# VarAFT: a variant annotation and filtration system for human next generation sequencing data

Jean-Pierre Desvignes, Marc Bartoli, Valérie Delague, Martin Krahn, Morgane Miltgen, Christophe Béroud, David Salgado

*Published online 31 May 2018* *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W545–W553***
*doi: 10.1093/nar/gky471*

# VarAFT: a variant annotation and filtration system for human next generation sequencing data

**Jean-Pierre Desvignes[1], Marc Bartoli[1], Valérie Delague[1], Martin Krahn[1,2], Morgane Miltgen[1], Christophe Béroud[1,2,*] and David Salgado[1,*]**

[1]Aix Marseille Univ, INSERM, MMG, 13005, Marseille, France and [2]APHM, Hôpital d'Enfants de la Timone, Département de Génétique Médicale et de Biologie Cellulaire, 13385 Marseille, France

Received January 29, 2018; Revised May 11, 2018; Editorial Decision May 14, 2018; Accepted May 16, 2018

## ABSTRACT

**With the rapidly developing high-throughput sequencing technologies known as next generation sequencing or NGS, our approach to gene hunting and diagnosis has drastically changed. In <10 years, these technologies have moved from gene panel to whole genome sequencing and from an exclusively research context to clinical practice. Today, the limit is not the sequencing of one, many or all genes but rather the data analysis. Consequently, the challenge is to rapidly and efficiently identify disease-causing mutations within millions of variants. To do so, we developed the VarAFT software to annotate and pinpoint human disease-causing mutations through access to multiple layers of information. VarAFT was designed both for research and clinical contexts and is accessible to all scientists, regardless of bioinformatics training. Data from multiple samples may be combined to address all Mendelian inheritance modes, cancers or population genetics. Optimized filtration parameters can be stored and re-applied to large datasets. In addition to classical annotations from dbNSFP, VarAFT contains unique features at the disease (OMIM), phenotypic (HPO), gene (Gene Ontology, pathways) and variation levels (predictions from UMD-Predictor and Human Splicing Finder) that can be combined to optimally select candidate pathogenic mutations. VarAFT is freely available at: http://varaft.eu.**

## INTRODUCTION

Massively parallel sequencing, also called NGS (next generation sequencing), led to a genetic revolution with the ability to sequence any human genome in a few hours. Nevertheless, despite the thousands of exomes and genomes that have been studied (Genome Aggregation Database (gnomAD) (1)), we still have only a limited vision of the human genome variability especially in the context of rare human genetic disease. Indeed, most disease-causing mutations are private, and the availability of functional tests is limited. Therefore, distinguishing neutral mutations from disease-causing ones is challenging. This is even more challenging for rare diseases, defined in Europe as conditions with a frequency below 1: 2000, most of them being very rare. A review from Orphanet (2), revealed that the majority of rare diseases are defined by a handful of published reports describing a few individuals with a previously unidentified genetic syndrome. It is now accepted that the limitation is no longer the sequencing of one, many or all genes but rather the data analysis. In addition, while scientists were previously experts for a limited number of genes, they are now facing the 'all genes data deluge'. This revolution has therefore resulted in a dependency on bioinformatics tools and methods to gather, store, analyze and mine the data flow. Indeed, NGS technologies typically result in the production of hundreds of millions to billions of reads per exome or genome, respectively. The analysis of these raw data can be divided into three steps as described by Gargis *et al.* (3). The primary analysis includes the production of sequence reads and assignment of base quality scores; the secondary analysis includes de-multiplexing, alignment of reads to a reference genome and variant calling; and the tertiary analysis is dedicated to the identification of disease-causing mutations. It involves the annotation and filtration of identified sequence variations. As reported by Salgado *et al.* (4) and Eilbeck *et al.* (5), the annotation includes various layers that should be combined in the filtration step to rapidly select a handful of candidate mutations. This filtration step can benefit from the combination of data from multiple samples as reported by Sawyer *et al.* (6) with Whole Exome Sequencing (WES) success rates ranging from 23% for singletons to 34% for families. To simplify this tedious process, various systems have been released such as QueryOR (7), VarElect (8), VCF-Miner (9) and BierApp (10). To annotate and prioritize mutations, these systems include mul-

*To whom correspondence should be addressed. David Salgado. Tel: +33 491324884; Fax: +33 491804319; Email: david.salgado@univ-amu.fr
Correspondence may also be addressed to Christophe Béroud. Tel: +33 491324488; Fax: +33 491804319; Email: christophe.beroud@inserm.fr

tiple annotations either captured through global systems such as ANNOVAR (11) and VEP (12) or individually retrieved. However, they only partially respond to users' needs and may require a preliminary annotation step performed by bioinformaticians. In addition, for web-based solutions, confidentiality issues may arise depending on national legislation (13). In this context, we designed a new system called VarAFT (Variant Annotation and Filtration Tool), that provides a full graphical interface and includes unique features to improve mutation annotation and prioritization. It combines classical data (phylogenetic, conservation and protein structures) with additional information at variant, gene and phenotype levels. In addition, it is one of the few systems able to combine small (single nucleotide variations, small insertion/deletions) and large rearrangements (copy number variations) to get a comprehensive picture of the individual genome.

