– Revista de –
MEDICINA DE LABORATORIO

www.revistamedicinadelaboratorio.es

Original

# Priorización de variantes de exoma mediante un sistema automático que emplea términos HPO

## *Prioritization of exome variants through an automatic system using HPO terms*

*José Miguel Lezana-Rosales[1], Diego Tuñón Le Poultel[1], Juan Francisco Quesada-Espinosa[1], Emma Soengas-Gonda[2], Ana Arteche-López[1], Carmen Palma-Milla[1], Irene Gómez-Manjón[1], María Isabel Álvarez-Mora[1], Rubén Pérez-de la Fuente[1], María Teresa Sánchez-Calvin[1], María José Gómez-Rodríguez[1], Marta Moreno García[1]*

[1]Hospital Universitario 12 de Octubre. Madrid. [2]Hospital Universitario Clínico San Carlos. Madrid

**Recibido:** 30/03/2020
**Aceptado:** 08/07/2020

**Correspondencia:** Diego Tuñón-Le Poultel. Hospital Universitario 12 de Octubre. Avda. de Córdoba, s/n. 28041 Madrid
e-mail: diegotunon@gmail.com

**Palabras clave:**
Genética. Secuenciación masiva. Exoma. Fenotipo. Ontología de gen.

## RESUMEN

**Introducción:** la secuenciación del exoma completo (WES) representa en la actualidad el estudio de primera elección para diagnosticar molecularmente enfermedades genéticas probablemente monogénicas con elevada heterogeneidad genética, o que solapan fenotípicamente con otras enfermedades y por ello requieren analizar multitud de genes para llegar al diagnóstico. Mediante WES se identifican miles de variantes, cuyo número final depende del tamaño de captura empleado y *pipeline* bioinformático aplicado. Esto requiere implementar técnicas de filtrado y priorización: una posibilidad es emplear paneles virtuales de genes (WES subpanelado), pero en paneles amplios aparecen multitud de variantes, lo que requiere una evaluación pormenorizada por el analista. El uso de términos de ontología de fenotipo (HPO) permite filtrar las variantes a través de asociaciones HPO-gen y establecer un sistema de priorización. El objetivo de este trabajo es el desarrollo de un sistema de priorización automática de variantes empleando términos HPO.

*Autoría: José Miguel Lezana-Rosales y Diego Tuñón-Le Poultel figuran como primeros firmantes de este artículo.*

*Conflicto de intereses: los autores declaran no tener ningún conflicto de interés.*

DOI: 10.20960/revmedlab.00077

Lezana-Rosales JM, Tuñón-Le Poultel D, Quesada-Espinosa JF, Soengas-Gonda E, Arteche-López A, Palma-Milla C, Gómez-Manjón I, Álvarez-Mora MI, Pérez-de la Fuente R, Sánchez-Calvin MT, Gómez-Rodríguez MJ, Moreno García M. Priorización de variantes de exoma mediante un sistema automático que emplea términos HPO. Rev Med Lab 2021;2(2):59-69

ISSN (electrónico): 2660-7638 - ISSN (papel): 2660-7484
Rev Med Lab 2021;2(2):59-69

**Material y métodos:** reanálisis de 33 pacientes diagnosticados previamente por WES subpanelado empleando un sistema de priorización basado en términos HPO y en las características inherentes a las variantes detectadas.

**Resultados:** tras el reanálisis se determinó que las variantes que explicaban el fenotipo clínico se encontraban en las primeras posiciones de la lista de variantes priorizadas (media: 1,43; SD: 0,87).

**Conclusión:** el sistema de priorización desarrollado permite la detección de variantes asociadas a las patologías estudiadas de forma más eficiente que por WES subpanelado, al encontrarse las variantes relacionadas con fenotipo ordenadas según su potencial patogenicidad. Este sistema representaría, por tanto, el primer abordaje en el análisis de variantes genéticas de WES.

**Keywords:**
Genetics. High-throughput nucleotide sequencing. Exome. Phenotype. Gene ontology.

## ABSTRACT

**Introduction:** whole exome sequencing (WES) currently represents the first-tier test for the diagnosis of genetic diseases that are probably monogenic with high genetic heterogeneity, or phenotypically overlapping with other diseases. Thus, they require the analysis of multiple genes to reach the diagnosis. Thousands of variants are identified by WES, the final number of which depends on the capture size, and bioinformatics pipeline used. This requires implementing filtering and prioritization techniques: one possibility is to use virtual gene panels (WES derived gene sub-panels), but in large panels many variants appear, which requires a detailed evaluation by the analyst. The use of human phenotype ontology (HPO) terms makes it possible to filter the variants through HPO-gene associations and establish a prioritization system. The objective of this work is the development of an automatic priority system of variants using HPO terms.

**Material and methods:** re-analysis of 33 patients previously diagnosed by WES derived gene sub-panels using a prioritization system based on HPO terms and the inherent characteristics of the detected variants.

**Results:** after re-analysis, it was determined that the variants that explained the clinical phenotype were in the first positions of the list of prioritized variants (mean: 1.43; SD: 0.87).

**Conclusion:** the developed prioritization system allows the detection of variants associated with the pathologies studied in a more efficient way than by WES derived gene sub-panels, since the variants related to phenotype are ordered according to their potential pathogenicity. This system would therefore represent the first approach in the analysis of genetic variants of WES.

## INTRODUCCIÓN

La secuenciación de nueva generación de ADN (*next-generation sequencing*, NGS) es una tecnología que, a diferencia de técnicas como la secuenciación Sanger, permite secuenciar en paralelo millones de fragmentos de ADN de una manera rápida y efectiva (1). La NGS permite el estudio de varios genes de manera simultánea mediante el uso de estrategias como los paneles de genes, la secuenciación de regiones codificantes e intrónicas flanqueantes de todos los genes o WES (*whole exome sequencing*) o el genoma completo (WGS). A día de hoy, hay trabajos publicados en la literatura científica que ponen de manifiesto que tanto el WES como el WGS son los estudios de primera elección en términos de coste-efectividad para el diagnóstico de enfermedades de etiología genética con fenotipos complejos o condiciones heterogéneas (2-4), aunque el estudio del WGS se circunscribe al uso en investigación en el artículo de Payne y cols. Sin embargo, la dificultad en la interpretación y el análisis del WGS hace que, generalmente, se estudie en aquellos pacientes sin diagnóstico por WES.

