USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESIS DE DOCTORADO

OPTIMIZATION OF STATISTICAL AND BIOINFORMATIC METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR RARE DISEASE DIAGNOSIS

Iria Roca Otero

ESCUELA DE DOCTORADO INTERNACIONAL

PROGRAMA DE DOCTORADO EN INVESTIGACIÓN CLÍNICA EN MEDICINA

SANTIAGO DE COMPOSTELA

2019

USC

UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR DE LA TESIS

OPTIMIZATION OF STATISTICAL AND BIOINFORMATIC METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR RARE DISEASE DIAGNOSIS

Dña. Iria Roca Otero

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1) *La tesis abarca los resultados de la elaboración de mi trabajo.*
2) *En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.*
3) *La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.*
4) *Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.*

En Santiago de Compostela, 01 de octubre de 2019

Fdo. Iria Roca Otero

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

OPTIMIZATION OF STATISTICAL AND BIOINFORMATIC METHODS FOR THE ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR RARE DISEASE DIAGNOSIS

Dª. María Rosaura Leis Trabazo y Dª. María Luz Couce Pico

INFORMAN:

Que la presente tesis, se corresponde con el trabajo realizado por Dª. ***Iria Roca Otero****, bajo mi dirección, y autorizo su presentación, considerando que reúne l os requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.*

De acuerdo con el artículo 41 del Reglamento de Estudios de Doctorado, declara también que la presente tesis doctoral es idónea para ser defendida en base a la modalidad de COMPENDIO DE PUBLICACIONES, en los que la participación del doctorando/a fue decisiva para su elaboración.

La utilización de estos artículos en esta memoria, está en conocimiento de los coautores, tanto doctores como no doctores. Además, estos últimos tienen conocimiento de que ninguno de los trabajos aquí reunidos podrá ser presentado en ninguna otra tesis doctoral.

En Santiago de Compostela, 01 de octubre de 2019

Fdo. María Rosaura Leis Trabazo

Fdo. María Luz Couce Pico

Esta tesis fue parcialmente financiada por el Instituto de Salud Carlos III de Madrid, en el marco de Proyectos de Investigación en Salud 2013 (FIS: PI13/02177).

El estudio fue aprobado por el Comité de Ética de la Investigación de Santiago-Lugo (número de registro: 2015/410).

AGRADECIMIENTOS

En primer lugar, a Ana: gracias por todos estos años, por enseñarme, formarme, y por haberme introducido en el mundo de la genética, sin tu apoyo esta tesis no existiría. A Rosanna, por ayudarme a entender un poco mejor la parte de laboratorio. Gracias a las dos por estar ahí en los buenos y sobre todo en los malos momentos, sin vosotras no lo habría conseguido.
Agradecer también al Dr. Tubío, por darnos la oportunidad de continuar nuestro trabajo a pesar de las dificultades.

Y, por supuesto, a mi familia: a Jaime, gracias por estar a mi lado y ser un apoyo constante, todo sería más difícil sin ti. Por último, y más importante, a mis padres, sin cuyo esfuerzo y cariño no estaría donde estoy ni sería quien soy.

TABLE OF CONTENTS

RESUMEN

El objetivo de este trabajo es la evaluación y optimización del análisis bioinformático de los datos generados por secuenciación masiva (NGS) aplicada al diagnóstico de enfermedades raras (EERR).

Las enfermedades raras (aquellas con un ratio de incidencia inferior a 1 de cada 2.000 personas), suponen, en su conjunto, un problema de primera magnitud para los sistemas sanitarios de todo el mundo, puesto que su prevalencia combinada es similar a la de algunas enfermedades más comunes como la diabetes (se estima que 1 de cada 7 personas desarrollarán una enfermedad rara a lo largo de su vida). El hecho de que el espectro fenotípico de estas enfermedades sea enormemente amplio, sumado a que los cuadros clínicos (en el caso por ejemplo de las enfermedades con afectación neurológica) son altamente solapantes, hace que su diagnóstico sea especialmente complicado y se retrase en el tiempo durante años. Esto conlleva mucho sufrimiento para pacientes y familias, una sobrecarga del sistema de salud y la incapacidad de proporcionar un diagnóstico genético adecuado en tiempo y forma. Cuando no existía la tecnología NGS, el análisis genético era el último paso en el proceso diagnóstico de este tipo de enfermedades. Durante este proceso, que habitualmente duraba años, se sometía al paciente a numerosas pruebas funcionales, bioquímicas, de imagen, anatomo-patológicas, etc. En base a los síntomas clínicos y los hallazgos en estas pruebas se interrogaban uno o varios genes consecutivamente sin alcanzar, en la mayor parte de los casos, un diagnostico etiológico definitivo.

La aparición de la NGS con su enorme potencia tiene la capacidad de modificar los protocolos diagnósticos. Esta herramienta permite el análisis simultáneo de miles de genes, incluso de todo el genoma, en un corto período de tiempo, y a un precio razonable, convirtiendo así al análisis genético en un test apto para ser considerado prueba de primera línea en el estudio de las enfermedades sospechosas de tener una base genética. Sin embargo, esta potencia tiene una contrapartida: la ingente cantidad de datos generados por la NGS supone un importante reto a la hora de filtrar y analizar los resultados. Este proceso, junto con la falta de personal entrenado y preparado para abordar estos complejos problemas, constituye en la actualidad el cuello de botella a la hora de aplicar estas nuevas tecnologías al campo de la clínica.

En la NGS, el genoma es fragmentado de manera aleatoria en pequeños trozos que son amplificados y secuenciados (leídos) en paralelo por las plataformas de secuenciación. Esto da lugar a millones de lecturas que tienen que ser posteriormente alineadas contra el genoma de referencia (alineamiento). Además de ser capaces de procesar esta inmensa cantidad de datos de manera óptima, los programas de alineamiento deben enfrentarse a dos grandes retos: (1) las posibles diferencias entre la secuencia leída en el paciente y la región del genoma de referencia (diferencias que pueden ser producidas por la propia variabilidad interindividual o por la existencia de variantes poblacionales, o incluso por los errores intrínsecos propios de cada plataforma), y (2) la presencia de secuencias repetitivas a lo largo del genoma. Estas regiones repetitivas (que suponen cerca de la mitad del ADN humano) pueden producir ambigüedad en el alineamiento, haciendo que los programas no puedan identificar con claridad el origen de la lectura. Así, los alineadores tienen que conseguir un equilibrio entre ser lo suficientemente permisivos como para poder alinear lecturas con pequeñas variaciones respecto del genoma de referencia, y lo suficientemente estrictos como para poder asignar unívocamente cada lectura a su posición original.

Una vez las lecturas se encuentran alineadas, se procede a la detección de variantes, es decir, identificar las diferencias que existen entre el genoma del paciente secuenciado y el genoma de referencia. En este trabajo, las agrupamos en tres grandes grupos, según su efecto sobre el genoma y la forma de detectarlas: (1) las variantes puntuales (SNVs por sus siglas en inglés *single nucleotide variants*) y las pequeñas deleciones e inserciones (INDELs), (2) las variantes en el número de copias (CNVs *copy number variants*), y (3) las variantes de reordenamiento (inversiones, translocaciones, y grandes inserciones *de novo*).

Existen tres grandes aproximaciones actualmente basadas en la tecnología NGS: los llamados "paneles génicos", en los que se secuencian en paralelo las zonas codificantes (exones) de una lista de genes (generalmente agrupados en función del fenotipo con el que estén relacionados); la secuenciación del exoma completo (WES), en el que se secuencian todas las regiones codificantes conocidas presentes en el genoma humano; y, por último, la secuenciación del genoma completo (WGS). A pesar de que el análisis WGS es mucho más completo (permite identificar variantes no detectables con las otras dos aproximaciones, como las variantes de reordenamiento o las variantes que no se encuentran en regiones codificantes), y los resultados obtenidos son generalmente más homogéneos (la distribución de las lecturas es mucho más uniforme), las dos primeras técnicas, englobadas bajo el término de secuenciación dirigida, son las más habituales en la práctica clínica, dado que suponen un menor coste que la WGS, y los datos obtenidos son mucho más manejables y fáciles de interpretar. Dado que el objetivo del presente estudio es la aplicación de la NGS al diagnóstico de EERR, nos hemos centrado en el análisis de datos procedentes de secuenciación dirigida. Las variantes detectables mediante secuenciación dirigida son las SNVs e INDELs, y los CNVs situados en regiones codificantes (aquellos que contienen 1 o más exones de un gen o varios genes).

Las SNVs son las variantes más sencillas de detectar (se trata simplemente de la substitución de un nucleótido por otro), y existen numerosas herramientas para su identificación. Por su parte, la detección de INDELs es más compleja, dado que su presencia supone una dificultad añadida para los programas de alineamiento (faltan o sobran nucleótidos de la secuencia de referencia). Sin embargo, el mayor reto que suponen ambos tipos de variantes es determinar su posible implicación con la enfermedad. Este proceso se conoce comúnmente como priorización de variantes y consiste en identificar, de toda la larga lista de variantes detectadas en el genoma de un paciente, las que más probablemente puedan estar implicadas en el fenotipo clínico a estudio. El primer paso en la priorización de variantes es filtrar las variantes comunes (con una frecuencia en bases de datos públicas superior al 1%, o incluso al 0,5%), puesto que esas frecuencias son incompatibles con la incidencia de las EERR. El siguiente paso es evaluar el impacto que pueden llegar a tener según el tipo de variante (*missense*, *nonsense*, *splicing*, *frameshift*, etc.) y la posición genómica en la que se encuentran. Para ello existen numerosas herramientas bioinformáticas que permiten la evaluación *in silico* de su impacto funcional. Estas pautas son una práctica común en el análisis de datos NGS, pero un paso imprescindible y no tan común en la priorización de variantes es la evaluación de la tolerancia de cada gen a las variantes *missense* (la tolerancia mutacional). Nuestra experiencia a lo largo de estos años en la aplicación de la NGS al diagnóstico de EERR es que algunos genes admiten una o incluso varias variantes *missense* raras sin que ello de lugar a ningún tipo de patología, mientras que en otros la presencia de una única variante puede ser catastrófica y determina un fenotipo clínico enormemente grave. Esto quiere decir que la selección purificadora negativa es mucho más fuerte para unos genes que para otros. Una forma de valorar la sensibilidad de cada gen a la variación es observando su número de variantes *missense* poblacionales frente a la suma de variantes totales (sinónimas + *missense*) de dicho gen. El ratio *missense* / *missense*+sinónimas nos da una idea de lo conservado que debe estar ese gen para que sea

funcional y de la selección negativa a la que está sometido. Cuando el ratio es muy alto implica que no existe una gran fuerza conservadora actuando para que la secuencia aminoacídica de la proteína permanezca inalterable. Cuanto más bajo es el ratio indica que la fuerza que actúa para conservar la secuencia original es más fuerte en ese gen, e implica que cualquier cambio puede afectar seriamente a la funcionalidad de la proteína codificada. Para evaluar dicha tolerancia mutacional, hemos aplicado el método propuesto por Petrovski y colaboradores (Petrovski et al. 2013) a las variantes comunes de 659 muestras de individuos con ascendencia europea extraídas del Proyecto 1000 Genomas presentes en 1.670 genes relacionados con EERR neurológicas y metabólicas. Para cada gen, hemos obtenido un z-score definido como el residuo estudentizado obtenido a partir de la regresión del número total de variantes *missense* comunes (con frecuencia >0.5%) contra el número total de variantes *missense* y sinónimas comunes presentes en cada gen. Un z-score en torno a cero indica que el gen tiene el número esperado de variantes *missense* dado su tasa mutacional. Los genes con un z-score negativo son aquellos que tienen menos variantes *missense* de las esperadas, es decir, son genes muy conservados en los que la evolución elimina cualquier variante porque afecta a la funcionalidad de la proteína codificada por el gen, y por lo tanto menos tolerantes a la presencia de estas variantes. Por su parte, los valores positivos de z-score pertenecen a los genes más tolerantes a las variaciones *missense*, es decir, a los menos conservados. Así, este parámetro permite identificar los genes en los que la presencia de variantes *missense* tienen una mayor probabilidad de resultar deletéreas. Sin embargo, la probabilidad de detectar variantes raras *missense* en un gen también depende de su tamaño, dado que a mayor número de bases nucleotídicas de un fragmento mayor es la probabilidad de que se produzca una mutación de manera aleatoria. Por lo tanto, además del parámetro de tolerancia mutacional, también es importante estimar la probabilidad de detección de variantes raras en el gen utilizando muestras control. Para ello, utilizando las mismas muestras que para el cálculo del z-score, calculamos la probabilidad de detectar una (en el caso de genes con herencia dominante o ligados al cromosoma X) o dos (en el caso de genes con herencia recesiva) variantes raras en cada gen según una distribución de Poisson de parámetro λ igual a la frecuencia de una/dos variantes raras (<0.5%) en dicho gen. Así, vemos que, en genes bien conservados, la probabilidad de contener variantes raras puede encontrarse en el mismo rango que en genes poco conservados debido al gran tamaño de dichos genes. Por lo tanto, es fundamental tener en cuenta estos dos parámetros a la hora de priorizar variantes.

La conservación específica del nucleótido donde se produce el cambio también es de vital importancia, dado que, si una variante *missense* con un alto impacto funcional teórico se encuentra situada en una región muy poco conservada dentro del gen, es muy posible que dicha variante no sea patogénica. Hay programas específicos para determinar la conservación de un nucleótido a lo largo de la evolución (GERP, SIFT…). Otro aspecto crucial es tener en cuenta la arquitectura mutacional del gen en el que se encuentran las variantes. Para algunos genes la presencia de incluso varias variantes *missense* no supone un problema, ya que solamente las variantes de truncamiento pueden afectar a la funcionalidad del mismo, como por ejemplo en el caso del gen *TTN* o de *SYNE1*. En otros casos, sin embargo, las variantes de truncamiento son menos deletéreas que las *missense*; por ejemplo, en el gen *KCNQ2* las variantes de truncamiento dan lugar a fenotipos mucho menos severos que las variantes *missense*. Existen genes (como *SETPB1* o *LMNA*) en los que la posición relativa de la variante dentro del gen, así como el tipo de variante (truncamiento vs *missense*), pueden dar lugar a fenotipos diferentes.

Una vez evaluadas todas estas características, y priorizadas las variantes más probablemente relacionadas con el fenotipo del paciente, el último paso en el estudio de enfermedades de herencia dominante o ligada al X es determinar si las variantes son *de novo*. Dado que estas variantes no han estado sujetas a selección negativa, son las más probablemente

patogénicas. Por supuesto, en el caso de enfermedades recesivas es imprescindible si encontramos dos variantes determinar que están en cromosomas opuestos. Por ello, el estudio familiar de las variantes priorizadas en los pacientes es esencial para una correcta interpretación de los resultados de un análisis NGS.

Los CNVs son variantes que han sido implicadas en multitud de EERR y en las enfermedades del neurodesarrollo en particular (epilepsia, autismo, esquizofrenia, discapacidad intelectual, ...). Sin embargo, este tipo de variantes han sido (y continúan siendo) infra-detectadas, especialmente las de menor tamaño, debido a que en el pasado la secuenciación clásica era insensible a ellas y las tecnologías utilizadas para su detección a gran escala (la hibridación genómica comparativa o *CGH array*, y los *arrays* de SNPs) únicamente permitían identificar CNVs de un tamaño superior a 30kb. Así, los CNVs de entre 1-30kb, que parecen estar asociados a numerosas patologías y enfermedades, han sido infra-detectados de forma sistemática a menos que se buscasen específicamente en un gen concreto con metodologías como el MLPA o PCR en tiempo real.

La aparición de la NGS trajo consigo la capacidad de detectar CNVs de menor tamaño, pero no existían herramientas bioinformáticas adecuadas para su detección, y su uso requería de expertos en bioinformática que no están presentes en muchos centros de diagnóstico molecular.

