How to identify pathogenic mutations among all those variations: Variant annotation and filtration in the genome sequencing era

*David Salgado[1, *¶], Matthew I. Bellgard[2,3], Jean-Pierre Desvignes[1] and Christophe Béroud[1,4]*

[1] Aix Marseille Univ, INSERM, GMGF, Marseille, France; [2] Centre for Comparative Genomics, Murdoch University, Perth, Western Australia; [3] Western Australian Neuroscience Research Institute, Perth, Western Australia; [4] APHM, Hôpital TIMONE Enfants, Laboratoire de Génétique Moléculaire, 13385, Marseille, France.

* Corresponding author

E-mail: david.salgado@univ-amu.fr

Abstract

High-throughput sequencing technologies have become fundamental for the identification of disease-causing mutations in human genetic diseases both in research and clinical testing contexts.

The cumulative number of genes linked to rare diseases is now close to 3,500 with more than 1,000 genes identified between 2010 and 2014 thanks to the early adoption of Exome Sequencing technologies. However, despite these encouraging figures, the success rate of clinical exome diagnosis remains low due to several factors including wrong variant annotation and non-optimal filtration practices which may lead to misinterpretation of disease-causing mutations.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23110.

In this review, we describe the critical steps of variant annotation and filtration processes to highlight a handful of potential disease-causing mutations for downstream analysis. We report the key annotation elements to gather at multiple levels for each mutation, and which systems are designed to help in collecting this mandatory information. We describe the filtration options, their efficiency and limits and provide a generic filtration workflow and highlight potential pitfalls through a use case.

Keywords

high-throughput sequencing, variant annotation, variant filtration, good practices, pathogenic mutation

1. Introduction

The identification of human disease-causing mutations has relied for decades on Sanger sequencing and pre-screening technologies such as Single Strand Conformational Polymorphism (SSCP) or Denaturating Gradient Gel Electrophoresis (DGGE). This process was very slow and costly especially when genetic heterogeneity occurred. More importantly, it was difficult to apply this approach to large genes such as *DMD* (79 exons) or *TTN* (363 exons). With the rapidly developing high throughput solid phase sequencing technologies also known as Next Generation Sequencing (NGS), the strategy of gene hunting and diagnosis has drastically improved. In fact, in less than ten years, these NGS technologies have moved from gene panel sequencing (100 Mb for the Roche GS FLX system) to whole genome sequencing (1500 Gb for the Illumina HiSeq4000) and from research context only to clinical practice. The limitation is no longer the sequencing of one, many or all genes, but rather the sequence analysis and interpretation. Traditionally, scientists were afforded the luxury to develop expertise in a limited number of disease genes over a significant period of time. Unfortunately, they are now facing the daunting "all genes data deluge" (Schatz and Langmead

2013) where the expectation is to understand and interpret the suite of genes and their network of interactions implicated in disease. This next generation sequencing revolution now relies heavily on the field of bioinformatics, its tools, methods and analysis strategies to gather, store, analyze and mine the data in order to make informed decisions.

Despite the tens of thousands of exomes and genomes that have been studied (for instance, Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org), we have only a limited understanding of the molecular mechanisms underpinning the human genome variability, especially in the context of rare human genetic diseases (see article from Collod-Béroud et al. in this issue). In fact, most disease-causing mutations are private (specific to a family) and the availability of functional tests to demonstrate their pathogenicity is limited. As such, distinguishing neutral mutations from disease-causing ones is challenging. This is even amplified for rare diseases, which are defined in Europe as conditions with a frequency below 1:2000 (Regulation (EC) N°141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products, http://ec.europa.eu/health/files/eudralex/vol- 1/reg_2000_141/reg_2000_141_en.pdf). It is estimated that more than 7,000 rare human genetic diseases exist (https://globalgenes.org/rare-diseases-facts-statistics/), most of them being very rare. A review from Orphanet (http://orphadata.org/data/xml/en_product2_prev.xml) revealed that the majority consist of a handful of published reports describing a few individuals with a previously unidentified genetic syndrome, see Figure 1.

Despite this apparent low number of affected individuals, it is estimated that all together the rare diseases account for up to 6-8% of the global population having a strong socio-economic impact (http://www.ema.europa.eu). To diagnose most of these rare diseases by 2020, the International Rare Diseases Research Consortium (IRDiRC) was launched in April 2011. It supports international collaboration and data sharing (Thompson et al. 2014) as well as large scale sequencing projects

(Turner et al. 2015). At the time of writing this manuscript, the cumulative number of new genes linked to rare diseases was 3,350 with 1,000 identified between 2010 and 2014, while 350 new rare diseases were described during this period (http://irdirc.org). Most of these genes have been identified as a result of both the revolution of NGS and advanced bioinformatics techniques.

In contrast, this success is tempered by clinical exome diagnosis which has a success rate of only 26% (Yang et al. 2013). This relatively low success rate may be due to a number of factors, namely: a) technical limitations such as the absence of the disease-causing mutation in the captured DNA; b) a poor capture of some exonic regions (GC-reach regions); c) a low sequencing depth; c) a poor sequencing quality of read extremities; d) a mutation type not compatible with NGS technologies such as triplet expansion or large structural genomic variation (Gilissen et al. 2012); e) limitation of the bioinformatics data analysis pipeline; f) presence of pseudogenes or repeated regions, which may lead to inadequate mapping and wrong calling of mutations (false positives); g) limitation of the mapping process as no "gold standard" exist and a compromise has to be made between speed and accuracy, or h) wrong annotation in databases which may lead to misinterpretation of a disease-causing mutation.

The objective of this paper is to review the critical steps of variant annotations and filtration in order to guide users to collect the most appropriate elements related to each mutation and apply the proper filtration options to rapidly select a handful of candidate disease-causing mutations for downstream validation.

2. Variants annotation

The variant annotation process places mutations identified by the variant calling step (see Beltran et al. in this issue) into their biological context. This step is a requirement for the identification of

variants of interest based upon a combined filtration of the collected data from one or multiple samples.

The main objective of this process is to gather substantial information at the variant and the gene levels. This will include the variants' data quality, their localization at the genomic, gene and transcripts levels, their genotype, their frequency in the general population, their impact at the mRNA and protein levels, the conservation among species of the affected protein residues, the variant pathogenicity prediction and reported associations with diseases. At the gene level, they include the gene function, its spatiotemporal expression pattern, its involvement in various pathways and its involvement in various phenotypes/diseases.

2.1 Annotations at the variant level

Currently, several methods are available for variant quality assessment depending on the variant calling tool such as UnifiedGenotyper GATK and HaplotypeCaller GATK (McKenna et al. 2010), SamTools (Li et al. 2009) or Platypus (Rimmer et al. 2014). These bioinformatics tools provide two scores: i) the variant quality score or the probability that this variation is real. It is provided as a Phred quality score (Q score) (Ewing and Green 1998) to assess the probability that a given base is called incorrectly; and ii) the genotype quality score, which is also a Phred quality score to assess the probability that the given genotype is incorrect.

