



HAL
open science

Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine

Kevin Yauy, François Lecoquierre, Stéphanie Baert-Desurmont, Detlef Trost, Aicha Boughalem, Armelle Luscan, Jean-Marc Costa, Vanna Geromel, Laure Raymond, Pascale Richard, et al.

► To cite this version:

Kevin Yauy, François Lecoquierre, Stéphanie Baert-Desurmont, Detlef Trost, Aicha Boughalem, et al.. Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine. *Genetics in Medicine*, 2022, 24 (6), pp.1316-1327. 10.1016/j.gim.2022.02.008 . hal-04001779

HAL Id: hal-04001779

<https://hal.science/hal-04001779>

Submitted on 23 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ARTICLE

Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine



Kevin Yauy^{1,2,*} , François Lecoquierre³, Stéphanie Baert-Desurmont³, Detlef Trost⁴, Aicha Boughalem⁴, Armelle Luscan⁴, Jean-Marc Costa⁴, Vanna Geromel⁵, Laure Raymond⁵, Pascale Richard⁶, Sophie Coutant³, Mélanie Broutin², Raphael Lanos², Quentin Fort², Stenzel Cackowski⁷, Quentin Testard^{1,5}, Abdoulaye Diallo², Nicolas Soirat², Jean-Marc Holder², Nicolas Duforet-Frebourg², Anne-Laure Bouge², Sacha Beaumeunier², Denis Bertrand², Jerome Audoux², David Genevieve⁸, Laurent Mesnard^{9,10}, Gael Nicolas³, Julien Thevenon¹, Nicolas Philippe²

ARTICLE INFO

Article history:

Received 30 November 2021

Received in revised form

7 February 2022

Accepted 7 February 2022

Available online 17 March 2022

Keywords:

ClinVar

Gene–phenotype associations

Sequencing reinterpretation

Variant pathogenicity

ABSTRACT

Purpose: Retrospective interpretation of sequenced data in light of the current literature is a major concern of the field. Such reinterpretation is manual and both human resources and variable operating procedures are the main bottlenecks.

Methods: Genome Alert! method automatically reports changes with potential clinical significance in variant classification between releases of the ClinVar database. Using ClinVar submissions across time, this method assigns validity category to gene–disease associations.

Results: Between July 2017 and December 2019, the retrospective analysis of ClinVar submissions revealed a monthly median of 1247 changes in variant classification with potential clinical significance and 23 new gene–disease associations. Re-examination of 4929 targeted sequencing files highlighted 45 changes in variant classification, and of these classifications, 89% were expert validated, leading to 4 additional diagnoses. Genome Alert! gene–disease association catalog provided 75 high-confidence associations not available in the OMIM morbid list; of which, 20% became available in OMIM morbid list. For more than 356 negative exome sequencing data that were reannotated for variants in these 75 genes, this elective approach led to a new diagnosis.

Conclusion: Genome Alert! (<https://genomealert.univ-grenoble-alpes.fr/>) enables systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine with limited human resource effect.

© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Gael Nicolas, Julien Thevenon, and Nicolas Philippe jointly supervised this work.

*Correspondence and requests for materials should be addressed to Kevin Yauy, Institute for Advanced Biosciences, UGA/Inserm U 1209/CNRS UMR 5309 joint research center, Site Santé-Allée des Alpes, 38700 La Tronche, France. *E-mail address:* kevin.yauy@univ-grenoble-alpes.fr

Affiliations are at the end of the document.

doi: <https://doi.org/10.1016/j.gim.2022.02.008>

1098-3600/© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Genetic tests are increasingly prescribed and included in health care pathways for diverse clinical indications.^{1,2} Several countries have developed population genomics organizations that are revolutionizing medical practices.^{3,4} However, many of these genomic analyses remain inconclusive owing to limitations in genomic and medical knowledge available at the time of analysis.

The American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) recommendations for variant classification aim at standardizing variant interpretation practices in genomic centers, in the context of medical interpretation.⁵ Recently, tools have been published to automatically classify genomic variants on the basis of these recommendations.⁶⁻⁸ Meanwhile, evolving medical knowledge and rapid adoption of clinical genome sequencing have influenced the standard practices and have created additional needs. A current and major preoccupation in this field is the definition of standards for periodic and prospective reanalysis of existing sequencing data. Indeed, reanalyzing existing genomic data improves diagnostic yield (7% increase per year).^{9,10}

In practice, such an in-depth reinterpretation is mainly manual and time-consuming, with major bottlenecks such as human and funding resources or lack of consistency between centers. Clinical recommendations from the American and European Societies of Human Genetics reinforce the need for a standardized and automated approach to the reinterpretation of genomic analyses.¹¹⁻¹⁴ Some companies offer paid black box services, with poorly detailed methods that cannot be reproduced.^{15,16}

Clinical knowledge of rare diseases is contained in expert-curated databases (such as OMIM¹⁷ or Clinical Genome Resource [ClinGen]¹⁸), peer-reviewed medical literature, and information sharing between health practitioners through community-based platforms (such as MatchMaker Exchange¹⁹ or ClinVar²⁰). Reliability and exhaustiveness of information vary widely across these data sources. Furthermore, careful monitoring of clinical knowledge by every laboratory represents an organizational challenge for a prospective reanalysis of acquired data. To enable a systematic, reproducible, and prospective genome interpretation, a collaborative approach for clinical knowledge aggregation combined with automated medical knowledge monitoring and curation is needed.

The main community-based repository of genomic knowledge is ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), a shared variant interpretation database that featured 1 million submissions in 2020. ClinVar is updated weekly with several thousands of modifications of variant classifications that could affect the diagnostic yield of previous analyses. There is currently no monitoring system that can highlight these changes at a scale for the complete database. Besides variant classification, gene-phenotype

association catalogs are crucial because they are commonly used to design phenotype-specific gene panels for dry-lab filtering and set the frontiers for clinical genome analysis.^{21,22} Although not their primary purpose, variant-centered databases could also theoretically provide a complementary resource to gather gene-phenotype knowledge.

In this article, we detail an automated method for the reassessment of variant pathogenicity and gene-phenotype associations through ClinVar follow-up. This procedure, called Genome Alert!, aims at performing a routine and systematic reinterpretation of existing genomic data. The procedure's effectiveness was evaluated through a 29-month multicentric series (2018-2019) of 5959 consecutive individuals screened using targeted sequencing (4929 individuals with hereditary cancers) and exome sequencing (1000 analyses including 356 undiagnosed individuals with suspected Mendelian disorders).

Materials and Methods

Genome Alert! standardized procedure

ClinVCF, Variant Alert!, and ClinVarome are a suite of tools that constitute the heart of the Genome Alert! standardized procedure.

ClinVCF: A ClinVar quality processing method

Before comparing different versions of the same source, data consistency needs to be verified. This first step is based on ClinVCF tool, and once every submission has been tracked, data will be processed for the next step.

