HGG Advances

ARTICLE

Long-read technologies identify a hidden inverted duplication in a family with choroideremia

Zeinab Fadaie,[1,2,7] Kornelia Neveling,[1,3,7] Tuomo Mantere,[1,4] Ronny Derks,[1] Lonneke Haer-Wigman,[1] Amber den Ouden,[1] Michael Kwint,[1,2] Luke O'Gorman,[1] Dyon Valkenburg,[2,5] Carel B. Hoyng,[2,5] Christian Gilissen,[1,4] Lisenka E.L.M. Vissers,[1,2] Marcel Nelen,[1] Frans P.M. Cremers,[1,2] Alexander Hoischen,[1,4,6,7,*] and Susanne Roosing[1,2,7,*]

Summary

The lack of molecular diagnoses in rare genetic diseases can be explained by limitations of current standard genomic technologies. Upcoming long-read techniques have complementary strengths to overcome these limitations, with a particular strength in identifying structural variants. By using optical genome mapping and long-read sequencing, we aimed to identify the pathogenic variant in a large family with X-linked choroideremia. In this family, aberrant splicing of exon 12 of the choroideremia gene *CHM* was detected in 2003, but the underlying genomic defect remained elusive. Optical genome mapping and long-read sequencing approaches now revealed an intragenic 1,752 bp inverted duplication including exon 12 and surrounding regions, located downstream of the wild-type copy of exon 12. Both breakpoint junctions were confirmed with Sanger sequencing and segregate with the X-linked inheritance in the family. The breakpoint junctions displayed sequence microhomology suggestive for an erroneous replication mechanism as the origin of the structural variant. The inverted duplication is predicted to result in a hairpin formation of the pre-mRNA with the wild-type exon 12, leading to exon skipping in the mature mRNA. The identified inverted duplication is deemed the hidden pathogenic cause of disease in this family. Our study shows that optical genome mapping and long-read sequencing have significant potential for the identification of (hidden) structural variants in rare genetic diseases.

Introduction

Choroideremia (CHM, OMIM: 303100) is a progressive, rare, X-linked form of chorioretinal degeneration with an estimated incidence of approximately 1:50,000 to 1:100,000 worldwide.[1–3] CHM affects the choroid and rod photoreceptors in the retina, leading to night blindness and impaired visual acuity in childhood, and subsequently leads to tunnel vision in the second and third decades of life, finally resulting in legal blindness.[4,5] Female carriers generally do not manifest significant visual impairment. Several cases are reported to develop severe visual impairment during adolescence due to skewed X-inactivation,[4,6–8] while the correlation between skewed X-inactivation and the clinical outcome is also debated.[7,9]

CHM is almost exclusively caused by nonsense variants, splice site variants, and pathogenic deletions and insertions in the *CHM* gene (NM_00390.3; OMIM: 300390), which spans 186 kb on chromosome Xq21.2 and encompasses 15 coding exons.[10,11] Due to the strong genotype-phenotype correlation for CHM, a genetic diagnosis can be made in approximately 94% of clinically diagnosed CHM individuals.[12,13] However, in one large Dutch CHM family, family A, described in 2003 by van den Hurk et al.,[14] the underlying genetic cause of disease could not be determined (Figure 1). RNA analysis in affected males showed an aberrant transcript lacking exon 12 (r.1414_1510del; p.Ser473Trpfs*4), supporting the clinical diagnosis of CHM (Figures S1A–S1C). The causative pathogenic variant on the DNA level, however, could not be identified.[14]

Nowadays, disease-causing variants are detected using short-read next-generation sequencing (NGS) techniques by analyzing gene panels, whole-exome sequencing (WES), or whole-genome sequencing (WGS).[15,16] These techniques are primarily applied for their high-throughput nature, the low per-base error rate, and their cost effectiveness compared to previous single-gene approaches.[17,18] However, short reads are inadequate when it comes to accurate mapping of highly repetitive regions, GC-rich regions, sequences with multiple homologous elements, and detection of structural variants (SVs).[19,20] Therefore, certain genetic conditions caused by rearrangements, large repeats, or balanced SVs, such as inversions and translocations, often remain hidden or not fully resolved when using short-read approaches.

Long-read sequencing technologies have been rapidly developing and seem to overcome the limitations of short

[1]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; [2]Donders Institute for Brain, Cognition, and Behavior, Radboud University Medical Center, Nijmegen, the Netherlands; [3]Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; [4]Radboud Institute of Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; [5]Department of Ophthalmology, Radboud University Medical Center, Nijmegen, the Netherlands; [6]Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, the Netherlands
[7]These authors contributed equally
*Correspondence: susanne.roosing@radboudumc.nl (S.R.), alexander.hoischen@radboudumc.nl (A.H.)
https://doi.org/10.1016/j.xhgg.2021.100046.

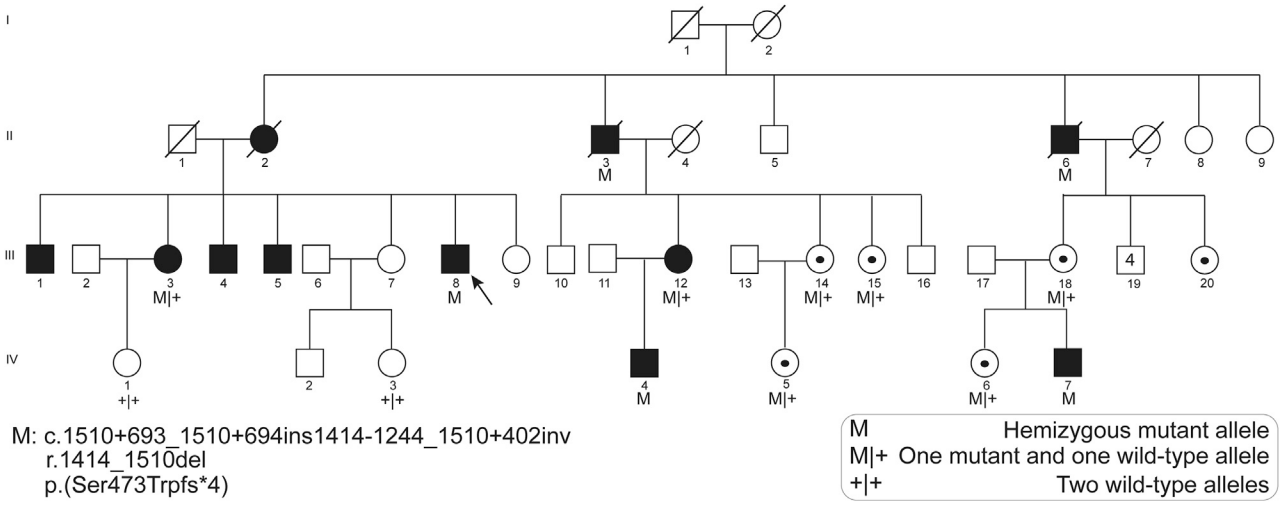

Figure 1. The pedigree of family A affected with choroideremia
The identified inverted duplication segregates with the disease in family A. The affected female individuals manifest the phenotype as well. DNA material of the affected male individual indicated by the arrow was utilized for optical mapping and long-read sequencing analysis.

reads in genetic research and molecular diagnoses of different human genetic diseases.[21–24] Their read length of several kilobases (1) simplifies the identification of SVs,[25–28] (2) simplifies the spanning of repeats and high GC-rich regions,[29,30] and (3) enables variant phasing.[24,31,32] Bionano optical genome mapping is a high-resolution cytogenetic technique and is based on ultra-high molecular weight (UHMW) DNA molecules that are fluorescently labeled at a 6-mer motif (CTTAAG).[33] Optical genome mapping can detect SVs as small as 500 bp, which is an approximately 10,000 times higher resolution compared to standard karyotyping, and therefore enables much more precise data analysis.[34]

The aim of the current study was to identify the genomic aberration in *CHM* leading to exon 12 skipping in the described CHM family.[14] To identify the underlying DNA defect, we first used conventional Sanger sequencing in order to identify potential pathogenic deep-intronic variants, preceding a combination of optical genome mapping and long-read sequencing.