The description of the localization of the mutation includes various stages. The first stage is the genomic coordinates of the variation and is dependent of the version of Human Genome assembly, currently GRCh37 or GRCh38. The second stage is the localization at the gene and transcripts level. It is dependent of the selected annotation of the human reference genome. It is provided by Ensembl, the University of California Santa Cruz (UCSC), or the National Center for Biotechnology Information (NCBI). As demonstrated by Zhao et al. these annotations strongly differ at the transcript level with

only 53% of junction reads mapping at the same genomic location depending of the used gene model (Zhao and Zhang 2015). It is thus usually recommended for protein coding genes to use the Consensus Coding Sequence (CCDS), which is the result of a collaborative effort to maintain a dataset of protein-coding regions that are identically annotated on the human and mouse reference genome assemblies. CCDS are consistently represented by the NCBI, Ensembl, and UCSC Genome Browsers (Pruitt et al. 2009). When combining data from different sources, it is highly recommended to use annotations performed using the same reference genome.

Once the genomic coordinates have been determined, the HGVS nomenclature is usually used to name the mutations at the cDNA level for all transcripts and protein level for coding genes. Once again, the translation from a genomic nomenclature to a cDNA nomenclature may result in differences based on the variant caller and the annotation tool. This is especially true for insertions and deletions in a repeated sequence. It may thus be interesting to control mutations nomenclature after the initial annotation step using a system able to correct such errors. One such tool is the Variant Effect Predictor from Ensembl (VEP) (Yates et al. 2015). In an attempt to evaluate the impact of functional annotation using various reference systems and tools, McCarthy et al. quantified the extent of differences in annotation of 80 million variants from a whole-genome sequencing study. They compared results from the ANNOVAR and VEP software using REFSEQ and ENSEMBL transcripts. They reported only 44% agreement in annotations for putative loss-of-function variants using ANNOVAR. They also support data from Zhao et al. (Zhao and Zhang 2015) showing that the splicing variants were the category with the greatest discrepancy. They concluded that the annotation step must be considered carefully, and that a conscious choice should be made to select transcript set and software for annotation (McCarthy et al. 2014).

Variant frequency

Another key annotation element is the frequency of the variant in the general population. If ideally this general population should match the sample's origin, it is usually not available and data is captured from large scale projects such as the 1000 genome project (1000 Genomes Project Consortium et al. 2015), dbSNP (Sherry et al. 2001), the EVS (http://evs.gs.washington.edu/EVS/) or the EXAC consortium (Lek et al. 2015). It is important to recognise that these datasets are not mutually exclusive as there is significant overlap and should therefore not be simply combined to extrapolate a global frequency. In addition, these datasets are not representative of a global population *per se* as they contain "assumed to be healthy" individuals and samples from selected individuals with a particular condition. Note that laboratories, which routinely perform high-though put sequencing usually build internal frequency databases to exclude potential artefacts.

Variant nomenclature and localization

The annotation of the impact of the DNA mutation at the mRNA and the protein levels is complex. If the consequence at the protein level is easy to predict and report using the HGVS nomenclature (p.), it is only a prediction and should be considered as such. For example, a frameshift mutation predicted to result in a premature termination codon and therefore to a shorter protein, usually does not exist in reality because of the nonsense mediated decay phenomenon (Miller and Pearce 2014). Another example is the prediction of a protein harboring a missense mutation that indeed does not exist as its primary impact is at the mRNA level as exemplified by the c.2167G>A (p.Asp723Asn) of the *FBN1* gene that lead to the exon 17 skipping (Evangelisti et al. 2010). The impact of mutations at the mRNA level is even less documented as it is annotated as splicing mutation if localized in the donor or acceptor splice site regions, while the splicing machinery recognizes many signals such as exonic splicing enhancers and silencers and is considered as one of the most complex process of the cell (Nilsen 2003). Only few annotation tools are now including data from the Human Splicing Finder system that provides predictions for the impact of any mutation on all splicing signals (Desmet et al. 2009). The ANNOVAR system provides splice sites effect predictions

by AdaBoost and Random Forest from dbscSNV (Wang et al. 2010). The VEP system integrates various modules allowing to run external algorithms such as MaxEntScan (Yeo and Burge 2004), GeneSplicer (Pertea et al. 2001) or the dbNSFP (Liu et al. 2016) for splice sites effect predictions.

Affected protein residue annotation

The next annotation for variations is the conservation of affected protein residues. This information is usually restricted to single nucleotide variations that may result in missense mutations. It could be used as the result of selection pressure to maintain a specific amino acid at a given position because of its importance for the structure or the function of the protein. Typically, the higher the conservation, the higher the probability that a missense could impact the protein function. Most annotators are using conservation data extracted from the dbNSFP (Liu et al. 2016) that colligate data from PhyloP (Pollard et al. 2010), PhastCons (Siepel et al. 2006), GERP++ (Davydov et al. 2010) and Siphy (Garber et al. 2009). This gives access to conservation data from 27, 46 and 100 species, respectively, when using the GRCh37/38 reference genome.

Variant pathogenicity

Variants are also annotated for their potential pathogenicity using multiple algorithms and systems. Most annotator systems provide access to the following predictors: SIFT (Sim et al. 2012) Polyphen2 (Adzhubei et al. 2010), LRT (Chun and Fay 2009), MutationTaster (Schwarz et al. 2014), Mutation Assessor (Reva et al. 2011), FATHMM (Shihab et al. 2013), MetaSVM and MetaLR (Dong et al. 2015), CADD (Kircher et al. 2014), VEST3 (Carter et al. 2013), PROVEAN (Choi et al. 2012), fitCons (Gulko et al. 2015), fathmm-MKL (Shihab et al. 2015), and DANN (Quang et al. 2015). Only the VarAFT annotator (http://varaft.eu) provides annotations from the most efficient predictor for cDNA substitutions: UMD-Predictor (Salgado et al. 2016).

The final annotation at the variant level corresponds to its association with diseases. This information is usually extracted from various databases such as ClinVar (Landrum et al. 2014),

COSMIC (Forbes et al. 2015), UNIPROT (UniProt Consortium 2014) or HGMD (Stenson et al. 2003). It is important to keep in mind that the quality of these associations is highly variable and that some resources are not freely available.