En el ámbito hospitalario, por tanto, la implementación del WES como técnica diagnóstica es cada día más frecuente. La ventaja del análisis de WES frente al análisis de paneles de genes cerrados radica en que puede llevarse a cabo un reanálisis bioinformático de

genes no incluidos en el estudio inicial, bien porque en el paciente han aparecido nuevas manifestaciones clínicas o bien porque posteriormente al estudio inicial otros genes se han asociado al fenotipo del paciente. Por el contrario, con el WES, la complejidad del procesamiento bioinformático aumenta, los requerimientos de espacio de almacenamiento son mayores, el análisis de los datos requiere un filtrado de variantes más eficiente y aumenta la posibilidad de encontrarse ante hallazgos incidentales.

Por tanto, identificar las variantes que explican el cuadro clínico del paciente estudiado, resulta un proceso de gran complejidad y es crítico el desarrollo de un adecuado método de filtrado y priorización de variantes.

La gran mayoría de los síndromes mendelianos humanos se han descrito en detalle en la base de datos de herencia mendeliana en el hombre (OMIM) y los sistemas jerárquicos basados en las descripciones clínicas de OMIM se han generado mediante minería de textos (proceso de análisis y derivación de información nueva a partir de textos) (5-7). Sin embargo, el análisis computacional de los datos contenidos en OMIM se encuentra limitado por la falta de un vocabulario estructurado y codificado que incluya anotaciones consistentes con relaciones bien definidas entre sí. Esta limitación motivó el desarrollo de la ontología del fenotipo humano (HPO), con el objetivo de describir las anomalías fenotípicas que caracterizan las enfermedades monogénicas humanas (8).

Esta ontología proporciona un vocabulario estandarizado de alteraciones fenotípicas en enfermedades humanas con base genética. Cada término ontológico describe una alteración (por ejemplo, el identificador *HP:0000256* se corresponde a macrocefalia). El sistema de ontología fenotípica HPO se está desarrollando actualmente utilizando la literatura médica, Orphanet, DECIPHER y OMIM. El proyecto HPO contiene actualmente más de 13.000 términos y más de 156.000 anotaciones a enfermedades hereditarias (9). Cada característica anotada puede tener como metadatos su edad típica de inicio y la frecuencia. Estos metadatos de anotación se pueden utilizar para mejorar la precisión de los algoritmos de coincidencia basados en HPO (10).

Como componente fundamental e integrador de la Iniciativa Monarch (11,12), HPO ha sido adoptado internacionalmente por numerosas organizaciones, tanto académicas como comerciales; estos incluyen el Proyecto 100.000 Genomas, el Programa y Red de Enfermedades No Diagnosticadas del National Institutes of Health, la Red Internacional de Enfermedades No Diagnosticadas (UDNI), RD-CONNECT, Solve RD y muchos otros (13,14).

La HPO ha alcanzado el estatus de recurso reconocido por el Consorcio Internacional de Investigación de Enfermedades Raras (*IRDiRC*) (15), ya que los conceptos que maneja son extremadamente valiosos para la integración, organización, búsqueda y análisis de datos, y está siendo utilizado por la Alianza Global para la Genómica y la Salud (16). Estas bases de datos no solo contribuyen al descubrimiento de genes y al diagnóstico de los pacientes incluidos en las plataformas, sino que también proporcionan datos fuente para muchos desarrollos computacionales.

Teniendo todo esto en cuenta, la integración de los términos HPO en el análisis de datos procedentes de WES permitiría detectar variantes asociadas a las patologías estudiadas de forma más eficiente frente al estudio mediante WES subpanelado. Esto se debe a que las variantes relacionadas con el fenotipo del paciente se encontrarían priorizadas en las primeras posiciones de la tabla de variantes, atendiendo a las relaciones gen-fenotipo. Es por ello que el objetivo del presente trabajo fue determinar si el sistema de priorización desarrollado en nuestro centro, que emplea términos HPO y que también tiene en cuenta las características inherentes a cada variante, es efectivo a la hora de ordenar por posible patogenicidad las variantes genéticas detectadas mediante WES, facilitando el diagnóstico genético. Se trató, además, de ver si esto representa una ventaja frente al sistema de filtrado de variantes mediante WES subpanelado.

## MATERIAL Y MÉTODOS

### Selección de pacientes

Se seleccionaron retrospectivamente 33 pacientes, en los cuales se habían identificado, mediante un estudio individualizado de WES subpanelado en el propio Servicio de Genética, variantes patogénicas y/o probablemente patogénicas asociadas a su cuadro clínico. En cuanto a la selección de los genes de los subpaneles analizados en estos pacientes se seleccionaron en base a consultas en bases de datos como OMIM, GeneReviews, DisGeNet, HGMD y PanelAPP, así como revisando la literatura científica.

Cabe decir, que el fenotipo de estos pacientes es variado, no se circunscribe a ningún grupo de patologías en concreto y son las enfermedades de herencia dominante o recesiva.

La clasificación de las variantes y diagnóstico genético se realizó por personal experto siguiendo los criterios establecidos por la American College of Medical Genetics and Genomics (ACMG) (17) . En algunos casos, y cuando fue posible, se llevó a cabo un estudio de segregación familiar para confirmar la patogenicidad de las variantes.

Todos los pacientes habían firmado el consentimiento informado para su estudio genético por parte del Servicio de Genética.

### Secuenciación

La extracción de ADN de sangre periférica en tubo EDTA de los probandos se llevó a cabo utilizando el *kit* de purificación de ADN en sangre *Maxwell* (*Pro-*

*mega*) de acuerdo con las instrucciones del fabricante. La concentración de la muestra extraída se analizó mediante fluorimetría con el aparato *Qubit 4.0 fluorometer* (*Invitrogen*). La fragmentación del ADN genómico se realizó con el sonicador *M220 Focused-ultrasonicator*. La cuantificación y validación de la biblioteca genómica se realizó utilizando un *Qubit 1X dsDNA HS Assay Kit 500rx* (*Invitrogen*) y un *Agilent High Sensitivity DNA Kit* (*Agilent Technologies, Santa Clara, CA, EEUU*). El ADN se capturó con el kit *IDT Exome Research Panel v1.0*, el cual abarca una región diana de 39 Mb que corresponde a 19,396 genes del genoma humano. El proceso de secuenciación se llevó a cabo en una plataforma *Illumina NextSeq 550* (2x75 *pb paired-end*).