Material and methods

The study adhered to the tenets of the Declaration of Helsinki and was approved by the local ethics committees of Radboud University Medical Center, Nijmegen, the Netherlands. Written informed consent was obtained from participants before inclusion to this study.

Bionano optical genome mapping

Bionano optical genome mapping was performed as described previously.[35,36] In brief, DNA was isolated from a lymphoblastoid cell line obtained from an affected male from family A (III-8; Figure 1), according to the manufacturer's instructions using the SP Blood & Cell Culture DNA Isolation Kit, (Bionano Genomics, San Diego, CA, USA). The isolated UHMW DNA was labeled for the CTTAAG sequence, using the DLS (Direct Label and Stain) DNA Labeling Kit (Bionano Genomics, San Diego, CA, USA), and was analyzed using a 3 × 1,300 Gb Saphyr chip (G2.3) on a Bionano Saphyr instrument, reaching 177× effective coverage with a label density of 14.37/100 kb and an average N50 of 232 kb. *De novo* assembly (using GRCh37) and variant annotation was performed using Bionano Solve version 3.4, which includes two different algorithms for SV (based on assembled maps) and copy number variant (CNV) (based on molecule coverage) calling. Annotated variants were filtered for rare events as described previously.[36] In addition, the region of interest around exon 12 of *CHM* was analyzed visually in Bionano Access version 1.4.3.

PacBio long-read sequencing

Long-read genome sequencing was performed using the SMRT sequencing technology (Pacific Biosciences, Menlo Park, CA, USA), using DNA isolated according to standard procedure.[37] In brief, library preparation was performed according to the manufacturer's instructions using the Procedure & Checklist – Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences). Size selection was performed using a BluePippin system (target fragments ± 15–18 kb). Sequence primer V2 and Polymerase 2.0 were used for binding. The SMRTbell complex was loaded onto an 8M SMRTcell and sequenced on a Sequel II instrument (Pacific Biosciences, Menlo Park, CA, USA), according to the manufacturer's instructions. Following sequencing, CCS (also called HiFi) reads were generated from the sequencing raw reads using SMRTLink 8.0.0 and mapped against the human genome (GRCh37). The region of interest, chrX:85,116,185–85,302,566 (based on NC_000023.10) was manually inspected in integrative genomics viewer (IGV) version 2.4. In addition, SVs were called using pbsv v.2.2.2 (SMRTLink v.8.0.0), and annotation was performed using an in-house SV pipeline[28] using public databases including Decipher,[38] Wellderly,[39] Genome of the Netherlands (GoNL),[40] 1000 Genomes Project,[41,42] Exome Aggregation Consortium (ExAC),[43] and Database of Genomic Variants (dgv.tcag.ca).[44] Moreover, GnomAD was assessed manually for SVs occurring in the region of interest.[45]

Sanger validation and breakpoint assessment and *in silico* interpretation

DNA material of available individuals from family A, along with DNA of an unrelated healthy female and male individual, were amplified and Sanger sequenced to validate the breakpoint junctions. Primer sequences and coordinates are listed in Table S1. Subsequently, to investigate the putative mechanism that mediated the SV to occur, the breakpoint regions were assessed using the Cluster Omega tool[46] for the presence of microhomology or repetitive elements as described previously.[47] Furthermore, the secondary structure between reference and mutant sequence was assessed for alterations underlying the exon 12 skipping using *in silico* tool RNAstructure version 6.0.1.[48] Due to size constraints of the predictive tool (<3 kb input), analyses were carried out by using a smaller region of wild-type CHM and by including both boundaries of the inverted duplication individually. In a first analysis, r.1414−1400 to r.1510+1510 of the wild type was used, where at r.1510+693 the first 200 bp of the 5′ side of the inverted duplication were inserted. In a second analysis, the 200 bp from the 3′ side were included at r.1510+693 in a total region from r.1414−1400 to r.1510+1510 of the wild type.

Results

Previous targeted RNA sequencing of family A revealed skipping of *CHM* exon 12[14] (Figures S1A–S1C); nevertheless, the underlying cause on the gDNA level could not be identified.[14] In the current study, we first screened DNA sequences 5 kb up- and downstream of *CHM* exon 12 by PCR and Sanger sequencing in affected cases from family A to identify putative non-coding pathogenic variants. Based on *in silico* prediction tools integrated in Alamut Visual software version 2.13 (Interactive Biosoftware, Rouen, France), no rare splice-altering variants with predicted pathogenic splice defects were determined.

Subsequently, optical genome mapping was carried out on genomic DNA of an affected male (III-8) of family A to investigate potential SVs that could have been missed by conventional Sanger sequencing. Optical genome mapping detected a total of 5,778 SVs, of which 29 were rare SVs, meaning that they were not identified in a control database comprising 107 samples (provided by Bionano Genomics).[33] Of these rare SVs, 15 (7 deletions, 6 insertions, and 2 intra-chromosomal translocations) were overlapping with genes, defined by Bionano Genomics as within a distance of 12 kb of a gene (Table S2). In addition, optical genome mapping called a total of 136 CNVs, which could be reduced to three rare CNVs when using filtering steps as described in Table S2.

We checked both the detected rare SVs and CNVs (after filtering) for variants overlapping with the suspected *CHM* locus and identified two SV calls at this locus. On visual inspection, it was determined that both variant calls identified the same insertion but with one label difference in two separately called sample maps. This insertion was called as 1,573 bp and 1,549 bp in length, respectively, within a genomic region of 15.9 kb in between label positions g.85,134,124 and g.85,150,032 (NC_000023.10) within the *CHM* gene (Figure 2A; Figure S1D).

To understand the origin of the inserted material, we aimed to perform long-range PCR followed by long-read sequencing. However, due to unsuccessful targeted amplification of the region of interest, suggesting a potential more complex event than anticipated, we performed long-read WGS on DNA of individual III-8 to identify the origin of the inserted material. We obtained an 8-fold average genome coverage with four HiFi reads spanning the X chromosome *CHM* locus in this male individual. SV analysis of these long-read data revealed an insertion of 1,752 bp downstream of *CHM* exon 12, between positions c.1510+693 and c.1510+694 (Figure 2B). According to the long-read WGS data, the insertion consisted of an inverted duplication of exon 12 whose 5′ breakpoint was located in intron 12 and 3′ breakpoint in intron 11 (c.1510+693_1510+694ins1414−1244_1510+402inv).