2.2 Annotations at the gene level

As many variations are new or lacking annotation, it is useful to access data at the gene level. Global functional information is provided by the Gene Ontology at the cellular component, molecular function and biological process levels (Gene Ontology Consortium 2015). Additionally, data from tissue expression can be obtained from the Genotype Tissue Expression resource (GTEx) (Carithers et al. 2015), the Gene Expression Atlas (Petryszak et al. 2016) and organism model databases such as the Mouse Genome Institute (Bult et al. 2016). Unfortunately, the ability to integrated this data is not available for most annotators and often requires developing specific plugins such as the VEP annotator GXA.pm plugin to gather data from the Gene Expression Atlas.

Another level of genes annotation is their involvement in various pathways, these data can be found in BioCarta (Nishimura 2001), the Pathway Interaction Database (PID) (Schaefer et al. 2009), the Reactome (Fabregat et al. 2016), the WikiPathways (Kutmon et al. 2016), and KEGG (Kanehisa et al. 2016). Despite their utility for gene hunting, the information captured in these tools are not automatically incorporated during the annotation process. They are only available through external links.

As for annotations at the variant level, it is important to capture the phenotypes associated to mutations from a particular gene. The Online Mendelian Inheritance in Man (OMIM) resource is providing such information (Amberger et al. 2014). Other resources such as ClinVar, HGMD might also be used as part of the annotation process.

2.3 Annotation software

Different types of annotation software are available either through command line, webservices or web interface. They require data in different format, most allowing direct annotation of VCF files. They are compatible with either a single or multiple release versions of the human reference genome and allow annotation of SNP, Indels and CNV for only a limited number. As discussed, no system is providing annotations at all levels (Table 1). To easily handle annotations at the variant and gene levels most of the systems rely on the dbNSFP database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants in the human genome (Liu et al. 2016).

3. Selection of potential disease-causing mutations

Once annotations at the gene and variant levels have been performed, the user needs to reduce the significant number of variations (usually in the tens of thousands) to a small, manageable number of putative candidate disease-causing mutations for further experimental validation. This filtering process is likely the most critical process of NGS analysis. The aim is to combine filtration criteria to exclude spurious variants by taking into account various parameters such as: the mode of inheritance, the disease frequency, the pathogenicity prediction of variations, the gene expression pattern, known relations between the observed phenotype and gene mutations, and so forth. Currently, as there is no gold standard describing an optimized filtration process for all situations, there are two options to proceed. Option one is to employ a semi-automatic prioritization systems such as eXtasy (Sifrim et al. 2013), OMIM Explorer (James et al. 2016) or Exomiser (Smedley et al. 2015). This approach is proven to be very useful in situations where the phenotype is clearly described (mostly available in clinical diagnostic context) using the proper ontology such as the Human Phenotype Ontology (HPO) (Köhler et al. 2014) or the disease name (Amberger et al. 2014), as well as for gene hunting for the most advanced systems such as Exomiser and Phenolyzer (Yang et al. 2015).

The alternate option is to adopt a fully manual prioritization procedure based on expert knowledge related to the disease phenotype and genes functions. This approach is greatly facilitated by the use of user friendly filtration tools such as VarAFT (http://varaft.eu), FM Filter (Akgün et al. 2016), VarSifter (Teer et al. 2012), ExomeSuite (Maranhao et al. 2014), wKGGSeq (Li et al. 2015) or customizable filtration tools that require advance knowledge in informatics as they must be used through command-line scripts such as ANNOVAR (Wang et al. 2010) or GEMINI (Paila et al. 2013).

The semi-automatic prioritization systems use different prioritization algorithms and data. As reported by Rehm et al., the choice of the filtering process may differ across case types and requires a high level of expertise in genetics and molecular biology (Rehm et al. 2013). We will here describe the various steps that may be combined for manual prioritization. Table 2 lists some filtration systems and available filtration options.

3.1 Mode of inheritance

The initial step is linked to the mode of inheritance allowing the selection of either homozygous, heterozygous or compound heterozygous mutations either inherited or *de novo*. This selection process is facilitated when multiple samples from the family (trio) or the patient (somatic events) are available. As reported by Farwell et al., the diagnostic rate is higher among families undergoing a trio Whole Exome Sequencing (WES) (37%) as compared to a singleton (21%) (Farwell et al. 2015).These data were confirmed by Sawyer et al., reporting WES success rates of 23% for singletons, 32% for sibling pairs, and 34% for families (Sawyer et al. 2016). For instance, in the case of recessive conditions, the disease-causing mutations mainly correspond to compound heterozygous mutations for non-consanguineous families except for situations where a mutations is frequent in the population such as the ΔF508 mutation (NM_000492.3: c.1521_1523del) of the *CFTR* gene (Alfonso-Sánchez et al. 2010) and to homozygous mutations for consanguineous families. The most advanced systems such as VarAFT allow supporting of these hypotheses in difficult missing data contexts where one parent might be not sufficiently covered in one specific region.

3.2 Mutation localization

The mutation localization is usually the second parameter. Variants from key genomic regions (exons, splice sites) are usually selected while variants from other genomic regions (3' and 5'UTR, intronic regions) are discarded. In fact, if pathogenic mutations have been reported in 3' and 5' UTR, they mainly correspond to trinucleotide repeat expansions as illustrated by Spinocerebellar Ataxia type12 (#604326) and Fragile X tremor/Ataxia syndrome (#300625) for 5'UTR and Spinocerebellar Ataxia type 8 (#608768), Myotonic Dystrophy 1 (#160900) or Huntington disease like 2 (#606438) for 3' UTR (Richards et al. 2013). Nevertheless, it is important to keep in mind that these repeat expansions are not captured by most NGS technologies such as Illumina or Proton that generate only short reads. Other variants from these regions may be subsequently analyzed when no meaningful result is obtained.

3.3 Mutation type

It is well recognized that not all mutation types might result in an equal effect on proteins and diseases. Thus, it is usually considered that nonsense mutations as well as frameshift mutations have a strong impact while at the other end of the spectrum, synonymous changes usually have no impact at the protein level but may affect mRNA maturation. If in the past, synonymous changes have been frequently filtered-out, they are today conserved during this selection process. It is recommended to remove only variants belonging to the "unknown" or "in-frame deletions and insertions" mutation types. Other variants will be excluded through other filtration process such as pathogenicity predictions and frequency.

3.4 Mutation frequency

As most genetic diseases are rare, mutation frequency could be used to exclude frequent variations (see paper from Collod et al. in this issue). Ideally, this information should be captured from a matched population. In practice it is rarely feasible and users rely on frequencies from 1000 genome project (1000 Genomes Project Consortium et al. 2015), dbSNP (Sherry et al. 2001), the EVS (http://evs.gs.washington.edu/EVS/) or the EXAC consortium (Lek et al. 2015) as mentioned in the annotation section. Additionally, some large scale sequencing analysis project, such as ExAC, allow at the same time to combine the frequency data with the level of coverage observed in order to provide an accurate reflection of a specific variant within a population. Depending on the mode of inheritance and the frequency of the disease, it is possible to calculate the theoretical threshold of the disease-causing mutation frequency under the assumption that all observed cases harbor a single mutation. For example, for an autosomal dominant disease with a frequency of 1:10000, the allele frequency threshold is 0.01% while for an autosomal recessive disease with the same frequency, the threshold is 1%. However, obtaining this information for genetically heterogeneous disorders with overlapping clinical phenotypes might be challenging.

