

1 Exome copy number variant detection, analysis and classification in a large cohort of 2 families with undiagnosed rare genetic disease

3
4 Gabrielle Lemire,^{1,2,3,4,5,30,#} Alba Sanchis-Juan,^{1,2,4,5,30} Kathryn Russell,^{1,2} Samantha Baxter,^{1,2}
5 Katherine R. Chao,^{1,2,5} Moriel Singer-Berk,^{1,2,5} Emily Gropman,^{1,2,3} Isaac Wong,^{2,5} Eleina
6 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}
7 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
8 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
9 Huang,² Kiran Garimella,² Samantha L. Kirkham,³ Jennifer E. Neil,^{3,7} Diane D. Shao,^{3,4,8}
10 Christopher A. Walsh,^{2,3,4,7} Emanuela Argili,^{9,10} Carolyn Le,^{9,10} Elliott H. Sherr,^{9,10} Joseph
11 Gleeson,^{11,12} Shirlee Shril,^{4,13} Ronen Schneider,^{4,13} Friedhelm Hildebrandt,^{4,13} Vijay G.
12 Sankaran,^{2,4,14} Jill A. Madden,^{3,15} Casie A. Genetti,^{3,15} Alan H. Beggs,^{2,3,4,15} Pankaj B.
13 Agrawal,^{2,3,4,15} Kinga M. Bujakowska,^{2,4,16} Emily Place,^{2,4,16} Eric A. Pierce,^{2,4,16} Sandra
14 Donkervoort,¹⁷ Carsten G. Bönnemann,¹⁷ Lyndon Gallacher,^{18,19} Zornitza Stark,^{18,19} Tiong
15 Tan,^{18,19} Susan M. White,^{18,19} Ana Töpf,²⁰ Volker Straub,²⁰ Mark D. Fleming,^{4,21} Martin R.
16 Pollak,^{4,22} Katrin Öunap,^{23,24} Sander Pajusalu,^{23,24} Kirsten A. Donald,^{25,26} Zandre Bruwer,^{25,26}
17 Gianina Ravenscroft,²⁷ Nigel G. Laing,²⁷ Daniel G. MacArthur,^{1,2,28,29} Heidi L. Rehm,^{1,2,4,5} Michael
18 E. Talkowski,^{1,2,4,5} Harrison Brand,^{1,2,4,5,31} Anne O’Donnell-Luria^{1,2,3,4,31,#}

19
20 ¹Broad Institute Center for Mendelian Genomics, Broad Institute of MIT and Harvard,
21 Cambridge, MA, USA

22 ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge,
23 MA, USA

24 ³Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA, USA

25 ⁴Harvard Medical School, Boston, MA, USA

26 ⁵Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

27 ⁶Department of Neurology, Brigham and Women’s Hospital, Boston, MA, USA

28 ⁷Howard Hughes Medical Institute, Boston Children’s Hospital, Boston, MA, USA

29 ⁸Department of Neurology, Boston Children’s Hospital, Boston, MA, USA

30 ⁹Department of Neurology, University of California, San Francisco, San Francisco, CA, USA

31 ¹⁰Institute of Human Genetics and Weill Institute for Neurosciences, University of California, San
32 Francisco, San Francisco, CA, USA

33 ¹¹Department of Neurosciences, University of California San Diego, La Jolla, CA, USA

34 ¹²Rady Children’s Institute for Genomic Medicine, San Diego, CA, USA

35 ¹³Department of Pediatrics, Boston Children’s Hospital, Boston, MA, USA

36 ¹⁴Division of Hematology/Oncology, Boston Children’s Hospital and Department of Pediatric
37 Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

38 ¹⁵The Manton Center for Orphan Disease Research, Boston Children’s Hospital, Boston, MA,
39 USA

40 ¹⁶Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear
41 Infirmary, Boston, MA, USA

42 ¹⁷Neuromuscular and Neurogenetic Disorders of Childhood Section, Neurogenetics Branch,
43 National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda,
44 MD, USA

45 ¹⁸Department of Paediatrics, University of Melbourne, Parkville, Victoria, Australia

46 ¹⁹Victorian Clinical Genetics Services, Murdoch Children’s Research Institute, Parkville, Victoria,
47 Australia

48 ²⁰John Walton Muscular Dystrophy Research Centre, Newcastle University and Newcastle
49 Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

50 ²¹Department of Pathology, Boston Children’s Hospital, Boston, MA, USA

51 ²²Division of Nephrology, Beth Israel Deaconess Medical Center, Boston, MA, USA
52 ²³Department of Clinical Genetics, Genetics and Personalized Medicine Clinic, Tartu University
53 Hospital, Tartu, Estonia
54 ²⁴Department of Clinical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University
55 of Tartu, Tartu, Estonia
56 ²⁵Department of Paediatrics and Child Health, Red Cross War Memorial Children's Hospital,
57 Cape Town, South Africa
58 ²⁶University of Cape Town, Cape Town, South Africa
59 ²⁷University of Western Australia, Harry Perkins Institute of Medical Research, QEII Medical
60 Centre, Nedlands, Australia
61 ²⁸Centre for Population Genomics, Garvan Institute, Sydney, Australia
62 ²⁹Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia
63 ³⁰These authors contributed equally
64 ³¹Senior authors
65 #Correspondance: glemiret@broadinstitute.org, odonnell@broadinstitute.org

66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96

97 Abstract

98

99 Copy number variants (CNVs) are significant contributors to the pathogenicity of rare genetic
100 diseases and with new innovative methods can now reliably be identified from exome
101 sequencing. Challenges still remain in accurate classification of CNV pathogenicity. CNV calling
102 using GATK-gCNV was performed on exomes from a cohort of 6,633 families (15,759
103 individuals) with heterogeneous phenotypes and variable prior genetic testing collected at the
104 Broad Institute Center for Mendelian Genomics of the GREGoR consortium. Each family's CNV
105 data was analyzed using the *seqr* platform and candidate CNVs classified using the 2020
106 ACMG/ClinGen CNV interpretation standards. We developed additional evidence criteria to
107 address situations not covered by the current standards. The addition of CNV calling to exome
108 analysis identified causal CNVs for 173 families (2.6%). The estimated sizes of CNVs ranged
109 from 293 bp to 80 Mb with estimates that 44% would not have been detected by standard
110 chromosomal microarrays. The causal CNVs consisted of 141 deletions, 15 duplications, 4
111 suspected complex structural variants (SVs), 3 insertions and 10 complex SVs, the latter two
112 groups being identified by orthogonal validation methods. We interpreted 153 CNVs as likely
113 pathogenic/pathogenic and 20 CNVs as high interest variants of uncertain significance. Calling
114 CNVs from existing exome data increases the diagnostic yield for individuals undiagnosed after
115 standard testing approaches, providing a higher resolution alternative to arrays at a fraction of
116 the cost of genome sequencing. Our improvements to the classification approach advances the
117 systematic framework to assess the pathogenicity of CNVs.

118

119

120

121

122

123

124

125

126 INTRODUCTION

127 Copy number variants (CNVs) are imbalances of genomic material compared with the reference
128 genome resulting in the addition (duplications and insertions) or removal (deletions) of genomic
129 segments. They vary in size but are defined as variants of more than 50 bp.^{1,2} CNVs are
130 significant contributors to rare genetic disease.^{3,4} Chromosomal microarrays (CMA) have been
131 the recommended first-line clinical test to investigate individuals with suspected rare genetic
132 diseases, especially for multiple congenital anomalies and intellectual disability disorders,^{5,6}
133 though practice is moving towards exome sequencing as a first-line test.⁷ Standard clinical
134 CMAs can only detect CNVs larger than 50-100 kilobases, so this low resolution precludes most
135 gene- and exon-level detection of CNVs. Due to technical challenges, CNVs have not
136 traditionally been identified by standard exome sequencing which typically focuses on single
137 nucleotides variants (SNVs) and indels.

