Emilia Zawadzka-Gosk
Krzysztof Wołk

# SEMANTIC-ENABLED HYBRID GENETIC DISEASE DIAGNOSTICS IN NEXT-GENERATION SEQUENCED DATA

**Abstract** *Next Generation Sequencing is a technology for genome sequencing used in genetics for the diagnosis of disease. NGS provides a list of all mutations in a genome, so identifying the one that causes a disease is not trivial. A number of applications for variant prioritization were developed, but the data they provide is a suggestion rather than a diagnosis; moreover, they suffer from issues such as identifying a nonpathogenic variant as a causal one or the inability to identify a causal gene. These issues inspired us to create a strategy for variant prioritization, which includes the use of the Exomiser and OMIM Explorer result sets improved by semantic analysis of abstracts and articles freely available from the PubMed and PubMed Central databases. For the wider scope of scientific articles, the Google Scholar repository will be used. The described approach enables us to present the latest and most accurate information about potential pathogenic variants.*

**Keywords** gene prioritization, variant prioritization, semantical text analysis

**Citation** Computer Science 19(2) 2018: 179–199

## 1. Introduction

The information about each living organism is programmed in its DNA. DNA is a molecule composed of two coiled strands called the double helix. A DNA molecule consists of smaller sections (nucleotides), which are built from sugar, a phosphate group, and a nucleobase such as cytosine (C), guanine (G), adenine (A), or thymine (T). The nucleobases create pairs according to the rule that C always pairs with G and A always pairs with T [25]. The DNA strand contains many genes that are formed in the chromosomes. Humans have 46 chromosomes. Half of them are received from one parent and the other half come from the other. Genes consist of some number of nucleotides and are the units of heredity. Genes can have different variants called alleles, which differ by one or more nucleotides. The process of creating new variations is called mutation. Mutations lead to population variety, but it can also be a cause of genetic diseases. Alleles can be dominant or recessive. The dominant allele is responsible for the dominant phenotype. This means that this allele determines the specific phenotype. For recessive phenotype manifestation, it is required that both copies of the allele be recessive. If both alleles (from the mother and father) of the same gene are the same for a particular trait they cause, the organism is called homozygous. If the alleles are different, then it is called heterozygous. Knowledge about genes and DNA is broadly used for genetic disease diagnostics. A contemporary technology such as Next-Generation Sequencing enables the sequencing of the entire human genome in one day [4]. The variety of methods allows for the selection of the most efficient and appropriate one, like sequencing an entire genome, coding genes (a whole exome), or analyzing only individual genes. The NGS analysis enables the gene variants responsible for genetic disorders to be located; however, due to the large amount of data, some additional analysis is required. A number of applications have been developed in order to facilitate the search for disease-causing variants. Although the programs employ sophisticated algorithms to compare patient data with the available genetic databases, the results can be treated more like a suggestion than a real diagnosis. The final list of provided pathogenic variants needs to be analyzed by a physician and considered important or not according to a geneticist's best knowledge. The analysis performed by several applications also uses animal genetic databases due to the lack of sufficient information in the human datasets. During the research performed with Warsaw Medical University, some examples showed the limitations of variant prioritization programs. An Exomiser prioritization tool indicated the WSF5 variant as a pathogenic one; however, according to the ClinVar tool, it is not a causal variant. Another example of an Exomiser application occurred in the analysis of a patient with a diagnosed SPATA5 variant. Although the mutation is described in PubMed and has an entry in Online Mendelian Inheritance in Man (OMIM) with the analyzed phenotype database, Exomiser did not indicate such a variant in its results. This issue demonstrates that complex algorithms may not always be suitable for simple analysis. An additional analysis that could confront variant prioritization application results with the latest scientific knowledge is needed. Such a solution can be implemented as a semantic analysis tool of the PubMed database containing a broad variety of the latest medical articles. Pointing to the latest research about a specific analyzed genetic variant may bring a considerable advantage to the diagnosis process.

## 2. State of the art

Coincident with the rise of genome-wide data for diagnostics has been the development of standards and catalogs for clinical sign-out [15,32,33]. Much of the focus has been on distinguishing clearly deleterious variants from other variants with less clear contributions to disease. Central to these efforts has been the development of compendia for matching the observed variations to well-vetted disease information [24,26]. Some variants cataloged as "deleterious" can also appear in unaffected individuals; therefore, additional tools have become necessary to identify the specific variants or variant combinations such as variant pairs for recessive diseases that may explain the observed phenotypes in affected individuals [47].

Parallel to the development of catalogs and standards for variant analysis has been the development of systematic tools for representing patient information. Initially constructed in 2008, the Human Phenotype Ontology (HPO) is a representation of the features of human diseases and the hierarchical relationships that exist among them [35].

### 2.1. Phenomizer

A key application of this work is Phenomizer, a software tool used for making comparisons between known diseases and patient phenotypes [20]. This tool uses semantic similarity methods to match patient characteristics as represented in the HPO to the OMIM disease catalog, which is also mapped to the ontology. Phenomizer returns candidates within the differential diagnosis as lists and tables, with the scores representing the quality of the match [20].

The goal of variant prioritization is to construct an ordered ranking of an observed genetic variation. This objective differs from that of a differential diagnosis, the fundamental purpose of Phenomizer. To bridge the gap between disease rankings and gene or variant rankings, extensions of this initial approach have been developed and applied to the genome-wide diagnostic data. Two such tools are PhenIX [26, 28, 47] and Phenomantics [28], which directly leverage the Phenomizer's semantic similarity calculation to consider the genome-wide genotypic data.

### 2.2. eXtasy

On the other hand, the eXtasy tool [38] takes a data-integration approach (genomic data fusion [1]) to variant prioritization. To generate an overall prediction of causality, ten different measures of variant deleteriousness from existing tools and databases along with a gene haploinsufficiency prediction score are combined with a phenotype-specific gene score. The phenotype-based method takes all disease genes known to be associated with a particular HPO term or terms from Phenomizer [20] and scores the similarity of each candidate gene in the exome of this gene set using the Endeavour algorithm [1]. Endeavour uses various measures of gene similarity, such as sequence similarity and co-expression, as well as involvement in the same protein – protein interactions or pathways.

A Random Forest algorithm is used to produce a single combined candidacy score from all of these sources of evidence. For variants that are missing data from any of the methods, an imputed score is calculated that ignores haploinsufficiency and uses median values across all variants for the missing deleteriousness scores.

