



## ARTICLE

# Optical genome mapping and revisiting short-read genome sequencing data reveal previously overlooked structural variants disrupting retinal disease—associated genes



### ARTICLE INFO

#### Article history:

Received 19 July 2022  
 Received in revised form  
 14 November 2022  
 Accepted 15 November 2022  
 Available online 16 December 2022

#### Keywords:

Inherited retinal diseases  
 Optical genome mapping  
 Next-generation sequencing  
 Short-read genome sequencing  
 Structural variants

### ABSTRACT

**Purpose:** Structural variants (SVs) play an important role in inherited retinal diseases (IRD). Although the identification of SVs significantly improved upon the availability of genome sequencing, it is expected that involvement of SVs in IRDs is higher than anticipated. We revisited short-read genome sequencing data to enhance the identification of gene-disruptive SVs.

**Methods:** Optical genome mapping was performed to improve SV detection in short-read genome sequencing—negative cases. In addition, reanalysis of short-read genome sequencing data was performed to improve the interpretation of SVs and to re-establish SV prioritization criteria.

**Results:** In a monoallelic *USH2A* case, optical genome mapping identified a pericentric inversion (173 megabase), with 1 breakpoint disrupting *USH2A*. Retrospectively, the variant could be observed in genome sequencing data but was previously deemed false positive. Reanalysis of short-read genome sequencing data (427 IRD cases) was performed which yielded 30 pathogenic SVs affecting, among other genes, *USH2A* ( $n = 15$ ), *PRPF31* ( $n = 3$ ), and *EYS* ( $n = 2$ ). Eight of these (>25%) were overlooked during previous analyses.

**Conclusion:** Critical evaluation of our findings allowed us to re-establish and improve our SV prioritization and interpretation guidelines, which will prevent missing pathogenic events in future analyses. Our data suggest that more attention should be paid to SV interpretation and the current contribution of SVs in IRDs is still underestimated.

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

## Introduction

Structural variants (SVs) are increasingly recognized as important causes of inherited diseases, and pathogenic variants have been described to be implicated in many diseases, including developmental disorders and sensory disorders.<sup>1–3</sup> SVs are defined as large (>1 kb) genomic

aberrations and can be subdivided into unbalanced (eg, deletions and duplications) and balanced (eg, inversions and translocations) rearrangements.<sup>4</sup> The number of identified pathogenic SVs has been growing rapidly, and SV identification has significantly improved with the arrival of genome sequencing technologies (Reurink et al unpublished).<sup>5,6</sup>

\*Correspondence and requests for materials should be addressed to Suzanne E. de Bruijn, Department of Human Genetics, Radboud University Medical Center, P.O. Box 9101, 6500, HB, Nijmegen, The Netherlands. E-mail address: [Suzanne.deBruijn@radboudumc.nl](mailto:Suzanne.deBruijn@radboudumc.nl)

A full list of authors and affiliations appears at the end of the paper.

doi: <https://doi.org/10.1016/j.gim.2022.11.013>

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

Inherited retinal diseases (IRDs) are a group of genetically heterogeneous disorders, and pathogenic variants in IRDs have been described in >270 genes (RetNet, <https://sph.uth.edu/retnet/>). The contribution of SVs to the mutational landscape of IRDs is currently being estimated to range between 5% and 15%.<sup>5,7-9</sup> Despite extensive sequencing efforts, in approximately one-third of IRD cases, no conclusive genetic diagnosis could be established.<sup>10</sup> Although literature reports suggest that this percentage can be improved by the implementation of (short-read) genome sequencing, there is still a significant degree of missing heritability (Reurink et al [unpublished]).<sup>5,6</sup> One of the main hypotheses for this missing heritability is the presence of pathogenic SVs that cannot be detected using short-read sequencing approaches. Several studies have reported the additive value of long-read sequencing or cytogenetic approaches that make use of (ultra)long DNA molecules for the identification of SVs.<sup>11-13</sup>

In this study, we showed that the power of SV detection from short-read data, and therefore the prevalence of pathogenic SVs, is still being underestimated in IRDs. Optical genome mapping (OGM) was performed to improve SV detection in genetically unexplained IRD cases and, surprisingly, revealed that several pathogenic SVs were overlooked during our previously performed short-read genome analyses. By revisiting genome sequencing data generated from established IRD cohorts and performing a focused SV reanalysis, several previously overlooked pathogenic SVs could be identified, including, but not limited to, large (pericentric) inversions and small intragenic deletions. Several lessons were learned during the process of data reanalysis, which allowed us to re-establish and optimize our SV prioritization protocols. We therefore advocate that more attention should be paid to SV interpretation during genome data analyses, and we believe that this will facilitate (partial) explanation for the missing heritability in IRDs and possibly in other inherited disorders as well.

## Materials and Methods

### Patient cohort

Genome sequencing data were collected from 427 IRD probands. This IRD study cohort included both genetically explained and unexplained samples that were incorporated in recent studies with 100 IRD cases described by Fadaie et al<sup>6</sup> and 100 Usher syndrome and monoallelic *USH2A*-associated recessive retinitis pigmentosa cases by Reurink et al (unpublished). Other samples included in the cohort were part of unpublished studies ( $n = 96$ ) focused on the analyses of genomic sequences of genetically unexplained IRD cases and a portion of samples ( $n = 131$ ) were not analyzed previously (Supplemental Figure 1). Before participation in these studies, all probands were prescreened

using either exome sequencing or targeted gene panel sequencing, which yielded no conclusive genetic diagnosis. Informed consent was obtained from all participants or their legal representatives.

### Genome sequencing

Genomic DNA was isolated from peripheral blood lymphocytes following standard procedures and analyzed through genome sequencing as described previously.<sup>6</sup> Sequencing was performed by BGI on a BGISEq500 using a 2x 100 basepair (bp) or 2x 150 bp paired-end module, with a minimal median coverage per genome of 30 fold. Read mapping to the Human Reference Genome build GRCh38/hg38 and single-nucleotide variant (SNV) calling were performed using Burrows-Wheeler Aligner V.0.78<sup>14</sup> and Genome Analysis Toolkit HaplotypeCaller (Broad Institute), respectively. SVs were called using Manta structural variant caller,<sup>15</sup> which is based on read-pair signals (split reads and discordant read pairs) and read-depth signals (copy number changes). In addition, copy number variant (CNV) detection was performed based on read-depth evidence using Canvas Copy Number Variant Caller.<sup>16</sup>

### OGM

For a single case, OGM (Bionano Genomics) was performed as previously described.<sup>12,13,17</sup> In brief, ultrahigh molecular weight DNA was isolated from whole peripheral blood (EDTA) using the SP Blood & Cell Culture DNA Isolation Kit (Bionano Genomics). DNA labeling was performed using the Direct Label and Stain (DLS) DNA Labeling Kit (Bionano Genomics), and the labeled sample was loaded on a 3x1300 Gb Saphyr chip (G2.3) on a Saphyr instrument (Bionano Genomics). Annotated de novo assembly using the genome build hg19 was performed using Bionano Solve version 3.6.1, which includes 2 separate algorithms for SV and CNV detection as described previously.<sup>12</sup> SV calls that were absent in a control OGM data set (>200 human population control samples) were prioritized. Identified candidate variants overlapping with an IRD-associated gene were visualized and investigated in Bionano Access version 1.6.1.

### Reanalysis of genome sequencing data and variant selection

All 278 IRD-associated genes listed on the RetNet webpage (<https://sph.uth.edu/retnet/>, accessed October 1, 2022) were investigated in this study, and genomic positions were extracted using the Ensembl genome browser.<sup>18</sup> All SVs and CNVs with at least 1 breakpoint within one of the listed IRD-associated genes ( $\pm 1$  megabase [Mb] flanking regions) were extracted. Extracted SVs and CNVs were combined for all samples followed by in-depth variant assessment.