### Pipeline *bioinformático* (KarMa)

El procesamiento de los datos procedentes del secuenciador se llevó a cabo a partir de un *pipeline* de desarrollo propio (*KarMa*), que cumple las especificaciones de validación establecidas por la Association for Molecular Pathology and the College of American Pathologists (AMP) (18).

En primer lugar, las secuencias en formato FASTQ de cada paciente fueron alineadas contra el genoma *hg19*, empleando una estrategia de doble alineamiento, usando *BWA-mem* (19) y *Bowtie2* (20). Se realizó también un doble genotipado de los archivos BAM resultantes con los programas *GATK* (21) y *VarDict* (22). Los archivos VCF obtenidos fueron anotados posteriormente con *Annovar* (23) y *VEP* (24), para dar lugar a un archivo tabulado con todas las variantes anotadas. Este archivo es el que posteriormente fue empleado como punto de inicio para el sistema de priorización.

Como parte de la rutina diagnóstica, y previamente al estudio mediante el sistema de priorización por HPO, se analizaron estos pacientes mediante WES subpanelado según la patología (Tabla I). La estrategia de filtrado de variantes considera los siguientes criterios que se tienen en cuenta en conjunto:

- *Frecuencia en poblaciones control* (*1000G* [25], *ExAC* [26], *GnomAD* [27]): menor del 3 %.
- *Frecuencia en base de datos propia* (*12OVar*): para eliminación de variantes artefactuales.
- *Tipo de variante y posición:* se aplica un filtrado por frecuencia más estricto dependiendo de si la variante es *missense,* sinónima, truncante, si está en una región canónica de *splicing*, etc.
- *Alertas de los genotipadores:* se eliminan las variantes que no pasan los criterios de calidad establecidos por los genotipadores.
- *Combinaciones de alineadores/genotipadores con las que se ha detectado:* se aceptan solo ciertas combinaciones, aunque se permiten otras bajo ciertas premisas.
- *Frecuencia alélica:* al menos 20 % de frecuencia del alelo alternativo, aunque se baja el umbral ante ciertos parámetros de anotación de la variante.
- *Cobertura:* se establecen puntos de corte de cobertura mínimos aceptables, bajo ciertas premisas del resto de criterios.
- *Clasificación de la variante en base de datos ClinVar y HGMD*: se tiene en cuenta, independientemente del resto de criterios.

Se eliminan del análisis final del panel las variantes que no cumplen los criterios establecidos, al no considerarse candidatas a explicar el fenotipo. Las variantes que sí superaron los filtros se clasificaron según lo establecido por la ACMG.

### Asignación de términos HPO

A cada uno de los pacientes se le asignó un mínimo de 3 términos HPO relacionados con sus características fenotípicas. Para ello se consultó la historia clínica y, mediante el buscador de términos (https://hpo.jax.org/app/), se asignaron los términos HPO correspondientes. En este buscador, los términos están relacionados de forma jerárquica, por lo que, cuando fue posible, se tomó el término que con más precisión refleja lo descrito en la historia clínica del paciente; puede verse un ejemplo de este proceso en la figura 1.

**Tabla I.**

| Paciente | Términos HPO | Panel | Genes panel | Genes HPO | Variantes filtradas |
|---|---|---|---|---|---|
| P1 | *HP:0001332, HP:0000729, HP:0002268, HP:0001250* | *Epilepsia Benigna del Lactante* | 7 | 1718 | 1 |
| P2 | *HP:0006863, HP:0000965, HP:0000154, HP:0000047, HP:0005109, HP:0000729* | *Trastornos del Espectro Autista* | 293 | 787 | 22 |
| P3 | *HP:0001644, HP:0025169, HP:0001653, HP:0005180, HP:0001659* | *Miocardiopatía Ampliado* | 121 | 272 | 12 |
| P4 | *HP:0006554, HP:0003256, HP:0001396, HP:0001511* | *Fallo Hepático* | 9 | 718 | 4 |
| P5 | *HP:0003236, HP:0007340, HP:0005109* | *Miopatías General* | 205 | 348 | 23 |

