] https://www.ncbi.nlm.nih.gov/genbank/
[11] *International Nucleotide Sequence Database Collaboration*, INSDC. http://www.insdc.org/
[12] Banco de ADN de Japón (*DNA DataBank of Japan, DDBJ)*
[13] Archivo de nucleótidos europeo *(European Nucleotide Archive, ENA*)
[14] Centro Nacional para la Información Biotecnológica (*National Center for Biotechnology Information*, NCBI)

Por otro lado, **Ensembl**[15], perteneciente al EBI[16], es un buscador de genomas de vertebrados, en total unas 80 especies diferentes, que apoya la investigación en la genómica comparativa, la evolución, la variación de secuencia y la regulación transcripcional.

De la misma manera, **RefSeq**[17] (también perteneciente a NCBI) proporciona un set de secuencias integradas, no redundantes y bien anotadas que incluyen información relacionada con: ADN genómico, transcritos y proteínas.

Por otro lado, existen bases de datos de proteínas como por ejemplo **BioGrid**[18] o **UniProtKB**[19]. BioGrid es una base de datos que archiva y divulga datos de interacción genética y proteica de organismos modelo[20] y humanos. Pero la base de datos principal de proteínas es UniProtKB, la cual incluye ontologías biológicas ampliamente aceptadas, clasificaciones y referencias cruzadas e indicaciones claras de la calidad de la anotación sobre la atribución de la evidencia de los datos experimentales y computacionales.

Las bases de datos de mutaciones son las más interesantes para el presente trabajo. Algunos ejemplos serían **dbSNP**, **dbGaP** y **ClinVar**, todas pertenecientes al NCBI. Como se ha mencionado en el capítulo anterior, ClinVar y dbGaP son bases de datos que contienen información acerca de variaciones genéticas y fenotipos. dbSNP es un repositorio central de variaciones genéticas que comprenden sustituciones simples de nucleótidos, llamados SNPs (Figura 5) y polimorfismos de inserciones y deleciones cortas.

[15] www.ensembl.org/index.html
[16] Instituto de Bioinfomática Europeo (*European Bioinformatics Institute*, EBI)
[17] www.ncbi.nlm.nih.gov/refseq/
[18] https://thebiogrid.org/
[19] www.uniprot.org/help/uniprotkb
[20] Especies no humanas que se estudian profundamente para entender fenómenos biológicos particulares con la expectativa de que los descubrimientos realizados en el modelo del organismo proporcionen una visión del funcionamiento de otros organismos.

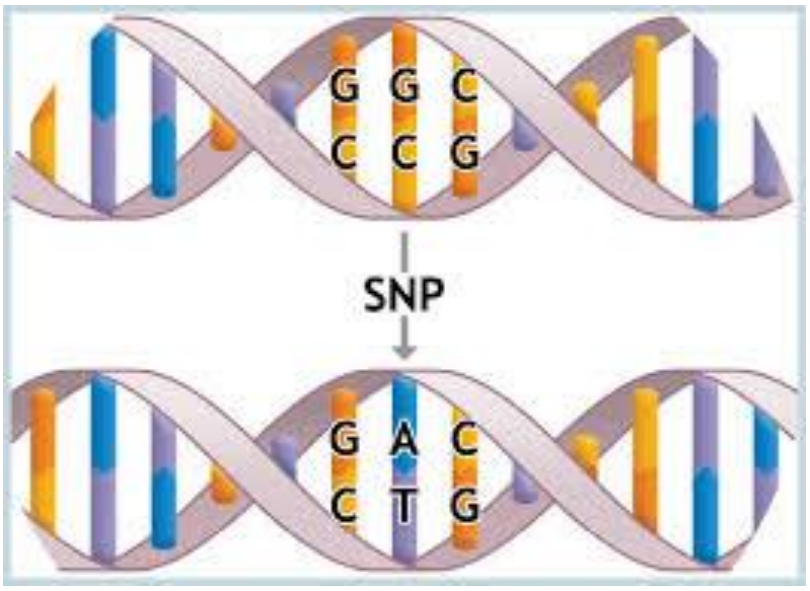


Figura 5. Ejemplo de Single-nucleotide polymorphism (SNP).

Además, existen otros repositorios como **Pubmed** que comprende más de 27 millones de citaciones para literatura biomédica procedente de MEDLINE[21], revistas científicas y libros online. Las citaciones suelen incluir enlaces al contenido del texto completo de la Central de Pubmed o al sitio web en que se ha publicado.

Se decide realizar la búsqueda en las bases de datos de mutaciones ya que en ellas se describen las relaciones genotipo-fenotipo para cada enfermedad. Como se ha mencionado antes, las consultas se realizan en los repositorios de datos genómicos de ClinVar y dbGaP debido a la magnitud e importancia de estos repositorios hoy en día, además de que ambos se componen de datos curados, es decir con un nivel de revisión que asegura una confianza en los datos. Se hará uso de otras bases de datos para la búsqueda de información complementaria y necesaria también para la creación de un repositorio completo sobre la enfermedad del Neuroblastoma. Estos otros repositorios se irán describiendo conforme el proceso de búsqueda de las variaciones los haya requerido.

[21] MEDLINE es una base bibliográfica autorizada que contiene citas y resúmenes para revistas biomédicas y de salud utilizadas por profesionales de la salud, enfermeras, clínicos e investigadores dedicados a la atención clínica, la salud pública y el desarrollo de políticas de salud.

4.1.1 ClinVar

En primer lugar, se realiza una primera búsqueda en ClinVar. Se accede desde NCBI y como término de búsqueda se utiliza "*neuroblastoma[Disease/Phenotype]",* se trata de una búsqueda avanzada en la que se indica que se quiere buscar por Neuroblastoma como enfermedad. De esta primera búsqueda en ClinVar se obtienen **359 variaciones**.

ClinVar permite hacer una descarga automática en formato de texto tabular, el cual se exporta al Excel. Para cada variación, en esta descarga automática, se obtienen los campos que se muestran en la Tabla 6.

	DESCRIPCIÓN	**EJEMPLO**
Name	Nombre de la variación en nomenclatura HGVS	NM_015074.3(KIF1B):c.-260C>T
Gene(s)	Nombre del gen	KIF1B
Condition(s)	Fenotipo/s a los que se asocia la variación	Neuroblastoma|Pheochromocytoma|Charcot-Marie-Tooth, Type 2
Frequency	Frecuencia global del alelo menor	GMAF:0.00280(T)
Clinical significance (Last reviewed)	Significancia clínica y cuándo ha sido evaluada por última vez	Likely benign(Last reviewed: Jun 14, 2016)
Review status	Estado de la revisión	criteria provided, single submitter
Chromosome	Cromosoma	1
Location	Posición de la variación dentro del cromosoma	10210698
Assembly	Identificador GRCH para la secuencia de referencia	GRCh38
VariationID	Identificador que asigna ClinVar a cada set de variaciones en las presentaciones	368793
AlleleID	Identificador único que asigna ClinVar a cada variación individual.	353022

Tabla 7. Datos obtenidos en la descarga automática de ClinVar.

Al hacer click en cada una de las variaciones, ClinVar presenta una serie de datos clasificados en tres (3) apartados:

1. **Interpretación**

El primer apartado corresponde al de "*interpretación*" (Figura 6), en él se muestra de nuevo a qué enfermedades está asociada la variación. Al lado del nombre de cada condición se ofrecen enlaces a otras bases de datos como MedGen (*https://www.ncbi.nlm.nih.gov/medgen/*) u OMIM (https://www.omim.org/) las cuales contienen información sobre las afecciones correspondientes.

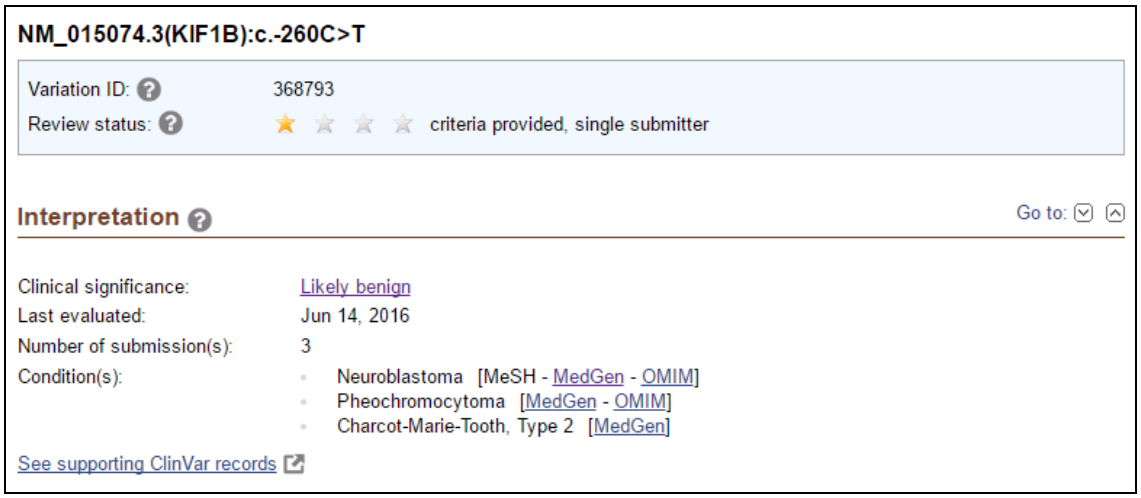


Figura 6. Ejemplo información del apartado interpretación de ClinVar.

2. **Alelos**

El apartado de "*Alelos*" contiene más información sobre la variación en sí (Figura 7).

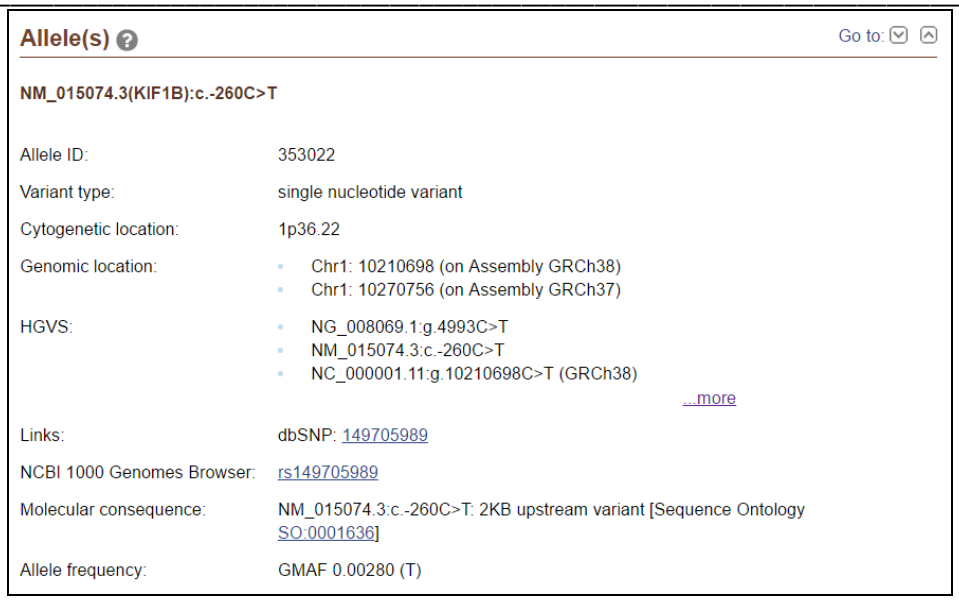


Figura 7. Ejemplo del apartado Alelos en ClinVar.

En primer lugar, aparece el número de identificación del alelo (**AlleleID**), seguida del tipo de variación. El **tipo de variaciones** que se pueden encontrar serían las que se muestran en la Tabla 8.

TIPO DE VARIACIÓN	DESCRIPCIÓN
Single nucleotide variant	Cambios en una sola base
Insertion	Inserción de 1 o varios nucleótidos
Deletion	Deleción de 1 o varios nucleótidos
Indel	Inserción o deleción que afectan a 2 o más nucleótidos
Inversion	Una secuencia continua de nucleótidos se invierte en la misma posición
Copy Number Gain	Secuencias repetidas cuyo número de repeticiones aumenta
Copy Number Loss	Secuencias repetidas cuyo número de repeticiones disminuye
Duplication	Duplicación
Tandem duplication	Duplicación en tándem
Microsatellite	Regiones cortas repetidas en tándems

Tabla 8. Tipos de variaciones.

Además, también ofrece la **localización citogenética**. Los genetistas utilizan este tipo de localización, la cual es una combinación de número y letras que proporciona la posición exacta en un cromosoma. Estas son las partes de las que se compone (Reference, 2017):

a) El *número o letra del cromosoma* en el cual se encuentra. Pueden ir de 1 a 22 que designan el cromosoma autosómico o, X e Y para designar los cromosomas sexuales.
b) El *brazo el cromosoma*. Cada cromosoma está dividido en dos brazos o secciones, el brazo más corto es llamado p mientras que el largo q. Esta largaría se basa en la distancia a la parte más estrecha del cromosoma llamada centrómero.
c) *Posición del gen* sobre el brazo p o q. La posición del gen se basa en un patrón distintivo de bandas claras u oscuras que aparecen cuando el cromosoma se tiñe de una manera específica lo que se conoce como bandeo G. Normalmente, la posición se designa con dos dígitos que representan la región y la banda. A veces, les siguen un punto decimal y uno o más dígitos adicionales que representan las sub-bandas en el área clara u oscura. El número que indica la posición del gen aumenta con la distancia al centrómero.

Un ejemplo sería el que se muestra en la Figura 8, donde 7q31.2 corresponde a la localización citogenética de un gen que se encuentra en el brazo largo del cromosoma 7, concretamente en la sub-banda 2 de la banda 1 dentro d la región 3.