3.5 Pathogenicity predictions

During the last few years, various systems have been developed to predict the pathogenicity of mutation from human genes. They contain predictions for synonymous and non-synonymous changes, mutations potentially affecting mRNA splicing motifs as well as regulatory regions including miRNA binding sites, transcription Factor binding sites (Boyle et al. 2012), chromatin states (Ward and Kellis 2016) and non-coding regions (Kircher et al. 2014; Ritchie et al. 2014). It is essential for a critical interpretation of results to understand the strengths and limitations of each system. For example, predictions of the impact of mutations on splice sites and branch points are very accurate,

while they are less efficient for splicing auxiliary splicing sequences (Desmet et al. 2010). Similarly, predictions of the pathogenicity of missense mutations could be performed with a wide range of systems with accuracy ranging from 72% to 85% on a dataset of 17,329 variants (Salgado et al. 2016). Some authors have proposed to integrate individual predictions into meta systems such as PON-P (Olatubosun et al. 2012) and Condell (González-Pérez and Lopez-Bigas 2011) but consensus is only achieved for a subset of mutations, which are often relatively easy to predict. Therefore, it is better to use a limited set of predictors for filtration rather than combining all predictions, which may result in many situations with discrepancies between predictors. For example the NM_022124.5:c.4488G>C (p.Gln1496His) mutation of the *CDH23* gene has been reported as a pathogenic mutation (Bolz et al. 2001). Only the CADD, UMD-Predictor and Mutation Taster systems predicted this variant as pathogenic while the SIFT, Polyphen2, Condel, Provean and Mutation Assessor systems predicted it as a non-pathogenic mutation. In Figure 1 a Venn diagram is presented highlighting the predictions from the most frequently used predictors on a subset of randomly chosen 5,000 variations from the Uniprot dataset (Salgado et al. 2016). The 6 predictors reach consensus for only 3074 (61.5%) of variants, while if using only the 2 most efficient systems (UMD-Predictor and CADD), consensus is achieved for 4275 (85.5%) variants.

3.6 Functional evidences

Numerous studies have resulted in functional annotations of genes in various species and a specific ontology has been developed to described these functional annotations: Gene Ontology (Gene Ontology Consortium 2015). In parallel, genes have been classified in various pathways (Nishimura 2001; Fabregat et al. 2016; Kanehisa et al. 2016) in order to capture relationships and facilitate gene hunting. Furthermore, spatiotemporal expression patterns have been established for many genes in

various species using a number of different high throughput techniques such as in situ hybridization (Tomancak et al. 2002), micro arrays (Petryszak et al. 2016) and RNA-seq (GTEx Consortium 2013). Unfortunately, despite the importance of these data they are usually not available for filtration.

3.7 Previous description in databases

As mentioned in the beginning of this section, the description of mutations might be helpful for variant prioritization. In addition to variants' frequency, the availability of curated and previously annotated data is of primary importance. Such data can be found in Locus Specific DataBases (LSDB) such as LOVD (Fokkema et al. 2005), UMD (Béroud et al. 2005) and others (http://www.hgvs.org/locus-specific-mutation-databases) or Core databases such as HGMD (Stenson et al. 2003), ClinVar (Landrum et al. 2014), OMIM (Amberger et al. 2014), Uniprot (UniProt Consortium 2014) or RDRF (Bellgard et al. 2014). These annotations are of different qualities due to different curation modes ranging from full curation by experts to direct submission without review. In this context it is important to do not consider annotations as definitive answers but rather as evidences of causality. These data are usually not available for filtration but as additional annotations.

3.8 Proposed Filtration Flowcharts

Figure 3 proposes a standard filtration flowchart to identify disease-causing mutations in a context of a trio analysis for the recessive mode of inheritance. The samples were described by Kamphans et al. (Kamphans et al. 2013) and correspond to a family with one daughter (sample ID #464) affected by Mabry syndrome and her two healthy parents (samples ID #466 and 467). The flowchart is composed of 5 steps, however, the fifth step could be divided into functional evidences and previous description in databases. As illustrated, the availability of the trio allows a drastic reduction of the

candidate variations/genes during the first step where the compound heterozygous hypothesis has been selected. Alternative hypothesis such as homozygous mutation in the daughter with heterozygous parents for the mutation as well as more complex hypothesis where one of the two parents has insufficient sequencing quality resulting in missing data, should also be investigated in parallel (see below). In the second step, only mutations localized in exons and their vicinity are conserved. This parameter corresponds to the most frequent situation and can be adjusted on a case by case basis especially for genes where mutations in UTR regions have been reported with a high frequency. The third step uses frequency information from the 3 most popular databases (EVS, 1000 genomes and ExAC) even if they overlap (see above). The frequency threshold can be adjusted if the disease frequency is known or arbitrarily fixed to 1% in case of a recessive condition. For the fourth step, we selected only predictions from the UMD-Predictor and CADD systems. Based on user experience, the selection of predictors might vary. Note that in this use case, one of the two disease-causing mutations from the *PIGO* gene is a missense variation (NM_032634, c.2869C>T, p.Leu957Phe) for which pathogenicity predictions are available, the second one being a frameshift deletion (NM_032634, c.2355dupC). It is important to note that in this case, the use of more predictions algorithms might have resulted in loss of this candidate gene as this mutation is predicted as being non-pathogenic by SIFT and Mutation Taster. The final step corresponds to the collection of additional evidence. Here only two candidate genes with compound heterozygous mutations were present: the *AFF1* and the *PIGO* genes. After collection of evidences from OMIM, the *PIGO* gene can be selected as the only hypothesis compatible with the phenotype.

In this use case the disease-causing mutations were efficiently captured which is only true for a limited number of situations as reported by the clinical exome diagnosis success rate of only 26% (Yang et al. 2013). For the negative cases it is important to consider the following hypothesis.

Hypothesis 1: poor sequence quality in one sample leading to missing data. Only few filtration tools allow the handling of missing data and might lead to candidate exclusion.

Hypothesis 2: wrong genotype. For instance, considering only compound heterozygous hypothesis in a case of recessive inheritance without consanguinity. The homozygous hypothesis should always be evaluated as well as the compound heterozygous situation with one *de novo* mutation. The recently released TADA (Transmission And De novo Association test) model is a Bayesian model that combines data from *de novo* mutations, inherited variants and standing variants in the population. This approach revealed a significant power increase for gene discovery, as demonstrated through the studies of exome data of Autism Spectrum Disorder (ASD) (He et al. 2013) and might be useful in other situations.

