Title

Exome copy number variant detection, analysis, and classification in a large cohort of families with undiagnosed rare genetic disease.

Permalink

https://escholarship.org/uc/item/3s20v4vs

Journal

American Journal of Human Genetics, 111(5)

Authors

Lemire, Gabrielle
Sanchis-Juan, Alba
Russell, Kathryn
et al.

Publication Date

2024-05-02

DOI

10.1016/j.ajhg.2024.03.008

Peer reviewed

Exome copy number variant detection, analysis, and classification in a large cohort of families with undiagnosed rare genetic disease

Authors

Gabrielle Lemire, Alba Sanchis-Juan,
Kathryn Russell, ..., Michael E. Talkowski,
Harrison Brand, Anne O'Donnell-Luria

Correspondence

glemiret@broadinstitute.org (G.L.),
odonnell@broadinstitute.org (A.O.-L.)

Lemire et al. applied copy number variant (CNV) detection on exome sequencing from a cohort of 6,633 families with undiagnosed rare genetic disorders. With the resolution provided by exome sequencing, they identified a causative CNV in 2.6% of families and assessed CNV pathogenicity by applying an advanced classification approach.

Lemire et al., 2024, The American Journal of Human Genetics *111*, 863–876
May 2, 2024 © 2024 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2024.03.008

ARTICLE

Exome copy number variant detection, analysis, and classification in a large cohort of families with undiagnosed rare genetic disease

Gabrielle Lemire,[1,2,3,4,5,30,*] Alba Sanchis-Juan,[1,2,4,5,30] Kathryn Russell,[1,2] Samantha Baxter,[1,2] Katherine R. Chao,[1,2,5] Moriel Singer-Berk,[1,2,5] Emily Groopman,[1,2,3] Isaac Wong,[1,2,5] Eleina England,[1,2,3] Julia Goodrich,[1,2,5] Lynn Pais,[1,2,3,5] Christina Austin-Tse,[1,2,5] Stephanie DiTroia,[1,2,3,5] Emily O'Heir,[1,2,3,5] Vijay S. Ganesh,[1,2,3,4,5,6] Monica H. Wojcik,[1,2,3,4,15] Emily Evangelista,[1,2] Hana Snow,[1,2] Ikeoluwa Osei-Owusu,[1,2,5] Jack Fu,[1,2,4,5] Mugdha Singh,[1,2,3,4,5] Yulia Mostovoy,[1,2,5] Steve Huang,[1,2] Kiran Garimella,[1,2] Samantha L. Kirkham,[3] Jennifer E. Neil,[3,7] Diane D. Shao,[3,4,8] Christopher A. Walsh,[2,3,4,7] Emanuela Argilli,[9,10] Carolyn Le,[9,10] Elliott H. Sherr,[9,10] Joseph G. Gleeson,[11,12] Shirlee Shril,[4,13] Ronen Schneider,[4,13] Friedhelm Hildebrandt,[4,13] Vijay G. Sankaran,[2,4,14] Jill A. Madden,[3,15] Casie A. Genetti,[3,15] Alan H. Beggs,[2,3,4,15] Pankaj B. Agrawal,[2,3,4,15] Kinga M. Bujakowska,[2,4,16] Emily Place,[2,4,16]

(Author list continued on next page)

Summary

Copy number variants (CNVs) are significant contributors to the pathogenicity of rare genetic diseases and, with new innovative methods, can now reliably be identified from exome sequencing. Challenges still remain in accurate classification of CNV pathogenicity. CNV calling using GATK-gCNV was performed on exomes from a cohort of 6,633 families (15,759 individuals) with heterogeneous phenotypes and variable prior genetic testing collected at the Broad Institute Center for Mendelian Genomics of the Genomics Research to Elucidate the Genetics of Rare Diseases consortium and analyzed using the seqr platform. The addition of CNV detection to exome analysis identified causal CNVs for 171 families (2.6%). The estimated sizes of CNVs ranged from 293 bp to 80 Mb. The causal CNVs consisted of 140 deletions, 15 duplications, 3 suspected complex structural variants (SVs), 3 insertions, and 10 complex SVs, the latter two groups being identified by orthogonal confirmation methods. To classify CNV variant pathogenicity, we used the 2020 American College of Medical Genetics and Genomics/ClinGen CNV interpretation standards and developed additional criteria to evaluate allelic and functional data as well as variants on the X chromosome to further advance the framework. We interpreted 151 CNVs as likely pathogenic/pathogenic and 20 CNVs as high-interest variants of uncertain significance. Calling CNVs from existing exome data increases the diagnostic yield for individuals undiagnosed after standard testing approaches, providing a higher-resolution alternative to arrays at a fraction of the cost of genome sequencing. Our improvements to the classification approach advances the systematic framework to assess the pathogenicity of CNVs.

Introduction

Copy number variants (CNVs) are imbalances of genomic material compared with the reference genome resulting in the addition (duplications and insertions) or removal (deletions) of genomic segments. CNVs and other types of structural variants (SVs) such as balanced translocations and inversions, can vary in size but have traditionally been defined as variants of more than 50 bp[1–3] and are significant contributors to rare genetic disease.[4,5] Chromosomal microarrays (CMAs) have been the recommended first-tier clinical test to investigate individuals with suspected rare genetic diseases, especially for multiple congenital anomalies and intellectual disability disorders,[6,7] though practice is moving toward exome sequencing as a first-line test.[8] Standard clinical CMAs typically only detect CNVs larger than 50–100 kilobases precluding detection of smaller gene- and exon-disrupting CNVs. Due to technical limitations from variable sequencing depth, CNVs are challenging to identify by standard exome sequencing,

[1]Broad Institute Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA; [4]Harvard Medical School, Boston, MA, USA; [5]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [6]Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA; [7]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA; [8]Department of Neurology, Boston Children's Hospital, Boston, MA, USA; [9]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA; [10]Institute of Human Genetics and Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA; [11]Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA; [12]Rady Children's Institute for Genomic Medicine, San Diego, CA, USA; [13]Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA; [14]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA; [15]The Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA; [16]Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear, Boston, MA, USA; [17]Neuromuscular and Neurogenetic Disorders of Childhood Section, Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA;

(Affiliations continued on next page)

Eric A. Pierce,[2,4,16] Sandra Donkervoort,[17] Carsten G. Bönnemann,[17] Lyndon Gallacher,[18,19] Zornitza Stark,[18,19] Tiong Yang Tan,[18,19] Susan M. White,[18,19] Ana Töpf,[20] Volker Straub,[20] Mark D. Fleming,[4,21] Martin R. Pollak,[4,22] Katrin Õunap,[23,24] Sander Pajusalu,[23,24] Kirsten A. Donald,[25,26] Zandre Bruwer,[25,26] Gianina Ravenscroft,[27] Nigel G. Laing,[27] Daniel G. MacArthur,[1,2,28,29] Heidi L. Rehm,[1,2,4,5] Michael E. Talkowski,[1,2,4,5] Harrison Brand,[1,2,4,5,31] and Anne O'Donnell-Luria[1,2,3,4,5,15,31,*]

which typically focuses on single-nucleotides variants (SNVs) and indels.

Traditionally, exome-based CNV algorithms[9–11] have relied on exome read depth to inform of the underlying copy number at a given locus. However, many factors influence exome read depth, so detecting CNVs from exome data is difficult due to the non-uniform distribution of captured reads caused by biases introduced by PCR amplification, exome capture, and mapping. These factors make it challenging to differentiate between a technical artifact and a bona fide CNV. The GATK-gCNV tool[12] uses a probabilistic framework to infer rare CNVs from read depth data in the presence of these systematic biases. The performance of GATK-gCNV has been benchmarked with genome sequencing; it achieved more than 95% sensitivity for rare CNVs when compared against matched genome sequencing in 7,962 samples.[12]

We used the GATK-gCNV algorithm to call CNVs across the Broad Institute Center for Mendelian Genomics (Broad CMG) exome cohort, a research center within the Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) consortium. The Broad CMG has performed exome sequencing on more than 6,000 families with a suspected genetic disease since 2016, representing a large cohort of individuals with heterogeneous phenotypes, including neurodevelopmental disorders, neuromuscular diseases, retinal disorders, blood disorders, kidney diseases, multiple malformations syndromes, and other conditions. Most individuals in this cohort have had prior testing by gene panels, exome sequencing, and/or clinical CMA, but the level of prior genetic testing is variable. Several molecular diagnostic laboratories and many research groups have incorporated CNV calling in their exome analysis, particularly in recent years. The reported additional diagnostic yield of CNV calling on exome data, most commonly used as a second-line test after CMA, on various cohorts of individuals with suspected rare genetic diseases varies between 1% and 2%.[13–17]