Coding SVs and CNVs were filtered and selected based on a minor allele frequency (MAF) of <1% in 1000 Genomes,<sup>19</sup> DECIPHER,<sup>20</sup> and our in-house SV database (consisting of 920 genomes of presumably healthy unrelated individuals). Inversion events were only considered when at least one of the breakpoints was located within an IRD-associated gene and therefore disrupt the gene.

After identification of a candidate variant, in-depth genome sequencing (re)analysis was performed including assessment of SNVs. SNVs were filtered based on an MAF of <1% (gnomAD V2.1.1<sup>21</sup> and our in-house SNV database [~15,000 alleles]). All SNVs in IRD-associated genes were evaluated. Missense variants were prioritized when a deleterious effect was predicted by at least 2 in silico tools: CADD-PHRED<sup>22</sup> ( $\geq 15$ , range = 0-48), Sorting Intolerant from Tolerant (SIFT)<sup>23</sup> ( $\leq 0.05$ , range = 0-1), Polymorphism Phenotyping (PolyPhen) version 2<sup>24</sup> ( $\geq 0.450$ , range = 0-1), or MutationTaster<sup>25</sup> (deleterious). Potential effects of missense, synonymous, or intronic variants on splicing were assessed using the deep-learning splice prediction algorithm SpliceAI<sup>26</sup> ( $\geq 0.2$ ) using default settings.

## Variant validation

Potentially pathogenic SNVs that were not previously identified were validated using Sanger sequencing. Identified SVs were validated by visualization using the Integrative Genomics Viewer (IGV) software V2.4<sup>27</sup> (Broad Institute), and breakpoints were confirmed using polymerase chain reaction (PCR) amplification and Sanger sequencing. Primer sequences for SV validations are listed in [Supplemental Table 1](#), and PCR conditions used are available upon request. Segregation analysis was performed when DNA of family members was available ([Supplemental Table 2](#)).

For one variant (a full-gene deletion of *MERTK*), validation was performed using quantitative real-time PCR (qPCR). qPCR was performed on the genomic DNA from the affected individual and unaffected unrelated controls ( $n = 2$ ). The experiment was performed using GoTaq qPCR Master Mix (Promega) on a Quantstudio 6 Flex Real-Time PCR System (Applied Biosystems). Primer pairs were designed to amplify parts of the genome in the suspected deleted region as well as regions outside the affected genome region as a reference for standard quantity. Primer sequences are listed in [Supplemental Table 1](#).

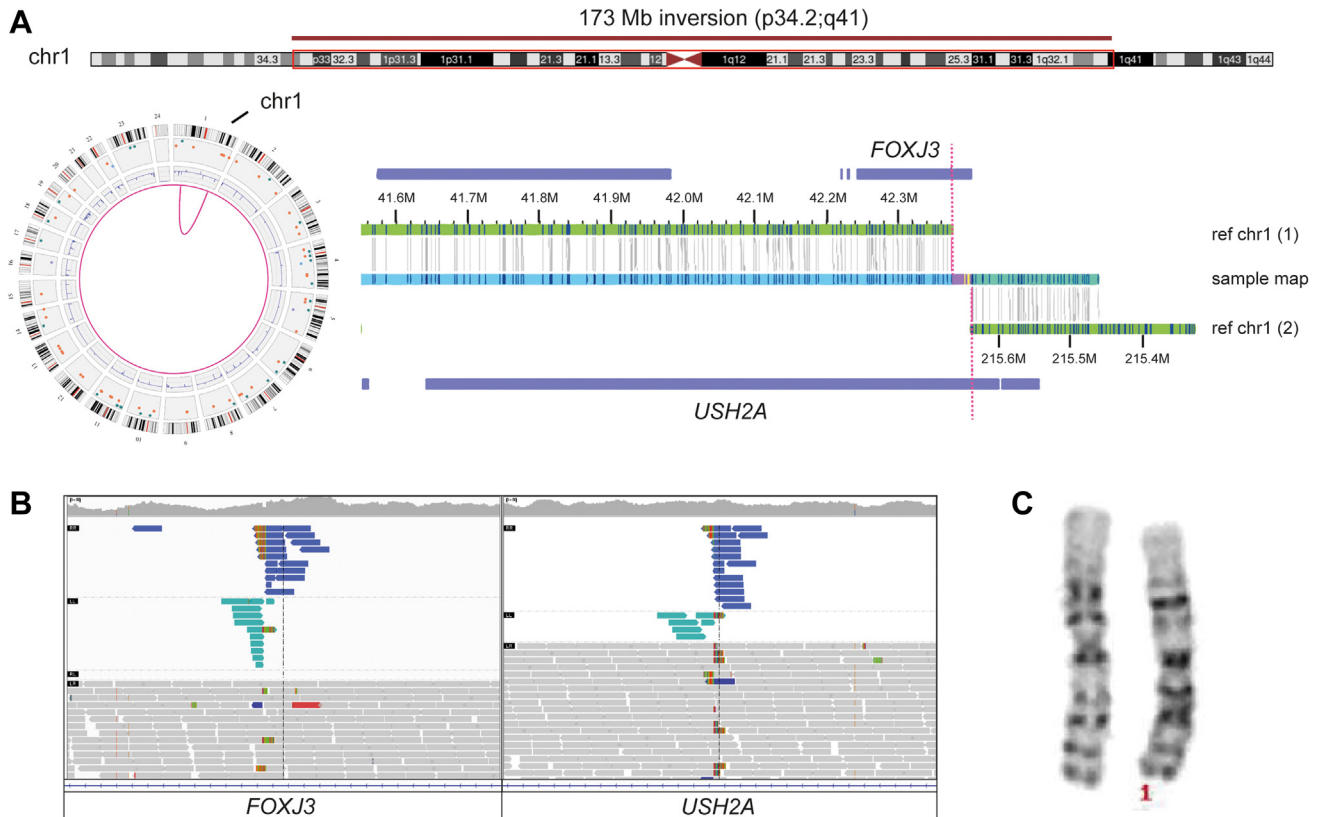
## Results

### OGM reveals a pathogenic 173 Mb pericentric inversion disrupting *USH2A*

The implementation of short-read genome sequencing technologies has significantly improved the diagnostic yield for inherited disorders; nonetheless, many affected

individuals remain genetically unexplained. For IRDs, about 30% of individuals lack a genetic diagnosis after genome sequencing has been performed, which indicates that a large diagnostic gap exists (Reurink et al [unpublished]).<sup>6</sup> One of the cases that remained genetically unexplained after genome sequencing was individual USH-44 (Reurink et al [unpublished]). This individual has been previously diagnosed with Usher syndrome type-II (OMIM 276901), a recessively inherited disorder characterized by retinitis pigmentosa and congenital hearing loss and associated with variants in several genes, the most important one being *USH2A*. Genome sequencing did reveal a heterozygous intragenic *USH2A* deletion (c.9258+2601\_9371+1539del, p.(?), NM\_206933.2), spanning exon 46, which is predicted to result in a frameshift, but the second pathogenic allele remained elusive (Reurink et al [unpublished]). As part of this study, we performed OGM: an innovative cytogenetics approach, which allows the efficient detection of SVs using ultralong DNA fragments.<sup>12,13</sup> OGM revealed a total of 5929 SV calls, of which 143 were absent in the OGM control cohort of 204 unrelated individuals. Two of the identified SVs overlapped with the *USH2A* gene, of which one corresponded to the previously identified heterozygous deletion spanning *USH2A* exon 46. The other SV supports a large approximately 173 Mb pericentric inversion event, which involves most of chromosome 1. The 3' breakpoint was predicted to be located within the *USH2A* gene ([Figure 1A](#)) and the 5' breakpoint within *FOXJ3*. The inversion was not present in any of the control samples, and none of the other SV calls were overlapping with any IRD-associated gene. Inspection of the inversion event using the Bionano Access software confirmed the presence of the *USH2A* inversion. In addition, the variant was confirmed using traditional karyotyping in patient-derived cells ([Figure 1](#)).

