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,[2,5] Eleina England,[1,2] 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,[1,2,5] Vijay S. Ganesh,[1,2,3,4,5,6] Monica H. Wojcik,[1,2,3,4] Emily Evangelista,[1,2] Hana Snow,[1,2] Ikeoluwa Osei-Owusu,[1,2,5] Jack Fu,[2,4,5] Mugdha Singh,[1,2,3,4,5] Yulia Mostovoy,[1,2,5] Steve Huang,[2] Kiran Garimella,[2] Samantha L. Kirkham,[3] Jennifer E. Neil,[3,7] Diane D. Shao,[3,4,8] Christopher A. Walsh,[2,3,4,7] Emanuela Argili,[9,10] Carolyn Le,[9,10] Elliott H. Sherr,[9,10] Joseph 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] Eric A. Pierce,[2,4,16] Sandra Donkervoort,[17] Carsten G. Bönnemann,[17] Lyndon Gallacher,[18,19] Zornitza Stark,[18,19] Tiong 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] Anne O'Donnell-Luria[1,2,3,4,31,#]

[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, Massachusetts, 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 Infirmary, 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
[18]Department of Paediatrics, University of Melbourne, Parkville, Victoria, Australia
[19]Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Parkville, Victoria, 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 Clinical Genetics, 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]University of Cape Town, Cape Town, South Africa
[27]University of Western Australia, Harry Perkins Institute of Medical Research, QEII Medical Centre, Nedlands, Australia
[28]Centre for Population Genomics, Garvan Institute, Sydney, Australia
[29]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia
[30]These authors contributed equally
[31]Senior authors
[#]Correspondance: glemiret@broadinstitute.org, odonnell@broadinstitute.org

Abstract

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 GREGoR consortium. Each family's CNV data was analyzed using the *seqr* platform and candidate CNVs classified using the 2020 ACMG/ClinGen CNV interpretation standards. We developed additional evidence criteria to address situations not covered by the current standards. The addition of CNV calling to exome analysis identified causal CNVs for 173 families (2.6%). The estimated sizes of CNVs ranged from 293 bp to 80 Mb with estimates that 44% would not have been detected by standard chromosomal microarrays. The causal CNVs consisted of 141 deletions, 15 duplications, 4 suspected complex structural variants (SVs), 3 insertions and 10 complex SVs, the latter two groups being identified by orthogonal validation methods. We interpreted 153 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. They vary in size but are defined as variants of more than 50 bp.[1,2] CNVs are significant contributors to rare genetic disease.[3,4] Chromosomal microarrays (CMA) have been the recommended first-line clinical test to investigate individuals with suspected rare genetic diseases, especially for multiple congenital anomalies and intellectual disability disorders,[5,6] though practice is moving towards exome sequencing as a first-line test.[7] Standard clinical CMAs can only detect CNVs larger than 50-100 kilobases, so this low resolution precludes most gene- and exon-level detection of CNVs. Due to technical challenges, CNVs have not traditionally been identified by standard exome sequencing which typically focuses on single nucleotides variants (SNVs) and indels.

Traditionally, exome-based CNV algorithms[8–10] 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 secondary to 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[11] 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 97% precision in detecting *de novo* CNVs captured by genome sequencing in 99 children from families with autism spectrum disorder and achieved more than 95% sensitivity for rare CNVs captured by genomes that span more than 4 exons, and more than 90% positive predictive value at all CNV sizes.[11]

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 have had prior gene panels, exome, 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 patients with suspected rare genetic diseases varies between 1 to 2%.[12–16]

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,[17] a reference dataset of structural variants (SV) from short-read genome sequencing of 10,847 individuals from the general population, helps 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 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. These standards did not intend to cover all curation scenarios and, for example, do not extend to guidance on how to score CNVs following an autosomal recessive or X-linked inheritance, 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.