138 Traditionally, exome-based CNV algorithms⁸⁻¹⁰ have relied on exome read depth to inform of
139 the underlying copy number at a given locus. However, many factors influence exome read
140 depth so detecting CNVs from exome data is difficult due to the non-uniform distribution of
141 captured reads secondary to biases introduced by PCR amplification, exome capture, and
142 mapping. These factors make it challenging to differentiate between a technical artifact and a
143 bona fide CNV. The GATK-gCNV tool¹¹ uses a probabilistic framework to infer rare CNVs from
144 read depth data in the presence of these systematic biases. The performance of GATK-gCNV
145 has been benchmarked with genome sequencing; it achieved 97% precision in detecting *de*
146 *novo* CNVs captured by genome sequencing in 99 children from families with autism spectrum
147 disorder and achieved more than 95% sensitivity for rare CNVs captured by genomes that span
148 more than 4 exons, and more than 90% positive predictive value at all CNV sizes.¹¹

149 We used the GATK-gCNV algorithm to call CNVs across the Broad Institute Center for
150 Mendelian Genomics (Broad CMG) exome cohort, a research center within the Genomics

151 Research to Elucidate the Genetics of Rare Diseases (GREGoR) consortium. The Broad CMG
152 has performed exome sequencing on more than 6,000 families with a suspected genetic
153 disease since 2016, representing a large cohort of individuals with heterogeneous phenotypes
154 including neurodevelopmental disorders, neuromuscular diseases, retinal disorders, blood
155 disorders, kidney diseases, multiple malformations syndromes, and other conditions. Most
156 individuals have had prior gene panels, exome, and/or clinical CMA but the level of prior genetic
157 testing is variable. Several molecular diagnostic laboratories and many research groups have
158 incorporated CNV calling in their exome analysis, particularly in recent years. The reported
159 additional diagnostic yield of CNV calling on exome data, most commonly used as a second-line
160 test after CMA, on various cohorts of patients with suspected rare genetic diseases varies
161 between 1 to 2%.¹²⁻¹⁶

162 The widespread implementation of CMA and exome/genome sequencing is expanding the types
163 and numbers of CNVs identified in both clinical and research settings, and it can be challenging
164 to determine the impact of these CNVs on human health. Several resources have been or are
165 being developed to address this challenge. For instance, high quality reference population data
166 such as gnomAD SV,¹⁷ a reference dataset of structural variants (SV) from short-read genome
167 sequencing of 10,847 individuals from the general population, helps determine the frequency of
168 a CNV in the population. Also, *in silico* prediction tools for CNVs are available including some
169 that have been developed with the goal of helping to distinguish deleterious CNVs from non-
170 deleterious CNVs. For example, the StrVCTVRE score is a predictive tool that incorporates
171 gene importance, conservation, coding sequence, and exon structure of the disrupted region
172 and can evaluate CNVs overlapping coding sequences.¹⁸ CADD-SV, another example, is a tool
173 developed using machine-learning random forest models to differentiate deleterious from
174 neutral SVs.¹⁹

175 Importantly, accurate classification of CNV pathogenicity requires a consistent and transparent
176 approach to be used across the human genetics field. Riggs *et al.* developed the American
177 College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource
178 (ClinGen) consensus standards to guide in the evaluation of germline CNVs and encourage
179 consistency in CNV interpretation across laboratories, technologies and specialties.²⁰ They
180 proposed a quantitative evidence-based evaluation framework to classify copy number loss and
181 copy number gain that follow an autosomal dominant inheritance. These standards did not
182 intend to cover all curation scenarios and, for example, do not extend to guidance on how to
183 score CNVs following an autosomal recessive or X-linked inheritance, CNVs with available
184 functional evidence, or SVs beyond deletions and duplications. Here, we developed and applied
185 additional evidence criteria to address these limitations and assess the pathogenicity of all
186 CNVs that were thought to be causal in the Broad CMG exome cohort.

187

188 METHODS

189 **Case selection**

190 The Broad CMG was established in 2016 as part of an initiative funded by the National Human
191 Genome Research Institute of the National Institutes of Health, with the goal of discovering the
192 variants and genes underlying Mendelian disease to increase diagnosis rates for individuals
193 with a suspected genetic condition.²¹⁻²³ The Broad CMG is now part of the NHGRI Genomics
194 Research to Elucidate the Genetics of Rare diseases (GREGoR) consortium, the focus of which
195 includes evaluating different approaches to improve rare disease diagnosis, such as CNV
196 calling on exome data. Families recruited and sequenced through the Broad CMG are enrolled
197 in research studies with local institutional review board (IRB) approval, including for sharing de-
198 identified samples for sequencing and analysis (MassGeneralBrigham 2013P001477).

199 Phenotypic information for the affected individuals in each family was provided using HPO
200 terms.²⁴

201
202 From February 2016 to May 2021 (5 years, 3 months), 6,633 families underwent CNV calling on
203 exome data through the Broad CMG (15,759 individuals). This cohort had heterogeneous
204 phenotypes including neurodevelopmental, neuromuscular, multiple congenital anomalies,
205 hematological, ocular or renal disorders. Most were enrolled due to an unrevealing prior genetic
206 diagnostic evaluation as many had a CMA, gene panel sequencing for known causes of
207 disease, or clinical exome prior to research exome through the CMG. The sequenced
208 individuals were submitted from a large number of studies and had variable levels of pre-
209 screening prior to enrollment (and this information was not systematically collected).

210 **Exome sequencing**

211 Exome sequencing was performed by the Genomics Platform at the Broad Institute of MIT and
212 Harvard. Libraries from DNA samples (>250 ng of DNA, at >2 ng/ul) were created with an
213 Illumina Nextera exome capture (37 Mb target) and sequenced (150 bp paired reads) to cover
214 >80% of targets at 20x and a mean target coverage of >80x from February 2016 through
215 January 2019 and then using a Twist exome capture (38 Mb target) and sequenced (150 bp
216 paired reads) to cover > 80% of targets at 20x and a mean target coverage of >60x thereafter.
217 Sample identity quality assurance checks were performed on each sample. The exome data
218 was de-multiplexed and each sample's sequence data were aggregated into a single Picard
219 CRAM file. The BWA aligner was used for mapping reads to the human genome build 38
220 (GRCh38). Single nucleotide variants and insertions/deletions (indels) were jointly called across
221 all samples using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 3.5.
222 Default filters were applied to SNV and indel calls using the GATK Variant Quality Score
223 Recalibration (VQSR) approach. Annotation was performed using Variant Effect Predictor

224 (VEP), during upload of the callset to *seqr*²⁵ for collaborative analysis between the Broad CMG
225 team and collaborating investigators.