Receiver operating characteristic (ROC) analysis was used to assess the ability of eXtasy to discriminate disease-causing variants from rare control variants or common polymorphisms. This analysis showed substantial improvement when compared to classical deleterious prediction methods such as PolyPhen, SIFT, MutationTaster, and CAROL. Currently, eXtasy only performs a prioritization of non-synonymous variants; however, when public datasets that are large enough for training become available, it will be expanded to include mitochondrial, noncoding, synonymous, and nonsense variants as well as mutations around the splice junction that affect splicing, insertion, and deletion of base mutations (indels). eXtasy does not perform filtering, so it is recommended that the exome is pre-filtered to remove off-target or common (MAF>1%) variants [38].

## 2.3. Phevor

Phevor [39] takes the outputs of variant-prioritization tools such as ANNOVAR or the Variant Annotation Analysis Search Tool (VAAST) [46] and then prioritizes the remaining genes using phenotype, gene function, and disease data. This knowledge comes from publicly available gene annotation sets from various biomedical ontologies such as the HPO, Mammalian Phenotype Ontology (MPO) [10,44], Disease Ontology (DO) [21], and Gene Ontology (GO) [7]. Users specify a list of terms from HPO, DO, MPO, GO, or OMIM [3] that characterize what is known about the patient. Phevor then generates a list of genes that have been annotated with these terms or their parent terms (if no gene annotations exist). Next, it identifies terms in the other ontologies that are annotated to these genes, and the process is repeated to expand the gene list. Thus, concepts in different ontologies are related through their annotation of the same gene. Finally, each gene receives a score based on the propagation of the seed nodes in each ontology and a combination procedure across the scores from the various ontologies. The final Phevor score combines the ranking information for the variant prioritization tool (or P-value from VAAST) with this gene score.

The benchmarking of Phevor on simulated disease exomes based on in-house generated exomes demonstrated a considerable improvement over variant prioritization methods such as ANNOVAR and VAAST, with 95–100% of the exomes having the causative variant in the top ten candidates. Three case studies where Phevor was used to identify disease-causing alleles have also been presented [30].

## 2.4. Phen-Gen

Phen-Gen [19] uses a Bayesian framework to compare the predicted deleterious variants in the patient's exome and known patient symptoms to the prior knowledge of human disease-gene associations and gene interactions. Coding variants are analyzed using a unifying framework to predict the damaging impact of the non-synonymous, splice-site, and indel variants. Phen-Gen also allows a genome-wide approach in which evolutionary conservation and the Encyclopedia of DNA Elements (ENCODE)--predicted functionality and proximity to the coding sequences are used to score noncoding variants. Any variant that has an MAF above 1% is removed from further analysis. Healthy individuals contain many damaging mutations, and the fact that this ability to tolerate mutations varies from gene to gene is also taken into account using a null model. This model uses the observed variants from the 1000 Genomes Project to generate a null distribution under either a dominant or recessive inheritance model for each gene. Genes are only retained for further analysis if the predicted damaging score for the variants exceeds that seen for 99% of the 1000 Genomes dataset. These remaining genes are then analyzed using the Phenomizer algorithm to semantically match the patient's phenotypes encoded using HPO to known disease-gene associations. The role of the novel (non-disease) genes is assessed by identifying functionally related genes using a random-walk-with-restart algorithm over a gene interaction network. Phenotype matches are distributed to these novel genes across the network such that the disease gene hub gets the majority of the score (90%) and the other genes each get a share of the remainder (according to their proximity to the disease gene). Benchmarking using the simulated exomes that were based on 1000 Genomes Project data showed that the correct disease variant was obtained as the top hit in 88% of the samples. Using a strategy in which the known associations were masked to simulate the discovery of novel associations, performance figures of 56% and 89% were obtained for the dominant and recessive disorders, respectively. In an evaluation using real patient data, 11 trios with recessive or X-linked intellectual disabilities were analyzed, and 81% of the reported genes were among the top ten candidates [19].

## 2.5. Exomiser

The original implementation of Exomiser [36] used semantic similarity comparisons between patient phenotypes and mouse phenotype data for each candidate gene in the exome. The PhenoDigm algorithm [43] is used to score each gene from 0 to 1, where 1 represents a perfect match, and genes with no data receive a default score of 0.6. This phenotype score is combined with a variant score that is based on allele rarity in the 1000 Genomes Project and ESP datasets together with the predictions of deleteriousness from PolyPhen, SIFT, and MutationTaster. Benchmarking on the simulated exomes based on 1000 Genomes Project data showed that 66% of the cases had the causative variant as the top hit under a dominant model and 83% under a recessive model [36]. Exomiser has been subsequently improved to include comparisons with human and fish phenotypes as well as the use of a random-walk with restart to score genes with no phenotype data (genes are scored based on their proximity to other genes in the StringDB interaction network that show phenotypic similarities to the patient data) [42].

### 2.6. PhenIX

PhenIX [47] [29] uses the same software framework as Exomiser; however, instead of using human, mouse, fish, and protein—protein association data, this tool is restricted to comparisons between patient phenotypes and the known disease gene phenotypes. This simplification is made because PhenIX is intended for diagnostic tasks in which only known the disease genes can be reported. In addition, the semantic similarity algorithm uses the Phenomizer algorithm [20]. Benchmarking on sequence files generated from a target-enrichment panel that was based on the known disease-associated genes revealed that 97% of the samples had the inserted variant as the top hit regardless of the inheritance model. The same performance was observed when using 1000 Genomes Project exomes [40].

### 2.7. OMIM Explorer

OMIM Explorer [17] introduces an interactive approach to variant prioritization. The implemented plots of the global and local visualizations enable the user to control the analysis results. The patient phenotype can be input in a free-text form and is transformed to the HPO annotation with a natural language processing tool (Bio-Lark Concept Recognizer [16]). The genotype data should be entered in the VCF format pre-filtered to <1% MAF (population minor allele frequency) or as a list of the main (rare) variant genes. OMIM Explorer uses a semantic similarity method to compare the patient phenotype to the OMIM catalog. To determine the similarity, two methods are used: the Resnik method [31] or ATO (ancestral term overlap). Variant frequency and pathogenicity are computed based on data from the ExAC database [11] and with the MutationTaster tool [37]. An analysis with autosomal dominant and autosomal recessive inheritance of the genetic disease models is available to the OE user. The default setup applies no filter for this. An algorithm for novel disease-causing gene discovery is also implemented in the application. First, the patient's phenotype is mapped to the OMIM records to receive the group of genes that causes diseases similar to the patient's phenotype. After that, the PINA 2.0 PPI network [8] is used to discover the patient's variants using the training set.