METHODS

Case selection

The Broad CMG was established in 2016 as part of an initiative funded by the National Human Genome Research Institute 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 Genomics Research to Elucidate the Genetics of Rare diseases (GREGoR) consortium, the focus of which includes evaluating different approaches to improve rare disease diagnosis, such as CNV calling on exome data. Families recruited and sequenced through the Broad CMG are enrolled in research studies with local institutional review board (IRB) approval, including for sharing de-identified samples for sequencing and analysis (MassGeneralBrigham 2013P001477).

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 families underwent CNV calling on exome data through the Broad CMG (15,759 individuals). 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 through the CMG. The sequenced individuals were submitted from a large number of studies and had variable levels of pre-screening 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 20x and a mean target coverage of >80x 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 20x and a mean target coverage of >60x 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). Single nucleotide variants 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[11], 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.[11]

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 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 scores (QS) (QS>50; developer recommendations are QS>50 for duplications, >100 for deletions, and >400 for homozygous deletions, see Babadi *et al*[11] for details), and their frequency in the Broad CMG callset. 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[17] 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 a SV in gnomAD to be considered as the same allele:

- same SV type (duplication, deletion, etc)
- either has sufficient reciprocal overlap (50% reciprocal overlap for large SV >5Kb; 10% reciprocal overlap for SV <5Kb).

Genes included in a CNV were evaluated for gnomAD gene constraint scores, ClinGen dosage sensitivity scores and disease association in OMIM; exons included in an intragenic CNV were evaluated for exon expression (pext score in gnomAD[26]) and conservation. 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 0-1, 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 patient'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 versus 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.*[11] for details) if:

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

CNV validation

CNV validations were performed by the investigator that contributed the sample by a variety of methods (FISH, karyotype, CMA, MLPA, Sanger sequencing, quantitative PCR, droplet digital (dd)PCR[27], genome sequencing) across different clinical or research laboratories, while some were validated by short read or long read genome sequencing performed at the Broad Genomics Platform (Table S1). Not all CNV identified by the gCNV pipeline were validated by another method, largely when samples were from historic cohorts where there was not a path to return results and often 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[28].

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

with moderate/strong supporting evidence by the ClinGen gene-disease validity criteria.[29] 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 for cases where 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.

Each CNV was evaluated and classified by two curators (GL and KR). 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.[29] 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 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 autosomal recessive inheritance

The current ACMG/ClinGen CNV standards do not yet provide guidance on how to score CNVs in genes for conditions that follow an autosomal recessive inheritance. To classify these variants within this project, we developed an approach, advancing the current framework.

- We applied category 2E and the PVS1 LoF flowchart[30] 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.
- 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.
- 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).
- 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[31])
    a) 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 X-linked inheritance

We developed the following flowchart to score points for CNVs with X-linked inheritance (Figure 2).

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[32]. 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 validation 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.*[33] which states that inversions can be evaluated as a 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 a LoF event. We applied the LoF PVS1 criteria[30] 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 calling using the GATK-gCNV algorithm was performed on exomes from the Broad CMG cohort of 6,633 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 (as defined in Babadi *et al.*[11]) unique variants (4,387 deletions and 5,543 duplications) were identified across 15,759 individuals from these 6,633 families (Figure 3A and Figure 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, some low-quality CNV calls were prioritized and ultimately interpreted as causal. Through the entire callset analysis, we have identified a causal variant in 173 previously undiagnosed families. CNV calling on existing exome data in this cohort thus resulted in an additional solve rate of 2.6% (173/6,633). The causal CNVs consisted of 144 deletion, 15 duplication, and 14 suspected complex (multiple CNVs on a chromosome) GATK-gCNV calls, which are currently resolved as 141 deletions, 15 duplications, 3 insertions, 10 complex SVs and 4 suspected complex SVs. Of the 10 validated complex SVs, three were initially deletion or duplication calls where a complex SV was identified on validation.

These CNVs mostly involved established genes/loci, but five families that are 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[34] or the literature. The disorder followed an autosomal dominant inheritance in 93 families, an autosomal recessive inheritance in 62 families and X-linked inheritance in 18 families (Figure 3B). The CNV was confirmed *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. 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 173 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 non-uniform across different phenotypes.