With VarAFT, users can easily annotate, filter and perform breadth and depth of coverage analysis from their data without computer programming skills and with limited hardware requirements, to efficiently identify disease-causing mutations as demonstrated in various situations (14–21).

## MATERIALS AND METHODS

VarAFT is a freely available application written in Java and can therefore be used on most computers. Various binaries are available to download for Mac, Windows and Linux operating systems. VarAFT does not require any specific hardware configuration, however performances are dependent on the number of CPU cores and the amount of available memory. For example, the annotation of one sample containing 58,783 variations takes 19 min with 1 CPU core and drops to 9 min with 4 CPU cores. The breadth and depth of coverage analysis of the same sample takes 21 min with 1 CPU core and 7 min with 4. This time could be reduced to, respectively, 27 and 6 s if the user limits the breadth and depth of coverage analysis to the ACMG actionable genes list (22). Specific versions have been created to allow the installation of VarAFT on Windows machines without administrative rights. The graphical user interface was created using the Java Swing library.

The 'coverage module' use BEDTools (23) to compute breadth and depth of coverage data for any sequencing experiment using BAM files. Tables and charts are respectively generated with the Swing JTable library from Oracle (http://oracle.com) and the JFreeChart library (http://jfree.org). The breadth and depth of coverage analysis is performed at the genomic level unless a BED file is provided to limit the analysis to regions of interest.

The 'annotation module' can collect variant data from various file formats including VCF/gVCF (single or multi-samples) or tabulated files. It combines all information from the dbNSFP (24) and ANNOVAR with unique features including OMIM (25), HPO (26), Gene Ontology (27), pathways (Reactome (28), KEGG (29), PID (30)) and predictions from UMD-Predictor (31) and HSF (Human Splicing Finder) (32). Note that ANNOVAR, KEGG, UMD-Predictor and HSF require a user registration to comply with their license.

Once annotated, data from any sample can be combined through the 'filtration module'. It allows users to combine data from multiple sources and layers. The display mode can be parametered to display a subset of available columns. Interactive filtration features allow the progressive reduction of the list of candidate mutations by combining the various annotations. An in-house mutation database can be generated by VarAFT or provided by users, to exclude frequent mutations reported in a specific population and/or platform-dependent artefacts. Once filtration steps have been defined and validated, they can be saved, reapplied and shared for subsequent analysis to ensure filtration standardization in a clinical diagnosis context or for large research networks. At any filtration step, selected data can be exported for downstream analysis or reporting. Moreover, the quality of each selected mutation can be viewed in its sequencing context using IGV (33) directly from VarAFT.

## RESULTS

### A highly integrative system to easily pinpoint candidate disease-causing variants

As previously reported, the ability to efficiently filter genetic variation to select candidate disease-causing mutations is improved by combining data at the variant, gene and phenotypic levels (4,5). Although multiple information are available at each level, no system was able to collect and combine all this information (4). VarAFT was therefore designed to aggregate a substantially larger amount of information including the ability to combine small and large genetic variants (Table 1). In parallel, to simplify the combination of data from multiple samples, the user is able to combine samples according to predefined transmission modes, autosomal recessive or dominant, and to take into account pedigree structure (consanguinity, *de novo* mutations) (Figure 1). To accommodate other types of scenario, such as analysis of somatic mutations or population genetics, a custom module is also available.

### Experiment quality control compatible with clinical use

The 'coverage analysis module' was designed to evaluate experiment quality. It provides the breadth and depth of coverage for any transcript or exon at the nucleotide level, either through dynamic histograms or tables (Figure 2). A report can be generated to rank genes and exons according to their breadth of coverage at a depth of 1, 5, 10, 20 or 30× and evaluate their quality in accordance with international guidelines (EuroGentest: www.eurogentest.org). In a clinical diagnosis context, BED files can be provided or generated through VarAFT to restrict the exome analysis to some transcripts or genes. Indeed, as indicated in the Eurogentest guidelines, to limit incidental findings it is recommended to focus on genes of interest for which a relationship between genotype and phenotype has been published and confirmed (34).