*(Continúa en la página siguiente)*

**Tabla I (cont.)**

| Paciente | Términos HPO | Panel | Genes panel | Genes HPO | Variantes filtradas |
|---|---|---|---|---|---|
| P6 | *HP:0001249, HP:0000252, HP:0004322, HP:0000708* | *Discapacidad Intelectual* | 945 | 2548 | 57 |
| P7 | *HP:0002612, HP:0000113, HP:0001409, HP:0001971, HP:0002040* | *Nefronoptisis y Multiquistosis Renal* | 45 | 136 | 2 |
| P8 | *HP:0025356, HP:0000729, HP:0001273, HP:0000256, HP:0006482, HP:0001263* | *Discapacidad Intelectual* | 945 | 1816 | 51 |
| P9 | *HP:0008689, HP:0004482, HP:0000325, HP:0000369, HP:0004322, HP:0011994, HP:0001642, HP:0001627* | *Rasopatías* | 28 | 1910 | 1 |
| P10 | *HP:0001250, HP:0001270, HP:0100543, HP:0009729, HP:0002463, HP:0009720* | *Esclerosis Tuberosa* | 3 | 1925 | 1 |
| P11 | *HP:0000044, HP:0000786, HP:0000938, HP:0010311* | *Hipogonadismo Hipogonadotropo* | 25 | 342 | 1 |
| P12 | *HP:0025615, HP:0009789, HP:0001369, HP:0002251* | *Enfermedad de Hirschpung* | 12 | 318 | 2 |
| P13 | *HP:0001263, HP:0001212, HP:0009882, HP:0001388, HP:0004322* | *Síndrome de Ehlers-Danlos e Hiperlaxitud* | 54 | 1965 | 4 |
| P14 | *HP:0001263, HP:0004322, HP:0004325, HP:0001642, HP:0001212, HP:0000378, HP:0009909* | *Discapacidad Intelectual* | 945 | 2264 | 136 |
| P15 | *HP:0003236, HP:0012548, HP:0001397, HP:0003198* | *Miopatías General* | 205 | 449 | 32 |
| P16 | *HP:0001670, HP:0001279, HP:0004904, HP:0001639* | *Miocardiopatía Hipertrófica* | 118 | 322 | 9 |
| P17 | *HP:0009062, HP:0002779, HP:0000105, HP:0000081, HP:0001723, HP:0002652* | *Discapacidad Intelectual* | 945 | 249 | 120 |
| P18 | *HP:0000729, HP:0001249, HP:0002463, HP:0001250, HP:0000256* | *Trastornos del Espectro Autista* | 293 | 2223 | 33 |
| P19 | *HP:0011344, HP:0001250, HP:0002490, HP:0001510, HP:0000252, HP:0001290* | *Discapacidad Intelectual Ampliado* | 1944 | 2475 | 251 |
| P20 | *HP:0000252, HP:0001320, HP:0001263, HP:0000243, HP:0000308, HP:0001770, HP:0012758, HP:0001250* | *Discapacidad Intelectual Ampliado* | 1944 | 2336 | 347 |
| P21 | *HP:0001596, HP:0001007, HP:0000141, HP:0002514, HP:0001263, HP:0001250, HP:0000107, HP:0002751* | *Esclerosis Tuberosa* | 3 | 2273 | 1 |
| P22 | *HP:0001634, HP:0001083, HP:0000767, HP:0000098, HP:0001488* | *Marfan y Aortopatías* | 34 | 456 | 7 |
| P23 | *HP:0004440, HP:0011315, HP:0011316, HP:0001357, HP:0030867* | *Craneosinostosis* | 56 | 86 | 13 |
| P24 | *HP:0001071, HP:0001653, HP:0001714, HP:0001712, HP:0001667* | *Miocardiopatía Hipertrófica* | 118 | 198 | 10 |
| P25 | *HP:0012265, HP:0002110, HP:0012050, HP:0002878, HP:0001935* | *Bronquiectasias y Discinesia Ciliar Primaria* | 67 | 293 | 47 |
| P26 | *HP:0003530, HP:0001332, HP:0007105* | *Aciduria Glutárica* | 5 | 409 | 3 |
| P27 | *HP:0001263, HP:0004322, HP:0000243, HP:0000347* | *Síndrome de Rubinstein-Taybi y Diagnóstico Diferencial* | 8 | 2016 | 2 |

*(Continúa en la página siguiente)*

**Tabla I (cont.)**

| Paciente | Términos HPO | Panel | Genes panel | Genes HPO | Variantes filtradas |
|---|---|---|---|---|---|
| P28 | *HP:0002474, HP:0000174, HP:0000290, HP:0000272, HP:0000689, HP:0008936, HP:0009062* | *Discapacidad Intelectual* | 945 | 1380 | 79 |
| P29 | *HP:0002123, HP:0001249, HP:0000708* | *Encefalopatía Epiléptica Ampliado* | 187 | 2160 | 15 |
| P30 | *HP:0001875, HP:0001935, HP:0001873, HP:0002037* | *Neutropenia Congénita y Vasculitis* | 60 | 404 | 6 |
| P31 | *HP:0000252, HP:0001250, HP:0001263, HP:0008936, HP:0002463, HP:0001999, HP:0000286, HP:0000494* | *Trastorno del Espectro Autista Sindrómico* | 181 | 2354 | 16 |
| P32 | *HP:0004322, HP:0000534, HP:0001263* | *Rasopatías** | 28 | 2002 | 1 |
| P33 | *HP:0003128, HP:0003348, HP:0000816* | *Acidosis Láctica* | 110 | 184 | 12 |

Términos HPO asignados a cada paciente tras evaluación de la historia clínica. Paneles aplicados a cada paciente previamente al reanálisis y número de genes evaluados en cada uno de ellos.

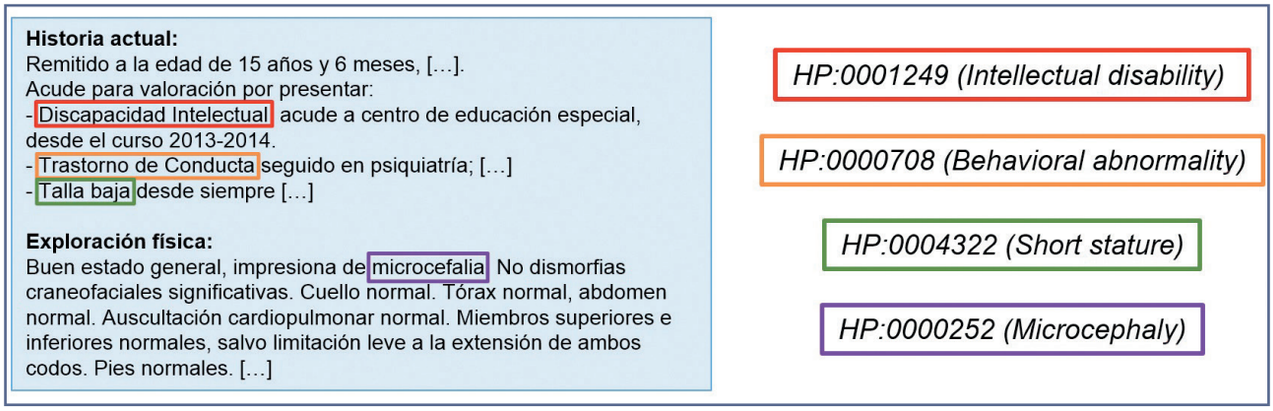


**Figura 1 –** *Fragmento de história clínica. Se seleccionan aquellos rasgos clínicos que tienen su traducción en forma de término de la base de datos HPO.*

## Desarrollo del sistema de priorización

Inicialmente, se descargó el archivo de relación término HPO-gen(es) del repositorio de HPO (http://compbio.charite.de/jenkins/job/hpo.annotations/lastSuccessfulBuild/artifact/util/annotation/phenotype_to_genes.txt). Se identificaron 4295 genes con asociación conocida con algún término HPO. Se descargó también del repositorio el archivo que relaciona jerárquicamente entre sí los términos HPO (https://raw.githubusercontent.com/obophenotype/human-phenotype-ontology/master/hp.obo). Cabe destacar que en este segundo archivo no se detallan todas las relaciones jerárquicas entre términos, ya que cada término solo se relaciona con su inmediatamente superior. Teniendo en cuenta lo anterior, se diseñó el árbol jerárquico de todas las relaciones taxonómicas de todos los términos HPO. El resultado fue un archivo de texto plano como el que se ve en la figura 2. El término más a la derecha es más específico, y a su izquierda están los términos padres menos específicos.