226

227 **CNV detection on exome data**

228 CNVs were detected from exome sequencing following GATK-gCNV best practices¹¹, as
229 follows: read coverage was first calculated for each exome using GATK CollectReadCounts.
230 After coverage collection, all samples were subdivided into batches of a median of 410 samples
231 (range:160-625) for gCNV model training and execution; these batches were determined based
232 on a principal components analysis (PCA) of sequencing read counts. After batching, one gCNV
233 model was trained per batch using GATK GermlineCNVCaller on a subset of training samples,
234 and the trained model was then applied to call CNVs for each sample per batch. Finally, all raw
235 CNVs were aggregated across all batches and post-processed using quality- and frequency-
236 based filtering to produce the final CNV callset. Methods are further described in Babadi et al.¹¹

237

238 **CNV analysis**

239 Each family's CNV data was manually analyzed in coordination with the SNV/indel data by
240 members of the Broad CMG analysis team using our in-house developed analysis platform,
241 *seqr*, an open-source, web-based tool for family-based monogenic disease analysis that
242 enables variant filtration, annotation and prioritization in addition to data sharing of candidate
243 disease genes (with variants and HPO terms) through the Matchmaker Exchange.²⁵ CNVs were
244 filtered based on their mode of inheritance, gCNV quality scores (QS) (QS>50; developer
245 recommendations are QS>50 for duplications, >100 for deletions, and >400 for homozygous
246 deletions, see Babadi *et al*¹¹ for details), and their frequency in the Broad CMG callset. For
247 autosomal dominant conditions, we filtered for CNVs with an allele frequency of <0.1% in the
248 Broad callset, and used <1% for autosomal recessive conditions. When analyzing each family,
249 factors used to help prioritize if a CNV was of clinical significance for a given individual included

250 the CNV size, its structural consequences (predicted loss-of-function (LoF) variant, copy gain),
251 its segregation pattern within the affected family, its frequency in the gnomAD-SV¹⁷ reference
252 population database, the number and characteristics of genes involved in the CNV, and *in silico*
253 prediction of pathogenicity tools. Of note, the following criteria needed to be met for a SV in
254 gnomAD to be considered as the same allele:

- 255 - same SV type (duplication, deletion, etc)
- 256 - either has sufficient reciprocal overlap (50% reciprocal overlap for large SV
257 >5Kb; 10% reciprocal overlap for SV <5Kb).

258 Genes included in a CNV were evaluated for gnomAD gene constraint scores, ClinGen dosage
259 sensitivity scores and disease association in OMIM; exons included in an intragenic CNV were
260 evaluated for exon expression (pext score in gnomAD²⁶) and conservation. If no promising
261 variants were found using our initial searches, we removed the QS filter to include low-quality
262 variants. We reviewed the StrVCTVRE score¹⁸ of candidate CNVs but did not use it to filter data
263 or rule out variants. The score ranges from 0-1, a score of 1 being more deleterious. In line with
264 the developer suggestions, CNVs with a score >0.37 were considered as having a higher
265 likelihood of being deleterious. To evaluate the quality of a given CNV, the patient's copy
266 number level was compared to any additional sequenced family members as well as a cluster of
267 other samples with similar read depth that act as controls. The copy number plot of each
268 compelling candidate was assessed to confirm an increase or decrease (corresponding to either
269 a gain or a loss) between the proband and the background cluster, and a difference in the
270 proband's copy number within versus outside the reported coordinates of the CNV (Figure 1).
271 We also visually inspected the read data of candidate CNVs using the Integrated Genomics
272 Viewer (IGV) to evaluate for sequencing artifacts (Figure 1).

273

274 A CNV is defined as high-confidence by GATK-gCNV (see Babadi *et al.*¹¹ for details) if:

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

281

282 **CNV validation**

283 CNV validations were performed by the investigator that contributed the sample by a variety of
284 methods (FISH, karyotype, CMA, MLPA, Sanger sequencing, quantitative PCR, droplet digital
285 (dd)PCR²⁷, genome sequencing) across different clinical or research laboratories, while some
286 were validated by short read or long read genome sequencing performed at the Broad
287 Genomics Platform (Table S1). Not all CNV identified by the gCNV pipeline were validated by
288 another method, largely when samples were from historic cohorts where there was not a path to
289 return results and often insufficient remaining DNA.

290

291 **Evaluation of CMA coverage for each causal CNV**

292 To evaluate how many causal CNVs could have been detected by a standard clinical CMA,
293 CNV detection sensitivity by CMA was assessed by evaluating the number of probes from the
294 Agilent GenetiSure Cyto CGH+SNP arrays (downloaded from <https://genome.ucsc.edu/> on May
295 23, 2023) included within the genomic coordinates of a given CNV. A minimum number of five
296 probes was required to consider that the CNV would confidently be called by CMA²⁸.

297

298 **Assessment of the pathogenicity of CNVs**

299 We considered a case solved if the CNV was classified as pathogenic or likely pathogenic and
300 conclusively explained the phenotype or if a variant was found involving a novel disease gene

301 with moderate/strong supporting evidence by the ClinGen gene-disease validity criteria.²⁹

302 Supporting genetic and/or experimental evidence were required to consider a CNV in a novel
303 gene as the diagnosis in a given family, most often by additional families identified through
304 Matchmaker Exchange. We also considered a case solved for cases where the analysis team
305 and referring provider, when relevant, considered the variant causative, even if a CNV was
306 technically a variant of uncertain significance (VUS) by ACMG/ClinGen CNV criteria.

307 Each CNV was evaluated and classified by two curators (GL and KR). In order to systematically
308 assess the pathogenicity of the SVs in this study, the ACMG/ClinGen standards for
309 interpretation and reporting of constitutional copy-number variants were applied.²⁰ For candidate
310 novel disease genes, the interpretation of gene-disease relationship was guided by the ClinGen
311 framework.²⁹ We developed an approach, including new curation criteria, to optimally capture
312 evidence for pathogenicity for the range of variants discussed in this article.

313

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

315 In order to score points from section 3 from the Riggs standards (“evaluation of gene number”),
316 we used OMIM gene number count (<https://genescout.omim.org/>), and have compared it to the
317 gene number count provided by the DECIPHER browser (<https://www.deciphergenomics.org/>)
318 and the ClinGen browser ([https://search.clinicalgenome.org/kb/gene-](https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=)
319 [dosage?page=1&size=25&search=](https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=)).

320

321 **Variants following autosomal recessive inheritance**

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

- 325 • We applied category 2E and the PVS1 LoF flowchart³⁰ for any intragenic CNV, or if a
326 CNV had a complete or partial overlap with a gene with an established gene-disease
327 relationship that follows an autosomal recessive inheritance.
- 328 • When the candidate CNV involved a gene with no established gene-disease
329 relationship, we did not score points from category 2, but rather used category 4 to build
330 up evidence for an established gene-disease relationship by finding additional cases
331 with overlapping variants from the literature.
- 332 • Points were awarded to the Broad CMG cases and published cases from the literature
333 using a similar system to that which is used when curating SNVs (the PM3 criteria)
334 [ClinGen Sequence Variant Interpretation Recommendation for in trans Criterion (PM3) -
335 Version 1.0 Working Group Page: <https://clinicalgenome.org/working-groups/sequence-variant-interpretation/>, Approved: May 2, 2019]. The point-based system suggested in
336 the PM3 criteria was translated into points of similar strength level in the Riggs
337 quantitative framework²⁰ (Table 1).
- 338 • We added 0.15 points when at least one individual with a unique phenotype (phenotype
339 is highly specific to disease, low genetic heterogeneity) has been reported by our study
340 or in the literature (equivalent of PP4 criteria in Richards et al³¹)
- 341 a) In some cases, we awarded 0.30 points when evidence was particularly strong.
342 This only applied for genetic diseases with a specific, unique phenotype, high
343 clinical sensitivity testing (e.g. biochemical assays, enzyme deficiency assays,
344 functional cytogenetic tests (e.g. chromosomal breakage study)), and consistent
345 family history. These additional points were only used one time per variant.
346

347

348 **Variants following X-linked inheritance**

349 We developed the following flowchart to score points for CNVs with X-linked inheritance (Figure
350 2).

351

352 **Complex SVs**

353 We defined a complex SV as a complex rearrangement typically composed of three or more
354 breakpoint junctions that cannot be characterized as a single canonical SV type³². Some
355 complex SVs were suspected on exome CNV analysis and/or identified after further validation.
356 As suggested by Riggs *et al.*²⁰, when classifying complex rearrangements (for example a paired
357 duplication inversion), we evaluated each CNV separately. The overall classification for the
358 event was defaulted to the most deleterious classification (for example, if the deletion portion
359 were classified as “pathogenic” and the duplication portion was classified as “uncertain
360 significance,” the entire SV was classified as “pathogenic”).