The widespread implementation of CMA and exome/genome sequencing is expanding the types and numbers of CNVs identified in both clinical and research settings, and it can be challenging to determine the impact of these CNVs on human health. Several resources have been or are being developed to address this challenge. For instance, high quality reference population data such as gnomAD SV v4[3] (a reference dataset of SVs from short-read genome sequencing of 63,046 individuals from the general population), and gnomAD CNV v4 (a reference dataset of CNVs from exome sequencing of 464,297 individuals from the general population) help determine the frequency of a CNV in the population. Also, *in silico* prediction tools for CNVs are available, including some that have been developed with the goal of helping to distinguish deleterious CNVs from non-deleterious CNVs. For example, the StrVCTVRE score is a predictive tool that incorporates gene importance, conservation, coding sequence, and exon structure of the disrupted region and can evaluate CNVs overlapping coding sequences.[18] CADD-SV, another example, is a tool developed using machine-learning random forest models to differentiate deleterious from neutral SVs.[19]

Importantly, accurate classification of CNV pathogenicity requires a consistent and transparent approach to be used across the human genetics field. Riggs et al. developed the 2020 American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen) consensus standards to guide in the evaluation of germline CNVs and encourage consistency in CNV interpretation across laboratories, technologies, and specialties.[20] They proposed a quantitative evidence-based evaluation framework to classify copy number loss and copy number gain that follow an autosomal-dominant inheritance. The curation process is divided into five sections: assessment of genomic content, overlap with established haploinsufficient or triplosensitive regions, evaluation of the number of genes in the CNV, evaluation of

[18]Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia; [19]Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Parkville, VIC, Australia; [20]John Walton Muscular Dystrophy Research Centre, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK; [21]Department of Pathology, Boston Children's Hospital, Boston, MA, USA; [22]Division of Nephrology, Beth Israel Deaconess Medical Center, Boston, MA, USA; [23]Department of Clinical Genetics, Genetics and Personalized Medicine Clinic, Tartu University Hospital, Tartu, Estonia; [24]Department of Genetics and Personalized Medicine, Institute of Clinical Medicine, Faculty of Medicine, University of Tartu, Tartu, Estonia; [25]Department of Paediatrics and Child Health, Red Cross War Memorial Children's Hospital, Cape Town, South Africa; [26]Neuroscience Institute, University of Cape Town, Cape Town, South Africa; [27]University of Western Australia Centre for Medical Research, Harry Perkins Institute of Medical Research, QEII Medical Centre, Nedlands, WA, Australia; [28]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW, Sydney, NSW, Australia; [29]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, VIC, Australia

[30]These authors contributed equally

[31]Senior authors

*Correspondence: glemiret@broadinstitute.org (G.L.), odonnell@broadinstitute.org (A.O.-L.)

https://doi.org/10.1016/j.ajhg.2024.03.008.

cases with the variant in the literature and databases, and scoring of the variant based on phenotype specificity and segregation in the family being studied. These standards did not intend to cover all curation scenarios and, for example, do not extend to guidance to evidence types used for SNVs and indels, such as how to score CNVs following an autosomal-recessive or X-linked inheritance pattern (allelic data), CNVs with available functional evidence, or SVs beyond deletions and duplications. Here, we developed and applied additional evidence criteria to address these limitations and assess the pathogenicity of all CNVs that were thought to be causal in the Broad CMG exome cohort.

Subjects, material, and methods

Case selection

The Broad CMG was established in 2016 as part of an initiative funded by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health with the goal of discovering the variants and genes underlying Mendelian disease to increase diagnosis rates for individuals with a suspected genetic condition.[21–23] The Broad CMG is now part of the NHGRI GREGoR consortium, the focus of which includes evaluating different approaches to improve rare disease diagnosis, such as CNV calling on exome data. Undiagnosed families recruited and sequenced through the Broad CMG are enrolled in research studies with local institutional review board approval, including for sharing de-identified samples for sequencing and analysis (MassGeneralBrigham 2013P001477). Informed consent was obtained for each recruited individual. Phenotypic information for the affected individuals in each family was provided using HPO terms.[24]

From February 2016 to May 2021 (5 years, 3 months), 6,633 undiagnosed families underwent CNV calling (integrated with SNV/indel analysis) on exome data through the Broad CMG (15,759 individuals, through multiple callsets). This cohort had heterogeneous phenotypes including neurodevelopmental, neuromuscular, multiple congenital anomalies, hematological, ocular, or renal disorders. Most were enrolled due to an unrevealing prior genetic diagnostic evaluation as many had a CMA, gene panel sequencing for known causes of disease, or clinical exome prior to research exome sequencing through the CMG. The sequenced individuals were submitted from a large number of studies and had variable levels of prescreening prior to enrollment (and this information was not systematically collected).

Exome sequencing

Exome sequencing was performed by the Genomics Platform at the Broad Institute of MIT and Harvard. Libraries from DNA samples (>250 ng of DNA, at >2 ng/ul) were created with an Illumina Nextera exome capture (37 Mb target) and sequenced (150 bp paired reads) to cover >80% of targets at 20× and a mean target coverage of >80× from February 2016 through January 2019 and then using a Twist exome capture (38 Mb target) and sequenced (150 bp paired reads) to cover >80% of targets at >20× and a mean target coverage of >60× thereafter. Sample identity quality assurance checks were performed on each sample. The exome data was de-multiplexed and each sample's sequence data were aggregated into a single Picard CRAM file. The BWA aligner was used for mapping reads to the human genome build 38 (GRCh38). SNVs and insertions/deletions (indels) were jointly called across all samples using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 3.5. Default filters were applied to SNV and indel calls using the GATK Variant Quality Score Recalibration (VQSR) approach. Annotation was performed using Variant Effect Predictor (VEP) during upload of the callset to seqr[25] for collaborative analysis between the Broad CMG team and collaborating investigators.

CNV detection on exome data

CNVs were detected from exome sequencing following GATK-gCNV best practices,[12] as follows: read coverage was first calculated for each exome using GATK CollectReadCounts. After coverage collection, all samples were subdivided into batches of a median of 410 samples (range: 160–625) for gCNV model training and execution; these batches were determined based on a principal components analysis (PCA) of sequencing read counts. After batching, one gCNV model was trained per batch using GATK GermlineCNVCaller on a subset of training samples, and the trained model was then applied to call CNVs for each sample per batch. Finally, all raw CNVs were aggregated across all batches and post-processed using quality- and frequency-based filtering to produce the final CNV callset. Methods are further described in Babadi et al.[12]

CNV analysis

Each family's CNV data was manually analyzed in coordination with the SNV/indel data by members of the Broad CMG analysis team using our in-house-developed analysis platform, *seqr,* an open-source, web-based tool for family-based monogenic disease analysis that enables project management, variant filtration, annotation, and prioritization in addition to data sharing of candidate disease genes (with variants and HPO terms) through the Matchmaker Exchange.[25] CNVs were filtered based on their mode of inheritance, gCNV quality score (QS) (QS > 50; see Babadi et al.[12] for details), and their frequency in the Broad CMG callset composed of 21,256 individuals (including the 15,759 individuals included for this study and additional samples). For autosomal-dominant conditions, we filtered for CNVs with an allele frequency of <0.1% in the Broad callset and used <1% for autosomal-recessive conditions. When analyzing each family, factors used to help prioritize if a CNV was of clinical significance for a given individual included the CNV size, its structural consequences (predicted loss-of-function [LoF] variant, copy gain), its segregation pattern within the affected family, its frequency in the gnomAD-SV (v2)[3] reference population database, the number and characteristics of genes involved in the CNV, and *in silico* prediction of pathogenicity tools. Of note, the following criteria needed to be met for an SV in gnomAD to be considered as the same allele:

(1) Same SV type (duplication, deletion, etc.).

(2) Either has sufficient reciprocal overlap (50% reciprocal overlap for large SV >5 Kb; 10% reciprocal overlap for SV <5 Kb).

Genes included in a CNV were evaluated for gnomAD gene constraint scores[26] (LOEUF, pLI), ClinGen dosage sensitivity scores, and disease association in OMIM; exons included in an intragenic CNV were evaluated for exon expression (pext score in gnomAD[27]) and conservation. The probability of a gene being dosage sensitive as defined by Collins et al. (haploinsufficiency [HI] and triplosensitivity [TS] scores) was also taken into account when evaluating genes included in a CNV.[28] If no promising variants were found using our initial searches, we removed the QS filter to include low-quality variants. We reviewed the StrVCTVRE score[18] of candidate CNVs but did not use it to filter data or rule out variants. The score ranges from 0 to 1, a score of 1 being more deleterious. In line with the developer suggestions, CNVs with a score >0.37 were considered as having a higher likelihood of being deleterious. To evaluate the quality of a given CNV, the proband's copy number level was compared to any additional sequenced family members as well as a cluster of other samples with similar read depth that act as controls. The copy number plot of each compelling candidate was assessed to confirm an increase or decrease (corresponding to either a gain or a loss) between the proband and the background cluster, and a difference in the proband's copy number within vs. outside the reported coordinates of the CNV (Figure 1). We also visually inspected the read data of candidate CNVs using the Integrated Genomics Viewer (IGV) to evaluate for sequencing artifacts (Figure 1).