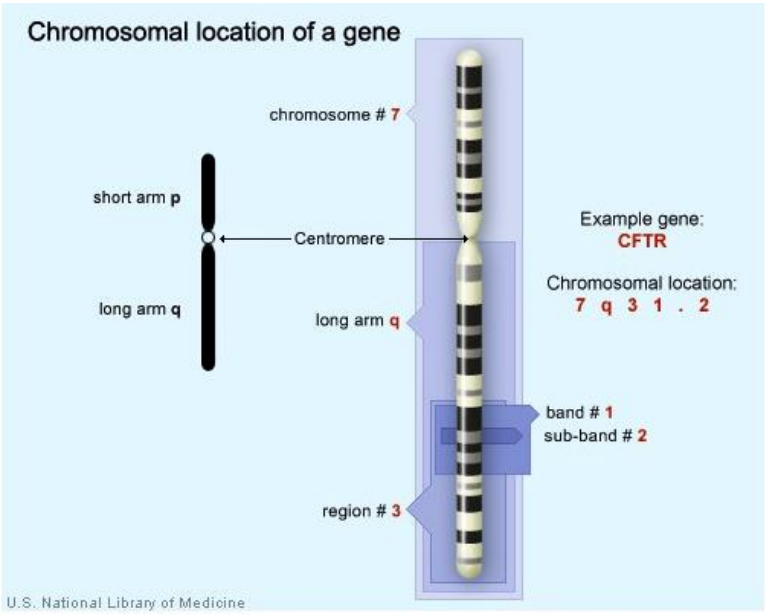
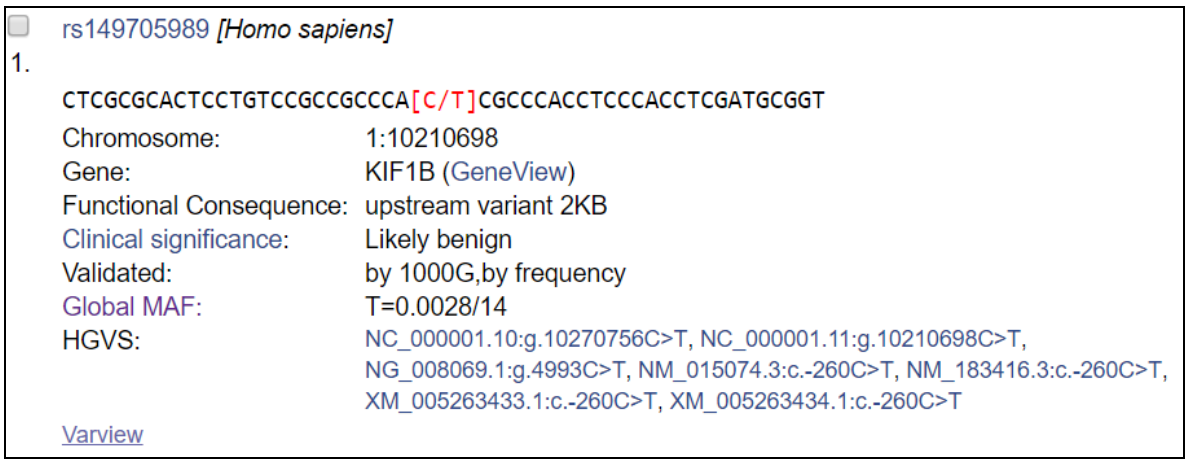


Figura 8. Ejemplo localización citogenética, 7q31.2 ("National Library of Medicine - National Institutes of Health", 2017).

Seguida de la localización citogenética se muestra la **localización genómica** (tanto respecto al genoma de referencia GRCh37 como al GRCh38) y los nombres en **nomenclatura HGVS**, además de la **consecuencia molecular** y la **frecuencia alélica**. También proporciona enlaces al repositorio de datos "dbSNP" utilizando el identificador "rs" que describe una variación de único nucleótido de manera unívoca. Entrando en el enlace a dbSNP se puede obtener información acerca del cual es el alelo de referencia y cual el alternativo, es decir, el que había antes y después de que se produjese la variación. También se encuentran las 25 bases por delante de la posición de la

variación y las 25 bases posteriores. Esta información se muestra como en la Figura 9. Además, si se vuelve a pulsar en esta ventana en el identificador "rs", aún se puede obtener más información, como, por ejemplo, la dirección de la cadena en la que se ha encontrado la variación, que puede ser hacia delante o hacia atrás.

El enlace a 1000Genomes Browser (https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) lleva a una web en la que el usuario puede explorar las variaciones, genotipos y las lecturas de alineamientos de secuencias que se han producido en un proyecto llamado 1000 Genomas. En este proyecto se pretendían alinear 1000 genomas a GRCh38, pero ya se han analizado más de 2500 secuencias genómicas ("1000 Genomes | A Deep Catalog of Human Genetic Variation", 2017).


rs149705989 *[Homo sapiens]*

1.

CTCGCGCACTCCTGTCCGCCGCCCA[C/T]CGCCCACCTCCCACCTCGATGCGGT

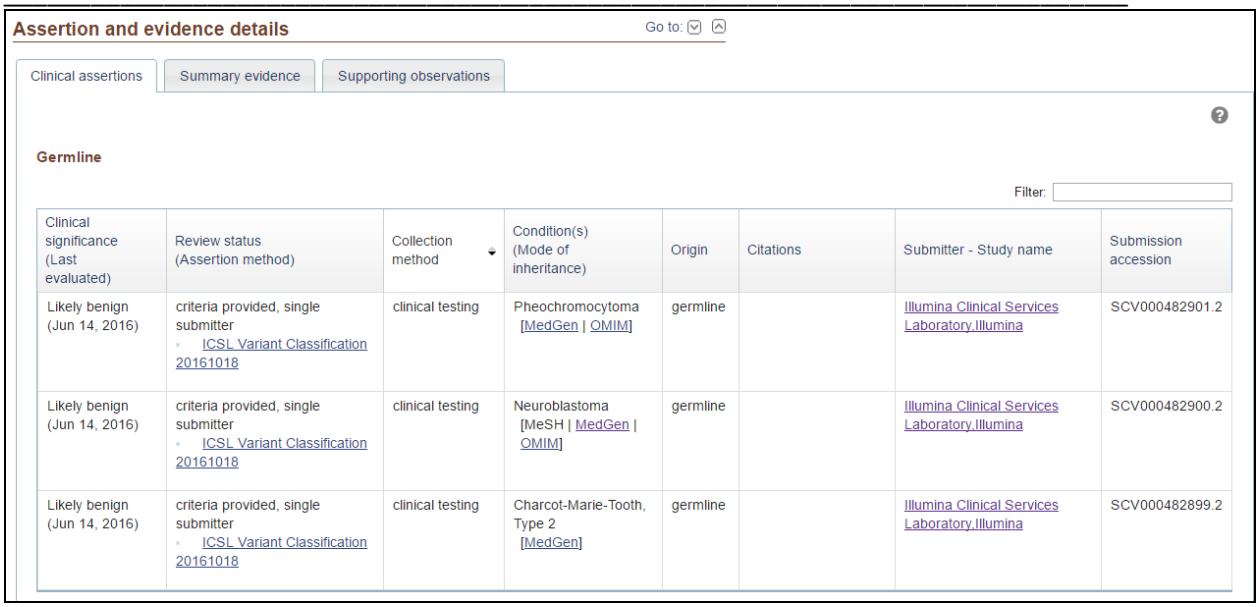
Chromosome:	1:10210698
Gene:	KIF1B (GeneView)
Functional Consequence:	upstream variant 2KB
Clinical significance:	Likely benign
Validated:	by 1000G,by frequency
Global MAF:	T=0.0028/14
HGVS:	NC_000001.10:g.10270756C>T, NC_000001.11:g.10210698C>T, NG_008069.1:g.4993C>T, NM_015074.3:c.-260C>T, NM_183416.3:c.-260C>T, XM_005263433.1:c.-260C>T, XM_005263434.1:c.-260C>T

Varview

Figura 9. Ejemplo de información que se proporciona en el enlace a dbSNP.

3. Assertion and evidence details

En el último apartado, llamado "*assertion and evidence details*" (Figura 10) aparecen todas las veces que ha sido validada dicha relación. Sobre cada validación aparece información como la fecha en la que fue realizada y la significancia clínica que se le dio, la condición a la que se asoció la variación o el nombre del *"submitter"* o del estudio con un enlace a él. Además, se muestra como los datos han sido obtenidos ("*Collection method*"). Este puede ser de distintos tipos, por ejemplo: testeo clínico ("*clinical testing*"), investigación ("*research*"), o solo a partir de literatura (*"literature only*").



Assertion and evidence details Go to:

Clinical assertions | Summary evidence | Supporting observations

Germline

Filter:

Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter - Study name	Submission accession
Likely benign (Jun 14, 2016)	criteria provided, single submitter ICSL Variant Classification 20161018	clinical testing	Pheochromocytoma [MedGen | OMIM]	germline		Illumina Clinical Services Laboratory,Illumina	SCV000482901.2
Likely benign (Jun 14, 2016)	criteria provided, single submitter ICSL Variant Classification 20161018	clinical testing	Neuroblastoma [MeSH | MedGen | OMIM]	germline		Illumina Clinical Services Laboratory,Illumina	SCV000482900.2
Likely benign (Jun 14, 2016)	criteria provided, single submitter ICSL Variant Classification 20161018	clinical testing	Charcot-Marie-Tooth, Type 2 [MedGen]	germline		Illumina Clinical Services Laboratory,Illumina	SCV000482899.2

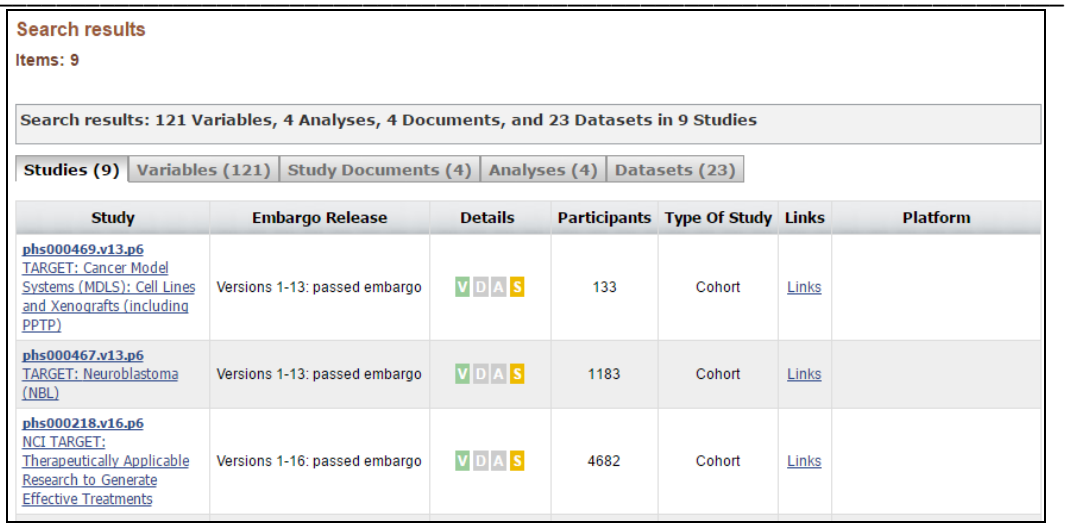
Figura 10. Ejemplo apartado "Assertion and evidence details".

Se han guardado los campos obtenidos en la descarga automática de ClinVar junto con otros datos relevantes sobre cada una de la variación, ya que lo que se pretende es caracterizar cada una de ellas de la manera más completa. Los campos que finalmente se almacenaron para la carga en la HGDB fueron los mismos que para las variaciones obtenidas en dbGaP y se describirán más adelante en la Tabla 10.

4.1.2 dbGaP

dbGaP es una base de datos que contiene **resultados de estudios**, los cuales han investigado la interacción entre genotipo y fenotipo, incluyendo estudios de asociaciones (GWAS, *Genome Wide Association Study*) de genoma, secuenciación médica, ensayos de diagnóstico molecular, así como asociaciones entre genotipo y características no clínicas. Fue desarrollada para archivar y distribuir los datos y los resultados de investigaciones científicas. Dichos estudios pueden facilitar la priorización de las variaciones y la generación de hipótesis biológicas (Tryka et al., 2013).

Es por esto que si se coloca directamente en el buscador "*Neuroblastoma*" lo que aparece no son las variaciones que están relacionadas con la enfermedad, sino que aparecen estudios en los que se ha investigado sobre asociaciones de genotipo y Neuroblastoma como se muestra en la Figura 11.



Search results

Items: 9

Search results: 121 Variables, 4 Analyses, 4 Documents, and 23 Datasets in 9 Studies

Studies (9) | Variables (121) | Study Documents (4) | Analyses (4) | Datasets (23)

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs000469.v13.p6 TARGET: Cancer Model Systems (MDLS): Cell Lines and Xenografts (including PPTP)	Versions 1-13: passed embargo	V D A S	133	Cohort	Links	
phs000467.v13.p6 TARGET: Neuroblastoma (NBL)	Versions 1-13: passed embargo	V D A S	1183	Cohort	Links	
phs000218.v16.p6 NCI TARGET: Therapeutically Applicable Research to Generate Effective Treatments	Versions 1-16: passed embargo	V D A S	4682	Cohort	Links	

Figura 11. Resultados en una primera búsqueda en dbGaP.

Se obtienen un total de 9 estudios en los que, de alguna manera u otra, se ha investigado sobre el Neuroblastoma. Aunque se haga uso de la búsqueda avanzada como en ClinVar introduciendo en el buscador "neuroblastoma [Disease]", se siguen obteniendo como resultados, estudios en los que se investiga la enfermedad del Neuroblastoma. En este caso, se reduce el número de estudios a 6 en los que se utiliza la palabra Neuroblastoma como para referirse a la enfermedad.

Afortunadamente, dbGaP posee una herramienta llamada "***Phenotype-Genotype Integrator***", que integra información de otras bases de datos como OMIM, dbSNP o Gene[22] (https://www.ncbi.nlm.nih.gov/gene), y que sí que muestra las variaciones relacionadas con enfermedades, cada una de ellas encontrada en un estudio. Esta herramienta permite buscar basándose en: localización cromosómica, gen, SNP o fenotipo.

Para la búsqueda de las variaciones, se introduce "*Neuroblastoma*" como término de búsqueda por fenotipo en dicha herramienta.

Se obtienen un total de **637 resultados**. Para el segundo paso de la metodología SILE habrá que tener en cuenta que en estos 637 resultados habrá variaciones repetidas ya que alguna variación puede haberse encontrado en estudios diferentes.

En la descarga automática que permite dbGaP, también en formato de texto tabular se obtienen los siguientes campos para cada variación mostrados en la Tabla 9.

[22] Repositorio que incluye gran cantidad de información sobre muchas especies, por ejemplo, secuencias de referencia, nomenclatura, variaciones, rutas biológicas, etc.