Hypothesis 3: involvement of a large rearrangement or Copy Number Variation (CNV). It is recognized that CNV are frequently involved in human genetic diseases with the archetype of Duchenne Muscular Dystrophy where they account for up to 60% of mutations (Bladen et al. 2015). These CNV might now be captured from WES data and various tools are available (Nam et al. 2016). It is therefore important to combine this information with SNV.

Hypothesis 4: some variations could be missed by the WES technologies because they are localized outside captured regions (deep intronic, regulatory regions).

Hypothesis 5: bioinformatics pipelines limitations. The alignment process could lead to wrong reads' mapping because of high homology between various genomic regions. This might end up with wrong variant calling resulting in false positives or false negatives. In addition, repeated sequences can not be aligned and prevent mutation identification as it is the case for trinucleotide repeats, which are responsible for many human diseases (Keogh and Chinnery 2013).

4. Discussion

High throughput sequencing technologies that include Whole Exome Sequencing generate a high number of sequence variations in all individuals. Identifying disease-causing mutations among this large amount of data is a significant challenge. In this paper, we described the annotation and filtration steps that are mandatory to rapidly end up with a handful of disease-causing candidate mutations for further analysis. Even if the success rate is still limited, these technologies have a strong potential to diagnose most human monogenic diseases. To do so, it is first critical to understand the advantages as well as the limits of the annotation and filtration systems as incorrect or incomplete annotations can cause scientists both to overlook potentially disease-relevant DNA mutations as well as potentially dilute interesting mutations into a pool of false positives. Additionally, the filtration process requires experts in order to efficiently define hypothesis and successfully apply filtration tools. Finally, a close collaboration with clinicians is also a pre-requisite to avoid misclassification of patients that often prevent the disease-causing mutation discovery as experienced by many research teams.

Despite the availability of many bioinformatics systems for annotation and filtration, there is no gold standard available to solve every situation. Some semi-automatic prioritization systems taking into account genotypes and phenotypes are now available and could be of interest for specific situations (Sifrim et al. 2013; Smedley et al. 2015; James et al. 2016). However, most users typically use a combination of annotation and filtration systems. This can be done manually through dedicated systems (Teer et al. 2012; Maranhao et al. 2014; Li et al. 2015; Akgün et al. 2016) or using frameworks such as the Galaxy (Goecks et al. 2010) and Yabi (Hunter et al. 2012). The latter is flexible enough to allow any combination of software tools and data into sophisticated analysis procedures. In addition, the generated workflows can be easily shared and adjusted. Despite these strong benefits, these systems require not only bioinformatics skills but also a full understanding of the parameters of each step to reach efficiency and are therefore not fully adopted.

With the future switch to whole genome sequencing, additional challenges will emerge such as the handling of very large datasets and the interpretation of new mutation types: i) deep intronic mutations; ii) regulatory region mutations; iii) mutations found in non-coding genes (lncRNA (Wapinski and Chang 2011), T-UCR, circular ncRNA, small nucleolar RNA and miRNA (Esteller 2011) and iv) mutations reported in extra-genic regions (Dickel et al. 2013). As reported by Berg et al., a significant obstacle to implementing WGS is the huge amount of information that will be generated even if they consider that only a small subset might be relevant for interpretation due to a lack of knowledge (Berg et al. 2011). Nevertheless, it is anticipated that with the global adoption of WGS, more data will be generated and will contribute to the development of new bioinformatics tools and systems to facilitate their interpretation. If people often consider sequencing costs as the major barrier for WGS adoption, it is important to not only consider these costs but also human resources required for data interpretation. In fact, human resource needs for full clinical interpretation of WGS data remain considerable as described by Dewey et al. who reported that approximately 100 variants should be manually evaluated in each patient and that candidate disease causing mutation curation required at least one hour per variant (Dewey et al. 2014).

It is reasonable to believe that during this WGS progressive adoption phase, most users will first benefit from better exonic regions coverage when compared to WES (Belkadi et al. 2015). In this situation they might directly benefit from all annotation and filtration systems developed for WES data analysis and described here before the availability of innovative annotation and filtration systems for non-coding mutations.

Acknowledgments

The authors gratefully acknowledge the combined support-in-part funding for this work. This includes the financial support of RD-Connect-European Union Seventh Framework Programme (FP7/2007–2013 program HEALTH. 2012.2. 1.1-1-C) under grant agreement number 305444: RD Connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research, the Australian National Health and Medical Research Council (APP1055319) under the NHMRC–European Union Collaborative Research Grants scheme, Bioplatforms Australia Pty. Ltd. funded through the National Collaborative Research Infrastructure Strategy Initiative.

The authors have no conflicts of interest to declare.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Meth 7: 248–249.

Akgün M, Faruk Gerdan Ö, Görmez Z, Demirci H. 2016. FMFilter: A fast model based variant filtering tool. Journal of Biomedical Informatics.

Alfonso-Sánchez MA, Pérez-Miranda AM, García-Obregón S, Peña JA. 2010. An evolutionary approach to the high frequency of the Delta F508 CFTR mutation in European populations. Medical hypotheses 74: 989–992.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2014. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res.

Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova J-L, Abel L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U.S.A. 112: 5473–5478.

Bellgard MI, Render L, Radochonski M, Hunter A. 2014. Second generation registry framework. Source Code Biol Med 9: 14.

Berg JS, Khoury MJ, Evans JP. 2011. Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time. Genetics in Medicine 13: 499–504.

Béroud C, Hamroun D, Collod-Béroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum. Mutat. 26: 184–191.

Bladen CL, Salgado D, Monges S, Foncuberta ME, Kekou K, Kosma K, Dawkins H, Lamont L, Roy AJ, Chamova T, Guergueltcheva V, Chan S, et al. 2015. The TREAT-NMD DMD Global database: Analysis of More Than 7000 Duchenne Muscular Dystrophy Mutations. Hum. Mutat.

Bolz H, Brederlow von B, Ramírez A, Bryda EC, Kutsche K, Nothwang HG, Seeliger M, del C-Salcedó Cabrera M, Vila MC, Molina OP, Gal A, Kubisch C. 2001. Mutation of CDH23, encoding a new member of the cadherin gene family, causes Usher syndrome type 1D. Nature Genetics 27: 108–112.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22: 1790–1797.

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group. 2016. Mouse genome database 2016. Nucleic Acids Res. 44: D840–7.

Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, et al. 2015. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation and biobanking 13: 311–319.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3: S3.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7: e46688.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19: 1553–1561.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6: e1001025.