To determine the exact breakpoints of the inversion event, we interrogated the implicated breakpoint regions in the available short-read genome sequencing data. Retrospectively, we noticed a 173.1 Mb inversion variant call that corresponded to the SV calls obtained from OGM. In addition, split reads spanning the breakpoints of the inversion event could be observed in IGV ([Figure 1B](#)). The variant was previously deemed to be false positive and overlooked, mainly because of the large size of the inverted region and the high number of large inversion events called in this sample. Using the genomic positions derived from the short-read data, breakpoints of the inversion event could be confirmed through PCR and Sanger sequencing (chr1:42320825-215677220delins42320846-215677215inv, hg38). The 3' breakpoint of the inversion is located in intron 62 of *USH2A* and thus disrupts the coding sequence of *USH2A*. Most likely, none or only truncated *USH2A* protein will be produced that lacks several protein domains, among which is the essential transmembrane domain. Based on these results, the variant was classified as pathogenic, and this individual with Usher syndrome was considered genetically solved.



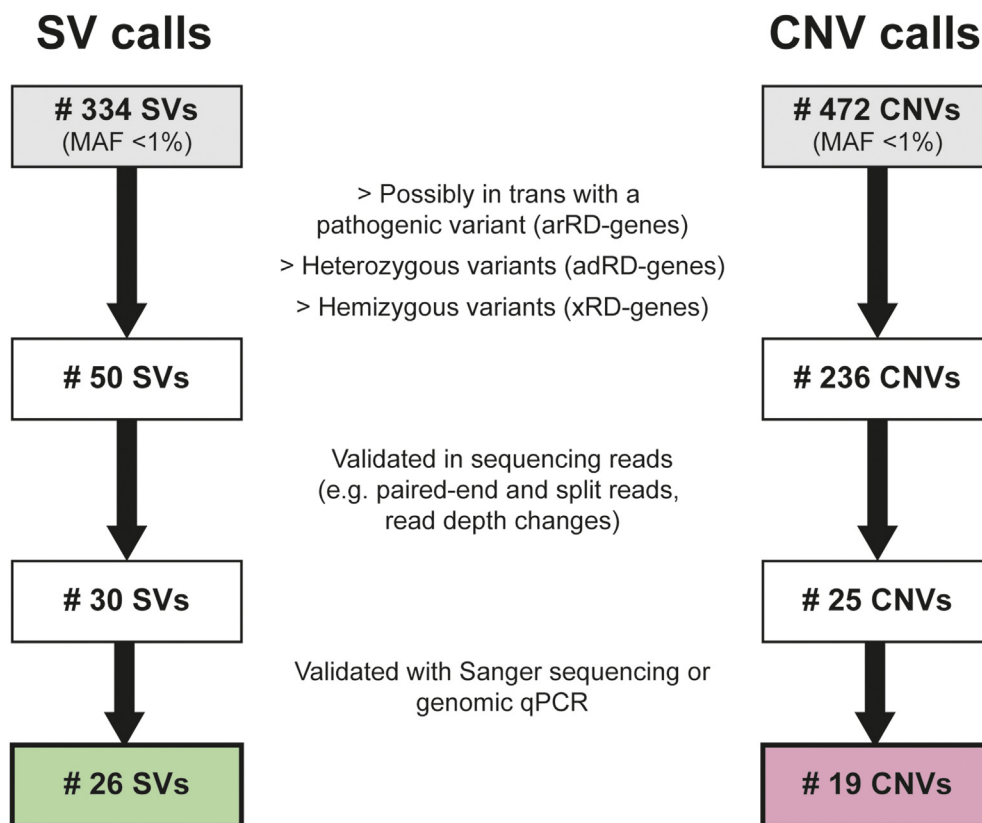
**Figure 1** Optical genome mapping reveals 173 Mb pericentric inversion on chromosome 1 with a breakpoint within the *USH2A* gene. A. Optical genome mapping (OGM) in individual USH-44 (Reurink et al, unpublished) predicted a large pericentric inversion of 173 Mb (*inv*[1]p34.2q41) on chr1, with the 3' breakpoint interrupting *USH2A*. The left panel displays a genome-wide circos plot, which illustrates the inversion event present on chromosome 1. The inversion is represented by a pink line at the inner ring connecting 2 distal regions. In the right panel, the sample genome map is mapped against the ref chr1, showing the inversion structural variant (SV) call and affected genes. The upper green sample genome maps to the 5' side of the breakpoint aligning to *FOXJ3*. The lower green sample maps to the 3' of the breakpoint aligning to *USH2A*. B. Interrogation of short-read genome sequencing data in Integrative Genomics Viewer (IGV) revealed split reads (green-blue colored reads) corresponding to the inversion breakpoints predicted through OGM. Retrospectively, an SV-call (Manta Structural Variant caller) matching with the inversion event was recognized and the exact SV breakpoints could be determined. Breakpoints were polymerase chain reaction amplified and validated using Sanger sequencing. C. The karyogram of this individual confirmed a pericentric inversion on chromosome 1. chr1, chromosome 1; Mb, megabase; ref chr1, reference genome.

## Revisiting genome sequencing data reveals previously overlooked SVs

The identification of the previously overlooked *USH2A* inversion event prompted us to reassess our available short-read genome sequencing data sets and optimize our variant prioritization protocol. We decided to merge SV and CNV data of all 427 genomes that were collected previously from unrelated individuals diagnosed with IRD and performed a comprehensive variant (re)analysis. SV and CNV data of genetically explained samples remained included in the analyses for control purposes. We refrained from filtering SVs based on size or quality criteria to allow the establishment of correct filtering criteria. Considering the large volume of data and the feasibility of these analyses, we decided to focus on rare coding SVs and CNVs only. After variant filtering, 334 SV calls and 472 CNV calls overlapping with an IRD-associated gene and a MAF of <1%

were selected and subjected to a detailed examination (Figure 2).

A total of 5 homozygous and 21 heterozygous candidate SVs, potentially in *trans* with a second heterozygous pathogenic allele, were identified in 25 samples. This included the 173 Mb *USH2A* inversion (variant 30) and an in-frame deletion of exons 22 to 24 in *USH2A*, which was identified in 2 samples (variant 18 and variant 22). Moreover, 1 hemizygous and 3 heterozygous potentially pathogenic variants were identified in IRD cases with an X-linked or a dominant mode of inheritance, respectively (Table 1, Supplemental Tables 2 and 3). All previously identified SVs ( $n = 15$ , Fadaie et al,<sup>6</sup> Reurink et al [unpublished], Velde et al,<sup>28</sup> unpublished data) were detected through this approach, which confirms the suitability of our method. Breakpoints of all SVs except one were validated, and segregation analysis was performed if possible (Supplemental Table 2). For an SV affecting *MERTK*,



**Figure 2 Interpretation and validation of SV and CNV calls.** After variant filtering, 50 SV calls and 236 CNVs calls overlapping with an inherited retinal disease–associated gene and a minor allele frequency of <1% were selected as potentially pathogenic. Of these, only 26 SV calls and 19 CNV calls could be validated by interrogating the raw sequencing data and validation of the variant calls using Sanger sequencing or genomic quantitative polymerase chain reaction. The other SV and CNV calls were considered false-positive calls. Most false-positive calls (98%) comprised deletion events partially spanning the X chromosome that were called in male probands. After interrogation of the sequencing data, it was concluded that these variants were not true hemizygous deletion events. adRD, autosomal dominant retinal disease; arRD, autosomal recessive retinal disease; CNV, copy number variant; SV, structural variant; xRD, X-linked retinal disease.

breakpoints could not be amplified, but the deletion event was validated through genomic qPCR (Supplemental Figure 2). Sizes of identified SVs ranged from 72 bp to 173 Mb, and variants included inversions ( $n = 3$ ), duplications ( $n = 3$ ), and deletions ( $n = 24$ ) (Figure 3). In concordance with previous studies of CNVs and SVs in IRDs,<sup>9</sup> variants in *USH2A* are the main contributor to the variants identified in this study (15/30 variants).

