Christopher Watson ORCID iD: 0000-0003-2371-1844

Long-read nanopore sequencing resolves a TMEM231 gene conversion event causing Meckel-Gruber syndrome

Christopher M. Watson[1,2], Philip Dean[1], Nick Camm[1], Jennifer Bates[1], Ian M. Carr[2], Carol A. Gardiner[3] David T. Bonthron[1,2]

1: Yorkshire Regional Genetics Service, St. James's University Hospital, Leeds, LS9 7TF, United Kingdom

2: Leeds Institute of Medical Research, University of Leeds, St. James's University Hospital, Leeds, LS9 7TF, United Kingdom

3: West of Scotland Regional Genetics Services, Queen Elizabeth University Hospital, 1345 Govan Road, Glasgow G51 4TF, United Kingdom

*Corresponding author:

Dr Christopher M. Watson

6.2 Clinical Sciences Building

Yorkshire Regional Genetics Service

St James's University Hospital

Leeds, LS9 7TF

United Kingdom

E-mail: c.m.watson@leeds.ac.uk

Tel: +44 (0) 113 206 5677

Fax: +44 (0) 113 343 8702

ABSTRACT

The diagnostic deployment of massively parallel short-read next generation sequencing (NGS) has greatly improved genetic test availability, speed and diagnostic yield, particularly for rare inherited disorders. Nonetheless, diagnostic approaches based on short-read sequencing have a poor ability to accurately detect gene conversion events. We report on the genetic analysis of a family in which 3 fetuses had clinical features consistent with the autosomal recessive disorder Meckel-Gruber syndrome (MKS). Targeted NGS of 29 known MKS-associated genes revealed a heterozygous *TMEM231* splice-donor variant c.929+1A>G. Comparative read-depth analysis, performed to identify a second pathogenic allele, revealed an apparent heterozygous deletion of *TMEM231* exon 4. To verify this result we performed single molecule long-read sequencing of a LR-PCR product spanning this locus. We identified 4 missense variants which were absent from the short-read dataset due to the preferential mapping of variant-containing reads to a downstream *TMEM231* pseudogene. Consistent with the parental segregation analysis, we demonstrate that the single-molecule long reads could be used to show that the variants are arranged in *trans*. Our experience shows that robust validation of apparent dosage variants remains essential to avoid the pitfalls of short-read sequencing, and that new "third-generation" long-read sequencing technologies can already aid routine clinical care.

Keywords:

Meckel-Gruber syndrome, TMEM231, gene conversion, nanopore sequencing

INTRODUCTION

The molecular diagnosis of rare genetic conditions has been revolutionised by the widespread adoption of short-read "next generation" sequencing (NGS). For urgent diagnosis (for example in paediatric intensive care units), it is now possible to perform whole genome sequencing in under two weeks (Mestek-Boukhibar et al., 2018). Nonetheless, most routine diagnostic analyses still focus on a subset of the genome; typically using a range of hybridisation capture reagents that enable enrichment of the coding and immediate flanking intronic sequences of disease-associated genes. For genetically heterogeneous disorders, the ability to concurrently sequence multiple disease genes, using a single reagent, has overcome the need to perform consecutive single-gene tests, and has thereby greatly improved the overall efficiency of testing pathways.

Although targeted gene capture panels remain popular, an increasingly hypothesis-free approach to gene selection means that testing is no longer predicated entirely on accurate clinical diagnosis. This is reducing the effects of ascertainment biases reported in the medical literature, and allowing refinement of the mutation spectrum for many disorders. In this way, the genotype-phenotype relationships for syndromes that have overlapping diagnostic features are being clarified. Two such conditions with considerable underlying genetic heterogeneity are Joubert (JBTS; MIM# 213300) and Meckel-Gruber (MKS; MIM# 249000) syndromes, which are ciliopathies associated with causative mutations in at least 27 and 17 genes respectively, 12 of which are reported to cause both conditions (Hartill et al., 2017). JBTS is characterised by a characteristic neuroradiological finding (the molar tooth sign) in conjunction with neurodevelopmental abnormalities, notably hypotonia and ataxia. There are frequently abnormalities of oculomotor function and breathing pattern, as well as a wide spectrum of associated features, including polydactyly, retinal dystrophy, renal disease and hepatic fibrosis. Although some children die in infancy, most survive, with varying developmental outcomes. MKS is a more severe perinatally lethal syndrome that is characterised by posterior cranial fossa abnormalities (typically occipital encephalocele), postaxial polydactyly, bilateral enlarged cystic kidneys and hepatic ductal plate malformation. Because of the genetic and phenotypic overlap between the two disorders, like most laboratories, we routinely deploy a gene panel diagnostic approach that examines all JBTS- or MKS-associated disease genes, irrespective of which of the two diagnostic categories the clinical features suggest.