The estimated sizes of causal CNVs by exome ranged from 293 bp to 80 Mb (Figure 3C). Twenty-two CNVs involved one exon and 14 CNVs involved two exons, which is below the benchmarked resolution of GATK-gCNV 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 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 (35/173 (20%)) in this cohort after looking at the copy number plot and/or validation methods.

We sought to evaluate how many of the causal CNVs could have been detected by 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.[28] Based on this, we estimate that 44% (76/173) of these CNVs are unlikely to have been detected by standard CMA.

More than half of the CNVs (105/173 (61%)) were validated 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 105 validated CNVs, 30 (29%) showed differences when comparing the initial results with the orthogonal validation results: 19 showed differences in gene/exon content and 11 showed differences in SV type. Importantly, the difference in gene or exon content identified in 19 families did not result in a change in the clinical interpretation of the CNV. Of note, only one of these 19 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 the *MAK* gene (OMIM #614181)[35] 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.

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.[32] A complex SV was suspected on the GATK-gCNV calls in 11 families (del/dup, paired dup, etc); seven of these were confirmed by genome, qPCR or CMA (Table S1) and four 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-four 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 (OMIM #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 (OMIM #606232)). The 17q12 deletion syndrome (OMIM #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, four individuals of European ancestry affected with cone rod dystrophy had a heterozygous 1-exon-deletion in *CLN3* (OMIM #204200) in trans with a pathogenic variant.[36] A founder variant in the Ashkenazi Jewish population, an Alu insertion in *MAK* (OMIM #614181)[35,37], was found in three affected individuals of this ancestry. Two individuals of different ancestries affected with retinitis pigmentosa were homozygous for the same 2-exon-deletion in *EYS* (OMIM #602772), a deletion previously reported in the literature.[36,38] Two individuals of European ancestry affected with retinitis pigmentosa had a heterozygous 4-exon-

478 deletion in *EYS* (OMIM #602772), a deletion reported in multiple affected individuals in the
479 literature [36,39–41] in trans with a pathogenic or likely pathogenic variant. Detailed information on
480 the CNV of each of these families is provided in Table S1.
481
482 The StrVCTVRE *in silico* score was evaluated across the cohort. This score was viewable on
483 each CNV within seqr during the initial analysis but was not used for filtering and not strongly
484 relied on in analysis (consistent with how other *in silico* scores are viewed in our analysis
485 pipeline). Sharo *et al.* reported that a 90% sensitivity is reached at a StrVCTVRE score of 0.37
486 (score ranges from 0-1, a score of 1 being more deleterious), which suggests that when used on
487 a collection of SVs called from a clinical cohort, this threshold may identify 90% of pathogenic
488 SVs while reducing the candidate SV list by 54%.[18] In this cohort, 161/168 unique causal CNVs
489 had a StrVCTVRE score greater than 0.37 (true positive rate of 0.96%), while this was the case
490 for 6162/10788 non-causal CNVs (false positive rate 0.57) (Table S2). The median score of the
491 161 unique causal CNVs was 0.77, and 0.42 for non-causal CNVs that had a StrVCTVRE score
492 calculated. One minor limitation of this analysis is that many large CNVs are fragmented, which
493 may result in lower StrVCTVRE scores for constituent parts than would be assigned for the
494 larger CNV event. While we manually reassembled and recalculated StrVCTVRE scores for
495 causal CNVs reported here (as it is appropriate to apply these scores to the entire CNV), non-
496 causal CNVs were not reassembled. We note that all CNVs greater than 3Mb size automatically
497 had a score of 1 demonstrating a correlation between the CNV size and the StrVCTVRE score
498 (Figure S2).
499
500 Using the 2020 ACMG/ClinGen CNV interpretation standards[20] and additional evidence criteria
501 that we developed (detailed in the Methods section), we interpreted 153 CNVs as likely
502 pathogenic/pathogenic and 20 CNVs as VUS of high interest, including the 5 in novel disease-
503 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 (148/173 (86%)) 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 versus DECIPHER resulted in a different final score for 24/148 (16%) CNVs, but this would only alter the final classification for one CNV, as points were awarded from other sections. That altered case was a 857kb *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 additional 2.6% solve rate of exome CNV calling identified in this cohort is comparable to previously reported diagnostic yield in other cohorts.[12–16] 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 within the introns. 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.