Los métodos para la detección de CNVs a partir de datos NGS varían según se estén analizando datos de secuenciación dirigida o del genoma completo. Mientras para la detección de CNVs en WGS existen múltiples herramientas, el número de ellas desarrolladas para secuenciación dirigida es mucho menor (aunque ha aumentado considerablemente en los últimos años). La mayor parte de estas herramientas se basan en la comparación de los patrones de profundidad de cobertura entre la muestra a estudiar y un conjunto de muestras control. La principal diferencia entre los distintos métodos radica en la modelización estadística en la que se basan (modelos ocultos de Markov, de Poisson, binomial negativa, etc.), y en el proceso de filtrado que aplican para reducir el número de falsos positivos. Cuando nos planteamos elegir una de estas herramientas para la gestión de nuestros datos nos encontramos con que el mayor hándicap para poder evaluar dichas herramientas era conseguir el suficiente número de muestras para utilizar como controles positivos de CNVs. Por ello, nos planteamos crear una amplia base de datos de muestras generadas artificialmente con CNVs de diferente tamaño y en diferentes posiciones. A la hora de elegir los programas para la simulación de datos artificiales, tuvimos en cuenta varias cosas. En primer lugar, la mayor parte de las herramientas de simulación existentes fueron creadas para imitar datos de WGS, y no son válidas para generar datos artificiales que simulen datos de secuenciación dirigida. Otro aspecto importante es que, en general, estas herramientas se dividen entre las que generan lecturas artificiales y las que permiten introducir variantes en dichas lecturas. Además, como los principales problemas asociados a la detección de estas variantes son los sesgos generados por el contenido GC, la presencia de secuencias repetitivas, el tipo de secuenciador utilizado, etc., es importante elegir un simulador de datos NGS que pueda reproducir esta variabilidad. Teniendo en cuenta estas limitaciones, concluimos que la aproximación óptima era utilizar Wessim (S. Kim, Jeong, and Bafna 2013) para la generación de lecturas simuladas que imiten las generadas en secuenciación dirigida, y RSVSim (Bartenhagen and Dugas 2013) para la introducción de CNVs en dichas lecturas. Con la combinación de ambas herramientas generamos 320 muestras simuladas con CNVs introducidos artificialmente (además de 20 muestras sin CNVs para ser utilizadas como controles negativos), a dos profundidades medias de cobertura diferentes (50X y 300X). Introdujimos duplicaciones y deleciones (tanto en heterocigosis como en homocigosis) de diferentes tamaños y en diferentes combinaciones. Con esta amplia base de datos NGS generada artificialmente, comparamos el rendimiento de 12 programas desarrollados para trabajar con datos de secuenciación dirigida: ExomeCNV (Sathirapongsasuti et al. 2011), ExomeCopy (Love et al. 2011), CONTRA (J. Li et al. 2012), ExomeDepth (Plagnol et al. 2012), CONIFER (Krumm et al. 2012), CANOES (Backenroth et al. 2014), CODEX (Jiang et al. 2015), CLAMMS (Packer et al. 2016), CoNVaDING (Johansson et al. 2016), DECoN (Fowler et al. 2016), CNVkit (Talevich et al. 2016), y SeqCNV (Chen et al. 2017). De los resultados obtenidos, sacamos las siguientes conclusiones: la primera fue que todas las herramientas mostraban un mejor rendimiento con mayores profundidades de cobertura media, lo que era de esperar, dado que una menor cobertura media implica un mayor número de zonas con poca profundidad de cobertura en las que la pérdida o ganancia de cobertura producida por deleciones o duplicaciones es similar a la variación generada por el ruido de fondo. La segunda conclusión fue que, en general, las deleciones son más sencillas de detectar que las duplicaciones, lo que también era de esperar de forma intuitiva, dado que la diferencia de coberturas es más sutil en las duplicaciones que en las deleciones. Otra conclusión fue que los CNVs que contienen un mayor número de exones son, en general, más fáciles de detectar que los de tamaño más reducido. Esto también era esperable, dado que cuanto mayor sea un CNV más difícil es que la diferencia en la profundidad de cobertura se pueda confundir con ruido de fondo. Encontramos que las herramientas que mejores resultados obtuvieron fueron DECoN, ExomeDepth, ExomeCNV, CANOES y CoNVaDING. Sin embargo, dado que ninguna de ellas consiguió un 100% de sensibilidad, quisimos identificar cuál sería la combinación óptima para conseguir eliminar los falsos negativos. Nuestra aproximación fue la de clasificar una región como CNV si al menos tres herramientas diferentes la categorizaban como tal. Los resultados obtenidos fueron bastante decepcionantes: para conseguir detectar todos los CNVs de las muestras simuladas, fue necesario combinar los resultados de al menos 9 herramientas diferentes, lo que supone un incremento considerable del tiempo y del coste computacional del análisis. Cabe resaltar que las muestras artificiales no pueden reproducir al 100% la complejidad de las muestras reales, por lo que estos resultados no son directamente extrapolables al análisis real; para lo que sirven es para identificar las tendencias generales (qué herramientas detectan mejor qué tipo de CNV, cuáles tienen menor número de falsos positivos, etc.).

A la vista de las carencias que tenían las herramientas existentes para la detección de estas variantes, decidimos desarrollar un programa de detección de CNVs enfocado en analizar datos de paneles génicos, y que fuese especialmente sensible a los CNVs de menor tamaño (aquellos que contengan un único exón). Para ello, realizamos primero un análisis exhaustivo de las posibles causas de variabilidad en los patrones de cobertura entre las muestras generadas en los análisis de secuenciación dirigida. Los resultados confirmaban algunos de los ya publicados: el contenido GC, la variabilidad técnica en la preparación de librerías y la secuenciación, las modificaciones en el diseño de paneles génicos, la integridad inicial del ADN etc., son factores que implican importantes cambios en la homogeneidad del perfil de cobertura entre muestras. Por lo tanto, a la hora de obtener resultados fiables con los programas basados en la comparación de patrones de cobertura, es crucial maximizar la homogeneidad entre las muestras. Esto se puede conseguir procesando las muestras en paralelo y de la misma forma, y maximizando la profundidad de cobertura media para aumentar la cobertura de las zonas con alto contenido GC.

El algoritmo de detección que desarrollamos (PattRec) aplica una normalización diferente dependiendo de si se están analizando exones o genes completos. En el caso de exones, el método utilizado es el siguiente: para cada nucleótido, la profundidad de cobertura se divide por la cobertura máxima del gen que lo contiene. Para evitar seleccionar erróneamente una duplicación como valor máximo, desarrollamos una subrutina para cada gen combinando la prueba de Chi-cuadrado para la detección de valores atípicos y el algoritmo de agrupamiento de las *k-means*. Una vez calculado el máximo, se calcula para cada nucleótido su log-ratio:

$$logratio_k = \log\left(\frac{normdoc_k(test)}{normdoc_k(cont)}\right)$$

donde: $normdoc_k(test)$ y $normdoc_k(cont)$ representan las coberturas normalizadas del test y del control (o de la media de los controles en el caso de que haya más de uno) en el nucleótido k, respectivamente. Estos log-ratios siguen una distribución normal $N(\mu_{exon}, \sigma_G)$, donde μ_{exon} es la media de todos los log-ratios del exón, y σ_G es la desviación típica de todos los log-ratios del gen. Los CNVs adyacentes del mismo tipo con un p-valor inferior a 0,05 y un porcentaje de subida/bajada similar se concatenan en una única región. El p-valor resultante es una modificación del método de Fisher para la combinación de probabilidades (corregido para pruebas dependientes, como está implementado en el paquete de R 'poolR'). Para reducir el número de falsos positivos, hacemos la regresión lineal de la cobertura media del test sobre la cobertura media normalizada del control para exón, y cada p-valor es penalizado en función de su distancia a los valores ajustados. Por último, aplicamos la corrección de Benjamini-Hochberg o la de Bonferroni en función del número de resultados obtenidos. En el archivo de salida se reportan los CNVs que tengan un p-valor <0,05 y un porcentaje de subida/bajada superior al 35% para deleciones y al 30% para duplicaciones (parámetros ajustables por el usuario). Para el análisis de genes completos se aplica el mismo método (a excepción de la penalización mediante la regresión), utilizando en este caso la siguiente normalización: si la muestra es de sexo femenino, la cobertura de cada nucleótido del gen se divide por la cobertura media global de la muestra. Si es de sexo masculino, los genes autosómicos se dividen por la cobertura media global de dichos genes, y los genes del cromosoma X se dividen por la media de dicho cromosoma.

Una vez optimizado PattRec, comparamos su rendimiento con el de 8 herramientas de detección de CNVs (ExomeDepth, ExomeCopy, ExomeCNV, CONTRA, CODEX, CLAMMS, SeqCNV y CNVkit), utilizando tanto datos de muestras con CNVs cedidas por otros laboratorios y secuenciadas de manera óptima (maximizando profundidad de cobertura, secuenciando al mismo tiempo test y controles, etc.), como datos de muestras de acceso libre (del Proyecto 1000 Genomas). En el caso de las muestras con CNVs reales secuenciadas en nuestro laboratorio, las herramientas con mayor sensibilidad fueron PattRec y ExomeCNV, seguidas de ExomeDepth y CNVkit, mientras que las muestras con mayor especificidad fueron PattRec, ExomeDepth y CONTRA, en ese orden. Los resultados obtenidos con las muestras del Proyecto 1000 Genomas fueron mucho peores (tanto en términos de sensibilidad como de especificidad), seguramente debido a la poca uniformidad existente entre los patrones de cobertura de las muestras (la media de correlación entre las profundidades de cobertura globales era inferior a 0,5), lo que resalta la importancia de minimizar la variabilidad en los patrones de cobertura entre la muestra a estudiar y los controles utilizados.

En nuestro primer manuscrito explicamos cómo siguiendo estas pautas, la lista de variantes raras detectadas puede restringirse de forma mucho más óptima a las que tienen más probabilidades de estar implicado en el fenotipo clínico del paciente, evitando en algunos casos la necesidad de realizar estudios funcionales que implican mucho tiempo y coste. En nuestro segundo manuscrito describimos una forma de generar datos NGS artificiales e introducir CNVs en ellos que permite evaluar el rendimiento de las herramientas existentes para la detección de CNVs. Con ellos comparamos 12 herramientas de detección de CNVs, evaluando los puntos fuertes y débiles de cada una. En el último manuscrito, presentamos un programa para la detección de CNVs, diseñado específicamente para trabajar con datos de paneles génicos, de fácil uso para laboratorios sin gran experiencia en bioinformática y especialmente sensible a los pequeños CNVs, y hemos comparado su rendimiento con el de otros programas existentes.

En resumen, con estos tres trabajos hemos pretendido optimizar al máximo el diagnóstico de las EERR mediante el uso de secuenciación dirigida. Somos conscientes de que quedan muchas lagunas que cubrir en el diagnóstico de EERR, como la detección fiable de mosaicismos, la detección de variantes fuera de regiones codificantes, el análisis de enfermedades atendiendo a su posible origen oligogénico, o las variantes de reordenamiento. Estos problemas serán abordados en el futuro inmediato, mediante la aplicación de la WGS al análisis de EERR.

ABSTRACT

The main goal of this thesis was to evaluate and optimize the bioinformatic analysis of data generated by next generation sequencing (NGS) technologies to facilitate the diagnosis of rare diseases (RDs).

RDs (diseases that affect fewer than 1 in 2,000 people) constitute a major problem for health systems around the world, with a combined prevalence comparable to that of more common diseases such as diabetes. Indeed, it is estimated that 1 in 7 people will develop a RD throughout their lives. The very broad phenotypic spectrum of these diseases, together with the significant overlap in clinical presentations (*e.g.,* diseases with neurological involvement) make diagnosis especially complex and time-consuming. The inability to provide patients with a timely genetic diagnosis results in protracted suffering for them and their families, as well as overload of health systems. Before the emergence of NGS technologies genetic analysis was considered the final step in the diagnostic process in RD patients. During this process, which usually lasted years, patients would undergo multiple tests (functional, biochemical, imaging, anatomic-pathological, etc.). Based on the results obtained and the patient's clinical signs, individuals with suspected genetic disorders were referred for analysis of a candidate gene by classical sequencing. In most cases this approach failed to establish a definitive etiological diagnosis.

NGS has important implications for the future of RD diagnosis. This tool allows rapid, cost effective, simultaneous analysis of thousands of genes, or even the entire genome, and has the potential to make genetic analysis a first-line test in the study of diseases with a suspected genetic component. However, users of NGS are faced with a new challenge: the huge amount of data generated poses major difficulties in when filtering and analyzing the results. This drawback, together with the lack of adequately trained personnel required to address these complex problems, has led to a bottleneck limiting the clinical application of these new technologies.

In NGS the genome is randomly fragmented into small pieces that are amplified and sequenced (read) in parallel by sequencing platforms. This results in millions of reads that must be subsequently aligned against a reference genome. Alignment programs must be capable of optimally processing huge amounts of data and addressing two key challenges: (1) possible differences that arise between the patient's sequence and the corresponding region in the reference genome (due to inter-individual variability, the existence of population variants, or error intrinsic to a given platform); and (2) the presence of repetitive sequences throughout the genome. These repetitive regions (which account for approximately half of all human DNA) can lead to ambiguities in the alignment data, hindering clear identification of the origin of the read. Sequence alignment tools must therefore strike a balance between being sufficiently permissive to ensure alignment of reads with small variations with respect to the reference genome, and being strict enough to be able to uniquely assign each read to its original position.

Once reads are aligned, the next step is variant detection, *i.e.,* identification of differences between the patient's sequenced genome and the reference genome. Variants can be clustered into three main types according to their effect on the genome and the manner in which they are detected: (1) single nucleotide variants (SNVs) and small deletions and insertions (INDELs); (2) copy number variants (CNVs); and (3) rearrangement variants (inversions, translocations, and large *de novo* insertions).

Three major NGS-based approaches are currently used: so-called “gene panels”, in which the coding areas (exons) of a list of genes are sequenced in parallel (usually grouped according to the phenotype with which they are related); whole-exome sequencing (WES), in which all known coding regions present in the human genome are sequenced; and whole-genome sequencing (WGS). WGS enables the most complete analysis, allowing identification of variants that cannot be detected using the other two approaches (*e.g.,* rearrangement variants and variants located outside of coding regions) and the results obtained are generally more homogeneous (read distribution is much more uniform). Nonetheless, the first two techniques, encompassed under the term “targeted sequencing”, are the most common in clinical practice: they are less costly than WGS and the data obtained is much more manageable and easier to interpret. Since the objective of this study was to apply NGS to the diagnosis of RDs, we have focused on the analysis of data produced by targeted sequencing approaches. The variants detectable by targeted sequencing are SNVs and INDELs, and CNVs located in coding regions (those that contain 1 or more exons of a gene or several genes).

SNVs are the result of the substitution of one nucleotide for another. They are therefore the simplest type of variant to detect, and there are a range of tools available that do so effectively. By contrast, the detection of INDELs is more complex: their presence results in an excess or deficit of nucleotides with respect to the reference sequence, creating added difficulties for alignment programs. However, for both SNVs and INDELs the greatest challenge is determining their possible involvement in the patient’s disease. This process is commonly known as variant prioritization and consists of identifying, from the entire list of variants detected in the patient’s genome, those most likely implicated in the clinical phenotype under study. The first step in prioritizing variants is to filter common variants (those with a frequency >1% in public databases, or even >0.5%): these frequencies are incompatible with the incidence of RD. The next step is to evaluate the impact according to the type of variant (missense, nonsense, splicing, frameshift, etc.) and its genomic position. There are numerous bioinformatics tools that enable *in silico* evaluation of a variant’s functional impact. While these are basic steps in the analysis of NGS data, another essential but less commonly performed task is to prioritize variants according to the tolerance of the gene to missense variants (mutational tolerance). In our experience over several years of applying NGS to RD diagnosis, some genes can tolerate one or even several rare missense variants with no pathological consequences, while in others the presence of a single variant can be catastrophic and give rise to a very severe clinical phenotype. This implies that the negative purifying selection is much stronger for some genes than for others. One way of assessing a gene’s sensitivity to variation is to compare the number of population missense variants for each gene with the sum of all variants (synonymous + missense) in that gene. The missense / missense + synonymous ratio provides an indication of how conserved a gene must be in order to remain functional and the level of negative selection to which it is subjected. A high ratio implies that the gene is not subjected to a high level of conservative force, and therefore the amino acid sequence of the encoded protein remains unchanged. A lower ratio indicates that the gene is subjected to strong forces acting to conserve its original sequence, and implies that any change can seriously affect the functionality of the encoded protein. To evaluate mutational tolerance, we applied the method proposed by Petrovski et al. (Petrovski et al. 2013) to common variants in 1,670 genes implicated in rare neurological and metabolic diseases in 659 individuals with European ancestry. These data were extracted from the 1000 Genomes Project. For each gene, we calculated a z-score, defined as the studentized residual obtained by regression of the total number of common missense variants (frequency> 0.5%) against the total number of common missense and common synonymous variants present in each gene. A z-score value close to zero indicates that the gene harbors the expected number of missense variants given its mutational rate. Genes with a negative z-score are those with fewer than expected missense variants. These are highly conserved genes that are less tolerant of the presence of these variants: evolution eliminates any variant owing to its effect on the functionality of the encoded protein. Conversely, a positive z-score indicates a less conserved gene, which is tolerant of missense variations. This parameter allows identification of genes in which the presence of missense variants is more likely to be deleterious. However, the probability of detecting rare missense variants in a gene also depends on its size: the greater the number of nucleotide bases in a fragment, the greater the probability that a mutation will randomly occur. Therefore, in addition to evaluating mutational tolerance it is also important to estimate the probability of detecting rare variants in the gene using control samples. To do this, using the same samples as for z-score calculation, we determined the probability of detecting 1 (in the case of dominantly inherited or X-linked genes) or 2 (in the case of recessively inherited genes) rare variants in each gene according to a Poisson distribution of parameter λ equal to the frequency of 1 or 2 rare variants (<0.5%) in that gene. We found that after accounting for gene size the probability of the presence of a rare variant can be similar in highly conserved and poorly conserved genes, underscoring the importance of taking these two parameters into account when prioritizing variants.

Another parameter that must be considered is the specific conservation of the nucleotide in which a given change occurs. If a missense variant with a theoretically high functional impact is located in a very poorly conserved region within the gene, it is very possible that it will have no pathogenic repercussions. Specific programs can evaluate the conservation of a nucleotide throughout evolution (GERP, SIFT, etc.). Another crucial parameter to consider is the mutational architecture of the gene in which a variant is located. In some genes (*e.g., TTN*, *SYNE1)* the presence of even several missense variants may have no deleterious effect, and gene functionality is only affected by truncating variants. In other cases, truncating variants are less deleterious than missense variants. For example, in *KCNQ2* truncating variants give rise to much less severe phenotypes than missense variants. In other genes (*e.g., SETPB1*, *LMNA*) the resulting phenotype is determined by both the type of variant (truncating or missense) and its relative position within the gene.