### Use case demonstrating VarAFT efficiency in various situations

VarAFT has been extensively used in both clinical diagnosis and research contexts. Disease-causing mutations were

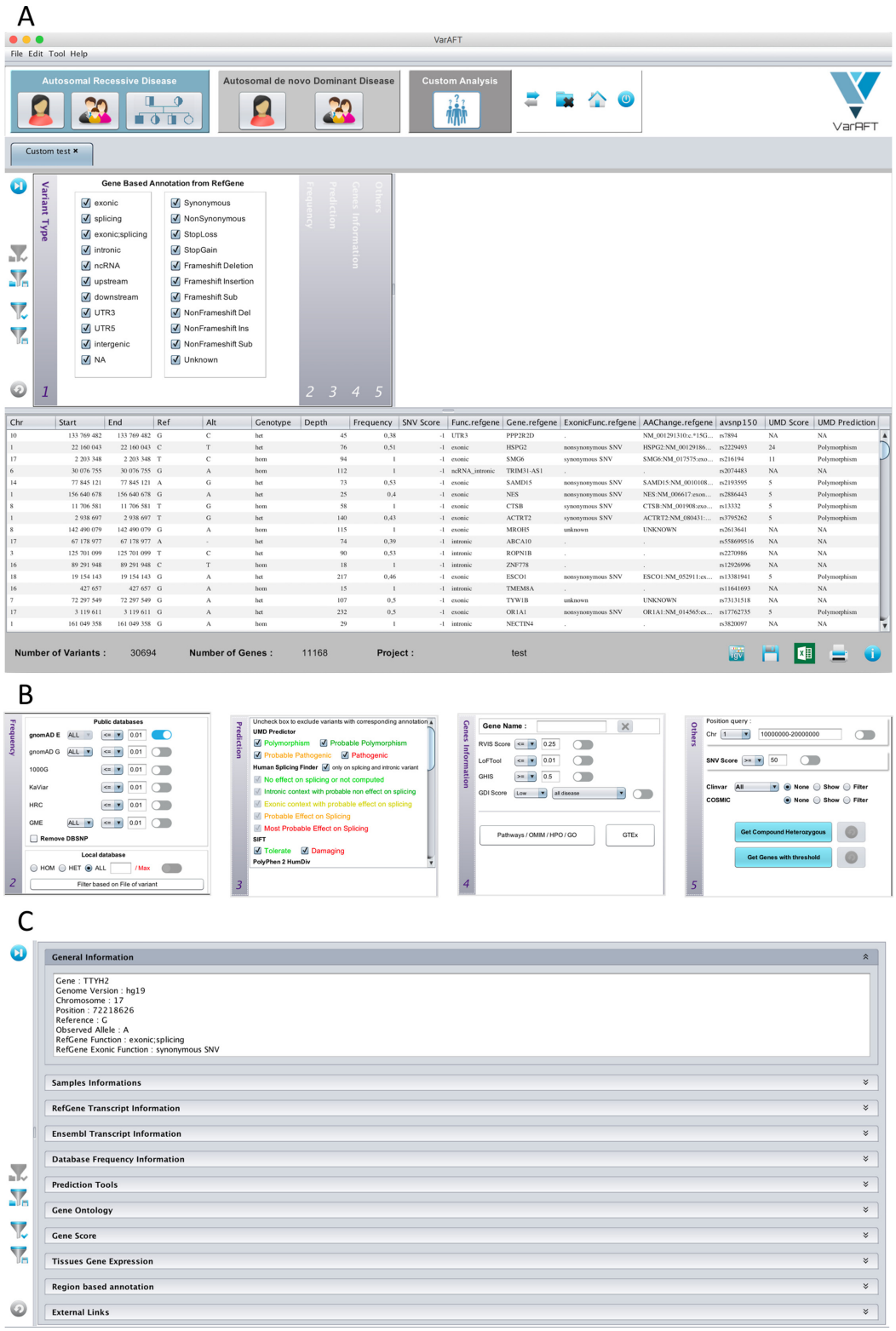


**Figure 1.** Analysis and Filtration module. (**A**) Top: main screen for filtration and analysis of variants. Top part: easy combination of samples (singleton, trio or any combination) for AD, ARD or other modes of inheritance. Middle: filtration parameters. Bottom: list of annotated variants. This list is dynamically updated based on filtration criteria. (**B**) Filtration criteria are divided into five sections (variant type, frequency, pathogenicity predictions, gene information and others) including multiple parameters that can be combined by the user. (**C**) The selection of one variant from the list gives access to additional data related to: general information linked to the variant itself, the presence of this variant in the analysed samples/patients, its impact on transcripts and related HGVS nomenclature (RefGene and Ensembl), its reported frequency in general populations, the prediction of its pathogenicity from multiple tools, the tolerance to loss of function of the related gene, the tissue expression pattern of the gene/transcripts, additional information related to chromosomal region as promotors, regulatory regions, etc. and access to various useful external websites.


**Figure 2.** Breadth and depth of coverage analysis of a WES experiment. (**A**) Top: main screen displaying the list of gene symbols and associated RefSeq transcripts, number of exons and coding sequence size with the following statistical values: mean depth ± standard deviation and breadth of coverage at 30×, 20×, 10×, 5× and 1× depth. Bottom-left: breadth of coverage for each exon of the selected transcript at a selected depth (20× for transcript NM_001244910 of the FCGR1B gene); Bottom-right: breadth of coverage at the nucleotide level of a selected exon (exon 1 of the NM_001244910 transcript). (**B**) Histogram displaying the percentage of transcripts with a breadth of coverage superior or equal to a depth of 1× (red), 5× (blue), 10× (green), 20× (yellow) and 30× (purple). (**C**) Breadth of coverage representation of the various exons from all transcripts of the selected gene (*MEN1*). Color code as for Figure 1A and C: red = <10× depth of coverage; yellow = 10–20× depth of coverage; blue ≥ 20× depth of coverage.