To validate the intragenic inverted duplication and its 5′ and 3′ breakpoints at the single-nucleotide level, PCR amplification was performed, and subsequent segregation analysis was carried out in seven additional family members. The PCR amplification of the 5′ and 3′ breakpoints of the inverted duplication confirmed the expected fragment for affected males and carrier females (Figures S2A and S2B). Moreover, a PCR amplification designed to span the 1,752 bp inverted duplication showed a larger fragment indicating the presence of the inverted duplication in the mutated hemizygous allele of the affected males (Figure S3). A wild-type fragment without the inverted duplication was observed for non-carrier females and unrelated individuals. Sanger sequence analysis confirmed that after a wild-type copy of exon 12 in the mutant allele, intron 12 was interrupted by an inverted duplication containing an additional copy of exon 12 (c.1510+402 to c.1414−1244 plus four additional nucleotides), inserted between c.1510+693 and c.1510+694 (Figure S2C).

To understand the possible mechanism leading to this inverted duplication, we assessed both breakpoints for the presence of microhomologies or repetitive elements. We observed an 8 bp microhomology region (CACAATTC) at positions c.1510+693 and c.1510+402, and a 4 bp (TGTG) microhomology region at positions c.1414−1244 and c.1510+703, respectively (Figure 3). Therefore, the origin of the SV may likely be explained by microhomology regions present at the breakpoints, as these may mediate SVs and resemble the previously suggested fork stalling and template switching/microhomology-mediated break-induced replication (FoSTeS/MMBIR) mechanism[49] (Figure 4).

Subsequently, we speculated that the inverted duplication may lead to skipping of exon 12 by disruption of the mRNA secondary structure. To examine our hypothesis, we assessed the differences of the RNA structure between the reference and the mutant sequences by including 200 bp from both the 5′ and 3′ end of the breakpoints of the inverted duplication using an *in silico* RNA structure tool. Combining these two predictions, the assessment predicted that the aberrant sequence nucleotides from position

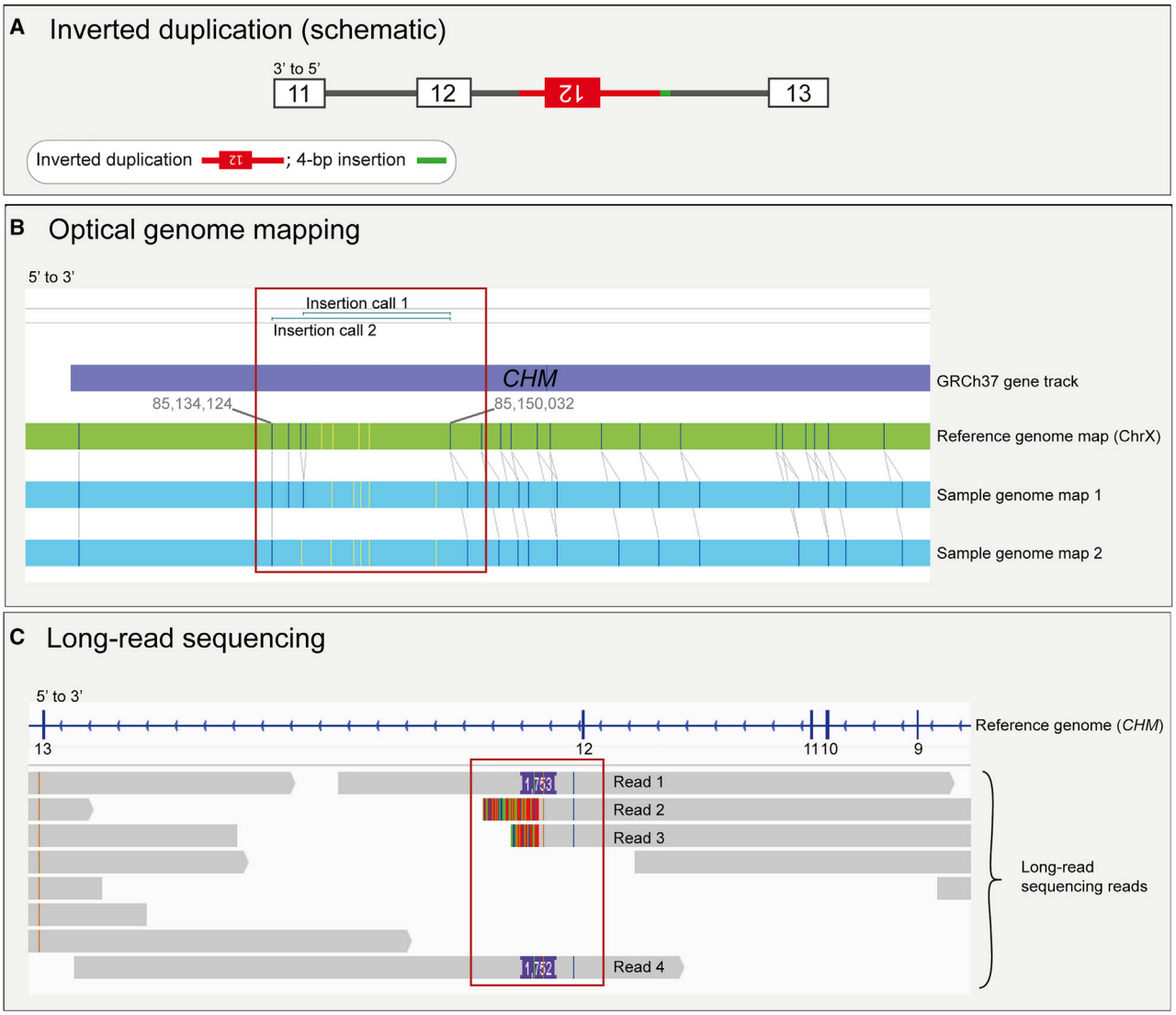

Figure 2. Identification of the intragenic inverted duplication through optical mapping and long-read sequencing
(A) Schematic representation of the inverted duplication in the *CHM* locus that has been identified in family A. (B) The result of optical genome mapping revealed an insertion of 1,573 and 1,549 bp within a 15.9 kb region upstream of *CHM* exon 12 in the affected individual compared to the reference genome. The green bar demonstrates the genome map of the reference genome. The blue bars show the genome maps of the affected individual; these two maps are only distinguished by one label difference. Both structural variant calls are shown on top, both calling an insertion within the region of interest. (C) By using long-read sequencing, the insertion first seen by optical mapping was identified as an intragenic inverted duplication. Two out of four reads covering this region span the inverted duplication completely (reads 1 and 4), whereas the two other reads (reads 2 and 3) do not span the entire event. *CHM* is located on the minus strand (3′ to 5′); however, the results shown in this figure are provided for the plus strand.

c.1414−1243 to c.1510+410 are able to generate a hairpin with the inverted duplication, confirming our hypothesis (Figure 5).