## 3. Technical review of Exomiser and OMIM Explorer tools

### 3.1. Exomiser

Exomiser [41] is an application for variant prioritization based on whole-exome sequencing results. The tool analyzes genetic data provided in a Variant Call Format (VCF) file and a set of Human Phenotype Ontology (HPO) terms. Exomiser's analysis consists of two main parts: filtering and prioritization. First, each of the input variants is annotated to the relative one from the hg19 database to gather information about them. After this, the variants not required for analysis are filtered out from the dataset. Exomiser enables an analysis to be performed with user-defined parameters. The two used most often are the minor allele frequency (MAF) filter (which enables us to find the rarest variants among all variants from the input dataset) and the expected inheritance pattern filter. Three options are available for inheritance pattern filtering: autosomal dominant (AD), autosomal recessive (AR), and X-linked (X). The filter restricts the output data for genes containing heterozygous variants (AD), genes containing homozygous or two heterozygous variants (AR), or X-chromosomal genes. Exomiser also enables the possibility of family-based filtering, which requires an additional input file (PED) containing family members' genes. Prioritization of the patient's variants is determined based on how rarely the variant occurs in the 1000 Genomes Project [40] and Exome Sequencing Project (ESP 6500) datasets. Four main prioritization methods are used in Exomiser. The PHIVE (Phenotypic Interpretation of Variants in Exomes) algorithm uses mouse gene data as a comparison to the human gene model. The mouse data comes from the Mouse Genome Database (MGD) [6] and the International Mouse Phenotyping Consortium [5]. The PhenIX algorithm is a tool for clinical diagnosis where only human data can be used. The algorithm computes variant ranking on the basis of the pathogenicity and semantic similarity of phenotypes from HPO that are connected to the Mendelian disease. The ExomWalker algorithm is used for discovering new causal genes by searching for the mutated genes that interact with the genes that are already implicated in a disease. The ExomWalker employs a random walk with restart algorithm. The random walk algorithm is a method usually used for image segmentation [14]. Having some number of predefined labeled pixels, the unlabeled pixels are assigned to the proper categories on the basis of the greatest probability of reaching the predefined pixel with a random walker. Exomiser uses a similar mechanism for discovering causal genes among closely interacting genes (a mutation of any gene in such a group will cause a similar phenotype). The user inputs the list of suspected genes, and the algorithm calculates which mutated gene is closely related to the inputted one in the protein-protein association network. The hiPHIVE method calculates variant ranking using human, mouse, and zebrafish data. Human data comes from OMIM and Orphanet databases. Zebrafish gene information comes from the Zebrafish Model Organism (ZFIN) database. To compute a variant's ranking, Exmiser uses two indicators: the variant and phenotype scores. The variant score is a measure that shows the pathogenicity and frequency of the variant. Pathogenicity is calculated on the basis of the scores from three sources: the Polyphen2, MutationTaster and 1 – SIFT scores. Frequency is determined with the use of the 1000 Genomes Project and ESP data. The variant score result is a compounding of these two scores. If no result can be established, the default value of 0.6 is assigned. The phenotypic score is calculated based on the semantic similarity of the patient's phenotype and phenotype-gene annotations for humans, mice, and zebrafish (for hiPHIVE). Besides this, a random-walk algorithm is used to compare the patient's phenotype to the related phenotypes of the nearby genes. A network for the computation is created with the use of STRING (Search Tool for the Retrieval of Interacting Genes/Proteins). The result of the algorithm is a probability vector with values between 0 and 1. The values from the vector (which are the proximity scores for the analyzed genes) are used as a weight for a phenotyping relevance score. The results are rescaled from 0 to 0.6. The final phenotypic score is computed as a maximum value from the semantic similarity analysis and random-walk algorithm. The variant and phenotypic scores are finally used to calculate the Exomiser Score, which is computed as follows:

$$ExomiserScore = \frac{1}{(1 + e^{(-(-13{,}96 + 11{,}61 \cdot PhenotypeScore + 11{,}61 \cdot VariantScore))})} \qquad (1)$$

## 3.2. OMIM Explorer

OMIM Explorer is an interactive web-based tool that was created for diagnostics based on a patient's phenotype information and genotype data. The application uses the OMIM database and HPO to integrate and analyze the medical information. OMIM [2] is an open-access medical database that contains human phenotypes and genes. It is updated daily based on the published biomedical literature. All of the entries in a database are numbered and marked according to their level of certainty (if they are reviewed or not). The tool also provides Morbid Map and Synopsis Map views of the relationships between the gene and the disease. HPO [35] is a database providing a structured set of phenotype abnormality terms. The ontology was created based on OMIM records and is broadly used by various medical applications. Genetic and phenotypic information is provided as input data for the OMIM Explorer application. The phenotype can be provided in a free-text form from which the HPO terms are extracted or in a list of HPO terms. Gene information might be uploaded as a VCF format file or as a list of genes. Based on the provided HPO terms, the diagnostic disease ranking is calculated as semantic similarity, a technique that computes matches between the queried term and the ontology. Two main methods are used by OE: Resnik [31] and ATO (but also, ATO weighted by the GO-Universal information and ATO weighted by annotation-based information content). The scores are calculated for all OMIM diseases. They could also be restricted to those present in an OMIM Morbidmap, the chosen genetic model (dominant or recessive), linked to the genetic variants, or set as required by a user. To compute the results for the inputted variant genes, the OE tool uses a transitive closure approach based on the phenotype and disease matches. The scores are calculated only for the diseases that the OMIM database maps for the input genes. The result is determined by function $F$, which computes the aggregation of similarity scores of the phenotype data diseases related to the genes. $F$ can be a maximum, mean, or sum. For comparison, the algorithm of the direct gene scoring approach is also used (the same as with Phenomantics) as well as the method of computing the unions of phenotypes related to input genes by the OMIM Morbidmap (the same as with PhenIX). The data can be analyzed as dominant or recessive genetic models. As OE is an interactive tool, two main visualizations are available. The global one is an interactive map of all available OMIM diseases distributed with the Multidimensional Scaling (MDS) [45] method. The position of the patient's disease on a chart is computed on the basis of the "m" nearest neighbors, where "m" can be chosen by the user. Local visualization presents the patient's query in the center and the most similar diseases placed on the radar according to their one-dimensional MDS results. The chart diseases are scaled based on their variant frequency in the ExAC database [11] and colored by their variant pathogenicity score calculated by the MutationTaster tool [37]. Visualization enables the interactive exclusions of selected diseases from the analysis. OE has an algorithm for novel gene and variant discovery that is generally based on the similarity of the patient's phenotype to OMIM entries in order to identify the set of genes mapped to the diseases that are most similar to the patient's query. After this step, the PINA 2.0 PPI [8] network is used to discover the candidate genes. OMIM Explorer also provides additional features (like phenotype suggestions) based on rare phenotypes that are not present in a patient's query but are annotated as most similar to the patient's disease.