The overwhelming majority of diagnostic NGS data is generated using sequencing-by-synthesis chemistry on instruments manufactured by Illumina. Although this method yields reads with high per-base consensus accuracy, the relatively short individual read length can prevent unambiguous mapping to the genomic reference sequence. Targets disproportionately affected by this limitation include families of gene and pseudogene homologs (Ebbert et al., 2019). Other genomic regions are intractable to analysis due to difficulties caused by the sequencing chemistry itself. These include regions of high GC content, which are difficult to amplify. “Third-generation” single-molecule sequencers promise to overcome some of these limitations by generating long sequencing reads (spanning multiple kilobases) and enabling targeted library preparation without a requirement for PCR amplification.

Here, we describe the use of “third-generation” methods to resolve an anomalous diagnostic result stemming directly from limitations inherent to the automated short-read pipeline approach. In a fetal sample referred for molecular genetic analysis of MKS, targeted variant screening identified a likely pathogenic *TMEM231* splice site mutation c.929+1A>G and an apparent heterozygous deletion of *TMEM231* exon 4. However, validation of the apparent deletion-containing allele, performed using single-molecule nanopore sequencing, revealed a gene conversion event (introducing 4 missense variants). This result had been masked from the short-read dataset due to the preferential mapping of gene conversion-containing reads to the downstream pseudogene.

MATERIALS AND METHODS

Consultant clinical geneticists reviewed post-mortem reports on the deceased subjects. Ethical approval for this study was given by the Leeds East Research Ethics Committee (18/YH/0070).

DNA was isolated using phenol/chloroform or bead-based extraction methods, following informed written consent. A custom SureSelect hybridisation enrichment reagent was used to create an Illumina-compatible sequencing library following manufacturer’s protocols (Agilent Technologies, Wokingham, UK). The probe set targeted coding sequences and immediate flanking intronic regions of 223 genes implicated in paediatric neurological disorders, of which 29 were associated with Meckel-Gruber or Joubert syndromes (see Supp. Data File 1). Eight such libraries were pooled in equimolar concentrations to create a “batch” which was combined with two other similar library batches and then sequenced on a NextSeq500 to produce paired-end 151-bp reads from a high-output cartridge (Illumina, San Diego, CA, USA).

Data processing was performed using a standard informatics pipeline. Raw data were converted from BCL to FASTQ.gz format using bcl2fastq (v.2.17.1.14). Adaptor sequences and low-quality bases (-q 10) were removed from reads using Cutadapt (v.1.9.1) (https://cutadapt.readthedocs.org) (Martin 2011) prior to alignment to an indexed human reference genome (hg19) using BWA MEM (v.0.7.13) (http://bio-bwa.sourceforge.net) (Li and Durbin, 2009). File manipulations, including sam-to-bam conversion, duplicate removal and bam file indexing, were performed using Picard (v.2.1.1) (http://broadinstitute.github.io/picard). Assembly-based realignment was performed at hybridisation capture target sites using ABRA2 v.0.03 (https://github.com/mozack/abra2) (Mose et al., 2014). The Genome Analysis Toolkit (GATK) (v.2.3-4Lite) was used to perform indel realignment, base quality score recalibration and variant calling, following the GATK best practice workflow (https://software.broadinstitute.org/gatk/) (DePristo et al., 2011). Each variant was annotated using Alamut Batch standalone (v.1.5.2; database version 2016.09.27) (Interactive Biosoftware, Rouen, France) before its clinical significance was assessed according to the Association for Clinical Genomic

Science best practice guidelines (Ellard et al., 2018). An in-house comparative read-depth method was performed to identify dosage variants. Normalised read counts for the control group were generated from corresponding intra-batch patient libraries. The number of reads mapping to each target base was assessed using the GATK walkers DepthOfCoverage, CallableLoci and CountReads. Aligned sequence reads were manually inspected using the Integrative Genome Viewer (v.2.4.10) (http://software.broadinstitute.org/software/igv/) (Thorvaldsdóttir et al., 2013). Reported variants have been deposited in LOVD3 (http://www.lovd.nl).