A CNV is defined as high-confidence by GATK-gCNV (see Babadi et al.[12] for details) if the following criteria are met:

(1) The CNV is present in a high-quality sample (with ≤200 autosomal raw CNV calls of which at least 35 have QS >20).
(2) The sample frequency of the call is ≤0.01 within the Broad callset.
(3) The number of overlapped exons is ≥3.
(4) The QS is equal or greater than the QS threshold as defined in Babadi et al.[12] (QS >50 for duplications, >100 for deletions, and >400 for homozygous deletions).

CNV confirmation

CNV confirmations were performed by the investigator that contributed the sample by a variety of methods (including FISH [fluorescent in situ hybridization], karyotype, CMA, MLPA [multiplex ligation-dependant probe amplification], Sanger sequencing, quantitative PCR, droplet digital [dd] PCR,[29] or genome sequencing) across different clinical or research laboratories, while some were confirmed by short-read or long-read genome sequencing performed at the Broad Genomics Platform (Table S1). Not all CNVs identified by the gCNV pipeline were confirmed by another method, largely when samples were from historic cohorts where there was not a path to return results or there was insufficient remaining DNA.

Evaluation of CMA coverage for each causal CNV

To evaluate how many causal CNVs could have been detected by a standard clinical CMA, CNV detection sensitivity by CMA was assessed by evaluating the number of probes from the Agilent GenetiSure Cyto CGH+SNP arrays (downloaded from https://genome.ucsc.edu/ on May 23, 2023) included within the genomic coordinates of a given CNV. A minimum number of five probes was required to consider that the CNV would confidently be called by CMA.[30]

Assessment of the pathogenicity of CNVs

We considered a case solved if the CNV was classified as pathogenic or likely pathogenic and conclusively explained the phenotype or if a variant was found involving a novel disease gene (here defined as a gene with no disease association in OMIM) with moderate/strong supporting evidence by the ClinGen gene-disease validity criteria.[31] Supporting genetic and/or experimental evidence were required to consider a CNV in a novel gene as the diagnosis in a given family, most often by additional families identified through Matchmaker Exchange. We also considered a case solved when the analysis team and referring provider, when relevant, considered the variant causative, even if a CNV was technically a variant of uncertain significance (VUS) by ACMG/ClinGen CNV criteria. For heterozygous CNVs in genes associated with a condition that follow an autosomal recessive inheritance pattern, the presence of a second variant involving that gene (confirmed or suspected compound heterozygous state) was required to consider the case solved.

Each CNV was evaluated and classified by two curators (G.L. and K.R.). In order to systematically assess the pathogenicity of the SVs in this study, the ACMG/ClinGen standards for interpretation and reporting of constitutional copy-number variants were applied.[20] For candidate novel disease genes, the interpretation of gene-disease relationship was guided by the ClinGen framework.[31] We developed an approach, including new curation criteria, to optimally capture evidence for pathogenicity for the range of variants discussed in this article.

Determination of the number of protein-coding genes included in a CNV

In order to score points from section 3 from the Riggs standards ("evaluation of gene number"), we used OMIM gene number count (https://genescout.omim.org/) and have

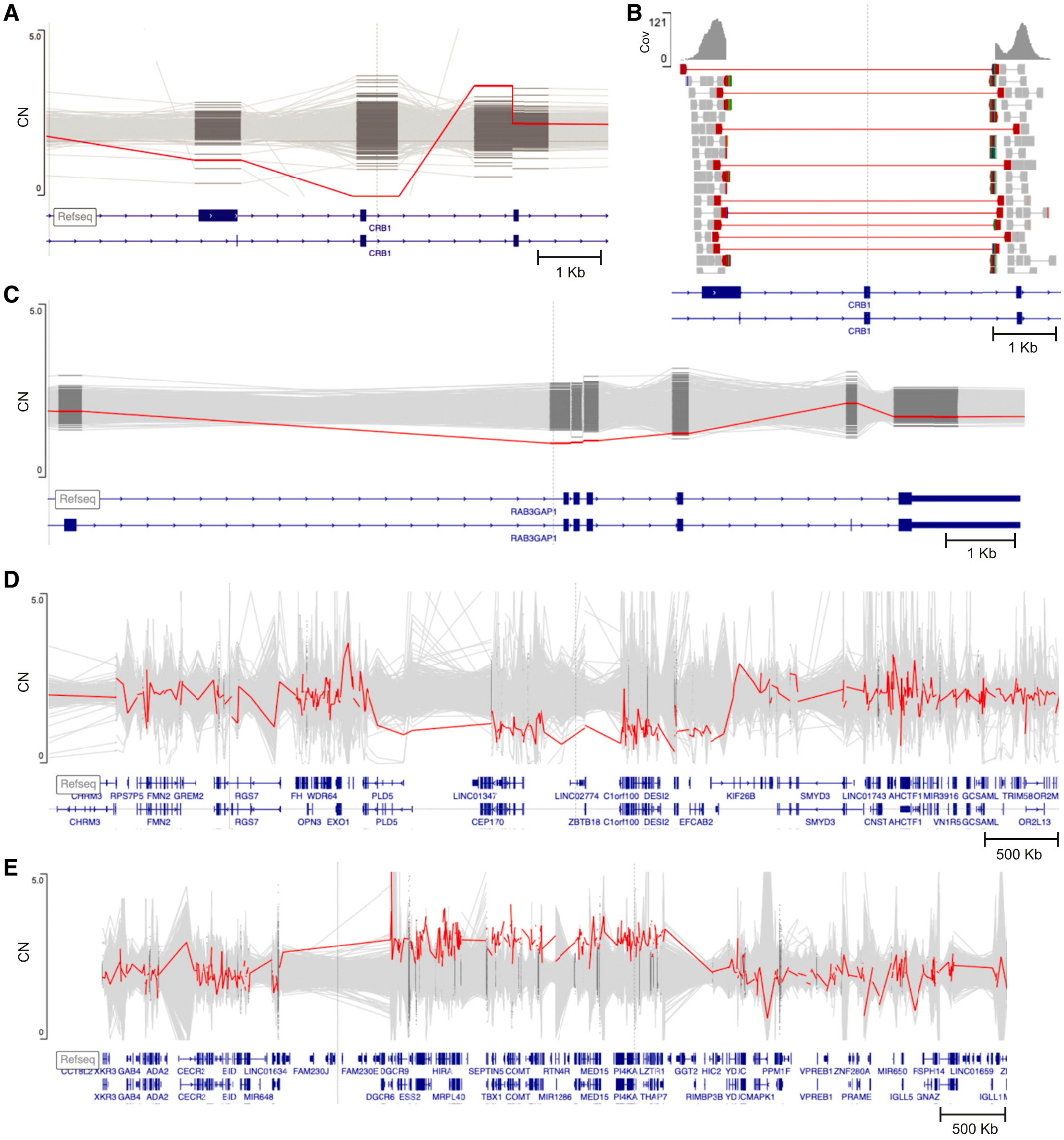


Figure 1. Exome copy number plot and reads visualization for examples of causal copy number variants (CNVs) in the Broad CMG cohort

(A) Individual affected with retinitis pigmentosa with a homozygous single-exon deletion in *CRB1* (chr1:197,438,450–197,439,442x0, quality score [QS] = 120) identified on exome. To evaluate the quality of the CNV, the patient's copy number (CN) level (in red) was compared to a cluster of other samples with similar read depth that act as controls. The proband's CN is decreased compared to the background cluster, compatible with a homozygous deletion. Y axis: CN.

(B) As breakpoints are located within the exome data, manual inspection of read data from the individual from (A) using the Integrated Genomics Viewer (IGV) showed discordant read pairs, split reads, and complete absence of coverage, compatible with a homozygous exon 10 deletion also including part of upstream exon 9 in *CRB1* (chr1:197,435,257–197,441,674x0 [GenBank: NM_201253.3]). Cov, coverage.

(C) Individual with multiple congenital anomalies and a heterozygous deletion of four exons in *RAB3GAP1* (Warburg micro syndrome) (red, chr2:135,162,318–135,164,794x1, QS = 92) in *trans* with a frameshift variant in *RAB3GAP1* (not shown, c.2393_2394del [GenBank: NM_012233.3] [p.Leu798ArgfsTer7]), both identified by exome. The presence of the deletion was validated by droplet digital PCR. Y axis: CN.

(D) Individual with a neurodevelopmental disorder with a *de novo* 2.6 Mb heterozygous 1q43q44 deletion (red, chr1:242,523,991–245,156,781x1, QS = 3077) identified on exome. The presence of this deletion was validated by quantitative PCR. Y axis: CN.