Once all these characteristics have been evaluated, and the variants most likely to be implicated in the patient's phenotype are prioritized, the last stage in the study of dominantly inherited or X-linked diseases is to determine whether the variants have arisen *de novo*. Because *de novo* variants have not been subjected to negative selection, they are most likely to be pathogenic. Of course, in the case of recessive diseases if two variants are detected it is essential to determine whether they are located on opposite chromosomes. A family study of the prioritized variants is thus essential for correct interpretation of the results of NGS analyses.

CNVs have been implicated in many RDs, and in neurodevelopmental diseases in particular (epilepsy, autism, schizophrenia, intellectual disability). However, because these variants cannot be detected using classical sequencing techniques, and because the technologies used for large-scale detection (comparative genomic hybridization [CGH] and single nucleotide polymorphism [SNP] arrays) can only identify CNVs >30 kb, these types of variants have been (and continue to be) under-detected. In particular, CNVs of 1–30 kb, which appear to be implicated in numerous diseases, have been systematically under-detected unless specifically searched for in a particular gene using specific methodologies such as multiplex ligation-dependent probe amplification (MLPA) or real-time polymerase chain reaction (PCR).

While the emergence of NGS has facilitated the detection of smaller CNVs, there remains a dearth of adequate bioinformatics tools for their detection, and their use requires expertise in bioinformatics not typically found in molecular diagnostics centers.

The methods used for CNV detection from NGS data vary depending on whether the data are derived from targeted sequencing or WGS approaches. While there are multiple tools available for the detection of CNVs from WGS data, fewer have been designed for use with targeted sequencing data (although the number of these tools has increased considerably in recent years). Most of these tools are based on the comparison of depth-of-coverage patterns between the study sample and a set of control samples. The main difference between the methods lies in the type of statistical modeling on which they are based (hidden Markov models, Poisson, negative binomial models, etc.), and the filtering process applied to reduce the number of false positives. When evaluating the utility of each of these tools for the management of our data, we found that the most challenging aspect of the evaluation process was obtaining enough samples to use as positive CNV controls. Therefore, we set about creating a large database of artificially generated samples containing CNVs of varying sizes and positions. When selecting programs to generate simulated data we took several factors into account. First, most existing simulation tools have been developed to mimic WGS data and are not valid for generating artificial reads that simulate targeted sequencing data. Second, in general these tools can be divided into those that generate artificial reads and those that allow the introduction of variants into artificial reads. Third, because one of the key problems associated with the detection of these variants is the generation of biases caused by GC content, the presence of repetitive sequences, and the type of platform used, among other factors, it is important to choose an NGS data simulator that can reproduce this variability. Taking into account these limitations, we concluded that the optimal approach was to use Wessim (S. Kim, Jeong, and Bafna 2013) to generate simulated reads that mimic those generated in targeted sequencing, and RSVSim (Bartenhagen and Dugas 2013) to introduce CNVs into those reads. Using this combination of tools, we generated 320 simulated samples with artificially introduced CNVs (plus 20 samples without CNVs that served as negative controls) at two different mean depths of coverage (50X and 300X). We introduced duplications and deletions (both heterozygous and homozygous) of different sizes and in different combinations. Using this large, artificially generated NGS database we compared the performance of 12 programs designed to work with targeted sequencing data: ExomeCNV (Sathirapongsasuti et al. 2011), ExomeCopy (Love et al. 2011), CONTRA (J. Li et al. 2012), ExomeDepth (Plagnol et al. 2012), CONIFER (Krumm et al. 2012), CANOES (Backenroth et al. 2014), CODEX (Jiang et al. 2015), CLAMMS (Packer et al. 2016), CoNVaDING (Johansson et al. 2016), DECoN (Fowler et al. 2016), CNVkit (Talevich et al. 2016), and SeqCNV (Chen et al. 2017). Based on the results obtained, we can draw several conclusions. First, all tools performed better with greater mean depth-of-coverage. This finding was unsurprising: lower mean depth-of-coverage implies a greater number of areas poorly covered in which the loss or gain of coverage caused by deletions or duplications is difficult to distinguish from the variation generated by background noise. Second, in general deletions are easier to detect than duplications. This finding was also expected, since the difference in coverage is more subtle in the case of duplications than deletions. Third, CNVs containing greater numbers of exons are, in general, easier to detect than those of smaller size. This was also expected, given that the larger the CNV the less likely the difference in depth-of-coverage is confused with background noise. We found that the tools that produced the best results were DECoN, ExomeDepth, ExomeCNV, CANOES, and CoNVaDING. However, given that none achieved 100% sensitivity, we sought to identify the optimal combination of tools to eliminate false negatives. To this end, we classified a region as a CNV if it was categorized as such by at least three different tools. The results obtained were disappointing: in order to detect all CNVs in the simulated samples it was necessary to combine the results of at least 9 different tools, which entailed a considerable increase in computational time and cost. It should be noted that artificial samples cannot completely reproduce the complexity of real samples, and therefore these results should not be directly extrapolated to real analyses. However, they do allow us to identify general trends (*e.g.,* which tools best detect which type of CNV, which tools produce the fewest false positives, etc).

In view of the deficiencies of existing tools for the detection of these variants, we developed a CNV detection program for the analysis of gene panel data, with particular sensitivity for small (single-exon) CNVs. First, we performed an exhaustive analysis of the possible causes of variability in the coverage patterns between the samples obtained by targeted sequencing analysis. The factors identified were in good agreement with those previously described in the literature: GC content, technical variability in the preparation of libraries and sequencing, modifications in the design of gene panels, and initial integrity of the DNA are all factors that result in significant changes in the homogeneity of coverage profiles across samples. In order to obtain reliable results with programs based on the comparison of coverage patterns it is therefore crucial to maximize homogeneity across samples. This can be achieved by processing the samples in parallel and in the same conditions, and by maximizing the mean depth-of-coverage to increase coverage in areas with high GC content.

The detection algorithm we have developed (PattRec) applies a different normalization algorithm depending on whether exons or whole genes are being analyzed. In the case of exon analysis, for each nucleotide the depth-of-coverage is divided by the maximum coverage of the gene in which it is located. To avoid erroneous selection of a duplication as a maximum value, we developed a subroutine for each gene by combining the Chi-squared test for the detection of outliers and the k-means clustering algorithm. Once the maximum is calculated, its log-ratio is calculated for each nucleotide as follows:

$$logratio_k = \log\left(\frac{normdoc_k(test)}{normdoc_k(cont)}\right)$$

where $normdoc_k(test)$ and $normdoc_k(cont)$ represent the normalized coverage of the test and control samples (or of the mean of the controls if there are more than one) in nucleotide k, respectively. These log-ratios follow a normal distribution $N(\mu_{exon}, \sigma_G)$, where μ_{exon} is the mean of all exon log-ratios, and σ_G is the standard deviation of all the log-ratios of the gene. Adjacent CNVs of the same type with a p-value <0.05 and a similar percentage of coverage's increase/decrease are concatenated in a single region. The resulting p-value is a modification of Fisher's method for the combination of probabilities (corrected for non-independent tests, as implemented in the R 'poolR' package). To reduce the number of false positives, we perform linear regression of the mean coverage of the test against the normalized mean coverage of the control, and each p-value is penalized based on its distance from the adjusted values. Finally, we apply a Benjamini-Hochberg or Bonferroni correction depending on the number of results obtained. The output file reports CNVs with a p-value of <0.05 and a percentage of increase/decrease >35% for deletions and >30% for duplications (user adjustable parameters). For the analysis of whole genes, the same method is applied (except for the regression penalty), in this case using the following normalization: for female samples the coverage of each nucleotide is divided by the global mean coverage of the sample; for male samples, the coverage of autosomal genes is divided by the overall mean coverage of the corresponding genes, and the coverage of X chromosome genes is divided by the mean coverage of the corresponding chromosome.

Once PattRec was optimized, we compared its performance with that of 8 CNV detection tools (ExomeDepth, ExomeCopy, ExomeCNV, CONTRA, CODEX, CLAMMS, SeqCNV and CNVkit), using CNV-containing samples provided by other laboratories and sequenced optimally (by maximizing depth-of-coverage and processing samples at the same time, in the same laboratory, using the same sequencing kit), as well as samples obtained from public databases (the 1000 Genomes Project). In the analysis of real samples sequenced in our laboratory the most sensitive tools were PattRec and ExomeCNV, followed by ExomeDepth and CNVkit, while the most specific tool was PattRec, followed by ExomeDepth and CONTRA. The results obtained with the 1000 Genome Project samples were much poorer (in terms of both sensitivity and specificity), probably due to the heterogeneity of coverage patterns across samples (the mean correlation between overall depths-of-coverage was <0.5), highlighting the importance of minimizing the variability in coverage patterns between the study sample and the controls used.

In our first article we explain how, following the aforementioned guidelines, lists of rare variants detected can be more optimally created to include only those most likely to be involved in the patient's clinical phenotype, thereby reducing the need to perform costly and time-consuming functional studies. Our second article describes a method to generate artificial targeted NGS data into which CNVs can then be introduced, allowing us to evaluate the performance of existing CNV detection tools. We used this method to compare 12 CNV detection tools, evaluating the strengths and weaknesses of each. In our third article we present a CNV-detection program that is specifically designed to work with gene panel data, can be easily used in laboratories without the need for extensive bioinformatics experience, and is especially sensitive to small CNVs, and we compare its performance with that of other existing programs.

In summary, the goal of each of the three articles presented here is to optimize the diagnosis of RD through the use of targeted sequencing data. There remain many shortcomings in the diagnosis of RDs, including the need for reliable methods to detect mosaic variants, variants located outside coding regions, diseases with a possible oligogenic origin, or rearrangement variants. These problems will be addressed in the near future with the application of WGS to the analysis of RDs.

1. INTRODUCTION

1.1 THESIS LAYOUT

This doctoral thesis is presented as a compendium of three papers published in peer-reviewed scientific journals (Chapter 3), each with its own abstract, main text, and references. A brief summary of each is presented below.

The article ***Prioritization of variants detected by next generation sequencing according to the mutation tolerance and mutational architecture of the corresponding genes*** (https://doi.org/10.3390/ijms19061584), which was published in the *International Journal of Molecular Sciences* (2018 JCR Impact Factor, 4.183) is presented in Section 3.1. This paper discusses several key concepts relating to variant prioritization in the diagnosis of rare diseases. The first concerns the "mutational tolerance" of genes in which variants are located (*i.e.,* the susceptibility of a given gene to any missense variation). This depends on the strength of the purifying selection acting against the variant. The second concept is the "mutational architecture" of each gene. This is the type and location of previously identified mutations in the gene and their association with different phenotypes or degrees of severity. The third concept concerns the type of inheritance (inherited vs. *de novo*) of the variants detected. Using real data, we show that genes, as opposed to variants, can be prioritized by calculating a specific mutational tolerance parameter for a given gene. The influence of mutational architecture on variant prioritization is also illustrated using five paradigmatic examples. Finally, the importance of the analysis of variants in the patient's family as an essential step in variant prioritization is also discussed.

The article ***Free-access copy-number variant detection tools for targeted next-generation sequencing data*** (https://doi.org/10.1016/j.mrrev.2019.02.005), published in *Mutation Research-Reviews in Mutation Research* (2018 JCR IF, 6.081), is presented in Section 3.2. This article describes a method to generate artificial targeted next-generation sequencing (NGS) data that simulate the data produced by sequencing platforms. Specifically, we focus on tools that allow us to reproduce the biases and variability in coverage patterns found in real samples. Furthermore, we review methods for the detection of copy number variants (CNVs) based on depth-of-coverage described in the current literature, and evaluate their effectiveness using the simulated data we have generated. We discuss the strengths and weaknesses of these detection methods when integrated into the daily workflow of a genetic diagnostic laboratory.

The article ***PattRec: An easy-to-use CNV detection tool optimized for targeted NGS assays with diagnostic purposes*** (https://doi.org/10.1016/j.ygeno.2019.07.011), which was published in *Genomics* (2018 JCR IF, 3.16) is presented in Section 3.3. This article presents a new CNV detection tool called PattRec, which is optimized for targeted NGS data and based on the comparison of coverage patterns between samples. The utility of this tool is evaluated using real data, including publicly available data (from the 1000 Genomes project) and data provided by other laboratories, and its performance is compared with that of existing CNV detection tools. The parameters that influence the reproducibility of coverage patterns between samples, including GC content, biases caused by differences in sample processing, and the use of different gene panel designs, are also evaluated.

1.2 THE DIAGNOSIS OF RARE DISEASES

Rare and ultra-rare diseases are defined as those with incidence rates of less than 1 in 2,000 and 1 in 100,000 people, respectively. Despite their low incidence, the large number of rare diseases (over 7,000 are described, and this number is continually growing) means that their combined prevalence is significant. According to EURORDIS (the European Organization of Rare Diseases) 6–8% of the European population will develop a rare disease throughout their lives. These prevalence rates are similar to those reported for common diseases such as diabetes and asthma. Rare diseases therefore constitute a major problem for doctors and have significant economic implications for health systems worldwide due to the difficulty in establishing a specific etiological diagnosis. Health services are generally unprepared to deal with diseases with such low incidences and variable phenotypic expression. Many of the clinical manifestations of these diseases overlap with those of more common diseases, and symptoms can appear late, even in adulthood. Approximately half of these diseases appear during childhood. Early diagnosis is therefore essential. However, months and even years can pass between the appearance of the first clinical signs and diagnosis. According to the Spanish Federation of Rare Diseases (FEDER) the mean time required to establish a diagnosis is 5 years ("Las Enfermedades Raras en cifras" n.d.).

Given their diagnostic complexity, together with the fact that over 80% of rare diseases have an identified genetic component, these diseases stand to benefit greatly from recent advances in the field of DNA sequencing. Until just a few years ago genetic analysis was considered the final stage of the diagnostic process in patients with rare diseases. After a process that typically lasted years and involved the documentation of clinical manifestations and successive biochemical, pathological, functional, and imaging tests, patients with suspected genetic disorders were referred for analysis of a candidate gene by classical sequencing. In most cases this would produce a negative result, and another candidate gene would be sequenced. This cycle would continue, increasing the time to diagnosis and in most cases ending without establishing a definitive diagnosis. The rate of diagnosis using this methodology was very low, except for certain diseases with well-defined clinical, biochemical, or pathognomonic characteristics and with low genetic heterogeneity (*e.g.,* phenylketonuria). The emergence of next generation sequencing technology (NGS) represented a turning point in our understanding of rare diseases, and in their diagnosis and treatment (Bacchelli and Williams 2016; Danielsson et al. 2014). The emergence of NGS approximately 15 years ago heralded the potential to radically change the diagnostic process by providing a fast, powerful, and low-cost alternative for the simultaneous genetic analysis of many genes early in the diagnostic process. Within a few weeks, NGS-based tools can close in on one or a small number of candidate genes and can help establish a rapid diagnosis in a considerable percentage of cases. This new diagnostic process can dramatically reduce waiting times and shorten the often endless search that many patients and their families had to endure before the advent of this technology. It is therefore unsurprising that the world's best healthcare systems have incorporated these powerful tools into their routine diagnostic processes.

NGS has also given rise to a new phenomenon in medicine known as reverse phenotyping. In some cases, the combined use of NGS and segregation analysis can identify a pathogenic mutation in a gene that is known to cause disease but was previously linked to a different phenotype. In such cases, retrospective clinical investigation of the patient and family members may reveal additional, previously unrecognized characteristics. In a review of more than 300 studies in which rare diseases had been investigated using whole-exome sequencing (WES), Boycott et al. found that approximately 25% of the genetic mutations related to a specific disorder were associated with a phenotype that was actually observed following clinical reevaluation of the patient after a genetic finding (Boycott et al. 2013). The recent literature includes many examples of reverse phenotyping. For example, Arif and colleagues identified a variant in *OPA3* (implicated in optic atrophy syndrome) in two affected members of a family in which no ophthalmological studies had previously been conducted, leading clinicians to reassess the phenotype of the patients and to ultimately establish a correct diagnosis (Arif et al. 2013). Graziano et al. identified a variant in *DDC* (which causes aromatic amino acid decarboxylase deficiency [AADC]) in three consanguineous patients with syndromic intellectual disability, thus expanding the AADC phenotype (Graziano et al. 2015). In their study, Zhang and coworkers detected two variants in heterozygosis in *SPG7* (which is implicated in hereditary spastic paraparesis) in a patient with slowly progressing and apparently sporadic ataxia whose symptoms included “emotional disconnection”, thereby adding neurobehavioral disorders to the phenotype of this disease (L. Zhang et al. 2017). These findings show that in rare diseases the sequence in which clinical signs appear, as well as their intensity, vary greatly from one patient to another. This helps explain why it can be so difficult to establish diagnosis. Just a few years ago, doctors had no choice but to observe and wait for further clinical signs to appear over time. However, NGS now provides doctors with powerful molecular tools that can uncover important clues early in the disease process and allow them to begin investigating manifestations that are not yet fully expressed or have not yet appeared. Many rare diseases can be caused by mutations in tens or hundreds of different genes. For example, the Bonne and Rivier team annually updates a list of genes associated with myopathies, the most recent version of which contains 535 genes (Bonne, Rivier, and Hamroun 2018). Wang and colleagues proposed an exhaustive list of 693 epilepsy-related genes, and another 284 genes potentially involved in this disease (Wang et al. 2017). The ability to sequence hundreds or thousands of genes in parallel allows analysis of genes that are implicated in the disease suspected to underlie the patient's phenotype, as well as genes associated with other diseases with overlapping phenotypes, without substantially increasing the cost of the test. This translates into an increase in the rate of diagnosis of diseases in which the complete clinical picture is difficult to identify or emerges slowly over time. Thanks to NGS, the number of genes associated with newly identified diseases has grown exponentially in all fields of medicine. The increase over the last 20 years in the number of phenotype entries in the Online Mendelian Inheritance in Man (OMIM) database for which the molecular basis of a particular phenotype is known is shown in Figure 1. This explosion of knowledge is a consequence to the use of NGS to rapidly sequence any region of the human genome, ranging from several genes to the entire genome, with a high degree of sensitivity.