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.[11] A small number of CNVs were nonetheless confirmed by genome sequencing by the Broad CMG as part of initial efforts to validate gCNV performance. Of the 23 deletions and two duplications identified by GATK-gCNV and validated by Broad CMG genomes, these were resolved as 22 deletions, one duplication and two complex SVs. We recommended that any candidate CNV variants be confirmed with an orthogonal method and generally validations are performed by the collaborating researcher who recruited the individual for sequencing. 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[11]; interestingly, 36 CNVs (36/173 (21%)) in this cohort involved fewer than 3 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 validated by various methods and validation 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 validated 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 on exome data, so seven suspected complex SVs and three initially unsuspected complex SVs in this cohort were identified by orthogonal validation methods. We likely underdetected complex SVs as 39% of the CNVs in this cohort were not validated and some validation methods would miss a more complex event, such as 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 used StrVCTVRE scores and we observed 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 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 database stands as a pivotal resource in human genetics but is currently 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 are 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 planned addition of CNV calling on gnomAD exomes 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[30] (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. 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

606 flowchart[30] for intragenic CNVs or CNVs including at least one gene that had an established
607 gene-disease relationship following an X-linked inheritance. Finally, to evaluate SVs other than
608 deletion and duplication, we took guidance from Collins *et al.*[33], which states that LoF can be
609 expected if there is an insertion within an exon, if an inversion breakpoint falls within a gene, or
610 if both inversion breakpoints fall within the same gene and span at least one exon. We thus
611 applied the PVS1 LoF flowchart here. Our approach refined multiple aspects of CNV
612 classification and advanced the systematic framework to assess the pathogenicity of CNVs.

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

627 CONCLUSION

628 CNV calling and analysis from existing exome data increases the solve rate by 2.6% in this
629 diverse and presumed monogenic cohort. This is a higher resolution alternative to arrays at a
630 fraction of the cost of genome sequencing and can be applied retrospectively to existing exome

datasets. We estimate that 44% of the 173 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.

Acknowledgements

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. V.S.G. was supported by NIH NHGRI T32 (T32HG010464) and M.H.W. by NIH/NICHD K23HD102589. Sequencing and analysis were provided by the Broad CMG, funded by the National Human Genome Research Institute grants UM1HG008900, U01HG0011755 and R01HG009141. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional funding came from NIH/NINDS grants R01NS032457, R01NS058721, NIH/NIDDK grant RC2DK122397, 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 (grant agreement No. 779257 (Solve-RD)), the Murdoch Children's Research Institute, the Harbig Foundation, the Victorian Government's Operational Infrastructure Support Program, TUBITAK ((The Scientific and Technological

Research Council of Turkey) Project No. 216S771), and Estonian Research Council grants PUT355, PRG471, PUTJD827, MOBTP175 and PSG774. We acknowledge the support provided by Samantha G. Beck, Yasmine Chahine, R. Sean Hill and Abbe Lai for ddPCR validation and variant interpretation. See supplemental for additional details.

Declaration of interests

H.L.R. has received support from Illumina and Microsoft to support rare disease gene discovery and diagnosis. A.O-D.L. has consulted for Tome Biosciences and Ono Pharma USA 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. receives research funding from Microsoft Inc, Illumina Inc and Levo Therapeutics. The remaining authors declare no competing interests.

Web resources

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

690 REFERENCES

691 1. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and
692 genotyping. Nat. Rev. Genet. *12*, 363–376.

693 2. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number
694 variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

695 3. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human
696 health, disease, and evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.

697 4. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of
698 genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*,
699 125–138.

700 5. Manning, M., Hudgins, L., and Professional Practice and Guidelines Committee (2010).
701 Array-based technology and recommendations for utilization in medical genetics practice
702 for detection of chromosomal abnormalities. Genet. Med. *12*, 742–745.