361

362 **Inversions and Insertions**

363 For variants initially called as deletion or duplication by GATK-gCNV in this cohort, some were
364 identified as including inversions or insertions by validation methods. The Riggs *et al.*²⁰
365 standards do not provide guidance on how to score inversions or insertions. Therefore, we took
366 guidance from Collins *et al.*³³ which states that inversions can be evaluated as a LoF event if
367 exactly one breakpoint falls within a gene, or both breakpoints fall within the same gene and
368 span at least one exon. Collins *et al.* also recommend evaluating a large insertion within an
369 exon as a LoF event. We applied the LoF PVS1 criteria³⁰ as appropriate for such cases.

370

371 **Variants with available functional evidence**

372 We added an additional 0.15 points for any variant with at least supporting functional evidence
373 of pathogenicity, either from the investigation of our cases or from the literature. Examples
374 included: expression assays (Western blot for protein expression, PCR for RNA expression),

375 RNA sequencing, cellular assays (impaired localization and/or function) or protein interaction
376 studies. If the evidence was stronger, the points were upgraded to “moderate” (0.30 points) or
377 strong (0.45 points). For example, RNA sequencing results showing a clear and significantly
378 decreased expression of a gene or an animal model with the exact variant recapitulating the
379 disease phenotype was given 0.45 point (strong evidence).

380

381 **RESULTS**

382 CNV calling using the GATK-gCNV algorithm was performed on exomes from the Broad CMG
383 cohort of 6,633 families with heterogeneous rare disease phenotypes and variable prior genetic
384 testing that typically included a gene panel, exome, and/or CMA. A total number of 9,930 high-
385 confidence (as defined in Babadi *et al.*¹¹) unique variants (4,387 deletions and 5,543
386 duplications) were identified across 15,759 individuals from these 6,633 families (Figure 3A and
387 Figure S1), 10,472 of the 15,759 samples had at least one rare (<1% frequency in the Broad
388 data callset) high-confidence CNV, and the median number identified was two (sd+-1.55) per
389 individual (Figure S1). The entire CNV callset for these individuals, with a total of 2,131,645
390 copy number calls (292,833 unique variants), was loaded into the seqr platform for analysis.
391 Many of these low-quality calls were likely artifacts, but by incorporating phenotype and allelic
392 variation (SNVs, indels, CNVs) in the analysis of each family, some low-quality CNV calls were
393 prioritized and ultimately interpreted as causal. Through the entire callset analysis, we have
394 identified a causal variant in 173 previously undiagnosed families. CNV calling on existing
395 exome data in this cohort thus resulted in an additional solve rate of 2.6% (173/6,633). The
396 causal CNVs consisted of 144 deletion, 15 duplication, and 14 suspected complex (multiple
397 CNVs on a chromosome) GATK-gCNV calls, which are currently resolved as 141 deletions, 15
398 duplications, 3 insertions, 10 complex SVs and 4 suspected complex SVs. Of the 10 validated
399 complex SVs, three were initially deletion or duplication calls where a complex SV was identified
400 on validation.

401

402 These CNVs mostly involved established genes/loci, but five families that are considered solved

403 had a CNV involving a novel disease gene candidate. Supporting genetic and/or experimental

404 evidence was required to consider a CNV in a novel gene as the explanation for a given family,

405 most often by additional families identified through Matchmaker Exchange³⁴ or the literature.

406 The disorder followed an autosomal dominant inheritance in 93 families, an autosomal

407 recessive inheritance in 62 families and X-linked inheritance in 18 families (Figure 3B). The

408 CNV was confirmed *de novo* in 70/93 (75%) of the families with an autosomal dominant

409 disorder, inherited from a parent in 3/93 families (one inherited from an affected parent, one

410 involving an imprinted locus, and one inherited from an unaffected parent for a condition known

411 to harbor incomplete penetrance/variable expressivity) and the inheritance was unknown in

412 20/93 families. The CNV was confirmed *de novo* in 7/18 (39%) of the families with an X-linked

413 disorder. Detailed information on the CNV of each family is provided in Table S1. The

414 predominant phenotype present in the 173 families was neurodevelopmental disorders (54%)

415 followed by neuromuscular disorders (15%), but the cohort with causal CNVs also included

416 individuals with multiple congenital anomalies, hematological, ocular, and renal phenotypes.

417 The degree of prescreening before research exome differed between individuals from different

418 sub-cohorts and was therefore non-uniform across different phenotypes.

419

420 The estimated sizes of causal CNVs by exome ranged from 293 bp to 80 Mb (Figure 3C).

421 Twenty-two CNVs involved one exon and 14 CNVs involved two exons, which is below the

422 benchmarked resolution of GATK-gCNV indicating it may be able to detect even smaller CNVs

423 when allowing for a higher false positive rate. Large CNVs were also identified as some

424 individuals did not have CMA prior to research enrollment. Large CNVs tend to be fragmented

425 into multiple small GATK-gCNV calls. We interpreted fragmented CNVs as being part of a larger

426 CNV event in 35 families (35/173 (20%)) in this cohort after looking at the copy number plot
427 and/or validation methods.

428
429 We sought to evaluate how many of the causal CNVs could have been detected by one of the
430 standard clinical CMAs, which is distinct from a high-density clinical array which often has one
431 or more probes per exon. Standard CMAs usually detect CNVs larger than 50-100 Kb but the
432 resolution varies across the genome and across different array designs as the probes are not
433 evenly spaced but are clustered around regions of clinical interest. CNV detection sensitivity by
434 a representative standard CMA was assessed based on the minimum number of probes
435 considered “sufficient” for CNV calling per target, which is defined as ≥ 5 probes for the Agilent
436 GenetiSure Cyto array.²⁸ Based on this, we estimate that 44% (76/173) of these CNVs are
437 unlikely to have been detected by standard CMA.

438
439 More than half of the CNVs (105/173 (61%)) were validated by various orthogonal methods,
440 such as CMA, PCR, FISH, karyotype, MLPA, Sanger across the CNV or breakpoints, or short or
441 long read genome sequencing. Of note, some of these methods did not provide breakpoints but
442 rather only confirmed the copy number change. Of the 105 validated CNVs, 30 (29%) showed
443 differences when comparing the initial results with the orthogonal validation results: 19 showed
444 differences in gene/exon content and 11 showed differences in SV type. Importantly, the
445 difference in gene or exon content identified in 19 families did not result in a change in the
446 clinical interpretation of the CNV. Of note, only one of these 19 CNVs was curated as a VUS
447 and the difference in the number of exons included in the CNV did not change the scoring and
448 classification of this CNV. The 11 cases with different SV type consisted of eight complex SVs
449 which were either incompletely characterized or not suspected by GATK-gCNV on the exome,
450 and a recurrent Alu insertion in the *MAK* gene (OMIM #614181)³⁵ identified in three individuals
451 with retinitis pigmentosa. This insertion was miscalled as a deletion by the GATK-gCNV

452 pipeline, but manual inspection of the exome reads showed discordant read pairs compatible
453 with an Alu insertion. Sanger sequencing resolved the nature of this event.

454
455 Overall, there were 10 confirmed complex SVs in this cohort. We defined a complex SV as a
456 complex rearrangement typically composed of three or more breakpoint junctions that cannot be
457 characterized as a single canonical SV type.³² A complex SV was suspected on the GATK-
458 gCNV calls in 11 families (del/dup, paired dup, etc); seven of these were confirmed by genome,
459 qPCR or CMA (Table S1) and four remained unvalidated. Two deletions and one duplication
460 identified by GATK-gCNV in three different families were revealed to be complex SVs (paired
461 deletion inversions and a paired inversion duplication) when validated by genome sequencing or
462 long-range PCR.