Three main NGS-based tests are used in the study of rare diseases. These tests can be ordered according to cost, ease of analysis, and scope, and include (1) parallel sequencing of coding sequences (exons) of gene groups in which mutations result in similar or overlapping phenotypes (gene panels); (2) whole-exome sequencing (WES), in which all known coding regions of the human genome are sequenced; and (3) whole-genome sequencing (WGS), which analyzes the entire human genome. In current clinical practice the most commonly used analysis is targeted sequencing, using either gene panels or WES (Lindy et al. 2018; Likar et al. 2018; Ortega-Moreno et al. 2017; Savarese et al. 2018). In recent years WES has predominated in studies of the genetic basis of rare diseases. This type of analysis covers only 1% (~30 Mb) of the human genome, and its main drawback is its inability to detect certain types of variants, such as those located in intronic or intergenic regions. Moreover, targeted sequencing offers much less uniform read distribution as a consequence of the enrichment of the areas to be studied, resulting in lower coverage in certain areas, especially those with high GC content (Meienberg et al. 2016). However, this type of analysis is very widely used thanks to the following features: compared with WGS, (1) the cost is much lower and (2) the amount of data generated is much more manageable, reducing both the time required for analysis and the complexity of the data obtained. Compared with WES, gene panels offer much faster response times, and fewer incidental findings (variants associated with a greater likelihood of developing a disease other than that being studied). Unfortunately, gene panels are unable to identify new disease-causing genes.

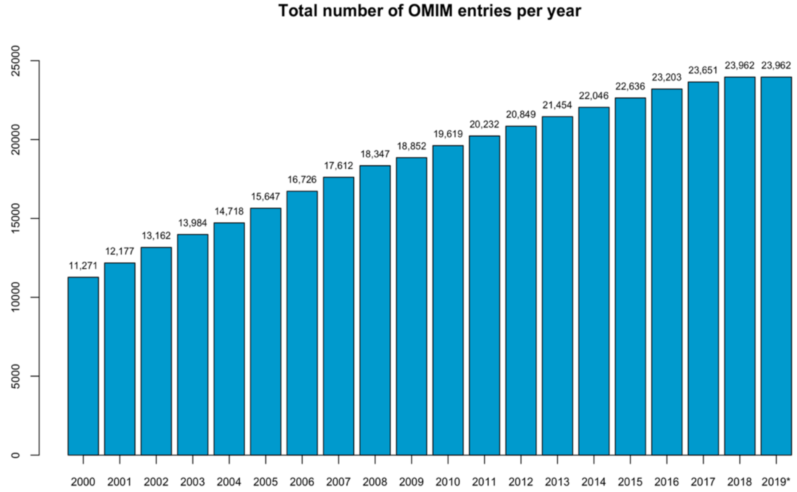


Figure 1. Total number of OMIM entries per year for the past 20 years. * data collected up to August 31st, 2019.

1.3 THE CHALLENGE OF GENOMIC DATA ANALYSIS: BIOINFORMATICS

The application of NGS to the study of diseases of genetic origin represents a tremendous step forward, but also presents a new challenge: the difficulty in filtering and interpreting the data produced. While the output of classical sequencing approaches is a single DNA sequence (with a well-defined genomic position), NGS data consists of millions of "pieces" of DNA sequence, the original position of which is not easily identifiable because all sequences from all the studied genes are mixed together. It is therefore necessary to develop specific bioinformatics programs to order the results obtained from sequencing platforms. The process of detecting variants (modifications in the nucleotide sequence with respect to the reference genome) is also less immediate than with classical sequencing. In fact, because NGS can produce errors and false positives, classical sequencing remains the reference method to confirm the presence of certain types of variants, especially in areas poorly covered by NGS.

1.3.1 Sequencing and alignment

NGS technologies randomly fragment the genome into small pieces that are amplified by PCR and subsequently "read" or sequenced in parallel. This sequencing consists of the reading of a certain number of bases of the fragment (the number of bases read usually ranges from 50 bp to 400 bp or more, depending on the platform used). In single-end sequencing, the fragment is read in only one direction, while in paired-end sequencing each fragment is read in both directions.

Among the most common sequencing methods (Salipante et al. 2014; Liu et al. 2012) are sequencing by synthesis (used by Illumina) (“Illumina | Sequencing and Array-Based Solutions for Genetic Research” n.d.), ligation sequencing (Thermofisher SOLiD) (“Life Technologies - ES” n.d.), semiconductor ion sequencing (Ion Torrent Systems Inc.) (“Ion Torrent - ES” n.d.), and pyrosequencing (created by Roche / 454 Life Sciences) (“Roche Life Science | Welcome” n.d.), although pyrosequencing has been obsolete since 2013 (“Bio-IT World” n.d.). Each of these processes produces millions of reads of 50–700 bp, which must then be aligned against the human reference genome to identify their genomic position. This alignment process is far from trivial: in addition to the difficulty in working with the massive amounts of data generated by the platform, each platform has its own intrinsic sequencing errors (error rates vary from 0.1–1%, depending on the platform) (Canzar and Salzberg 2017). In Illumina sequencing, the most frequent errors are single-nucleotide substitutions, in Ion Torrent and 454 the most common errors are small deletions, and SOLiD produces A-T biases (Fox and Reid-Bayliss 2014; Goodwin, McPherson, and McCombie 2016). Furthermore, the alignment process can be further complicated by the presence of variants (common or rare) in the sequenced sample that may cause reads in a given region to differ from the reference genome. However, the most challenging aspect of the alignment process is the presence of repetitive sequences in the reference genome, *i.e.,* pieces of DNA that are repeated (the exact same sequence or small variations thereof), even hundreds of times, at different locations within the genome. These sequences account for approximately half of all human DNA (Batzer and Deininger 2002) and pose a great challenge for aligners, particularly sequences that share a high degree of similarity. Alignment algorithms seek the best possible match between the reads and the reference genome, and generally achieve up to 80% unique alignments, since most of the repetitive sequences present in the genome differ sufficiently from one other so as not to pose a problem. However, those that share a higher percentage of similarity can result in ambiguities in the alignment data. Moreover, the potential presence of population variants in these regions further complicates the process. Aligners must therefore choose whether to discard reads that fall in these repetitive regions, prioritize better aligned reads, or report all possible alignments (assigning a penalty according to the number of mismatched bases) (Treangen and Salzberg 2011).

The algorithms used must therefore be strict enough to uniquely assign each read to its corresponding genomic position (taking into account the presence of repetitive sequences in the genome), but sufficiently permissive to be able to align reads with discrepancies relative to the reference genome. The two most used methods for alignment are (1) seed and associative matrix (seed / hash) methods, which search for matches in sub-sequences (seeds) assuming that at least one will match perfectly with the reference (hash); and (2) methods based on the Burrows-Wheeler transformation (Burrows and Wheeler, n.d.), which index the reference genome so that the search for matches is computationally much less expensive (Flicek and Birney 2009; Heng Li and Homer 2010). The best known algorithms that use the Burrows-Wheeler transformation are BWA (H. Li and Durbin 2009) and Bowtie (Langmead et al. 2009). BWA-mem (Li, Heng 2013) applies the seed / hash method to find matches between the seeds and the reference, and assigns each an alignment “suitability” value using the Smith-Waterman method (Smith and Waterman 1981). Burrows-Wheeler transformation-based methods are generally faster but less sensitive than seed / hash methods, and are more suitable for shorter read lengths (<70 bp) than seed / hash methods (which are recommended for read lengths ≥70 bp) (“Burrows-Wheeler Aligner” n.d.).

1.3.2 Variant detection

Once the millions of reads generated by the platform are aligned, the next step is to detect the variations in the sample with respect to the reference genome. While most of these variations are inherited from the parents, they can also occur *de novo* in the germ cells of the parents or at some point during embryonic development. Specific detection methodologies are used for different variant types, each of them has unique features that make them more or less easy to detect. Furthermore, not all variants are detectable using all types of sequencing.

Variants can be classified into three main groups according to the type of modification they produce in the genome and how they are detected: single-nucleotide variants and small insertions/deletions; copy number variations; and genomic rearrangements (**Figure *2***).

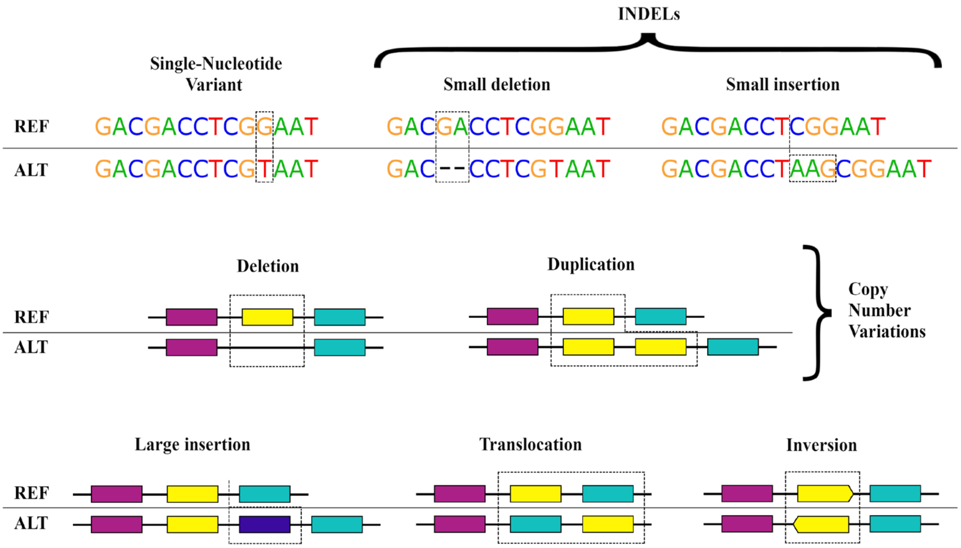


Figure 2. Types of variants according to the modification produced into the genome.

1.3.2.1 Single-nucleotide variants and small insertions/deletions

Single-nucleotide variants (SNVs) are variations in a single nucleotide in the DNA chain, while small insertions and deletions (INDELs) are defined as losses or gains of a small number of nucleotides (≤50 bp). These variants are the most common in the genome (99.9% of all variants are SNVs or INDELs) (1000 Genomes Project Consortium et al. 2015; Katsonis et al. 2014). They can be categorized as genic (those that occur within a gene) or non-genic (those located in intergenic regions). Genic variants can be subgrouped as follows:

1. Coding variants: variants that occur in coding regions of the gene (exons).
    a. SNVs: depending on the nucleotide change produced and the position within the amino acid, these variants can be classified as follows:

        i. Synonymous: the resulting amino acid remains the same (*e.g.,* the change TG**C**→TG**T** still produces a cysteine).
        ii. Missense: the nucleotide change produces a different amino acid (*e.g.,* the change TG**C**→TG**G** produces a tryptophan instead of a cysteine).
        iii. Nonsense: the nucleotide change produces a stop codon (*e.g.,* the change TG**C**→TG**A** produces a stop codon instead of a cysteine).
    b. INDELs: these are divided into two subgroups, depending on whether the number of nucleotides inserted or deleted is a multiple of three:
        i. Non-frameshift*:* if the number of inserted or deleted bases is three or a multiple of three, one or more new amino acids are generated (or deleted) but the rest of the sequence is unaffected.
        ii. Frameshift: if the number of bases is not a multiple of three, the reading frame of the gene is altered beginning at the location of the variant. In most cases this results in the appearance of a premature stop codon in which the mRNA is subject to a process known as nonsense-mediated mRNA decay (NMD), resulting in its elimination (Lin et al. 2017). In other cases, a premature stop codon is not produced and the INDEL causes a change in the amino acid sequence, resulting in a protein longer than the original encoded protein (the stop codon emerges downstream of the original codon).

2. Noncoding variants: these occur in introns (intronic variants) or in cis-regulatory regions (5 'UTR and 3' UTR).

3. Splice-site variants: these variants affect the consensus regions necessary for correct splicing of exons. Different types of consensus sequences are involved in the splicing process. The most conserved are those found on the border between an exon and an intron (splice acceptor and/or donor sites), while less conserved ones can be located both in intronic regions (branch site) or within exons: splicing enhancers (ESE) and splicing silencers (ESS) (Anna and Monika 2018).

Of these variants, those that typically most affect gene functionality are nonsense, frameshift, and splice-site variants. These usually result in premature termination of transcription, and are thus known as truncation variants (Ng et al. 2008). Variants that result in substitution of one amino acid for another (missense) can be totally harmless (common polymorphisms with no effect on the protein) or can lead to gain or loss of function of the encoded protein, with potential pathological repercussions. The latter have low frequencies and are often studied as potential causative mutations in rare diseases.

Many tools are used to detect SNVs (the most easily detected variant type). Two main approaches are used, depending on the type of variant sought: (1) those designed to detect germline variants (*i.e.,* which are inherited from parents or arise *de novo* in the parents' germ cells); and (2) those designed to detect mosaic variants (*i.e.,* which appear at some point during embryonic development and are therefore not present in all tissues) or somatic variants (which arise after birth in a specific tissue). To detect germline variants, variant callers usually apply a Bayesian approach based on the expected number of reads of the variants (50% for heterozygous variants, 100% for homozygous variants). Thus, all variants with an allelic frequency outside the ranges permitted by each tool are discarded as false positives. The most widely used tools that apply this Bayesian approach are SAMtools (Heng Li 2011), FreeBayes (Garrison and Marth 2012) and GATK (McKenna et al. 2010). Detection of somatic and mosaic variants is more complex, as they can have different allelic frequencies. The most common detection methods are based either on comparison of affected with healthy tissue (the approach most often used in cancer studies, in which tumor tissue can be compared with blood), or on the use of a set of samples from control individuals (*i.e.,* healthy individuals) to filter out common germline variants. The algorithms applied can be Bayesian-based (*e.g.,* SomaticSniper (Larson et al. 2012) and Strelka (Saunders et al. 2012)), or heuristic (*e.g.,* VarScan2 (Koboldt et al. 2012)).

The most important factors that can affect SNV detection include intrinsic errors in sequencing platforms (which can lead to false positives), the type of sample sequenced (*e.g.,* formalin-fixed paraffin embedded (FFPE) samples are usually quite degraded, and this increases the probability of false positives and even false negatives), and, above all, a lack of sufficient coverage in the region studied (Spencer, Zhang, and Pfeifer 2015).

Detection of INDELs is challenging for several reasons. First, correct alignment of the sequence reads to the reference genome is more difficult when there are either more or fewer nucleotides with respect of the reference sequence. Second, even when the reads are correctly placed, alignment at the nucleotide level is usually incorrect due to repetitive local structures, partial overlapping, or insufficient high-quality sequence flanking the INDEL. Third, while Illumina's short sequence reads have a low overall INDEL error rate, systematic INDEL errors can occur, particularly in homopolymers (Albers et al. 2011; Montgomery et al. 2013). Repetitions of all kinds complicate the mapping process, as they introduce ambiguity as regards the true position of a read, potentially reducing the sensitivity with which we can detect INDELs or other mutations. If not analyzed correctly, repetitions can also introduce false positives by suggesting the presence of artificial INDELs between repetitive elements and decreasing the specificity of variant calling. In particular, simple tandem repeats (STRs) are especially difficult genomic sequences to sequence and analyze: they have a sequencing error rate substantially higher than that of other sequences and are prone to polymerase slippage, which can artificially extend or contract the length of the repetitive element (Narzisi and Schatz 2015).

1.3.2.2 Copy number variants

In general humans carry two copies of each genomic region (one inherited from each parent). A CNV is the result of an alteration in this number, caused by losses or gains of genetic material (resulting in no copies, one copy, or three or more copies). These variants can arise as a consequence of several different mechanisms, one of which is homologous recombination during meiosis between repeated sequences of low copy numbers (LCRs), specific to the region. The type of DNA rearrangement resulting from these events is a function of the orientation of repeated sequences that serve as substrates for homologous recombination. Recombination between direct repetitions can lead to elimination and/or duplication of the genetic material located between the repetitions, while recombination between inverted repetitions results in inversion of the intermediate genomic sequence (J. R. Lupski 1998). CNVs present in the human genome cover a greater number of nucleotides and arise *de novo* more frequently than SNVs (Stankiewicz and Lupski 2010). They exert a greater influence than SNVs on human evolution and genetic diversity among individuals, and have been implicated in susceptibility to several rare diseases, including autism and schizophrenia. Locus-specific mutation rates for CNVs are in the range 10^{-4}–10^{-5} (*i.e.,* 1000–10000 times greater than the corresponding rate for SNVs) (James R. Lupski 2007).