Adicionalmente, se adjudicó una puntuación a cada variante identificada, teniendo en cuenta las siguientes consideraciones:

1. *Número de los términos HPO utilizados que están relacionados con el gen en el que se localiza la variante* (Fig. 3): los términos más específicos se ponderaron más en la puntuación que los más generales.
2. *Tipo de variante:* mayor puntuación para las variantes radicales y en los sitios canónicos de *splicing*, seguido de las variantes de cambio de sentido y, por último, las sinónimas e intrónicas profundas.
3. *Predicciones de patogenicidad:* mayor puntuación cuanto mayor es el número de predictores de patogenicidad *in silico* que apoyan un efecto deletéreo de la variante genética.
4. *Frecuencia global de la variante en la base de datos poblacional ExAC (26):* más puntuación a las variantes menos frecuentes en esta base de datos.

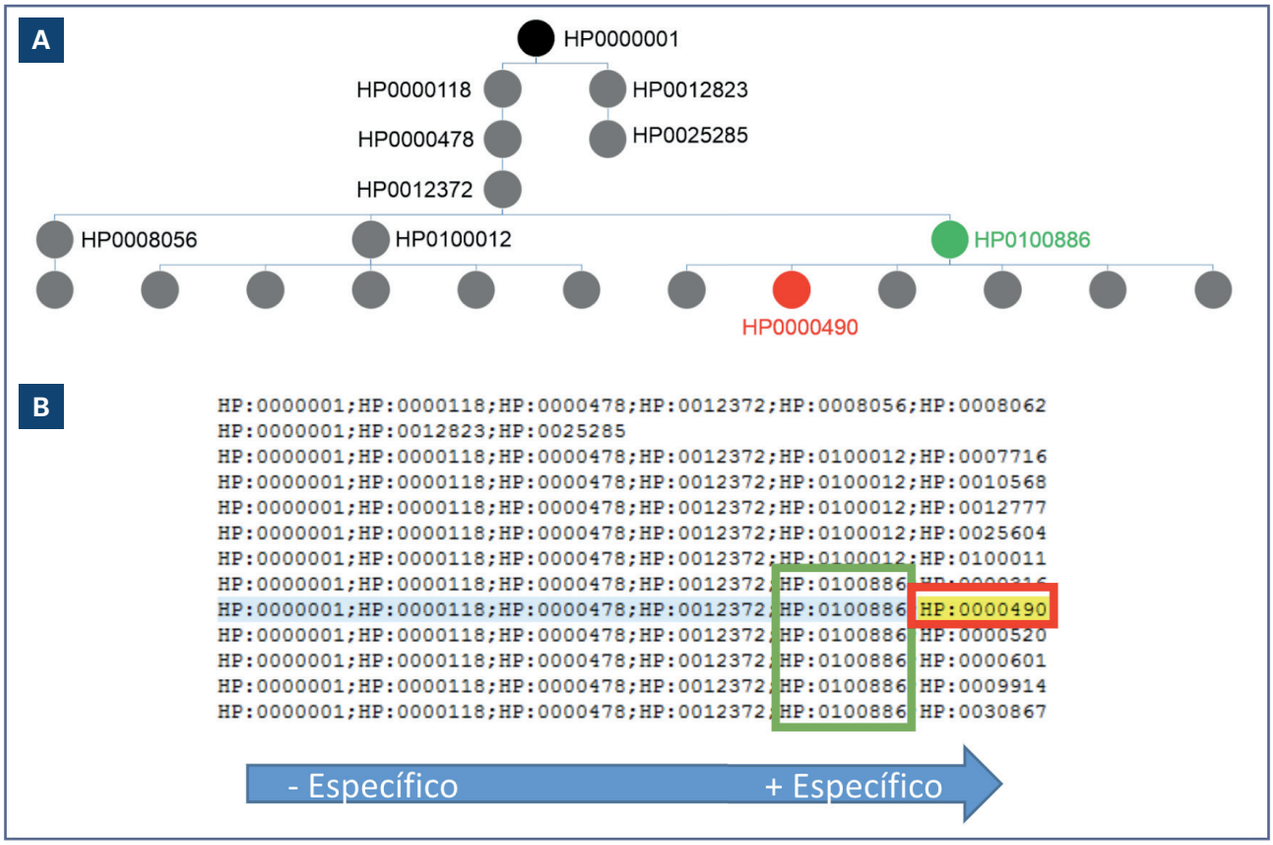


**Figura 2 –** *A. Ejemplo de relación jerárquica de términos HPO inferida a partir del archivo original de relaciones. B. Fragmento de texto plano del que deriva la recreación taxonómica de la figura superior.*

| | | | |
|---|---|---|---|
| HP:0001814 | Deep-set nails | 3265 | HRAS |
| HP:0001814 | Deep-set nails | 2146 | EZH2 |
| HP:0001814 | Deep-set nails | 64324 | NSD1 |
| HP:0001814 | Deep-set nails | 8726 | EED |
| HP:0001814 | Deep-set nails | 23512 | SUZ12 |
| HP:0000490 | Deeply set eye | 1281 | COL3A1 |
| HP:0000490 | Deeply set eye | 84992 | PIGY |
| HP:0000490 | Deeply set eye | 6913 | TBX15 |
| HP:0000490 | Deeply set eye | 2563 | GABRD |
| HP:0000490 | Deeply set eye | 91147 | TMEM67 |
| HP:0000490 | Deeply set eye | 2314 | FLII |
| HP:0000490 | Deeply set eye | 4621 | MYH3 |
| HP:0000490 | Deeply set eye | 6925 | TCF4 |

**Figura 3 –** *Extracto de relación término HPO/genes. Cada término HPO (primera columna) que refiere a una característica fenotípica (segunda columna) se asocia con ciertos genes (cuarta columna). Los números en la tercera columna corresponden al identificador de los genes.*

Teniendo en cuenta lo anterior, se determinaron 4 puntuaciones diferentes:

- *Puntuación absoluta:* se generó solo teniendo en cuenta el porcentaje de términos HPO que se relaciona con cada gen asociado a la variante.
- *Puntuación absoluta filtrada:* como la anterior, pero eliminando variantes artefactuales teniendo en cuenta su aparición reiterada, sin corresponder a posiciones polimórficas, en las diferentes muestras. Para ello se empleó la base de datos de variantes *12OVar*.
- *Puntuación combinada:* se generó teniendo en cuenta el porcentaje de términos HPO (dando más peso si los términos son más específicos), el tipo de variante, los predictores y la frecuencia global de la variante en ExAC.
- *Puntuación combinada filtrada:* como el anterior, pero eliminando variantes artefactuales.