	DESCRIPCIÓN	EJEMPLO
Trait	Condición a la que se asocia la variación	Neuroblastoma
SNP rs	Identificador rs único	6435862
ContextGene	Contexto en el que se encuentra la asociación (intrónica o intergénica)	intron
Gene	Nombre del gen	BARD1
Gene ID	Identificador asignado a cada gen por Gene (BD también perteneciente a NCBI)	580
Gene 2	Sinónimo del gen (si no tiene sinónimos se escribe el mismo nombre)	BARD1
Gene 2 ID	Identificador asignado al gen 2 por Gene (BD también perteneciente a NCBI)	580
Chromosoma	Cromosoma	2
Location	Posición dentro del cromosoma	215672546
P-Value	Nivel de evidencia de la relación genotipo-fenotipo asignada por el estudio	9,00E-18
PubMed	Identificador de la referencia bibliográfica en la base de datos PubMed	19412175
Analysis ID	Identificador asignado al análisis por dbGaP	2895
Study ID	Identificador asignado al estudio por dbGaP	124
Study Name	Nombre del estudio	*Genome-Wide Association Study of Neuroblastoma*

Tabla 9. Campos obtenidos en la descarga automática de dbGaP.

Además de esta información, los resultados incluyen otros apartados como una tabla con un resumen de los SNPs encontrados y de los estudios (o un visor dinámico de la secuencia genómica), aunque estos apartados resultan de menor relevancia para el trabajo.

4.1.3 Resultados tras el paso 1: "*Search" (Búsqueda)*

Tras una primera búsqueda, se han obtenido **359 variaciones de Clinvar** y **637 de dbGaP**, lo que suma un total de **996**, adjuntadas en el Anexo 1. De estas variaciones habrá que tener en cuenta que muchas podrían estar repetidas y otras podrían no ser lo suficientemente relevantes.

Se decide obtener un conjunto de datos para cada variación, tanto las procedentes de ClinVar como las de dbGaP de manera que se puedan caracterizar de manera unívoca cada una de ellas. Los campos finalmente recolectados para cada variación son los que se muestran en la Tabla 10.

Los identificadores numéricos y símbolos de cada gen han sido extraídos de **HUGO *"Gene Nomenclature Committee* (HGNC)"**[23]. Este organismo es el responsable de aprobar símbolos y nombres únicos para permitir una comunicación científica no ambigua. Por otro lado, el identificador HG se refiere a la versión del Genoma Humano, en este caso para todas las variaciones será hg38 ya que este es el identificador de la última versión y la que se ha usado.

Refseq contiene una colección de secuencias las cuales incluyen: ADN genómico, transcritos y proteínas. A partir de RefSeq se obtienen las secuencias de referencia genómicas ("NG_IDENTIFIER") y del transcrito ("NM_IDENTIFIER"). Dado que existen diversos tipos de secuencia de referencia, una misma variación puede recibir distintos nombres en función de la secuencia de referencia utilizada. Por ello, los distintos nombres en notación HGVS que cada variación puede adquirir se recogen en "HGVS_NM", "HGVS_NG" y "HGVS_NC" según la secuencia a la que la posición se refiere sea un transcrito, genómica o cromosómica**.** La estructura es la siguiente:

<ID_SECUENCIA>:<TIPO_SECUENCIA>.<POSICIÓN><CAMBIO>

Un ejemplo de esta nomenclatura sería: NM_015074.3(KIF1B):c.-260C>T, donde también se incluye entre paréntesis el gen.

Cabe remarcar que los símbolos de las bases al describir el cambio son letras mayúsculas en secuencias de ADN y minúsculas en RNA. Los aminoácidos en secuencias de proteínas pueden ser designados mediante el código de una o tres letras.

De **dbSNP** se obtiene el alelo de referencia y el alternativo, así como las 25 bases antes y después de la variación. También se recoge la dirección en la que se ha encontrado la variación que puede ser *Forward* o *Backward.*

El tipo de variación, el fenotipo y el número de identificación de la referencia en PubMed se obtiene de ClinVar o dbSNP según la procedencia de la variación. La significancia clínica y el

[23] http://www.genenames.org/

estado de revisión se obtienen solamente de las variaciones procedentes de ClinVar, ya que dbGaP no dispone de esta información (*actualmente*).

El título y autores de la referencia se obtienen de PubMed haciendo uso del identificador ID de Pubmed (**PMID**). Por último, como ID de la variación que la identifica de manera unívoca se ha elegido el identificador "rs" de dbSNP.

	DESCRIPCIÓN	**PROCEDENCIA**	**EJEMPLO**
ID_SYMBOL	Identificador del gen	HGNC	16636
ID_HUGO	Identificador oficial del gen		KIF1B
OFFICIAL_NAME	Nombre oficial del gen		Kinesin family member 1B
GENE_SYNONIM	Símbolos alternativos del gen		CMT2, CMT2A, CMT2A1, HMSNII, KLP, NBLST1
CHROMOSOME	Número del cromosoma	ClinVar o dbGaP	1
HG_IDENTIFIER	ID que representa la versión del genoma	Proyecto Genoma Humano	Hg38
NG_IDENTIFIER	Identificador de la secuencia de referencia genómica y versión	RefSeq	NG_008069.1
NM_IDENTIFIER	Identificador de la secuencia de referencia del transcrito y versión		NM_015074.3
HGVS_NM	Nombre HGVS referido al transcrito	ClinVar o dbSNP	NM_015074.3(KIF1B):c.-260C>T
HGVS_NC	Nombre HGVS referido al cromosoma		NC_000001.11:g.10210698C>T
HGVS_NG	Nombre HGVS referido al gen		NG_008069.1:g.4993C>T
GRCh38_POSITION	Posición de la variación mapeada en el genoma de referencia GRCh38		10210698
REF_ALLELE	Alelo de referencia		C
ALT_ALLELE	Alelo alternativo		T
FLAKING_LEFT	Fragmento de DNA de las 25 bases antes de la variación		CTCGCGCACTCCTGTCCGCCGCCCA
FLAKING_RIGHT	Fragmento de DNA de		CGCCCACCTCCCACCTCGATGCGGT

	las 25 bases después de la variación	dbSNP	
STRAND	Sentido en el que la variación ha sido encontrada		F
TYPE	Tipo de variación	ClinVar o dbGaP	Single nucleotide variant
PHENOTYPE	Enfermedad/es a la/s que se asocia		Neuroblastoma|Pheochromocytoma |Charcot-Marie-Tooth, Type 2
SIGNIFICANCE	Efecto clínico de la variación	ClinVar	Likely benign(Last reviewed: Jun 14, 2016)
REVIEW_STATUS	Nivel de evidencia que apoya la asociación		criteria provided, single submitter
PMID	Identificador de la citación en PubMed	ClinVar o dbGaP	18463370
AUTHORS	Autores de la referencia	PubMed	Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH
TITLE	Título de la referencia		Chromosome 6p22 locus associated with clinically aggressive neuroblastoma.
DB_VARIATION_ID	Identificador de la variación en dbSNP	ClinVar o dbGaP	rs1561277

Tabla 10. Campos finalmente recolectados para la base de datos.

Ambas bases de datos poseen datos curados por biólogos o expertos genetistas, es por ello por lo que todas las variaciones encontradas con menor o mayor grado de evidencia están relacionadas con el Neuroblastoma y todas ellas con una referencia bibliográfica o estudio que lo valida.

4.2 *IDENTIFICATION*

El principal objetivo de este paso es preparar los datos para la carga en la HGDB quitando posibles redundancias u otros problemas de calidad de los datos.

Las 996 variaciones que llegan a este paso ya han superado una primera validación realizada por los clínicos y genetistas que han proporcionado estos datos en ClinVar y dbGaP. A continuación, se describen los pasos que han seguido los datos para que puedan ser cargados en la base de datos final.

De las 359 variaciones obtenidas de ClinVar, 2 de ellas tuvieron que ser descartadas ya que se presentaban en ClinVar como grandes anomalías que afectaban a muchos genes por lo que no

tenían una posición exacta y no podían ser cargadas en la base de datos como tal. De los 357 restantes, se vio que 9 posiciones estaban repetidas (2 entradas por cada posición) y otra por 3 entradas. Se revisó cada una de ellas para ver si podían ser representadas por una única entrada aquellas que tenían la misma localización. Examinando la estructura de la base de datos del genoma humano, solo se pueden contraer aquellas entradas en una sola si la única diferencia entre ellas es el alelo alternativo. De estas 10 posiciones duplicadas o triplicadas, solo dos parejas cumplían dicho requisito. Cada una de estas dos parejas fueron anotadas como una sola entrada y se guardaron los dos alelos alternativos. Finalmente, se obtuvieron un total de 355 de ClinVar.

Se contempló la posibilidad de filtrar según la significancia clínica ya que los posibles valores que se le pueden asignar a cada variación son los que se muestran a continuación (Richards et al., 2015):

1. *Benign*: Variación benigna.
2. *Likely benign:* Variación con una alta probabilidad de ser benigna (mayor del 90%).
3. *Uncertain significance:* Cuando una variación no cumple los criterios para clasificarse como benigno o patógeno, o cuando la evidencia de benigno y patógeno es conflictiva.
4. *Pathogenic:* Variación patógena, causante de la enfermedad.
5. *Likely pathogenic:* Variación con alta probabilidad de causar la enfermedad (mayor del 90%).
6. *Risk factor*: Para las variaciones que se interpretan no como para causar un trastorno sino para aumentar el riesgo de padecerlo.
7. *Conflicting interpretations*: varios "*submitters"* asignan a la variación una significancia clínica diferente.

Podría haberse filtrado solo aquellas variaciones con nivel de significancia *clínica: "likely pathogenic", "pathogenic"* y "*risk factor"*. Sin embargo, la herramienta que se usará para la explotación de la base de datos, "*VarSearch*", como se ha mencionado en capítulos anteriores, compara las variaciones de un fichero VCF o FASTA con las que contiene la base de datos, y las clasifica en "*variaciones encontradas"* y "*variaciones no encontradas"*. En el caso que un paciente tenga una variación que se conoce con una significancia clínica benigna, si no se hubiesen cargado las variaciones con esta condición, la variación en cuestión se clasificaría como "*no encontrada"*. En este caso, se precisaría de un análisis posterior para determinar su efecto, análisis innecesario ya que este ya se conocía. Por esta razón, se decide cargar todas las variaciones sea cual sea su significancia clínica.

De las 637 variaciones en dbGaP se hizo una amplia revisión ya que, como se ha mencionado antes, podían aparecer muchas repetidas ya que podían ser encontradas en estudios diferentes. La mayoría de ellas aparecían dos veces con la única diferencia el nombre del estudio. Se eliminaron aquellas repetidas y en el título del estudio se escribieron los dos estudios que la describían. Después de este filtro, quedaron un total de 321 variaciones.

dbGaP almacena las asociaciones genotipo-fenotipo derivadas de estudios GWAS[24]. La estrecha relación entre una variación genética y la enfermedad son representadas mediante el *p-valor*[25]. Este parámetro estadístico es inversamente proporcional al nivel de asociación. Cuanto menor es el p-valor mayor es la evidencia que apoya la asociación variación-enfermedad. Haciendo uso de esta información se decide realizar un segundo filtro con el fin de obtener aquellas de las que se tiene mayor evidencia. Se dice que aquellas que tengan un p-valor inferior a 10^{-8} se consideran reales. Por el contrario, aquellas con p-valores más modestos (por ejemplo, 10^{-5}-10^{-6}) representarán falsos positivos; un p-valor de 10^{-5} equivale a una verdadera asociación menor del 1% (Galvan, Ioannidis and Dragani, 2010). Por este motivo se deciden rechazar todas aquellas variaciones con un p-valor mayor de 10^{-8}, superando el filtro 29 variaciones.

Estas 29 variaciones se volvieron a revisar por evitar duplicaciones y se vio que una de ellas había sido incluida en dos estudios que le habían asociado un *p-valor* distinto. Como el *p-valor* es un campo que no iba a ser cargado en la base de datos se colapsaron estas dos entradas en una sola.

Finalmente, se juntaron estas 28 variaciones con las 355 procedentes de ClinVar y se vieron posibles duplicaciones. Ninguna variación había sido encontrada en ambas bases de datos. Por tanto, el total de variaciones -filtradas-para el Neuroblastoma se finalizó con 383, adjuntadas en el Anexo 2.

4.3 LOAD

El proceso de carga se basa en la introducción de la información identificada en el proceso anterior ("*Identification*") a la base de datos del genoma humano (HGDB). Dicho proceso se subdivide en 3 subprocesos como se menciona en la Tesis de J. F. Reyes (Reyes, 2013):

- *Extracción*: Consiste en la obtención de la información de las variaciones de un repositorio externo en un formato que se pueda almacenar en un servidor, para luego poder manejarlo localmente.
- *Transformación*: Tras la extracción selectiva, se debe tratar la información obtenida para adecuarla a la base de datos del genoma humano.
- *Carga*: Finalmente se realiza el proceso completo de carga, en el que se introduce la información ya transformada desde un fichero local a la base de datos.

[24] Los estudios de asociación en todo el genoma buscan en el genoma pequeñas variaciones, llamadas polimorfismos de un solo nucleótido o SNPs, que ocurren con más frecuencia en personas con una enfermedad particular que en personas sin la enfermedad.

[25] Probabilidad que mide la evidencia contra la hipótesis nula. En este caso, la probabilidad de que una determinada variación sea causante de un determinado fenotipo tomando como hipótesis nula que la variación no tiene ningún efecto sobre el fenotipo.

Como se ha mencionado antes, la carga que se va a realizar es "*selectiva*" ya que la base de datos posee sólo datos estudiados y validados para generar un repositorio genómico con datos de calidad.

De todo el esquema conceptual de la base de datos del genoma humano, "*VarSearch"* actualmente implementa la vista de "*Variaciones*" (Reyes Román, J. F. and Pastor López, Ó., 2016). Es por esto que el trabajo se centra solo en los atributos y tablas que pertenecen a esta parte del esquema (Figura 12).