To date, two main tools have been used to detect CNVs: comparative genomic hybridization (CGH array) and single-nucleotide polymorphism (SNP) array. These allow detection of CNVs of a minimum size of 30 kb, which are not detectable by chromosomal banding. Because the aforementioned tools cannot detect smaller CNVs (1–30 kb), their rates are likely underestimated (Redon et al. 2006). Crucially, it is very likely that CNVs within this size range play key roles in the development of certain rare diseases. For example, Poultney et al. reported that up to 7% of autism cases harbor exon deletions of 1–30 kb that potentially contribute to their disease (Poultney et al. 2013). Detection of CNVs by NGS is therefore of utmost importance, as it is the only methodology capable of detecting small CNVs. Furthermore, the ability of NGS to detect CNVs means that CNV identification can be incorporated into routine gene panels and WES analyses, which can increase the diagnostic rate up to 6% without increasing the cost of the analysis (Pfundt et al. 2017).

The methods applied to detect CNVs differ depending on whether we are working with WGS or targeted sequencing data. Four main approaches can be distinguished: (1) paired-end mapping; (2) split read mapping; (3) depth-of-coverage; and (4) *de novo* local assembly. In paired-end sequencing the DNA fragments are read at both ends, with a fixed separation between reads (known as insert size). When these reads fall near the breakpoint (*i.e.,* the beginning or end) of a deleted area of the genome, the size of the insert is larger than stipulated, while in cases of duplications both reads have a much smaller distance, and can even overlap. Methods based on this approach therefore look for reads with an insert size distinct from that expected for the detection of CNVs (Korbel et al. 2007). Split-read methods also use paired-end reads, but are based on a different principle: the objective of this approach is to detect breakpoints by looking for reads with partners that are not mapped, or only partially mapped, against the reference genome (Z. D. Zhang et al. 2011). Algorithms based on depth-of-coverage are based on the premise that the number of reads in a genomic region is proportional to the original number of copies in that region. Therefore, coverage in deleted areas is lower (reduced by approximately half if one copy of the alleles is missing and near zero when two copies are missing) than for the rest of the genome, and is higher in duplicate areas (approximately 1.5 times higher if one allele is duplicated). These three methodologies start with the reads already aligned or mapped against the reference genome. By contrast, in *de novo* assembly methods DNA fragments are reconstructed from the reads generated by the platform by assembling the reads that overlap one another. Subsequently, these assembled fragments are compared with the reference genome to identify regions with CNVs (Alkan, Coe, and Eichler 2011).

Not all detection methods used in WGS are applicable to targeted sequencing. Because CNVs usually contain both coding and noncoding regions, breakpoints generally fall outside the areas sequenced in these analyses, and therefore paired-end, split read, and *de novo* assembly methodologies are not valid. In such cases the only appropriate methodologies are those based on depth-of-coverage. However, unlike WGS the coverage is not uniformly distributed throughout the sequenced regions. Additional measures are therefore required to overcome this problem, the most common of which is to compare coverage patterns between the sample and a set of control samples sequenced under the same conditions.

1.3.2.3 Rearrangement variants

Rearrangement variants are those in which the amount of genetic material remains constant but is relocated throughout the genome. This category includes inversions (chromosomal rearrangements in which the orientation of a segment is altered), translocations (chromosomal segments that move from one genomic position to another, either within the same chromosome or in another), and large *de novo* insertions (≥50 bp). The later can be subclassified as follows, depending on the type of sequence inserted: mobile element insertions (MEIs); nuclear mitochondrial DNA insertions (NUMTs); viral element insertions (VEIs); and insertions of unspecified sequence (Kosugi et al. 2019).

Detection of these variants is based on identification of their breakpoints using techniques such as paired-end mapping, split-read mapping, and *de novo* assembly. Most of these variants are not detectable by targeted sequencing, as the breakpoints tend to lie in noncoding areas of the genome. Moreover, the rearrangement variants can be present in the form of complex rearrangements, which are composed of several of these 'canonical' variants, making their detection and identification even more difficult (Sanchis-Juan et al. 2018).

1.3.3 Variant priorization

Once the variants have been detected, the next step is to identify those most likely related to the patient's phenotype. Most of the variants detected in the human genome are not directly implicated in any disease, at least in the context of the study of rare diseases. The vast majority of variants are common (*i.e.,* are found at high frequencies in the general population): only 1–4% of genome variants are rare (*i.e.,* have a frequency in the population of less than 0.5%) (1000 Genomes Project Consortium et al. 2015). This applies not only to single-nucleotide variants and small insertions and deletions. For example, *NEB* contains a region of 8 exons (exons 82–89, 90–97, and 98–105) that is triplicated in the general population (*i.e.*, the normal copy number of such region is six) (Kiiski et al. 2016). In fact, Conrad and colleagues estimated that there are 3,797 CNVs with frequencies >5% (size >450 bp) in the European population (Redon et al. 2006). Therefore, the first step when performing variant analysis is usually to filter the common variants using multiple existing public databases, such as the 1000 Genomes Project ("1000 Genomes | A Deep Catalog of Human Genetic Variation" n.d.), the Exome Aggregation Consortium ("ExAC Browser" n.d.), or the Genome Aggregation Database ("GnomAD" n.d.).

Even if we focus solely on rare variants, their pathogenicity cannot be ensured, since many of them may have no impact, or no harmful impact, on gene expression. The results of the 1000G, gnomAD, and ExAC projects provide many examples of these types of scenarios. For example, of the 60,706 exomes analyzed in ExAC, 54% of the variants detected were singletons (*i.e.,* variants that appear only once in the entire database) (Lek et al. 2016). The variants with the greatest functional impact, in addition to CNVs and rearrangement variants, are those that modify the reading pattern of the gene and/or the amino acids it encodes. However, both synonymous variants and those located in noncoding areas may be related to the patient's phenotype. A growing number of studies associate these variants with clinical phenotypes (Sauna and Kimchi-Sarfaty 2013; Dixit, Kumar, and Mohapatra 2019; J. E. Miller et al. 2018; Sharma et al. 2019). While synonymous variants do not result in amino acid modifications, they are found in the coding areas of the gene and can lead to the appearance or disappearance of consensus sequences involved in mRNA splicing, the stability of which is consequently altered (Sauna and Kimchi-Sarfaty 2011). The same applies to noncoding gene variants: until recently intergenic DNA was known as junk DNA, but we now know that it contains sequences essential for the differential regulation of space-time gene expression (Barrett, Fletcher, and Wilton 2012).

The chromosomal position of a variant within the gene is also important. For example, 25% of cases of idiopathic dilated cardiomyopathy are caused by truncation variants in *TTN*, and yet truncation variants in this gene have also been found in about 3% of healthy individuals. The difference between deleterious and nondeleterious truncation variants is their location: the former are mainly located in the A-band of the gene, while the latter are located outside of that band (Ehsan et al. 2017). There are a variety of tools used to predict the impact of variants on gene expression, according to their position and the type of change they cause. Most of these tools are designed for missense variants (CONDEL (González-Pérez and López-Bigas 2011), MutationTaster2 (Schwarz et al. 2014), PoliPhen-2 (Adzhubei, Jordan, and Sunyaev 2013), FATHMM (Shihab et al. 2013)); splice-site variants (GeneSplicing (Pertea, Lin, and Salzberg 2001), Human Splicing Finder (Desmet et al. 2009)), and INDELs (PROVEAN (Choi and Chan 2015)).

There is general consensus regarding the classification of variants according to their potential pathogenicity. The American College of Medical Genetics (ACMG) classifies variants into five groups according to their relationship with a specific disease: pathogenic, likely pathogenic, benign, likely benign, and of uncertain significance (Richards et al. 2015). In general, in order to classify a variant as pathogenic it must be a protein-truncating variant (nonsense, frameshift, CNV, or other type of splicing variant) or a missense variant that produces an amino acid change previously associated with the disease, with a very low frequency in databases, not inherited from healthy parents, and located in a gene for which a relationship with the disease is well documented. Although there are some specific guidelines for CNV classification (Kearney et al. 2011), the majority of existing guides for variant classification are oriented towards SNVs and INDELs, as these are the most widely studied variants.

On the other hand, several studies have linked the presence of CNVs with changes in the expression of genes located within or near the CNV (Henrichsen, Chaignat, and Reymond 2009). CNVs are implicated in numerous rare diseases, including autism spectrum disorder (Kushima et al. 2018; Yingjun et al. 2017; Pinto et al. 2014), schizophrenia (Marshall et al. 2017; Sriretnakumar et al. 2019; Avramopoulos 2018), intellectual disability (Cooper et al. 2011; Gilissen et al. 2014), and several neurodevelopment diseases (Thygesen et al. 2018; Hehir-Kwa et al. 2011; Takumi and Tamada 2018). Pfundt and collaborators performed CNV analyses on 2,603 samples from patients with various diseases of genetic origin (neurodevelopmental, movement, metabolic disorders, etc.), and detected clinically relevant CNVs in 123 samples (Pfundt et al. 2017). An estimated 15% to 20% of cases of neurodevelopmental disorders, including intellectual disability and autism spectrum disorder, can be attributed to CNVs (D. T. Miller et al. 2010). In fact, analysis of CNVs by chromosomal microarray (MCA) is considered a first-line test for the clinical diagnosis of patients with intellectual disability of unknown cause (Moeschler, Shevell, and Genetics 2014). However, a growing number of studies indicate that WES and WGS are of greater diagnostic utility than CMA: a meta-analysis conducted by Clark et al. reported that the probability of establishing diagnosis using WES or WGS is up to 8.3 times higher than that with CMA, suggesting that these approaches should be considered first-line tests for the diagnosis of diseases of genetic origin (Clark et al. 2018). It should be noted that the presence of CNVs is not always associated with disease. In the genomes of healthy individuals Zarrei et al. identified 107 coding genes from which at least 85% of exons were deleted in homozygosis, suggesting that removal of these genes has no phenotypic consequences (Zarrei et al. 2015). This highlights another problem encountered in such analyses: although a variant may have a significant impact on the gene, not all genes are equally sensitive to variation. Certain genes can tolerate large variations in their structure with no pathological consequences, while in others much smaller changes can lead to disease.

So far, we have focused primarily on Mendelian diseases, in which variants in a single gene give rise to disease (also known as monogenic diseases). However, not all diseases are caused by variations in a single gene: some arise from combinations of variants in different genes (oligogenic diseases). The simplest forms of oligogenic disease are digenic diseases, of which several types are described: classic, pseudo-digenic, or combinations of two different Mendelian diseases (Deltas 2018). Classic digenic diseases are those in which the disease only manifests when the patient carries two variants in two distinct genes. Tang and colleagues described one such scenario in their study of a family of Chinese origin, two members of which had early-onset Parkinson's disease. The affected family members each carried two variants, one in *DJ-1* (also known as *PARK7*) and another in *PINK1*, while other family members who carried only one of the variants were unaffected (Tang et al. 2006). Pseudo-digenic diseases are those in which a variant in one gene produces the disease, while a variant in another gene modifies the phenotype. For example, in cystic fibrosis patients who are homozygous for the Phe508del variant in *CFTR*, the presence of a variant in *TGFβ1* modifies the disease phenotype, increasing the risk of developing severe lung disease (Drumm et al. 2005). Finally, combinations of two different Mendelian diseases caused by variants in distinct genes can also be considered digenic. In a retrospective study of 7,374 patients, Posey and coworkers reported 97 cases of combinations of two distinct diseases caused by variants in different genes. These diseases can have very different phenotypes. One such example concerned a patient carrying variants in *ARID1B and G6PD*, which cause two diseases with very different clinical characteristics: Coffin-Siris syndrome 1 and hemolytic anemia, respectively. Alternatively, the phenotypes of the two diseases can overlap, as observed in another case in which a patient carried variants in *KCNQ2* and *SCN8A*, which respectively gave rise to two types of epileptic encephalopathy: epileptic encephalopathy, early infantile, 7; and epileptic encephalopathy, early infantile, 13 (Posey et al. 2017).

Kim and collaborators have proposed a method to detect diseases with oligogenic inheritance (A. Kim et al. 2019), and have used this approach to identify genes involved in holoprosencephaly, demonstrating that the appearance of this disease is a consequence of the combined effects of multiple variants. First, the authors did not prioritize variants according to existing guidelines for the identification of pathogenic or likely pathogenic variants, as these are oriented towards Mendelian diseases and generally rule out all variants that cannot alone give rise to the disease. Secondly, they focused their analysis on all genes potentially (even remotely) related to a phenotype similar to that of disease under study, or those with expression patterns that resemble that of the disease of interest. Finally, using a large cohort of both patients and healthy controls, they looked for sets of two or more rare variants in the prioritized genes in patients (either variants inherited from each parent, or *de novo* variants) that did not appear in the same combination in the controls.

2. OBJECTIVES

2.1 MAIN OBJECTIVE

The main goal of this thesis is the development of specific tools for the analysis of data produced by targeted NGS technologies. These tools would facilitate genetic diagnosis of rare diseases, optimize the detection and prioritization of SNVs, INDELs, and CNVs, and minimize the occurrence of false negatives and false positives.

2.2 SPECIFIC OBJECTIVES

To achieve this main objective, the following specific goals were defined:

a. Identify genes with the highest and lowest tolerance to SNVs and INDELs, and implement a methodology for the prioritization of these variants based on their potential pathogenicity (3.1).

b. Evaluate currently existing algorithms and tools for CNV detection that are applicable to targeted NGS data (3.2).

c. Identify the possible causes of variability in the coverage patterns between the samples obtained by targeted sequencing analysis, and develop of a method for CNV detection based on the comparison of such coverage patterns between samples (3.3).

3. RESULTS

3.1 ARTICLE 1

Iria Roca, Ana Fernández-Marmiesse, Sofía Gouveia, Marta Segovia, and María L. Couce. 2018. "Prioritization of Variants Detected by Next Generation Sequencing According to the Mutation Tolerance and Mutational Architecture of the Corresponding Genes". *International Journal of Molecular Sciences* 19(6): E1584.
DOI: http://dx.doi.org/10.3390/ijms19061584

3.2 ARTICLE 2

Iria Roca, Lorena González-Castro, Helena Fernández, Mª Luz Couce, and Ana Fernández-Marmiesse. 2019. "Free-access copy-number variant detection tools for targeted next-generation sequencing data". *Mutation Research-Reviews in Mutation Research* 779: 114-125. DOI: https://doi.org/10.1016/j.mrrev.2019.02.005

3.3 ARTICLE 3

Iria Roca, Lorena González-Castro, Joan Maynou, Lourdes Palacios, Helena Fernández, Mª Luz Couce, and Ana Fernández-Marmiesse. 2019. "PattRec: An easy-to-use CNV detection tool optimized for targeted NGS assays with diagnostic purposes". *Genomics*.
DOI: https://doi.org/10.1016/j.ygeno.2019.07.011

4. GENERAL DISCUSSION

The analysis of NGS output data is not trivial: for each type of variant there are different limitations in terms of detection and prioritization. The easiest variants to identify are SNVs and INDELs, but prioritization of these variants remains problematic. To determine the possible pathogenicity of the variants detected, it is essential to evaluate their potential functional impact, taking into account the type of variant in question and their genomic position. This type of evaluation is common practice in the analysis of SNVs and INDELs, as evidenced by the large number of available tools for their evaluation *in silico* (CONDEL, Human Splicing Finder, MutationTaster2, PoliPhen-2, FATHMM, GeneSplicer, PROVEAN, etc.). Evaluation of the mutational tolerance of each gene is equally important but much less common. Not all genes are equally tolerant. In some the presence of a variant can be the sole cause of a disease, while in others the presence of one or more variants has no pathological consequences. In our article *Prioritization of Variants Detected by Next Generation Sequencing According to the Mutation Tolerance and Mutational Architecture of the Corresponding Genes* (3.1), we presented an approach for the evaluation of mutational tolerance based on the ratio of missense variants to the total number of missense and synonymous variants detected in each gene. In principle, missense variants have a greater impact on the gene than synonymous variants. Therefore, in genes with lower mutational tolerance the proportion of the former is expected to be smaller, since negative selection acts on these genes to restrict the perpetuation of variants with a greater functional impact. Similarly, genes with greater mutational tolerance will contain a larger proportion of missense than synonymous variants, since the presence of the former does not affect gene functionality and therefore negative selection does not occur. While knowledge of mutational tolerance is highly valuable for variant prioritization, it should be borne in mind that gene size affects the probability of randomly detecting a rare variant. Thus, although two genes may be equally tolerant of mutations, there will be a greater probability of encountering rare missense variants in the larger gene. Therefore, to analyze variants in each gene we must take into account the probability of detecting rare variants in the gene in question (based on their frequency in a control population) as well as the gene's tolerance to the presence of missense variants. Conservation of the nucleotide in which the variant is found is, in turn, of vital importance in determining its possible pathogenicity. The presence of missense variants located in very poorly conserved regions of genes should be interpreted with caution, as it is possible that their presence has no pathogenic consequences. Conversely, variants that theoretically have no deleterious effects (*e.g.,* synonymous variants) but are located in highly conserved regions should be studied in greater detail, as they may impair correct expression of the gene. Another important factor when prioritizing the variants detected is the mutational architecture of the gene in which they are found. Not all genes are equally sensitive to all types of mutations. Certain genes (*e.g., TTN, SYNE1*) can tolerate multiple rare missense variants, but undergo alterations in functionality in the presence of only one or two truncation (frameshift, nonsense) variants. In other genes (*e.g., KCNQ2*) truncation variants are not especially deleterious, yet the presence of a theoretically less harmful variant (*e.g.,* a missense variant) can give rise to a very severe phenotype. In some genes (*e.g., TCF4*) different types of variants can produce different phenotypes, ranging from mild to severe. Therefore, in-depth knowledge of the genes being analyzed, their tolerance to different types of variants, and the relationship between the genomic position of variants and the different resulting phenotypes is crucial when prioritizing and interpreting the results obtained from NGS assays. After prioritization of the variants that are most likely implicated in the patient's phenotype, it is necessary to determine whether they are inherited or have arisen *de novo* (in the case of dominant or X-linked inheritance genes). *De novo* variants have not been subject to negative selection, and are the most likely to contribute to the patient's phenotype, particularly dominantly inherited genes. Therefore, a family study of the prioritized variants (analysis of the parents and, where possible, the siblings of the index case) is essential for correct interpretation of the results obtained. In cases in which there is a previous family medical history it is essential to evaluate cosegregation of the variant with the phenotype under study.