Discussion

Since the introduction of NGS, diagnostic yields for rare genetic diseases have significantly increased.[15,16] However, the diagnostic success rate for short-read WES or WGS is still limited to 30%–70%.[16,50,51] We speculate that the remaining genetic defects in unresolved cases can be partially explained by hidden SVs that could not be detected by short-read sequencing technologies, rather than purely by thus-far unidentified disease-associated genes.[28]

In the current study, we aimed to unravel a genetic mystery in family A, for which the disease locus and the resulting consequence at the RNA level was known for >15 years,[14] but the disease-causing and molecularly proven pathogenic variant remained undetected thus far. The possibilities considered were splice-altering pathogenic variants that may affect the splicing of *CHM* exon 12 or thus-far hidden structural variants that lead to a splice defect on the RNA level. Here, we provide evidence that the latter was the case.

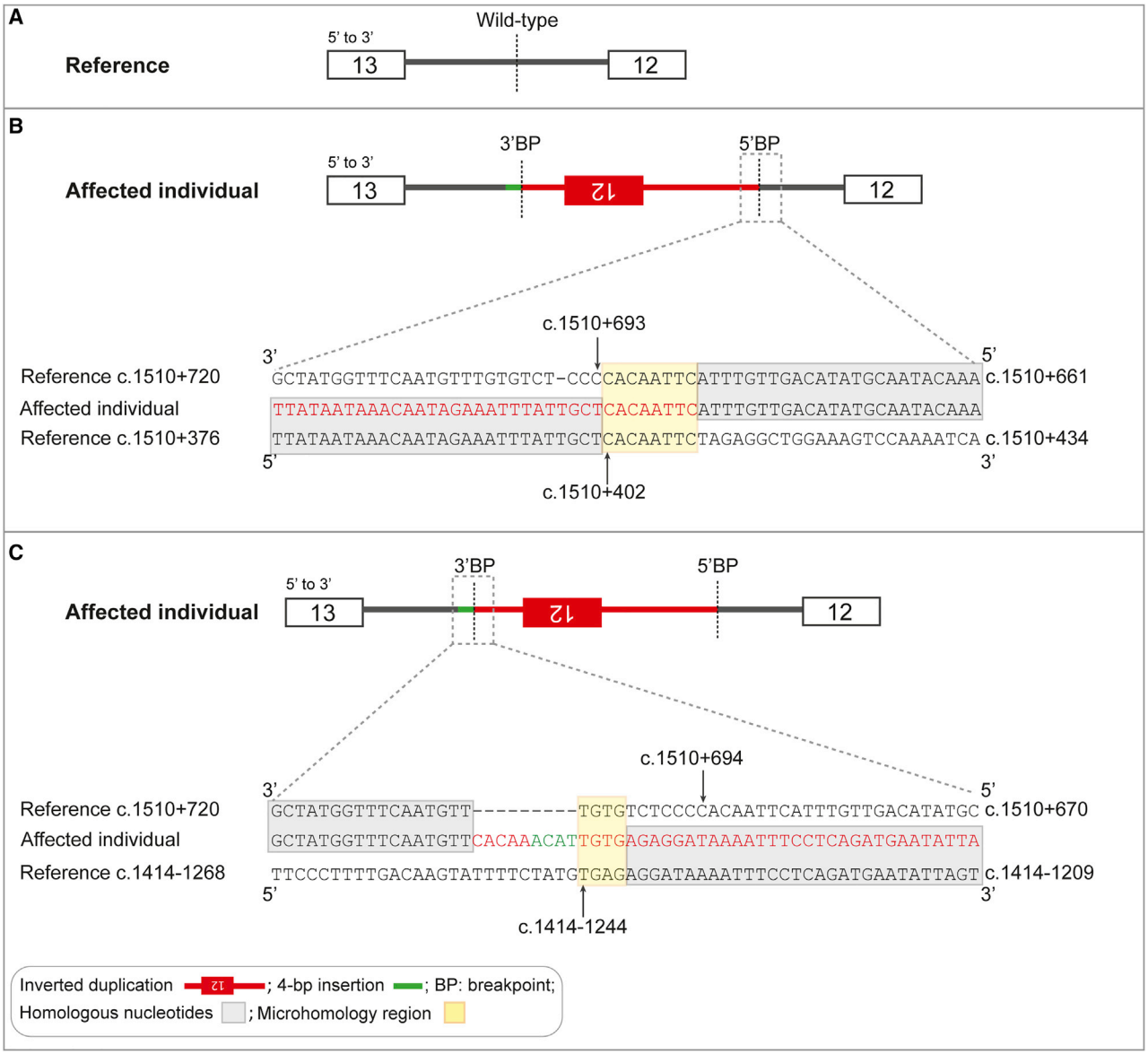

Figure 3. Assessment of microhomology at 5′ and 3′ breakpoints

(A) A schematic representation of the genomic region of exon 13 to exon 12 *CHM* (5′→3′) (B and C) The 5′ and 3′ breakpoint regions of the inverted duplication event were assessed for the presence of microhomology using multiple sequence alignment of the Cluster Omega tool. (B) Analysis of the reference fragment spanning the insertion site c.1510+693 and c.1510+694 (upper sequence) and the reference sequence spanning c.1510+402 (BP-5′, lower sequence) region showed a microhomology region of 8 nucleotides. (C) Analysis of the reference fragment spanning the insertion site c.1510+693 and c.1510+694, (upper sequence) and the reference sequence spanning c.1414−1244 (BP-3′, lower sequence) showed microhomology of 4 nucleotides. 60 bp reference sequences spanning each position were used as input. The start and end positions of the assessed sequences are provided. The reference sequence is indicated in black; the observed sequence as in family A is marked in red and green. Homology between the reference and observed sequence is shown with a vertical black line, and the regions of microhomology are highlighted in the yellow boxes.

Due to the strong genotype-phenotype association in choroideremia and the already known splice aberration,[14] Sanger sequencing of 5 kb surrounding exon 12 rather than WES or WGS was performed in the studied family. However, since no putative pathogenic variant was identified by Sanger sequencing, next we assumed that the aberrant splicing may occur due to a complex genomic structural variant instead. Therefore, we utilized optical genome mapping and long-read WGS to fully unravel the underlying event. Using these approaches, we identified a 1,752 bp inverted duplication downstream of *CHM* exon 12 as the thus-far hidden SV in family A. Although the inverted duplication was within the prescreened region using Sanger sequencing efforts in both males and females from family A, the event was still not detected previously, likely due to overlapping content of the sequence of the SV and the wild-type sequence. A targeted long-read amplicon sequencing would have been sufficient to confirm the SV detected by optical genome mapping. However, this effort failed due to a >18 kb predicted