To characterise an apparent *TMEM231* heterozygous exon 4 deletion (NM_001077416.2), a 6163-bp long-range PCR amplicon was sequenced using a MinION long-read sequencer (ONT: Oxford Nanopore Technologies, Oxford, UK). Each PCR reaction consisted of 1 µL of genomic DNA (~50 ng/µL), 14.24 µL of nuclease-free H_2O, 2 µL of 10 × SequalPrep™ reaction buffer (Invitrogen, Paisley, UK), 0.36 µL of 5 U/µL SequalPrep™ long polymerase (Invitrogen), 0.4 µL of DMSO (Invitrogen), 1µL of 10 × SequalPrep™ Enhancer A (Invitrogen), and 0.5 µL each of 10 pmol/µL forward (dCAGGAAACAGCTATGACCACCACCTGATTCTGAAACACG) and reverse (dTGTAAAACGACGGCCAGTGCCAGTGAACATTTGAAAGGC) primers. Oligonucleotides were purchased from Thermo Fisher Scientific (Waltham, MA, USA). Thermocycling conditions are recorded in Supp. Table S1. Both primers contained universal sequencing tags (underlined) to enable confirmation of amplicon specificity by Sanger sequencing on an ABI3730, following manufacturer's protocols (Applied Biosystems, Paisley, UK). Amplification products were subsequently resolved on a 1% tris-borate EDTA agarose gel, before being extracted and purified using a QIAquick column (Qiagen GmbH, Hilden, Germany). To create a MinION-compatible sequencing library, an end-repair and nickase treatment reaction was first performed. The reaction consisted of 3.5 μL Ultra™ II end prep reaction buffer (New England Biolabs, Ipswich, MA, USA), 3.5 μL FFPE DNA repair buffer (New England Biolabs), 3.0 μL Ultra™ II end prep enzyme mix (New England Biolabs), 2.0 μL FFPE DNA repair mix (New England Biolabs), 32.4 μL nuclease-free H_2O and 15.6 μL (1.5 μg) of the gel-extracted 6.2-kb PCR product (95.9 ng/μL). The reaction was incubated at 20°C for 5 minutes then 65°C for 5 minutes. Following an AMPure XP bead clean-up reaction (Beckman Coulter, Indianapolis, IN, USA), sequencing adaptors were ligated to the termini of double-stranded DNA. The reaction comprised 60 μL of PCR products, 25 μL of LNB (ONT), 10 μL of Quick Ligase (New England Biolabs) and 5 μL of AMX (ONT) and was incubated at room temperature for 10 minutes. A further AMPure XP bead clean-up was performed, using SFB (ONT) to wash the beads, and 15 μL of EB (ONT) to elute the sample. A MinION FLO-MIN106 Rev D flowcell was prepared for loading. Briefly, 30 μL of FLT (ONT) was combined with a vial of FLB (ONT) to create a flowcell priming mix, from which 800 μL was loaded through the flowcell priming port. The eluted library (12 μL) was combined with 37.5 μL of SQB (ONT) and 25.5 μL of LB (ONT), then loaded into the flowcell through the SpotON port. A 48-hour MinION sequencing run was initiated using MinKNOW software (v.3.1.13) (ONT).

Offline basecalling was performed using Guppy (v.2.1.3) to convert raw data from fast5 to FASTQ format (http://nanoporetech.com). Adaptor sequences were trimmed from the resulting reads using Porechop (v.0.2.3) (https://github.com/rrwick/Porechop) before NanoFilt (v.2.2.0) (https://github.com/wdecoster/nanofilt) removed low-quality reads (-q 4) (De Coster et al., 2018). Alignment to the human reference genome (build hg19) was performed using minimap2 (v.2.10) (https://github.com/lh3/minimap2) (Li, 2018). In view of the excessive read depth generated from the full dataset (mean coverage of 1.19 million ×), raw data were down-sampled to comprise only the first 20 fast5 files produced by the MinION (corresponding to 80,000 reads). Summary sequencing metrics are recorded in Supp. Table S2. Variant calling was performed using NanoPolish (v.0.11.0) (https://nanopolish.readthedocs.io) to generate a VCF file, used in combination with WhatsHap (v.0.17) (https://whatshap.readthedocs.io/) (Martin et al., 2016) to group individual reads based on their variant-defined haplotype. Sam-to-bam file conversion, read sorting and bam file indexing was performed using samtools (v.1.8) (Li et al., 2009). Read metrics were generated using NanoStat (v.1.1.0) (https://github.com/wdecoster/nanostat) (De Coster et al., 2018).