(E) Individual with a neurodevelopmental disorder with a *de novo* 2.1 Mb 22q11.2 duplication (red, chr22:18,985,739–21,081,116x3, QS = 3077) identified on exome. The presence of this duplication was validated by chromosomal microarray (CMA). Y axis: CN. All coordinates on GRCh38.

compared it to the gene number count provided by the DECIPHER browser (https://www.deciphergenomics.org/) and the ClinGen browser (https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=).

Variants following an autosomal-recessive inheritance pattern

The current ACMG/ClinGen CNV standards do not yet provide guidance on how to score CNVs in genes for

Table 1. **Adapted PM3 table to score CNVs in genes for conditions that follow an autosomal-recessive inheritance pattern**

	Points per proband	
Variant classification/ zygosity	**Confirmed in *trans***	**Phase unknown**
Second variant is pathogenic (P) or likely pathogenic (LP)	0.30	0.15 (P) 0.08 (LP)
Homozygous occurrence of this variant (max 0.30 point)	0.15	N/A
Second variant is a variant of uncertain significance (max 0.16 point)	0.08	0.0

conditions that follow an autosomal-recessive inheritance pattern. To classify these variants within this project, we developed an approach, advancing the current framework.

(1) We applied category 2E and the PVS1 LoF flowchart[32] for any intragenic CNV or if a CNV had a complete or partial overlap with a gene with an established gene-disease relationship that follows an autosomal recessive inheritance pattern.

(2) When the candidate CNV involved a gene with no established gene-disease relationship, we did not score points from category 2 but rather used category 4 to build up evidence for an established gene-disease relationship by finding additional cases with overlapping variants from the literature.

(3) Points were awarded to the Broad CMG cases and published cases from the literature using a similar system to that which is used when curating SNVs (the PM3 criteria) (ClinGen Sequence Variant Interpretation Recommendation for *in trans* Criterion [PM3] - Version 1.0 Working Group Page: https://clinicalgenome.org/working-groups/sequence-variant-interpretation/, Approved: May 2, 2019). The point-based system suggested in the PM3 criteria was translated into points of similar strength level in the Riggs quantitative framework[20] (Table 1).

(4) We added 0.15 points when at least one individual with a unique phenotype (phenotype is highly specific to disease, low genetic heterogeneity) has been reported by our study or in the literature (equivalent of PP4 criteria in Richards et al.[33]). In some cases, we awarded 0.30 points when evidence was particularly strong. This only applied for genetic diseases with a specific, unique phenotype, high clinical sensitivity testing (e.g., biochemical assays, enzyme deficiency assays, functional cytogenetic tests [e.g., chromosomal breakage study]), and consistent family history. These additional points were only used one time per variant.

Variants following an X-linked inheritance pattern

We developed the following flowchart to score points for CNVs with an X-linked inheritance pattern (Figure 2). We incorporated sex of proband, parental genotype, and parental affected status to score both the proband in which the X-linked variant was identified and, if applicable, any individual in the published literature or public databases that had variants of similar genomic content to the variant of interest. The points for each case could be increased or decreased based on phenotype specificity, by increments of 0.15 points and up to 0.45 points.

Complex SVs

We defined a complex SV as a complex rearrangement typically composed of three or more breakpoint junctions that cannot be characterized as a single canonical SV type.[34] Some complex SVs were suspected on exome CNV analysis and/or identified after further validation. As suggested by Riggs et al.,[20] when classifying complex rearrangements (for example a paired duplication inversion), we evaluated each CNV separately. The overall classification for the event was defaulted to the most deleterious classification (for example, if the deletion portion were classified as "pathogenic" and the duplication portion was classified as "uncertain significance," the entire SV was classified as "pathogenic").

Inversions and insertions

For variants initially called as deletion or duplication by GATK-gCNV in this cohort, some were identified as including inversions or insertions by confirmation methods. The Riggs et al.[20] standards do not provide guidance on how to score inversions or insertions. Therefore, we took guidance from Collins et al.,[35] which states that inversions can be evaluated as an LoF event if exactly one breakpoint falls within a gene or both breakpoints fall within the same gene and span at least one exon. Collins et al. also recommend evaluating a large insertion within an exon as an LoF event. We applied the ACMG/AMP PVS1 LoF criteria[32] as appropriate for such cases.

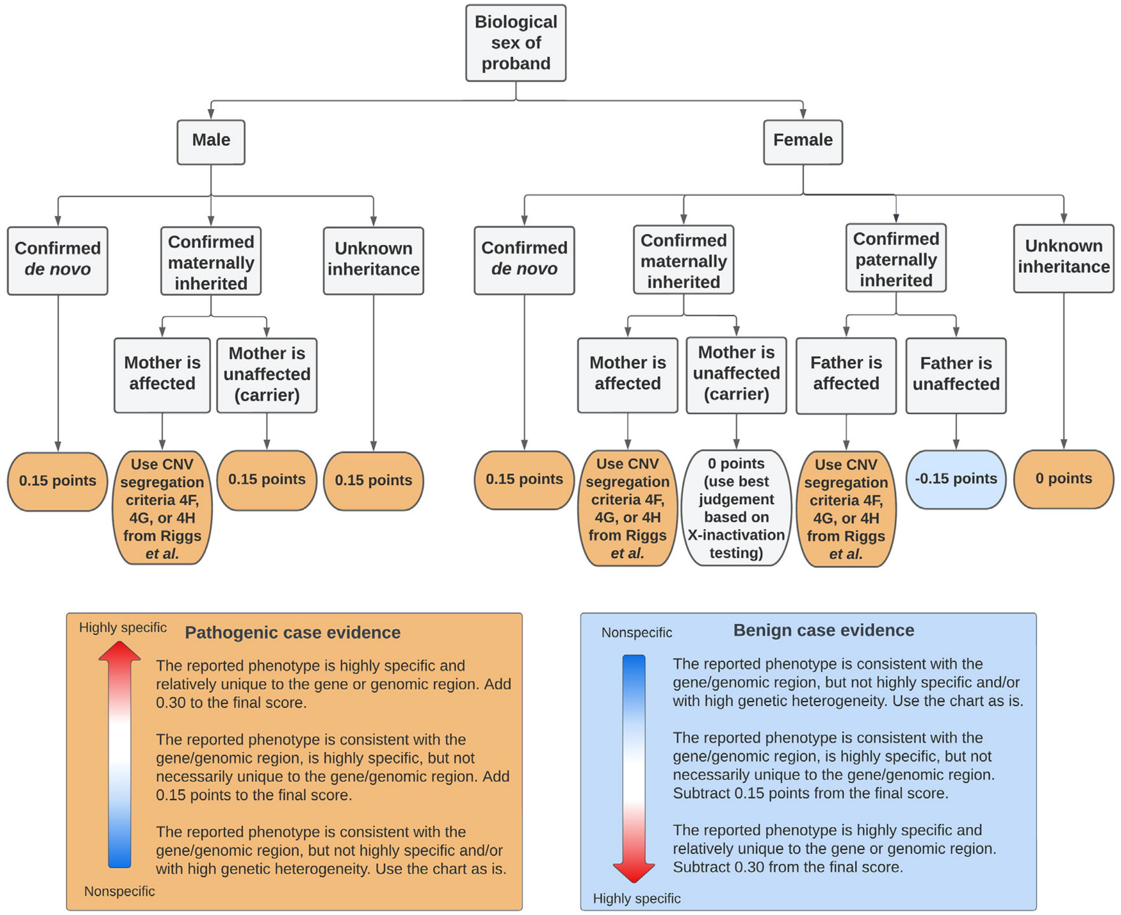
Variants with available functional evidence

We added an additional 0.15 points for any variant with at least supporting functional evidence of pathogenicity, either from the investigation of our cases or from the literature. Examples included expression assays (western blot for protein expression, PCR for RNA expression), RNA sequencing, cellular assays (impaired localization and/or function), or protein interaction studies. If the evidence was stronger, the points were upgraded to moderate (0.30 points) or strong (0.45 points). For example, RNA sequencing results showing a clear and significantly decreased expression of a gene or an animal model with the exact variant recapitulating the disease phenotype was given 0.45 point (strong evidence).

Results

CNV detection and analysis

CNV calling using the GATK-gCNV algorithm was performed, in parallel to SNV/indel analysis, on exomes



Biological sex of proband
Male
Female
Confirmed *de novo*
Confirmed maternally inherited
Unknown inheritance
Confirmed *de novo*
Confirmed maternally inherited
Confirmed paternally inherited
Unknown inheritance
Mother is affected
Mother is unaffected (carrier)
Mother is affected
Mother is unaffected (carrier)
Father is affected
Father is unaffected
0.15 points
Use CNV segregation criteria 4F, 4G, or 4H from Riggs *et al.*
0.15 points
0.15 points
0.15 points
Use CNV segregation criteria 4F, 4G, or 4H from Riggs *et al.*
0 points (use best judgement based on X-inactivation testing)
Use CNV segregation criteria 4F, 4G, or 4H from Riggs *et al.*
-0.15 points
0 points
Highly specific
Pathogenic case evidence
Nonspecific
Nonspecific
Benign case evidence
Highly specific

The reported phenotype is highly specific and relatively unique to the gene or genomic region. Add 0.30 to the final score.