## 4. Authors' solutions to Next Generation Sequencing results analysis

### 4.1. Implementation of Exomiser as Cloud Service

The main goal of this research was to provide Warsaw Medical University (WUM) with a usable genetic diagnostic tool. Because most research conducted on scientific NGS was not necessary, the main problem was the data filtration in order to provide robust and accurate results that would be highly related to only potential human diseases. It was not possible to return adequate information about gene mutation in other species. First, the Exomiser tool was used and a special tool implemented in order to carry on user queries in an automatic manner. Unfortunately, the results provided by Exomiser were not accurate, and a lot of noisy data was provided. The experiments were conducted on eight random and anonymous sequenced gene samples provided by the WUM. Table 1 presents the number of results provided by the Exomiser tool without any modifications.

As presented in Table 1, such an enormous number of potential results makes it very hard or even impossible for a doctor to make the right diagnosis in an affordable amount of time. Also, the accuracy of such results is very low. This is why we were able to adjust adequate filtering strategies for such an analysis in an empirical study based on the patient's HPOs and doctor's suspicions. First of all, as a pre-analysis step, we removed the synonymous gene variants, intron variants, and intergenic variants (coding and non-coding). Because disease-causing mutations are rare, we also filtered the sequenced input variants by frequency.

**Table 1**

Number of results provided by Exomiser with adjusted filtering strategy

| Sample ID | Number of results |
|---|---|
| 1 | 18 |
| 2 | 15 |
| 3 | 15 |
| 4 | 23 |
| 5 | 19 |
| 6 | 17 |
| 7 | 12 |
| 8 | 18 |

When conducting an autosomal dominant analysis, the maximum frequency was set to 0.0001; for the autosomal recessive 0.3; and for the X recessive 0.0003 [27]. We also removed the results with a prioritization score lower than 0.5 using the HiPhive prioritization algorithm [18]. With this strategy, we greatly reduced the number of results (as presented in Table 2).

**Table 2**

Number of results provided by Exomiser without modifications

| Sample ID | Number of results |
|---|---|
| 1 | 1462 |
| 2 | 1375 |
| 3 | 2034 |
| 4 | 1653 |
| 5 | 3261 |
| 6 | 1742 |
| 7 | 1274 |
| 8 | 1563 |

Because not all of the results were known to be pathogenic, we annotated the results in accordance to the ClinVar database that archives and aggregates information about relationships among variations and human health [24]. Unfortunately, the results were only partially satisfactory. While some results correlated with the doctor's judgments and some were potentially interesting because of the correspondence with other species, they were not analyzed enough to be treated as diagnoses. Because of this, a trial was made to use the most recent OMIM Explorer tool. OMIM Explorer did not encounter as many resultant pathogenic variants as Exomiser. The analysis was performed with the default application settings. No inheritance model was chosen. The input VCF [9] file for OMIM Explorer needs to be pre-filtered with MAF. Files that contain more than 100,000 rows are not processed by the program. The experiment was provided with the same set of patient data as used in Exomiser. In Table 3, we present the number of results provided by OMIM Explorer.

Pathogenic variants returned by OE-overlapped Exomiser only results in a small range, with a maximum number of two common variants and a minimum of zero similar results.

**Table 3**
Number of results provided by OMIM Explorer

| Sample ID | Number of results |
|---|---|
| 0 | 16 |
| 1 | 19 |
| 2 | 12 |
| 3 | 15 |
| 4 | 14 |
| 5 | 19 |
| 6 | 15 |
| 7 | 14 |
| 8 | 13 |

The OMIM Explorer tool is a web application that requires direct user engagement in the analysis process. To simplify the use of OE, the web crawler tool was developed. A script enables the VCF file to be provided as well as the phenotype as a list of HPO terms. Such an approach will also facilitate an automatic comparison of the different variant prioritization application results. Both OMIM Explorer and Exomiser suffered very similar drawbacks. We also discovered that some recently discovered diseases causing gene mutations were not or gene mutations that no longer are considered to be problematic for human health were within the prioritization results. In a real-life diagnosis, such mistakes should not appear and the databases should be as current as possible.

## 4.2. Semantical human-genetic diagnoser (HGD)

The problems described in Section 4.1 made us prepare more-sophisticated diagnosis strategies. First, we interpolated the results of Exomiser and OMIM Explorer. Second, we developed a semantic text analysis tool accompanied with web crawlers in order to analyze scientific research articles. For this purpose, a sophisticated search engine (ElasticSearch) was used [13]. At first, our web crawlers downloaded abstracts from the PubMed repository (only abstracts, which are freely available) [22], and facilitated articles and supplementary files freely available in the PubMed Central database [34]. Third, using the Python tool we implemented with the ElasticSearch engine, we queried those repositories using the diagnosed HPO's and genes prioritized with Exomiser and OMIM Explorer. Our search engine semantically analyzed abstracts, articles, and supplementary files. Not only was the HPO-gene relationship queried, but the publication release date, number of keyword hits, and number of citations were also taken into account [23]. In addition, we also queried the Google Scholar repository (which has a wider scope) and interpolated both search results [12]. Lastly, having the search results gathered, we altered the prioritization results in accordance with the findings. The most-relevant results were at the top of the list, whereas irrelevant or out-of-date results were removed. If no very accurate answer could be given to the user, the tool provided him/her with a list of articles that correlate with the disease for his/her own judgment. Such a methodology allows users to receive up to five relevant results in most cases. The current version of our semantic text analysis provides a web browser searching interface. Simple logical operations of conjunction and disjunction can be applied to a query. Moreover, a user-defined slop is available for use. Slop describes how the searched terms can be distant from each other in an article. a slop with a value of 1 allows for at most one additional word placed between the searched words. The default value for the slop in our tool is 5. A returned-results score is presented as a graphical star ratio.

### 4.3. Authors' motivation

The aim of our work is to support the process of diagnosing patients with genetic diseases. The solution we developed facilitates two difficulties in the diagnoses: the huge amount of data provided by the Next Generation Sequencing process, and the great number of scientific articles published in the genetic area. We use stable and reliable tools to provide the most-probable causal variants from the input data. It provides suggestions for the doctor regarding which variants needs to be analyzed first. The list of causal variants is verified with the scientific articles from the PubMed repository. The doctor can decide whether the proposed variant is valuable for the diagnosis and the knowledge about the variant could be completed. The solution allows for an automatic data analysis. The process of diagnosis can be faster and more efficient, which brings benefits to the doctors and patients because the treatment can be started earlier. The doctor does not need to analyze a huge amount of data. Supported by knowledge from scientific papers, the diagnosis could be more precise and accurate.