Surprisingly, >25% of the newly identified variants ( $n = 8$ ) were identified in previously analyzed samples, indicating that these variants were previously overlooked. One of these variants is another large inversion event that disrupts the *USH2A* gene (variant 28). The variant was identified in sample USH-42 (Reurink et al [unpublished]), an individual diagnosed with Usher syndrome type-II and heterozygous for a known pathogenic variant in *USH2A*. This individual was previously included in the genome sequencing study performed by Reurink et al (unpublished) but remained genetically unexplained after genome analysis was performed. Only by reassessing the (size) filtering criteria, we were able to pick up this large inversion that

affects *USH2A*. To find a possible explanation why several other variants were misinterpreted during previous analyses, an overview of the respective SV and CNV calls and corresponding quality scores are summarized in Supplemental Table 4.

For all 29 cases in which a candidate SV was identified, in-depth genome sequencing (re)analysis, including SNV analysis, was performed to exclude the presence of other (likely) pathogenic variants in IRD-associated genes that could have been misinterpreted in earlier analyses as well. No additional potentially pathogenic homozygous or compound heterozygous (recessive inheritance), hemizygous (X-linked inheritance), or heterozygous variants (dominant inheritance) were revealed in the affected individuals. All 30 variants are classified as (likely) pathogenic according to the American College of Medical Genetics and Genomics classification system<sup>29,30</sup> and expected to be part of the underlying causative genetic defects responsible for the IRD phenotype in the respective individuals. A detailed summary of the number of genetically explained samples in this and previous studies can be found in Supplemental Figure 3.

**Table 1** Overview of structural variants identified in this study

Variant	Gene	SV	Zyg	Chr	Genomic Positions (hg38)	Consequence	Source
1	<i>ARSG</i>	del	het	17	g.68364608_68373476del	fs, exons 7-8 del	Velde et al <sup>28</sup>
2	<i>CDHR1</i>	inv	hom	10	g.84204183_84262987delins 84204828_84261690inv	fs, exon 9 del, exons 10-17 inv	This study <sup>a</sup>
3	<i>EYS</i>	del	hom	6	g.64986218_65013355del	fs, exon 14 del	This study
4	<i>EYS</i>	del	hom	6	g.64388690_64388840del	fs, exon 29 del	This study
5	<i>HGSNAT</i>	del	het	8	g.43140524_43140595del	In-frame, exon 1 partial del	This study <sup>a</sup>
6	<i>MERTK</i>	del	het	2	g.(?_111986574)_ (112008773_?)del <sup>b</sup>	Complete gene del	This study
7	<i>MFRP</i>	del	hom	11	g.119346419_119352600del	fs, start lost	This study <sup>a</sup>
8	<i>NMNAT1</i>	del	het	1	g.9970211_9972447del	fs, exon 2 del	This study
9	<i>PRPF31</i>	del	het	19	g.54106454_54133135del	Complete gene del	Fadaie et al <sup>6</sup>
10	<i>PRPF31</i>	del	het	19	g.54115080_54121762del	fs, exons 1-3 del	This study <sup>a</sup>
11	<i>PRPF31</i>	del	het	19	g.54116703_54121868del	fs, exons 2-4 partial del	This study <sup>a</sup> , Reurink et al (unpublished)
12	<i>RP2</i>	del	hem	X	g.46859569_46864195del	fs, exon 3 del	This study
13	<i>RPGRIP1</i>	del	hom	14	g.21331505_21335028del	fs, exon 21 del	Fadaie et al <sup>6</sup>
14	<i>SPATA7</i>	del	het	14	g.88428046_88429358del	fs, exon 8 partial del	This study
15	<i>TTLL5</i>	del	het	14	g.75776985_75785341del	fs, exons 24-26 del	This study <sup>a</sup>
16	<i>USH2A</i>	dup	het	1	g.216245168_216388984dup	fs, exons 4-13 dup	Reurink et al (unpublished)
17	<i>USH2A</i>	del	het	1	g.215876077_215877784del	In-frame, exon 43 partial del	Reurink et al (unpublished)
18	<i>USH2A</i>	del	het	1	g.216086060_216149818del	In-frame, exons 22-24 del	Reurink et al (unpublished)
19	<i>USH2A</i>	del	het	1	g.216120207_216246942del	fs, exons 21-22 partial del	Reurink et al (unpublished)
20	<i>USH2A</i>	del	het	1	g.216363578_216364956del	fs, exon 4 partial del	This study
21	<i>USH2A</i>	del	het	1	g.215836478_215843033del	fs, exon 47 del	Reurink et al (unpublished)
22	<i>USH2A</i>	del	het	1	g.216086060_216149818del	In-frame, exons 22-24 del	Reurink et al (unpublished)
23	<i>USH2A</i>	del	het	1	g.216144058_216207338del	In-frame, exons 16-21 partial del	Reurink et al (unpublished)
24	<i>USH2A</i>	dup	het	1	g.215650104_216220580dup	fs, exons 15-65 dup	Reurink et al (unpublished)
25	<i>USH2A</i>	del	het	1	g.215829927_215838026del	fs, exon 47 partial del	Reurink et al (unpublished)
26	<i>USH2A</i>	del	het	1	g.216291967_216334290del	In-frame, exons 5-10 del	Reurink et al (unpublished)
27	<i>USH2A</i>	dup	het	1	g.215965562_216251419dup	In-frame, exons 12-36 dup	This study
28	<i>USH2A</i>	inv	het	1	g.209815568_215637482inv	fs, exons 70-72 inv	This study <sup>a</sup> , Reurink et al (unpublished)
29	<i>USH2A</i>	del	het	1	g.215836452_215841693del	fs, exon 46 del	Reurink et al (unpublished)
30	<i>USH2A</i>	inv	het	1	g.42320825_215677220delins 42320846_215677215inv	fs, exons 63-72 inv	This study <sup>a</sup> , Reurink et al (unpublished)

Source column represents whether the samples were previously analyzed in published or unpublished studies or the samples are novel and included in this study. More detailed variant data can be found in [Supplemental Tables 2 and 3](#).

*chr*, chromosome; *del*, deletion; *dup*, duplication; *fs*, frameshift; *hem*, hemizygous; *het*, heterozygous; *hom*, homozygous; *inv*, inversion; *SV*, structural variant; *Zyg*, zygosity.

<sup>a</sup>Genome sequencing data were previously analyzed, and variant was not identified (published or unpublished studies).