## RESULTADOS

### Paneles empleados en el análisis previo

Los paneles virtuales para el diagnóstico de los pacientes, previo a la priorización por HPO incluyeron 329,94 genes como promedio (SD = 525,27). El número de variantes promedio por panel tras la aplicación del filtrado de *KarMa*, sin aplicar priorización, fue de 40,03 (SD = 75,44). Los paneles empleados y el número de variantes consideradas para el análisis se detallan en la tabla I.

## Términos HPO por paciente

Tras la revisión de la historia clínica, se asignaron 5,06 términos HPO de media por paciente (SD = 1,56). Los términos empleados en cada paciente se detallan en la tabla I.

## Orden de las variantes causales en cada paciente

Las variantes causales identificadas en el análisis previo de los pacientes y su clasificación se ilustran en la tabla II.

Como parte de este reanálisis, previamente a la priorización, se filtraron las variantes de potencial bajo impacto considerando diferentes criterios explicados previamente en la sección de materiales y métodos. A continuación, se realizó la priorización según lo detallado en el apartado del desarrollo del sistema de priorización en material y métodos. Las posiciones alcanzadas por las variantes tras la priorización fue la siguiente:

- *Score* absoluto: media = 7,81; SD = 11,86
- *Score* absoluto filtrado: media = 4,47; SD = 7,47
- *Score* combinado: media = 3,125; SD = 5,42
- *Score* combinado filtrado: media = 1,43; SD = 0,87

En todos los pacientes, salvo un paciente (P3), la posición de la variante causal ocupaba en todos los casos las primeras posiciones de la lista (Tabla II). El caso del P3 es debido a que el gen cuya variante explicaba el fenotipo no estaba presente en la lista de asociación original de término HPO-gen. De manera gráfica, los resultados anteriores se pueden representar en forma de diagrama de caja (Fig. 4).

**Tabla II.**

| Paciente | Gen | Variante(s) causal(es) | dbSNP | Clasificación según ACMG | *Score* absoluto | *Score* absoluto filtrado | *Score* combinado | *Score* combinado filtrado |
|---|---|---|---|---|---|---|---|---|
| P1 | *PRRT2* | NM_145239.2:c.649dupC; p.(Arg217Profs*8) | rs587778771 | P | *1* | 1 | 1 | 1 |
| P2 | *MED13L* | NM_015335.4:c.2110C>T; p.(Q704*) | - | P | *2* | 2 | 2 | 2 |
| P3 | *EMD* | NM_000117,2:c.77T>C; p.(V26A) | rs727505029 | P | - | - | - | - |
| P4 | *LARS1* | NM_020117.10:c.1237C>T; p.(R413*) / NM_020117.10:c.1292T>A; p.(V431D) | rs773819736 / rs150429680 | P / P | *4* | 1 | 4 | 1 |
| P5 | *FHL1* | NM_001159702.2:c.486C>G; p.(C162W) | rs377693754 | PP | *1* | 1 | 1 | 1 |
| P6 | *KMT2A* | NM_001197104.1:c.3461G>A; p.(R1154Q) | - | PP | *4* | 4 | 1 | 1 |
| P7 | *PKHD1* | NM_138694.3:c.5485C>T; p.(Q1829*) / NM_138694.3:NM_138694.3: c.10909C>T; p.(p.R3637C) | rs774759689 / rs141349745 | P / VSCI | *1 y 2* | 1 y 2 | 1 | 1 |
| P8 | *H1-4* | NM_005321.2:c.446_447insT; p.K149Nfs*46 | - | P | *25* | 3 | 20 | 2 |
| P9 | *PTPN11* | NM_002834.4:c.922A>G; p.N308D | rs28933386 | P | *1* | 1 | 1 | 1 |
| P10 | *TSC2* | NM_000548.4:c.1831C>G; p.R611G | rs45469298 | P | *2* | 2 | 1 | 1 |
| P11 | *FGFR1* | NM_023110.2:c.830G>T; p.(C277F) | - | P | *1* | 1 | 1 | 1 |
| P12 | *RET* | NM_020975.4:c.1711G>A; p.D571N | rs750958377 | PP | *25* | 22 | 6 | 3 |
| P13 | *ARID1B* | NM_020732.3:c.2248C>T; p.R750* | rs797045272 | P | *1* | 1 | 1 | 1 |