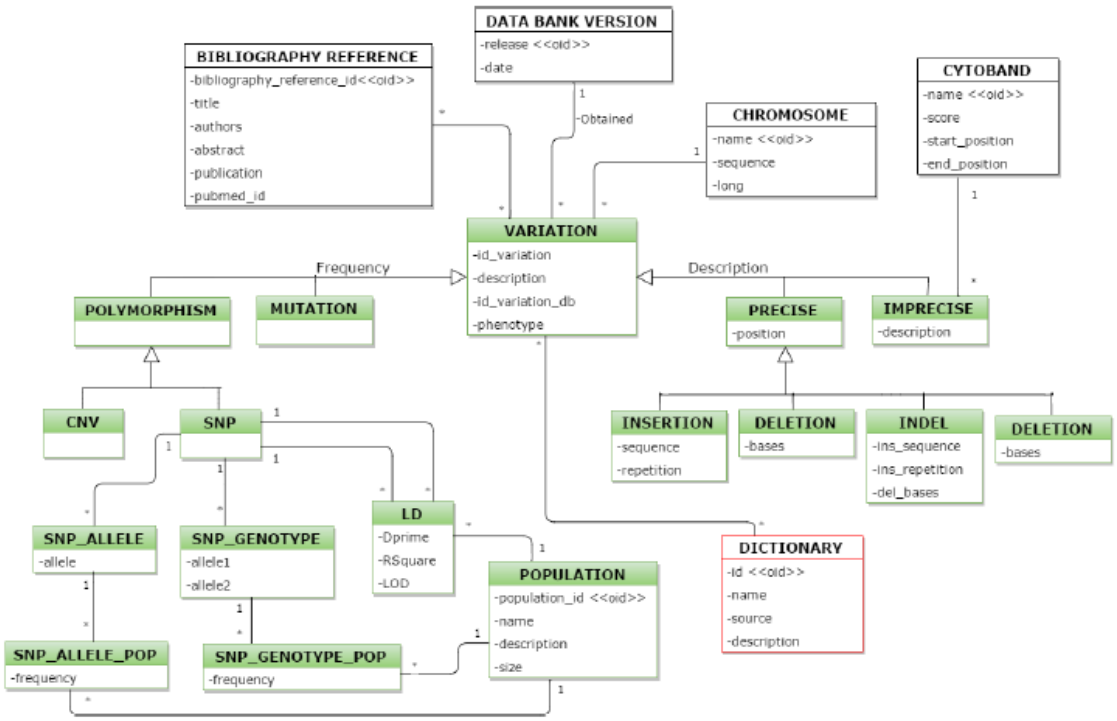


Figura 12. Vista de variaciones, versión v2 del CSHG (Pastor López, O et al. 2016).

Primero, se realizó un mapeado de los campos recogidos con los que tenía la HGDB como se muestran en la Figura 13.

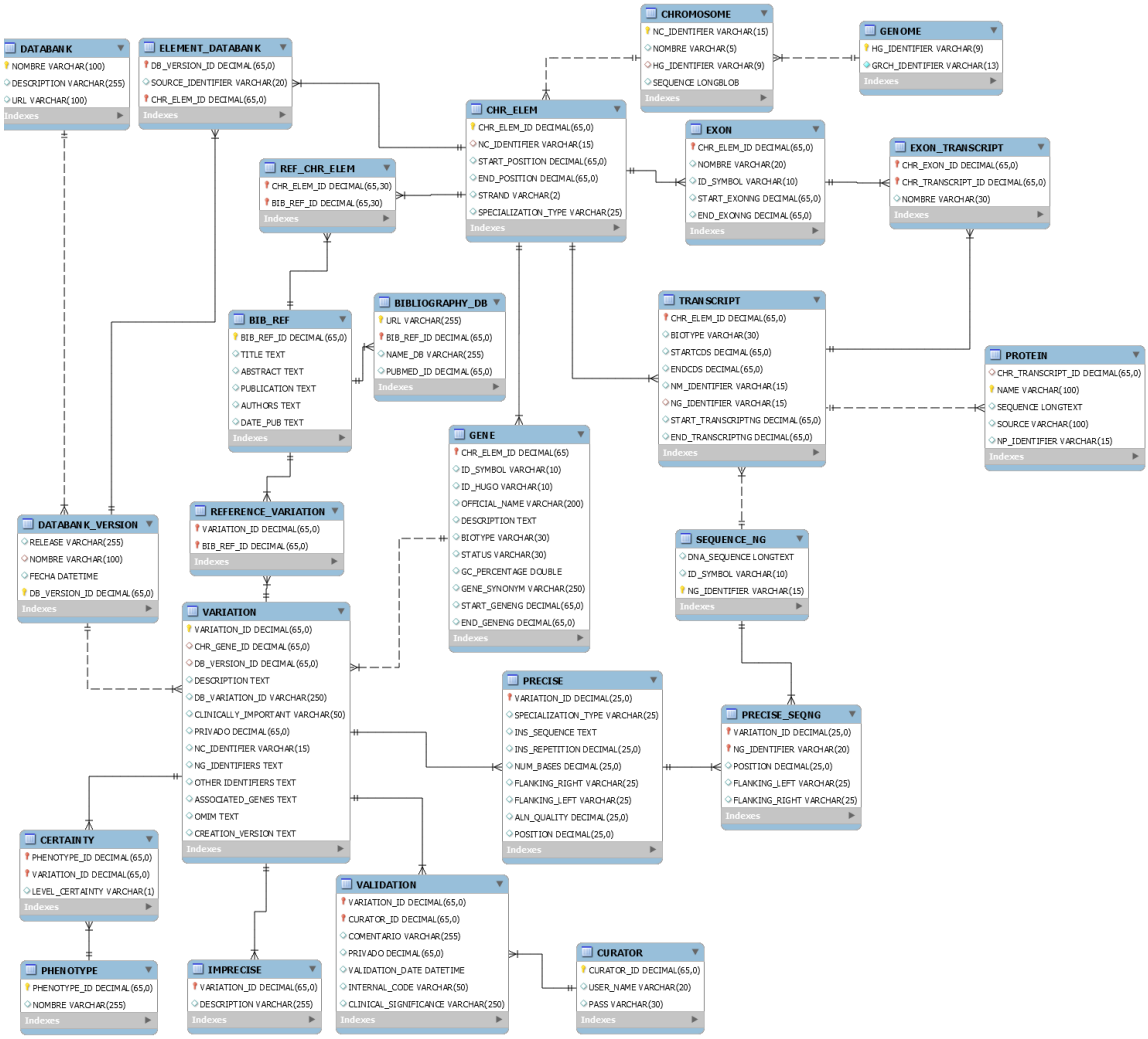


Figura 13. Esquema de la base de datos Human Genome Database (HGDB).

Se creó una plantilla *.csv (utilizando la herramienta de "Excel"), en la que se mapearon los datos recogidos según a qué tabla pertenecían, cada tabla en una pestaña del Excel. Se rellenaron las columnas con los datos de los que se disponía, otros quedaron "NULL" temporalmente por la inexistencia o falta de disponibilidad de dicha información. Además, se hizo un resumen de los atributos que presenta HGDB para ver cuáles eran obligatorios y cuáles no, para un correcto funcionamiento de la herramienta "*VarSearch*".

Finalmente, de todos los atributos que aparecen en la Figura 13, se cargaron los que se muestran en la Tabla 11. En esta tabla, se especifica el tipo de dato permitido para cada atributo, así como un ejemplo para mejor entendimiento.

TABLA	ATRIBUTOS	TIPO DE DATO	EJEMPLO EN LA BD
GENOME	HG_IDENTIFIER	VARCHAR (9)	hg38
	GRCH_IDENTIFIER	VARCHAR (13)	GRCh38.p10
CHROMOSOME	NC_IDENTIFIER	VARCHAR (15)	NC_000001.11
	NOMBRE	VARCHAR (5)	chr1
	HG_IDENTIFIER	VARCHAR (9,0)	hg38
CHR_ELEM	CHROMOSOME_ELEMENT_ID	DECIMAL (65,0)	3
	NC_IDENTIFIER	VARCHAR (15)	NC_000003.12
	STRAND	VARCHAR (2)	+
	SPECIALIZATION_TYPE	VARCHAR (25)	transcriptable element
GENE	CHR_ELEM_ID	DECIMAL (65,0)	1
	ID_SYMBOL	VARCHAR	KIF1B
	ID_HUGO	VARCHAR	16636
	OFFICIAL_NAME	VARCHAR	kinesin family member 1B
	GENE_SYNONYM	VARCHAR	CMT2,CMT2A,CMT2A1,HMSNII,KLP,NBLST1
SEQUENCE_NG	ID_SYMBOL	VARCHAR (10)	KIF1B
	NG_IDENTIFIER	VARCHAR (15)	NG_008069.1
TRANSCRIPT	CHR_ELEM_ID	DECIMAL (65,0)	15
	NM_IDENTIFIER	VARCHAR (15)	NM_015074.3
	NG-IDENTIFIER	VARCHAR (15)	NG_008069.1
DATABANK	NOMBRE	VARCHAR (100)	ClinVar
	URL	VARCHAR (100)	https://www.ncbi.nlm.nih.gov/clinvar/
DATABANK_VERSION	RELEASE	VARCHAR (255)	ClinVarFullRelease_2017-01
	NOMBRE	VARCHAR (100)	ClinVar
	DB_VERSION_ID	DECIMAL (65,0)	1
ELEMENT_DATABANK	DB_VERSION_ID	DECIMAL (65,0)	1
	CHR_ELEM_ID	DECIMAL (65,0)	1
VARIATION	VARIATION_ID	DECIMAL (65)	1
	CHR_GENE_ID	DECIMAL (65)	2
	DB_VERSION_ID	DECIMAL (65)	1
	DB_VARIATION_ID	VARCHAR (250)	78174819
	CLINICALLY_IMPORTANT	VARCHAR (50)	Benign
	NC_IDENTIFIER	VARCHAR (15)	NC_000002.12
	NG_IDENTIFIERS	TEXT	NG_009445.1
	OTHER_IDENTIFIERS	TEXT	NM_004304.4(ALK):c.4836G>A,NC_000002.12:g.29193251C>T,NG_009445.1:g.733316G>A

PRECISE	VARIATION_ID	DECIMAL (25,0)	1
	SPECIALIZATION_TYPE	VARCHAR (25)	Indel
	INS_SEQUENCE	TEXT	T
	INS_REPETITION	DECIMAL (25,0)	2
	NUM_BASES	DECIMAL (25,0)	1
	FLANKING_RIGHT	VARCHAR (25)	TTGCTTTTCAGAATGGTATCCTCGT
	FLANKING_LEFT	VARCHAR (25)	AGGGCCCAGGCTGGTTCATGCTATT
	POSITION	DECIMAL (25,0)	29193251
BIB_REF	BIB_REF_ID	DECIMAL (65,0)	2
	TITLE	TEXT	The kinesin KIF1Bbeta acts downstream from EglN3 to induce apoptosis and is a potential 1p36 tumor suppressor.
	AUTHORS	TEXT	Schlisio S, Kenchappa RS, Vredeveld LC, George RE, Stewart R, Greulich H, Shahriari K, Nguyen NV, Pigny P, Dahia PL, Pomeroy SL, Maris JM, Look AT, Meyerson M, Peeper DS, Carter BD, Kaelin WG Jr.
BIBLIOGRAPHY_DB	URL	VARCHAR (255)	https://www.ncbi.nlm.nih.gov/pubmed/18334619
	BIB_REF_ID	DECIMAL (65,0)	2
	PUBMED_ID	DECIMAL (65,0)	18334619
REFERENCE_VARIATION	VARIATION_ID	DECIMAL (65,0)	1
	BIB_REF_ID	DECIMAL (65,0)	8
PHENOTYPE	PHENOTYPE_ID	DECIMAL (65,0)	1
	NOMBRE	VARCHAR (255)	Neuroblastoma
CERTAINTY	PHENOTYPE_ID	DECIMAL (65,0)	
	VARIATION_ID	DECIMAL (65,0)	
	LEVEL_CERTAINTY	VARCHAR (1)	

Tabla 11. Atributos de cada variación cargados en la base de datos.

La mayoría de todos estos atributos ya habían sido recogidos, pero algunos otros no se habían tenido en cuenta. Por ejemplo, se añadieron las URLs en la tabla de la bibliografía, o también se buscaron las versiones de las bases de datos utilizadas.

El mapeado más costoso fue el de los elementos de la tabla "**PRECISE**". ClinVar y dbGaP utilizan nombres para los tipos de variaciones diferentes a los que son aceptados por *VarSearch*.

VarSearch se basa en el Esquema Conceptual del Genoma Humano (ECGH) (Reyes, J.F. et al., 2016), el cual solo contempla 4 tipos de variaciones posibles: "*insertion", "deletion", "indel"* e "*inversion"*, como se muestra en la Figura 12. El significado de estos 4 tipos de variaciones se ha explicado en la Tabla 8 en el paso *Search*.

Por tanto, se realizó una correspondencia, conforme la Tabla 12, entre los tipos de variaciones representados en ClinVar y dbGaP con los tipos de variaciones que tiene en cuenta la base de datos de VarSearch.

TÉRMINOS CLINVAR O DBGAP	TÉRMINOS ACEPTADOS
Insertion	Insertion
Deletion	Deletion
Indel	Indel
Inversion	Inversion
Single Nucleotide Variant	Indel
Copy Number Gain	Insertion
Copy Number Loss	Deletion
Duplication	Insertion
Tandem Duplication	Insertion
Microsatellite	Insertion

Tabla 12. Correspondencias entre ClinVar y dbGaP y Varsearch.

Además, para cada tipo de variación se requiere una información específica que ha tenido que ser añadida. En la siguiente tabla 13 se muestran con una cruz los campos básicos necesarios según el tipo de variación representada en VarSearch:

- *Position*: Posición donde empieza la variación, en coordenadas cromosómicas.
- *Ins_repetition*: Número de veces que se repite la secuencia insertada.
- *Ins_sequence*: Secuencia de nucleótidos que se inserta.
- *Num_bases*: Número de bases borradas.

TIPO	POSITION	INS_REPETITION	INS_SEQUENCE	NUM_BASES
Insertion	X	X	X	
Deletion	X			X
Indel	X		X	X
Inversion	X			X

Tabla 13. Campos básicos necesarios según el tipo de variación.

Una vez se dispuso de todos los atributos separados según tablas, se creó un fichero *.csv por cada tabla que incluía cada uno de los atributos en columnas y en el mismo orden que se representa en la HGDB. La carga de datos se ha realizado sobre una base de datos vacía, de modo que todas las variaciones que posee están relacionadas con el Neuroblastoma.

Para la carga de estos archivos se decidió usar el gestor de bases de datos ***HeidiSQL***[26]. Se trata de una herramienta útil de uso gratuito que permite navegar y editar datos, crear y editar tablas de manera sencilla.