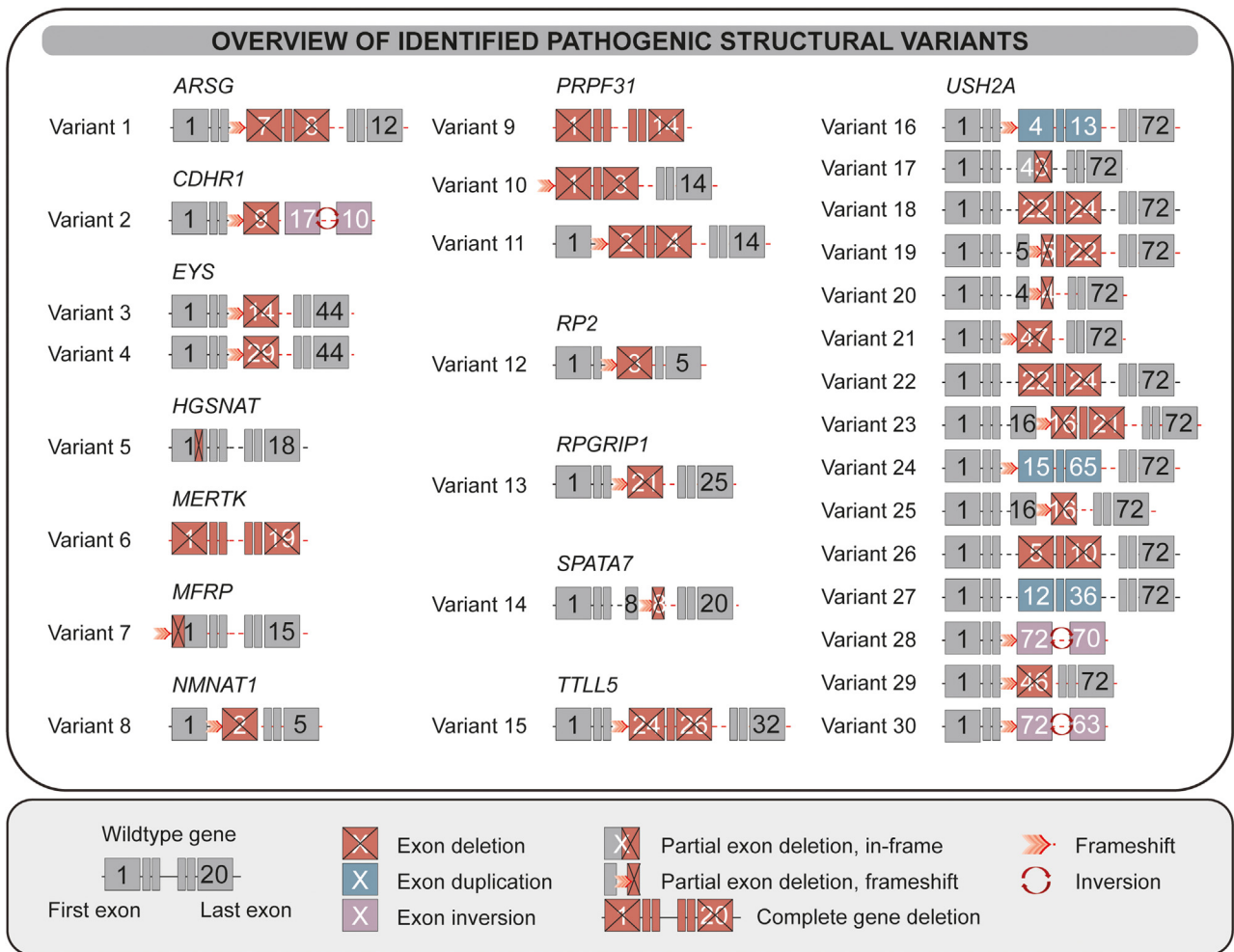
<sup>b</sup>Breakpoints could not be confirmed through Sanger sequencing; SV has been validated using genomic quantitative polymerase chain reaction only.

## Discussion

SVs are considered as important contributors to the mutational landscape of inherited disorders, including IRDs (Reurink et al [unpublished]).<sup>5,6</sup> Although SV detection did improve upon the arrival of short-read genome sequencing, the process of SV detection is still not trivial and SV interpretation is a complex process. On average, 10,000 SVs are called in a sample, of which >50 SVs overlap with the coding regions of an IRD-associated gene including several false-positive calls. There is an increased need for using standard prioritization protocols for the interpretation of SV data. In addition, SV algorithms are still continuously

developing to improve the accurate detection of SVs. Despite these efforts, it is expected that many pathogenic SVs escape detection through short-read sequencing or are misinterpreted, which would suggest that these types of variants might well be one of the most important sources of the reported missing heritability for IRDs.

In this study, OGM was performed to identify large pathogenic SVs in a case that was heterozygous for a pathogenic *USH2A* variant. A large *USH2A*-disruptive pericentric inversion on chromosome 1 was identified, which genetically explained this case after decades of research. In hindsight, interrogation of the short-read genome sequencing data did reveal the presence of a large

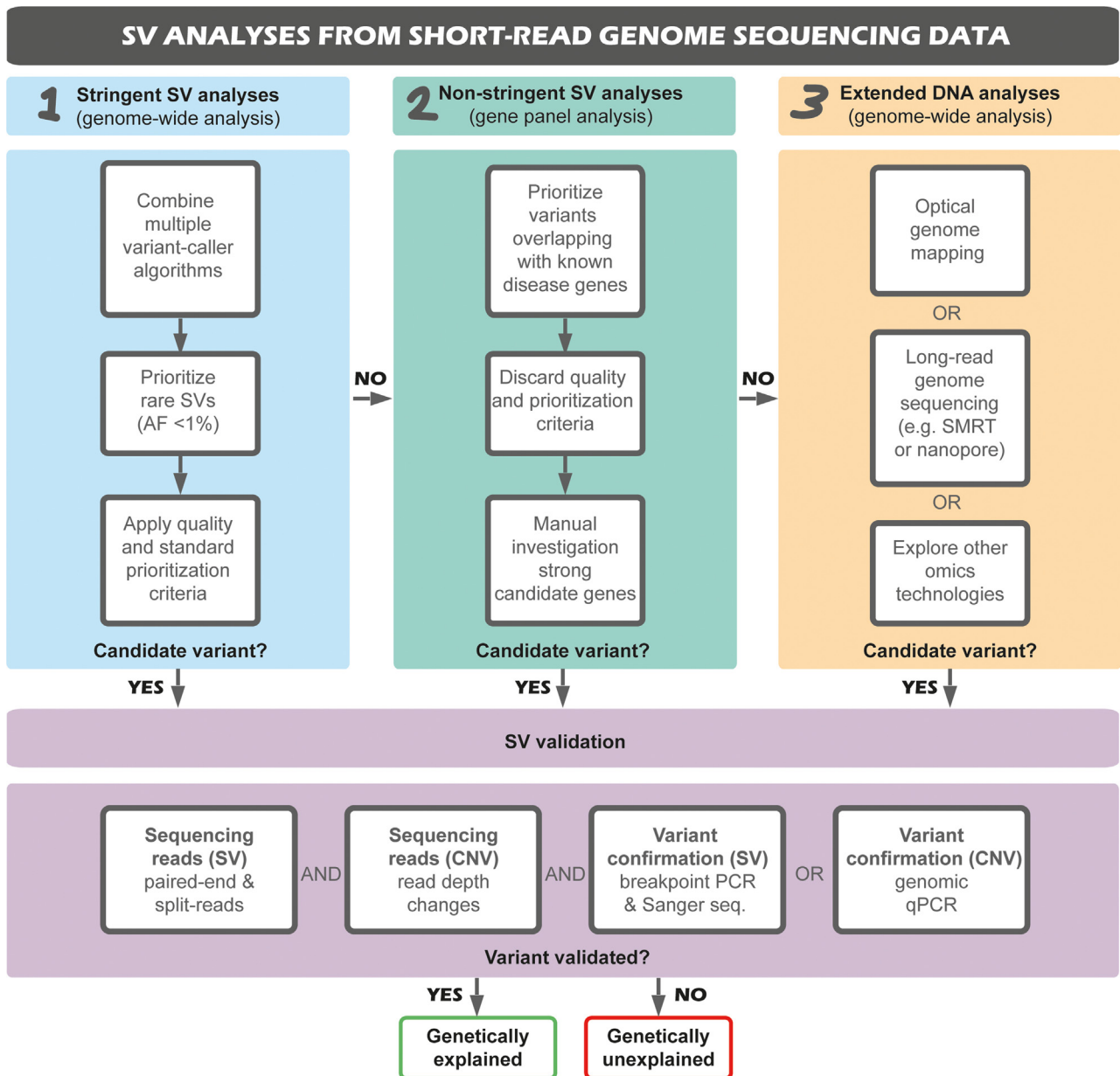


**Figure 3** Genome sequencing reanalysis reveals 30 (previously overlooked) structural variants. Schematic overview of all identified structural variants (SVs) in this study. SVs included 3 inversions, 3 duplications, and 24 deletions events. All SVs disrupted the protein-coding regions of an inherited retinal disease–associated gene and are classified as (likely) pathogenic. More detailed variant information can be found in [Supplemental Tables 2 and 3](#).

inversion event as well as a corresponding variant call. Split reads could be observed at the predicted inversion breakpoint junctions. This suggests that the *USH2A* inversion was overlooked and misinterpreted during previous analysis.