Desmet FO, Hamroun D, Collod-Beroud G, Claustres M. 2010. Bioinformatics identification of splice site signals and prediction of mutation effects. Recent advances in nucleic ….

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 37: e67–e67.

Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP, Pakdaman N, Ormond KE, et al. 2014. Clinical Interpretation and Implications of Whole-Genome Sequencing. JAMA 311: 1035–1045.

Dickel DE, Visel A, Pennacchio LA. 2013. Functional anatomy of distant-acting mammalian enhancers. Philosophical Transactions of the Royal Society B-Biological Sciences 368:.

Dong C, Wei P, Jian X, Gibbs R. 2015. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human molecular ….

Esteller M. 2011. Non-coding RNAs in human disease. Nature Reviews. Genetics 12: 861–874.

Evangelisti L, Lucarini L, Attanasio M, Lapini I, Giusti B, Porciani C, Gensini GF, Abbate R, Pepe G. 2010. A single heterozygous nucleotide substitution displays two different altered mechanisms in the FBN1 gene of five Italian Marfan patients. European journal of medical genetics 53: 299–302.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, et al. 2016. The Reactome pathway Knowledgebase. Nucleic Acids Res. 44: D481–7.

Farwell KD, Shahmirzadi L, El-Khechen D, Powis Z, Chao EC, Tippin Davis B, Baxter RM, Zeng W, Mroske C, Parra MC, Gandomi SK, Lu I, et al. 2015. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genetics in medicine : official journal of the American College of Medical Genetics 17: 578–586.

Fokkema IFAC, Dunnen den JT, Taschner PEM. 2005. LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum. Mutat. 26: 63–68.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, et al. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43: D805–11.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25: i54–62.

Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43: D1049–56.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. 20: 490–497.

Goecks J, Nekrutenko A, Taylor J, Team G. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11:.

González-Pérez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. 88: 440–

449.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45: 580–585.

Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nature Genetics 47: 276–283.

He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, et al. 2013. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet 9: e1003671.

Hunter AA, Macgregor AB, Szabo TO, Wellington CA, Bellgard MI. 2012. Yabi: An online research environment for grid, high performance and cloud computing. Source Code Biol Med 7: 1.

James RA, Campbell IM, Chen ES, Boone PM, Rao MA, Bainbridge MN, Lupski JR, Yang Y, Eng CM, Posey JE, Shaw CA. 2016. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Medicine 8: 13.

Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, Robinson PN, Parkhomchuk D, Krawitz PM. 2013. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS ONE 8: e70151.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44: D457–62.

Keogh MJ, Chinnery PF. 2013. Next generation sequencing for neurological diseases: new hope or new hype? Clinical neurology and neurosurgery 115: 948–953.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 1–8.

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, et al. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 42: D966–74.

Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Mélius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, et al. 2016. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 44: D488–94.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42: D980–5.

Lek M, Karczewski K, Minikel E, Samocha K, Banks E. 2015. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv.

Li H, Li H, Handsaker B, Handsaker B, Wysoker A, Wysoker A, Fennell T, Fennell T, Ruan J, Ruan J,

Homer N, Homer N, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li MJ, Deng J, Wang P, Yang W, Ho SL, Sham PC, Wang J, Li M. 2015. wKGGSeq: A Comprehensive Strategy-Based and Disease-Targeted Online Framework to Facilitate Exome Sequencing Studies of Inherited Disorders. Hum. Mutat. 36: 496–503.

Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. 37: 235–241.

Maranhao B, Biswas P, Duncan JL, Branham KE, Silva GA, Naeem MA, Khan SN, Riazuddin S, Hejtmancik JF, Heckenlively JR, Riazuddin SA, Lee PL, et al. 2014. exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels. Genomics 103: 169–176.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier J-B, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine 6: 26.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Miller JN, Pearce DA. 2014. Nonsense-mediated decay in genetic disease: friend or foe? Mutation research. Reviews in mutation research 762: 52–64.

Nam J-Y, Kim NKD, Kim SC, Joung J-G, Xi R, Lee S, Park PJ, Park W-Y. 2016. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics 17: 185–192.

Nilsen TW. 2003. The spliceosome: the most complex macromolecular machine in the cell? Bioessays.

Nishimura D. 2001. BioCarta. Biotech Software & Internet Report: The ….

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. Hum. Mutat. 33: 1166–1174.

Paila U, Chapman BA, Kirchner R, Quinlan AR. 2013. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput. Biol. 9: e1003153.

Pertea M, Lin X, Salzberg SL. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 29: 1185–1190.

Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AM-P, Jupp S, Koskinen S, Mannion O, Huerta L, et al. 2016. Expression Atlas update-an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 44: D746–52.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–121.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 19: 1316–1323.

Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31: 761–763.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E, Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. 2013. ACMG clinical laboratory standards for next-generation sequencing. Genetics in medicine : official journal of the American College of Medical Genetics 15: 733–747.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39: e118.

Richards RI, Samaraweera SE, van Eyk CL, O'Keefe LV, Suter CM. 2013. RNA pathogenesis via Toll-like receptor-activated inflammation in expanded repeat neurodegenerative diseases. Frontiers in molecular neuroscience 6: 25.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics 46: 912–918.

Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. Nat Meth 11: 294–296.

Salgado D, Desvignes J-P, Rai G, Blanchard A, Miltgen M, Pinard A, Lévy N, Collod-Béroud G, Béroud C. 2016. UMD-Predictor: a High Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution. Hum. Mutat. n–a–n–a.

Sawyer SL, Hartley T, Dyment DA, Beaulieu CL, Schwartzentruber J, Smith A, Bedford HM, Bernard G, Bernier FP, Brais B, Bulman DE, Warman Chardon J, et al. 2016. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clinical genetics 89: 275–284.

Schaefer CF, Anthony K, Krupa S, Buchoff J. 2009. PID: the pathway interaction database. Nucleic acids ….

Schatz MC, Langmead B. 2013. The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. IEEE spectrum 50: 26–33.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Meth 11: 361–362.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29: 308–311.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. 34: 57–65.

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31: 1536–1543.

Siepel A, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection. Research in Computational Molecular ….

Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y. 2013. eXtasy: variant prioritization by genomic data fusion. Nat Meth 10: 1083–1084.

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. 40: W452–7.

Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, et al. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature Protocols 10: 2004–2015.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. 21: 577–581.

Teer JK, Green ED, Mullikin JC, Biesecker LG. 2012. VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer. Bioinformatics 28: 599–600.

Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, Hansson MG, 't Hoen P-BA, Patrinos GP, Dawkins H, Ensini M, Zatloukal K, et al. 2014. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. 29 Suppl 3: S780–7.

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biology 3: research0088.1.