*(Continúa en la página siguiente)*

**Tabla II (cont.)**

| Paciente | Gen | Variante(s) causal(es) | dbSNP | Clasificación según ACMG | *Score* absoluto | *Score* absoluto filtrado | *Score* combinado | *Score* combinado filtrado |
|---|---|---|---|---|---|---|---|---|
| P14 | *ZEB2* | NM_014795.3:c.1027C>T; p.R343* | rs786204815 | P | *3* | 1 | 2 | 1 |
| P15 | *RYR1* | NM_000540.2:c.7038_7040del; p.(E2348del) | rs121918596 | P | *4* | 2 | 3 | 2 |
| P16 | *MYBPC3* | NM_000256.3:c.2308+1G>A | rs112738974 | P | *1* | 1 | 1 | 1 |
| P17 | *ACTB* | NM_001101.3:c.269_271del; p.(F90del) | - | P | *8* | 4 | 6 | 3 |
| P18 | *KMT2E* | NM_018682.3:c.71+1G>T | - | P | *31* | 3 | 24 | 2 |
| P19 | *NEXMIF* | NM_001008537.2:c.2692C>T; p.(Q898*) | - | P | *6* | 1 | 4 | 1 |
| P20 | *EFTUD2* | NM_004247.3:c.2405dupA; p.(H802Qfs*83) | - | P | *2* | 2 | 2 | 2 |
| P21 | *TSC2* | NM_000548.4:c.4977_4978insA; p.(G1660Rfs*44) | - | PP | *2* | 1 | 2 | 1 |
| P22 | *FBN1* | NM_000138.4:c.364C>T; p.(R122C) | rs137854467 | P | *1* | 1 | 1 | 1 |
| P23 | *FGFR3* | NM_001163213.1:c.749C>G; p.(P250R) | rs4647924 | P | *1* | 1 | 1 | 1 |
| P24 | *GLA* | NM_000169.2:c.713G>A; p.(S238N) | rs730880450 | P | *1* | 1 | 1 | 1 |
| P25 | *TAP2* | NM_001290043.1:c.404_405del: p.K135fs / NM_018833.2:c.404_405del: p.K135Sfs*31 | . | P | *42* | 24 | 2 | 1 |
| P26 | *GCDH* | NM_000159.3:c.395G>A; p.(R132Q) / NM_000159.3:c.1204C>T; p.(R402W) | rs200639270 / rs121434369 | P / P | *1 y 2* | 1 y 2 | 1 | 1 |
| P27 | *CREBBP* | NM_004380.2:c.3610-2A>G | - | P | *38* | 2 | 30 | 1 |
| P28 | *ANKRD11* | NM_001256183.1:c.2398_2401del; p.E800Nfs*61 | rs797045027 | P | *14* | 3 | 10 | 2 |
| P29 | *KCNB1* | NM_004975.3:c.1041C>G; p.(S347R) | - | P | *1* | 1 | 1 | 1 |
| P30 | *SRP54* | NM_003136.3:c.342_344del; p.(T117del) | - | P | *1* | 1 | 1 | 1 |
| P31 | *FOXG1* | NM_005249.4:c.553A>G; p.S185G | - | PP | *23* | 8 | 18 | 5 |
| P32* | *DPH1* | NM_001383.4:c.229+1G>A / NM_001383.4:c.15_29del; p.Met6_Val10del | - | P / VSCI | *1 y 2* | 1 y 2 | 1 | 1 |
| P33 | *PDHA1* | NM_001173454.1:c.620C>T; p.Ala207Val | rs863224150 | P | *1* | 1 | 1 | 1 |

Variante(s) causal(es) detectada(s) en los pacientes previamente a la reevaluación con sistema de priorización y posición según los diferentes *scores*. Las variantes se clasificaron según los criterios establecidos por la ACMG. P: patogénica; PP: probablemente patogénica; VSCI: variante de significado clínico incierto. *En el paciente P32 la variante causal se detectó en un gen que no estaba incluido en el panel.

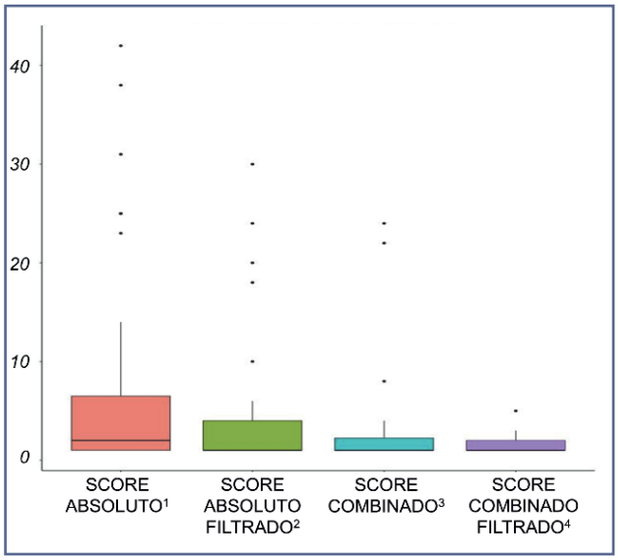


**Figura 4 –** *Distribución de posiciones de las variantes causales tras aplicar priorizaciones según los diferentes* scores.

## DISCUSIÓN

En este trabajo se ha empleado un sistema de priorización para ordenar las variantes generadas en estudios de exoma, y se presentan unos resultados prometedores, ya que en el 69,69 % de los casos la variante o variantes causales estaban en la primera posición de la lista de variantes y en el 93,93 % de los casos estaban dentro de las tres primeras posiciones. Esto, como se discutirá más adelante, supone una ventaja a la hora de abordar los estudios genéticos de WES.

En el contexto del estudio que se presenta, hay trabajos como los presentados por Jezela-Stanek y cols. (28), en el cual se emplea una estrategia basada en términos HPO, obteniendo rendimientos diagnósticos del 38 %, pero no emplean una comparación respecto a un abordaje de WES subpanelado ni cotejan cómo quedan las variantes causales en una lista priorizada, como sí se hace en el presente estudio. También hay iniciativas como Exomiser (29) que generan excelentes resultados en la priorización de variantes: en una evaluación independiente con 134 exomas (30) los resultados obtenidos con la configuración más favorable fueron algo inferiores (media = 2,1; SD = 5) a los presentados en este trabajo (media = 1,43; SD = 0,87).

En cualquier caso, el estudio del WES sin ningún tipo de filtrado por genes candidatos (ya sea por paneles WES subpanelado o priorizando variantes empleando términos HPO) y solo atendiendo a tipología de las variantes (tipo de cambio, frecuencias poblacionales, predictores de patogenicidad, etc.) es más complejo y requiere mayor tiempo de evaluación. Cabe indicar además, en relación con el estudio mediante WES subpanelado, que este implica tener que actualizar los paneles con relativa frecuencia y, si los paneles son muy extensos, el número de variantes a evaluar crece dificultando el análisis. Por tanto, a nivel práctico, resulta más sencillo actualizar periódicamente los archivos de relaciones gen-término HPO y jerarquía de HPO, que actualizar cientos de paneles.

Una ventaja sustancial de este sistema de priorización es el ahorro de tiempo de análisis por parte del personal experto en diagnóstico genético, ya que evaluar las primeras variantes, en la mayoría de los casos, sería suficiente para diferenciar aquellas variantes que son verdaderamente candidatas a las que no lo son. Sin embargo, hay que tener en cuenta una limitación que se ha puesto de manifiesto en el trabajo, la cual ocurre cuando las relaciones gen-HPO no están explicitadas en la lista de asociación original de término HPO-gen (paciente P3). Es por ello que, pese a que este sistema se demuestra eficaz a la hora de establecer el diagnóstico molecular de los pacientes, como se ha podido comprobar, ha de emplearse una segunda comprobación de las variantes mediante WES subpanelado cuando el resultado no sea concluyente. Por el contrario, en pacientes como el P32, el diagnóstico molecular no se habría alcanzado de no ser por el empleo del sistema de priorización, ya que originalmente el gen donde se encontraba la variante causal afectaba a un gen no incluido en el panel.