To confirm and genotype the *TMEM231* intron 3/exon 4 gene conversion, a 1,187-bp amplicon (based on the assay reported in Maglic et al., 2016) was optimised for use with our routine diagnostic Sanger sequencing workflow (as previously described). Each PCR comprised 0.5 μL of genomic DNA (~50 ng/μL), 14.1 μL of Megamix (Microzone Ltd, Haywards Heath, UK), 0.2 μL of 10 pmol/μL forward primer (dCAGGAAACAGCTATGACCTGACTACGGCCTTGAACTCA) and 0.2 μL of 10 pmol/μL reverse primer (dTGTAAAACGACGGCCAGTGTAACACTGTAGATGCTTTTAG). Thermocycling conditions are recorded in Supp. Table S3. An internal sequencing primer (dCCTTGAAAGTCCTGAGCAGC) was used to generate the Sanger sequencing chromatograms. To confirm and validate the *TMEM231* intron 5 (NM_001077416.2) splicing variant, a 517-bp amplicon was optimised for Sanger sequencing. Each reaction consisted of 0.5 µL of genomic DNA (~50 ng/µL), 11.3 µL of Megamix, 0.1 µL of 10 pmol/µL forward primer (dTGTAAAACGACGGCCAGTTGCTTTAAGAGGCAGGGTCT) and 0.1 µL of 10 pmol/µL reverse primer (dCAGGAAACAGCTATGACCGCCAGAGTCAAACAGCCATC). Thermocycling conditions are recorded in Supp. Table S3. Chromatograms from all Sanger sequencing reactions were analysed using 4Peaks software (v.1.8) (http://nucleobytes.com/4peaks/index.html).

RESULTS

DNA from a male fetus of Scottish origin was referred for molecular genetic testing of 29 Meckel-Gruber and Joubert syndrome-associated disease genes. The pregnancy had been terminated in the thirteenth week of gestation, following identification of enlarged multicystic dysplastic kidneys and occipital encephalocoele. Additional post-mortem findings included hepatic ductal plate malformation and postaxial polydactyly of both hands. Other dysmorphic features, including small low set ears, a wide neck with broad shoulders and bilateral talipes equinovarus, were also noted. The kidneys were of increased weight and brain weight was markedly reduced. Two previous abnormal pregnancies were also recorded; in the first, the fetus had occipital encephalocoele and bilateral enlarged cystic kidneys, and the second ended in a first trimester miscarriage.

Two possibly relevant DNA sequence variants were initially identified, and their clinical significance was interpreted according to criteria recommended by the Association for Clinical Genomic Science (Ellard et al., 2018). The first of these was the heterozygous variant *KIAA0586* c.4589-6A>C (NM_001244189.1). Standard *in silico* tools suggested that this variant was unlikely to be pathogenic and did not provide an explanation for the presenting clinical features. The second variant, *TMEM231* (c.929+1A>G) (NM_001077416.2), which was heterozygous and located at the intron 5 splice donor site, was classified as likely pathogenic. This variant had not been previously reported on ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) or the LOVD3 *TMEM231* gene-specific variant database (https://databases.lovd.nl/shared/variants/TMEM231). Consistent with a rare recessive inheritance pattern, there were only two recorded observations of this sequence variant in the gnomAD v.2.1 dataset (among 244,308 total alleles), and these had been identified in separate individuals (Lek et al., 2016). The zygosity and segregation of the c.929+1A>G variant were confirmed by Sanger sequencing; the proband's mother was determined to be a heterozygous carrier (Supp. Figure S1).

We subsequently performed copy number analysis, using a comparative read depth method, in an attempt to identify a second pathogenic allele that would confirm a molecular diagnosis of Meckel-Gruber syndrome. These data revealed an apparent *TMEM231* exon 4 deletion (NM_001077416.2), with a dosage quotient (0.63) suggestive of a heterozygous deletion (Supp. Figure S2A). To verify and better delineate this observation, a 6.2-kb long-range PCR amplicon across the region was generated. We opted to sequence this using a long-read single-molecule sequencer (Oxford Nanopore MinION), because the two alleles could not be resolved on agarose gel electrophoresis for conventional Sanger sequencing. The expected exon 4 deletion could not be validated by this analysis, but we instead observed four clustered substitution variants (c.742-1G>A, c.742A>G, c.747C>T and c.752T>C) that were present in the long-read, but not the short-read, datasets (Figure 1). The same clustered variants have previously been observed and attributed to a gene conversion event occurring between *TMEM231* and a pseudogene located approximately 40 kb downstream (centromeric) from the functional gene (Maglic et al., 2016). It therefore appeared that during alignment, reassignment of short reads containing pseudogene-derived variants, from TMEM231 to the more closely-matching pseudogene locus, was the explanation for the apparent loss of read depth at exon 4. We confirmed the presence of the conversion event by Sanger sequencing a 1187-bp amplicon derived from the gene-specific intron 3/exon 4 locus, and showed that it was paternally inherited (Figure 2A). Two flanking pseudogene-derived variants (c.742-49G>A and c.823+4A>G) reported by Maglic et al. (2016) were not detected on the converted allele in our patient (data not shown); suggesting that the allele we have observed reflects a recurrent gene conversion event, rather than an ancestral founder allele.

To assess the relative number of copies of the *TMEM231* exon 4 pseudogene, we performed comparative read depth analysis across the corresponding downstream pseudogene region. We calculated a normalised dosage quotient of

1.20, suggestive of a total of 3 copies of the pseudogene exon (Supp. Figure S2). We determined that short-read sequences that mapped to *TMEM231* pseudogene exon 4 (chr16:75,536,434-75,536,515) have a mean MAPQ value of 54, indicating high alignment specificity.