ClinVCF imports monthly updated ClinVar Xtensible Markup Language (XML) files. XML format was preferred over VCF mainly because of better consistency and traceability across versions for the ClinVar Variation ID, the history of changes in each variant classification, and the additional gene-phenotype data availability in XML. ClinVCF considers an automatic reclassification of variants with at least 4 submissions and conflicting interpretations of pathogenicity status. Consensus classification according to ClinVar policies sets the conflicting interpretations of pathogenicity status when at least 1 conflict in submission is observed, except if an expert consortium (as ClinGen) has defined classification (details available in [Supplemental Method 1](#)). On the basis of the provided classifications transformed from literal transcription (eg, likely pathogenic) to class number (eg, class 4), if ≥ 4 submissions are available, a new consensus is proposed after outlier submissions removal according to the 1.5* Interquartile Range (IQR) Tukey method.²³ We only reclassify variants from conflicting status to likely pathogenic or pathogenic status. ClinVCF provides a 3-tier reclassification confidence score detailed in [Supplemental Figure 1](#). As an output, ClinVCF writes a Variant Calling File (VCF) v4.2 file.

Variant Alert!: A variant knowledge monitoring tool

Variant Alert! tool aims at identifying changes in variant classification across 2 versions of the database. Changes were defined as (1) a modification in the classification of an existing variant and (2) the creation or suppression of a variant entry.

Stratification of the consequences in classification modification was proposed (Supplemental Table 1). Major classification modification was defined as a change that may affect the clinical management of a patient (eg, uncertain significance to likely pathogenic status). Minor classification modification was defined as a change that may not affect the clinical management of a patient (eg, pathogenic to likely pathogenic status).

Variant Alert! writes 2 files: (1) the list of variants that were modified, added, or removed and (2) the list of genes that were added to or removed from the database. This gene list is notably used by ClinVarome.

ClinVarome: A method for automated gene–disease association evaluation

ClinVarome tool aims to periodically and automatically evaluate gene–disease association in the ClinVar database. To differentiate genes on the basis of their clinical validity, the work from European Molecular Biology Laboratory–European Bioinformatics Institute Gene2Phenotype,²⁴ ClinGen,¹⁸ and Genomic England PanelApp²⁵ were first compared. Although theoretically comparable, their rationales and contents were partially overlapping and with conflicting classifications. To discriminate candidate genes from definitive gene–disease associations, we decided to use an unsupervised clustering model. Only the genes with at least 1 likely pathogenic or pathogenic variant (single nucleotide variant or indel affecting a single gene) in ClinVar were considered in a list called ClinVarome. As a consensus criterion, we chose to assess the strength of a gene–disease association through the quantification of 4 variables: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification (CLNSIG, likely pathogenic or pathogenic), (3) highest ClinVar review variant confidence (CLNREVSTAT, from 0 to 4 stars), and (4) time interval between the first and the last pathogenic variant submission (replication of the gene–disease association event). For these 4 variables, values were gathered through periodic monitoring of changes in the database following the ClinVCF and Variant Alert! tool procedures. Clustering variants according to these variables allowed us to define clusters of genes according to their clinical validity. The scikit-learn Agglomerative Clustering tool (parameters: Euclidean affinity, ward linkage) was used, and t-distributed stochastic neighbor embedding representation (parameters: 2 components, perplexity 150, 2000 iterations, and 1000 iterations without progress) was performed. Gene–disease validity classification was computed per gene but not per disease. The Gene Curation Coalition (GenCC) (<https://thegencc.org/>) database was released recently and was used to evaluate ClinVarome. To compare ClinVarome

clusters and GenCC classification, GenCC submissions were summarized into 3 categories (Green, Orange, Red) (Supplemental Methods 2).

Study design and participants

To evaluate the clinical impact of Genome Alert!, we collected 5929 consecutive germline sequencing data samples from 3 centers in France between July 2017 and December 2019 as part of their routine genetic investigation: (1) a variant database gathering all class 3 (uncertain significance), class 4 (likely pathogenic), and class 5 (pathogenic) variants identified in a colon cancer–targeted sequencing (14 genes) sequenced in 2540 individuals in the Rouen University Hospital; (2) a cancer-targeted sequencing data set of 2389 individuals by the Cerba laboratory (66 genes); and (3) exome sequencing data of individuals with developmental disorders, rare kidney diseases, or other rare diseases as follows: 108 probands from the Rouen University Hospital, 477 probands (with 356 negative analysis) from the Cerba laboratory, and 415 probands from the Eurofins Biomnis laboratory. Patient samples, together with a basic phenotype description and molecular diagnosis (when available), were anonymized. Two main clinical evaluations were performed: (1) variant-centered reanalysis, which aims at matching individuals that carry exact variants with potential clinical significance reported by Genome Alert!, and (2) gene-centered reanalysis, which aims at matching individuals who carry candidate variants in high-confidence clinical genes referenced in ClinVarome and not in OMIM. Initial analyses were performed between 0 and 2 years before this reanalysis.

Selection of variants with potential clinical significance

All sequencing data were systematically reinterpreted according to Genome Alert!'s report and compared with the initial variant interpretation. For targeted sequencing and exome reanalysis, genomic positions of variants with major changes in classification were queried in the existing patient's variant calling files (variant-centered analysis). For exome data, we performed a reanalysis of variants in VCF with the following criteria: (1) among 75 ClinVarome morbid genes, which were not available in OMIM, and with a second event of gene–disease validation (including a likely pathogenic or pathogenic variant with ClinVar review confidence ≥ 2 stars and a likely pathogenic or pathogenic variant entry subsequent to the initial entry); (2) variant not shared with another individual in the series; (3) sufficient sequencing quality (variant allele fraction $> 25\%$ and read depth > 20 reads); (4) rare in Genome Aggregation Database²⁶ population (frequency $< 10^{-5}$ if heterozygous genotype or 10^{-4} if homozygous genotype); and (5) protein consequence among nonsense, frameshift, missense (missense are selected with Combined Annotation