703 6. Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P.,
704 Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus statement:
705 chromosomal microarray is a first-tier clinical diagnostic test for individuals with
706 developmental disabilities or congenital anomalies. Am. J. Hum. Genet. *86*, 749–764.

707 7. Manickam, K., McClain, M.R., Demmer, L.A., Biswas, S., Kearney, H.M., Malinowski, J.,
708 Massingham, L.J., Miller, D., Yu, T.W., Hisama, F.M., et al. (2021). Exome and genome
709 sequencing for pediatric patients with congenital anomalies or intellectual disability: an
710 evidence-based clinical guideline of the American College of Medical Genetics and
711 Genomics (ACMG). Genet. Med. *23*, 2029–2037.

712 8. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M.,
713 Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery
714 and statistical genotyping of copy-number variation from whole-exome sequencing depth.
715 Am. J. Hum. Genet. *91*, 597–607.

716 9. Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W.,
717 Hambleton, S., Burns, S.O., Thrasher, A.J., et al. (2012). A robust model for read count
718 data in exome sequencing experiments and implications for copy number variant calling.
719 Bioinformatics *28*, 2747–2754.

720 10. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome
721 Sequencing Project, Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number
722 variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–
723 1532.

724 11. Babadi, M., Fu, J.M., Lee, S.K., Smirnov, A.N., Gauthier, L.D., Walker, M., Benjamin, D.I.,
725 Zhao, X., Karczewski, K.J., Wong, I., et al. (2023). GATK-gCNV enables the discovery of
726 rare copy number variants from exome sequencing data. Nat. Genet. 10.1038/s41588-023-
727 01449-0.

728 12. Rajagopalan, R., Murrell, J.R., Luo, M., and Conlin, L.K. (2020). A highly sensitive and
729 specific workflow for detecting rare copy-number variants from exome sequencing data.

Genome Med. *12*, 14.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

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. A *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. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline and somatic genomes. Trends Genet. *28*, 43–53.

33. 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.

34. 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.

35. 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. Proc. Natl. Acad. Sci. U. S. A. *108*, E569–

E576.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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*. 10.3390/genes8070178.

FIGURE LEGENDS

Figure 1. Exome copy number plot and reads visualization for examples of causal CNVs in the Broad CMG cohort. (A) Individual affected with retinitis pigmentosa with a homozygous single exon deletion in *CRB1* (chr1:197438450-197439442x0, 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 fell 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* (chr1:197435257-197441674x0 (NM_201253.3)). Cov= coverage. (C) Individual with multiple congenital anomalies and a heterozygous deletion of 4 exons in *RAB3GAP1* (Warburg micro syndrome) (red, chr2:135162318-135164794x1, QS =92) in trans with a frameshift variant in *RAB3GAP1* (not shown, NM_012233.3: c.2393_2394del, 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, chr1:242523991-245156781x1, 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* 2.1Mb 22q11.2 duplication (red, chr22:18985739-21081116x3, QS =3077) identified on exome. The presence of this duplication was validated by chromosomal microarray. Y axis: CN. All coordinates on GRCh38.