463
464 Twenty-four unrelated families with causal CNVs had a recurrent CNV that was identified in
465 more than one other unrelated family in this cohort. The recurrent 22q11.2 microdeletion
466 syndrome (OMIM #188400) was identified in nine individuals with neurodevelopmental disorders
467 in this cohort. Two individuals with a neurodevelopmental disorder were diagnosed with 22q13.3
468 deletion syndrome (Phelan-McDermid syndrome (OMIM #606232)). The 17q12 deletion
469 syndrome (OMIM #614527) was identified in two individuals with renal cystic disease. There
470 were multiple recurrent CNVs identified in the subgroup of individuals with retinal disorders in
471 this cohort. Indeed, four individuals of European ancestry affected with cone rod dystrophy had
472 a heterozygous 1-exon-deletion in *CLN3* (OMIM #204200) in trans with a pathogenic variant.³⁶
473 A founder variant in the Ashkenazi Jewish population, an Alu insertion in *MAK* (OMIM
474 #614181)^{35,37}, was found in three affected individuals of this ancestry. Two individuals of
475 different ancestries affected with retinitis pigmentosa were homozygous for the same 2-exon-
476 deletion in *EYS* (OMIM #602772), a deletion previously reported in the literature.^{36,38} Two
477 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 standards²⁰ 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

504 the number of protein-coding genes included in each CNV and compared that number to three
505 different reference databases: OMIM (<https://genescout.omim.org/>), DECIPHER browser
506 (<https://www.deciphergenomics.org/browser>) and ClinGen browser
507 (<https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=>) (Table S1). The
508 vast majority of CNVs (148/173 (86%)) showed differences in gene number between these
509 three commonly used databases). Using the 2020 ACMG/ClinGen CNV interpretation
510 standards²⁰, different points are scored based on the number of genes included in a CNV
511 (section 3 of the standards). For example, 0 points are given for a deletion with 0-24 genes,
512 0.45 points for a deletion of 25-34 genes, and 0.9 points for a deletion of more than 35 genes.
513 For copy gain, 0 points are given for 0-34 genes, 0.45 points are given for 35-49 genes and 0.9
514 points for more than 50 genes. We used the number of genes provided by the OMIM database
515 to perform the curation. Using the OMIM database versus DECIPHER resulted in a different
516 final score for 24/148 (16%) CNVs, but this would only alter the final classification for one CNV,
517 as points were awarded from other sections. That altered case was a 857kb *de novo* 22q13
518 duplication which would be classified as a VUS if we use the gene number provided by OMIM
519 (28 protein-coding genes) but would be classified as pathogenic if we had used DECIPHER
520 browser (35 protein-coding genes). Detailed information on the CNV curation of each family is
521 provided in Table S1.

522

523 DISCUSSION

524 We present the analysis and curation results from CNV calling on exome data across a large
525 and phenotypically heterogeneous cohort. The additional 2.6% solve rate of exome CNV calling
526 identified in this cohort is comparable to previously reported diagnostic yield in other cohorts.¹²⁻
527 ¹⁶ In this cohort, most causal CNVs were deletions. Duplications were more common in the
528 callset, but are less likely to disrupt gene function and also typically require more functional
529 investigation to confirm a deleterious effect. Our callset contains many candidate duplications

530 (and deletions) that could potentially elucidate additional affected families, but their
531 pathogenicity remains uncertain and has not been further investigated.

532
533 Similar to using the probes on a microarray to estimate CNV size, the size of a CNV from
534 exome analysis is an estimate based on which exons have an abnormal copy number, but the
535 breakpoints typically occur somewhere within the introns. In addition, some exons have more
536 heterogeneous coverage and the deletion or duplication may involve more or fewer exons than
537 predicted. This can also result in a large CNV being called as multiple smaller events, but when
538 the data is reviewed, it can often be assembled into a larger event.

539
540 In this study, we did not attempt to validate and map all the CNV breakpoints, and we did not
541 assess the validation rate of GATK-gCNV as this has been done previously.¹¹ A small number
542 of CNVs were nonetheless confirmed by genome sequencing by the Broad CMG as part of
543 initial efforts to validate gCNV performance. Of the 23 deletions and two duplications identified
544 by GATK-gCNV and validated by Broad CMG genomes, these were resolved as 22 deletions,
545 one duplication and two complex SVs. We recommended that any candidate CNV variants be
546 confirmed with an orthogonal method and generally validations are performed by the
547 collaborating researcher who recruited the individual for sequencing. The sensitivity of GATK-
548 gCNV decays greatly for CNVs smaller than three exons (e.g. only ~50% for CNVs involving 1
549 exon), but the precision is relatively stable¹¹; interestingly, 36 CNVs (36/173 (21%)) in this
550 cohort involved fewer than 3 exons, highlighting the benefit of reviewing the full dataset with the
551 context of the patient's phenotype and for some cases, a pathogenic variant in trans, can
552 highlight small or poor quality CNV calls that warrant further attention. More than half of CNVs
553 were validated by various methods and validation is either underway or may not be possible for
554 the remainder of the identified CNVs. Importantly, the difference in size and in gene/exon
555 content for validated CNVs did not lead to a change in the interpretation of any of the CNVs

556 initially identified as causal, but it is possible that some interesting CNVs in this cohort were
557 overlooked for that reason.

558
559 GATK-gCNV can only call deletions or duplications on exome data, so seven suspected
560 complex SVs and three initially unsuspected complex SVs in this cohort were identified by
561 orthogonal validation methods. We likely underdetected complex SVs as 39% of the CNVs in
562 this cohort were not validated and some validation methods would miss a more complex event,
563 such as droplet digital or quantitative PCR which only confirm the abnormal copy number
564 without mapping the breakpoints.

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

574 High-quality reference population data is essential for effective CNV analysis. The gnomAD SV
575 database stands as a pivotal resource in human genetics but is currently limited to sequencing
576 data from short-read genomes. We used the database to evaluate if a given CNV was present
577 in the general population, which we found was useful for variant analysis and prioritization.

578 There are a myriad of technical differences between genome and exome sequencing and, while
579 studies have shown high overlap between CNV calling between the two techniques, the planned
580 addition of CNV calling on gnomAD exomes is anticipated to improve clinical CNV interpretation

581 since they will be more analogous from a technical standpoint. As the gnomAD SV dataset
582 expands in terms of size (incorporating both exome and genome data) and ancestral diversity,
583 its utility as an invaluable tool for both rare disease diagnosis and broader genetic studies will
584 only increase.

585 Standards for CNV classification are an important yet challenging area requiring ongoing
586 development. We proposed new evidence criteria to enable the assessment of the pathogenicity
587 of all CNVs that were thought to be causal in our cohort. We identified four areas that needed
588 additions or refinements. First, we suggested that functional data, including expression assays
589 (Western blot, PCR, RNA sequencing) and cellular assays (localization/function), be
590 incorporated as evidence at the supporting level of 0.15 points, and could be increased in
591 weight as appropriate. For example, abnormalities observed in RNA sequencing data or an
592 animal model with the same variant recapitulating the phenotype could be scored 0.3 or 0.45
593 points, respectively. Given the increasing availability of RNA sequencing, we suggest that
594 incorporating scoring for functional evidence is essential for CNV classification. Second, to
595 score CNVs involving genes associated with disorders with autosomal recessive inheritance, we
596 proposed an approach inspired by the ACMG/AMP criteria PM3 used for SNVs by incorporating
597 phase and classification of the second variant (Table 1). The point-based system suggested in
598 the PM3 criteria was translated into points of similar strength level in the Riggs quantitative
599 framework. We also used the PVS1 flowchart³⁰ (or criteria 2E in Riggs *et al.*²⁰) for intragenic
600 CNVs or CNVs including at least one gene that had an established gene-disease relationship
601 following an autosomal recessive inheritance. Additional points were added based on
602 phenotype specificity and familial segregation. Third, to classify CNVs that follow an X-linked
603 inheritance pattern, we developed a scoring system based on biological sex of the proband,
604 parental genotype, and affected status of the transmitting parent (Figure 2). Points were
605 upgraded by one or two strength levels based on phenotype specificity. We also used the PVS1

606 flowchart³⁰ 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

631 datasets. We estimate that 44% of the 173 causal CNVs may not have been detected by
632 standard clinical CMAs. In classifying these variants, we advanced the current standards to take
633 into account additional types of evidence contributing to the systematic framework to assess the
634 pathogenicity of CNVs.