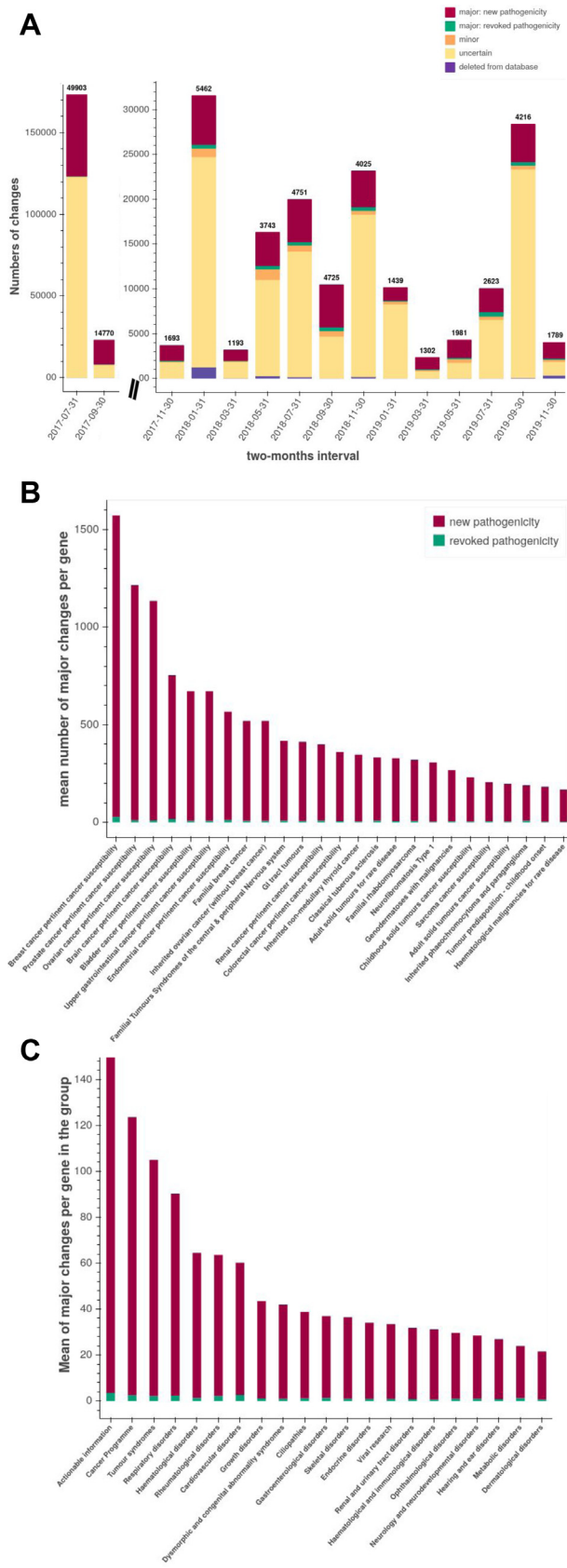


Figure 1 ClinVar variant classification monitoring between July 2017 and December 2019. A. Bar chart distribution of every 2 months of changes in variant classification. The bar chart was

Dependent Depletion²⁷ score > 30 and MetaSVM²⁸ = D), or splice variants (based on dbcsnv RF²⁹ predicted impact score > 0.6) (gene-centered reanalysis).

Results

ClinVar knowledge dynamics

To get insights into variant classification and gene–disease association and to estimate the amount of new clinically relevant information in the ClinVar database available through time, a retrospective analysis of ClinVar submissions over 29 months was performed (July 2017 [included] to December 2019). Of note, VCF genomic positions in ClinVar were introduced in July 2017 and probably are associated with the largest injection in the ClinVar database.

The number of variants with ACMG/AMP classification⁵ increased from 144,943 to 491,838. Among modifications in the database, the count of major changes was 107,167 in ACMG/AMP classification, and among these, 103,615 resulted in a pathogenicity status, which was previously unreported, whereas 3552 resulted in the revocation of a previously established pathogenicity (Figure 1A). These changes varied significantly according to disease groups the between gene panels (according to Genomics England PanelApp), in which the oncogenetic panels were on top of the list of panels. The panels and disease groups presenting most of the changes per gene are presented in Figure 1B and C and Supplemental Table 2. Clinical gene entries in ClinVar were also monitored. A median of 23 ClinVar morbid genes per month that were newly associated with Mendelian disease was observed (Figure 2).

Changes in variant classification

To evaluate the robustness of clinical variant information, the consistency of variant classification was explored and is described in Supplemental Table 3. Among 144,943

split for better readability. Bold numbers and dark red color represent new (likely) pathogenic variant entries, green represents number of revoked (likely) pathogenic variants, orange represents number of minor change variants (eg, pathogenic to likely pathogenic), yellow represents number of changes with uncertain clinical impact (VUS or conflict entry), and purple represents number of changes leading to variant disappearance. B. Bar chart of top panels with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants. C. Bar chart of top disease group with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants.

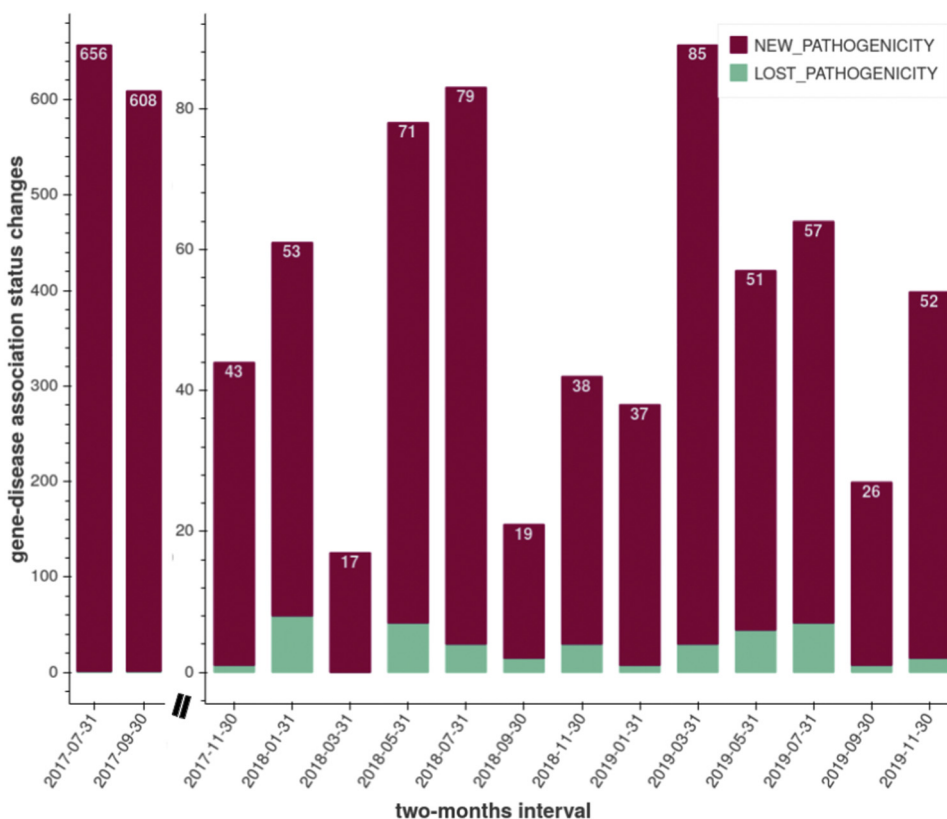


Figure 2 ClinVar clinical genes entries associated with new or deprecated Mendelian disease (morbid status) distribution between December 2017 and December 2019. The bar chart was split for better readability. Dark red represents morbid genes entries (first variant with likely pathogenic or pathogenic status), and green represents revoked morbid genes. White numbers represents number of new morbid gene entries by 2 months.

variants available in July 2017, 10,254 (7%) were reclassified between July 2017 and December 2019, ie, we observed only a small portion of variants being reclassified over time. These reclassifications included automatically reclassified variants with conflicting interpretations. More precisely, among the 11,417 likely pathogenic variants, 1125 (9.94 %) variants were reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretations of pathogenicity.