This finding prompted us to reanalyze our genome sequencing data sets that were collected over time and to perform a comprehensive reanalysis focused on the identification of pathogenic coding SVs overlapping with and disrupting an IRD-associated gene. This approach was very successful because 30 (likely) pathogenic SVs in 29 IRD probands were identified. Most remarkably, 8 of the identified pathogenic variants (>25%) were overlooked during initial analysis. We have critically evaluated our findings, and we found several explanations why these variants were not recognized previously. This allowed us to re-establish our SV prioritization protocols and analysis pipeline. Based on this, we would like to share the lessons that we learned from this reanalysis to facilitate and improve future analyses and interpretation of SV data sets ([Figure 4](#)).

First, it is generally accepted that SV calling from short-read sequencing data is not optimal yet and it is still difficult to distinguish true variant calls from false-positive calls. Especially large SVs (>2 Mb) are often considered as noise and deemed false positive and are frequently excluded from analyses to simplify the prioritization process.<sup>31</sup> However, in this study, 2 of the identified variants exceed this size threshold. Both of these variants were not identified in previously performed genome analysis because of this reason. The identified large inversion events encompass hundreds of genes but exhibit an apparent breakpoint in *USH2A* and were validated through breakpoint PCR. This demonstrates that it is warranted to critically evaluate breakpoints of all variants that are called, especially when an SV has a breakpoint in a disease gene of interest. In our updated SV annotation pipeline, SV breakpoint information has now been implemented for all variant calls. In addition, we have noticed that the usage of multiple variant databases (both global and local) is crucial to allow a better discrimination between true and false-positive calls. Especially, the



**Figure 4 Proposed flowchart for structural variant analysis from short-read genome sequencing data.** A recommended workflow for the identification of potential causal structural variants (SVs) from short-read genome sequencing data. Initially, a stringent structural variant analysis of patient samples should be performed. In this stringent analysis, standard quality and filtering criteria (eg, a minor allele frequency of <1%) should be applied to allow fast and efficient identification of causal variants and a genome-wide analysis should be performed. Potential candidate variants should be validated by analyzing the sequencing reads as well as confirmed through breakpoint PCR and Sanger sequencing. Duplications and deletions can also be confirmed through a genomic qPCR. An extended analysis of the data should be performed in case when no causal variants were identified during the first stringent analyses. Nonstringent variant prioritization can be performed, focusing on variants identified in disease-associated genes (gene panel analysis) minimizing the number of variants to be interrogated in detail. In addition, a manual inspection of strong candidate genes should be performed to identify indications for possible structural variations (eg, read-depth changes or split reads) in regions of interest. When nonstringent SV analysis does not yield any potential candidate variants, genome-wide follow-up studies based on other DNA technologies such as long-read genome sequencing or optical genome mapping should be considered. AF, allele frequency; CNV, copy number variant; PCR, polymerase chain reaction; qPCR, quantitative PCR; seq, sequencing; SMRT, single molecule real time; SV, structural variant.

implementation of an in-house SV database allowed us to exclude a significant number of false-positive calls from analysis that were introduced by technical errors in our sequencing pipeline.

Second, an important step in the SV calling pipeline is the quality assessment of a variant. Quality assessment is based on different aspects, such as the number of reads that are supporting the structural event or lack of paired



sequencing reads supporting the event. A variant is granted a quality score, which indicates whether a variant has failed or passed this assessment. Quality scores are considered easy tools for variant filtering and an efficient way to reduce potential candidate variants. Nevertheless, because SVs usually are complex variants resulting from recombination in homologous or repetitive regions, coverage of these events are generally low, which could consequently result in a low quality score. Therefore, low quality scores without the filter, pass, should also be considered during SV analyses. In our study, 6 of 30 identified SVs received a quality warning by at least 1 variant caller. Five of these variants were not picked up in earlier genome analysis and previously overlooked, suggesting that filtering on quality scores is an important issue and could possibly explain why pathogenic variants are still missed during analysis. All of the variants that received a quality warning could be validated using Sanger sequencing or genomic qPCR. An important eye-opener based on this result is that quality scoring should not be used for prioritization or exclusion for SVs but could be used as supportive evidence only.

Third, our results show that it is of utmost importance that multiple SV caller tools, based on different lines of evidence, are combined to improve data interpretation. SV detection can be based on either read-pair evidence (split reads and discordant read pairs; eg, Manta structural variant caller) or read-depth evidence (copy number changes; eg, Manta structural variant caller and Canvas Copy Number Variant Caller). Generally, read-pair evidence allows the detection of shorter (>50 bp) variants both balanced and unbalanced, whereas read-depth evidence allows the detection of longer (>1 kb) CNVs and can also be applied to identify terminal deletions or duplications. To allow efficient SV detection, it is crucial that both types of evidence are combined for SV calling. In this study, we identified 27 unbalanced rearrangements, ie, duplication and deletion events. Only 15 of 27 variants were identified by both SV caller (Manta structural variant caller<sup>15</sup>) and CNV caller (Canvas Copy Number Variant Caller<sup>16</sup>). Although the importance of combining SV callers has been recognized for some time, there are still reports available in which only a single SV caller is employed in SV pipelines. Moreover, SV caller algorithms are still continuously improving, and therefore SV detection pipelines should be frequently evaluated or updated. A striking example is variant 1 described in this study: a partial heterozygous deletion of *ARSG* identified in an individual diagnosed with Usher syndrome type-IV. This individual was previously described in a recent study by Velde et al,<sup>28</sup> in which the heterozygous *ARSG* deletion was identified only after visual inspection of the sequencing reads in IGV. In the study described by Velde et al,<sup>28</sup> CNV calling was performed using Control-FREEC<sup>32</sup> and the *ARSG* deletion was not called. In this study, SV and CNV analysis was performed using an updated pipeline, in which the Canvas Copy Number Variant caller<sup>16</sup> has been implemented for CNV calling. This time, the *ARSG* deletion was called and recognized by Canvas.

Finally, although our findings encourage the use of short-read sequencing for the identification of SVs, results and variant calls should always be treated with caution. Not all variant calls prioritized in our reanalysis pipeline could be observed in the raw sequencing reads or validated using either breakpoint PCR or genomic qPCR. For one of these variants, an inversion event disrupting the *EYS* gene, the presence of split reads could be clearly observed in the short-read sequencing data, whereas the variant could not be validated using long-read genome sequencing yet (data not shown). Therefore, variant validation remains a crucial step of the prioritization process.

With these data, we have shown that short-read sequencing data in terms of SV detection are more powerful than assumed, however interpretation of SVs remains challenging. Also, SVs affecting repetitive regions or regions of high complexity most likely still remain undetected. There is ample evidence of SVs that could be detected using long-read sequencing approaches only because they are based on de novo assembly and improved mapping of highly homologous regions.<sup>33,34</sup> It was shown that long-read sequencing detects about 2.5-fold more SVs than multiple short-read SV detection algorithms combined.<sup>11</sup> Nevertheless, the results of our study indicate that it is worthwhile to invest additional effort and time in SV analyses and interpretation using short-read data. First, a genome-wide stringent SV analysis should be performed, which allows the rapid identification of known pathogenic variants but also allows the identification of possibly pathogenic SVs in genes that have not been associated with disease before. After an initial stringent analysis is performed, we would like to advocate that a nonstringent SV analysis should be applied that prioritizes SVs that affect known candidate disease genes. In addition, manual inspection of sequencing reads overlapping with strong candidate genes (eg, in cases with a monoallelic pathogenic variant or a strong genotype–phenotype correlation) should be performed to identify hints of possible structural variation (eg, coverage changes or presence of split reads) in the region of interest. In this way, we hypothesize that previously hidden pathogenic SVs can still be exposed from existing data sets before engaging expensive long-read methods.