**Table 1.** Different annotations available within VarAFT and their availability for filtration

| LEVEL | ANNOTATION | FILTRATION |
|---|---|---|
| VARIANT | Localization | |
| | RefGene (Gene, Transcript, Function, Nomenclature) | X |
| | Ensembl (Gene, Transcript, Function, Nomenclature) | X |
| | Frequency | |
| | 1000 Genomes | X |
| | DbSNP | X |
| | Genome Aggregation Database (gnomAD) | X |
| | Known VARiants (KaViar) | X |
| | Haplotype Reference Consortium (HRCR1) | X |
| | Great Middle East Database (GME) | X |
| | Prediction | |
| | UMD-Predictor | X |
| | Human Splicing Finder | X |
| | SIFT | X |
| | Polyphen 2 | X |
| | LRT | X |
| | Mutation Taster | X |
| | Mutation Assessor | X |
| | FATHMM | |
| | PROVEAN | X |
| | VEST3 | |
| | MetaSVM | |
| | MetaLR | |
| | M-CAP | X |
| | CADD | X |
| | DANN | X |
| | FATHMM-MKL | |
| | Eigen | X |
| | GERP++ | X |
| | Conservation | |
| | phyloP100way | |
| | phyloP20way | |
| | phastCons100way | |
| | phastCons20way | |
| | SiPhy_29way | |
| | wgRNA | |
| | predicted microRNA targets (target ScanS) | |
| | genomicSuperDups | |
| | gwasCatalog | |
| | wgEncodeBroadHmmGm12878HMM | |
| GENE | Expression | |
| | GTEx | X |
| | Pathways/GO | |
| | KEGG | X |
| | Pathway Interactome Database (PID) | X |
| | REACTOME | X |
| | Gene Ontology | X |
| | Score Tolerance | |
| | Residual Variation Intolerance Score (RVIS) | X |
| | Gene Damaging Index (GDI) | X |
| | Lost Of Function Tool (LOF) | X |
| | genome-wide haploinsufficiency score (GHIS) | X |
| DISEASE PHENOTYPE | Ontology | |
| | Human Phenotype Ontology (HPO) | X |
| | Database | |
| | Online Mendelian Inheritance in Man (OMIM) | X |
| | Catalogue Of Somatic Mutations In Cancer (Cosmic) | X |
| | ClinVar | X |

X = annotation available for filtration.

identified from trio, cohorts and individual cases from autosomal dominant and recessive diseases, such as dystonia, neuromuscular disorders, mental retardation and premature aging. For example, VarAFT was recently used to analyze 306 genes in a cohort of distal myopathy patients (35). The software was also evaluated as a prioritization system to highlight mutations involved in cancers (14) and other situations (14–21).

To demonstrate VarAFT usefulness and efficiency, we chose the following use cases:

**Use case #1.** The first dataset was extracted from Kamphans *et al.* and contains VCF files from four individuals (464, 465, 466 and 467) (36) from a family with an autoso-

mal recessive disease. Each sample contained respectively 20 708, 20 560, 20 552 and 20 547 variants. The VarAFT processing for this family included the following five steps: (i) combination of data from the various samples taking into account the mode of inheritance (AR) to select compound heterozygous in the affected individuals (464 and 465) and present in only one parent (466 and 467); (ii) selection of variants localized in exons or bordering introns (±4 bp); (iii) exclusion of variants with a frequency in general populations above 1%; (iv) selection of variants predicted as pathogenic by CADD (37) and (v) selection of variants predicted as pathogenic by the UMD-Predictor system. Note that steps can be conducted in any order and will lead to the same result. The two remaining variants correspond to the two mutations, c.2869C>T (p.Leu957Phe) and c.2355dupC (p.Gly785fs), in the *PIGO* gene, identified by the authors as the disease-causing mutations in this family (Figure 3).

**Use case #2** was extracted from Miltgen *et al.* and contains VCF files from four patients from a multigenerational family from Flemish origin with Craniocervical Dystonia and an autosomic dominant mode of inheritance (20). Each sample contained, respectively, 66 215, 58 783, 58 959 and 59 495 variants. The VarAFT processing for this family included the following seven steps: (i) combination of data from the various samples taking into account the mode of inheritance (AD) to select heterozygous variants in the affected individuals (D7, D8 and D11) and absent in D10; (ii) selection of variants localized in exons or bordering introns (±4 bp); (iii) exclusion of variants with a frequency in general populations above 1%; (iv) selection of variants predicted as pathogenic by CADD; (v) selection of variants predicted as pathogenic by the UMD-Predictor system; (vi) selection of genes expressed in the affected tissue (brain) and (vii) selection of genes associated with at least one ('OR' option) HPO term describing symptoms found in patients (dystonia, Blepharospasm and torticollis). Note that steps can be conducted in any order and will lead to the same result. The two remaining variants correspond to the mutation c.240+1G>T from the *SURF1* gene and c.1969G>A (p.Ala657Thr) from the *ANO3* gene. The *SURF1* mutation was excluded as this gene is only involved in autosomal recessive diseases: Charcot-Marie-Tooth disease, type 4K (MIM #616624) and Leigh syndrome, due to COX IV deficiency (MIM #516000). In contrast, mutations from the *ANO3* gene have been reported in the autosomal dominant Dystonia 24 and this mutation was identified by the authors as the disease-causing mutations in this family (Figure 4).