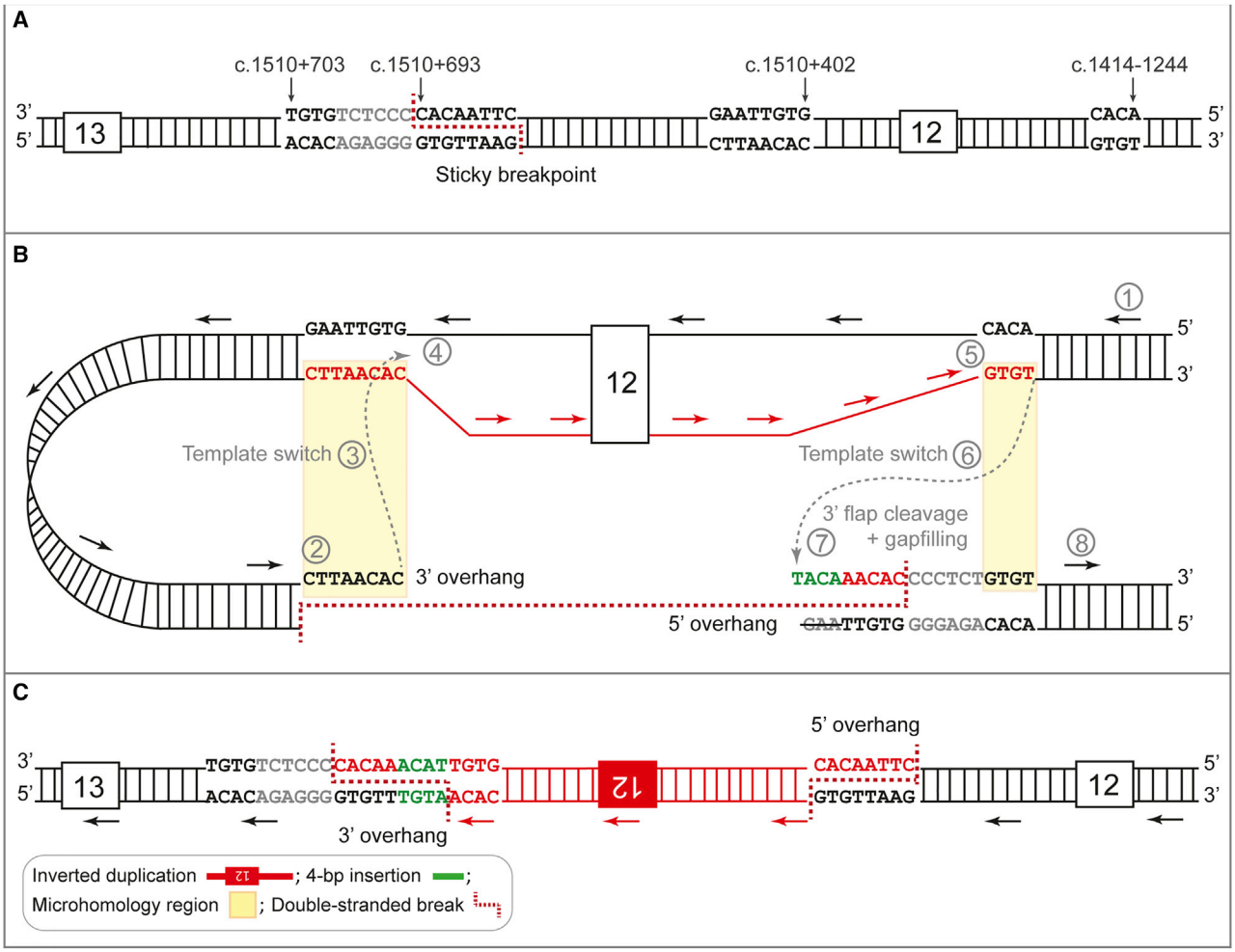

Figure 4. A proposed FoSTeS/MMBIR mechanism underlying the origin of inverted duplication
(A) A schematic representation of the genomic region of intron 11 until intron 13 of *CHM* is shown as present on the reverse strand. The relevant nucleotides for the proposed model are depicted. A red dotted line represents the location of the sticky end break. (B) The proposed mechanism of the SV is illustrated, i.e., (1) the DNA polymerase synthesized the DNA from 5′ to 3′, (2) the polymerase stalls due to the sticky end break at position c.1510+693, and (3) template switching to the forward strand of *CHM* (indicated in red) occurs due to the presence of 8 bp microhomology. (4) The polymerase continues DNA replication of the strand and thereby generates the inverted duplication containing a second copy of exon 12. (5) A 4 bp microhomology region at position c.1414−1244 in the forward strand stalls the DNA replication, and (6) template switching occurs to the reverse strand. (7) The DNA mismatch repair mechanism completes the 3′ sticky overhang by 3′ flap cleavage and fill in synthesis leading to a 4 bp random nucleotide insertion. From there, (8) DNA replication continues in the original strand. (C) The resulting *CHM* allele, specific for family A, containing the inverted duplication is shown, occurring through the FoSTeS/MMBIR mechanism.

amplicon size encompassing a 1.5 kb event within a 15.9 kb region. We also cannot exclude that short-read WGS or WES would have been able to detect the copy number gain, but it is unlikely that is would have unraveled the exact duplicated inversion. However, generally coverage-depth-based CNV algorithms for WES data are less sensitive to copy number gains than copy number losses, and single-exon CNVs remain challenging for multiple algorithms, of which several only detect copy number events of two exons or larger.[52,53] Short-read WGS could possibly detect the copy number gain in case that sufficient coverage was achieved for the locus; however, it remains speculative whether the exact inverted duplication would have been identified, or whether WGS would require additional analyses and PCR validations to confirm the exact nature of the SV. The current approach not only shows that optical genome mapping and long-read WGS confirm identified SVs orthogonally but also showcases the complete unraveling of SV details by long-read WGS compared to copy number inferences from coverage-based NGS approaches.

Studying the breakpoints of SVs at single-nucleotide resolution is fundamental to deduce the mutational mechanisms underlying the SV origin.[54] We postulate that the SV in family A has originated through a FoSTeS/MMBIR mechanism (Figure 4). The FosTes/MMBIR mechanism was first described by Zhang et al.[55] and suggested to contribute to SV rearrangements in the human genome on a diverse scale, from several megabases to a single gene or only one exon. These microhomology-mediated