## 5. Results and conclusions

In an experiment with our HGD tool, we used nine anonymous VCF files [9] that were first analyzed with the Exomiser and OMIM Explorer applications. The Exomiser settings and filters were set with the values described in Section 4.1 of this article. The OMIM Explorer analysis was run with default OE settings. The same HPO terms were provided to both tools. Exomiser usually returned a slightly larger number of suggested pathogenic variants than OMIM Explorer did. The comparison of the two resultant sets of each sample revealed that some variant suggestions appeared in both resultant sets. The number of overlapping results was usually not greater than two; for one sample, the application returned divergent results. Table 4 presents detailed information about the number of returned results for the analyzed samples.

**Table 4**

Number of variants returned by Exomiser and OMIM Explorer for analyzed samples

| Sample ID | Exomiser | OMIM Explorer | Overlapping variants |
|---|---|---|---|
| 0 | 24 | 16 | 2 |
| 1 | 16 | 19 | 1 |
| 2 | 21 | 12 | 1 |
| 3 | 22 | 15 | 0 |
| 4 | 25 | 14 | 2 |
| 5 | 24 | 19 | 2 |
| 6 | 30 | 15 | 1 |
| 7 | 30 | 14 | 1 |
| 8 | 21 | 13 | 1 |

Each of the resultant variants was provided to the semantic text analysis tool in order to retrieve the related PubMed articles. The retrieval was performed according to three patterns: with the variant name only, with the variant name and one or more of the patient's phenotype terms, or with the variant name and all of the phenotype terms. Table 5 presents the average and medium numbers of returned articles for each sample based on the set of variants outputted by Exomiser and OMIM Explorer. The calculation was created taking into account the numbers of articles found in the PubMed and PubMed Central databases for each variant (returned by Exomiser or OE) for each sample.

**Table 5**

Average and median number of PubMed articles returned by HGD from Exomiser and OMIM Explorer resultant variants

| ID | Exomiser | | | | | | OMIM Explorer | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variant only | | Variant and any HPO | | Variant and all HPO | | Variant only | | Variant and any HPO | | Variant and all HPO | |
| | Avg | Med | Avg | Med | Avg | Med | Avg | Med | Avg | Med | Avg | Med |
| 0 | 497 | 365 | 10 | 4 | 2 | 0 | 3433 | 314 | 77 | 12 | 7 | 2 |
| 1 | 1591 | 485 | 29 | 4 | 2 | 1 | 926 | 312 | 39 | 20 | 6 | 2 |
| 2 | 1183 | 419 | 18 | 7 | 2 | 1 | 911 | 582 | 62 | 27 | 8 | 2 |
| 3 | 1114 | 290 | 18 | 4 | 2 | 0 | 535 | 171 | 51 | 23 | 7 | 2 |
| 4 | 740 | 317 | 20 | 5 | 3 | 1 | 1383 | 496 | 112 | 36 | 19 | 5 |
| 5 | 2367 | 338 | 38 | 5 | 3 | 0 | 946 | 299 | 33 | 17 | 3 | 1 |
| 6 | 408 | 249 | 7 | 3 | 0 | 0 | 1772 | 255 | 61 | 19 | 11 | 3 |
| 7 | 652 | 276 | 18 | 4 | 3 | 0 | 789 | 386 | 62 | 32 | 11 | 3 |
| 8 | 381 | 145 | 8 | 2 | 1 | 0 | 1147 | 898 | 74 | 36 | 12 | 2 |

The more accurate the query returns, the smaller the number of articles. Adding phenotype requirements to the computation decreased the number of returned results.

Some cases showed a surprisingly large number of returned articles, even with precise queries. This was usually caused by the fact that the genetic variant name was the same as that of another medical term, which increased the result set.

The analysis that required the presence of the variant name and all pathogenic terms in the article returned the smallest number of PubMed articles. In this analysis pattern, the majority of the cases had fewer than five related papers. Only about 20% of the variants provided by Exomiser (Figure 1) for a chosen sample and about 30% in OMIM Explorer (Figure 2) had five or more articles found in PubMed, which may indicate that they could be analyzed first in the diagnostic process.

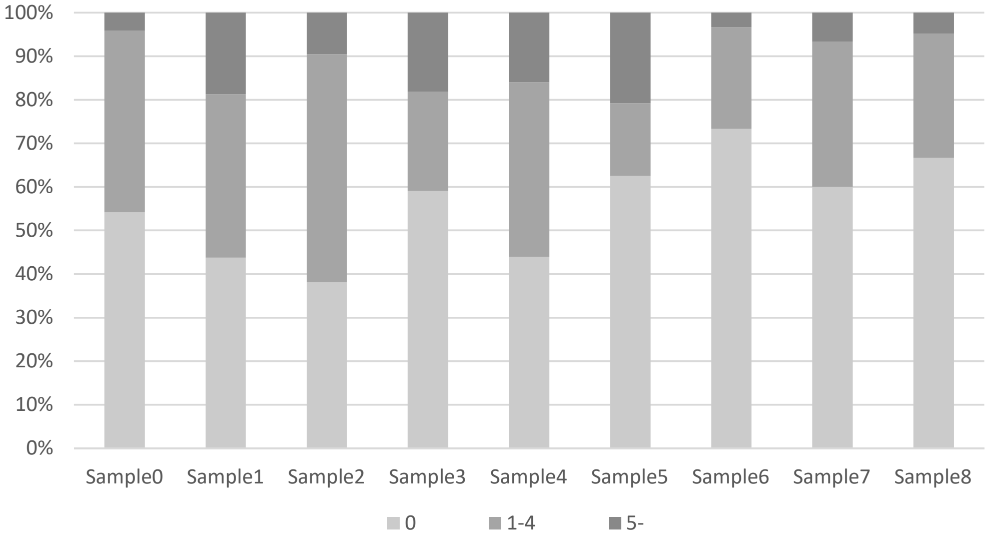

**Figure 1.** Relationship of number of PubMed articles returned by HGD analysis with variant and all phenotype terms for Exomiser resultant variants

A similar operation was performed with the Google Scholar repository. As was expected, Google Scholar returned many more articles than our HGD semantic search module. The average and median calculations of the article numbers for the analyzed samples are presented in Table 6.

The differences between the number of found articles by the search tools are quite large. Google Scholar searches for a given expression in a wide variety of sources, not only among openly available articles but also in scientific journals where a subscription is necessary. Besides that, Google Scholar uses a mechanism that improves searching by looking for similar-looking words, which is not always desirable in the case of genetic variant names because it generates a lot of article suggestions that are not related to the searched subjects. Google Scholar searches more than just medical journals, which also contributes to the increase in the number of suggested articles.