Targeted NGS enables the detection of CNVs, in addition to SNVs and INDELs. While SNVs and INDELs are relatively easy to detect (especially SNVs), CNV detection is more complex. Multiple tools have been developed to detect this type of variant. Most are WGS-based. Those specially designed to work with targeted NGS data typically compare coverage patterns between samples to detect areas in which genetic material has been lost or gained. The main difference between the various methods lies in the methodology used to make these comparisons, and the process used to filter the results and rule out false positives. Consequently, not all CNV detection tools are equally sensitive for all variant types. Some detect deletions better than duplications, while others are more sensitive to larger variants (in terms of the number of exons covered and the number of base pairs). It is essential to identify the strengths and weaknesses of each tool in order to select the most appropriate tool for the specific analysis to be performed. Obtaining a sufficient number of samples with clearly identified CNVs in order to test a given tool poses a significant challenge. In our article *Free-access copy-number variant detection tools for targeted next-generation sequencing data* (3.2) we presented a methodology for the generation of simulated CNV-containing samples, which we then used to evaluate the most commonly used CNV detection tools. A notable problem that arises when attempting to detect variants by comparing coverage patterns is the presence of biases in coverage patterns (*e.g.,* biases generated by GC content, the presence of repetitive sequences, or the type of sequencer used). It is therefore important to choose a simulator that can reproduce this variability. However, it is impossible to fully reproduce the complexity of real samples, and therefore results obtained with artificial samples should be considered a best-case scenario (CNV detection methods will generally produce poorer results with real samples than with simulated samples). This does not mean that the results obtained with artificial samples cannot be extrapolated. These data can be used to deduce general trends. If a tool detects deletions much better than duplications, or has problems detecting small CNVs, the same effect can be expected with real samples. What cannot be assumed is that the sensitivity and specificity are comparable in both scenarios. Based on the results obtained with the simulated samples that we generated, we can reach several global conclusions. First, the best results are obtained with greater mean depth-of-coverage. This is unsurprising: the lower the coverage the greater the likelihood that the area containing the CNV is poorly covered (increasing the likelihood of false negatives). Poorly covered areas can also produce false positives. Second, in general CNVs that are larger (in terms of exon number) are easier to detect than smaller CNVs. This may be due to the fact that the lower the number of exons contained in a CNV, the greater the likelihood that they will not be detected due to background noise and high variability, among other factors. Another potential explanation is that most tools prioritize larger CNVs to reduce the likelihood of false positives, most of which are single-exon variants. Another conclusion we can draw from these comparisons is that most tools detect deletions better than duplications. Because

CNV detection is based on differences in coverage it is unsurprising that duplications (which generally consist of 3 instead of 2 copies, resulting in a ratio between the duplicated area and the normal of around 1.5) are more difficult to detect than deletions (in which there is 1 copy instead of 2, and thus the ratio between deleted and normal zone is 0.5), as the differences are more subtle. Finally, the presence of more than one CNV in the same sample (duplications and/or deletions in different genes) has varying effects depending on the tool used: in general "compound" CNVs (multiple CNVs in the same sample) are easier (or equally easy) to detect than "simple" CNVs (one CNV per sample). It should be noted that combinations of several factors can bias the results obtained. Both compound and simple CNVs can include small-and large-sized CNVs, duplications, and deletions, and therefore the results obtained may be influenced by parameters other than CNV type (compound or simple). Because none of the tools analyzed can detect all the variants in the simulated samples, and in order to increase detection sensitivity, we combined several tools in an attempt to reduce the number of false negatives obtained. To this end, we considered a region to be positive if it was detected by at least three distinct tools. The results obtained were not very encouraging: we found that to achieve maximum sensitivity it was necessary to combine at least 9 different tools, significantly increasing the computational cost of the analysis.

To address the multiple shortcomings of existing tools for CNV detection (difficulty detecting small CNVs and duplications, high numbers of false positives, etc.), we have developed a program for CNV detection based on the comparison of coverage patterns between samples. This tool is described in the article *PattRec: An easy-to-use CNV detection tool optimized for targeted NGS assays with diagnostic purposes* (3.3). PattRec is specifically designed to analyze data obtained from gene panel sequencing, and its main objective is to detect CNVs that have either arisen *de novo* or have low frequencies in the population. PattRec offers several advantages over existing tools: it is not necessary to separate samples from female and male patients when analyzing genes located on the X chromosome (due to the normalization process used); users can opt to exclude from the analysis regions containing known polymorphic CNVs; the program generates a database in which the results of the different analyses performed are stored (allowing rapid identification of regions that contain large numbers of positives and are therefore likely regions with high variability, as well as polymorphic CNVs in the population); the false positive rate is reduced by performing the same analysis on several copies of the test and control samples at slightly less than mean coverage (although this considerably increases the computational time); the default output file is in *xlsx* format with a color code to facilitate interpretation (although users can choose plain text files if they wish to use the data as input for another program); and the program features an intuitive, user-friendly graphical user interface (GUI), which allows pre-analysis adjustment of various parameters (*e.g.,* minimum coverage required for control samples in each region, percentage required to define a duplication or a deletion). To compare the performance of PattRec with existing tools we analyzed samples from public repositories (the 1000 Genomes project) as well as those provided by other laboratories in which CNVs had been identified using other methods. PattRec showed slightly greater sensitivity than other tools, and more effectively detected small (single-exon) CNVs. In analyses run using publicly available samples, all the tools performed poorly. This is because the "internal" samples and their respective controls were sequenced in the best possible conditions to reduce inter-sample variability (*i.e.,* were processed at the same time, in the same laboratory, using the same sequencing kit), which is not possible in the case of samples from public repositories. This highlights the importance of minimizing factors that can generate bias and variability between samples, since all CNV detection programs based on the comparison of coverage patterns require a high degree of similarity between samples in order to produce optimal results. Although the results obtained in analyses of samples sequenced under optimal conditions are very promising, the results should always be corroborated using orthogonal methods (*e.g.,* qPCR), since false positives due to intrinsic sequencing-related issues cannot be ruled out, especially in cases of single-exon CNVs. The greatest limitation of the PattRec method is that it was not created for use with large gene panels, and therefore it cannot be recommended for use with WES data. Because the main objective when creating the program was to achieve the highest sensitivity possible (especially for small CNVs, even single-exon variants), its application to WES data results in many false positives, since WES does not offer the same degree of stability (in terms of coverage patterns) as gene panels.

In summary, the goal of each of the three articles presented here was to optimize the diagnosis of rare diseases through the use of targeted sequencing data, providing a global methodology for the analysis of these kind of data.

5. CONCLUSIONS AND FUTURE RESEARCH

5.1 Conclusions

This thesis presents a global methodology for the analysis of NGS data to facilitate diagnosis of rare diseases. Given that gene panels are the most commonly used analytical approach for diagnostic studies of rare diseases, the main objective was to optimize the identification and filtering of all variants that can be identified using this type of sequencing. Because there are already a variety of tools designed for the detection of SNVs and INDELs that produce good results, we chose to work with existing tools and focused on optimizing the filtering process, in terms of both the variant and the gene in which it is located. Currently available tools for the detection of the other variant type found in these analyses, CNVs, are less well optimized for targeted NGS data. These tools are relatively new and, as discussed above, have various shortcomings that make them inappropriate for use in clinical diagnosis. To address this need, we have created a specific tool for the detection of CNVs in the context of gene panel analysis. This tool offers greater sensitivity, especially for the detection of small CNVs, which are more likely implicated in the development of rare diseases.

5.2 Future research

The main drawback of the methodology presented here is the applicability to clinical practice of the results obtained. Because gene panels are the most commonly used analytical tool for the study of rare diseases, the results presented here focus on variants that can be detected using gene panels. However, WES is increasingly used in routine practice in centers that study these types of diseases, and although far from widespread a growing number of centers perform WGS, which allows the identification of variants not detectable using other types of analysis.

The following are the next steps required to optimize the genetic diagnosis of rare diseases:

1. Identification of non-Mendelian (oligogenic) inheritance

Although this is theoretically possible using gene panel analysis, we believe that for correct identification of these genes it is essential to work with data produced by WES (or ideally WGS). This would enable analysis of the possible roles in rare diseases of all genes, not just those previously linked to a disease in the literature. Because very large sample sizes are required to perform this type of study the diseases that can be studied are limited. The first objective is the study of epilepsy, for which large sample sizes can be attained relatively easily. Moreover, evidence suggests a digenic or oligogenic origin for many forms of epilepsy (Hempelmann et al. 2006; Marini et al. 2004).

2. Role of mosaicism in the development of rare diseases

The studies presented here do not address the analysis of mosaic variants, owing to the difficulty detecting and subsequently confirming the presence of these variants. The presence of mosaic variants has been linked to several rare diseases in recent studies. Stosser et al. detected mosaic pathogenic variants with a frequency of 3.5% in 9 epilepsy-associated genes (Stosser et al. 2018). In their study, Cao and coworkers estimated that 1.5% of diagnoses established for approximately 12,000 samples could be attributed to mosaic variants (Cao et al.

2019). Confirmation of the presence of these variants requires analysis of affected tissue. Therefore, one of our future objectives is to optimize detection of this type of variant and to perform the necessary analyses only in cases of patients in which the presence of mosaic variants is strongly suspected.

3. Rearrangement variants and variants located in noncoding areas of the genome.

Many variants are only detectable by WGS analysis (noncoding regulatory variants and rearrangement variants). The two main problems with the application of WGS to clinical practice are its cost and difficulties associated with data analysis. The growing number of laboratories and companies employing this technique, together with the gradually decreasing cost of analysis, leaves no doubt that its use will be widespread in the not too distant future. It is therefore important to optimize the analysis of data produced using this methodology to enable simultaneous identification of SNVs, INDELs, CNVs, and rearrangement variants. Studies of neurodevelopmental disorders (Soden et al. 2014), or in early infantile epileptic encephalopathy (Ostrander et al. 2018) , among other diseases, have already reported increases in the percentage of cases diagnosed through the use of WGS. However, this increase in diagnostic rate is limited by difficulties in determining the pathogenicity of rearrangement and intronic variants (Alfares et al. 2018). It is therefore essential that we broaden our knowledge through the study of these variants. This will require the analysis of as many patients as possible, and creation of a solid pipeline for the detection of these variants.

6. REFERENCES

"1000 Genomes | A Deep Catalog of Human Genetic Variation." n.d. Accessed June 30, 2019. http://www.internationalgenome.org/.

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. 2013. "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2." Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.] 0 7 (January): Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76.

Albers, Cornelis A., Gerton Lunter, Daniel G. MacArthur, Gilean McVean, Willem H. Ouwehand, and Richard Durbin. 2011. "Dindel: Accurate Indel Calls from Short-Read Data." Genome Research 21 (6): 961–73. https://doi.org/10.1101/gr.112326.110.

Alfares, Ahmed, Taghrid Aloraini, Lamia Al Subaie, Abdulelah Alissa, Ahmed Al Qudsi, Ahmed Alahmad, Fuad Al Mutairi, et al. 2018. "Whole-Genome Sequencing Offers Additional but Limited Clinical Utility Compared with Reanalysis of Whole-Exome Sequencing." Genetics in Medicine 20 (11): 1328–33. https://doi.org/10.1038/gim.2018.41.

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. "Genome Structural Variation Discovery and Genotyping." Nature Reviews Genetics 12 (5): 363–76. https://doi.org/10.1038/nrg2958.

Anna, Abramowicz, and Gos Monika. 2018. "Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation." Journal of Applied Genetics 59 (3): 253–68. https://doi.org/10.1007/s13353-018-0444-7.

Arif, Beenish, Kishore R. Kumar, Philip Seibler, Franca Vulinovic, Amara Fatima, Susen Winkler, Gudrun Nürnberg, et al. 2013. "A Novel OPA3 Mutation Revealed by Exome Sequencing: An Example of Reverse Phenotyping." JAMA Neurology 70 (6): 783–87. https://doi.org/10.1001/jamaneurol.2013.1174.

Avramopoulos, Dimitrios. 2018. "Recent Advances in the Genetics of Schizophrenia." Molecular Neuropsychiatry 4 (1): 35–51. https://doi.org/10.1159/000488679.

Bacchelli, Chiara, and Hywel J. Williams. 2016. "Opportunities and Technical Challenges in Next-Generation Sequencing for Diagnosis of Rare Pediatric Diseases." Expert Review of Molecular Diagnostics 16 (10): 1073–82. https://doi.org/10.1080/14737159.2016.1222906.

Backenroth, Daniel, Jason Homsy, Laura R. Murillo, Joe Glessner, Edwin Lin, Martina Brueckner, Richard Lifton, Elizabeth Goldmuntz, Wendy K. Chung, and Yufeng Shen. 2014. “CANOES: Detecting Rare Copy Number Variants from Whole Exome Sequencing Data.” Nucleic Acids Research 42 (12): e97. https://doi.org/10.1093/nar/gku345.

Barrett, Lucy W., Sue Fletcher, and Steve D. Wilton. 2012. “Regulation of Eukaryotic Gene Expression by the Untranslated Gene Regions and Other Non-Coding Elements.” Cellular and Molecular Life Sciences 69 (21): 3613–34. https://doi.org/10.1007/s00018-012-0990-9.

Bartenhagen, Christoph, and Martin Dugas. 2013. “RSVSim: An R/Bioconductor Package for the Simulation of Structural Variations.” Bioinformatics (Oxford, England) 29 (13): 1679–81. https://doi.org/10.1093/bioinformatics/btt198.

Batzer, Mark A., and Prescott L. Deininger. 2002. “Alu Repeats and Human Genomic Diversity.” Nature Reviews Genetics 3 (5): 370–79. https://doi.org/10.1038/nrg798.

“Bio-IT World.” n.d. Accessed June 24, 2019. https://www.bio-itworld.com.

Bonne, Gisèle, François Rivier, and Dalil Hamroun. 2018. “The 2019 Version of the Gene Table of Neuromuscular Disorders (Nuclear Genome).” Neuromuscular Disorders 28 (12): 1031–63. https://doi.org/10.1016/j.nmd.2018.09.006.

Boycott, Kym M., Megan R. Vanstone, Dennis E. Bulman, and Alex E. MacKenzie. 2013. “Rare-Disease Genetics in the Era of next-Generation Sequencing: Discovery to Translation.” Nature Reviews. Genetics 14 (10): 681–91. https://doi.org/10.1038/nrg3555.

Burrows, M, and David J Wheeler. n.d. “A Block-Sorting Lossless Data Compression Algorithm,” 24.

“Burrows-Wheeler Aligner.” n.d. Accessed June 24, 2019. http://bio-bwa.sourceforge.net/.

Canzar, Stefan, and Steven L. Salzberg. 2017. “Short Read Mapping: An Algorithmic Tour.” Proceedings of the IEEE 105 (3): 436–58. https://doi.org/10.1109/JPROC.2015.2455551.

Cao, Ye, Mari J. Tokita, Edward S. Chen, Rajarshi Ghosh, Tiansheng Chen, Yanming Feng, Elizabeth Gorman, et al. 2019. “A Clinical Survey of Mosaic Single Nucleotide Variants in Disease-Causing Genes Detected by Exome Sequencing.” Genome Medicine 11 (1): 48. https://doi.org/10.1186/s13073-019-0658-2.

Chen, Yong, Li Zhao, Yi Wang, Ming Cao, Violet Gelowani, Mingchu Xu, Smriti A. Agrawal, et al. 2017. “SeqCNV: A Novel Method for Identification of Copy Number Variations in Targeted next-Generation Sequencing Data.” BMC Bioinformatics 18 (1): 147. https://doi.org/10.1186/s12859-017-1566-3.

Choi, Yongwook, and Agnes P. Chan. 2015. “PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels.” Bioinformatics 31 (16): 2745–47. https://doi.org/10.1093/bioinformatics/btv195.

Clark, Michelle M., Zornitza Stark, Lauge Farnaes, Tiong Y. Tan, Susan M. White, David Dimmock, and Stephen F. Kingsmore. 2018. “Meta-Analysis of the Diagnostic and Clinical Utility of Genome and Exome Sequencing and Chromosomal Microarray in Children with Suspected Genetic Diseases.” Npj Genomic Medicine 3 (1): 1–10. https://doi.org/10.1038/s41525-018-0053-8.