635

636 **Data and code availability**

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

640

641 **Acknowledgements**

642 We thank the families who participated in this study for sharing their samples and medical data,
643 along with all the research groups who collaborate with the Broad CMG. G.L. was supported by
644 the Fonds de recherche en santé du Québec (FRQS). A.S.J. was supported by a
645 Massachusetts General Hospital (MGH) Fund for Medical Discovery Research Award. V.S.G.
646 was supported by NIH NHGRI T32 (T32HG010464) and M.H.W. by NIH/NICHD K23HD102589.
647 Sequencing and analysis were provided by the Broad CMG, funded by the National Human
648 Genome Research Institute grants UM1HG008900, U01HG0011755 and R01HG009141. The
649 content is solely the responsibility of the authors and does not necessarily represent the official
650 views of the National Institutes of Health. Additional funding came from NIH/NINDS grants
651 R01NS032457, R01NS058721, NIH/NIDDK grant RC2DK122397, Sanofi Genzyme, Ultragenyx,
652 LGMD2I Research Fund, Samantha J. Brazzo Foundation, LGMD2D Foundation, Kurt+Peter
653 Foundation, Muscular Dystrophy UK, Coalition to Cure Calpain 3, European Union's Horizon
654 2020 research and innovation programme (grant agreement No. 779257 (Solve-RD)), the
655 Murdoch Children's Research Institute, the Harbig Foundation, the Victorian Government's
656 Operational Infrastructure Support Program, TUBITAK ((The Scientific and Technological

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

661

662 **Declaration of interests**

663 H.L.R. has received support from Illumina and Microsoft to support rare disease gene discovery
664 and diagnosis. A.O-D.L. has consulted for Tome Biosciences and Ono Pharma USA Inc. D.G.M
665 is a paid advisor to GlaxoSmithKline, Insitro, Variant Bio and Overtone Therapeutics, and has
666 received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Google, Merck,
667 Microsoft, Pfizer, and Sanofi-Genzyme. C.A.W. is a paid advisor to Maze Therapeutics. M.E.T.
668 receives research funding from Microsoft Inc, Illumina Inc and Levo Therapeutics. The
669 remaining authors declare no competing interests.

670

671 **Web resources**

672 seqr, <https://seqr.broadinstitute.org/>
673 GATK-gCNV, <https://app.terra.bio/#workspaces/help-gatk/Germline-CNVs-GATK4>
674 DECIPHER, <https://www.deciphergenomics.org/>
675 OMIM, <https://www.omim.org/>, <https://genescout.omim.org/>
676 ClinGen, [https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=
677 gnomAD, <https://gnomad.broadinstitute.org/>
678 ClinVar, <https://www.ncbi.nlm.nih.gov/clinvar/>
679 MatchMaker Exchange, <https://www.matchmakerexchange.org>
680 StrVCTVRE, <https://strvctvre.berkeley.edu/>](https://search.clinicalgenome.org/kb/gene-dosage?page=1&size=25&search=)

681

682

683

684

685

686

687

688

689

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.

- 730 Genome Med. 12, 14.
- 731 13. Pfundt, R., Del Rosario, M., Vissers, L.E.L.M., Kwint, M.P., Janssen, I.M., de Leeuw, N.,
732 Yntema, H.G., Nelen, M.R., Lugtenberg, D., Kamsteeg, E.-J., et al. (2017). Detection of
733 clinically relevant copy-number variants by exome sequencing in a large cohort of genetic
734 disorders. *Genet. Med.* 19, 667–675.
- 735 14. Marchuk, D.S., Crooks, K., Strande, N., Kaiser-Rogers, K., Milko, L.V., Brandt, A., Arreola,
736 A., Tilley, C.R., Bizon, C., Vora, N.L., et al. (2018). Increasing the diagnostic yield of exome
737 sequencing by copy number variant analysis. *PLoS One* 13, e0209185.
- 738 15. Bergant, G., Maver, A., Lovrecic, L., Čuturilo, G., Hodzic, A., and Peterlin, B. (2018).
739 Comprehensive use of extended exome analysis improves diagnostic yield in rare disease:
740 a retrospective survey in 1,059 cases. *Genet. Med.* 20, 303–312.
- 741 16. Testard, Q., Vanhoye, X., Yauy, K., Naud, M.-E., Vieville, G., Rousseau, F., Dauriat, B.,
742 Marquet, V., Bourthoumieu, S., Geneviève, D., et al. (2022). Exome sequencing as a first-
743 tier test for copy number variant detection: retrospective evaluation and prospective
744 screening in 2418 cases. *J. Med. Genet.* 59, 1234–1240.
- 745 17. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V.,
746 Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for
747 medical and population genetics. *Nature* 581, 444–451.
- 748 18. Sharo, A.G., Hu, Z., Sunyaev, S.R., and Brenner, S.E. (2022). StrVCTVRE: A supervised
749 learning method to predict the pathogenicity of human genome structural variants. *Am. J.*
750 *Hum. Genet.* 109, 195–209.
- 751 19. Kleinert, P., and Kircher, M. (2022). A framework to score the effects of structural variants
752 in health and disease. *Genome Res.* 32, 766–777.
- 753 20. Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G.,
754 Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the
755 interpretation and reporting of constitutional copy-number variants: a joint consensus
756 recommendation of the American College of Medical Genetics and Genomics (ACMG) and
757 the Clinical Genome Resource (ClinGen). *Genet. Med.* 22, 245–257.
- 758 21. Bamshad, M.J., Shendure, J.A., Valle, D., Hamosh, A., Lupski, J.R., Gibbs, R.A.,
759 Boerwinkle, E., Lifton, R.P., Gerstein, M., Gunel, M., et al. (2012). The Centers for
760 Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare
761 Mendelian conditions. *Am. J. Med. Genet. A* 158A, 1523–1525.
- 762 22. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir,
763 Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al. (2019). Insights into
764 genetics, human biology and disease gleaned from family based genomic studies. *Genet.*
765 *Med.* 21, 798–812.
- 766 23. Baxter, S.M., Posey, J.E., Lake, N.J., Sobreira, N., Chong, J.X., Buyske, S., Blue, E.E.,
767 Chadwick, L.H., Coban-Akdemir, Z.H., Doheny, K.F., et al. (2022). Centers for Mendelian
768 Genomics: A decade of facilitating gene discovery. *Genet. Med.* 24, 784–797.
- 769 24. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The