To further demonstrate the clinical utility of long-read sequencing, we performed variant calling across the long-range PCR amplicon and used the resulting genotypes to group reads according to their likely haplotype (Figure 2B). These data were consistent with the pathogenic *TMEM231* variants being arranged in *trans*; an observation that was consistent with the parental inheritance studies. DNA samples were not available for testing from the previous two affected pregnancies, but given the clinical similarities and the genetic findings above, it seems highly likely that they too were compound heterozygous for the intron 5 splice donor mutation and the exon 4 gene conversion.

By using the apparent TMEM231 exon 4 deletion as a surrogate marker for the gene conversion allele, we retrospectively analysed 45 cases referred for JBTS/MKS testing, and 377 cases referred for analysis of non-JBTS/MKS associated gene (but sequenced using the same hybridisation capture reagent). We were unable to identify any further gene conversion alleles, suggesting that their population frequency is less than 1/844.

DISCUSSION

NGS workflows that target curated lists of disease-associated genes are now deeply embedded into routine practice. For some heterogeneous conditions (such as the Meckel-Gruber/Joubert spectrum of ciliopathies discussed here) there remains debate concerning the set of genes that should be analysed. For practicality, many laboratories rely on hybridisation capture to enable enrichment of target loci, but capture reagents frequently become outdated as gene lists and reported mutation spectra expand. Whole genome sequencing has the potential to circumvent these limitations, and its introduction into routine practice has now begun (Turnbull et al., 2018). Irrespective of these different approaches, though, the overwhelming majority of NGS-based diagnoses are based on short-read Illumina technology, for which there are well-documented limitations (van Dijk et al., 2018). In recent years, in contrast, the use of "third-generation" long-read sequencers has significantly increased our knowledge of genetic variation, particularly with respect to the frequency and complexity of structural variants (Huddleston and Eichler, 2016).

We routinely supplement our diagnostic variant calling with comparative read-depth analyses. Although it is recognised that the sensitivity of this varies across different target loci (due to factors such as GC content affecting capture efficiency) false positive deletion-containing variants are rarely identified. This led us to query our inability to validate an apparent *TMEM231* exon 4 deletion.

The finding that four variants were present in the long-read (but not the short-read) dataset alerted us to the presence of a previously-reported gene conversion event involving a downstream *TMEM231* pseudogene (Maglic et al., 2016). The implicated source exon contains pseudogene-specific variants, and these result in loss of TMEM231 function (as a result of aberrant splicing) following the gene conversion.

When alignment to the reference genome is performed, the variants introduced at the functional locus by gene conversion direct alignment of the short reads to the corresponding pseudogene locus. This has two effects: it “hides” the existence of these sequence variants (from both automated variant callers and manual interrogation by genome analysts) and reduces the apparent read depth over the target exon, mimicking a deletion. It may be noted that while it is not uncommon to encounter reads mapping to pseudogene sequences, the analyst is typically alerted to such loci when read alignment metrics (defined by the MAPQ score) are reduced 0; this was not the case here.

From a molecular diagnostic perspective, we have previously discussed how the technical limitations of existing NGS workflows can be ascertained by further investigating patients in whom only a single pathogenic variant is initially discovered (Watson et al., 2016). Although mutations in *TMEM231* have thus far reportedly accounted for only a limited proportion of the MKS mutation spectrum (Hartill et al., (2017) it remains unclear how many patients have been identified with only a single pathogenic variant and in whom a second pathogenic variant, if present, remains refractory to analysis. To our knowledge, this is the second report of a *TMEM231* exon 4 gene conversion, and our observation that the c.742-49G>A and c.823+4A>G pseudogene-specific variants are absent from the gene-conversion allele, suggest that this mutation is recurrent, rather than ancestral, in origin. Further characterisation of the gene-specific and pseudogene loci, to define the minimum and maximum extent of the gene conversion event, was not possible, due to limited biological material.

The recognition of recurrent *TMEM231* gene conversion events leads us to recommend that such events should be ruled out in all cases for whom a single pathogenic mutation has thus far been identified. In view of the route by which we identified the gene conversion event, we suggest that laboratories performing comparative read-depth analysis of *TMEM231* exon 4 also assess the relative dosage of the corresponding pseudogene exon (chr16:75,536,434-75,536,515). Despite this recommendation, it remains unclear how neutral polymorphisms affecting the functional gene, or pseudogene-specific variants that are not represented in existing reference assemblies (but may be frequent and as-yet uncharacterised in the general population), would affect the sensitivity of this approach.