Cooper, Gregory M., Bradley P. Coe, Santhosh Girirajan, Jill A. Rosenfeld, Tiffany Vu, Carl Baker, Charles Williams, et al. 2011. “A Copy Number Variation Morbidity Map of Developmental Delay.” Nature Genetics 43 (9): 838. https://doi.org/10.1038/ng.909.

Danielsson, Krissi, Liew Jun Mun, Amanda Lordemann, Jimmy Mao, and Cheng-Ho Jimmy Lin. 2014. “Next-Generation Sequencing Applied to Rare Diseases Genomics.” Expert Review of Molecular Diagnostics 14 (4): 469–87. https://doi.org/10.1586/14737159.2014.904749.

Deltas, C. 2018. “Digenic Inheritance and Genetic Modifiers.” Clinical Genetics 93 (3): 429–38. https://doi.org/10.1111/cge.13150.

Desmet, François-Olivier, Dalil Hamroun, Marine Lalande, Gwenaëlle Collod-Béroud, Mireille Claustres, and Christophe Béroud. 2009. “Human Splicing Finder: An Online Bioinformatics Tool to Predict Splicing Signals.” Nucleic Acids Research 37 (9): e67–e67. https://doi.org/10.1093/nar/gkp215.

Dixit, Ritu, Ashok Kumar, and Bhagyalaxmi Mohapatra. 2019. “Implication of GATA4 Synonymous Variants in Congenital Heart Disease: A Comprehensive in-Silico Approach.” Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 813 (January): 31–38. https://doi.org/10.1016/j.mrfmmm.2018.12.002.

Drumm, Mitchell L., Michael W. Konstan, Mark D. Schluchter, Allison Handler, Rhonda Pace, Fei Zou, Maimoona Zariwala, et al. 2005. “Genetic Modifiers of Lung Disease in Cystic Fibrosis.” New England Journal of Medicine 353 (14): 1443–53. https://doi.org/10.1056/NEJMoa051469.

Ehsan, Mehroz, He Jiang, Kate L.Thomson, and Katja Gehmlich. 2017. “When Signalling Goes Wrong: Pathogenic Variants in Structural and Signalling Proteins Causing Cardiomyopathies.” Journal of Muscle Research and Cell Motility 38 (3): 303–16. https://doi.org/10.1007/s10974-017-9487-3.

“ExAC Browser.” n.d. Accessed June 30, 2019. http://exac.broadinstitute.org/.

Flicek, Paul, and Ewan Birney. 2009. “Sense from Sequence Reads: Methods for Alignment and Assembly.” Nature Methods 6 (S11): S6–12. https://doi.org/10.1038/nmeth.1376.

Fowler, Anna, Shazia Mahamdallie, Elise Ruark, Sheila Seal, Emma Ramsay, Matthew Clarke, Imran Uddin, et al. 2016. “Accurate Clinical Detection of Exon Copy Number Variants in a Targeted NGS Panel Using DECoN." Wellcome Open Research 1 (November): 20. https://doi.org/10.12688/wellcomeopenres.10069.1.

Fox, Edward J, and Kate S Reid-Bayliss. 2014. "Accuracy of Next Generation Sequencing Platforms." Journal of Next Generation Sequencing & Applications 01 (01). https://doi.org/10.4172/2469-9853.1000106.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-Based Variant Detection from Short-Read Sequencing." ArXiv:1207.3907 [q-Bio], July. http://arxiv.org/abs/1207.3907.

Gilissen, Christian, Jayne Y. Hehir-Kwa, Djie Tjwan Thung, Maartje van de Vorst, Bregje W. M. van Bon, Marjolein H. Willemsen, Michael Kwint, et al. 2014. "Genome Sequencing Identifies Major Causes of Severe Intellectual Disability." Nature 511 (7509): 344–47. https://doi.org/10.1038/nature13394.

"GnomAD." n.d. Accessed June 30, 2019. https://gnomad.broadinstitute.org/.

González-Pérez, Abel, and Nuria López-Bigas. 2011. "Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel." American Journal of Human Genetics 88 (4): 440–49. https://doi.org/10.1016/j.ajhg.2011.03.004.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." Nature Reviews Genetics 17 (6): 333–51. https://doi.org/10.1038/nrg.2016.49.

Graziano, Claudio, Anita Wischmeijer, Tommaso Pippucci, Carlo Fusco, Chiara Diquigiovanni, Margit Nõukas, Martin Sauk, et al. 2015. "Syndromic Intellectual Disability: A New Phenotype Caused by an Aromatic Amino Acid Decarboxylase Gene (DDC) Variant." Gene 559 (2): 144–48. https://doi.org/10.1016/j.gene.2015.01.026.

Hehir-Kwa, Jayne Y., Benjamín Rodríguez-Santiago, Lisenka E. Vissers, Nicole de Leeuw, Rolph Pfundt, Jan K. Buitelaar, Luis A. Pérez-Jurado, and Joris A. Veltman. 2011. "De Novo Copy Number Variants Associated with Intellectual Disability Have a Paternal Origin and Age Bias." Journal of Medical Genetics 48 (11): 776–78. https://doi.org/10.1136/jmedgenet-2011-100147.

Hempelmann, Anne, Kirsten P. Taylor, Armin Heils, Susanne Lorenz, Jean-Francois Prud'homme, Rima Nabbout, Olivier Dulac, et al. 2006. "Exploration of the Genetic Architecture of Idiopathic Generalized Epilepsies." Epilepsia 47 (10): 1682–90. https://doi.org/10.1111/j.1528-1167.2006.00677.x.

Henrichsen, Charlotte N., Evelyne Chaignat, and Alexandre Reymond. 2009. "Copy Number Variants, Diseases and Gene Expression." Human Molecular Genetics 18 (R1): R1-8. https://doi.org/10.1093/hmg/ddp011.

"Illumina | Sequencing and Array-Based Solutions for Genetic Research." n.d. Accessed June 23, 2019. https://emea.illumina.com/?langsel=/es/.

"Ion Torrent - ES." n.d. Accessed June 23, 2019. https://www.thermofisher.com/es/es/home/brands/ion-torrent.html.

Jiang, Yuchao, Derek A. Oldridge, Sharon J. Diskin, and Nancy R. Zhang. 2015. "CODEX: A Normalization and Copy Number Variation Detection Method for Whole Exome Sequencing." Nucleic Acids Research 43 (6): e39. https://doi.org/10.1093/nar/gku1363.

Johansson, Lennart F., Freerk van Dijk, Eddy N. de Boer, Krista K. van Dijk-Bos, Jan D. H. Jongbloed, Annemieke H. van der Hout, Helga Westers, et al. 2016. "CoNVaDING: Single Exon Variation Detection in Targeted NGS Data." Human Mutation 37 (5): 457–64. https://doi.org/10.1002/humu.22969.

Katsonis, Panagiotis, Amanda Koire, Stephen Joseph Wilson, Teng-Kuei Hsu, Rhonald C. Lua, Angela Dawn Wilkins, and Olivier Lichtarge. 2014. "Single Nucleotide Variations: Biological Impact and Theoretical Interpretation." Protein Science: A Publication of the Protein Society 23 (12): 1650–66. https://doi.org/10.1002/pro.2552.

Kearney, Hutton M., Erik C. Thorland, Kerry K. Brown, Fabiola Quintero-Rivera, and Sarah T. South. 2011. "American College of Medical Genetics Standards and Guidelines for Interpretation and Reporting of Postnatal Constitutional Copy Number Variants." Genetics in Medicine 13 (7): 680–85. https://doi.org/10.1097/GIM.0b013e3182217a3a.

Kiiski, Kirsi, Vilma-Lotta Lehtokari, Ari Löytynoja, Liina Ahlstén, Jenni Laitila, Carina Wallgren-Pettersson, and Katarina Pelin. 2016. "A Recurrent Copy Number Variation of the NEB Triplicate Region: Only Revealed by the Targeted Nemaline Myopathy CGH Array." European Journal of Human Genetics 24 (4): 574–80. https://doi.org/10.1038/ejhg.2015.166.

Kim, Artem, Clara Savary, Christèle Dubourg, Wilfrid Carré, Charlotte Mouden, Houda Hamdi-Rozé, Hélène Guyodo, et al. 2019. "Integrated Clinical and Omics Approach to Rare Diseases: Novel Genes and Oligogenic Inheritance in Holoprosencephaly." Brain: A Journal of Neurology 142 (1): 35–49. https://doi.org/10.1093/brain/awy290.

Kim, Sangwoo, Kyowon Jeong, and Vineet Bafna. 2013. "Wessim: A Whole-Exome Sequencing Simulator Based on in Silico Exome Capture." Bioinformatics (Oxford, England) 29 (8): 1076–77. https://doi.org/10.1093/bioinformatics/btt074.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." Genome Research 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, et al. 2007. "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome." Science 318 (5849): 420–26. https://doi.org/10.1126/science.1149504.

Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. "Comprehensive Evaluation of Structural Variation

Detection Algorithms for Whole Genome Sequencing." Genome Biology 20 (1): 117. https://doi.org/10.1186/s13059-019-1720-5.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. 2012. "Copy Number Variation Detection and Genotyping from Exome Sequence Data." Genome Research 22 (8): 1525–32. https://doi.org/10.1101/gr.138115.112.

Kushima, Itaru, Branko Aleksic, Masahiro Nakatochi, Teppei Shimamura, Takashi Okada, Yota Uno, Mako Morikawa, et al. 2018. "Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights." Cell Reports 24 (11): 2838–56. https://doi.org/10.1016/j.celrep.2018.08.022.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." Genome Biology 10 (3): R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Larson, David E., Christopher C. Harris, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. 2012. "SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data." Bioinformatics (Oxford, England) 28 (3): 311–17. https://doi.org/10.1093/bioinformatics/btr665.

"Las Enfermedades Raras en cifras." n.d. Accessed June 22, 2019. https://enfermedades-raras.org/index.php/enfermedades-raras/enfermedades-raras-en-cifras.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." Nature 536 (7616): 285–91. https://doi.org/10.1038/nature19057.

Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." Bioinformatics 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." Bioinformatics (Oxford, England) 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." ArXiv:1303.3997 [q-Bio], March. http://arxiv.org/abs/1303.3997.

Li, Heng, and Nils Homer. 2010. "A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing." Briefings in Bioinformatics 11 (5): 473–83. https://doi.org/10.1093/bib/bbq015.

Li, Jason, Richard Lupat, Kaushalya C. Amarasinghe, Ella R. Thompson, Maria A. Doyle, Georgina L. Ryland, Richard W. Tothill, Saman K. Halgamuge, Ian G. Campbell, and

Kylie L. Gorringe. 2012. “CONTRA: Copy Number Analysis for Targeted Resequencing.” Bioinformatics (Oxford, England) 28 (10): 1307–13. https://doi.org/10.1093/bioinformatics/bts146.

“Life Technologies - ES.” n.d. Accessed June 23, 2019. https://www.thermofisher.com/es/es/home.html.

Likar, Tina, Mensuda Hasanhodžić, Nataša Teran, Aleš Maver, Borut Peterlin, and Karin Writzl. 2018. “Diagnostic Outcomes of Exome Sequencing in Patients with Syndromic or Non-Syndromic Hearing Loss.” PLOS ONE 13 (1): e0188578. https://doi.org/10.1371/journal.pone.0188578.

Lin, Maoxuan, Sarah Whitmire, Jing Chen, Alvin Farrel, Xinghua Shi, and Jun-tao Guo. 2017. “Effects of Short Indels on Protein Structure and Function in Human Genomes.” Scientific Reports 7 (1): 9313. https://doi.org/10.1038/s41598-017-09287-x.

Lindy, Amanda S., Mary Beth Stosser, Elizabeth Butler, Courtney Downtain-Pickersgill, Anita Shanmugham, Kyle Retterer, Tracy Brandt, Gabriele Richard, and Dianalee A. McKnight. 2018. “Diagnostic Outcomes for Genetic Testing of 70 Genes in 8565 Patients with Epilepsy and Neurodevelopmental Disorders.” Epilepsia 59 (5): 1062–71. https://doi.org/10.1111/epi.14074.

Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. “Comparison of Next-Generation Sequencing Systems.” Journal of Biomedicine and Biotechnology 2012: 1–11. https://doi.org/10.1155/2012/251364.

Love, Michael I., Alena Myšičková, Ruping Sun, Vera Kalscheuer, Martin Vingron, and Stefan A. Haas. 2011. “Modeling Read Counts for CNV Detection in Exome Sequencing Data.” Statistical Applications in Genetics and Molecular Biology 10 (1). https://doi.org/10.2202/1544-6115.1732.

Lupski, J. R. 1998. “Genomic Disorders: Structural Features of the Genome Can Lead to DNA Rearrangements and Human Disease Traits.” Trends in Genetics: TIG 14 (10): 417–22.

Lupski, James R. 2007. “Genomic Rearrangements and Sporadic Disease.” Nature Genetics 39 (7 Suppl): S43-47. https://doi.org/10.1038/ng2084.

Marini, Carla, Ingrid E. Scheffer, Kathryn M. Crossland, Bronwyn E. Grinton, Fiona L. Phillips, Jacinta M. McMahon, Samantha J. Turner, et al. 2004. “Genetic Architecture of Idiopathic Generalized Epilepsy: Clinical Genetic Analysis of 55 Multiplex Families.” Epilepsia 45 (5): 467–78. https://doi.org/10.1111/j.0013-9580.2004.46803.x.

Marshall, Christian R., Daniel P. Howrigan, Daniele Merico, Bhooma Thiruvahindrapuram, Wenting Wu, Douglas S. Greer, Danny Antaki, et al. 2017. “Contribution of Copy Number Variants to Schizophrenia from a Genome-Wide Study of 41,321 Subjects.” Nature Genetics 49 (1): 27–35. https://doi.org/10.1038/ng.3725.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A

MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” Genome Research 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Meienberg, Janine, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. 2016. “Clinical Sequencing: Is WGS the Better WES?” Human Genetics 135 (3): 359–62. https://doi.org/10.1007/s00439-015-1631-9.

Miller, David T., Margaret P. Adam, Swaroop Aradhya, Leslie G. Biesecker, Arthur R. Brothman, Nigel P. Carter, Deanna M. Church, et al. 2010. “Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies.” The American Journal of Human Genetics 86 (5): 749–64. https://doi.org/10.1016/j.ajhg.2010.04.006.

Miller, Jason E., Manu K. Shivakumar, Shannon L. Risacher, Andrew J. Saykin, Seunggeun Lee, Kwangsik Nho, Dokyoon Kim, and for the Alzheimer’s Disease Neuroimaging Initiative (ADNI). 2018. “Codon Bias among Synonymous Rare Variants Is Associated with Alzheimer’s Disease Imaging Biomarker.” In Biocomputing 2018, 365–76. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC. https://doi.org/10.1142/9789813235533_0034.

Moeschler, John B., Michael Shevell, and Committee On Genetics. 2014. “Comprehensive Evaluation of the Child With Intellectual Disability or Global Developmental Delays.” Pediatrics 134 (3): e903–18. https://doi.org/10.1542/peds.2014-1839.

Montgomery, Stephen B., David L. Goode, Erika Kvikstad, Cornelis A. Albers, Zhengdong D. Zhang, Xinmeng Jasmine Mu, Guruprasad Ananda, et al. 2013. “The Origin, Evolution, and Functional Impact of Short Insertion-Deletion Variants Identified in 179 Human Genomes.” Genome Research 23 (5): 749–61. https://doi.org/10.1101/gr.148718.112.

Narzisi, Giuseppe, and Michael C. Schatz. 2015. “The Challenge of Small-Scale Repeats for Indel Discovery.” Frontiers in Bioengineering and Biotechnology 3 (January). https://doi.org/10.3389/fbioe.2015.00008.

Ng, Pauline C., Samuel Levy, Jiaqi Huang, Timothy B. Stockwell, Brian P. Walenz, Kelvin Li, Nelson Axelrod, Dana A. Busam, Robert L. Strausberg, and J. Craig Venter. 2008. “Genetic Variation in an Individual Human Exome.” Edited by Nicholas J. Schork. PLoS Genetics 4 (8): e1000160. https://doi.org/10.1371/journal.pgen.1000160.

Ortega-Moreno, Laura, Beatriz G. Giráldez, Victor Soto-Insuga, Rebeca Losada-Del Pozo, María Rodrigo-Moreno, Cristina Alarcón-Morcillo, Gema Sánchez-Martín, et al. 2017. “Molecular Diagnosis of Patients with Epilepsy and Developmental Delay Using a Customized Panel of Epilepsy Genes.” PloS One 12 (11): e0188978. https://doi.org/10.1371/journal.pone.0188978.

Ostrander, Betsy E. P., Russell J. Butterfield, Brent S. Pedersen, Andrew J. Farrell, Ryan M. Layer, Alistair Ward, Chase Miller, et al. 2018. “Whole-Genome Analysis for Effective Clinical Diagnosis and Gene Discovery in Early Infantile Epileptic Encephalopathy.” Npj Genomic Medicine 3 (1): 1–10. https://doi.org/10.1038/s41525-018-0061-8.