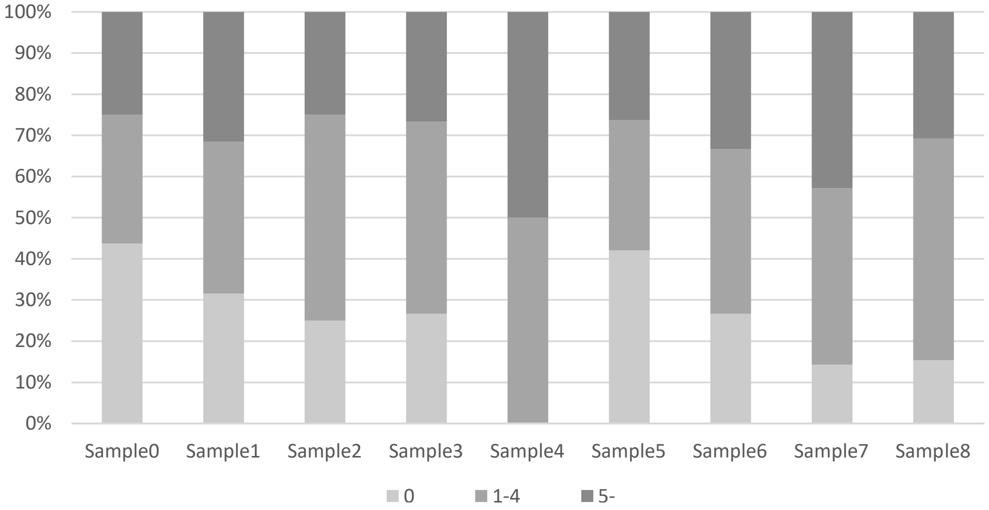

**Figure 2.** Relationship of number of PubMed articles returned by HGD analysis with variant and all phenotype terms for OMIM Explorer resultant variants

**Table 6**

Average and median number of PubMed articles returned by HGD semantic search mechanism from Exomiser and OMIM Explorer resultant variants

| ID | Exomiser | | | | | | OMIM Explorer | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variant only | | Variant and any HPO | | Variant and all HPO | | Variant only | | Variant and any HPO | | Variant and all HPO | |
| | Avg | Med | Avg | Med | Avg | Med | Avg | Med | Avg | Med | Avg | Med |
| 0 | 41707 | 5370 | 1677 | 81 | 1301 | 12 | 692694 | 4595 | 5774 | 216 | 1609 | 54 |
| 1 | 169751 | 6255 | 1850 | 254 | 1464 | 43 | 78622 | 5510 | 1858 | 413 | 1947 | 95 |
| 2 | 95970 | 7010 | 1031 | 130 | 1345 | 40 | 98434 | 6085 | 2383 | 422 | 1552 | 108 |
| 3 | 52511 | 7125 | 1305 | 76 | 1001 | 19 | 49284 | 2130 | 2038 | 413 | 1250 | 120 |
| 4 | 25056 | 5140 | 433 | 133 | 1925 | 40 | 136539 | 6105 | 3707 | 437 | 1800 | 180 |
| 5 | 152441 | 4600 | 1253 | 89 | 186 | 17 | 15184 | 4780 | 705 | 303 | 536 | 95 |
| 6 | 53116 | 5510 | 354 | 75 | 1247 | 17 | 241357 | 6370 | 1323 | 467 | 1965 | 125 |
| 7 | 12184 | 5700 | 673 | 80 | 535 | 18 | 99423 | 4415 | 758 | 349 | 364 | 74 |
| 8 | 71355 | 2270 | 582 | 60 | 836 | 8 | 202310 | 6020 | 3294 | 543 | 3534 | 157 |

After combining the Exomiser and OMIM Explorer results with the articles returned by HGD and GoogleScholar, it is noticeable that the variants present in both variant prioritization tool's result sets have quite a large number of returned articles. The numbers are usually above the medians presented in Tables 5 and 6. This indicates that the known and well-studied variants were proposed by the tool.

For each analyzed sample, the variants present in both the OMIM Explorer and Exomiser resultant sets were provided to semantic search tools (HGD and Google Scholar) with a patient's phenotype terms. As the input for the search, two approaches were chosen: variant name with all phenotype terms required and variant name with at least one of the provided phenotype words required. The results are presented in Table 7.

**Table 7**

Number of articles found by HGD and GoogleScholar for variants present in OE and Exomiser result sets for provided samples

| SampleID | Variant name | HGD | | GoogleScholar | |
|---|---|---|---|---|---|
| | | Variant and any HPO | Variant and all HPO | Variant and any HPO | Variant and all HPO |
| 0 | WFS1 | 76 | 14 | 677 | 211 |
| | PLIN1 | 13 | 4 | 159 | 40 |
| 1 | CDKN1C | 51 | 7 | 625 | 164 |
| 2 | CYP21A2 | 16 | 1 | 409 | 125 |
| 3 | – | – | – | – | – |
| 4 | ABCC8 | 211 | 45 | 2200 | 1370 |
| | GNAS | 41 | 4 | 543 | 157 |
| 5 | ALMS1 | 31 | 7 | 303 | 75 |
| | CPT2 | 58 | 5 | 761 | 242 |
| 6 | MC2R | 33 | 2 | 563 | 133 |
| 7 | ABCC8 | 211 | 45 | 2200 | 1370 |
| 8 | GNAS | 41 | 4 | 543 | 157 |

The strategy where the doctor searches for the pathogenic variant name and at least one of the phenotype terms is the most likely one. Therefore, this case is presented in Figure 3. The median number of articles found for one variant in the 'OR strategy' by Google Scholar is 137, but half of the results are numbers between 45 and 622. Reading and manually analyzing all of the available articles might be difficult. As its repository is much smaller, HGD returns a median number of 7 articles, but the values of the first and third quartiles are 2 and 22, respectively. Reading or inspecting such a number of articles is time-consuming. If the doctor decides to verify all of the variants returned by the algorithms, the presented numbers need to be multiplied by the median number of returned variants by Exomiser and OMIM Explorer (which is 39). The overall number of articles found by the geneticist in GoogleScholar will exceed a thousand. HGD will return about 300 or more articles; thus, a manual analysis of such a large number of papers is not possible. A strategy for prioritizing, filtering, and choosing valuable articles is required. The results of the experiment show that the proper filtering strategy for variant prioritization can bring a great reduction in the number of potentially causal variants. The diagnosis is mostly made based on the doctor's experience and knowledge, but the information and discoveries delineated in the scientific articles is also important. An analysis of the scientific articles is important. Due to the large number of sources, the fast development and increased research in the area of a manual analysis of every potentially valuable paper might be difficult.