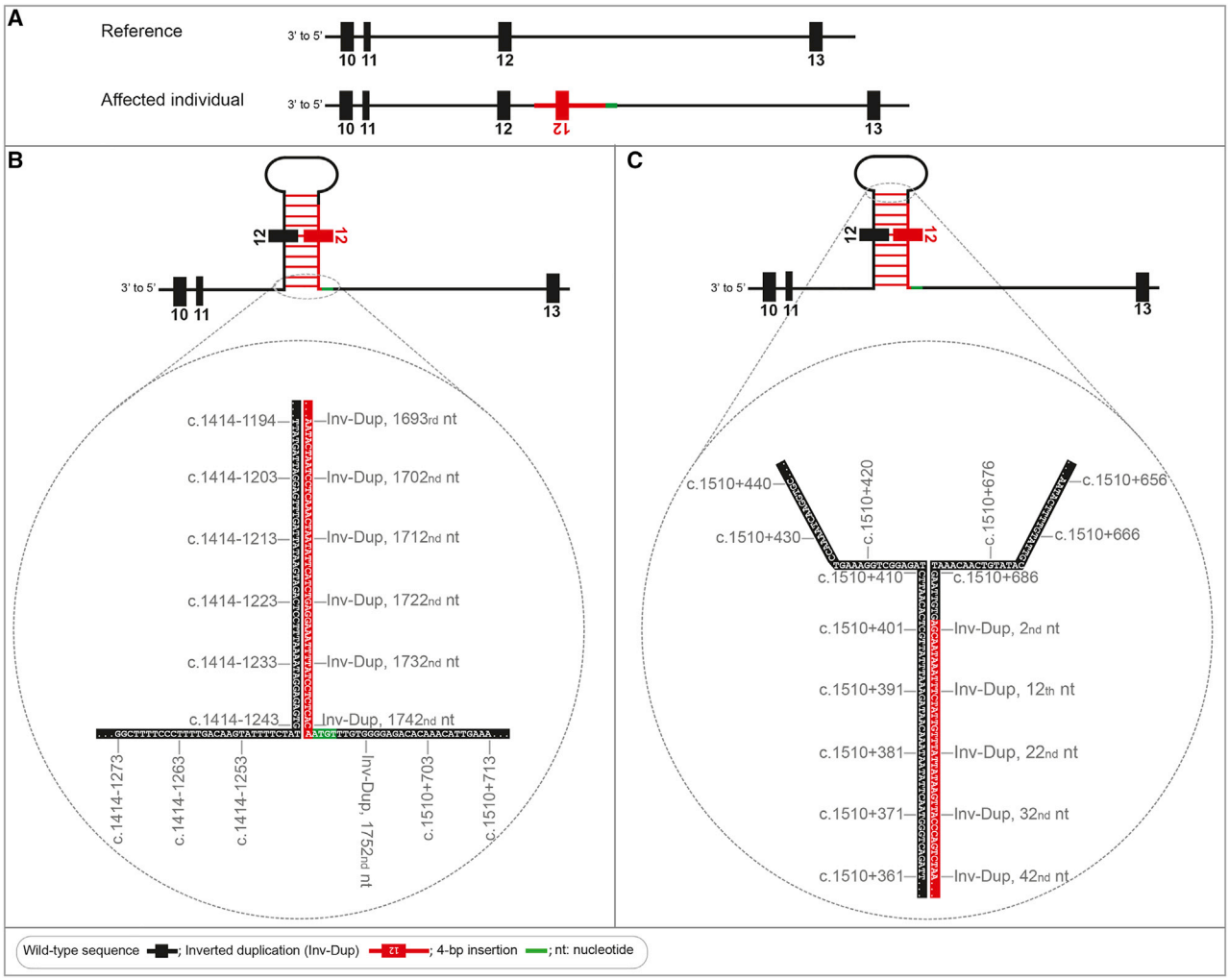

Figure 5. Hairpin formation putatively underlying the observed *CHM* exon 12 skipping in mature mRNA
(A) Schematic representation of *CHM* exons 10 to 13 of the reference genome and the affected individual with the intragenic inverted duplication downstream of exon 12. (B) Enlargement of the hairpin stem at the basal part at the nucleotide level. The first base pair of the hairpin stem is assembled from c.1414−1243 to the 1,742nd nucleotide on the inverted duplication. The last nucleotide of the inverted duplication and 4-bp inserted sequence (highlighted in green) do not contribute to the hairpin stem. (C) Enlargement of the hairpin stem at the top part at the nucleotide level. The hairpin stem is terminated by the last base pair from c.1510+410 to c.1510+686 of wild-type sequence. The 274-nucleotide single-strand RNA starting from c.1510+411 till c.1510+685 is on the loop part of the hairpin structure.

mechanisms have provided new insights for deciphering fundamental pathogenic and evolutionary changes in the human genome.[54,56–59]

In order to understand how this inverted duplication upstream of exon 12 leads to skipping of wild-type exon 12 on the RNA level, we speculated whether the mechanism underlying this splice defect may be explained by an alteration of the mRNA secondary structure, which is investigated broadly in existing literature.[60–63] In a recent study, Masson et al.[64] investigated the disease-causing mechanism of an *Alu*-element insertion in the 3′ UTR of the gene *SPINK1*. By a full-gene expression assay, they confirmed that the inserted *Alu* element is in the opposite orientation with an existing *Alu* element in *SPINK1* intron 3, which disrupted splicing by forming an altered RNA secondary structure leading to severe infantile isolated exocrine pancreatic insufficiency. Likewise, we hypothesized that the wild-type exon 12 and the inverted duplicated exon 12 create a hairpin structure in the pre-mature mRNA. The predicted hairpin could likely interfere with the process of splicing for exon 12 in the mRNA of affected cases of family A, as such validating the splicing defect described for this family in 2003. This phenomenon could prevent binding of essential regulatory splicing elements, such as the spliceosome small nuclear ribonucleoprotein particle (snRNP) complex and exonic enhancer elements, and potentially lead to exon 12 skipping in the mature mRNA as previously shown for this family.[14] The *in silico* RNA structure analysis confirmed that in the aberrant sequence, nucleotides from position c.1414−1243 to c.1510+410 likely generate a hairpin with the inverted duplication (Figure 5). The resulting mRNA defect leads to an out-of-frame skipping of exon

12 and thereby is predicted to lead to a truncated protein after four amino acids (p.Ser473Trpfs*4). Due to the already large-sized wild-type introns 11 and 12 (6 kb and 15 kb), functional validation of this phenomenon was not feasible. Our hypothesis is, however, matching the observed mRNA outcome of exon 12 skipping as observed in family A and may thereby emphasize the underappreciated role of RNA secondary structures in regulating splicing processes, contributing to rare and poorly described disease-causing mechanisms in human cells. A quantification of remaining mRNA in carrier females from family A will provide evidence on a potential correlation of levels of wild-type mRNA and clinical severity (D.V., C.B.H., and R.W.J. Collin, personal communication).

Pathogenic intragenic inverted duplications are not widely reported as genetic causes in diseases, whereas single-exon duplications are found more frequently.[65] One reason can be explained by the inability of genome-wide CNV microarrays to identify the location and orientation of gained genomic material within a gene.[65] Therefore, one may speculate that other gains detected by CNV microarrays or coverage-based NGS tools may underlie similar mutational mechanisms, as these intragenic inverted duplications remain unresolved in less comprehensively studied cases.

The family described in this manuscript has been studied over many years, using various complementary technologies to identify a genetic cause of disease. The necessity of using multiple technologies in order to get the full set of personal genetic variants has been proven especially for structural variants, as recently also described by the 1000 Genomes SV consortium.[66,67] We foresee that long-read sequencing technologies may deliver a near-perfect genome analysis in the future and may then be used as generic stand-alone technology. Although these data are already very promising compared to short-read sequencing technologies, they still come with relatively low throughput and relatively high costs for relatively low coverage, limiting their broad usage today. Optical genome mapping instead offers a relatively high genome coverage, with a straightforward analysis, for comparably low costs. We have recently shown that optical genome mapping presents with 100% sensitivity and >80% positive predictive value for both constitutional as well as somatic structural aberrations.[36] Therefore, it may be considered as a first-tier test for (research) indications where structural variants are suspected to be causative. Optical genome mapping will never be able to replace a sequencing technology. However, until (high-coverage) long-read genomes will be able to replace all other technologies, we argue that optical genome mapping and (short-read or low-coverage long-read) WGS complement each other, as presented in this study highlighting the promise for solving the unsolved rare disease cases.