The reported phenotype is consistent with the gene/genomic region, is highly specific, but not necessarily unique to the gene/genomic region. Add 0.15 points to the final score.

The reported phenotype is consistent with the gene/genomic region, but not highly specific and/or with high genetic heterogeneity. Use the chart as is.

The reported phenotype is consistent with the gene/genomic region, but not highly specific and/or with high genetic heterogeneity. Use the chart as is.

The reported phenotype is consistent with the gene/genomic region, is highly specific, but not necessarily unique to the gene/genomic region. Subtract 0.15 points from the final score.

The reported phenotype is highly specific and relatively unique to the gene or genomic region. Subtract 0.30 from the final score.

Figure 2. Flowchart illustrating how points were scored for copy number variants (CNVs) that followed an X-linked inheritance pattern

We incorporated sex of proband, parental genotype, and parental affected status to score both the proband in which the X-linked variant was identified and, if applicable, any individual in the published literature or public databases that had variants of similar genomic content to the variant of interest. The points for each case could be increased or decreased based on phenotype specificity, up to 0.45 points.

from the Broad CMG cohort of 6,633 undiagnosed families with heterogeneous rare disease phenotypes and variable prior genetic testing that typically included a gene panel, exome, and/or CMA. A total number of 9,930 high-confidence unique variants (4,387 deletions and 5,543 duplications) were identified across 15,759 individuals from these 6,633 families (Figures 3A and S1), 10,472 of the 15,759 samples had at least one rare (<1% frequency in the Broad data callset) high-confidence CNV, and the median number identified was two (SD + −1.55) per individual (Figure S1). The entire CNV callset for these individuals, with a total of 2,131,645 copy number calls (292,833 unique variants), was loaded into the seqr platform for analysis. Many of these low-quality calls were likely artifacts, but by incorporating phenotype and allelic variation (SNVs, indels, CNVs) in the analysis of each family, 26 low-quality CNV calls were prioritized and ultimately interpreted as causal. Through the entire callset analysis, we have identified a causal variant in 171 previously undiagnosed families. CNV calling on existing exome data in this cohort thus resulted in an additional solve rate of 2.6% (171/6,633). The causal CNVs consisted of 143 deletion, 15 duplication, and 13 suspected complex (multiple CNVs on a chromosome) GATK-gCNV calls, which are currently resolved as 140 deletions, 15 duplications, 3 insertions, 10 complex SVs, and 3 suspected complex SVs. Of the 10 validated complex SVs, three were initially deletion or duplication calls where a complex SV was identified on validation.

Causal CNVs results

The CNVs mostly involved established genes/loci, but five families that were considered solved had a CNV involving a novel disease gene candidate. Supporting genetic and/or experimental evidence was required to consider a CNV in a novel gene as the explanation for a given family, most often by additional families identified through Matchmaker Exchange[36] or the literature. Four of the five CNVs involving a novel gene included at least one haploinsufficient gene, as defined by Collins et al.[28] and gnomAD gene constraint scores.[26] The disorder followed an autosomal dominant inheritance in 93 families, an autosomal recessive inheritance in 60 families and X-linked inheritance in 18 families (Figure 3B). The CNV was confirmed

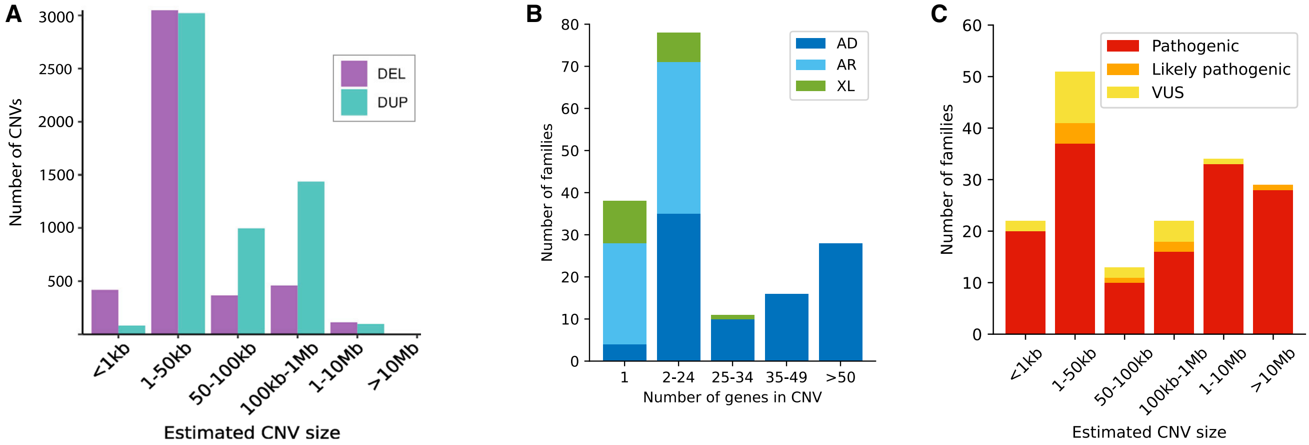


Figure 3. Characteristics of copy number variants (CNVs) across the entire callset and the subset of causal CNVs
(A) Number of high-confidence CNVs by estimated size that were identified in the Broad CMG exome callset of 6,633 families sequenced between 2016 and 2021. Large CNVs tend to be fragmented into multiple small GATK-gCNV calls, accounting for why there are no CNVs in the >10 Mb category of the graph. These CNVs were interpreted as being part of the same underlying event when looking at the copy number plot and/or validation methods and are presented that way in Figures 3B and 3C. DEL, deletion; DUP, duplication.
(B) Mode of inheritance and number of genes involved in each CNV in 171 families in which the CNV was interpreted as causal. The number of genes included in each interval was chosen based on cutoffs suggested for CNV scoring in section 3 of the Riggs et al. ACMG/ClinGen standards.[20]AD, autosomal dominant; AR, autosomal recessive; XL, X-linked.
(C) CNV classification by estimated size in 171 families in which the CNV was interpreted as causal by the multidisciplinary team. The causal CNVs consisted of 140 deletions, 15 duplications, 3 insertions (miscalled as deletion by GATK-gCNV), and 13 complex structural variants (SVs). We interpreted 151 CNVs as likely pathogenic/pathogenic and 20 CNVs as variants of uncertain significance (VUSs).

de novo in 70/93 (75%) of the families with an autosomal dominant disorder, inherited from a parent in 3/93 families (one inherited from an affected parent, one involving an imprinted locus, and one inherited from an unaffected parent for a condition known to harbor incomplete penetrance/variable expressivity), and the inheritance was unknown in 20/93 families. Of the 60 families having a disorder that follow an autosomal recessive inheritance pattern, the CNV was homozygous in 39 families, was in a confirmed compound heterozygous state with an SNV in seven families, and was in a presumed compound heterozygous state with an SNV for 14 families. The CNV was confirmed *de novo* in 7/18 (39%) of the families with an X-linked disorder. Detailed information on the CNV of each family is provided in Table S1. The predominant phenotype present in the 171 families was neurodevelopmental disorders (54%) followed by neuromuscular disorders (15%), but the cohort with causal CNVs also included individuals with multiple congenital anomalies, hematological, ocular, and renal phenotypes. The degree of prescreening before research exome differed between individuals from different sub-cohorts and was therefore nonuniform across different phenotypes.

The estimated sizes of causal CNVs by exome ranged from 293 bp to 80 Mb (Figure 3C). One-fifth (21%) of solved cases had a CNV below the benchmarked resolution of GATK-gCNV (22 one-exon and 14 two-exon CNVs), indicating it may be able to detect even smaller CNVs when allowing for a higher false positive rate. Large CNVs were also identified as some individuals did not have CMA prior to research enrollment. Large CNVs of more than 5–10 Mb tend to be fragmented into multiple small GATK-gCNV calls. We interpreted fragmented CNVs as being part of a larger CNV event in 35 families (35/171, 20%) in this cohort after looking at the copy number plot and/or confirmation methods.

Evaluation of CMA coverage for each causal CNV

We sought to evaluate how many of the causal CNVs could have been detected by evaluating probes on one of the standard clinical CMAs, which is distinct from a high-density clinical array, which often has one or more probes per exon. Standard CMAs usually detect CNVs larger than 50–100 Kb, but the resolution varies across the genome and across different array designs as the probes are not evenly spaced but are clustered around regions of clinical interest. CNV detection sensitivity by a representative standard CMA was assessed based on the minimum number of probes considered “sufficient” for CNV calling per target, which is defined as ≥5 probes for the Agilent GenetiSure Cyto array.[30] Based on this, we estimate that 44% (75/171) of these CNVs are unlikely to have been detected by standard CMA.