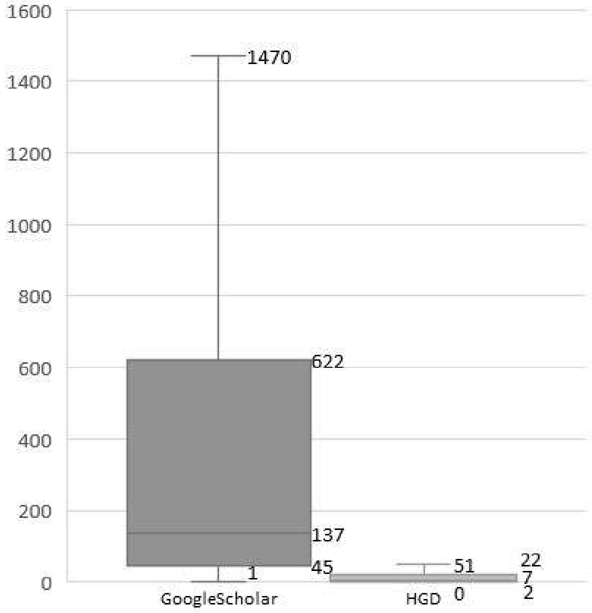

**Figure 3.** Numbers of articles returned by Google Scholar and HGD for one variant and at least one phenotype term

The research on the HGD tool presented in this article is a strategy that is going to be developed. The solution requires additional medical and technical analyses with diagnosed samples. Currently, the analyzed samples require final medical diagnosis, which will also verify the HGD application approach and review the idea of variant scoring.The performed experiments revealed a number of challenges to meet in further HGD research. Including ExomiserScore and OMIM Explorer, the calculated measures will improve the calculation of most highly pathogenic variants, especially for those samples where the results returned by Exomiser and OE do not have common values. A semantic search performed by HGD requires us to include scientific article publication dates into the scoring. The solution should also exclude the articles that are not from the genetic field and are presented in the result list because of the similarity between medical terms.

## References

[1] Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L.C., De Moor B., Marynen P., Hassan B., Carmeliet P., Moreau Y.: Gene prioritization through genomic data fusion, *Nature Biotechnology*, vol. 24(6), pp. 537–544, 2006.

[2] Amberger J., Bocchini C.A., Scott A.F., Hamosh A.: McKusick's online Mendelian inheritance in man (OMIM®), *Nucleic Acids Research*, vol. 37(suppl 1), pp. D793–D796, 2009.

[3] Amberger J.S., Bocchini C.A., Schiettecatte F., Scott A.F., Hamosh A.: OMIM. org: Online Mendelian Inheritance in Man (OMIM®). An online catalog of human genes and genetic disorders, *Nucleic Acids Research*, vol. 43(D1), pp. D789–D798, 2015.

[4] Behjati S., Tarpey P.S.: What is next generation sequencing? *Archives of Disease in Childhood. Education & Practice Edition*, vol. 98(6), pp. 236–238, 2013.

[5] Brown S.D., Moore M.W.: The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping, *Mammalian Genome*, vol. 23(9-10), pp. 632–640, 2012.

[6] Bult C.J., Eppig J.T., Kadin J.A., Richardson J.E., Blake J.A., Mouse Genome Database Group: The Mouse Genome Database (MGD): mouse biology and model systems, *Nucleic Acids Research*, vol. 36(suppl 1), pp. D724–D728, 2008.

[7] Consortium G.O.: Gene ontology consortium: going forward, *Nucleic Acids Research*, vol. 43(D1), pp. D1049–D1056, 2015.

[8] Cowley M.J., Pinese M., Kassahn K.S., Waddell N., Pearson J.V., Grimmond S.M., Biankin A.V., Hautaniemi S., Wu J.: PINA v2.0: mining interactome modules, *Nucleic Acids Research*, pp. D862–D865, 2011.

[9] Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T.: The variant call format and VCFtools, *Bioinformatics*, vol. 27(15), pp. 2156–2158, 2011.

[10] Eppig J.T., Blake J.A., Bult C.J., Kadin J.A., Richardson J.E., Mouse Genome Database Group: The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease, *Nucleic Acids Research*, vol. 43(D1), pp. D726–D736, 2015.

[11] ExAC Browser. `http://exac.broadinstitute.org/`.

[12] Falagas M.E., Pitsouni E.I., Malietzis G.A., Pappas G.: Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses, *The FASEB Journal*, vol. 22(2), pp. 338–342, 2008.

[13] Gormley C., Tong Z.: *Elasticsearch: The Definitive Guide*, O'Reilly Media, Inc., 2015.

[14] Grady L.: Random walks for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(11), pp. 1768–1783, 2006.

[15] Green R.C., Berg J.S., Grody W.W., Kalia S.S., Korf B.R., Martin C.L., McGuire A.L., Nussbaum R.L., O'Daniel J.M., Ormond K.E.: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing, *Genetics in Medicine*, vol. 15(7), pp. 565–574, 2013.

[16] Groza T., Köhler S., Doelken S., Collier N., Oellrich A., Smedley D., Couto F.M., Baynam G., Zankl A., Robinson P.N.: Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora, *Database*, vol. 2015, bav005, pp. 1–13, 2015.

[17] James R.A., Campbell I.M., Chen E.S., Boone P.M., Rao M.A., Bainbridge M.N., Lupski J.R., Yang Y., Eng C.M., Posey J.E., Shaw C.A.: A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics, *Genome Medicine*, vol. 8(1), pp. 1–17, 2016.

[18] Jamsheer A., Olech E.M., Kozłowski K., Niedziela M., Sowińska-Seidler A., Obara-Moszyńska M., Latos-Bieleńska A., Karczewski M., Zemojtel T., Shaw C.A.: Exome sequencing reveals two novel compound heterozygous XYLT1 mutations in a Polish patient with Desbuquois dysplasia type 2 and growth hormone deficiency, *Journal of Human Genetics*, vol. 61(7), pp. 577–583, 2016.

[19] Javed A., Agrawal S., Ng P.C.: Phen-Gen: combining phenotype and genotype to analyze rare disorders, *Nature Methods*, vol. 11(9), pp. 935–937, 2014.

[20] Köhler S., Schulz M.H., Krawitz P., Bauer S., Dölken S., Ott C.E., Mundlos C., Horn D., Mundlos S., Robinson P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies, *The American Journal of Human Genetics*, vol. 85(4), pp. 457–464, 2009.

[21] Kibbe W.A., Arze C., Felix V., Mitraka E., Bolton E., Fu G., Mungall C.J., Binder J.X., Malone J., Vasant D.: Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data, *Nucleic Acids Research*, vol. 43(D1), pp. D1071–D1078, 2015.