In conclusion, we identified likely pathogenic SVs in 29 of 427 (6.8%) probands (genetically explained and unexplained samples). The current contribution of SVs to the mutational landscape of IRDs is estimated to be 5% to 15%,<sup>5,7-9</sup> which is in line with this percentage. However, samples included in this study were prescreened using exome sequencing or targeted gene panel sequencing, which already included CNV analysis. This strongly suggests that the true contribution of SVs is higher than the currently anticipated percentages. In addition, >25% of the pathogenic SVs identified in this study were overlooked and/or misinterpreted during initial sequencing analysis. Through an initial discovery using OGM, we have

successfully explained a portion of missing heritability in our short-read genome sequencing cohort through re-establishing our protocols and pipelines. For the feasibility of this study, it was decided to focus on coding SVs only. It is clearly established that both SNVs and SVs can also have pathogenic consequences via noncoding mechanisms by affecting regulatory elements, topologically associated domain organization, splicing mechanisms, or untranslated region disruption.<sup>2,35,36</sup> It is likely that more pathogenic SVs are present in our data set, and follow-up analyses are warranted. This study highlights that SVs are an underestimated cause of IRDs and demand a sophisticated approach and more attention to facilitate detection during genome analyses.

## Data Availability

Data are available upon request. All pathogenic and likely pathogenic variants identified in this study have been submitted to the Leiden Open (source) Variation Database (LOVD) (<http://www.lovd.nl>). All other genome sequencing data are subject to controlled access because they may compromise the privacy of research participants. These data may become available upon a data transfer agreement approved by the local ethics committee and can be obtained after contacting the corresponding author (S.E.d.B.) upon request.

## Acknowledgments

The authors would like to thank Ellen Kater-Baats, Ronald van Beek, and Michiel Oorsprong for performing optical genome mapping and Brigitte Faas, Dominique Smeets, and Guillaume van de Zande for their help with karyotyping analysis. We thank the Department of Human Genetics and the Radboud Genome Technology Center for infrastructural and computational support. We would also like to thank Saskia van der Velde-Visser and Marlie Jacobs-Camps for DNA sample preparation and administration and Manar Salameh for technical assistance.

## Funding

This work has been funded by the European Union's Horizon 2020 Research and Innovation Programme under the EJP RD COFUND-EJP N° 825575 (to F.P.M.C. and S.R.). The work of K.R. and S.R. was funded by the Foundation Fighting Blindness (FFB)-career development award (CD-GE-0621-0809-RAD) (to S.R.). The work of A.H., L.E.L.M.V., and C.G. was funded by the Solve-RD project of the European Union's Horizon 2020 research and

innovation programme (No. 779257). The work of K.R. and L.W. was supported by grant awards from Fighting Blindness Ireland (FB Irl; FB16FAR, FB18CRE, FB20DOC) (to F.P.M.C., S.R., and G.J.F.), The Health Research Board of Ireland (HRB;POR/2010/97) (to G.J.F.) in conjunction with Health Research Charities Ireland (HRCI; MRCG-2013-8, MRCG-2016-14) (to G.J.F.), the Irish Research Council (IRC; GOIPG/2017/1631) (to G.J.F.), and Science Foundation Ireland (SFI; 16/1A/4452) (to G.J.F.). The work of J.R. was supported by the VELUX Stiftung (to H.K., F.P.M.C. and S.R.). Á.F.K. was supported by the Hungarian Scientific Research Fund OTKA PD\_21 138521 grant. This research was also supported by the Algemene Nederlandse Vereniging ter Voorkoming van Blindheid, Oogfonds, Landelijke Stichting voor Blinden en Slechtienden, Rotterdamse Stichting Blindenbelangen, Stichting Blindenhulp, Stichting tot Verbetering van het Lot der Blinden, and Stichting Blinden-Penning (to S.R. and F.P.M.C.).

## Author Information

Conceptualization: S.E.d.B., F.P.M.C., S.R.; Data Curation: J.C., M.R.N., L.E.L.M.V., C.G.; Formal Analysis: S.E.d.B., K.R., K.N., J.R., R.J.H.-M., L.H.-W., A.H., S.R.; Funding Acquisition: K.N., F.P.M.C., A.H., S.R.; Investigation: S.E.d.B., K.R., K.N., T.B.-Y., J.R., H.K., L.W., A.S.P., W.B., G.J.F., Á.F.K., I.F., N.W., M.E.W., D.S., R.J.E.P., L.H.-W., C.B.H.; Methodology: S.E.d.B., K.R., K.N., J.C., C.G., A.H.; Project Administration: K.N., F.P.M.C., A.H., S.R.; Supervision: H.K., G.J.F., R.J.E.P., M.R.N., L.E.L.M.V., C.G., F.P.M.C., A.H., S.R.; Validation: S.E.d.B., K.R.; Visualization: S.E.d.B., K.R.; Writing-original draft: S.E.d.B., S.R.; Writing-review and editing: S.E.d.B., K.R., K.N., J.C., T.B.-Y., J.R., H.K., L.W., A.S.P., W.B., G.J.F., Á.F.K., R.J.H.-M., I.F., N.W., M.E.W., D.S., R.J.E.P., L.H.-W., C.B.H., M.R.N., L.E.L.M.V., L.I.v.d.B., C.G., F.P.M.C., A.H., S.R.

## Ethics Declaration

The study adhered to the tenets of the Declaration of Helsinki and was approved by the local ethics committees of the Radboud University Medical Center (Nijmegen, The Netherlands); the Rotterdam Eye Hospital (Rotterdam, The Netherlands); Amsterdam UMC (Amsterdam, The Netherlands) (NL34152.078.10), (MEC-2010-359; OZR protocol no. 2009-32); Department of Ophthalmology, The Royal Victoria Eye and Ear Hospital (Dublin, Ireland) (13-06-2011: HRA-POR201097); Ramabam Health Care Campus (Haifa, Israel); and the University Hospital of Tübingen (349/2003V and 116/2015B02). Written informed consent was obtained from patients before DNA analysis and inclusion in this study.


## Conflict of Interest

The authors declare no conflicts of interest.

## Additional Information

The online version of this article <https://doi.org/10.1016/j.gim.2022.11.013> contains supplementary material, which is available to authorized users.

## Authors

Suzanne E. de Bruijn<sup>1,2,\*</sup> , Kim Rodenburg<sup>1,2</sup>, Jordi Corominas<sup>1</sup>, Tamar Ben-Yosef<sup>3</sup>, Janine Reurink<sup>1,2</sup>, Hannie Kremer<sup>1,2,4</sup>, Laura Whelan<sup>5</sup>, Astrid S. Plomp<sup>6</sup>, Wolfgang Berger<sup>7,8,9</sup>, G. Jane Farrar<sup>5</sup>, Árpád Ferenc Kovács<sup>10</sup>, Isabelle Fajardy<sup>11,12</sup>, Rebekkah J. Hitti-Malin<sup>1,2</sup>, Nicole Weisschuh<sup>13</sup>, Marianna E. Weener<sup>14</sup>, Dror Sharon<sup>15</sup>, Ronald J.E. Pennings<sup>2,4</sup>, Lonke Haer-Wigman<sup>1</sup>, Carel B. Hoyng<sup>2,16</sup>, Marcel R. Nelen<sup>1</sup>, Lisenka E.L.M. Vissers<sup>1,2</sup>, L. Ingeborgh van den Born<sup>17</sup>, Christian Gilissen<sup>1,18</sup>, Frans P.M. Cremers<sup>1,2</sup>, Alexander Hoischen<sup>1,18,19</sup>, Kornelia Neveling<sup>1</sup>, Susanne Roosing<sup>1,2</sup>