In conclusion, this study demonstrates the great opportunities of optical genome mapping and long-read sequencing to unravel previously hidden SVs in so-far unsolved diseases. The combined approach of optical genome mapping and long-read sequencing used in this study was beneficial due to the strong correlation between the CHM phenotype and the *CHM* gene. Both approaches appear to be capable of identifying hidden structural variants that remained refractory to standard techniques and may lead to finding new disease mechanisms. As such, they are revealed to be powerful complementary technologies for the molecular diagnoses of previously unsolved rare disease cases.

Data and code availability

All data generated and analyzed during this study are included in this published article and its supplementary information files. Individuals' data cannot be made publicly available due to local regulation. However, specific requests that can be sent to the corresponding authors. All software is commercially available via Bionano Genomics Inc. All filter settings suggested here can be reproduced in the available Bionano Genomics software suite. The pathogenic structural variant has been submitted to the "Global Variome shared LOVD" and can be accessed through LOVD.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100046.

Acknowledgments

We would like to show our gratitude to the affected individuals and their family members in this study. We thank Ellen Kater-Baats, Michiel Oorsprong, and Ronald van Beek for performing optical genome mapping. We further thank the Department of Human Genetics and the Radboud Genome Technology Center for infrastructural and computational support. The work of Z.F. is funded by the Foundation Fighting Blindness USA Project Program Award, grant no. PPA-0517–0717-RAD (to F.P.M.C., S.R., and C.B.H.). The research was supported 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 Algemene Nederlandse Vereniging ter voorkoming van Blindheid, Oogfonds, Landelijke Stichting voor Blinden en Slechtzienden; Rotterdamse Stichting Blindenbelangen, Stichting Blindenhulp, and Stichting Steunfonds Uitzicht (to S.R.). C.G., L.E.L.M.V., and A.H. were supported by the Solve-RD Project. The Solve-RD project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement N° 779257. This research was part of the Netherlands X-omics Initiative and partially funded by NWO (the Netherlands Organization for Scientific Research; project 184.034.019). T.M. was supported by the Sigrid Jusélius Foundation.

Declaration of interests

The authors declare no competing interests.

Received: April 22, 2021
Accepted: July 1, 2021

Web resources

LOVD, https://www.lovd.nl/CHM

OMIM, https://www.omim.org
RNAstructure, https://rna.urmc.rochester.edu/RNAstructure.html
https://genomics.scripps.edu/

References

1. Harris, G.S., and Miller, J.R. (1968). Choroideremia. Visual defects in a heterozygote. Arch. Ophthalmol. *80*, 423–429.
2. van den Hurk, J.A., Schwartz, M., van Bokhoven, H., van de Pol, T.J., Bogerd, L., Pinckers, A.J., Bleeker-Wagemakers, E.M., Pawlowitzki, I.H., Rüther, K., Ropers, H.H., and Cremers, F.P. (1997). Molecular basis of choroideremia (CHM): mutations involving the Rab escort protein-1 (REP-1) gene. Hum. Mutat. *9*, 110–117.
3. MacDonald, I.M., Binczyk, N., Radziwon, A., and Dimopoulos, I. (2020). Choroideremia. In Hereditary Chorioretinal Disorders, G. Cheung, ed. (Springer), pp. 99–106.
4. Kärnä, J. (1986). Choroideremia. A clinical and genetic study of 84 Finnish patients and 126 female carriers. Acta Ophthalmol. Suppl. *176*, 1–68.
5. Pameyer, J.K., Waardenburg, P.J., and Henkes, H.E. (1960). Choroid eremia. Br. J. Ophthalmol. *44*, 724–738.
6. Radziwon, A., Arno, G., K Wheaton, D., McDonagh, E.M., Baple, E.L., Webb-Jones, K., G Birch, D., Webster, A.R., and MacDonald, I.M. (2017). Single-base substitutions in the CHM promoter as a cause of choroideremia. Hum. Mutat. *38*, 704–715.
7. Perez-Cano, H.J., Garnica-Hayashi, R.E., and Zenteno, J.C. (2009). CHM gene molecular analysis and X-chromosome inactivation pattern determination in two families with choroideremia. Am. J. Med. Genet. A. *149A*, 2134–2140.
8. Fahim, A.T., and Daiger, S.P. (2016). The Role of X-Chromosome Inactivation in Retinal Development and Disease. Adv. Exp. Med. Biol. *854*, 325–331.
9. Edwards, T.L., Groppe, M., Jolly, J.K., Downes, S.M., and MacLaren, R.E. (2015). Correlation of retinal structure and function in choroideremia carriers. Ophthalmology *122*, 1274–1276.
10. van Bokhoven, H., van den Hurk, J.A., Bogerd, L., Philippe, C., Gilgenkrantz, S., de Jong, P., Ropers, H.H., and Cremers, F.P. (1994). Cloning and characterization of the human choroideremia gene. Hum. Mol. Genet. *3*, 1041–1046.
11. Cremers, F.P.M., van de Pol, D.J., van Kerkhoff, L.P., Wieringa, B., and Ropers, H.H. (1990). Cloning of a gene that is rearranged in patients with choroideraemia. Nature *347*, 674–677.
12. Simunovic, M.P., Jolly, J.K., Xue, K., Edwards, T.L., Groppe, M., Downes, S.M., and MacLaren, R.E. (2016). The Spectrum of CHM Gene Mutations in Choroideremia and Their Relationship to Clinical Phenotype. Invest. Ophthalmol. Vis. Sci. *57*, 6033–6039.
13. Ramsden, S.C., O'Grady, A., Fletcher, T., O'Sullivan, J., Hart-Holden, N., Barton, S.J., Hall, G., Moore, A.T., Webster, A.R., and Black, G.C. (2013). A clinical molecular genetic service for United Kingdom families with choroideraemia. Eur. J. Med. Genet. *56*, 432–438.
14. van den Hurk, J.A., van de Pol, D.J., Wissinger, B., van Driel, M.A., Hoefsloot, L.H., de Wijs, I.J., van den Born, L.I., Heckenlively, J.R., Brunner, H.G., Zrenner, E., et al. (2003). Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. Hum. Genet. *113*, 268–275.
15. Tucker, T., Marra, M., and Friedman, J.M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. *85*, 142–154.
16. Carss, K.J., Arno, G., Erwood, M., Stephens, J., Sanchis-Juan, A., Hull, S., Megy, K., Grozeva, D., Dewhurst, E., Malka, S., et al.; NIHR-BioResource Rare Diseases Consortium (2017). Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am. J. Hum. Genet. *100*, 75–90.
17. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. USA *106*, 19096–19101.
18. Teer, J.K., and Mullikin, J.C. (2010). Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet. *19* (R2), R145–R151.
19. Salzberg, S.L., and Yorke, J.A. (2005). Beware of mis-assembled genomes. Bioinformatics *21*, 4320–4321.
20. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.
21. Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. *19* (R2), R227–R240.
22. Chaisson, M.J., Wilson, R.K., and Eichler, E.E. (2015). Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. *16*, 627–640.
23. van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet. *34*, 666–681.
24. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front. Genet. *10*, 426.
25. Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
26. Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K.S., et al. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet. Med. *20*, 159–163.
27. Mizuguchi, T., Suzuki, T., Abe, C., Umemura, A., Tokunaga, K., Kawai, Y., Nakamura, M., Nagasaki, M., Kinoshita, K., Okamura, Y., et al. (2019). A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J. Hum. Genet. *64*, 359–368.
28. Pauper, M., Kucuk, E., Wenger, A.M., Chakraborty, S., Baybayan, P., Kwint, M., van der Sanden, B., Nelen, M.R., Derks, R., Brunner, H.G., et al. (2021). Long-read trio sequencing of individuals with unsolved intellectual disability. Eur. J. Hum. Genet. *29*, 637–648.
29. Loomis, E.W., Eid, J.S., Peluso, P., Yin, J., Hickey, L., Rank, D., McCalmon, S., Hagerman, R.J., Tassone, F., and Hagerman, P.J. (2013). Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. *23*, 121–128.
30. Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., et al. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with