[22] Kilicoglu H., Shin D., Fiszman M., Rosemblat G., Rindflesch T.C.: SemMedDB: a PubMed-scale repository of biomedical semantic predications, *Bioinformatics*, vol. 28(23), pp. 3158–3160, 2012.

[23] Kononenko O., Baysal O., Holmes R., Godfrey M.W.: Mining modern repositories with elasticsearch. In: *Proceedings of the 11th Working Conference on Mining Software Repositories*, pp. 328–331. ACM, 2014.

[24] Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., Maglott D.R.: ClinVar: public archive of relationships among sequence variation and human phenotype, *Nucleic Acids Research*, vol. 42(D1), pp. D980–D985, 2014.

[25] Learn.Genetics. `http://learn.genetics.utah.edu/`.

[26] Li M.X., Kwan J.S.H., Bao S.Y., Yang W., Ho S.L., Song Y.Q., Sham P.C.: Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies, *PLoS Genetics*, vol. 9(1), p. e1003143, 2013.

[27] Little R.D., Folz C., Manning S.P., Swain P.M., Zhao S.C., Eustace B., Lappe M.M., Spitzer L., Zweier S., Braunschweiger K.: A mutation in the LDL receptor-related protein 5 gene results in the autosomal dominant high-bone-mass trait, *The American Journal of Human Genetics*, vol. 70(1), pp. 11–19, 2002.

[28] Masino A.J., Dechene E.T., Dulik M.C., Wilkens A., Spinner N., Krantz I.D., Pennington J.W., Robinson P.N., White P.S.: Clinical phenotype-based gene prioritization: An initial study using semantic similarity and the human phenotype ontology, *BMC Bioinformatics*, vol. 15(1), pp. 1–11, 2014.

[29] PhenIX – Phenotypic Interpretation of eXomes. `http://compbio.charite.de/PhenIX/`.

[30] Phevor: Phenotype Driven Variant Ontological Re-ranking tool. `http://weatherby.genetics.utah.edu/cgi-bin/Phevor/PhevorWeb.html`.

[31] Resnik P.: Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 1, pp. 448–453, 1995.

[32] Richards C.S., Bale S., Bellissimo D.B., Das S., Grody W.W., Hegde M.R., Lyon E., Ward B.E., the Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee: ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007, *Genetics in Medicine*, vol. 10(4), pp. 294–300, 2008.

[33] Richards S., Aziz N., Bale S., Bick D., Das S., Gastier-Foster J., Grody W.W., Hegde M., Lyon E., Spector E., Voelkerding K., Rehm H.L., ACMG Laboratory Quality Assurance Committee: 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*, vol. 17(5), pp. 405–423, 2015.

[34] Roberts R.J.: PubMed Central: The GenBank of the published literature. In: *Proceedings of the National Academy of Sciences*, vol. 98(2), pp. 381–382, 2001.

[35] Robinson P.N., Köhler S., Bauer S., Seelow D., Horn D., Mundlos S.: The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, *The American Journal of Human Genetics*, vol. 83(5), pp. 610–615, 2008.

[36] Robinson P.N., Köhler S., Oellrich A., Wang K., Mungall C.J., Lewis S.E., Washington N., Bauer S., Seelow D., Krawitz P.: Improved exome prioritization of disease genes through cross-species phenotype comparison, *Genome Research*, vol. 24(2), pp. 340–348, 2014.

[37] Schwarz J.M., Rödelsperger C., Schuelke M., Seelow D.: MutationTaster evaluates disease-causing potential of sequence alterations, *Nature Methods*, vol. 7(8), pp. 575–576, 2010.

[38] Sifrim A., Popovic D., Tranchevent L.C., Ardeshirdavani A., Sakai R., Konings P., Vermeesch J.R., Aerts J., De Moor B., Moreau Y.: eXtasy: variant prioritization by genomic data fusion, *Nature Methods*, vol. 10(11), pp. 1083–1084, 2013.

[39] Singleton M.V., Guthery S.L., Voelkerding K.V., Chen K., Kennedy B., Margraf R.L., Durtschi J., Eilbeck K., Reese M.G., Jorde L.B.: Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families, *The American Journal of Human Genetics*, vol. 94(4), pp. 599–610, 2014.

[40] Siva N.: 1000 Genomes project, *Nature Biotechnology*, vol. 26(3), pp. 256–256, 2008.

[41] Smedley D., Jacobsen J.O., Jäger M., Köhler S., Holtgrewe M., Schubach M., Siragusa E., Zemojtel T., Buske O.J., Washington N.L., Bone W.P., Haendel M.A., Robinson P.N.: Next-generation diagnostics and disease-gene discovery with the Exomiser, *Nature Protocols*, vol. 10(12), pp. 2004–2015, 2015.

[42] Smedley D., Köhler S., Czeschik J.C., Amberger J., Bocchini C., Hamosh A., Veldboer J., Zemojtel T., Robinson P.N.: Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases, *Bioinformatics*, vol. 30(22), pp. 3215–3222, 2014.

[43] Smedley D., Oellrich A., Köhler S., Ruef B., Westerfield M., Robinson P., Lewis S., Mungall C.: PhenoDigm: analyzing curated annotations to associate animal models with human diseases, *Database*, vol. 2013, p. bat025, 2013.

[44] Smith C.L., Goldsmith C.A., Eppig J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, *Genome Biology*, vol. 6(R7), pp. R7.1–R7.9, 2004.

[45] Wickelmaier F.: An introduction to MDS. In: *Sound Quality Research Unit*, Aalborg University, Denmark, vol. 46, 2003.

[46] Yandell M., Huff C., Hu H., Singleton M., Moore B., Xing J., Jorde L.B., Reese M.G.: A probabilistic disease-gene finder for personal genomes, *Genome Research*, vol. 21(9), pp. 1529–1542, 2011.

[47] Zemojtel T., Köhler S., Mackenroth L., Jäger M., Hecht J., Krawitz P., Graul-Neumann L., Doelken S., Ehmke N., Spielmann M., Oien N.C., Schweiger M.R., Krüger U., Frommer G., Fischer B., Kornak U., Flöttmann R., Ardeshirdavani A., Moreau Y., Lewis S.E., Haendel M., Smedley D., Horn D., Mundlos S., Robinson P.N.: Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome, *Science Translational Medicine*, vol. 6(252), pp. 252ra123, 2014.

## Affiliations

**Emilia Zawadzka-Gosk**

Polish-Japanese Academy of Information Technology, Warsaw, Poland, e.zawadzka@pja.edu.pl

**Krzysztof Wołk**

Polish-Japanese Academy of Information Technology, Warsaw, Poland, kwolk@pja.edu.pl

**Received**: 30.01.2017
**Revised**: 01.04.2018
**Accepted**: 03.04.2018