Automatic variant reclassification with conflicting interpretations

A criticism of the ClinVar database is the misclassification of pathogenic variants, such as the well-known *HFE* pathogenic variant NM_000410.3:c.845G>A. We observed that it was mostly due to a unique outlier submission with a classification for a distinct condition (eg, cutaneous photosensitivity porphyria phenotype). We evaluated our method to remove such outlier submissions. Among all the variants available in ClinVar in December 2019, 22,973 of a total of 503,994 (4.5%) variants were classified with a conflicting interpretation of pathogenicity. Genome Alert! automatic reclassification method proposes to detect outlier submissions to suggest a consensus classification. This

allowed the reclassification of 188 variants from conflict to likely pathogenic or pathogenic classification in 135 genes and 1625 variants in 436 genes from conflict to likely benign or benign classification (Supplemental Table 4, Supplemental Figures 1 and 2).

Variants automatically reclassified as likely pathogenic or pathogenic in cancer ($n = 9$) and cardiogenetic disease ($n = 11$) were presented to French National experts in the field. Of these 20 automatic reclassifications, 17 were confirmed as accurate by experts and 3 remained as variants of uncertain significance, lacking evidence of pathogenicity for our experts.

Clinical impact of changes in variant classification

To assess the clinical impact of Genome Alert!'s changes in variant classification, previously analyzed cancer-predisposition targeted sequencing data were assessed (4929 individuals from 2 genetic centers) (variant-centered reanalysis, Figure 3). Among all variants detected in this cohort, this method highlighted 45 variants with major changes between the time of analysis and December 2019, which were proposed for manual review by their referring geneticists (Supplemental Tables 5 and 6).

Among the 45 variants, 30 had been already manually reported by the clinical geneticists as likely pathogenic or

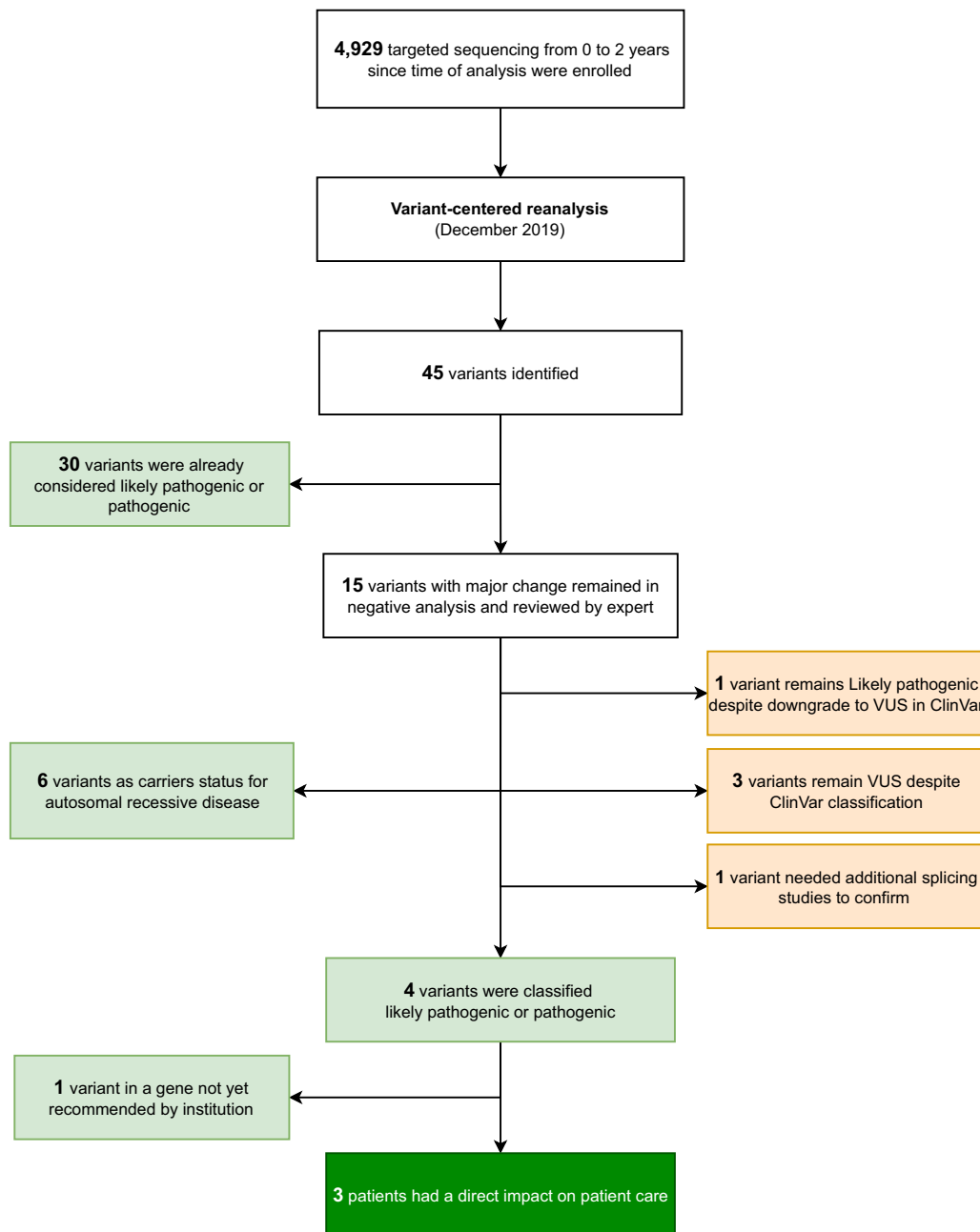


Figure 3 Experimental design of the variant-centered reanalysis. Flow charts describing how the sequencing data were reinterpreted according to variant reclassification only. Green box represents new diagnosis. Light green boxes represent confirmed variant classification. Orange boxes represent excluded variants. VUS, variant of uncertain significance.

pathogenic at the initial time of analysis, meaning that these classifications were ahead of the ClinVar database. The 15 unreported variants were manually curated, looking for additional diagnoses. Among them, 14 variants were newly classified as likely pathogenic or pathogenic and 1 was downgraded as a variant of uncertain significance (VUS) in ClinVar. The manual curation of these 14 variants led to the conclusion that 6 corresponded to a carrier status for a recessive disorder, 3 were manually classified as VUS, and 5 were submitted to a multidisciplinary meeting for external review. Finally, 4 of these latter 5 were classified as likely

pathogenic or pathogenic by experts leading to additional diagnoses. One variant remained classified as a VUS, and complementary studies on the patient's messenger RNA were proposed before conclusion (*PALB2*, NC_000016.9(NM_024675.3):c.3350+4A>G). Finally, an 89% validation rate (40 of 45) of major changes were observed. This variant reclassification tracking system allowed an additional diagnosis per 1000 analyses.