Turner C, Brice A, Bushby K, Riess O, Hanna M, van Ommen G, Muntoni F, Klockgether T, Wirth B, Lochmueller H, Timmerman V, Schoells L, et al. 2015. NeurOmics: EU-funded-omics research for diagnosis and therapy in rare neuromuscular and neurodegenerative diseases. Neuromuscular Disorders 25: S298–S299.

UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 42: D191–8.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. 38: e164.

Wapinski O, Chang HY. 2011. Long noncoding RNAs and human disease. Trends in cell biology 21: 354–361.

Ward LD, Kellis M. 2016. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 44: D877–81.

Yang H, Robinson PN, Wang K. 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Meth 12: 841–843.

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, et al. 2013. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med 369: 1502–1511.

Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, Ruffier M, Taylor K, Vullo A, Flicek P. 2015. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics 31: 143–145.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology 11: 377–394.

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16: 97.

Legends:

Figure 1: Distribution of rare diseases according to their estimated prevalence. Data were extracted from the rare disease epidemiological data from Orphadata (http://orphadata.org/data/xml/en_product2_prev.xml).

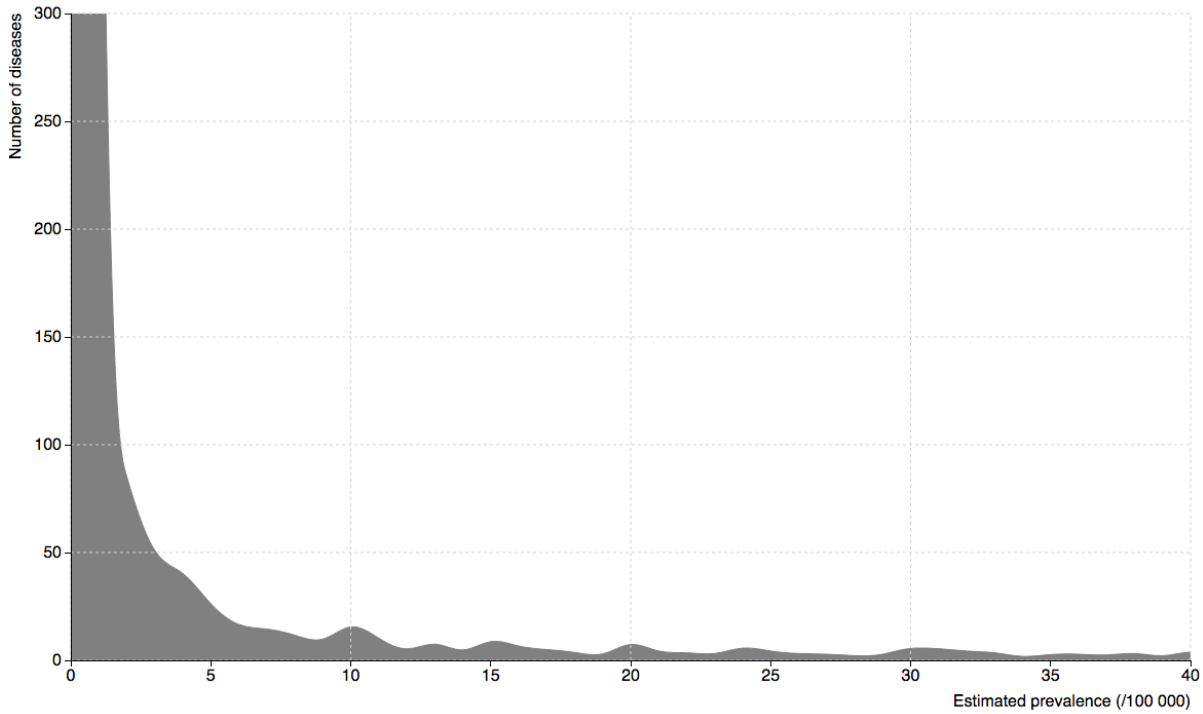

Figure 2: Venn diagram of pathogenicity predictions of 5,000 variants from Uniprot using CADD (Kircher et al. 2014), SIFT (Sim et al. 2012), Polyphen 2 (PPH2) (Adzhubei et al. 2010), Provean, Mutation Taster (MutTaster) (Schwarz et al. 2014) and UMD-Predictor (UMD-Pred) (Salgado et al. 2016) pathogenicity prediction systems.