**Use case #3** is an artificial VCF created by inserting the *SH3TC2* compound heterozygous c.[279G>A]; [805+2T>C] disease-causing mutations identified by Piscosquito *et al.*, 2016 (38) into a personal exome (https://personalgenomics.zone). The sample contained 37 694 variants. The VarAFT processing for this sample included the following five steps: (i) mode of inheritance (AR) to select only compound heterozygous variants; (ii) selection of variants localized in exons or bordering introns (±4 bp); (iii) exclusion of variants with a frequency in general populations above 1%; (iv) selection of variants predicted as pathogenic by the UMD-Predictor system; and (v) selection of genes associated with at least one ('OR' option) general HPO term associated with the pathology ('Decreased number of large peripheral myelinated nerve fibers' HP:0003387 or 'Segmental peripheral demyelination' HP:0007107). Note that steps can be conducted in any order and will lead to the same result. The two remaining variants correspond to the mutations c.2860C>T (p.Arg954*) and c.279G>A (p.Lys93Lys) from the *SH3TC2* gene and identified by the authors as the disease-causing mutations in this family (Figure 5). Note that this mutation was predicted as pathogenic only by the UMD-Predictor system as it impacts the donor splice site. This was confirmed by predictions from the HSF system.

As illustrated in the three use cases, VarAFT was flexible enough to apply optimal filtration criteria taking into account the mode of inheritance and the available phenotypic information. In each case, the process resulted in the identification of the disease-causing mutations (Figures 3–5). Only a subset of information was used in the processes and additional features (pathways, tissue expression, etc.) are available for more complex situations.

## DISCUSSION

VarAFT is a multiplatform freely available software that allows the simultaneous annotation, filtration, and breadth and depth of coverage analysis of WES, WGS and targeted sequencing experiments from any sequencing platform. Its graphical user interface, various modules and unique features, such as pathogenicity predictions from UMD-Predictor and HSF, allow untrained users to rapidly highlight disease-causing mutations in multiple genetic scenarios. In addition, VarAFT allows visualization of data quality (breadth and depth of coverage) through direct access to BAM files. The nucleotides and genotypes can also be easily accessed through IGV.

As reported by Salgado *et al.* (4), on one hand, automatic prioritization systems are now available to ensure a homogeneous treatment of samples. However, these systems are based on previously established links between genotype and phenotype and can only solve a limited number of diagnosis and research problems. On the other hand, manual systems are numerous and heterogeneous in their content and filtration features. VarAFT was tailored to overcome identified limitations such as the access to multiple layers of information, the combination of small (SNV) and large (CNV) mutations in a single analysis and the accessibility for all scientists, regardless of bioinformatics skill level. So, users can rapidly end up with shorter and more accurate lists of candidate disease-causing mutations, facilitating downstream validation, gene discovery and genetic counseling. As a standalone application, it can be used for clinical diagnosis as data are processed locally avoiding network privacy issues.

VarAFT uses recognized resources, formats and ontologies making it suitable for integration in any NGS environment. VarAFT was presented to the scientific community in various international training courses (RD-Connect, 3Gb-Test, Variant Effect Predictor training course from the Human Variome Project, ELIXIR training course on variants analysis) and was rapidly adopted by 800 users from more than 50 countries. At last, VarAFT was instrumental