Replication of the variant-centered reanalysis was performed in the exome sequencing cohort, looking for variant exact match. Selective reanalysis in previous exome

sequencing analysis (1000 individuals in 3 genomic centers) highlighted <1 variant per exome (only 297 variants) with major changes between the time of analysis and December 2019. These 297 variants were then explored by clinical geneticists. Among all 297 variants, 1 variant (*POLG*, NM_002693.2:c.2243G>C) was automatically reclassified as pathogenic by our IQR outlier submission method and was initially reported as VUS, thus helping us to confirm the diagnosis. Compound heterozygosity was observed for a pathogenic variant (*POLG*, NM_002693.3:c.1399G>A). Exome sequencing reanalysis with the variant-centered reanalysis also provides an additional diagnosis per 1000 analyses.

Monitoring ClinVar gene–disease association knowledge

A focus has been toward exploring rarely explored gene–disease association in ClinVar data. To discriminate candidate genes from definitive gene–disease associations in ClinVarome, unsupervised clustering was performed on the basis of the following criteria: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification, (3) highest ClinVar review variant confidence, and (4) time interval between the first and the last pathogenic variant submission. According to distances between clusters and model dendrogram, the number of clusters was set to 4 (Figure 4). Careful observation of these clusters identified objective patterns to understand the classification. We observed that all genes in the first and second clusters had a reproducibility event (a new likely pathogenic or pathogenic variant entry, the confirmation of the likely pathogenic or pathogenic classification by another submitter or expert panel) in pathogenicity status, thus giving them strong confidence. Genes from the first cluster hold pathogenic variants with ClinVar's ≥ 2 stars of review confidence and the second cluster genes include pathogenic variants with different entry dates and <2 stars of review confidence. Genes in the third cluster had 1 strong argument for pathogenicity but needed another event to be fully confirmed (the third cluster genes contained at least 1 pathogenic variant and all pathogenic entries were added at the same date). Because genes in the fourth cluster were only likely pathogenic variants, their gene–disease association remained to be confirmed (Supplemental Table 7).

To assess the exhaustivity of the ClinVarome, a comparison with the OMIM database was performed. In December 2019, there was a 95% overlap (3675/3858) between OMIM morbid clinical genes and ClinVarome morbid genes. Overall, 365 genes were referenced only in OMIM and not in ClinVarome. We observed patterns that were not available in ClinVar. These patterns include nonconfirmation of a disorder as a genuine Mendelian disorder (only 1 publication or isolated patient reports), susceptibility to multifactorial disorders or infection, referencing of genes belonging to

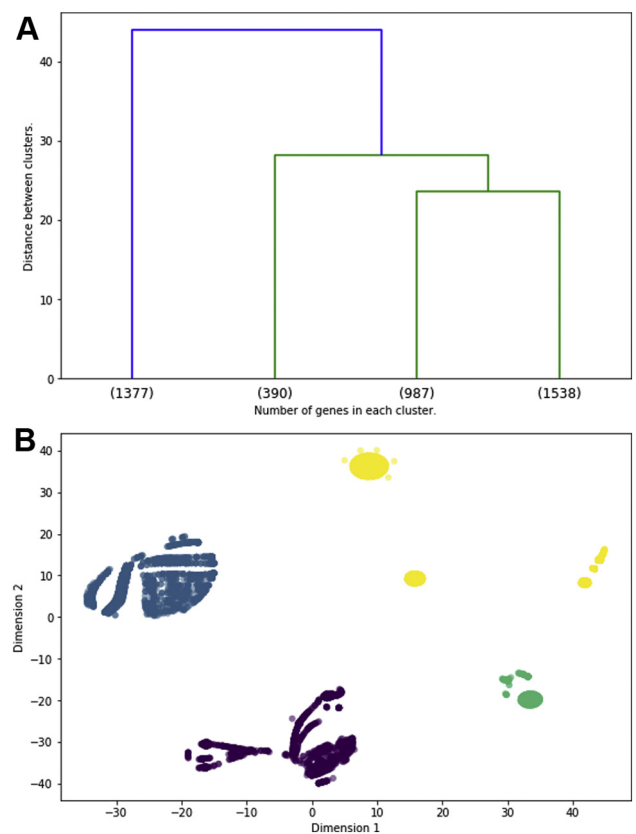


Figure 4 ClinVarome morbid genes exploration and gene–disease validity classification. A. Agglomerative clustering dendrogram of ClinVarome in December 2019. B. t-distributed stochastic neighbor embedding representation of ClinVarome 4 variables by gene data. Green represents fourth cluster (390 genes), yellow represents third cluster (987 genes), blue represents second cluster (1538 genes), and purple represents first cluster (1377 genes).

molecular mechanism distinctive from a single gene disorder as microdeletion or microduplication syndromes, Mendelian traits that are not diseases, epigenetic loci, genes with targeted pathogenic complex variants, and very recently described diseases. The evaluation focused on these 519 specific genes, referenced only in ClinVar and not in OMIM, to assess their potential value in additional diagnoses.

Among the 519 ClinVarome only genes in December 2019, 15 genes were in the first cluster, 60 genes were in the second cluster (ie, 75 high-confidence genes), 140 genes were in the third cluster, and 304 genes were in the fourth cluster. Then, we monitored their inclusion in the OMIM morbid list in the upcoming months. Among the 519 genes exclusively referenced in ClinVarome in December 2019, 55 were reported OMIM morbid 8 months later in August 2020, including 15 of the 75 (20%) initial high-confidence genes. Moreover, 125 of the 140 OMIM morbid genes additional entries between December 2019 and August 2020 were also referenced in ClinVarome release of August 2020. This observation suggested that candidate genes in

ClinVarome may be considered as diagnostic genes before the OMIM validation of the gene–disease causality.

Clinical impact of ClinVarome morbid genes not available in OMIM

We evaluated the relevance of this approach by performing a selective reanalysis of a subsample of the new entries in the ClinVarome, focusing only on the 75 genes that were absent from OMIM morbid list and were referenced in ClinVarome’s first and second clusters (gene-centered reanalysis). This experiment highlighted 42 variants in 356

negative exome sequencing data. In this data set, 42 variants were prioritized and were proposed for further interpretation. Among them, 39 were excluded by the expert. The experts’ arguments included the presence of variants unrelated to the disease phenotype or a single case series available in the literature. A total of 3 variants were further explored with Sanger sequencing validation, of which 2 were excluded because of artifact status or discordant inheritance pattern (Figure 5).