A critical long-standing consideration, when undertaking PCR-based amplification of loci known to have pseudogene counterparts, has been the specificity of amplification primers. With the advent of "third generation" sequencing, this concern may, to a greater or lesser extent, be alleviated by accurate mapping of long-read sequences from a mixed pool of functional and pseudogene-derived amplification products. The extent to which this will work is likely to depend on both the achieved read length and the sequence divergence between gene and pseudogene; this will be unpredictable in the context of the whole genome, and require locus-by-locus assessment. Fortunately, the possibility of performing PCR-free target enrichment and long-read sequencing, from bulk genomic DNA (Gabrieli et al., 2018), may eventually obviate this question of primer specificity entirely.

Our ability to determine phase information directly, showing that the pathogenic variants were on opposite parental alleles, highlights a further advantage of long-read sequencing. This capability is likely to prove crucial whenever it is not possible to undertake parental segregation studies. Additional technological advances, in direct and synthetic long-read sequencing, present possibilities for creating contiguous haplotypes linking gene-specific regions to their pseudogene counterparts; this could enable more precise characterisation of the minimal and maximal extents of identified gene conversion events. However, the requirement for high-molecular weight DNA must be borne in mind for future diagnostic work-flows, and may prove to be a significant challenge, particularly in the analysis of degraded fetal tissue or very small diagnostic biopsies.

Finally, this report also cautions that dosage variants identified by comparative read-depth methods should always be validated by another method. Aside from the quality assurance provided by this approach, precise characterisation of clinically relevant structural variants often allows the creation of facile diagnostic tests that can be used for extended family testing (Watson et al., 2014).

Conflict of interest statement

Dr Watson has received travel expenses to speak at a conference organised by Oxford Nanopore Technologies organised conference.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34,* 2666-2669. *doi:*10.1093/bioinformatics/bty149

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics 43,* 491-498. *doi:*10.1038/ng.806

Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, Keene D, Larson E, Crane P, Asmann YW, Ertekin-Taner N, Younkin SG, Ross OA, Rademakers R, Petrucelli L, Fryer JD. (2019). Systematic analysis of dark and camouflaged genes: disease-relevant genes hiding in plain sight. *bioRxiv* 9th January. *doi:*10.1101/514497

Ellard S, Baple EL, Owens M, Eccles DM, Turnbull C, Abbs S, Scott R, Deans ZC, Lester T, Campbell J, Newman WG, McMullan DJ. ACGS Best Practice Guidelines for Variant Classification. (2018). *Association for Clinical Genomic Science*. Retrieved from http://www.acgs.uk.com/media/1140458/uk_practice_guidelines_for_variant_classification_2018_v1.0.pdf

Gabrieli T, Sharim H, Fridman D, Arbib N, Michaeli Y, Ebenstein Y. (2018). Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Research, 46,* e87. *doi:*10.1093/nar/gky411.

Hartill V, Szymanska K, Sharif SM, Wheway G, Johnson CA. (2017). Meckel-Gruber Syndrome: An Update on Diagnosis, Clinical Management, and Research Advances. *Frontiers In Pediatrics 5,* 244. *doi:*10.3389/fped.2017.00244

Huddleston J and Eichler EE. (2016). An Incomplete Understanding of Human Genetic Variation. *Genetics, 202,* 1251-1254. *doi:*10.1534/genetics.115.180539

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536,* 285-291. *doi:*10.1038/nature19057

Li H and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics, 25,* 1754-1760. *doi:*10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25,* 2078-2079. *doi:*10.1093/bioinformatics/btp352

Li H. Minimap2: pairwise alignment for nucleotide sequences. (2018). *Bioinformatics 34,* 3094-3100. *doi:*10.1093/bioinformatics/bty191

Maglic D, Stephen J, Malicdan MC, Guo J, Fischer R, Konzman D, NISC Comparative Sequencing Program, Mullikin JC, Gahl WA, Vilboux T, Gunay-Aygun M. (2016). TMEM231 Gene Conversion Associated with Joubert and Meckel-Gruber Syndromes in the Same Family. *Human Mutation, 37,* 1144-1148. *doi:*10.1002/humu.23054

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal 17,* 10-12. *doi:*10.14806/ej.17.1.200

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schoenhuth A, Marschall T. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv* 14th November. *doi:*10.1101/085050

Mestek-Boukhibar L, Clement E, Jones WD, Drury S, Ocaka L, Gagunashvili A, Le Quesne Stabej P, Bacchelli C, Jani N, Rahman S, Jenkins L, Hurst JA, Bitner-Glindzicz M, Peters M, Beales PL, Williams HJ. (2018). Rapid Paediatric