CNV confirmation

More than half of the CNVs (116/171, 68%) were confirmed by various orthogonal methods, such as CMA, PCR, FISH, karyotype, MLPA, Sanger across the CNV or breakpoints, or short- or long-read genome sequencing. Of note, some of these methods did not provide breakpoints but rather only confirmed the copy number change. Of the 116 confirmed CNVs, 35 (30%) showed differences when comparing the initial results with the orthogonal confirmation results: 24 showed differences in gene/exon content and 11 showed differences in SV type. For example, GATK-gCNV initially detected a one-exon deletion in *CLN3* (MIM: 607042) in the exome of five individuals, which was later corrected to a two-exon deletion by genome sequencing or Sanger sequencing. That single-exon deletion was in frame, and the addition of a second exon resulted in that CNV disrupting the reading frame, but neither CNV was classified as pathogenic. Importantly, the difference in gene or exon content identified in 24 families did not result in a change in the clinical interpretation of the CNV. Of note, only one of these 24 CNVs was curated as a VUS, and the difference in the number of exons included in the CNV did not change the scoring and classification of this CNV. The 11 cases with different SV type consisted of eight complex SVs, which were either incompletely characterized or not suspected by GATK-gCNV on the exome, and a recurrent Alu insertion in *MAK* (MIM: 154235)[37] identified in three individuals with retinitis pigmentosa. This insertion was miscalled as a deletion by the GATK-gCNV pipeline, but manual inspection of the exome reads showed discordant read pairs compatible with an Alu insertion. Sanger sequencing resolved the nature of this event.

Special categories of CNVs

Overall, there were 10 confirmed complex SVs in this cohort. We defined a complex SV as a complex rearrangement typically composed of three or more breakpoint junctions that cannot be characterized as a single canonical SV type.[34] A complex SV was suspected on the GATK-gCNV calls in 10 families (del/dup, paired dup, etc); seven of these were confirmed by genome, qPCR, or CMA (Table S1), and three remained unvalidated. Two deletions and one duplication identified by GATK-gCNV in three different families were revealed to be complex SVs (paired deletion inversions and a paired inversion duplication) when validated by genome sequencing or long-range PCR.

Twenty-five unrelated families with causal CNVs had a recurrent CNV that was identified in more than one other unrelated family in this cohort. The recurrent 22q11.2 microdeletion syndrome (MIM: 188400) was identified in nine individuals with neurodevelopmental disorders in this cohort. Two individuals with a neurodevelopmental disorder were diagnosed with 22q13.3 deletion syndrome (Phelan-McDermid syndrome [MIM: 606232]). The 17q12 deletion syndrome (MIM: 614527) was identified in two individuals with renal cystic disease. There were multiple recurrent CNVs identified in the subgroup of individuals with retinal disorders in this cohort. Indeed, five individuals of European ancestry affected with non-syndromic retinal degeneration had a heterozygous 2-exon-deletion in *CLN3*, a common variant in that gene,[38] in *trans* with a pathogenic variant.[39] A founder variant in the Ashkenazi Jewish population, an Alu insertion in *MAK*,[37,40] was found in three affected individuals of this ancestry. Two individuals of different ancestries affected with retinitis pigmentosa were homozygous for the same two-exon deletion in *EYS* (MIM: 612424), a deletion previously reported in the literature.[39,41] Two individuals of European ancestry affected with retinitis pigmentosa had a heterozygous four-exon deletion in *EYS*, a deletion reported in multiple affected individuals in the literature,[39,42–44] in *trans* with a pathogenic or likely pathogenic variant. Detailed information on the CNV of each of these families is provided in Table S1.

StrVCTVRE in silico score

The StrVCTVRE *in silico* score was evaluated across the cohort. This score was viewable on each CNV within seqr during the initial analysis but was not used for filtering and not strongly relied on in analysis (consistent with how other *in silico* scores are viewed in our analysis pipeline). Sharo et al. reported that a 90% sensitivity is reached at a StrVCTVRE score of 0.37 (score ranges from 0 to 1, a score of 1 being more deleterious) and observed on a collection of SVs called from a clinical cohort that this threshold may identify 90% of pathogenic SVs while reducing the candidate SV list by 54%.[18] In this cohort, 158/165 unique causal CNVs had a StrVCTVRE score greater than 0.37 (true positive rate of 96%), while this was the case for 6,162/10,788 non-causal CNVs (false positive rate 57%) (Table S2). The median score of the 158 unique causal CNVs was 0.78 and 0.42 for non-causal CNVs that had a StrVCTVRE score calculated. One minor limitation of this analysis is that many large CNVs are fragmented, which may result in lower StrVCTVRE scores for constituent parts than would be assigned for the larger CNV event. While we manually reassembled and recalculated StrVCTVRE scores for causal CNVs reported here (as it is appropriate to apply these scores to the entire CNV), non-causal CNVs were not reassembled. We note that all CNVs greater than 3 Mb size automatically had a score of 1, demonstrating a correlation between the CNV size and the StrVCTVRE score (Figure S2).

CNV classification

Using the 2020 ACMG/ClinGen CNV interpretation standards[20] and additional evidence criteria that we developed (detailed in subjects, material, and methods), we interpreted 151 CNVs as likely pathogenic/pathogenic and 20 CNVs as VUSs of high interest, including the five in novel disease-gene candidates (Figure 3C). When evaluating the pathogenicity of each CNV, we determined the number of protein-coding genes included in each CNV and compared that number to three different reference databases: OMIM (https://genescout.omim.org/), DECIPHER browser (https://www.deciphergenomics.org/browser), and ClinGen browser (https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=) (Table S1). The vast majority of CNVs (146/171, 85%) showed differences in gene number between these three commonly used databases. Using the 2020 ACMG/ClinGen CNV interpretation standards,[20] different points are scored based on the number of genes included in a CNV (section 3 of the standards). For example, 0 points are given for a deletion with 0–24 genes, 0.45 points for a deletion of 25–34 genes, and 0.9 points for a deletion of more than 35 genes. For copy gain, 0 points are given for 0–34 genes, 0.45 points are given for 35–49 genes, and 0.9 points for more than 50 genes. We used the number of genes provided by the OMIM database to perform the curation. Using the OMIM database vs. DECIPHER resulted in a different final score for 24/146 (16%) CNVs, but this would only alter the final classification for one CNV, as points were awarded from other sections. That altered case was an 857 kb *de novo* 22q13 duplication, which would be classified as a VUS if we use the gene number provided by OMIM (28 protein-coding genes) but would be classified as pathogenic if we had used DECIPHER browser (35 protein-coding genes). Detailed information on the CNV curation of each family is provided in Table S1.

Discussion

We present the analysis and curation results from CNV calling on exome data across a large and phenotypically heterogeneous cohort. The sequenced individuals in our cohort were submitted from a large number of studies and had variable levels of testing prior to enrollment. This is a limitation of the current study as the likelihood of identifying a causal CNV varied among individuals. The additional 2.6% solve rate of exome CNV calling identified in this cohort is nonetheless comparable to previously reported diagnostic yields of 1%–2% in other studies.[13–17] In this cohort, most causal CNVs were deletions. Duplications were more common in the callset but are less likely to disrupt gene function and also typically require more functional investigation to confirm a deleterious effect. Our callset contains many candidate duplications (and deletions) that could potentially elucidate additional affected families, but their pathogenicity remains uncertain and has not been further investigated.

Similar to using the probes on a microarray to estimate CNV size, the size of a CNV from exome analysis is an estimate based on which exons have an abnormal copy number, but the breakpoints typically occur somewhere intronic or intergenic. In addition, some exons have more heterogeneous coverage, and the deletion or duplication may involve more or fewer exons than predicted. This can also result in a large CNV being called as multiple smaller events, but when the data is reviewed, it can often be assembled into a larger event. Based on the probe set from a representative clinical CMA, we estimated that 44% of the causal CNVs in our cohort were unlikely to have been detected by standard CMA. CMAs with higher density probe coverage can often detect small exonic CNVs (depending on probe placement) and array-based methods will be more effective at detecting intronic and intergenic CNVs as these regions are not well covered in exome data.