Overall, this method could ascertain a new diagnosis from the 356 negative exome sequencing data. A nonsense *DLG4* variant NM_001128827.1:c.1840C>T was reported

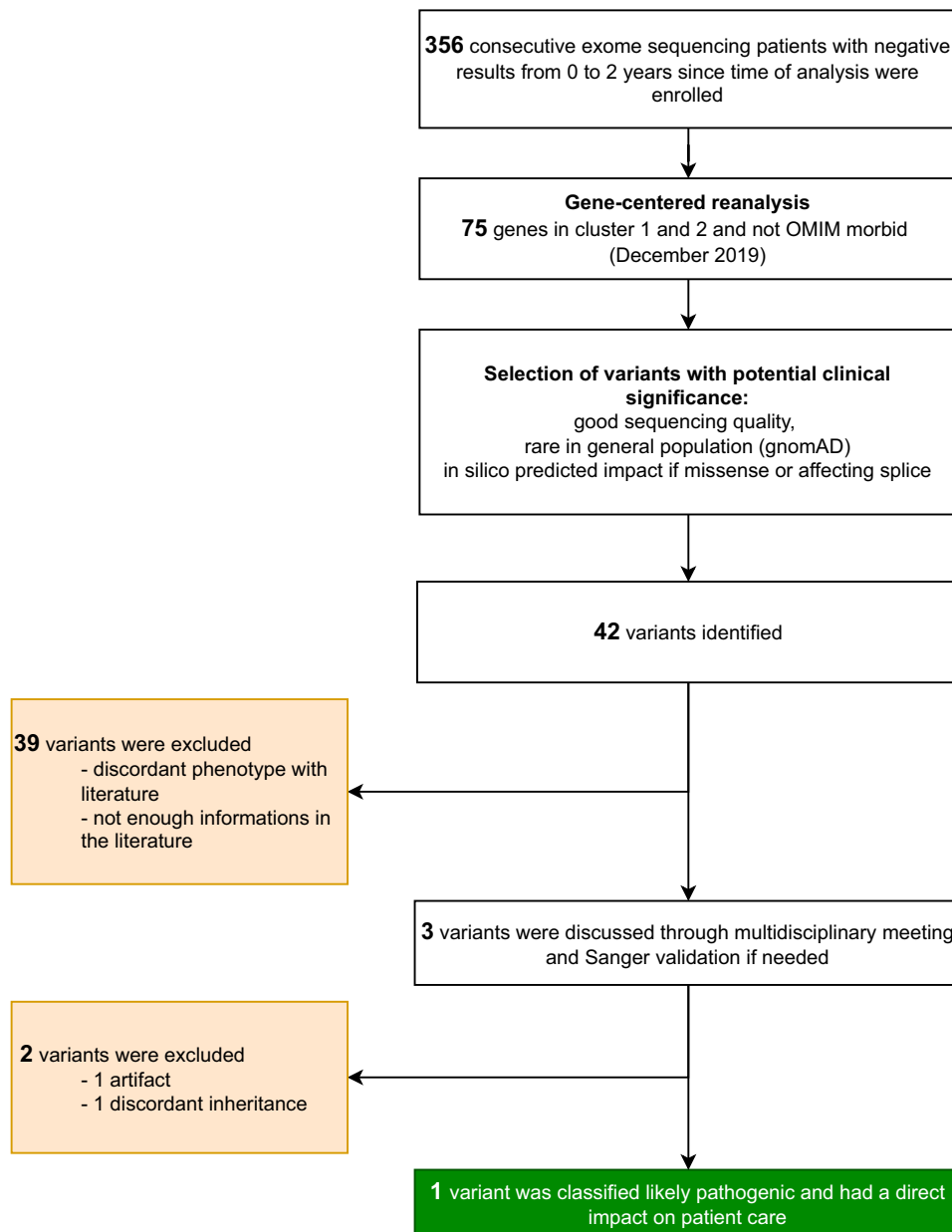


Figure 5 Experimental design for a targeted gene-centered reanalysis. These 75 genes were reported in ClinVarome and not in OMIM and classified as related to a disease (clusters 1 and 2). This list of 75 genes was used for the reinterpretation of negative exome sequencing data ($n = 346$). Green box represents new diagnosis. Orange boxes represent excluded variants. gnomAD, Genome Aggregation Database.

as likely pathogenic, responsible for the patient's phenotype (intellectual disability and microcephaly). Although the first report of *DLG4* association to intellectual developmental disorder was described back in 2016, this gene–disease association was added to the OMIM database only in February 2020.

ClinVarome comparison with the GenCC database

A comparison of gene–disease validity confidence and exhaustivity of ClinVarome with the GenCC database was performed. In October 2021, there was a 65% (3332 of 5187) gene overlap between the 2 databases. Nonoverlapping genes represent mostly the uncertain gene–disease associations from these 2 databases. Exclusive genes in GenCC ($n = 334$) were significantly enriched in orange and red genes (151 of 745 orange genes [$P < .0001$], 158 of 252 red genes [$P < .0001$]). Exclusive genes in ClinVarome ($n = 1471$) were significantly enriched in third and fourth cluster genes (407 of 501 third cluster genes [$P < .0001$], 448 of 743 fourth cluster genes [$P < .0001$]). The 2 databases present a high concordance in gene–disease association confidence (Supplemental Table 8).

Discussion

With the increasing amount of genetic testing performed in health care, there is a critical need for standardized methods to enable prospective genomic data reinterpretation in clinical routine. Through the reassessment of variant pathogenicity and gene–phenotype associations in ClinVar, Genome Alert!'s data mining method proposes the automatic report of a handful of variants that can reasonably be manually interpreted. Our method was applied to a multicentric series of 4929 sequencing tests with various local bioinformatic systems. Genome Alert! successfully allowed new diagnoses in targeted and exome sequencing through query of laboratory's VCFs or variant database and proposed a portable and open-source framework for an automated reanalysis of sequencing data.

Retrospective monitoring of the cutting-edge medical literature on existing genomic data is a major concern for paving the way to genomic medicine.³⁰ There are numerous technical and medical challenges in setting up a routine procedure for reanalysis. This work explored the dynamics of change across all fields of genomic medicine in ClinVar.

Several medical indications for genomic testing were noticed to bear numerous changes in variant classification. Retrospective analysis of the ClinVar database provided an estimation of new clinically relevant information reported each month, which may lead to additional diagnoses in the existing data.³¹ Overall, 9.94 % (1125) of likely pathogenic variants were eventually downgraded and reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretation of

pathogenicity in ClinVar over the study period (Supplemental Table 3). This analysis highlights the required carefulness in returning results to the families for likely pathogenic variants because such information could be used for genetic counseling and patient management.

Genome Alert! methods are based on the processing of submissions from the ClinVar full XML release, with no distinction made between submissions with different contexts (eg, somatic or germline status and distinct conditions). Besides, Genome Alert! attributes a unique variant ID on the basis of VCF nomenclature. As such, these variants with potential clinical significance reported by Genome Alert! should be queryable a priori in each genomic center. However, VCF nomenclature is not easy to use with complex variation, which could lead to errors. A switch to the Variation Representation specification from the Global Alliance for Genomics and Health could provide an interesting improvement step.