Sequencing (RaPS): comprehensive real-life workflow for rapid diagnosis of critically ill children. *Journal of Medical Genetics, 55,* 721-728. *doi:*10.1136/jmedgenet-2018-105396

Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. (2014). ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics, 30,* 2813-2815. *doi:*10.1093/bioinformatics/btu376

Thorvaldsdóttir H, Robinson JT, Mesirov JP. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics, 14,* 178-192. *doi:*10.1093/bib/bbs017

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereau A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100 000 Genomes Project. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ, 361,* k1687. *doi:*10.1136/bmj.k1687

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics, 34,* 666-681. *doi:*10.1016/j.tig.2018.05.008

Watson CM, Crinnion LA, Tzika A, Mills A, Coates A, Pendlebury M, Hewitt S, Harrison SM, Daly C, Roberts P, Carr IM, Sheridan EG, Bonthron DT. (2014). Diagnostic whole genome sequencing and split-read mapping for nucleotide resolution breakpoint identification in CNTNAP2 deficiency syndrome. *American Journal of Medical Genetics Part A, 164A,* 2649-2655. *doi:*10.1002/ajmg.a.36679

Watson CM, Crinnion LA, Berry IR, Harrison SM, Lascelles C, Antanaviciute A, Charlton RS, Dobbie A, Carr IM, Bonthron DT. (2016). Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. *BMC Medical Genetics, 17,* 1.

*doi:*10.1186/s12881-015-0265-z

FIGURES

Figure 1: A representative comparison of read alignments between the short-read NextSeq and long-read MinION datasets. Reads identifying the gene conversion event (containing four non-reference nucleotides) were not visible in in short-read data. The *y*-axis scale for each cumulative read-depth plot is labelled. Non-reference bases with a minor allele fraction exceeding 0.35 are highlighted. Non-reference bases in variant-containing reads are coloured with respect to the sense strand (T: red, C: blue, A: green and G: brown). The IGV's "quick-consensus mode" is enabled. *TMEM231* is encoded on the antisense strand. Sense strand nucleotides are shaded grey.