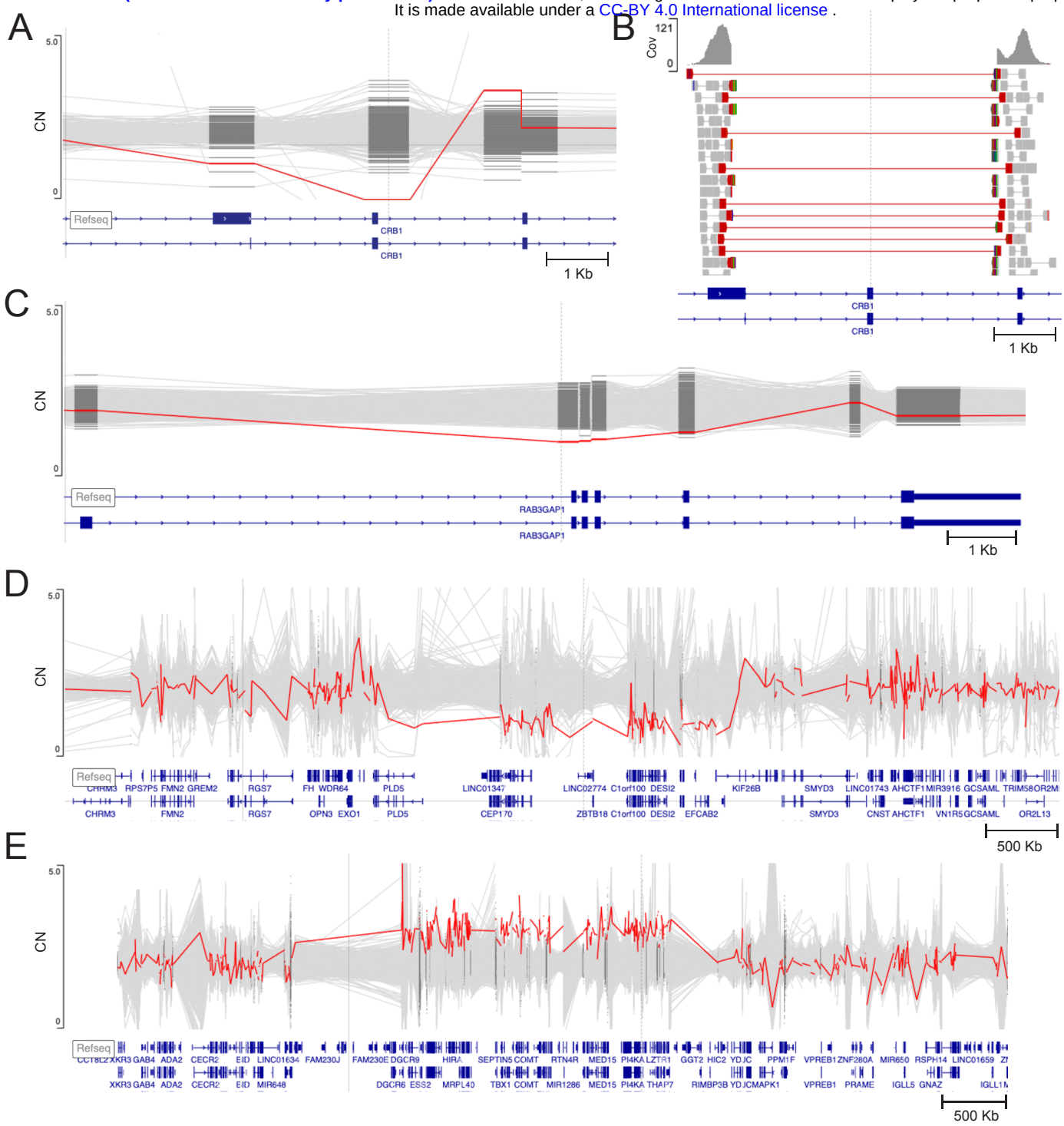
Figure 2. Flowchart illustrating how points were scored for CNVs that followed a X-linked inheritance. We incorporated sex of proband, parental genotype and parental affected status to