Clinical effect of changes in variant classification (variant-centered reanalysis) provided in our targeted and exome sequencing cohort provided an additional diagnosis per 1000 analyses. Because time from initial analysis varies from 0 to 2 years, this diagnostic yield will certainly increase with time. This automated system is better for large cohorts of targeted sequencing, with a low number of variants to reinterpret and reaching 10% diagnostic yield in the re-examined variants. Recent literature emphasizes the importance of a standardized procedure adapted for sequencing data reanalysis for considering few candidate variants after an accurate annotation of new gene–phenotype associations and filtering procedure.³⁰

A particular effort was made to evaluate confidence in the reported information to reach a consensus across multiple annotations. The prospective reassessment of ClinVar highlighted numerous conflicts in variant classification. Although our system rarely reclassifies variants with conflicting interpretations, this automatic reclassification method aims to at least remove these potential errors. The expert review of ClinVCF automatic reclassification validates this method on the basis of outlier submission removal using the IQR method, and succeeds in reclassifying abnormalities such as the *HFE* pathogenic variant NM_000410.3:c.845G>A. This work highlights the value of the persistence over time of a classification for relevant genomic information. This work specifically focused on oncogenetics and cardiogenetics, fields in which variant interpretations are particularly conflicting and shifting.^{32,33} Overall, in the ClinVar database, 188 variants could be reclassified in 29 months (ranging from 2017 to 2019). After 8 months, in August 2020, a total of 307 variants were reclassified, highlighting the importance of a systematic and partially automated variant reassessment (Supplemental Figure 2).

Existing literature for gene-centered reanalysis has emphasized the importance of OMIM as an updated resource but not exhaustive.³⁴ To explore and evaluate specifically the ClinVar database for gene-centered reanalysis, we chose to focus our reanalysis on 75 high-confidence ClinVarome

morbid genes (first and second clusters) not available in OMIM morbid genes list. Complementary to OMIM morbid genes, these high-confidence ClinVarome morbid genes from the first and second clusters could provide additional diagnoses in exome or genome sequencing analysis (gene-centered reanalysis). One additional diagnosis was identified with this tight subsampling of variants among the 356 negative exomes, validating the proof of concept. Additional experiments could be performed to fully evaluate the ClinVarome, such as reanalysis with the full list of ClinVarome morbid genes not found in OMIM, additional cohorts, or an extended analysis considering the variants with different phenotypes not reported in the literature.

On the basis of literature data and feature engineering processes from all ClinVarome features during clustering model development, we identified 4 discriminative features for gene–disease clinical validity available in ClinVarome data. Overall, the evaluation relies mainly on the amount of knowledge but also on reported review confidence and more importantly on the time-scale of entries. The Genome Alert! gene-curation via machine learning methods provides an original attempt for automated evaluation of gene confidence in disease. Genome Alert! proposes a standardized clinical validity confidence score that could allow a prospective gene–phenotype association assessment. As such, this approach could be useful to update *in silico* gene panels. This procedure proposes a complementary approach to the aggregation of multiple expert-reviewed databases such as DDG2P, Genomic England PanelApp, or ClinGen gene–disease validity available in the GenCC database.³⁵ However, ClinVarome gene–disease validity confidence is defined for all diseases associated with a gene, which is less precise than curations submitted to the GenCC database. As ClinVarome is a more exhaustive database, this resource could prioritize genes to be curated by GenCC submitters, particularly in the first and second clusters.

In summary, Genome Alert! highlights changes with potential clinical significance and provides a large retrospective study of a partially automated system for sequencing data reinterpretation. This procedure enables the systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine, with a limited human resource effect and a diagnostic yield improvement. Genome Alert! provides an open-source accessible framework to the community, thus hoping to be applicable in every genetic center.

Data Availability

Software summary

Project name: Genome Alert!

Project home page: <https://genomealert.univ-grenoble-alpes.fr/>

Operating system(s): UNIX (Mac, Linux)

Programming language: Nim, Python, R

License: Apache Licence 2.0

Any restrictions to use by nonacademics: No

Genome Alert! results are publicly available at <https://genomealert.univ-grenoble-alpes.fr/>. Relevant data used to generate Genome Alert! results are available from ClinVar FTP (all monthly ClinVar full XML release data were downloaded from <https://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/>) and in the following resources: OMIM (<https://omim.org/>), Genomic England PanelApp (<https://panelapp.genomicsengland.co.uk/>), and RefSeq annotation (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz). All codes for generating Genome Alert! procedures are available at public GitHub repositories: ClinVCF tool for ClinVar XML full release processing and extraction to VCF format (<https://github.com/SeqOne/clinvcf>), Variant Alert! tool to compare ClinVCF release (https://github.com/SeqOne/variant_alert), ClinVarome tool to evaluate clinical validity of ClinVar morbid genes (<https://github.com/SeqOne/clinvarome>), and the Genome Alert! shiny app (https://github.com/SeqOne/GenomeAlert_app).

Acknowledgments

We sincerely thank all patients, clinicians, biologists, and bioinformaticians involved in this project. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

Author Information

Conceptualization: K.Y., F.L., S.Ca., D.B., A.-L.B., J.A., G.N., J.T., N.P.; Data Curation: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R.; Formal Analysis: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.C., Q.F., J.A.; Funding Acquisition: J.T., N.P.; Methodology: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Ca., Q.F., J.A.; Project Administration: A.-L.B., J.A., G.N., J.T., N.P.; Resources: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Software: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Supervision: A.-L.B., J.A., G.N., J.T., N.P.; Validation: J.A., G.N., J.T., N.P.; Visualization: A.-L.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.; Writing-original draft: K.Y., F.L., Q.F., Q.T., J.-M.H., D.B., G.N., J.T., N.P.; Writing-review and editing: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Co., M.B., R.L., Q.F., S.Ca., Q.T., A.D., N.S., J.-M.H., N.D.-F., A.-L.B., S.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.

Ethics Declaration

Patients referred to the Eurofins Biomnis laboratory, Cerba laboratory, and CHU de Rouen Molecular Genetics laboratory

provided written consent for analysis of their DNA using next-generation sequencing, including research analysis for the purpose of obtaining a molecular diagnosis. Sequencing samples were de-identified. Local Ethics Committee of the CHU Grenoble-Alpes approved the study. Patients or legal guardians provided informed written consent for genetic analyses in a medical setting. This research conforms to the principles of the Helsinki Declaration.

Conflict of Interest

K.Y., M.B., R.L., Q.F., A.D., N.S., D.B., A.-L.B., and N.D.-F. are partially or fully employed by SeqOne Genomics; J.M.-H., S.B., J.A., and N.P. hold shares in SeqOne Genomics; D.T., A.B., A.L., and J.-M.C. are partially or fully employed by Laboratoire Cerba. V.G. and L.R. are partially or fully employed by Laboratoire Eurofins Biomnis. All other authors declare no conflicts of interest.