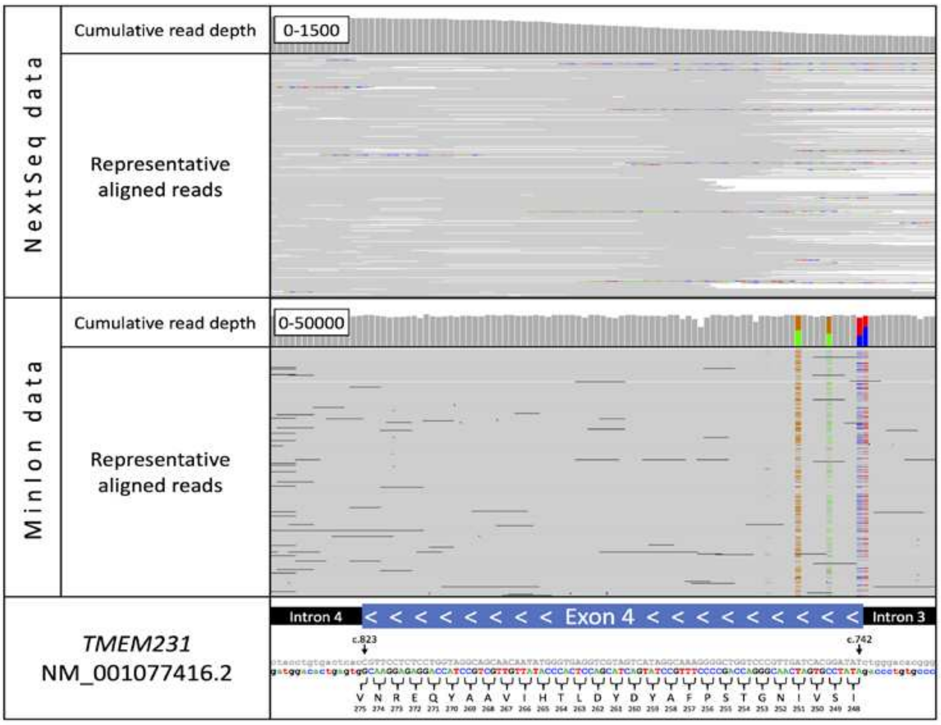
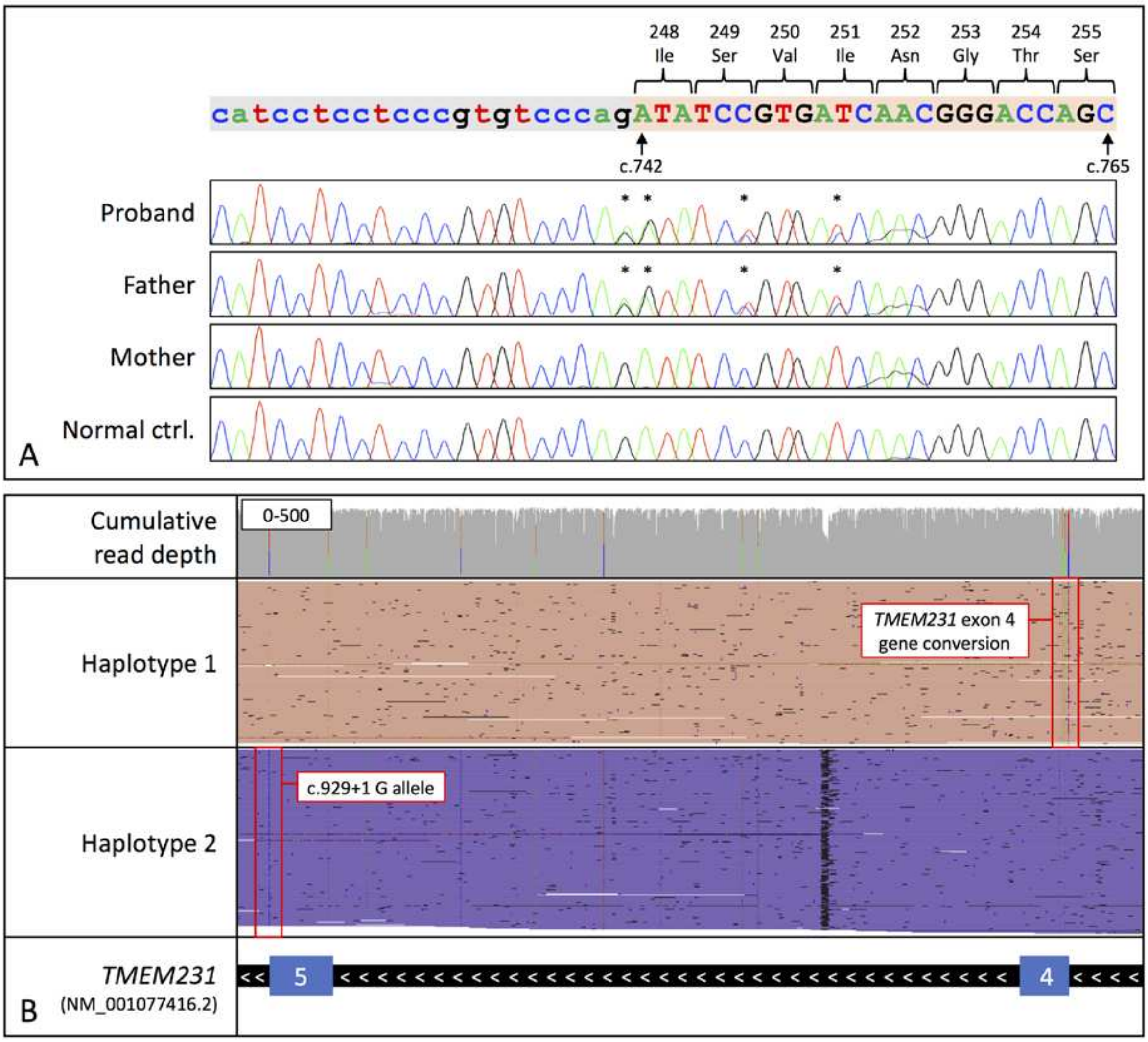


Figure 2: (A) Sanger sequencing chromatograms confirmed the intron 3/exon 4 gene conversion in the proband and established paternal inheritance. A "*" denotes each of the four variants (c.742-1G>A, c.742A>G, c.747C>T and c.752T>C) that characterise the gene conversion event. Intron 3 sequence is displayed in lowercase and shaded grey. Exon 4 sequence is displayed in uppercase with orange highlighting. Numbering is according to transcript NM_001077416.2. Chromatograms have been reverse-complemented to aid their interpretation in the context of the *TMEM231* gene sequence, which is encoded on the antisense strand. **(B)** Long-read sequencing demonstrated that that the heterozygous gene conversion and c.929+1A>G mutations (red boxes) are arranged in *trans* on separate parental haplotypes. To aid visualisation, 1% of the analysed reads were displayed using the IGV's "quick consensus mode". The *y*-axis scale for each cumulative read-depth plot is labelled.


Ile
Ser
Val
Ile
Asn
Gly
Thr
Ser
catcctcctcccgtgtcccagATATCCGTGATCAACGGGACCAGC
c.742
c.765
Proband
Father
Mother
Normal ctrl.
A
Cumulative read depth
Haplotype 1
TMEM231 exon 4 gene conversion
Haplotype 2
c.929+1 G allele
TMEM231
(NM_001077416.2)
B