Packer, Jonathan S., Evan K. Maxwell, Colm O'Dushlaine, Alexander E. Lopez, Frederick E. Dewey, Rostislav Chernomorsky, Aris Baras, John D. Overton, Lukas Habegger, and Jeffrey G. Reid. 2016. "CLAMMS: A Scalable Algorithm for Calling Common and Rare Copy Number Variants from Exome Sequencing Data." Bioinformatics (Oxford, England) 32 (1): 133–35. https://doi.org/10.1093/bioinformatics/btv547.

Pertea, Mihaela, Xiaoying Lin, and Steven L. Salzberg. 2001. "GeneSplicer: A New Computational Method for Splice Site Prediction." Nucleic Acids Research 29 (5): 1185–90.

Petrovski, Slavé, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, and David B. Goldstein. 2013. "Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes." PLoS Genetics 9 (8): e1003709. https://doi.org/10.1371/journal.pgen.1003709.

Pfundt, Rolph, Marisol Del Rosario, Lisenka E. L. M. Vissers, Michael P. Kwint, Irene M. Janssen, Nicole de Leeuw, Helger G. Yntema, et al. 2017. "Detection of Clinically Relevant Copy-Number Variants by Exome Sequencing in a Large Cohort of Genetic Disorders." Genetics in Medicine: Official Journal of the American College of Medical Genetics 19 (6): 667–75. https://doi.org/10.1038/gim.2016.163.

Pinto, Dalila, Elsa Delaby, Daniele Merico, Mafalda Barbosa, Alison Merikangas, Lambertus Klei, Bhooma Thiruvahindrapuram, et al. 2014. "Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders." American Journal of Human Genetics 94 (5): 677–94. https://doi.org/10.1016/j.ajhg.2014.03.018.

Plagnol, Vincent, James Curtis, Michael Epstein, Kin Y. Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W. Wood, et al. 2012. "A Robust Model for Read Count Data in Exome Sequencing Experiments and Implications for Copy Number Variant Calling." Bioinformatics (Oxford, England) 28 (21): 2747–54. https://doi.org/10.1093/bioinformatics/bts526.

Posey, Jennifer E., Tamar Harel, Pengfei Liu, Jill A. Rosenfeld, Regis A. James, Zeynep H. Coban Akdemir, Magdalena Walkiewicz, et al. 2017. "Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation." New England Journal of Medicine 376 (1): 21–31. https://doi.org/10.1056/NEJMoa1516767.

Poultney, Christopher S., Arthur P. Goldberg, Elodie Drapeau, Yan Kou, Hala Harony-Nicolas, Yuji Kajiwara, Silvia De Rubeis, et al. 2013. "Identification of Small Exonic CNV from Whole-Exome Sequence Data and Application to Autism Spectrum Disorder." American Journal of Human Genetics 93 (4): 607–19. https://doi.org/10.1016/j.ajhg.2013.09.001.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." Nature 444 (7118): 444–54. https://doi.org/10.1038/nature05329.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. "Standards and Guidelines for the Interpretation of Sequence

Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." Genetics in Medicine 17 (5): 405–23. https://doi.org/10.1038/gim.2015.30.

"Roche Life Science | Welcome." n.d. Accessed June 24, 2019. https://www.lifescience.roche.com/en_es.html.

Salipante, Stephen J., Toana Kawashima, Christopher Rosenthal, Daniel R. Hoogestraat, Lisa A. Cummings, Dhruba J. Sengupta, Timothy T. Harkins, Brad T. Cookson, and Noah G. Hoffman. 2014. "Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S RRNA-Based Bacterial Community Profiling." Applied and Environmental Microbiology 80 (24): 7583–91. https://doi.org/10.1128/AEM.02206-14.

Sanchis-Juan, Alba, Jonathan Stephens, Courtney E. French, Nicholas Gleadall, Karyn Mégy, Christopher Penkett, Olga Shamardina, et al. 2018. "Complex Structural Variants in Mendelian Disorders: Identification and Breakpoint Resolution Using Short- and Long-Read Genome Sequencing." Genome Medicine 10 (1): 95. https://doi.org/10.1186/s13073-018-0606-6.

Sathirapongsasuti, Jarupon Fah, Hane Lee, Basil A. J. Horst, Georg Brunner, Alistair J. Cochran, Scott Binder, John Quackenbush, and Stanley F. Nelson. 2011. "Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV." Bioinformatics (Oxford, England) 27 (19): 2648–54. https://doi.org/10.1093/bioinformatics/btr462.

Sauna, Zuben E., and Chava Kimchi-Sarfaty. 2011. "Understanding the Contribution of Synonymous Mutations to Human Disease." Nature Reviews. Genetics 12 (10): 683–91. https://doi.org/10.1038/nrg3051.

Sauna, Zuben E., and Chava Kimchi-Sarfaty. 2013. "Synonymous Mutations as a Cause of Human Genetic Disease." In ELS. American Cancer Society. https://doi.org/10.1002/9780470015902.a0025173.

Saunders, Christopher T., Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. 2012. "Strelka: Accurate Somatic Small-Variant Calling from Sequenced Tumor-Normal Sample Pairs." Bioinformatics (Oxford, England) 28 (14): 1811–17. https://doi.org/10.1093/bioinformatics/bts271.

Savarese, Marco, Annalaura Torella, Olimpia Musumeci, Corrado Angelini, Guja Astrea, Luca Bello, Claudio Bruno, et al. 2018. "Targeted Gene Panel Screening Is an Effective Tool to Identify Undiagnosed Late Onset Pompe Disease." Neuromuscular Disorders: NMD 28 (7): 586–91. https://doi.org/10.1016/j.nmd.2018.03.011.

Schwarz, Jana Marie, David N. Cooper, Markus Schuelke, and Dominik Seelow. 2014. "MutationTaster2: Mutation Prediction for the Deep-Sequencing Age." Nature Methods 11 (4): 361–62. https://doi.org/10.1038/nmeth.2890.

Sharma, Yogita, Milad Miladi, Sandeep Dukare, Karine Boulay, Maiwen Caudron-Herger, Matthias Groß, Rolf Backofen, and Sven Diederichs. 2019. "A Pan-Cancer Analysis of Synonymous Mutations." Nature Communications 10 (1): 2569. https://doi.org/10.1038/s41467-019-10489-2.

Shihab, Hashem A, Julian Gough, David N Cooper, Peter D Stenson, Gary L A Barker, Keith J Edwards, Ian N M Day, and Tom R Gaunt. 2013. "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions Using Hidden Markov Models." Human Mutation 34 (1): 57–65. https://doi.org/10.1002/humu.22225.

Smith, T.F., and M.S. Waterman. 1981. "Identification of Common Molecular Subsequences." Journal of Molecular Biology 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Soden, Sarah E., Carol J. Saunders, Laurel K. Willig, Emily G. Farrow, Laurie D. Smith, Josh E. Petrikin, Jean-Baptiste LePichon, et al. 2014. "Effectiveness of Exome and Genome Sequencing Guided by Acuity of Illness for Diagnosis of Neurodevelopmental Disorders." Science Translational Medicine 6 (265): 265ra168. https://doi.org/10.1126/scitranslmed.3010076.

Spencer, David H., Bin Zhang, and John Pfeifer. 2015. "Chapter 8 - Single Nucleotide Variant Detection Using Next Generation Sequencing." In Clinical Genomics, edited by Shashikant Kulkarni and John Pfeifer, 109–27. Boston: Academic Press. https://doi.org/10.1016/B978-0-12-404748-8.00008-3.

Sriretnakumar, Venuja, Clement C. Zai, Syed Wasim, Brianna Barsanti-Innes, James L. Kennedy, and Joyce So. 2019. "Copy Number Variant Syndromes Are Frequent in Schizophrenia: Progressing towards a CNV-Schizophrenia Model." Schizophrenia Research 209 (July): 171–78. https://doi.org/10.1016/j.schres.2019.04.026.

Stankiewicz, Paweł, and James R. Lupski. 2010. "Structural Variation in the Human Genome and Its Role in Disease." Annual Review of Medicine 61: 437–55. https://doi.org/10.1146/annurev-med-100708-204735.

Stosser, Mary Beth, Amanda S. Lindy, Elizabeth Butler, Kyle Retterer, Caitlin M. Piccirillo-Stosser, Gabriele Richard, and Dianalee A. McKnight. 2018. "High Frequency of Mosaic Pathogenic Variants in Genes Causing Epilepsy-Related Neurodevelopmental Disorders." Genetics in Medicine: Official Journal of the American College of Medical Genetics 20 (4): 403–10. https://doi.org/10.1038/gim.2017.114.

Takumi, Toru, and Kota Tamada. 2018. "CNV Biology in Neurodevelopmental Disorders." Current Opinion in Neurobiology 48: 183–92. https://doi.org/10.1016/j.conb.2017.12.004.

Talevich, Eric, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. 2016. "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing." PLoS Computational Biology 12 (4): e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

Tang, Beisha, Hui Xiong, Ping Sun, Yuhu Zhang, Danling Wang, Zhengmao Hu, Zanhua Zhu, et al. 2006. “Association of PINK1 and DJ-1 Confers Digenic Inheritance of Early-Onset Parkinson’s Disease.” Human Molecular Genetics 15 (11): 1816–25. https://doi.org/10.1093/hmg/ddl104.

Thygesen, Johan H., Kate Wolfe, Andrew McQuillin, Marina Viñas-Jornet, Neus Baena, Nathalie Brison, Greet D’Haenens, et al. 2018. “Neurodevelopmental Risk Copy Number Variants in Adults with Intellectual Disabilities and Comorbid Psychiatric Disorders.” The British Journal of Psychiatry: The Journal of Mental Science 212 (5): 287–94. https://doi.org/10.1192/bjp.2017.65.

Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions.” Nature Reviews. Genetics 13 (1): 36–46. https://doi.org/10.1038/nrg3117.

Wang, Jie, Zhi-Jian Lin, Liu Liu, Hai-Qing Xu, Yi-Wu Shi, Yong-Hong Yi, Na He, and Wei-Ping Liao. 2017. “Epilepsy-Associated Genes.” Seizure, 25th Anniversary Issue, 44 (January): 11–20. https://doi.org/10.1016/j.seizure.2016.11.030.

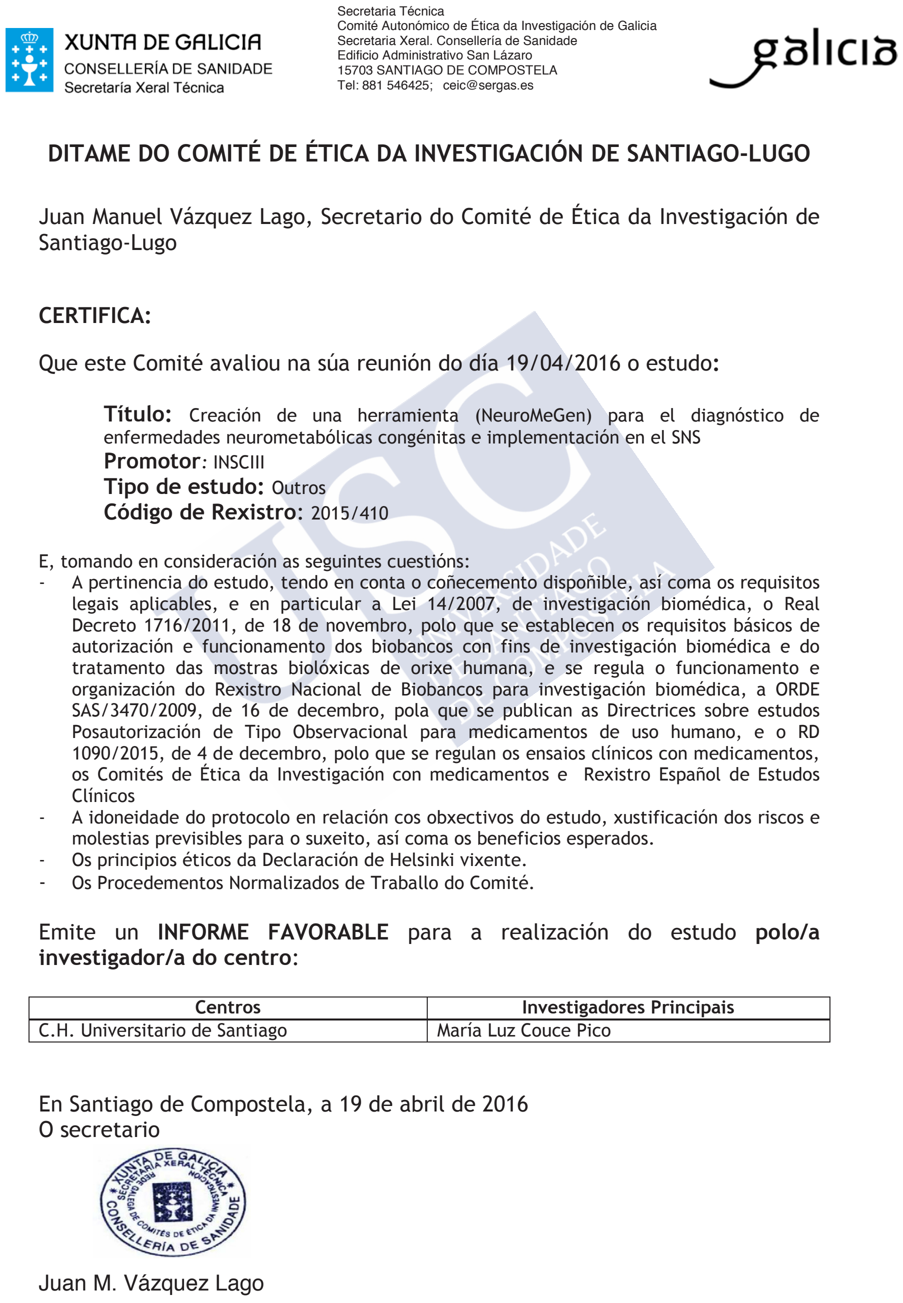
Yingjun, Xie, Yuan Haiming, Wang Mingbang, Zhong Liangying, Zhou Jiaxiu, Song Bing, Yin Qibin, and Sun Xiaofang. 2017. “Copy Number Variations Independently Induce Autism Spectrum Disorder.” Bioscience Reports 37 (4). https://doi.org/10.1042/BSR20160570.

Zarrei, Mehdi, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. 2015. “A Copy Number Variation Map of the Human Genome.” Nature Reviews. Genetics 16 (3): 172–83. https://doi.org/10.1038/nrg3871.

Zhang, Linwei, Karen N. McFarland, S. H. Subramony, Kenneth M. Heilman, and Tetsuo Ashizawa. 2017. “SPG7 and Impaired Emotional Communication.” Cerebellum (London, England) 16 (2): 595–98. https://doi.org/10.1007/s12311-016-0818-5.

Zhang, Zhengdong D, Jiang Du, Hugo Lam, Alex Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. 2011. “Identification of Genomic Indels and Structural Variations Using Split Reads.” BMC Genomics 12 (1). https://doi.org/10.1186/1471-2164-12-375.

ANNEX: Ethics Committee Approval



XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Secretaria Técnica
Comité Autonómico de Ética da Investigación de Galicia
Secretaria Xeral. Consellería de Sanidade
Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Tel: 881 546425; ceic@sergas.es

galicia

DITAME DO COMITÉ DE ÉTICA DA INVESTIGACIÓN DE SANTIAGO-LUGO

Juan Manuel Vázquez Lago, Secretario do Comité de Ética da Investigación de Santiago-Lugo

CERTIFICA:

Que este Comité avaliou na súa reunión do día 19/04/2016 o estudo:

Título: Creación de una herramienta (NeuroMeGen) para el diagnóstico de enfermedades neurometabólicas congénitas e implementación en el SNS
Promotor*:* INSCIII
Tipo de estudo: Outros
Código de Rexistro: 2015/410

E, tomando en consideración as seguintes cuestións:

- A pertinencia do estudo, tendo en conta o coñecemento dispoñible, así coma os requisitos legais aplicables, e en particular a Lei 14/2007, de investigación biomédica, o Real Decreto 1716/2011, de 18 de novembro, polo que se establecen os requisitos básicos de autorización e funcionamento dos biobancos con fins de investigación biomédica e do tratamento das mostras biolóxicas de orixe humana, e se regula o funcionamento e organización do Rexistro Nacional de Biobancos para investigación biomédica, a ORDE SAS/3470/2009, de 16 de decembro, pola que se publican as Directrices sobre estudos Posautorización de Tipo Observacional para medicamentos de uso humano, e o RD 1090/2015, de 4 de decembro, polo que se regulan os ensaios clínicos con medicamentos, os Comités de Ética da Investigación con medicamentos e Rexistro Español de Estudos Clínicos
- A idoneidade do protocolo en relación cos obxectivos do estudo, xustificación dos riscos e molestias previsibles para o suxeito, así coma os beneficios esperados.
- Os principios éticos da Declaración de Helsinki vixente.
- Os Procedementos Normalizados de Traballo do Comité.

Emite un **INFORME FAVORABLE** para a realización do estudo **polo/a investigador/a do centro**:

Centros	Investigadores Principais
C.H. Universitario de Santiago	María Luz Couce Pico

En Santiago de Compostela, a 19 de abril de 2016
O secretario

XUNTA DE GALICIA SECRETARÍA XERAL TÉCNICA REDE GALEGA DE COMITÉS DE ÉTICA DA INVESTIGACIÓN CONSELLERÍA DE SANIDADE

Juan M. Vázquez Lago