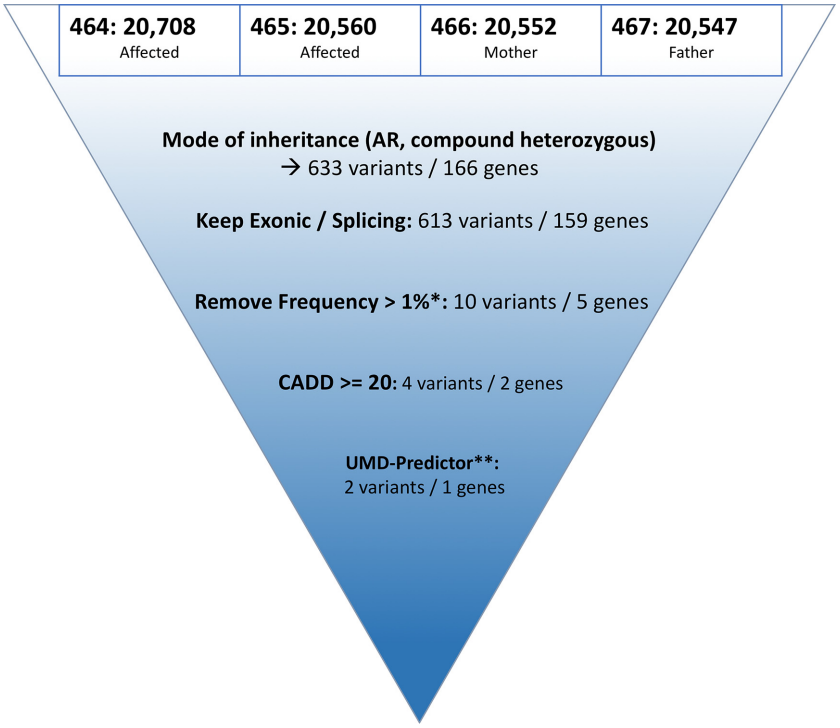


**Figure 3.** Use case#1. The identification of the disease-causing mutations from the AR family described by Kamphans *et al.* (36) including data from four members. It was performed in five steps using mainly data from the variant annotation layer. * gnomAD; 1000 genomes; KaViar; HRCR1 and GME databases; ** polymorphism and probably polymorphism were excluded.

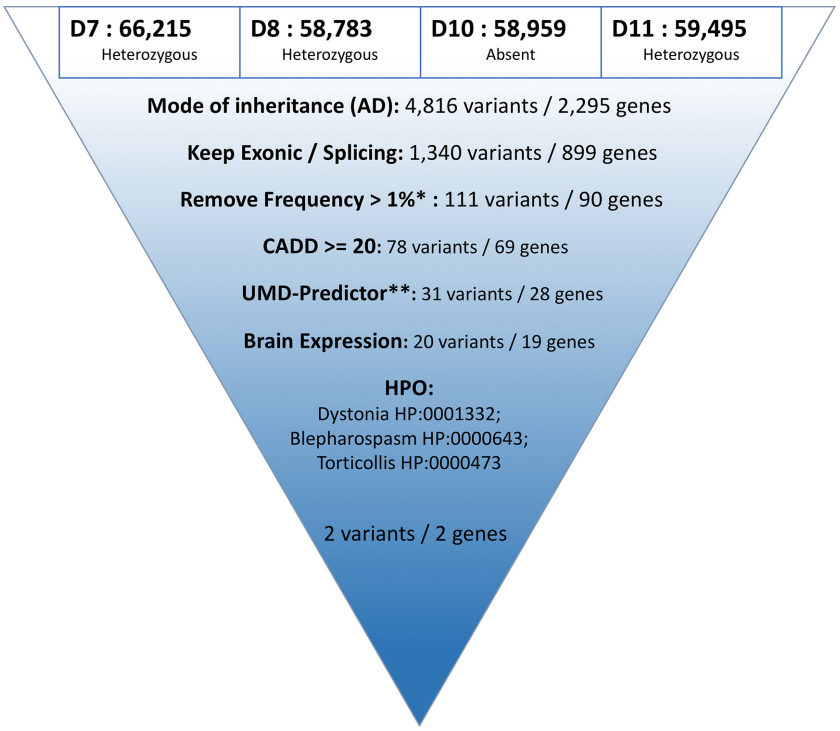


**Figure 4.** Use case#2. The identification of the disease-causing mutations from the AD family described by Miltgen *et al.* (20) including data from four members. It was performed in seven steps using data from the variant annotation and the phenotype layers. * gnomAD; 1000 genomes; KaViar; HRCR1 and GME databases; ** polymorphism and probably polymorphism were excluded.