Additional Information

The online version of this article (<https://doi.org/10.1016/j.gim.2022.02.008>) contains supplementary material, which is available to authorized users.

Affiliations

¹Institute for Advanced Biosciences, Centre de recherche UGA / Inserm U 1209 / CNRS UMR 5309, Grenoble, France; ²SeqOne Genomics, Montpellier, France; ³Department of Genetics and Reference Center for Developmental Disorders, Normandy Center for Genomic and Personalized Medicine, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, F 76000, Rouen, France; ⁴Laboratoire Cerba, Saint-Ouen-l'Aumône, France; ⁵Laboratoire Eurofins Biomnis, Lyon, France; ⁶Unité Fonctionnelle de Cardiogénétique et Myogénétique, Centre de Génétique, Hôpitaux Universitaire Pitié Salpêtrière-Charles Foix, Paris, France; ⁷Grenoble Institut Neurosciences, GIN, Inserm U1216, Université de Grenoble Alpes, Grenoble, France; ⁸Medical Genetic Department for Rare Diseases and Personalized Medicine, Montpellier University Hospital, Montpellier, France; ⁹Soins Intensifs Néphrologiques et Rein Aigu, Hôpital Tenon, Assistance Publique des Hôpitaux de Paris, Paris, France; ¹⁰UMR_S1155, INSERM, Sorbonne Université, Paris, France

References

- Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med*. 2018;379(14):1353–1362. <http://doi.org/10.1056/NEJMra1711801>.
- Shendure J, Findlay GM, Snyder MW. Genomic medicine—progress, pitfalls, and promise. *Cell*. 2019;177(1):45–57. <http://doi.org/10.1016/j.cell.2019.02.003>.
- Dollfus H. Le plan France Médecine Génomique 2025 et les maladies rares. *Med Sci (Paris)*. 2018;34(Hors série n°1):39–41. <http://doi.org/10.1051/medsci/201834s121>.
- Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*. 2020;583(7814):96–102. <http://doi.org/10.1038/s41586-020-2434-2>.
- Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–424. <http://doi.org/10.1038/gim.2015.30>.
- Nykamp K, Anderson M, Powers M, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105–1117. Published correction appears in *Genet Med*. 2020;22(1):240–242. <https://doi.org/10.1038/gim.2017.37>.
- Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*. 2018;20(9):1054–1060. <http://doi.org/10.1038/gim.2017.210>.
- Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978–1980. <http://doi.org/10.1093/bioinformatics/bty897>.
- Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet Med*. 2018;20(6):645–654. <http://doi.org/10.1038/gim.2017.162>.
- Wright CF, McRae JF, Clayton S, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*. 2018;20(10):1216–1223. <http://doi.org/10.1038/gim.2017.246>.
- Bombard Y, Brothers KB, Fitzgerald-Butt S, et al. The responsibility to recontact research participants after reinterpretation of genetic and genomic research results. *Am J Hum Genet*. 2019;104(4):578–595. <http://doi.org/10.1016/j.ajhg.2019.02.025>.
- Clayton EW, Appelbaum PS, Chung WK, Marchant GE, Roberts JL, Evans BJ. Does the law require reinterpretation and return of revised genomic results? *Genet Med*. 2021;23(5):833–836. <http://doi.org/10.1038/s41436-020-01065-x>.
- Carrieri D, Howard HC, Benjamin C, et al. Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics. *Eur J Hum Genet*. 2019;27(2):169–182. <http://doi.org/10.1038/s41431-018-0285-1>.
- Deignan JL, Chung WK, Kearney HM, et al. Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2019;21(6):1267–1270. <http://doi.org/10.1038/s41436-019-0478-1>.
- Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. *N Engl J Med*. 2019;380(25):2478–2480. <http://doi.org/10.1056/NEJMc1812033>.
- James KN, Clark MM, Camp B, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ Genom Med*. 2020;5:33. <http://doi.org/10.1038/s41525-020-00140-1>.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(Database issue):D789–D798. <http://doi.org/10.1093/nar/gku1205>.
- Rehm HL, Berg JS, Brooks LD, et al. ClinGen—the clinical genome resource. *N Engl J Med*. 2015;372(23):2235–2242. <http://doi.org/10.1056/NEJMs1406261>.
- Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36(10):915–921. <http://doi.org/10.1002/humu.22858>.

20. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(Database issue):D980–D985. <http://doi.org/10.1093/nar/gkt1113>.
21. Tumienė B, Maver A, Writzl K, et al. Diagnostic exome sequencing of syndromic epilepsy patients in clinical practice. *Clin Genet.* 2018;93(5):1057–1062. <http://doi.org/10.1111/cge.13203>.
22. Pengelly RJ, Ward D, Hunt D, Mattocks C, Ennis S. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Sci Rep.* 2020;10(1):3235. <http://doi.org/10.1038/s41598-020-60215-y>.
23. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;1(1):73–79. <http://doi.org/10.1002/widm.2>.
24. Thomann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun.* 2019;10(1):2373. <http://doi.org/10.1038/s41467-019-10016-3>.
25. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51(11):1560–1565. <http://doi.org/10.1038/s41588-019-0528-2>.
26. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434–443. Published correction appears in *Nature.* 2021;590(7846):E53. Published correction appears in *Nature.* 2021;597(7874):E3–E4. <https://doi.org/10.1038/s41586-020-2308-7>.
27. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886–D894. <http://doi.org/10.1093/nar/gky1016>.
28. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 2020;12(1):103. <http://doi.org/10.1186/s13073-020-00803-9>.
29. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 2014;42(22):13534–13544. <http://doi.org/10.1093/nar/gku1206>.
30. Matalonga L, Hernandez-Ferrer C, Piscia D, et al. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur J Hum Genet.* 2021;29(9):1337–1347. Published correction appears in *Eur J Hum Genet.* 2021;29(9):1466–1469. <https://doi.org/10.1038/s41431-021-00852-7>.
31. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. *Hum Mutat.* 2018;39(11):1623–1630. <http://doi.org/10.1002/humu.23641>.
32. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med.* 2016;375(7):655–665. <http://doi.org/10.1056/NEJMsa1507092>.
33. Li D, Shi Y, Li A, et al. Retrospective reinterpretation and reclassification of BRCA1/2 variants from Chinese population. *Breast Cancer.* 2020;27(6):1158–1167. <http://doi.org/10.1007/s12282-020-01119-7>.
34. Bruel AL, Nambot S, Quéré V, et al. Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *Eur J Hum Genet.* 2019;27(10):1519–1531. <http://doi.org/10.1038/s41431-019-0442-1>.
35. Lazo de la Vega L, Yu W, Machini K, et al. A framework for automated gene selection in genomic applications. *Genet Med.* 2021;23(10):1993–1997. <http://doi.org/10.1038/s41436-021-01213-x>.