## Affiliations

<sup>1</sup>Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; <sup>2</sup>Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands; <sup>3</sup>Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel; <sup>4</sup>Hearing and Genes, Department of Otorhinolaryngology, Radboud University Medical Center, Nijmegen, The Netherlands; <sup>5</sup>The School of Genetics and Microbiology, Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland; <sup>6</sup>Department of Human Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; <sup>7</sup>Institute of Medical Molecular Genetics, University of Zurich, Schlieren, Switzerland; <sup>8</sup>Neuroscience Center Zurich (ZNZ), University and ETH Zurich, Zurich, Switzerland; <sup>9</sup>Zurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland; <sup>10</sup>2nd Department of Paediatrics, Faculty of Medicine, Semmelweis University, Budapest, Hungary; <sup>11</sup>Division of Maternal Malnutrition, Department of Perinatal Environment and Health, Lille University, Lille, France; <sup>12</sup>Division Biochemistry and Molecular Biology, Biology and Pathology Center, Lille, France; <sup>13</sup>Center for Ophthalmology, Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany; <sup>14</sup>Clinical Research Center, Ophthalmic, CRO,

Moscow, Russia; <sup>15</sup>Division of Ophthalmology, Hadassah University Medical Center, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel; <sup>16</sup>Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands; <sup>17</sup>The Rotterdam Ophthalmic Institute, The Rotterdam Eye Hospital, Rotterdam, The Netherlands; <sup>18</sup>Radboud Institute of Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; <sup>19</sup>Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands

## References

- Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161(5):1012-1025. <http://doi.org/10.1016/j.cell.2015.04.004>
- de Bruijn SE, Fiorentino A, Ottaviani D, et al. Structural variants create new topological-associated domains and ectopic retinal enhancer-gene contact in dominant retinitis pigmentosa. *Am J Hum Genet*. 2020;107(5):802-814. <http://doi.org/10.1016/j.ajhg.2020.09.002>
- Shearer AE, Kolbe DL, Azaiez H, et al. Copy number variants are a common cause of non-syndromic hearing loss. *Genome Med*. 2014;6(5):37. <http://doi.org/10.1186/gm554>
- National Library of Medicine. dbVar. National Center for Biotechnology Information. Accessed June 1, 2022. <https://www.ncbi.nlm.nih.gov/dbvar/>
- Carss KJ, Arno G, Erwood M, et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *Am J Hum Genet*. 2017;100(1):75-90. <http://doi.org/10.1016/j.ajhg.2016.12.003>
- Fadaie Z, Whelan L, Ben-Yosef T, et al. Whole genome sequencing and in vitro splice assays reveal genetic causes for inherited retinal diseases. *NPJ Genom Med*. 2021;6(1):97. <http://doi.org/10.1038/s41525-021-00261-1>
- Biswas P, Villanueva AL, Soto-Hermida A, et al. Deciphering the genetic architecture and ethnographic distribution of IRD in three ethnic populations by whole genome sequence analysis. *PLoS Genet*. 2021;17(10):e1009848. <http://doi.org/10.1371/journal.pgen.1009848>
- Zampaglione E, Kinde B, Place EM, et al. Copy-number variation contributes 9% of pathogenicity in the inherited retinal degenerations. *Genet Med*. 2020;22(6):1079-1087. <http://doi.org/10.1038/s41436-020-0759-8>
- Van Schil K, Naessens S, Van de Sompele S, et al. Mapping the genomic landscape of inherited retinal disease genes prioritizes genes prone to coding and noncoding copy-number variations. *Genet Med*. 2018;20(2):202-213. Published correction appears in *Genet Med*. 2019;21(8):1998. <http://doi.org/10.1038/gim.2017.97>
- Haer-Wigman L, van Zelst-Stams WA, Pfundt R, et al. Diagnostic exome sequencing in 266 Dutch patients with visual impairment. *Eur J Hum Genet*. 2017;25(5):591-599. <http://doi.org/10.1038/ejhg.2017.9>
- Chaisson MJP, Sanders AD, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019;10(1):1784. <http://doi.org/10.1038/s41467-018-08148-z>
- Mantere T, Neveling K, Pebrel-Richard C, et al. Optical genome mapping enables constitutional chromosomal aberration detection. *Am J Hum Genet*. 2021;108(8):1409-1422. <http://doi.org/10.1016/j.ajhg.2021.05.012>
- Neveling K, Mantere T, Vermeulen S, et al. Next-generation cytogenetics: comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. *Am J Hum Genet*. 2021;108(8):1423-1435. <http://doi.org/10.1016/j.ajhg.2021.06.001>

14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. <http://doi.org/10.1093/bioinformatics/btp324>
15. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222. <http://doi.org/10.1093/bioinformatics/btv710>
16. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*. 2016;32(15):2375-2377. <http://doi.org/10.1093/bioinformatics/btw163>
17. Fadaie Z, Neveling K, Mantere T, et al. Long-read technologies identify a hidden inverted duplication in a family with choroideremia. *HGG Adv*. 2021;2(4):100046. <http://doi.org/10.1016/j.xhgg.2021.100046>
18. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995. <http://doi.org/10.1093/nar/gkab1049>
19. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. <http://doi.org/10.1038/nature15393>
20. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet*. 2009;84(4):524-533. <http://doi.org/10.1016/j.ajhg.2009.03.010>
21. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. Published correction appears in *Nature*. 2021;590(7846):E53. Published correction appears in *Nature*. 2021;597(7874):E3-E4. <http://doi.org/10.1038/s41586-020-2308-7>
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. <http://doi.org/10.1038/ng.2892>
23. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc*. 2016;11(1):1-9. <http://doi.org/10.1038/nprot.2015.123>
24. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249. <http://doi.org/10.1038/nmeth0410-248>
25. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11(4):361-362. <http://doi.org/10.1038/nmeth.2890>
26. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. *Cell*. 2019;176(3):535-548.e24. <http://doi.org/10.1016/j.cell.2018.12.015>
27. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26. <http://doi.org/10.1038/nbt.1754>
28. Velde HM, Reurink J, Held S, et al. Usher syndrome type IV: clinically and molecularly confirmed by novel ARSG variants. *Hum Genet*. 2022;141(11):1723-1738. <http://doi.org/10.1007/s00439-022-02441-0>
29. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. <http://doi.org/10.1038/gim.2015.30>
30. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat*. 2018;39(11):1517-1524. <http://doi.org/10.1002/humu.23626>
31. Wu Z, Jiang Z, Li T, et al. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun*. 2021;12(1):6501. <http://doi.org/10.1038/s41467-021-26856-x>
32. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423-425. <http://doi.org/10.1093/bioinformatics/btr670>
33. Bedoni N, Quinodoz M, Pinelli M, et al. An Alu-mediated duplication in NMNAT1, involved in NAD biosynthesis, causes a novel syndrome, SHILCA, affecting multiple tissues and organs. *Hum Mol Genet*. 2020;29(13):2250-2260. <http://doi.org/10.1093/hmg/ddaa112>
34. Reiner J, Pisani L, Qiao W, et al. Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a Bardet-Biedl syndrome 9 (BBS9) deletion. *NPJ Genom Med*. 2018;3:3. <http://doi.org/10.1038/s41525-017-0042-3>
35. Ellingford JM, Ahn JW, Bagnall RD, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med*. 2022;14(1):73. <http://doi.org/10.1186/s13073-022-01073-3>
36. Van de Sompele S, Small KW, Cicek MB, et al. Multi-omics approach dissects cis-regulatory mechanisms underlying North Carolina macular dystrophy, a retinal enhanceropathy. *Am J Hum Genet*. 2022;109(11):2029-2048. <http://doi.org/10.1016/j.ajhg.2022.09.013>