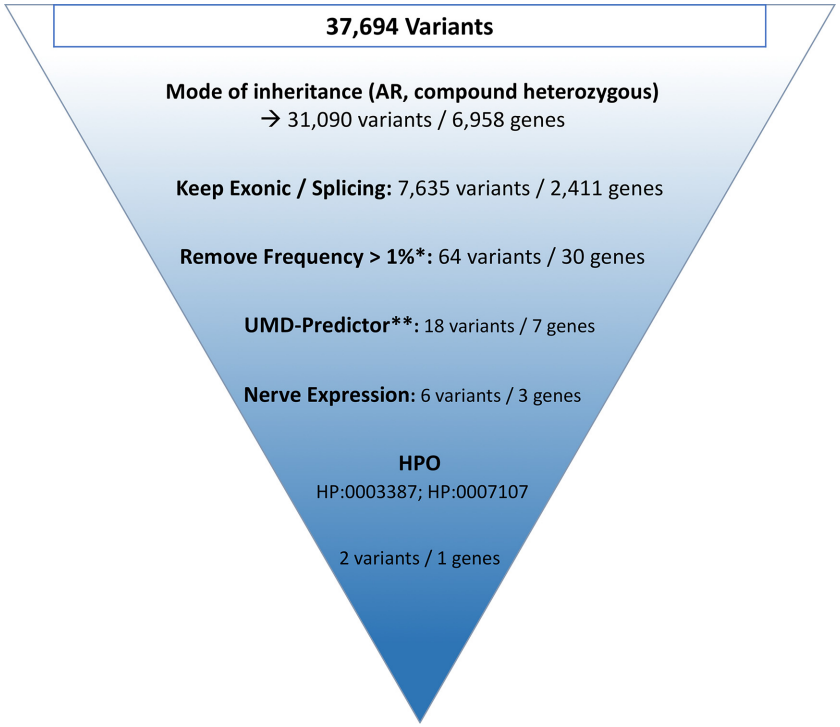


**Figure 5.** Use case#3. The identification of the disease-causing mutations from an artificial single VCF corresponding to a proband with an AR. It was performed in five steps using data from the variant annotation and the phenotype layers. * gnomAD; 1000 genomes; KaViar; HRCR1 and GME databases; ** polymorphism and probably polymorphism were excluded.

in the creation of the RD-Connect Genome-Phenome analysis platform (39).

## ACKNOWLEDGEMENTS

We are grateful to all users for their continuous support and feedback. We thank the members of the AP-HM Medical Genetics department, which were the first adopters of the system in a clinical diagnosis context and all members of the Marseille Medical Genetics research unit for useful input. We acknowledge our RD-Connect and NeurOmics partners for their adoption of the VarAFT system. Additionally, we thank the French ELIXIR Node, the Institut Français de Bioinformatique (IFB), for hosting a cloud appliance of the VarAFT system used for trainings. At last, we would like to thank Michael J. Mitchell for manuscript editing.

## FUNDING

European Union Seventh Frame-work Program (FP7/2007–2013) [305,444 (RD-Connect)]; Association Française contre les Myopathies (AFM-TELETHON). Funding for open access charge: Academic Grants.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
2. Weinreich,S.S., Mangon,R., Sikkens,J.J., Teeuw,M.E.E. and Cornel,M.C. (2008) [Orphanet: a European database for rare diseases]. *Ned. Tijdschr. Geneeskd.*, **152**, 518–519.
3. Gargis,A.S., Kalman,L., Bick,D.P., da Silva,C., Dimmock,D.P., Funke,B.H., Gowrisankar,S., Hegde,M.R., Kulkarni,S., Mason,C.E. *et al.* (2015) Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat. Biotechnol.*, **33**, 689–693.
4. Salgado,D., Bellgard,M.I., Desvignes,J.-P. and Béroud,C. (2016) How to identify pathogenic mutations among all those variations: variant annotation and filtration in the genome sequencing era. *Hum. Mutat.*, **37**, 1272–1282.
5. Eilbeck,K., Quinlan,A. and Yandell,M. (2017) Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.*, **18**, 599–612.
6. Sawyer,S.L., Hartley,T., Dyment,D.A., Beaulieu,C.L., Schwartzentruber,J., Smith,A., Bedford,H.M., Bernard,G., Bernier,F.P., Brais,B. *et al.* (2016) Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin. Genet.*, **89**, 275–284.
7. Bertoldi,L., Forcato,C., Vitulo,N., Birolo,G., De Pascale,F., Feltrin,E., Schiavon,R., Anglani,F., Negrisolo,S., Zanetti,A. *et al.* (2017) QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics*, **18**, 225.
8. Stelzer,G., Plaschkes,I., Oz-Levi,D., Alkelai,A., Olender,T., Zimmerman,S., Twik,M., Belinky,F., Fishilevich,S., Nudel,R. *et al.* (2016) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*, **17**(Suppl. 2), 444.
9. Hart,S.N., Duffy,P., Quest,D.J., Hossain,A., Meiners,M.A. and Kocher,J.-P. (2015) VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. *Brief. Bioinform.*, **17**, 346–351.
10. Alemán,A., Garcia-Garcia,F., Salavert,F., Medina,I. and Dopazo,J. (2014) A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Res.*, **42**, W88–W93.

11. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
12. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
13. Niemiec,E. and Howard,H.C. (2016) Ethical issues in consumer genome sequencing: use of consumers' samples and data. *Appl. Transl. Genom.*, **8**, 23–30.
14. Jallades,L., Baseggio,L., Sujobert,P., Huet,S., Chabane,K., Callet-Bauchu,E., Verney,A., Hayette,S., Desvignes,J.-P., Salgado,D. *et al.* (2017) Exome sequencing identifies recurrent BCOR gene alterations and the absence of KLF2, TNFAIP3 and MYD88 mutations in splenic diffuse red pulp small B-cell lymphoma. *Haematologica*, **102**, 1758–1766.
15. Saultier,P., Vidal,L., Canault,M., Bernot,D., Falaise,C., Pouymayou,C., Bordet,J.-C., Saut,N., Rostan,A., Baccini,V. *et al.* (2017) Macrothrombocytopenia and dense granule deficiency associated with FLI1 variants: ultrastructural and pathogenic features. *Haematologica*, **102**, 1006–1016.
16. Elouej,S., Beleza-Meireles,A., Caswell,R., Colclough,K., Ellard,S., Desvignes,J.-P., Béroud,C., Lévy,N., Mohammed,S. and De Sandre-Giovannoli,A. (2017) Exome sequencing reveals a de novo POLD1 mutation causing phenotypic variability in mandibular hypoplasia, deafness, progeroid features, and lipodystrophy syndrome (MDPL). *Metab. Clin. Exp.*, **71**, 213–225.
17. Marquet,S., Bucheton,B., Reymond,C., Argiro,L., El-Safi,S.H., Kheir,M.M., Desvignes,J.-P., Béroud,C., Mergani,A., Hammad,A. *et al.* (2017) Exome sequencing identifies two variants of the alkylglycerol monooxygenase gene (AGMO) as a cause of relapses in visceral leishmaniasis in children, in Sudan. *J. Infect. Dis.*, **216**, 22–28.
18. Cerino,M., Gorokhova,S., Laforêt,P., Ben Yaou,R., Salort-Campana,E., Pouget,J., Attarian,S., Eymard,B., Deleuze,J.-F., Boland,A. *et al.* (2017) Genetic Characterization of a French Cohort of GNE-mutation negative inclusion body myopathy patients with exome sequencing. *Muscle Nerve*, **56**, 993–997.
19. Galant,D., Gaborit,B., Desgrouas,C., Abdesselam,I., Bernard,M., Lévy,N., Merono,F., Coirault,C., Roll,P., Lagarde,A. *et al.* (2016) A heterozygous ZMPSTE24 mutation associated with severe metabolic syndrome, ectopic fat accumulation, and dilated cardiomyopathy. *Cells*, **5**, E21.
20. Miltgen,M., Blanchard,A., Mathieu,H., Kreisler,A., Desvignes,J.-P., Salgado,D., Roubertie,A., Barré,L., Raï,G., Blanck,V. *et al.* (2016) Novel heterozygous mutation in ANO3 responsible for craniocervical dystonia. *Mov. Disord.*, **31**, 1251–1252.
21. Rapetti-Mauss,R., Lacoste,C., Picard,V., Guitton,C., Lombard,E., Loosveld,M., Nivaggioni,V., Dasilva,N., Salgado,D., Desvignes,J.-P. *et al.* (2015) A mutation in the Gardos channel is associated with hereditary xerocytosis. *Blood*, **126**, 1273–1280.
22. Kalia,S.S., Adelman,K., Bale,S.J., Chung,W.K., Eng,C., Evans,J.P., Herman,G.E., Hufnagel,S.B., Klein,T.E., Korf,B.R. *et al.* (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.*, **19**, 249–255.
23. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
24. Liu,X., Wu,C., Li,C. and 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.
25. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2014) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
26. Köhler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H.V., Bailleul-Forestier,I., Black,G.C.M., Brown,D.L., Brudno,M., Campbell,J. *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.
27. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
28. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
29. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
30. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
31. Salgado,D., Desvignes,J.-P., Raï,G., Blanchard,A., Miltgen,M., Pinard,A., Lévy,N., Collod-Beroud,G. and Béroud,C. (2016) UMD-Predictor: a High throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. *Hum. Mutat.*, **37**, 439–446.
32. Desmet,F.-O., Hamroun,D., Lalande,M., Collod-Beroud,G., Claustres,M. and Béroud,C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, **37**, e67.
33. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
34. Matthijs,G., Souche,E., Alders,M., Corveleyn,A., Eck,S., Feenstra,I., Race,V., Sistermans,E., Sturm,M., Weiss,M. *et al.* (2016) Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.*, **24**, 2–5.
35. Sevy,A., Cerino,M., Gorokhova,S., Dionnet,E., Mathieu,Y., Verschueren,A., Franques,J., Maues de Paula,A., Figarella-Branger,D., Lagarde,A. *et al.* (2015) Improving molecular diagnosis of distal myopathies by targeted next-generation sequencing. *J. Neurol. Neurosurg. Psychiatr.*, **87**, 340–342.
36. Kamphans,T., Sabri,P., Zhu,N., Heinrich,V., Mundlos,S., Robinson,P.N., Parkhomchuk,D. and Krawitz,P.M. (2013) Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PLoS One*, **8**, e70151.
37. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
38. Piscosquito,G., Saveri,P., Magri,S., Ciano,C., Gandioli,C., Morbin,M., Bella,D.D., Moroni,I., Taroni,F., Pareyson,D. *et al.* (2016) Screening for SH3TC2 gene mutations in a series of demyelinating recessive Charcot-Marie-Tooth disease (CMT4). *J. Peripher. Nerv. Syst.*, **21**, 142–149.
39. Thompson,R., Johnston,L., Taruscio,D., Monaco,L., Béroud,C., Gut,I.G., Hansson,M.G., 't Hoen,P.-B.A., Patrinos,G.P., Dawkins,H. *et al.* (2014) RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J. Gen. Intern. Med.*, **29**(Suppl. 3), S780–S787.