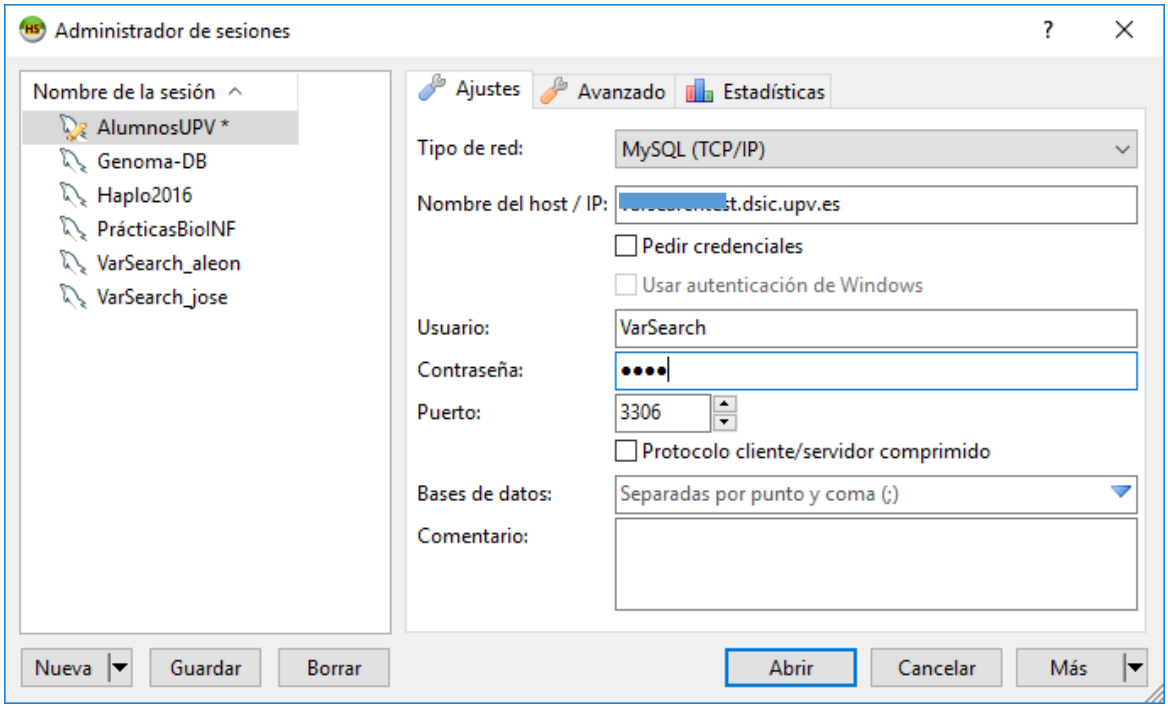


Figura 14. Configuración en Heidi para el acceso a la HGDB.

Al entrar en la aplicación, se puede navegar por cada una de las tablas y consultar cada atributo. Un ejemplo sería como se muestra en la Figura 15, donde se muestran a la izquierda las diferentes tablas y si por ejemplo se entra en "BIBLIOGRAPHY_DB" se pueden ver detalles de cada uno de sus atributos como por ejemplo el tipo de datos permitido.

[26] https://www.heidisql.com/

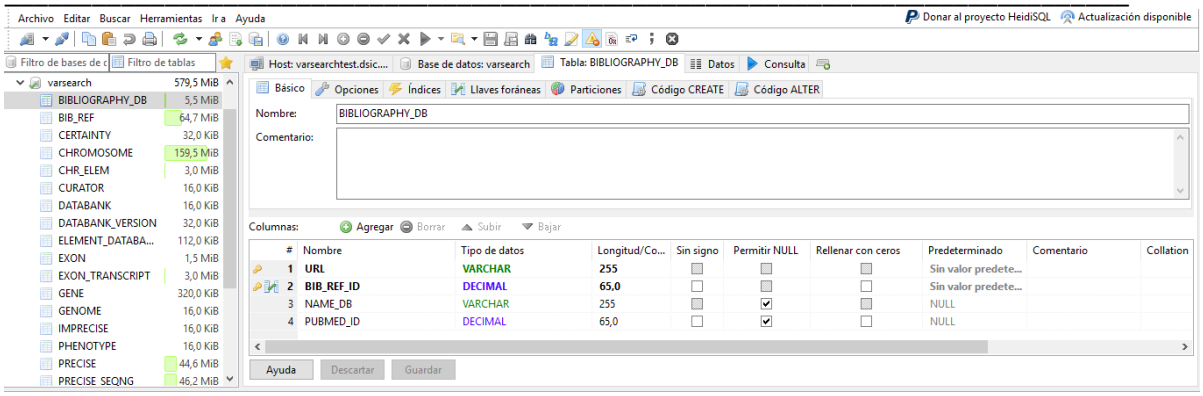


Figura 15. Visualización desde HeidiSQL.

HeidiSQL posee un apartado dentro de "Herramientas" que permite importar archivos con formato *.csv. Un ejemplo sería el que se muestra en la Figura 16, donde se está importando el fichero*.csv correspondiente a la tabla "BIBLIOGRAPHY_DB". Se elige el tipo de codificación (*en este caso el idioma será latin 1*), se ignora la primera fila ya que corresponde a la descripción de los datos de cada columna, se elige la base de datos destino y se especifica a qué tabla corresponde en ésta. Los ficheros *.csv se crearon con la misma nomenclatura que posee la base de datos para evitar y/o prevenir conflictos. De esta misma manera es como se ha ido realizando la carga del resto de tablas.

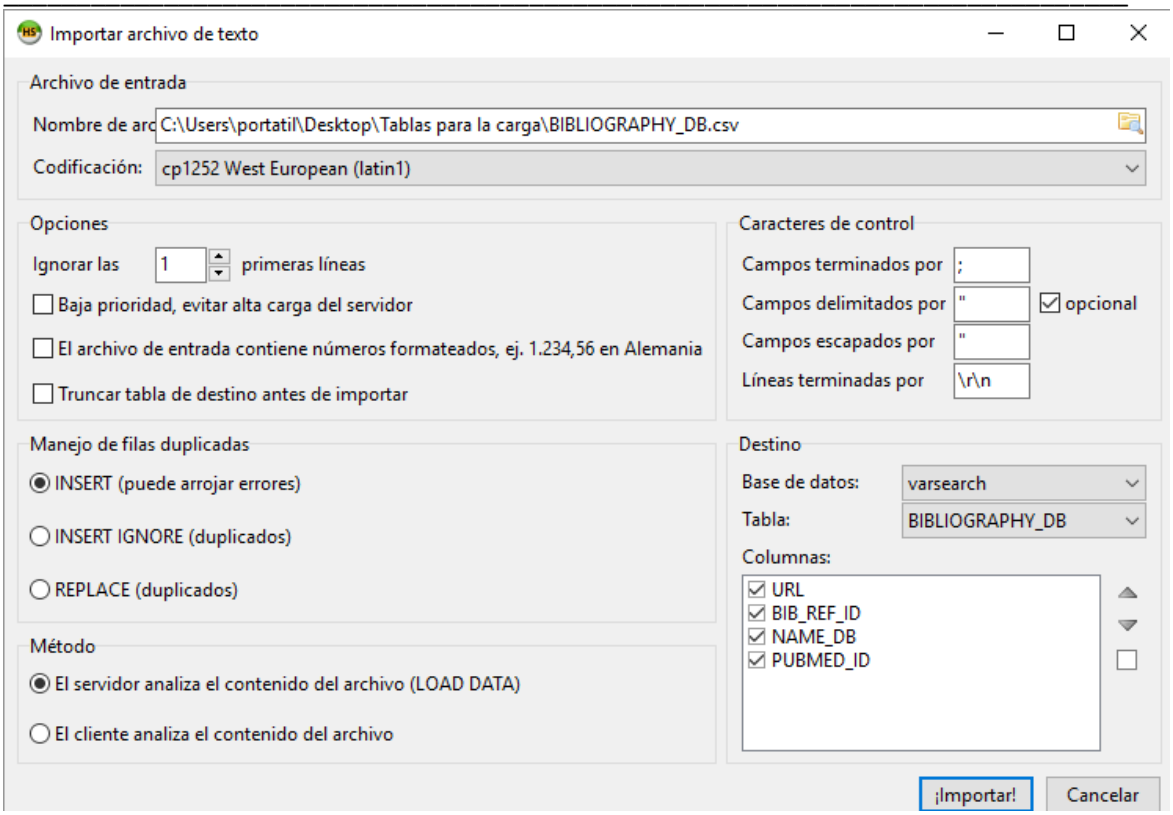


Figura 16. Vista de la importación de archivos CSV en HeidiSQL.

Una vez realizada la carga se han seguido unos pasos de revisión con el fin de garantizar la integridad y calidad de los datos. Se ha revisado que en cada tabla se han cargado correctamente todos los atributos y que no ha habido inconsistencia o incongruencia en los datos.

4.4 *EXPLOTAITION*

El principal objetivo del último paso de la metodología SILE es la creación de conocimiento a partir del correcto manejo de los datos. En el caso de este trabajo, este último paso consiste en comparar las muestras de pacientes con las variaciones genéticas validadas.

A partir de las tecnologías de nueva generación (NGS) (Rodríguez-Santiago and Armengol, 2012) se pueden determinar las variaciones genéticas extraídas de una muestra de tumor. Estos resultados se presentan en archivos *.VCF, los cuales pueden ser analizados por *Varsearch*, con el objetivo de contrastar cuáles de las variaciones están también almacenadas en la base de datos de dicha herramienta.

El funcionamiento de "VarSearch" es el que se ve en la Figura 17. Se introduce un fichero VCF o Sanger, la herramienta lo analiza y proporciona como resultado las variaciones que ha

encontrado en el fichero que se encuentran también en la base de datos. Además, permite al usuario añadir una validación a cualquier variación encontrada.

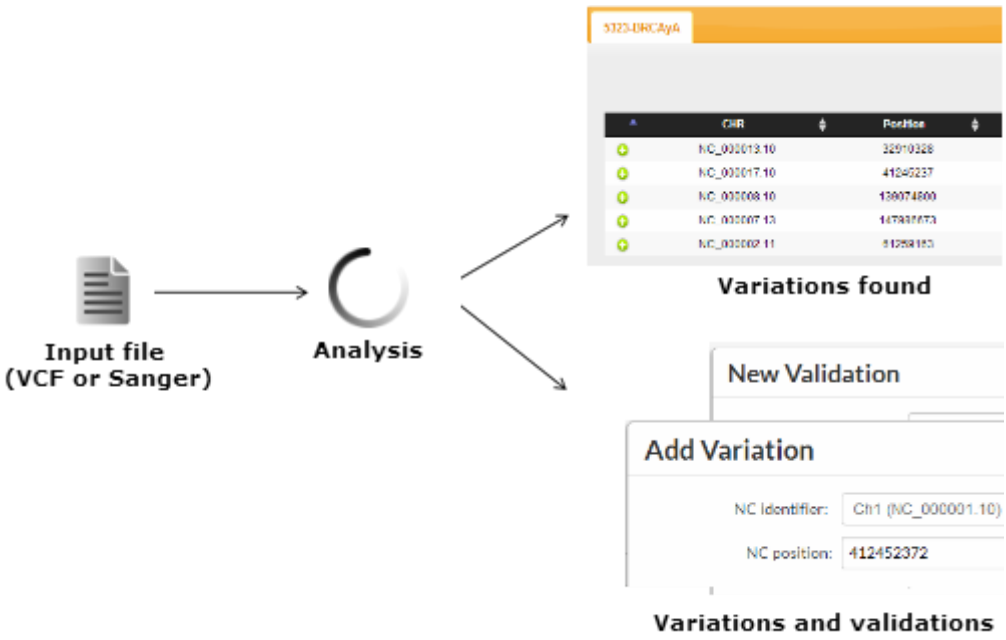


Figura 17. Funcionamiento de "VarSearch".

Se realizaron diversas pruebas con ficheros VCF para validar el correcto funcionamiento de *VarSearch* con las variaciones cargadas. Para ello, se generó un nuevo fichero *.VFC en el que se insertaron variaciones de otra enfermedad más las variaciones relacionadas con el Neuroblastoma. Dicho fichero fue importado a *VarSearch* como se muestra en la Figura 18.

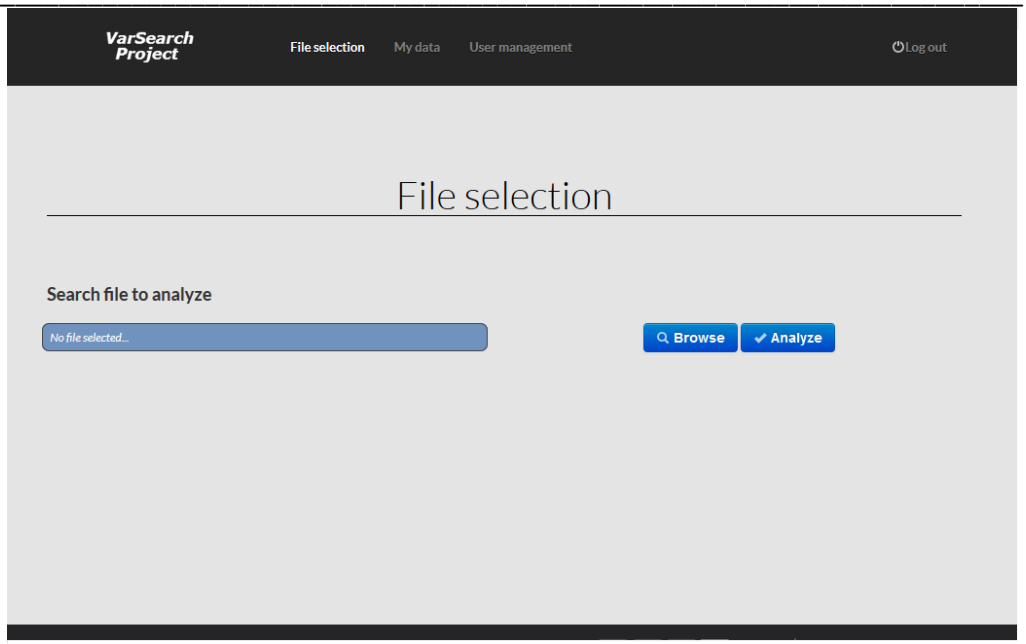


Figura 18. Elección de fichero en VarSearch.