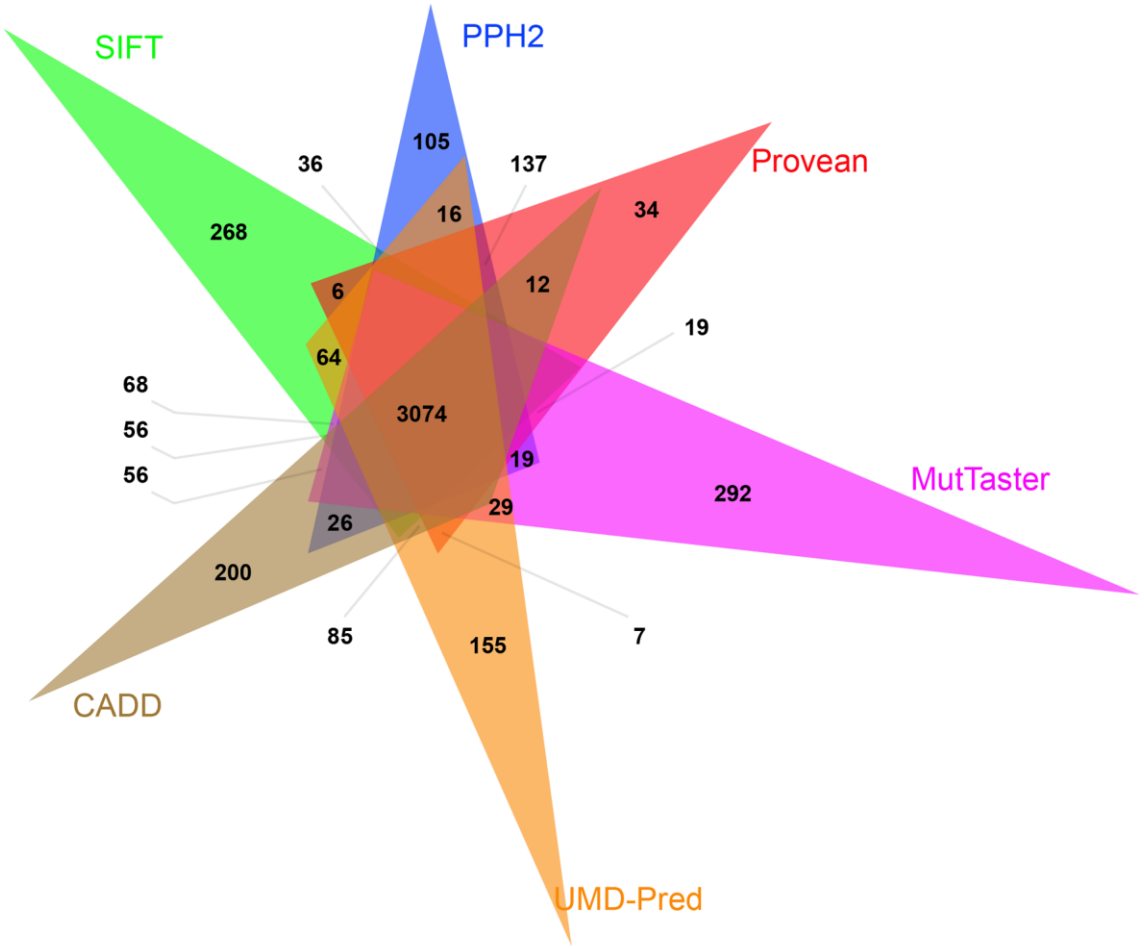

Figure 3: Filtration flowchart for a recessive disease from a trio (father, mother and affected daughter) (Kamphans et al. 2013). 1st step = mode of inheritance. Only genes with compound heterozygous mutations found in the daughter and transmitted by the two parents are selected; 2nd step = mutation localization. Only mutations present in the exons and intronic regions +/-8 nucleotides from the exon are conserved; 3rd step = frequency. Mutations with a reported frequency in ESP, 1000 genomes or ExAC above 1% are removed; 4th = predictions. Only mutations predicted as pathogenic or probably pathogenic by UMD-Predictor and CADD are conserved; 5th = other evidences. Genes of interest are analyzed using data from OMIM to select genes with a compatible impact on phenotype.

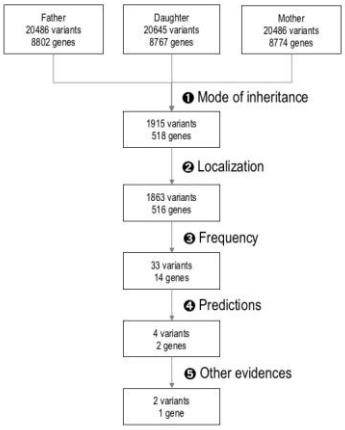

Table 1: Non-exhaustive list of annotation systems for WES. VCF = Variant Call Format; TSV = Tab-separated values.

		Annovar	SNPeff	Ensembl VEP	SeattleSeq	AnnTools	Oncotator	Vanno	Variant Annotation Tools
	Availability	Command line	Command line	Command line Webservice Web	web	command line	Command line Web	Web	Command line
	url	http://annovar.openbioinformatics.org/en/latest/	http://snpeff.sourceforge.net/	http://www.ensembl.org/info/docs/tools/vep/index.html	http://snp.gs.washington.edu/SeattleSeqAnnotation144/HelpAbout.jsp	http://anntools.sourceforge.net/	https://www.broadinstitute.org/oncotator/	http://cgts.cgu.edu.tw/vanno/	http://variantttools.sourceforge.net/Annotation/HomePage
	Input file	Multiple formats	VCF	Multiple formats	Multiple formats	Multiple formats	TSV	Multiple formats	Multiple formats
	Output	TSV	VCF	TSV	SeattleSeq, VCF	TSV	TSV	TSV	Multiple formats
	Genome version	Any	Any	GRCh37, GRCh38	GRCh37, GRCh38	GRCh37	GRCh37	GRCh38	GRCh36, GRCh37
	Annotation types	SNP, Indels, CNV	SNP, Indels	SNP, Indels	SNP, Indels	SNP, Indels, CNV	SNP, Indels	SNP, Indels	SNP, Indels
level	Variant quality	Yes	Yes	Yes	-	Yes	-	Yes	Yes

	Variant localisation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Gene/transcript annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Genotype	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes
	Population frequency	Yes	-	Yes	Yes	Yes	Yes	Yes	Yes
	Impact at the RNA level	Yes	Yes	Yes	-	-	-	-	-
	Impact at the protein level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Conservation	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes
	Reported impact	-	-	Yes	Yes	-	Yes	Yes	Yes
	Predicted pathogenicity	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes
Gene level	Gene ontology	-	-	-	-	-	Yes	Yes	-
	Pathways	-	-	-	Yes	-	-	Yes	Yes
	Tissue expression	-	-	-	-	-	-	-	-

Table 2: Non-exhaustive list of filtration systems for WES. VCF = Variant Call Format; TSV = Tab-separated values; ped file = pedigree file format; * = select variants based on the mutation genotype only; ** = only use mutation frequency from the "Born in Bradford sequencing project".

	Software name	availability	Input	inheritance	Recessive	Dominant	heterozygotes	de novo	analysis	localization	Mutation type	frequency	predictions	evidences	Clinical report	score
Manual filtration	**ANNOVAR**	Command line	VCF	-	-	-	-	-	-	-	-	Yes	Yes	-	Yes	-
	BIERApp	Web interface	VCF	Ye	Ye	Ye	Ye	Ye	Ye	Yes	Yes	Yes	-	-	-	-

				s	s	s	s	s	s							
	FILTUS	Standalone graphical user interface	VCF	Yes	Yes	Yes	Yes	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	-
	FMFilter	Standalone graphical user interface	VCF	Yes	Yes	Yes	Yes	Yes	-	Yes - if provided	Yes - if provided	Yes - if provided	-	-	-	-
	Gemini	Command line Web interface	VCF, ped file	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
	Vanno	Web interface	VCF	-	-	-	-	-	-	Yes	Yes	Yes	No but provided	Yes	Yes	-
	VarAFT	Standalone graphical user interface	VCF	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No - But provided	-
	VarSifter	Command line Web interface	VCF	Yes	Yes	Yes	Yes	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	-
	VCF-MINER	Local Web interface	VCF	Yes	Yes	Yes	Yes	Yes	Yes	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	Yes - if provided	-
Automatic	**ExomeWalker**	Web interface	VCF, ped	Yes	Yes	Yes	Yes	Yes	-	-	-	Yes	No but provi	No but provi	Yes	Yes

			file										ded	ded		
	Exomiser	Command line	VCF, ped file	Yes	Yes	Yes	Yes	Yes	-	-	-	Yes	Yes	-	Yes	Yes
	eXtasy	Command line Web interface	VCF	-	-	-	-	-	-	-	-	-	No but provided	No but provided	-	Yes
	MirTRIOS	Web interface	VCF, family file	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes	Yes	-	No - But provided	Yes
	OMIM Explorer	Web interface	VCF	Yes*	-	-	-	-	-	-	-	-	-	-	Yes	Yes
	OVA	Web interface	VCF	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes**	-	Yes	Yes	Yes
	wKG GSeq	Web interface	VCF, ped file	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes	Yes	Yes	Yes	Yes