In this study, we did not attempt to validate and map all the CNV breakpoints, and we did not assess the validation rate of GATK-gCNV as this has been done previously.[12] A small number of causal CNVs in this cohort were nonetheless confirmed by genome sequencing by the Broad CMG as part of initial efforts to validate gCNV performance. Of the 30 deletions and two duplications identified by GATK-gCNV and confirmed by genome sequencing, these were resolved as 29 deletions, one duplication, and two complex SVs. We recommended that any candidate CNV variants identified by exome be validated with an orthogonal method to confirm the event and clarify the breakpoints if possible (though note that breakpoints from CMA are not the true breakpoints either). This is particularly important for CNVs with QSs below the recommended thresholds or if altering the CNV call by an exon or two at either side of the event would change its interpretation (more likely to be an issue for smaller CNVs where such adjustments might affect whether the resulting deletion or duplication leads to an in-frame or out-of-frame transcript). The sensitivity of GATK-gCNV decays greatly for CNVs smaller than three exons (e.g., only ~50% for CNVs involving 1 exon), but the precision is relatively stable.[12] We might thus have underdetected some small causal CNVs in our cohort. Still, CNVs in 36 families (36/171, 21%) in this cohort involved fewer than three exons, highlighting the benefit of reviewing the full dataset with the context of the patient's phenotype and for some cases, a pathogenic variant in *trans* can highlight small or poor quality CNV calls that warrant further attention. More than half of CNVs were confirmed by various methods (including 85% [22/26] of causal CNVs that did not meet the standards to be considered a high-quality variant), and confirmation is either underway or may not be possible for the remainder of the identified CNVs. Importantly, the difference in size and in gene/exon content for confirmed CNVs did not lead to a change in the interpretation of any of the CNVs initially identified as causal, but it is possible that some interesting CNVs in this cohort were overlooked for that reason.

GATK-gCNV can only call deletions or duplications, so seven suspected complex SVs and three initially unsuspected complex SVs in this cohort were identified by orthogonal confirmation methods. We likely underdetected complex SVs as 32% of the CNVs in this cohort were not confirmed, and some confirmation methods would miss a more complex event, such as CMA, droplet digital, or quantitative PCR, which only confirm the abnormal copy number without mapping the breakpoints.

There are only a few *in silico* prediction tools available for CNV interpretation. Our group applied StrVCTVRE scores, and we observed that it was a useful tool to consider when prioritizing CNVs in this cohort. Generally, we use *in silico* predictions as accessory annotations for review when considering a variant rather than using it to filter out variants, even more so because large CNVs may be represented by multiple smaller fragmented calls. More data on analysis of cohorts of patients with rare diseases is needed to determine its utility overall and comparison to other available SV predictors. Of note, StrVCTVRE only provides a prediction score for CNVs overlapping a coding region, which was not a factor for this cohort given that it was exome based, but this is a limitation of the score when considering genome sequencing and noncoding SVs.

High-quality reference population data is essential for effective CNV analysis. The gnomAD SV v2 dataset stands as a pivotal resource in human genetics but is limited to sequencing data from short-read genomes. We used the database to evaluate if a given CNV was present in the general population, which we found was useful for variant analysis and prioritization. There is a myriad of technical differences between genome and exome sequencing and, while studies have shown high overlap between CNV calling between the two techniques, the recently released gnomAD CNV v4 dataset with GATK-gCNV calls on >400,000 individuals is anticipated to improve clinical CNV interpretation since they will be more analogous from a technical standpoint. As the gnomAD SV dataset expands in terms of size (incorporating both exome and genome data) and ancestral diversity, its utility as an invaluable tool for both rare disease diagnosis and broader genetic studies will only increase.

Standards for CNV classification are an important yet challenging area requiring ongoing development. We proposed new evidence criteria to enable the assessment of the pathogenicity of all CNVs that were thought to be causal in our cohort. We identified four areas that needed additions or refinements. First, we suggested that functional data, including expression assays (western blot, PCR, RNA sequencing) and cellular assays (localization/function), be incorporated as evidence at the supporting level of 0.15 points and could be increased in weight as appropriate. For example, abnormalities observed in RNA-sequencing data or an animal model with the same variant recapitulating the phenotype could be scored 0.3 or 0.45 points, respectively. Given the increasing availability of RNA sequencing, we suggest that incorporating scoring for functional evidence is essential for CNV classification. Second, to score CNVs involving genes associated with disorders with autosomal-recessive inheritance, we proposed an approach inspired by the ACMG/AMP criteria PM3 used for SNVs by incorporating phase and classification of the second variant (Table 1). The point-based system suggested in the PM3 criteria was translated into points of similar strength level in the Riggs quantitative framework. We also used the PVS1 flowchart[32] (or criteria 2E in Riggs et al.[20]) for intragenic CNVs or CNVs including at least one gene that had an established gene-disease relationship following an autosomal-recessive inheritance pattern. Additional points were added based on phenotype specificity and familial segregation. Third, to classify CNVs that follow an X-linked inheritance pattern, we developed a scoring system based on biological sex of the proband, parental genotype, and affected status of the transmitting parent (Figure 2). Points were upgraded by one or two strength levels based on phenotype specificity. We also used the PVS1 flowchart[32] for intragenic CNVs or CNVs including at least one gene that had an established gene-disease relationship following an X-linked inheritance pattern. Finally, to evaluate SVs other than deletion and duplication, we took guidance from Collins et al.,[35] which states that LoF can be expected if there is an insertion within an exon, if an inversion breakpoint falls within a gene, or if both inversion breakpoints fall within the same gene and span at least one exon. We thus applied the PVS1 LoF flowchart here. Our approach refined multiple aspects of CNV classification and advanced the systematic framework to assess the pathogenicity of CNVs.

An important step in CNV classification involves determining the number of protein-coding genes it contains. We observed some significant differences in gene number in CNVs evaluated in this cohort depending on which database was queried, the OMIM database being the most conservative. OMIM's gene count results from manual curation of published references while DECIPHER extracts this information directly from the Ensembl GRCh38 genome. OMIM might thus underestimate the real number of genes present in a CNV and DECIPHER might overestimate it. Even though different points were scored for several CNVs, the choice of which database to use did not affect the final classification except for one duplication in this cohort. For that duplication, the genes that were missing in OMIM but included in DECIPHER consisted of seven protein-coding genes. Our group opted for a conservative approach and used the OMIM database, but this question needs to be further studied as this can lead to confusion during the curation process. In addition, a sliding scale to score progressive points based on the increasing number of genes in a given CNV could be used instead of fixed cutoffs, and features such as LoF constraint, HI, and TS scores could be incorporated.

Conclusion

CNV calling and analysis from existing exome data increases the solve rate by 2.6% in this diverse and presumed monogenic cohort. This is a higher resolution alternative to CMA at a fraction of the cost of genome sequencing and can be applied retrospectively to existing exome datasets. We estimate that 44% of the 171 causal CNVs may not have been detected by standard clinical CMAs. In classifying these variants, we advanced the current standards to take into account additional types of evidence contributing to the systematic framework to assess the pathogenicity of CNVs.

Data and code availability

The CNVs that were interpreted as causal in this cohort were submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) (submitter ID 506627, Broad Rare Disease Group). The ClinVar accession numbers of each CNV are listed in Table S1.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.03.008.

Acknowledgments

We thank the families who participated in this study for sharing their samples and medical data along with all the research groups who collaborate with the Broad CMG. G.L. was supported by the Fonds de recherche en santé du Québec (FRQS). A.S.-J. was supported by a Massachusetts General Hospital (MGH) Fund for Medical Discovery Research Award. Recruitment, sequencing, analysis, and staffing were supported by grants from T32HG010464, UM1HG008900, U01HG0011755, R00DE026824, R01DE031261, R01HD081256, R01MH115957, R01HD105266, P50HD104224, R01HG009141, K23HD102589, R01NS035129, R01NS058721, RC2DK122397, R01EY012910, R01EY026904, and P30EY014 104, Autism Speaks, the Foundation Fighting Blindness (EGI-GE-1218-0753-UCSD), Iraty Award, Lions Foundation, Research to Prevent Blindness, Sanofi Genzyme, Ultragenyx, LGMD2I Research Fund, Samantha J. Brazzo Foundation, LGMD2D Foundation, Kurt+Peter Foundation, Muscular Dystrophy UK, Coalition to Cure Calpain 3, European Union's Horizon 2020 research and innovation programme (779257 [Solve-RD]), Murdoch Children's Research Institute, Australian NHMRC Investigator Grant (APP2007769), Harbig Foundation, the Victorian Government's Operational Infrastructure Support Program, TUBITAK (project no. 216S771), and Estonian Research Council grants PUT355, PRG471, PUTJD827, MOBTP175, and PSG774. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or other funding agencies. We acknowledge the support provided by Samantha G. Beck, Yasmine Chahine, R. Sean Hill, and Abbe Lai for ddPCR validation and variant interpretation. We thank Stojan Peric, Nagia Fahmi, John Vissing, Jordi Diaz Manera, Jaume Colomer, Anna Kostera-Pruszczyk, Tiziana Mongini, and Shahriar Nafissi. See supplemental information for additional details.

Declaration of interests

H.L.R. and M.E.T. have received support from Illumina and Microsoft to support rare disease gene discovery and diagnosis. A.O'D.-L. is on the scientific advisory board for Congenica Inc. D.G.M. is a paid advisor to GlaxoSmithKline, Insitro, Variant Bio, and Overtone Therapeutics and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Google, Merck, Microsoft, Pfizer, and Sanofi-Genzyme. C.A.W. is a paid advisor to Maze Therapeutics. M.E.T. has also received reagents and/or research support from Levo Therapeutics, Pacific Biosciences, Oxford Nanopore, and Ionis Pharmaceuticals.