- 770 Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.
771 *Am. J. Hum. Genet.* **83**, 610–615.
- 772 25. Pais, L.S., Snow, H., Weisburd, B., Zhang, S., Baxter, S.M., DiTroia, S., O’Heir, E.,
773 England, E., Chao, K.R., Lemire, G., et al. (2022). seqr: A web-based analysis and
774 collaboration tool for rare disease genomics. *Hum. Mutat.* **43**, 698–707.
- 775 26. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk,
776 M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O’Donnell-Luria, A.H., et al. (2020).
777 Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**,
778 452–458.
- 779 27. Tai, A.C., Parfenov, M., and Gorham, J.M. (2018). Droplet Digital PCR with EvaGreen
780 Assay: Confirmational Analysis of Structural Variants. *Curr. Protoc. Hum. Genet.* **97**, e58.
- 781 28. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C.,
782 Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive
783 assessment of array-based platforms and calling algorithms for detection of copy number
784 variants. *Nat. Biotechnol.* **29**, 512–520.
- 785 29. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight,
786 S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the
787 Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed
788 by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100**, 895–906.
- 789 30. Abou Tayoun, A.N., Pesaran, T., DiStefano, M.T., Oza, A., Rehm, H.L., Biesecker, L.G.,
790 Harrison, S.M., and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI)
791 (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant
792 criterion. *Hum. Mutat.* **39**, 1517–1524.
- 793 31. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde,
794 M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of
795 sequence variants: a joint consensus recommendation of the American College of Medical
796 Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**,
797 405–424.
- 798 32. Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline
799 and somatic genomes. *Trends Genet.* **28**, 43–53.
- 800 33. Collins, R.L., Brand, H., Redin, C.E., Hanscom, C., Antolik, C., Stone, M.R., Glessner, J.T.,
801 Mason, T., Pregno, G., Dorrani, N., et al. (2017). Defining the diverse spectrum of
802 inversions, complex structural variation, and chromothripsis in the morbid human genome.
803 *Genome Biol.* **18**, 36.
- 804 34. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M.,
805 Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a
806 platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921.
- 807 35. Tucker, B.A., Scheetz, T.E., Mullins, R.F., DeLuca, A.P., Hoffmann, J.M., Johnston, R.M.,
808 Jacobson, S.G., Sheffield, V.C., and Stone, E.M. (2011). Exome sequencing and analysis
809 of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated
810 kinase (MAK) as a cause of retinitis pigmentosa. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E569–

- 811 E576.
- 812 36. Zampaglione, E., Maher, M., Place, E.M., Wagner, N.E., DiTroia, S., Chao, K.R., England,
813 E., Cmg, B., Catomeris, A., Nassiri, S., et al. (2022). The importance of automation in
814 genetic diagnosis: Lessons from analyzing an inherited retinal degeneration cohort with the
815 Mendelian Analysis Toolkit (MATK). *Genet. Med.* 24, 332–343.
- 816 37. Venturini, G., Koskiniemi-Kuendig, H., Harper, S., Berson, E.L., and Rivolta, C. (2015). Two
817 specific mutations are prevalent causes of recessive retinitis pigmentosa in North American
818 patients of Jewish ancestry. *Genet. Med.* 17, 285–290.
- 819 38. Pieras, J.I., Barragán, I., Borrego, S., Audo, I., González-Del Pozo, M., Bernal, S., Baiget,
820 M., Zeitz, C., Bhattacharya, S.S., and Antiñolo, G. (2011). Copy-number variations in EYS:
821 a significant event in the appearance of arRP. *Invest. Ophthalmol. Vis. Sci.* 52, 5625–5631.
- 822 39. Bujakowska, K.M., Fernandez-Godino, R., Place, E., Consugar, M., Navarro-Gomez, D.,
823 White, J., Bedoukian, E.C., Zhu, X., Xie, H.M., Gai, X., et al. (2017). Copy-number variation
824 is an important contributor to the genetic causality of inherited retinal degenerations. *Genet.*
825 *Med.* 19, 643–651.
- 826 40. Ellingford, J.M., Campbell, C., Barton, S., Bhaskar, S., Gupta, S., Taylor, R.L.,
827 Sergouniotis, P.I., Horn, B., Lamb, J.A., Michaelides, M., et al. (2017). Validation of copy
828 number variation analysis for next-generation sequencing diagnostics. *Eur. J. Hum. Genet.*
829 25, 719–724.
- 830 41. McGuigan, D.B., Heon, E., Cideciyan, A.V., Ratnapriya, R., Lu, M., Sumaroka, A., Roman,
831 A.J., Batmanabane, V., Garafalo, A.V., Stone, E.M., et al. (2017). EYS Mutations Causing
832 Autosomal Recessive Retinitis Pigmentosa: Changes of Retinal Structure and Function with
833 Disease Progression. *Genes* 8. 10.3390/genes8070178.
- 834
- 835
- 836
- 837
- 838
- 839
- 840
- 841
- 842
- 843
- 844
- 845

846 **FIGURE LEGENDS**

847 **Figure 1. Exome copy number plot and reads visualization for examples of causal CNVs**

848 **in the Broad CMG cohort.** (A) Individual affected with retinitis pigmentosa with a homozygous

849 single exon deletion in *CRB1* (chr1:197438450-197439442x0, Quality score (QS) = 120)

850 identified on exome. To evaluate the quality of the CNV, the patient's copy number (CN) level

851 (in red) was compared to a cluster of other samples with similar read depth that act as controls.

852 The proband's CN is decreased compared to the background cluster, compatible with a

853 homozygous deletion. Y axis: CN. (B) As breakpoints fell within the exome data, manual

854 inspection of read data from the individual from (A) using the Integrated Genomics Viewer (IGV)

855 showed discordant read pairs, split reads and complete absence of coverage, compatible with a

856 homozygous exon 10 deletion also including part of upstream exon 9 in *CRB1*

857 (chr1:197435257-197441674x0 (NM_201253.3)). Cov= coverage. (C) Individual with multiple

858 congenital anomalies and a heterozygous deletion of 4 exons in *RAB3GAP1* (Warburg micro

859 syndrome) (red, chr2:135162318-135164794x1, QS =92) in trans with a frameshift variant in

860 *RAB3GAP1* (not shown, NM_012233.3: c.2393_2394del, p.Leu798ArgfsTer7), both identified by

861 exome. The presence of the deletion was validated by droplet digital PCR. Y axis: CN. (D)

862 Individual with a neurodevelopmental disorder with a *de novo* 2.6 Mb heterozygous 1q43q44

863 deletion (red, chr1:242523991-245156781x1, QS =3077) identified on exome. The presence of

864 this deletion was validated by quantitative PCR. Y axis: CN. (E) Individual with a

865 neurodevelopmental disorder with a *de novo* 2.1Mb 22q11.2 duplication (red, chr22:18985739-

866 21081116x3, QS =3077) identified on exome. The presence of this duplication was validated by

867 chromosomal microarray. Y axis: CN. All coordinates on GRCh38.

868

869 **Figure 2. Flowchart illustrating how points were scored for CNVs that followed a X-linked**

870 **inheritance.** We incorporated sex of proband, parental genotype and parental affected status to

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

875

876 **Figure 3. Characteristics of CNVs across the entire callset and the subset of causal**

877 **CNVs.** (A) Number of high-confidence CNVs by estimated size that were identified in the Broad
878 CMG exome callset of 6,633 families sequenced between 2016 and 2021. Large CNVs tend to
879 be fragmented into multiple small GATK-gCNV calls, accounting for why there are no CNVs in
880 the >10 Mb category of the graph. These CNVs were interpreted as being part of the same
881 underlying event when looking at the copy number plot and/or validation methods and are
882 presented that way in Figure 3B and 3C. DEL: deletion; DUP: duplication. (B) Mode of
883 inheritance and number of genes involved in each CNV in 173 families in which the CNV was
884 interpreted as causal. The number of genes included in each interval was chosen based on
885 cutoffs suggested for CNV scoring in section 3 of the Riggs *et al.* ACMG/ClinGen standards.²⁰
886 (C) CNV classification by estimated size in 173 families in which the CNV was interpreted as
887 causal by the multidisciplinary team. The causal CNVs consisted of 141 deletions, 15
888 duplications, 3 insertions (miscalled as deletion by GATK-gCNV), and 14 complex structural
889 variants (SV). We interpreted 153 CNVs as likely pathogenic/pathogenic and 20 CNVs as VUS.