Una vez elegido el fichero, al pulsar en analizar, empieza un proceso el que la herramienta busca las coincidencias de las variaciones contenidas en el fichero con las de la base de datos, como se muestra en la Figura 19.

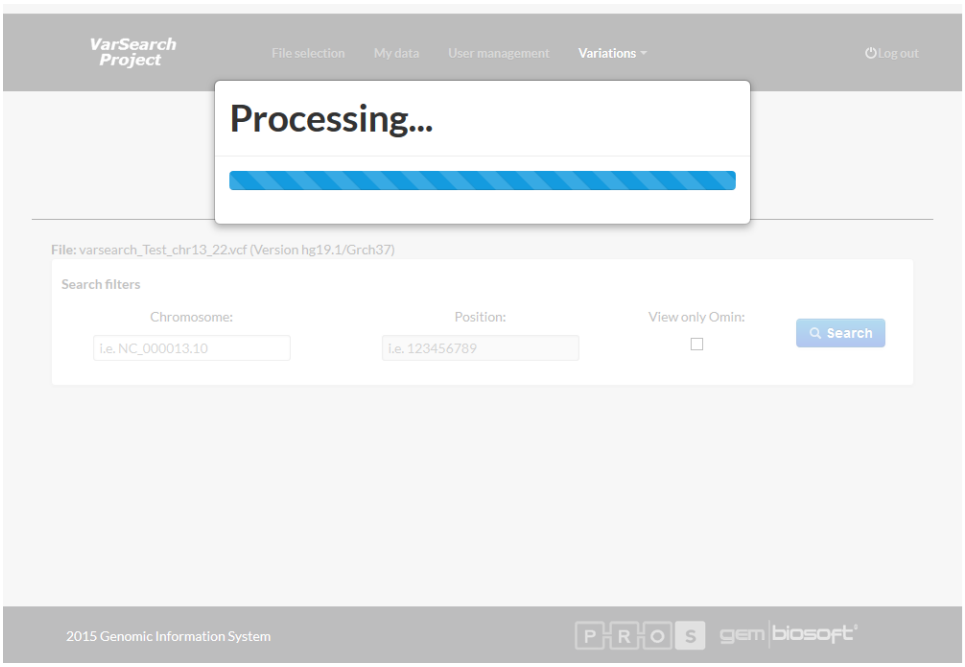


Figura 19. Procesando el fichero por VarSearch.

Los resultados se distribuyen en "*variaciones encontradas*" y "*variaciones no encontradas*". Como todas las variaciones que posee la base de datos son del Neuroblastoma, todas las encontradas estarán relacionadas con la enfermedad.

Las "*variaciones encontradas*" se presentan en una lista al usuario como se muestra en la Figura 20.

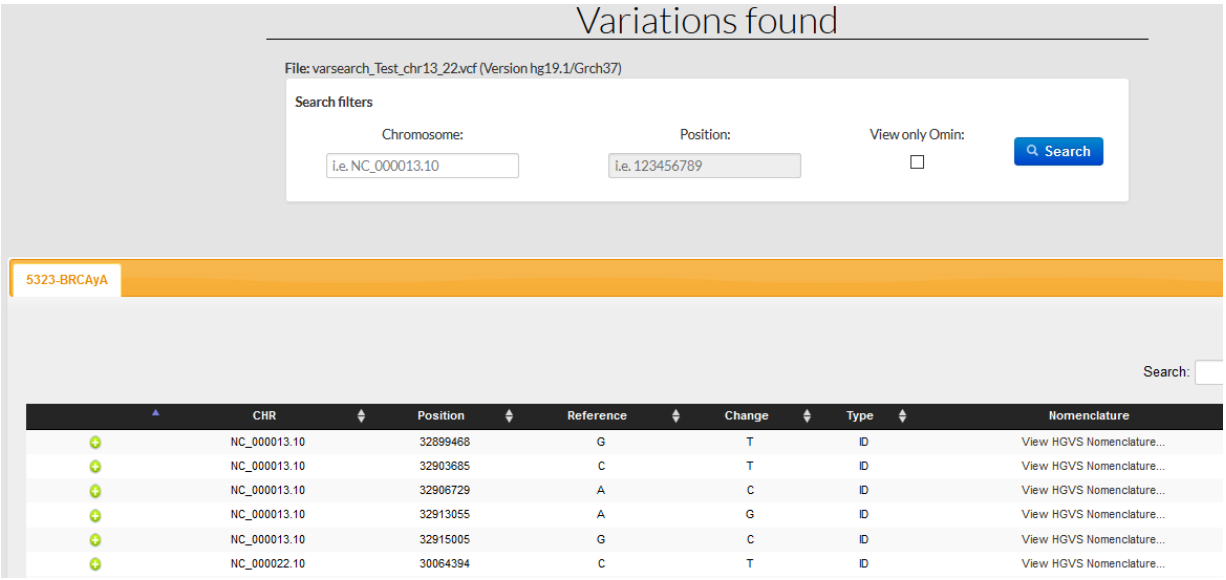


Figura 20. Variaciones encontradas en el fichero.

Además, VarSearch proporciona más información que aparece al hacer "*click*" en el + de la izquierda de cada variación encontrada. En este caso, se muestra información adicional como la significancia clínica, las referencias o validaciones (Figura 21). El usuario puede añadir una validación por cada variación y puede elegir si será pública o privada, en este caso, solo el usuario que poseedor de la variación podrá ver y guardar la validación.

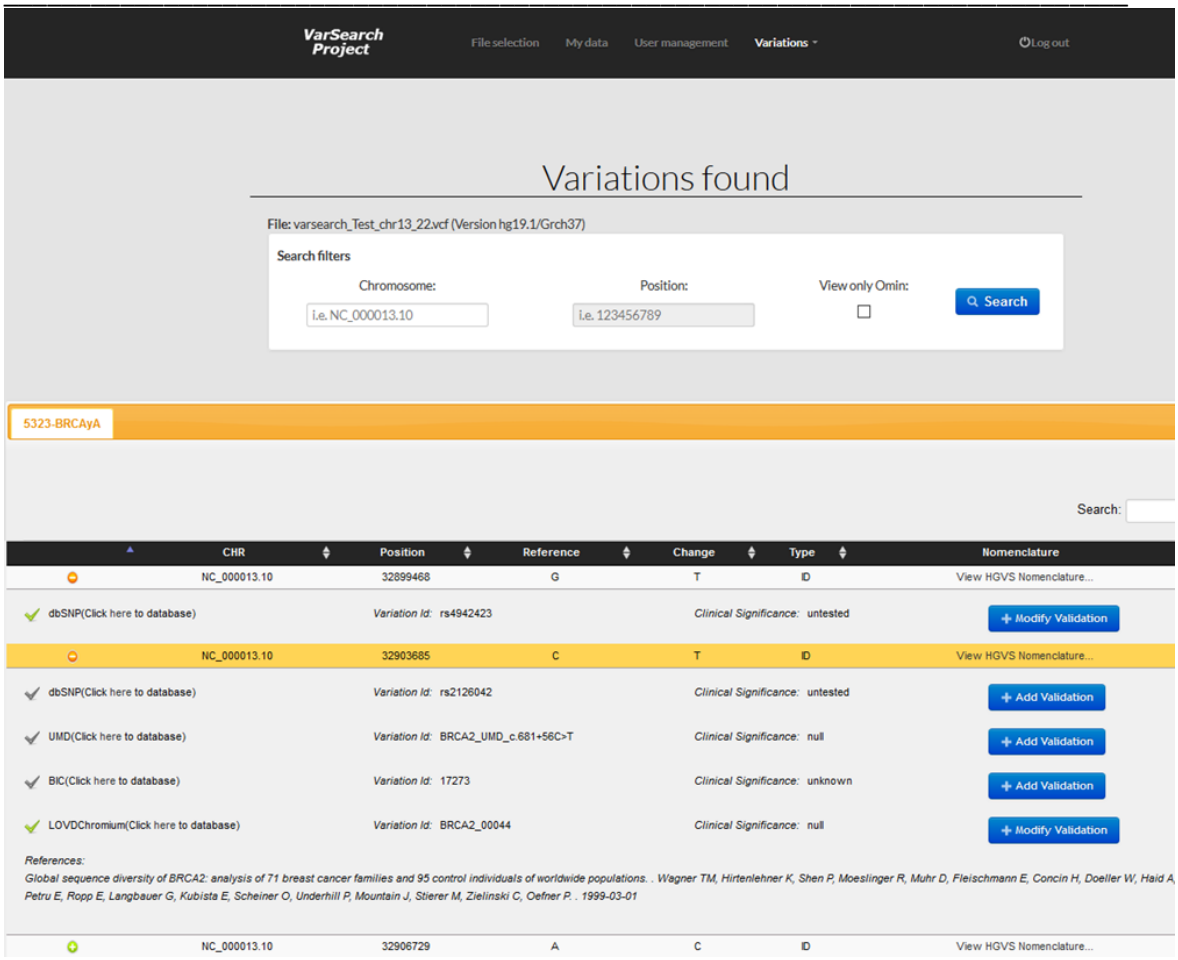


Figura 21. Análisis del fichero usando VarSearch.

Por otro lado, se muestran también las "*variaciones no encontradas*", es decir, aquellas contenidas en el fichero, pero no guardas en la base de datos, por tanto, no relacionadas con la enfermedad (Figura 22).

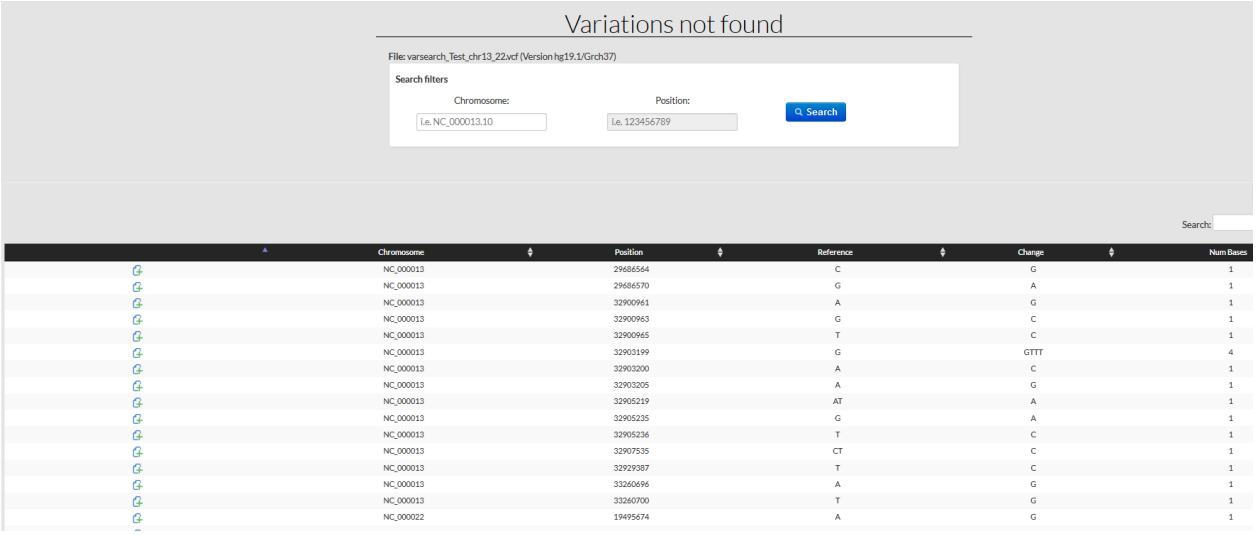


Figura 22. Variaciones no encontradas en el fichero.

Los estudios de estas enfermedades/variaciones tienen como objetivo poder ser facilitados mediante *Test Genéticos Directos al Consumidor (TGDC)* como es el caso del Proyecto en desarrollo *GenesLove.Me* (GLM) (Figura 23), los cuales proporcionan tests para enfermedades concretas como la alopecia androgénica, la intolerancia a la lactosa, la sensibilidad al alcohol o el dupuytren[27] (Reyes Román, J. F., Iñiguez-Jarrín, C. and Pastor López, Ó., 2017).

Figura 23. Pantalla principal de GenesLove.Me.

[27] Acúmulo de tejido fibrótico en la mano que ocasiona una retracción de los tendones flexores y provoca que los dedos quedan en flexión.

CAPÍTULO 5. CONCLUSIONES Y TRABAJO FUTURO

El presente trabajo tenía como finalidad la creación de un *Sistema de Información Genómico* (GeIS) para la enfermedad del Neuroblastoma. Debido a la gran heterogeneidad de los datos disponibles útiles para la creación de un diagnóstico correcto de la enfermedad, se ha decidido seguir una serie de pasos que extraigan conocimiento de aquellos repositorios que mejor se adapten a las necesidades del proyecto, evitando falta de calidad en los datos.

El Modelo Conceptual del Genoma Humano permite una caracterización del genoma humano y se ha utilizado como estructura fuerte sobre la que se ha creado el GeIS. Se ha hecho una amplia revisión de las fuentes de datos disponibles para la enfermedad con el fin de elegir aquellas más óptimas. Como metodología se ha decidido seguir SILE creada por el Grupo de Investigación PROS, ya que ésta proporciona los resultados deseados.

Se ha conseguido obtener una base de datos para el Neuroblastoma que incluye las variaciones validadas y directamente relacionadas con la enfermedad. Sirve para la creación de diagnósticos genómicos ya que mediante *VarSearch* se ha conseguido relacionar muestras de pacientes con las variaciones almacenadas en la base de datos. Esto proporciona a los médicos y genetistas un diagnóstico mucho más preciso ya que se sabe exactamente cuáles son las variaciones que posee el paciente y están involucradas en la aparición de la enfermedad.

Aunque hay mucho trabajo realizado y conocimiento obtenido, esta base de datos tendrá que ser actualizada constantemente ya que cada día hay nuevos descubrimientos y habrá que ir adaptando la base de datos a las nuevas variaciones.

Por otro lado, sólo se han utilizado bases de datos curadas, pero podría añadirse información de otras bases de datos que podrían ser útil para los médicos en su proceso de tomar una decisión. Por tanto, como trabajo futuro se podrían explorar en análisis posteriores otros repositorios con el fin de incluir biomarcadores[28] como miRNA[29], medicamentos dirigidos a variaciones concretas u otra información relevante.