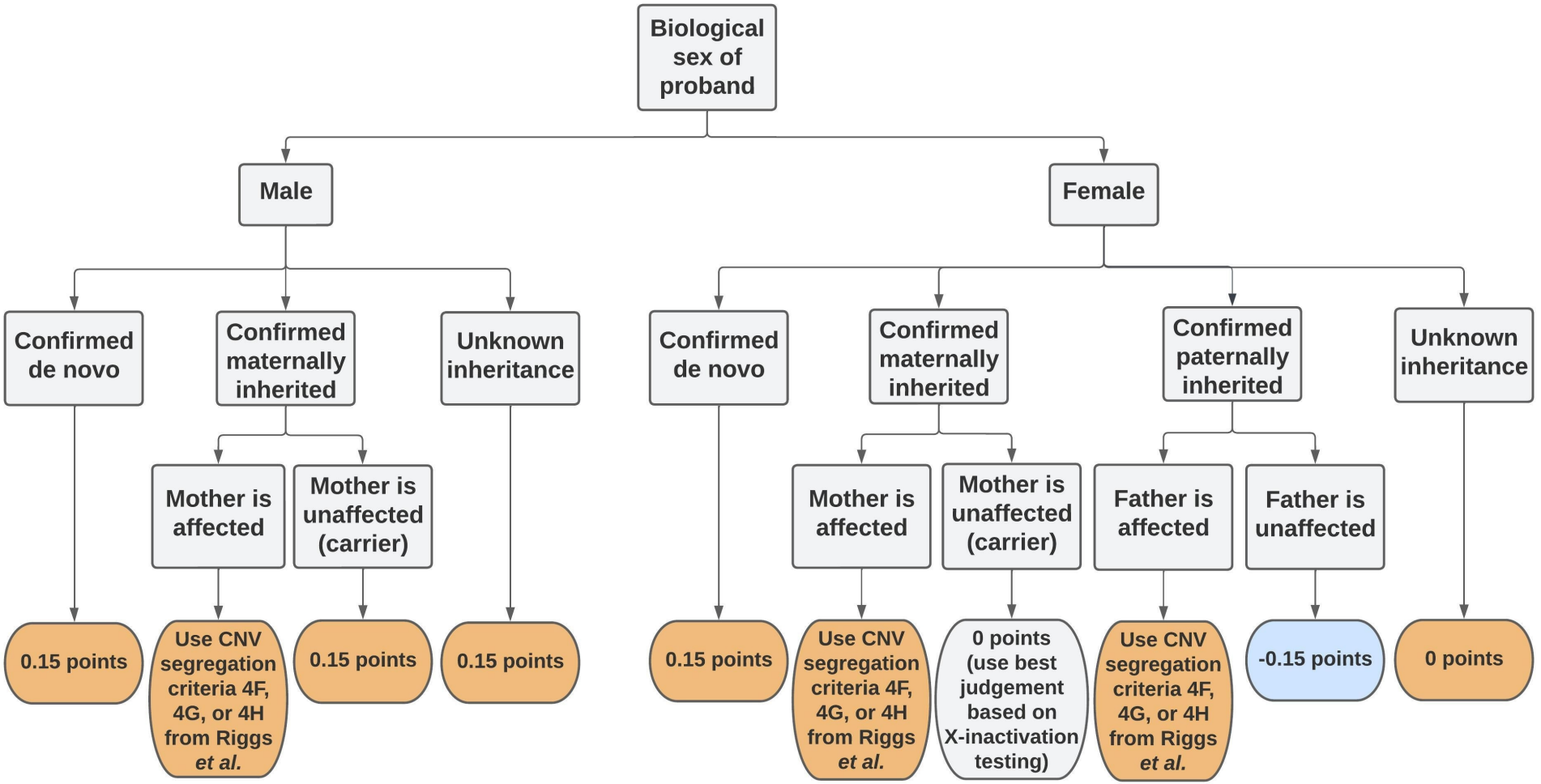
score both the proband in which the X-linked variant was identified and, if applicable, any individual in the published literature or public databases that had variants of similar genomic content to the variant of interest. The points for each case could be increased or decreased based on phenotype specificity,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

Figure 3. Characteristics of 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 Figure 3B and 3C. DEL: deletion; DUP: duplication. (B) Mode of inheritance and number of genes involved in each CNV in 173 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] (C) CNV classification by estimated size in 173 families in which the CNV was interpreted as causal by the multidisciplinary team. The causal CNVs consisted of 141 deletions, 15 duplications, 3 insertions (miscalled as deletion by GATK-gCNV), and 14 complex structural variants (SV). We interpreted 153 CNVs as likely pathogenic/pathogenic and 20 CNVs as VUS.

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

Variant classification/zygosity	Points per Proband	
	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





Highly specific

Pathogenic case evidence

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.

Nonspecific

Nonspecific

Benign case evidence

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.

Highly specific