Cabe señalar que la cohorte de pacientes empleada en el trabajo es limitada en cuanto a su número, aunque los resultados obtenidos apoyan la eficacia del sistema de priorización desarrollado, incluso ante estudios de etiología genética diversa. En un futuro, cuando se disponga de un número mayor de pacientes, se podrían refinar los parámetros empleados para puntuar las variantes y establecer a partir de qué número de términos HPO se obtienen mejores resultados de priorización. Al respecto de los términos HPO dependería, en todo caso, de la especificidad del propio término y cuán complejo es el fenotipo de cada paciente. Hay que destacar que este sistema de priorización podría solventar este problema en cuanto a que se aplica una ponderación en la puntuación teniendo en cuenta si el término es más o menos generalista.

Con vistas a la mayor automatización, es de sumo interés el desarrollo de un "lector automático" de historias clínicas que extraigan todos los términos HPO posibles. Como alternativa a esto, podría ser una buena praxis comenzar a codificar las historias clínicas con estos términos HPO, ya que estandarizar el fenotipo de los pacientes mediante el uso de estos términos permite su explotación para múltiples fines, más allá del empleado en este trabajo.

En conclusión, hemos desarrollado un sistema automático de priorización de variantes que emplea términos HPO, que por su eficiencia permite usarse como primer abordaje para el diagnóstico molecular de enfermedades monogénicas, especialmente en contextos de enfermedades con elevada heterogeneidad genética. La integración de este sistema en un *pipeline* bioinformático de análisis de variantes genéticas reduciría el tiempo de respuesta de los procesos diagnósticos.

## BIBLIOGRAFÍA


1. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing:from basic research to diagnostics. Clin Chem 2009;55:641-58. DOI: 10.1373/clinchem.2008.112789
2. Stark Z, Schofield D, Alam K, Wilson W, Mupfeki N, Macciocca I, et al. Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet Med 2017;19(8):867-74. DOI: 10.1038/gim.2016.221
3. Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. JAMA Pediatr 2017;171(9):855-62. DOI: 10.1001/jamapediatrics.2017.1755
4. Payne K, Gavan SP, Wright SJ, Thompson AJ. Cost-effectiveness analyses of genetic and genomic diagnostic tests. Nat Rev Genet 2018;19(4):235-46. DOI: 10.1038/nrg.2017.108
5. Masseroli M, Galati O, Manzotti M, Gibert K, Pinciroli F. Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. BMC Bioinformatics 2005;6(Suppl 4):S18. DOI: 10.1186/1471-2105-6-S4-S18
6. Bajdik CD, Kuo B, Rusaw S, Jones S, Brooks-Wilson A. CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinformatics 2005;6:78. DOI: 10.1186/1471-2105-6-78
7. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet 2006;14(5):535-42. DOI: 10.1038/sj.ejhg.5201585
8. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83(5):610-5. DOI: 10.1016/j.ajhg.2008.09.017
9. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019;47(D1):D1018-27. DOI: 10.1093/database/bay026
10. Köhler S. Improved ontology-based similarity calculations using a study-wise annotation model. Database [Internet]. 2018 Jan 1;2018. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bay026/4953405
11. Mungall CJ, Washington NL, Nguyen-Xuan J, Condit C, Smedley D, Köhler S, et al. Use of model organism and disease databases to support matchmaking for human disease gene discovery. Hum Mutat 2015;36(10):979-84. DOI: 10.1002/humu.22857
12. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 2017;45(D1):D712-22. DOI: 10.1093/nar/gkw1128
13. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. Am J Hum Genet 2017;100(2):185-92.
14. Taruscio D, Groft SC, Cederroth H, Melegh B, Lasko P, Kosaki K, et al. Undiagnosed Diseases Network International (UDNI): White paper for global actions to meet patient needs. Mol Genet Metab 2015;116(4):223-5. DOI: 10.1016/j.ymgme.2015.11.003
15. Lochmüller H, Torrent I Farnell J, Le Cam Y, Jonker AH, Lau LP, Baynam G, et al. The International Rare Diseases Research Consortium: Policies and Guidelines to maximize impact. Eur J Hum Genet 2017;25(12):1293-302. DOI: 10.1038/s41431-017-0008-z
16. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet 2017;100(5):695-705. DOI: 10.1016/j.ajhg.2017.04.003
17. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. Genet Med 2015;17(5):405-24. DOI: 10.1038/gim.2015.30
18. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn 2018;20(1):4-27. DOI: 10.1016/j.jmoldx.2017.11.003
19. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16 [cited 2016 Mar 7];3. Available from: http://arxiv.org/abs/1303.3997
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357-9. DOI: 10.1038/nmeth.1923
21. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma 2013;(Supl.43). DOI: 10.1002/0471250953.bi1110s43
22. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res 2016;44(11):108. DOI: 10.1093/nar/gkw227
23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164. DOI: 10.1093/nar/gkq603
24. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17(1):122. DOI: 10.1186/s13059-016-0974-4
25. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74. DOI: 10.1038/nature15393
26. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 2017;45(D1):D840-5. DOI: 10.1093/nar/gkw971
27. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581(7809):434-43. DOI: 10.1038/s41586-020-2308-7
28. Jezela-Stanek A, Ciara E, Jurkiewicz D, Kucharczyk M, Jędrzejowska M, Chrzanowska KH, et al. The phenotype-driven computational analysis yields clinical diagnosis for patients with atypical manifestations of known intellectual disability syndromes. Mol Genet Genomic Med 2020;8(9). DOI: 10.1002/mgg3.1263
29. Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 2014;24(2):340-8. DOI: 10.1101/gr.160325.113
30. Cipriani V, Pontikos N, Arno G, Sergouniotis PI, Lenassi E, Thawong P, et al. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. Genes (Basel) 2020;11(4):460. DOI: 10.3390/genes11040460