Además, este proceso de obtención de una base de datos homogénea para la enfermedad del Neuroblastoma podría aplicarse como método para el desarrollo de Sistemas de Información

[28] Substancias que indican un estado biológico sean químicos, fisiológicos o morfológicos. Es decir, indicadores que pueden medirse objetivamente, el cual indicaría que un proceso biológico es normal o patológico.

[29] MicroRNA: moléculas pequeñas de RNA con importantes funciones regulatorias a nivel postranscripcional. Recientemente, se ha observado que reúnen muchas características de los buenos biomarcadores (son estables en muchos fluidos corporales y la mayoría de las secuencias se conservan entre especies) (Mar, F. et al., 2015).

Genómicos para cualquier enfermedad de origen genético proporcionando una estructura estable y estandarizada.

El uso eficiente de los datos genómicos ayudará al entendimiento de las bases moleculares de las enfermedades raras. Es necesaria una comprensión profunda de ellos para descubrir biomarcadores que permitan la obtención de un diagnóstico precoz y el desarrollo de medicamentos dirigidos a una población específica. Por este motivo, es necesario una adecuada gestión de la gran variedad de datos.

BIBLIOGRAFÍA

1000 Genomes | A Deep Catalog of Human Genetic Variation. (2017). *Internationalgenome.org*. Retrieved 7 June 2017, from http://www.internationalgenome.org/

Auton, A., Abecasis, G., Altshuler, D., Durbin, R., Abecasis, G., Bentley, D. et al. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74. http://dx.doi.org/10.1038/nature15393

Benjamín Rodríguez-Santiago (2012). Next generation sequencing technology in pre- and postnatal genetic diagnosis. *Quantitative Genomic Medicine Laboratories*, 23(2).

Brisse, H., McCarville, M., Granata, C., Krug, K., Wootton-Gorges, S., & Kanegawa, K. et al. (2011). Guidelines for Imaging and Staging of Neuroblastic Tumors: Consensus Report from the International Neuroblastoma Risk Group Project. *Radiology*, *261*(1), 243-257. http://dx.doi.org/10.1148/radiol.11101352

Burriel, V., Reyes, J. F., Heredia, A., Iñiguez-Jarrín, C. and León, A. (2017). GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma. *IEEE 11th International Conference on Research Challenges in Information Science (RCIS 2017),* 451-452.

Cushing, H. and Wolbach, S. B. (1927). The Transformation of a Malignant Paravertebral Sympathicoblastoma into a Benign Ganglioneuroma. *The American Journal of Pathology*, *3*(3), 203–216.7.

Everson, T., and Cole, W. (1966). *Spontaneous regression of cancer* (1st ed.). Philadelphia: Saunders.

Galvan, A., Ioannidis, J., and Dragani, T. (2010). Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends In Genetics*, *26*(3), 132-141. http://dx.doi.org/10.1016/j.tig.2009.12.008

Historia del Neuroblastoma. (2017). News-Medical.net. Retrieved 25 May 2017, from http://www.news-medical.net/health/Neuroblastoma-History-(Spanish).aspx

Home - dbGaP - NCBI. (2017). *Ncbi.nlm.nih.gov*. Retrieved 13 June 2017, from https://www.ncbi.nlm.nih.gov/gap

Hyun Lee, S., Kim, J., Zheng, S., Huse, J., Seol Bae, J., & Won Lee, J. et al. (2017). ARID1B alterations identify aggressive tumors in neuroblastoma. *Oncotarget*. http://dx.doi.org/10.18632/oncotarget.17500

Landrum, M., Lee, J., Riley, G., Jang, W., Rubinstein, W., Church, D., and Maglott, D. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1), D980-D985. http://dx.doi.org/10.1093/nar/gkt1113

León, A. (2017). Apuntes de la asignatura "Bioinformática". Presentación, Escuela Técnica Superior de Ingenieros Industriales (ETSII).

London, W., Castleberry, R., Matthay, K., Look, A., Seeger, R., & Shimada, H. et al. (2005). Evidence for an Age Cutoff Greater Than 365 Days for Neuroblastoma Risk Group Stratification in the Children's Oncology Group. *Journal Of Clinical Oncology*, *23*(27), 6459-6465. http://dx.doi.org/10.1200/jco.2005.05.571

Mar, F., Morales, M. R., Rodríguez, C. and Reséndez, D. (2015). Detección de microRNA extracelulares y su potencial como biomarcadores moleculares. *Universidad Autónoma de Nuevo León, FCN*, 71.

Mason, G., Hart-Mercer, J., Joan Millar, E., Strang, L. and Wynne, N. (1957). Adrenaline-secreting neuroblastoma in an infant. *The Lancet*, *270*(6990), 322-325. http://dx.doi.org/10.1016/s0140-6736(57)92211-0

Monclair, T., Brodeur, G., Ambros, P., Brisse, H., Cecchetto, G., & Holmes, K. et al. (2009). The International Neuroblastoma Risk Group (INRG) Staging System: An INRG Task Force Report. *Journal Of Clinical Oncology*, *27*(2), 298-303. http://dx.doi.org/10.1200/jco.2008.16.6876

National Library of Medicine - National Institutes of Health. (2017). Nlm.nih.gov. Retrieved 27 June 2017, from https://www.nlm.nih.gov/

Neuroblastoma. (2017). *National Cancer Institute*. Retrieved 25 May 2017, from https://www.cancer.gov/espanol/tipos/neuroblastoma/paciente/tratamiento-neuroblastoma-pdq

Pastor López, O.; Reyes Román, JF.; Valverde Giromé, F. (2016). Conceptual Schema of the Human Genome (CSHG). http://hdl.handle.net/10251/67297.

Pryse-Phillips, W. (2009). *Companion to clinical neurology* (1st ed.). Oxford: Oxford University Press.

Reference, G. (2017). How do geneticits indicate the location of a gene?. *Genetics Home Reference*. Retrieved 7 June, from https://ghr.nlm.nih.gov/primer/howgeneswork/genelocation

Reservados, I. (2017). *Orphanet: Neuroblastoma*. *Orpha.net*. Retrieved 25 May 2017, from http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=ES&Expert=635

Reyes, J. (2013). Integración de Haplotipos al Modelo Conceptual del Genoma Humano utilizando la Metodología SILE (TFM). Universitat Politècnica de València (UPV).

Reyes, J. F., Pastor, O., Casamayor, J.C. and Valverde, F. (2016). Applying Conceptual Modeling to Better Understand the Human Genome. *The 35th International Conference on Conceptual Modeling, Springer International Publishing*, 404-412. http://dx.doi.org/10.1007/978-3-319-46397-1_31

Reyes Román, J., León Palacio, A., and Pastor López, Ó. (2017). Software Engineering and Genomics: The Two Sides of the Same Coin?. Proceedings Of The 12Th International Conference On Evaluation Of Novel Approaches To Software Engineering. http://dx.doi.org/10.5220/0006368203010307

Reyes Román, J. F., Iñiguez-Jarrín, C. and Pastor López, Ó. (2017). GenesLove.Me: A Model-based Web-application for Direct-to-consumer Genetic Tests. *ENASE 2017*, 133-143.

Reyes Román, J. F. and Pastor López, Ó. (2016). Use of GeIS for Early Diagnosis of Alcohol Sensitivity. *Proceedings Of The 9Th International Joint Conference On Biomedical Engineering Systems And Technologies,* 3, 284-289, http://dx.doi.org/10.5220/0005822902840289

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., & Gastier-Foster, J. 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, 17(5), 405-423. http://dx.doi.org/10.1038/gim.2015.30

Rodríguez-Santiago, B., & Armengol, L. (2012). Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal. *Diagnóstico Prenatal*, *23*(2), 56-66. http://dx.doi.org/10.1016/j.diapre.2012.02.001

Rothenberg, A., Berdon, W., D'Angio, G., Yamashiro, D., and Cowles, R. (2008). Neuroblastoma—remembering the three physicians who described it a century ago: James Homer Wright, William Pepper, and Robert Hutchison. *Pediatric Radiology*, *39*(2), 155-160. http://dx.doi.org/10.1007/s00247-008-1062-z

Sanger, F., Nicklen, S. and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467. http://dx.doi.org/10.1073/pnas.74.12.5463

Silván, C. (2017). *Catecolaminas: Síntesis, Liberación y Funciones - Lifeder*. *Lifeder*. Retrieved 9 June 2017, from https://www.lifeder.com/catecolaminas/

Sistema nervioso simpático - Definición. (2017). *CCM Salud*. Retrieved 9 June 2017, from http://salud.ccm.net/faq/9880-sistema-nervioso-simpatico-definicion

Theruvath, J. et al. (2016). Next-generation sequencing reveals germline mutations in an infant with synchronous ocurrence of nephro- and neuroblastoma. *Pediatric Hematology and Oncology*, 33(4), 264-275. http://dx.doi.org/10.1080/08880018.2016.1184362

Tryka, K., Hao, L., Sturcke, A., Jin, Y., Wang, Z., and Ziyabari, L. et al. (2013). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, *42*(D1), D975-D979. http://dx.doi.org/10.1093/nar/gkt1211

Van Dijk, E., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9), 418-426. http://dx.doi.org/10.1016/j.tig.2014.07.001

Zettler, P., Sherkow, J., & Greely, H. (2014). 23andMe, the Food and Drug Administration, and the Future of Genetic Testing. *JAMA Internal Medicine*, *174*(4), 493. http://dx.doi.org/10.1001/jamainternmed.2013.14706

PRESUPUESTO

DOCUMENTO II

"Diseño de un Sistema de Información Genómica para el Diagnóstico del Neuroblastoma"

Clara Soler Pellicer

Curso 2016/2017

PRESUPUESTO

1. OBJETIVO DEL PRESUPUESTO

El objetivo del presupuesto es valorar económicamente el trabajo realizado.

A continuación, se va a detallar el presupuesto necesario para el desarrollo del Trabajo Fin de Grado (TFG) expuesto en la Memoria. El presupuesto contiene los costes de: *personal*, *software* y *hardware*. Los costes desglosados no incluyen el IVA, este se aplica al presupuesto total final.

Para el cálculo de la amortización en los costes de software y hardware, se ha utilizado la siguiente fórmula:

$$Coste\ imputable\ (\text{sin}\ IVA) = t \cdot \frac{C}{T}$$

Donde:

- *t* es el tiempo de uso del equipo (en meses)
- *C* es el coste del equipo/licencia
- *T* es el tiempo de amortización (en meses)

2. PRESUPUESTO DESGLOSADO

2.1. Costes de personal

El presupuesto del presente trabajo proviene principalmente de los costes de personal, debido a las tareas que han tenido que realizarse. En este caso, a parte del coste del ingeniero biomédico que ha realizado el trabajo, también se contabilizan las horas que han invertido los miembros del Grupo de Investigación PROS en sus tareas de apoyo. En la Tabla 1 se computan el total de horas de dedicación al proyecto por parte del ingeniero biomédico y de los ingenieros de PROS. Además, se detalla el coste unitario, las horas realizadas y a partir de éstos el coste imputable.

COSTES DE PERSONAL			
Perfil	Coste unitario	Horas realizadas	Coste imputable
Ingeniero biomédico	40€/hora	300 horas	12.000,00€
Ingeniero informático (PROS)	70€/hora	20 horas	1.400,00€
COSTES DE PERSONAL TOTALES			13.400,00€

Tabla 14.Costes de personal para la elaboración del proyecto.

2.2. Costes de software

En esta parte del presupuesto se detallan las licencias necesarias, así como el coste de las aplicaciones utilizadas. Esta parte del presupuesto incluye el coste de las licencias del Sistema Operativo y Microsoft Office, además de la utilización del gestor de bases de datos HeidiSQL y la herramienta "*VarSearch*", ambos gratuitos.

COSTES DE SOFTWARE					
Programa	Coste de la licencia	Número de licencias	Periodo de uso	Duración de la licencia	Coste imputable
Sistema Operativo Microsoft Windows 10 Pro	230,58€	1	6 meses	Indefinida (uso de 3 años)	38,43€
Microsoft Office Hogar y Estudiantes 2016	123,14€	1	6 meses	Indefinida (uso de 3 años)	20,52€
HeidiSQL	0,00€	1	6 meses	Indefinida	0,00€
VarSearch	0,00€	1	6 meses	Indefinida	0,00€
COSTES DE SOFTWARE TOTALES					58,95€

Tabla 15. Costes de software para la elaboración del proyecto.

2.3. Costes de hardware

A continuación, se detalla el coste de hardware. En el presente trabajo se ha utilizado el ordenador portátil *Toshiba Satellite C855-2DR*.

COSTES DE HARDWARE					
Equipo	Coste del equipo (sin IVA)	Unidades	Periodo de amortización	Periodo de uso	Coste imputable (sin IVA)
Toshiba Satellite C855-2DR	550,00€	1	3 años	6 meses	91,67€
COSTES DE HARDWARE TOTALES					91,67€

Tabla 16. Costes de hardware para la elaboración del proyecto.

3. PRESUPUESTO TOTAL

En primer lugar, se calcula el *Presupuesto de Ejecución Material* que corresponde a la suma de los presupuestos parciales mencionados anteriormente (costes de personal, software y hardware). A continuación, se calculan los gastos generales (13% del presupuesto de ejecución) y el beneficio industrial (6% del presupuesto de ejecución). Los gastos generales se refieren a los gastos necesarios para llevar a cabo la actividad de la empresa pero que no están directamente relacionados con el producto o servicio que esta ofrece, es decir, que no aumentan el beneficio de la empresa. Y el beneficio industrial, se corresponde con lo que se gana verdaderamente al hacer el proyecto. La suma del presupuesto de ejecución, los gastos generales y el beneficio industrial forman el *Presupuesto de Ejecución por Contrata*. Por último, se le añade el IVA para el cálculo del presupuesto total final.

CÁLCULO DEL PRESUPUESTO TOTAL	
Costes de personal totales	13.400,00€
Costes de software totales	58,95€
Costes de hardware totales	91,67€
Total del presupuesto de Ejecución Material	**13.550,62€**
Gastos generales (13%)	1.761,58€
Beneficio industrial (6%)	813,04€
Total del presupuesto de Ejecución por Contrata	**16.125,24€**
IVA (21%)	3.386,30€
PRESUPUESTO TOTAL	**19.511,54€**

Tabla 17. Cálculo del presupuesto total desglosado.