Received: October 5, 2023
Accepted: March 11, 2024
Published: April 1, 2024

Web resources

ClinGen, https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=
ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
DECIPHER, https://www.deciphergenomics.org/
GATK-gCNV, https://app.terra.bio/#workspaces/help-gatk/Germline-CNVs-GATK4
gnomAD, https://gnomad.broadinstitute.org/
GREGoR Consortium, https://gregorconsortium.org/
MatchMaker Exchange, https://www.matchmakerexchange.org
OMIM, https://www.omim.org/, https://genescout.omim.org/
seqr, https://seqr.broadinstitute.org/
StrVCTVRE, https://strvctvre.berkeley.edu/

References

1. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.
2. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.
3. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.
4. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. Annu. Rev. Genom. Hum. Genet. *10*, 451–481.
5. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138.
6. Manning, M., Hudgins, L.; Professional Practice and Guidelines Committee; and Guidelines Committee (2010). Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. Genet. Med. *12*, 742–745.
7. Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. *86*, 749–764.
8. Manickam, K., McClain, M.R., Demmer, L.A., Biswas, S., Kearney, H.M., Malinowski, J., Massingham, L.J., Miller, D., Yu, T.W., Hisama, F.M.; and ACMG Board of Directors (2021). Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *23*, 2029–2037.
9. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. *91*, 597–607.
10. Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W., Hambleton, S., Burns, S.O., Thrasher, A.J., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics *28*, 2747–2754.
11. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome Sequencing Project, Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–1532.

12. Babadi, M., Fu, J.M., Lee, S.K., Smirnov, A.N., Gauthier, L.D., Walker, M., Benjamin, D.I., Zhao, X., Karczewski, K.J., Wong, I., et al. (2023). GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. Nat. Genet. *55*, 1589–1597. https://doi.org/10.1038/s41588-023-01449-0.
13. Rajagopalan, R., Murrell, J.R., Luo, M., and Conlin, L.K. (2020). A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. Genome Med. *12*, 14.
14. Pfundt, R., Del Rosario, M., Vissers, L.E.L.M., Kwint, M.P., Janssen, I.M., de Leeuw, N., Yntema, H.G., Nelen, M.R., Lugtenberg, D., Kamsteeg, E.-J., et al. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. Genet. Med. *19*, 667–675.
15. Marchuk, D.S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L.V., Brandt, A., Arreola, A., Tilley, C.R., Bizon, C., Vora, N.L., et al. (2018). Increasing the diagnostic yield of exome sequencing by copy number variant analysis. PLoS One *13*, e0209185.
16. Bergant, G., Maver, A., Lovrecic, L., Čuturilo, G., Hodzic, A., and Peterlin, B. (2018). Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. Genet. Med. *20*, 303–312.
17. Testard, Q., Vanhoye, X., Yauy, K., Naud, M.-E., Vieville, G., Rousseau, F., Dauriat, B., Marquet, V., Bourthoumieu, S., Geneviève, D., et al. (2022). Exome sequencing as a first-tier test for copy number variant detection: retrospective evaluation and prospective screening in 2418 cases. J. Med. Genet. *59*, 1234–1240.
18. Sharo, A.G., Hu, Z., Sunyaev, S.R., and Brenner, S.E. (2022). StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants. Am. J. Hum. Genet. *109*, 195–209.
19. Kleinert, P., and Kircher, M. (2022). A framework to score the effects of structural variants in health and disease. Genome Res. *32*, 766–777.
20. Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet. Med. *22*, 245–257.
21. Bamshad, M.J., Shendure, J.A., Valle, D., Hamosh, A., Lupski, J.R., Gibbs, R.A., Boerwinkle, E., Lifton, R.P., Gerstein, M., Gunel, M., et al. (2012). The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am. J. Med. Genet. *158A*, 1523–1525.
22. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet. Med. *21*, 798–812.
23. Baxter, S.M., Posey, J.E., Lake, N.J., Sobreira, N., Chong, J.X., Buyske, S., Blue, E.E., Chadwick, L.H., Coban-Akdemir, Z.H., Doheny, K.F., et al. (2022). Centers for Mendelian Genomics: A decade of facilitating gene discovery. Genet. Med. *24*, 784–797.
24. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. *83*, 610–615.
25. Pais, L.S., Snow, H., Weisburd, B., Zhang, S., Baxter, S.M., DiTroia, S., O'Heir, E., England, E., Chao, K.R., Lemire, G., et al. (2022). seqr: A web-based analysis and collaboration tool for rare disease genomics. Hum. Mutat. *43*, 698–707.
26. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
27. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458.
28. Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.-M., Ulirsch, J., et al. (2022). A cross-disorder dosage sensitivity map of the human genome. Cell *185*, 3041–3055.e25.
29. Tai, A.C., Parfenov, M., and Gorham, J.M. (2018). Droplet Digital PCR with EvaGreen Assay: Confirmational Analysis of Structural Variants. Curr. Protoc. Hum. Genet. *97*, e58.
30. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. *29*, 512–520.
31. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am. J. Hum. Genet. *100*, 895–906.
32. Abou Tayoun, A.N., Pesaran, T., DiStefano, M.T., Oza, A., Rehm, H.L., Biesecker, L.G., Harrison, S.M.; and ClinGen Sequence Variant Interpretation Working Group ClinGen SVI (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum. Mutat. *39*, 1517–1524.
33. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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. *17*, 405–424.
34. Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline and somatic genomes. Trends Genet. *28*, 43–53.
35. Collins, R.L., Brand, H., Redin, C.E., Hanscom, C., Antolik, C., Stone, M.R., Glessner, J.T., Mason, T., Pregno, G., Dorrani, N., et al. (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. Genome Biol. *18*, 36.
36. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. *36*, 915–921.
37. Tucker, B.A., Scheetz, T.E., Mullins, R.F., DeLuca, A.P., Hoffmann, J.M., Johnston, R.M., Jacobson, S.G., Sheffield, V.C., and Stone, E.M. (2011). Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene

male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa*108*, pp. E569–E576.

38. Smirnov, V.M., Nassisi, M., Solis Hernandez, C., Méjécase, C., El Shamieh, S., Condroyer, C., Antonio, A., Meunier, I., Andrieu, C., Defoort-Dhellemmes, S., et al. (2021). Retinal Phenotype of Patients With Isolated Retinal Degeneration Due to CLN3 Pathogenic Variants in a French Retinitis Pigmentosa Cohort. JAMA Ophthalmol. *139*, 278–291.
39. Zampaglione, E., Maher, M., Place, E.M., Wagner, N.E., DiTroia, S., Chao, K.R., England, E., Cmg, B., Catomeris, A., Nassiri, S., et al. (2022). The importance of automation in genetic diagnosis: Lessons from analyzing an inherited retinal degeneration cohort with the Mendelian Analysis Toolkit (MATK). Genet. Med. *24*, 332–343.
40. Venturini, G., Koskiniemi-Kuendig, H., Harper, S., Berson, E.L., and Rivolta, C. (2015). Two specific mutations are prevalent causes of recessive retinitis pigmentosa in North American patients of Jewish ancestry. Genet. Med. *17*, 285–290.
41. Pieras, J.I., Barragán, I., Borrego, S., Audo, I., González-Del Pozo, M., Bernal, S., Baiget, M., Zeitz, C., Bhattacharya, S.S., and Antiñolo, G. (2011). Copy-number variations in EYS: a significant event in the appearance of arRP. Invest. Ophthalmol. Vis. Sci. *52*, 5625–5631.
42. Bujakowska, K.M., Fernandez-Godino, R., Place, E., Consugar, M., Navarro-Gomez, D., White, J., Bedoukian, E.C., Zhu, X., Xie, H.M., Gai, X., et al. (2017). Copy-number variation is an important contributor to the genetic causality of inherited retinal degenerations. Genet. Med. *19*, 643–651.
43. Ellingford, J.M., Campbell, C., Barton, S., Bhaskar, S., Gupta, S., Taylor, R.L., Sergouniotis, P.I., Horn, B., Lamb, J.A., Michaelides, M., et al. (2017). Validation of copy number variation analysis for next-generation sequencing diagnostics. Eur. J. Hum. Genet. *25*, 719–724.
44. McGuigan, D.B., Heon, E., Cideciyan, A.V., Ratnapriya, R., Lu, M., Sumaroka, A., Roman, A.J., Batmanabane, V., Garafalo, A.V., Stone, E.M., et al. (2017). EYS Mutations Causing Autosomal Recessive Retinitis Pigmentosa: Changes of Retinal Structure and Function with Disease Progression. Genes *8*, 178. https://doi.org/10.3390/genes8070178.