890

891

892

893

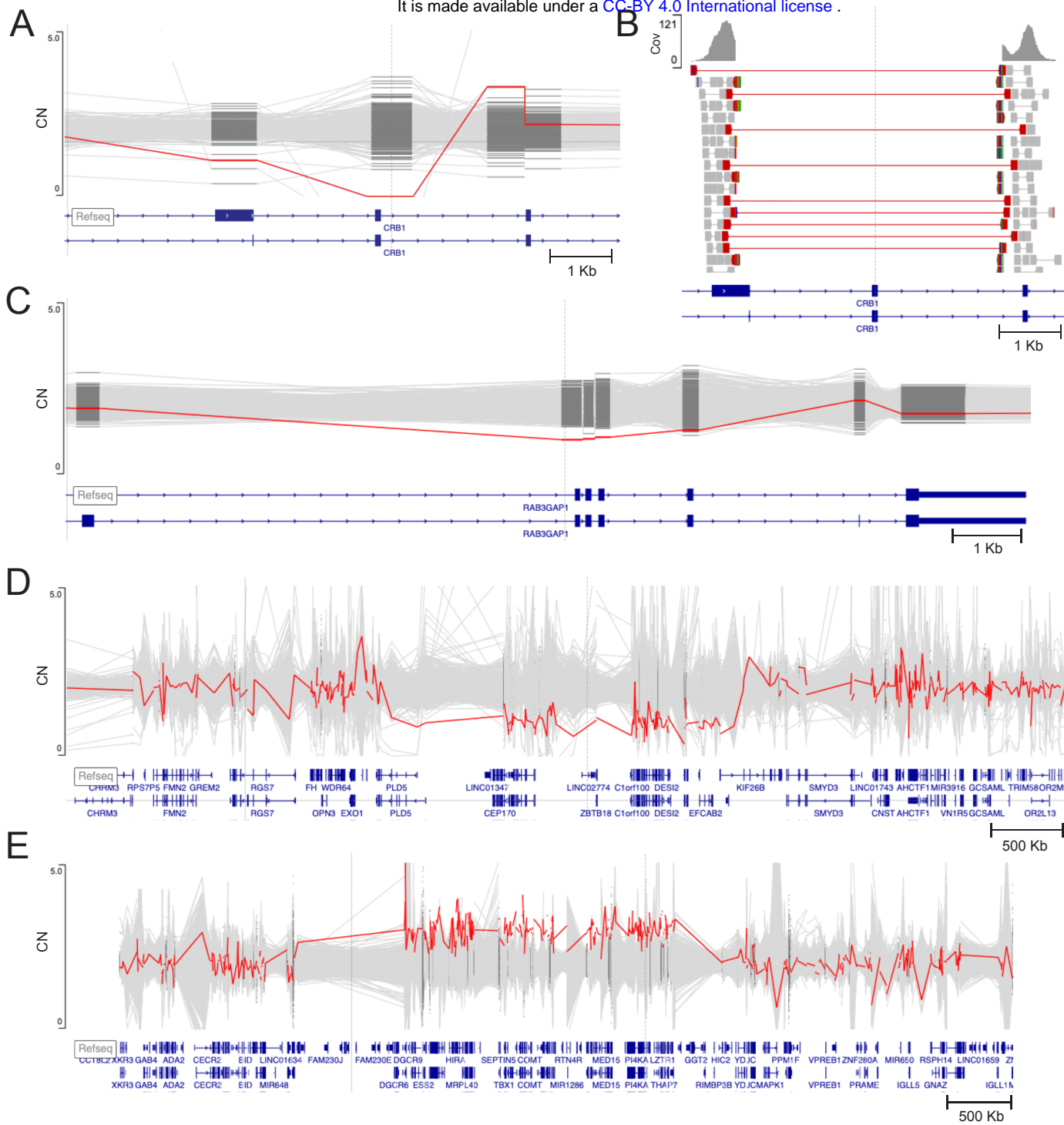
894

895

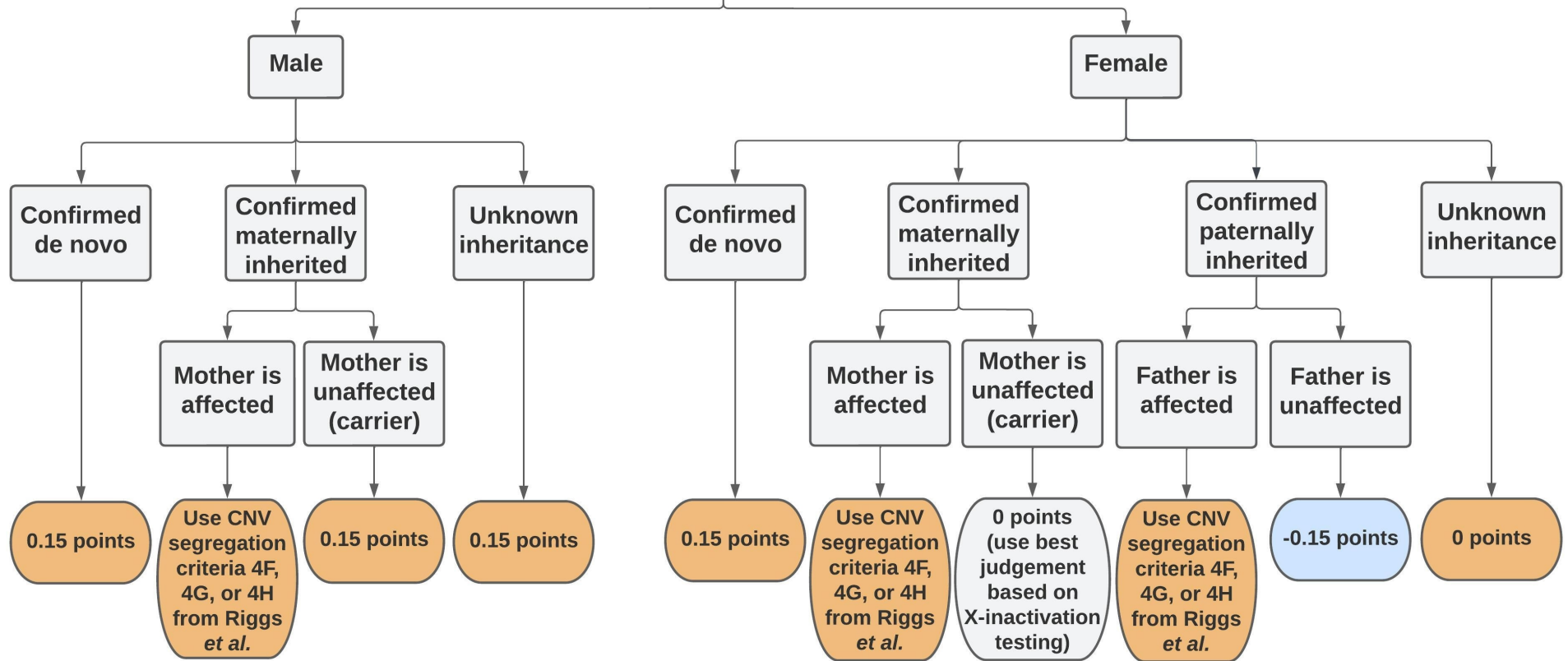
896 **Table 1.** Adapted PM3 table to score CNVs in genes for conditions that follow an autosomal
897 recessive inheritance

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

898
899
900
901



Biological sex of proband



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