El presupuesto total asciende a **diecinueve mil quinientos once con cincuenta y cuatro euros (19.511,54€).**

ANEXOS

DOCUMENTO III

“Diseño de un sistema de información genómica para el diagnóstico del Neuroblastoma”

Clara Soler Pellicer

Curso 2016/2017

ANEXO 1. TOTAL DE VARIACIONES ENCONTRADAS (DESCRITAS SEGÚN SU IDENTIFICADOR RS DE DBSNP).

dbGaP					ClinVar		
6435862	4719687	1858111	7152233	13001220	149705989	376175333	113994091
4712653	11037575	2121283	17862054	9653034	146663377	35228363	576431612
110419	11037575	2121283	17862054	9653034	867021284	145780832	200585833
6939340	6782527	313528	6441188	2480775	181820595	2293564	878854661
4712653	6782527	313528	6441188	2480775	528568887	56165377	878854656
4712653	1989838	977867	2830495	2991769	140240544	35093491	147102592
6939340	1989838	977867	2830495	2991769	886044965	149968229	755556501
6939340	2974129	17047538	3794370	16849204	112765394	150966028	138589984
7587476	2974129	17047538	3794370	16849204	749389756	2246745	756986057
7587476	17842780	2733332	6778370	12029359	143654307	115392387	536284304
9295536	17842780	2733332	6778370	12029359	530566864	77677701	372440265
9295536	4487594	17012789	1946295	2178907	138324955	201490095	773447647
3768716	4487594	17012789	1946295	2178907	12402052	4358080	749418931
3768716	932206	731956	4566595	9863794	144889528	1063611	878854654
3768716	932206	731956	4566595	9863794	114084418	530550940	878854653
17487792	7837606	4234840	2340388	17578222	771929965	11723860	878854659
17487792	7837606	4234840	2340388	17578222	886044966	59260453	74774946
10498025	16967789	343166	2256432	1013812	886044967	6826373	773367495
10498025	16967789	343166	2256432	1013812	139613776	62412180	540427775
110419	6716793	2885663	7639867	162096	755866386	180795407	773380015
110419	6716793	2885663	7639867	162096	886044975	560413438	762395127
6939340	12189640	4505549	2078786	9287300	886044976	118046131	878854657
6939340	12189640	4505549	2078786	9287300	377570278	535962589	147858673
2592232	4632732	16841387	10936142	9682099	763679404	397840867	776228721
2592232	4632732	16841387	10936142	9682099	886044977	544491872	397706189
4712653	10248903	1123848	8068445	1519337	17034660	73139116	182561050
4712653	10248903	1123848	8068445	1519337	3753037	577950819	80088378
6715570	7557557	4237255	895459	4438469	142881321	201654270	186421480
6715570	7557557	4237255	895459	4438469	121908161	186778106	1881422
9295536	7679676	196048	204938	11830829	117525287	114290493	141010693
9295536	7679676	196048	204938	11830829	140015591	75913938	372612147
6712055	2164210	12455846	2835340	7584646	145846362	17885864	74716434
6712055	2164210	12455846	2835340	7584646	121908162	747626591	56146053
6435862	1489550	4609161	1714524	3828112	755314640	17882335	146074150
6435862	1489550	4609161	1714524	3828112	150831576	751829128	543328121
10498026	2053710	4954564	10505975	12505133	150904940	757355779	138406372
10498026	2053710	4954564	10505975	12505133	149566646	17879258	150292405
4758051	3134899	2745636	1925035	454212	2297881	17885216	57881134
4758051	3134899	2745636	1925035	454212	200470260	771616235	560163657
3790171	1783596	907548	17210989	3010493	3215996	886059409	576431612
3790171	1783596	907548	17210989	3010493	574168097	886059410	56270786
7272481	749873	2268225	17105232	9393227	140229905	886059411	141444957
7272481	749873	2268225	17105232	9393227	147318592	781647693	72857538
1561277	10229203	1921246	4779984	6073330	868389032	886059412	72857540

1561277	10229203	1921246	4779984	6073330	745380720	745503233	886055928
1703940	12660382	4930431	1374230	11950562	121908163	558416040	886055929
1703940	12660382	4930431	1374230	11950562	121908163	886059413	144437923
9892996	1117751	3795857	939114	12941604	886044985	770841700	79339096
9892996	1117751	3795857	939114	12941604	114266141	776498322	886055930
4755731	11677798	11601325	1680730	13420111	76519832	775569375	55941323
4755731	11677798	11601325	1680730	13420111	140769726	886059414	780536554
309160	8068707	351408	2619046	7641536	116089798	886059415	368581969
309160	8068707	351408	2619046	7641536	760253167	774951337	143916398
1027702	11185668	341933	2991770	6014573	772429569	104893855	367674546
1027702	11185668	341933	2991770	6014573	149417293	104893856	886055931
1427065	4487594	10840002	1952224	11714239	12125492	192127241	367712624
1427065	4487594	10840002	1952224	11714239	121908164	56247462	886055932
6441201	6817411	2279927	11648088	9648623	147066476	373764155	200868013
6441201	6817411	2279927	11648088	9648623	751084365	75895956	886055933
11681160	9350051	9832305	2094596	957459	77172218	74830770	780271684
11681160	9350051	9832305	2094596	957459	145969842	61754933	760041708
309137	2994598	11759745	11727649	186895	72867431	749286400	753267950
309137	2994598	11759745	11727649	186895	143669846	878854655	756782371
110419	3820449	16951664	9827781	1454909	146436697	373605278	767822322
110419	3820449	16951664	9827781	1454909	75413741	147241767	886055934
2224536	1915304	4620711	6714748	12151161	11121552	375646347	886055935
2224536	1915304	4620711	6714748	12151161	61999305	56249474	578166431
13150445	10090288	3768716	12295951	4657121	886055925	367642503	886055936
13150445	10090288	3768716	12295951	4657121	1728828	142126984	572286447
6430612	2415603	11773271	6735489	1218885	886055926	372471601	755231246
6430612	2415603	11773271	6735489	1218885	547955328	183314518	886055937
6744811	5752592	757558	7114014	6024619	74669215	780746584	886055938
6744811	5752592	757558	7114014	6024619	886055927	776604754	547587734
2710638	6774870	885769	8083994	12210008	78174819	755251438	886055939
2710638	6774870	885769	8083994	12210008	1881421	878854660	1536262
4084113	313519	17861942	6719500	218626	139437088	878854658	1002076
4084113	313519	17861942	6719500	218626	1881420	779178799	1138791
4696715	10753713	704225	1063178	1455576	1670283	761628666	3748581
4696715	10753713	704225	1063178	1455576	56132472	2246745	886044987
1714518	10456051	7637803	5326	1604144	138827116	371592787	148690591
1714518	10456051	7637803	5326	1604144	56181542	76150405	78490707
4770073	1001555	7637803	1218897	1604144	55772745	577768628	146717943
4770073	1001555	7637803	1218897	1604144	201768549	863225282	7544928
16980240	4755731	4267812	6760875	1604144	17007646	863225285	4240912
16980240	4755731	4267812	6760875	1604144	3738869	281864720	4240913
6929659	11037575	4971342	2017620	1604144	56247462	863225281	2847347
6929659	11037575	4971342	2017620	1604144	56071005	281864719	557129908
1012646	6945242	4834308	16876210	1604144	3795850	863225284	181454124
1012646	6945242	4834308	16876210	1604144	4622670	863225283	567547345
2991769	7486178	4978813	2340388	12203592	2293563	863225281	6694522
2991769	7486178	4978813	2340388	12203592	571451087	185749715	886045009
1892577	6911198	2011946	17322289	12145940	192312673	190108168	762218807
1892577	6911198	2011946	17322289	12145940	886045011	574858597	756198031
2991770	11591988	7632059	10932572	1656377	77474900	886044999	192963286
2991770	11591988	7632059	10932572	1656377	886045012	886045000	536529721

4689963	563916	8017950	12207718	11616409	72867441	750420134	549049444
4689963	563916	8017950	12207718	11616409	150497684	886045001	886045010
4719687	1858111	7152233	13001220	7765425	201136295	886045002	886045013
11131168	2401711	11716327	16872251	3759387	886044992	886045003	141688466
11131168	2401711	11716327	16872251	3759387	554239975	369817908	886045014
10864463	1193510	10089685	2059320	3759387	886044994	886045004	535976639
10864463	1193510	10089685	2059320	3759387	549614550	546229540	886045015
17047538	11562721	6992153	4613118	10091852	189075267	886045005	775958997
664576	6069551	6992153	6495001	10091852	886044995	886045006	886045016
664576	6069551	13359418	6495001	8181023	182518399	41301987	886044989
10936131	7730742	13359418	2829156	8181023	763386544	886045007	886044986
10936131	7730742	11562721	2829156	4613118	886044996	886045008	886044988
7765425	6038449	2822524	1883628	12434331	755083691	765149352	886044986
1463142	5752592	4942925	1523993	7980416	886044997	559140260	886044986
1463142	5752592	4942925	1523993	7980416	886044998	551392566	886044991
10035707	10181051	10832869	882362	7134594	886044990	55733526	113994088
10035707	10181051	10832869	882362	7134594	886044993	77102810	41310365
13432282	9466269	10511979	2058804	7624303	571589510	113994090	139210251
13432282	9466269	10511979	2058804	7624303	2256740	113994087	78868210
2209761	2822612	4850679	11611238	16852600	145271283	113994089	553404052
2209761	2822612	4850679	11611238	16852600	138668699	281864719	144047666
6038449	2822524	1883628	12434331	2897834	150557024	113994092	3748580
2897834	7129386	9857883	8073596	16842717	61754865	281864720	184232081
17065417	949455	1056927	6446700	1538972	35073634	115172954	281864719
17065417	949455	1056927	6446700	1538972	GRCh38/hg38 7q21.13(chr7:88520043-90241919)x3 *	GRCh38/hg38 3q28(chr3:191319392-191543941)x1 *	
7298565	10886492	4594848	12514981	4336470			
7298565	10886492	4594848	12514981	4336470			
245852	6682554	2822556	4134468	4902467			
245852	6682554	2822556	4134468	4902467			
7129386	9857883	8073596	16842717	12745240			
12745240	4780959	867595	3915772	4769414			
4780959	867595	3915772	17047538	4769414			
1489802	1489802						

Tabla 18. Variaciones encontradas.

*Estas dos variaciones aparecen como anomalías que afectan a varios genes y a varias posiciones distintas. No tienen número de identificación rs, así que se expresan como aparecen en la base de datos ClinVar. Estas variaciones acaban siendo filtradas en el paso *Identification* de la metodología SILE como se explica en el Capítulo 4.

ANEXO 2. VARIACIONES VALIDADAS/FILTRADAS (DESCRITAS SEGÚN SU IDENTIFICADOR RS DE DBSNP).

Total de variaciones validadas					
149705989	11121552	118046131	371592787	144437923	886044995
146663377	61999305	535962589	76150405	79339096	182518399
867021284	1561277	397840867	577768628	886055930	763386544
181820595	309160	544491872	863225282	55941323	886044996
528568887	11681160	73139116	863225285	780536554	755083691
140240544	3768716	577950819	281864720	368581969	886044997
886044965	17487792	201654270	863225281	143916398	886044998
112765394	7587476	186778106	281864719	367674546	185749715
749389756	6712055	114290493	863225284	886055931	190108168
143654307	6435862	75913938	863225283	367712624	574858597
530566864	6715570	17885864	863225281	886055932	886044999
138324955	886055925	747626591	281864719	200868013	886045000
12402052	1728828	17882335	281864719	886055933	750420134
144889528	886055926	751829128	113994092	780271684	886045001
114084418	547955328	757355779	281864720	760041708	886045002
771929965	74669215	17879258	113994090	753267950	886045003
886044966	886055927	17885216	113994087	756782371	369817908
886044967	78174819	1427065	113994089	767822322	886045004
139613776	1881421	11037575	113994088	886055934	546229540
755866386	139437088	4084113	113994091	886055935	886045005
886044975	1881420	110419	576431612	578166431	886045006
886044976	1670283	16980240	200585833	886055936	41301987
377570278	56132472	1012646	878854661	572286447	886045007
763679404	138827116	3790171	878854656	755231246	886045008
886044977	56181542	771616235	147102592	886055937	765149352
17034660	55772745	886059409	755556501	886055938	559140260
3753037	201768549	886059410	138589984	547587734	551392566
142881321	17007646	886059411	756986057	886055939	886045009
121908161	3738869	781647693	536284304	1536262	762218807
117525287	56247462	886059412	372440265	1002076	756198031
140015591	56071005	745503233	773447647	1138791	192963286
145846362	3795850	558416040	749418931	3748581	536529721
121908162	4622670	886059413	878854654	886044987	549049444
755314640	2293563	770841700	878854653	148690591	886045010
150831576	35073634	776498322	878854659	78490707	886045013
150904940	138668699	775569375	74774946	146717943	141688466
149566646	150557024	886059414	773367495	7544928	886045014
2297881	61754865	886059415	540427775	4240912	535976639
200470260	2256740	774951337	773380015	4240913	886045015
3215996	145271283	104893855	762395127	2847347	775958997
574168097	55733526	104893856	878854657	557129908	886045016
140229905	77102810	192127241	147858673	181454124	886044989
147318592	376175333	56247462	776228721	567547345	886044986
868389032	35228363	373764155	397706189	6694522	886044988

745380720	145780832	75895956	182561050	115172954	886044986
121908163	2293564	74830770	80088378	144047666	886044986
121908163	56165377	61754933	186421480	3748580	886044991
886044985	35093491	749286400	1881422	184232081	886044990
114266141	149968229	878854655	141010693	41310365	886044993
76519832	150966028	373605278	372612147	139210251	571589510
140769726	2246745	147241767	74716434	78868210	4712653
116089798	115392387	375646347	56146053	553404052	6939340
760253167	77677701	56249474	146074150	571451087	9295536
772429569	201490095	367642503	543328121	192312673	4758051
149417293	4358080	142126984	138406372	886045011	10498026
12125492	1714518	372471601	150292405	77474900	2592232
121908164	13150445	183314518	57881134	886045012	7272481
147066476	1063611	780746584	560163657	72867441	10498025
751084365	530550940	776604754	576431612	150497684	75413741
77172218	11723860	755251438	56270786	201136295	560413438
145969842	59260453	878854660	141444957	886044992	2246745
72867431	6826373	878854658	72857538	554239975	886055929
143669846	62412180	779178799	72857540	886044994	189075267
146436697	180795407	761628666	886055928	549614550	

Tabla 19. Variaciones validadas/filtradas.