neuronal intranuclear inclusion disease. Nat. Genet. *51*, 1215–1221.
31. Seo, J.S., Rhie, A., Kim, J., Lee, S., Sohn, M.H., Kim, C.U., Hastie, A., Cao, H., Yun, J.Y., Kim, J., et al. (2016). De novo assembly and phasing of a Korean human genome. Nature *538*, 243–247.
32. Porubsky, D., Garg, S., Sanders, A.D., Korbel, J.O., Guryev, V., Lansdorp, P.M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. Nat. Commun. *8*, 1293.
33. Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A.K.Y., McCaffrey, J., Young, E., Lam, E.T., Hastie, A.R., Wong, K.H.Y., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat. Commun. *10*, 1025.
34. Chan, S., Lam, E., Saghbini, M., Bocklandt, S., Hastie, A., Cao, H., Holmlin, E., and Borodkin, M. (2018). Structural Variation Detection and Analysis Using Bionano Optical Mapping. Methods Mol. Biol. *1833*, 193–203.
35. Neveling, K., Mantere, T., Vermeulen, S., Oorsprong, M., van Beek, R., Kater-Baats, E., Pauper, M., van der Zande, G., Smeets, D., Weghuis, D.O., et al. (2021). Next generation cytogenetics: comprehensive assessment of 48 leukemia genomes by genome imaging. Am J Hum Genet *108*, 1423–1435.
36. Mantere, T., Neveling, K., Pebrel-Richard, C., Benoist, M., van der Zande, G., Kater-Baats, E., Baatout, I., van Beek, R., Yammine, T., Oorsprong, M., et al. (2021). Next generation cytogenetics: genome-imaging enables comprehensive structural variant detection for 100 constitutional chromosomal aberrations in 85 samples. Am J Hum Genet *108*, 1409–1422.
37. Diekstra, A., Bosgoed, E., Rikken, A., van Lier, B., Kamsteeg, E.J., Tychon, M., Derks, R.C., van Soest, R.A., Mensenkamp, A.R., Scheffer, H., et al. (2015). Translating sanger-based routine DNA diagnostics into generic massive parallel ion semiconductor sequencing. Clin. Chem. *61*, 154–162.
38. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.
39. Erikson, G.A., Bodian, D.L., Rueda, M., Molparia, B., Scott, E.R., Scott-Van Zeeland, A.A., Topol, S.E., Wineinger, N.E., Niederhuber, J.E., Topol, E.J., and Torkamani, A. (2016). Whole-Genome Sequencing of a Healthy Aging Cohort. Cell *165*, 1002–1011.
40. Francioli, L.C., Menelaou, A., Pulit, S.L., Van Dijk, F., Palamara, P.F., Elbers, C.C., Neerincx, P.B., Ye, K., Guryev, V., Kloosterman, W.P.; and Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.
41. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
42. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
43. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
44. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.
45. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
46. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.
47. de Bruijn, S.E., Fiorentino, A., Ottaviani, D., Fanucchi, S., Melo, U.S., Corral-Serrano, J.C., Mulders, T., Georgiou, M., Rivolta, C., Pontikos, N., et al. (2020). Structural Variants Create New Topological-Associated Domains and Ectopic Retinal Enhancer-Gene Contact in Dominant Retinitis Pigmentosa. Am. J. Hum. Genet. *107*, 802–814.
48. Mathews, D.H. (2006). RNA secondary structure analysis using RNAstructure. Curr. Protoc. Bioinformatics Chapter 12, Unit 12.16.
49. Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. PathoGenetics *1*, 4.
50. Haer-Wigman, L., van Zelst-Stams, W.A., Pfundt, R., van den Born, L.I., Klaver, C.C., Verheij, J.B., Hoyng, C.B., Breuning, M.H., Boon, C.J., Kievit, A.J., et al. (2017). Diagnostic exome sequencing in 266 Dutch patients with visual impairment. Eur. J. Hum. Genet. *25*, 591–599.
51. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353.
52. Pfundt, R., Del Rosario, M., Vissers, L.E.L.M., Kwint, M.P., Janssen, I.M., de Leeuw, N., Yntema, H.G., Nelen, M.R., Lugtenberg, D., Kamsteeg, E.J., et al. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. Genet. Med. *19*, 667–675.
53. Retterer, K., Scuffins, J., Schmidt, D., Lewis, R., Pineda-Alvarez, D., Stafford, A., Schmidt, L., Warren, S., Gibellini, F., Kondakova, A., et al. (2015). Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. Genet. Med. *17*, 623–629.
54. Abyzov, A., Li, S., Kim, D.R., Mohiyuddin, M., Stütz, A.M., Parrish, N.F., Mu, X.J., Clark, W., Chen, K., Hurles, M., et al. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat. Commun. *6*, 7256.
55. Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. *41*, 849–853.
56. Parks, M.M., Lawrence, C.E., and Raphael, B.J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. Genome Biol. *16*, 72.
57. Lupski, J.R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. *14*, 417–422.

58. Gu, S., Yuan, B., Campbell, I.M., Beck, C.R., Carvalho, C.M., Nagamani, S.C., Erez, A., Patel, A., Bacino, C.A., Shaw, C.A., et al. (2015). Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. Hum. Mol. Genet. *24*, 4061–4077.
59. Ottaviani, D., LeCain, M., and Sheer, D. (2014). The role of microhomology in genomic structural variation. Trends Genet. *30*, 85–94.
60. Saha, K., England, W., Fernandez, M.M., Biswas, T., Spitale, R.C., and Ghosh, G. (2020). Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing. Nucleic Acids Res. *48*, 6294–6309.
61. Rubtsov, P.M. (2016). [Role of pre-mRNA secondary structures in the regulation of alternative splicing]. Mol. Biol. (Mosk.) *50*, 935–943.
62. Jin, Y., Yang, Y., and Zhang, P. (2011). New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. RNA Biol. *8*, 450–457.
63. Solnick, D. (1985). Alternative splicing caused by RNA secondary structure. Cell *43*, 667–676.
64. Masson, E., Maestri, S., Cooper, D.N., Férec, C., and Chen, J.-M. (2020). RNA secondary structure mediated by Alu insertion as a novel disease-causing mechanism. bioRxiv, 2020.2001.2030.926790.
65. Lupski, J.R. (2015). Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ. Mol. Mutagen. *56*, 419–436.
66. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.
67. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117.