

METHOD

Phasing of de novo mutations using a scaled-up multiple amplicon long-read sequencing approach

Giles S. Holt¹  | Lois E. Batty¹ | Bilal K. S. Alobaidi¹ | Hannah E. Smith¹ |
Manon S. Oud²  | Liliana Ramos³ | Miguel J. Xavier¹  | Joris A. Veltman¹ 

¹Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK

²Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, The Netherlands

³Department of Obstetrics and Gynecology, Division of Reproductive Medicine, Radboudumc, Nijmegen, The Netherlands

Correspondence

Joris A. Veltman, Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.

Email: joris.veltman@newcastle.ac.uk

Funding information

Wellcome Trust; Royal Society and Wolfson Foundation award

Abstract

De novo mutations (DNMs) play an important role in severe genetic disorders that reduce fitness. To better understand their role in disease, it is important to determine the parent-of-origin and timing of mutational events that give rise to these mutations, especially in sex-specific developmental disorders such as male infertility. However, currently available short-read sequencing approaches are not ideally suited for phasing, as this requires long continuous DNA strands that span both the DNM and one or more informative single-nucleotide polymorphisms. To overcome these challenges, we optimized and implemented a multiplexed long-read sequencing approach using Oxford Nanopore technologies MinION platform. We focused on improving target amplification, integrating long-read sequenced data with high-quality short-read sequence data, and developing an anchored phasing computational method. This approach handled the inherent phasing challenges of long-range target amplification and the normal accumulation of sequencing error associated with long-read sequencing. In total, 77 of 109 DNMs (71%) were successfully phased and parent-of-origin identified. The majority of phased DNMs were prezygotic (90%), the accuracy of which is highlighted by an average mutant allele frequency of 49.6% and standard error of 0.84%. This study demonstrates the benefits of employing an integrated short-read and long-read sequencing approach for large-scale DNM phasing.

KEYWORDS

DNM, long-read sequencing, parent-of-origin, phasing, targeted sequencing, WES

1 | INTRODUCTION

De novo mutations (DNMs) arise from mutational events that occur during gametogenesis in either parent germ cells, or postzygotically in both somatic and germ cells of the individual carrying them. On average, one to two DNMs can be found in the coding region of a

person's genome (Durbin et al., 2010; O'roak et al., 2011; Xu et al., 2011). DNMs are of particular significance due to their contribution to many diseases and genetic disorders, notably those affecting individual fitness such as intellectual disability and male infertility (Acuna-Hidalgo et al., 2016; Awadalla et al., 2010; Gilissen et al., 2014; Oud et al., 2022; Taylor et al., 2019; Veltman & Brunner,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Human Mutation* published by Wiley Periodicals LLC.

2012). It has been shown that approximately 80% of DNMs are of paternal origin (Goldmann et al., 2016; Kong et al., 2012; Oud et al., 2022; Yuen et al., 2016). A major factor known to contribute to an increase in DNMs in individuals is advanced parental age at the time of conception, particularly paternal age (Goldmann et al., 2016; Kong et al., 2012). Investigating the parental origin and timing of DNMs provides not only biological insight into the generation and ability of these DNMs to underlie genetic disorders, it has also been shown to be important for determining the recurrence risk of these disorders (Almobarak et al., 2021; Campbell et al., 2014).

Phasing analysis interrogates the diploid genome, allowing allele separation of the parental chromosomes. This helps not only to determine the parental origin and timing of DNMs, but is also critical to identify compound heterozygous mutations and look into allele-specific expression, linked variants, and structural variation (Ebert et al., 2021; Soifer et al., 2020; Tewhey et al., 2011). With short-read whole genome sequencing (WGS) of parent-offspring trios, 15%–20% of DNMs can be successfully phased and parent-of-origin called (Goldmann et al., 2016). However, this percentage is expected to be even lower in whole exome sequencing (WES). Phasing challenges can be attributed to the limited sequencing read lengths, the presence of intronic gaps, and the reduced amount of genetic variation in the exonic regions compared to intronic regions (Frigola et al., 2017). By definition, germline DNMs need to be absent in the parental somatic cells, requiring trio-based exome or genome sequencing of parent-offspring trios for discovery. In the next step, the parent-of-origin and zygosity of a DNM can be identified by targeted amplification and long-read sequencing of a region spanning the DNM as well as one or more parentally informative single-nucleotide polymorphism (iSNP). While this appears straightforward, long-read sequencing has both random and positional error, which may result in false variants used for phasing, reducing the reliability of downstream analysis (Magi et al., 2018; Watson & Warr, 2019).

There are numerous methodologies to target genomic regions for enrichment before sequencing, with the majority being divided into PCR- or CRISPR-based approaches (Gilpatrick et al., 2020; Hafford-Tear et al., 2019; Player et al., 2020), in addition, a novel “adaptive sampling” method has recently been developed (Loose et al., 2016; Payne et al., 2021). Adaptive sampling takes advantage of the real-time individual pore control, sequencing, and basecalling of the Oxford nanopore technologies (ONT) platform to selectively sequence DNA. Briefly, adaptive sampling uses a known library to interpret each DNA molecule as it is ratcheted through a protein pore, reversing the polarity through the pore, and rejecting the molecule if it fails to match its reference library. Comparative to PCR enrichment, adaptive sampling or “ReadUntil” is limited in coverage, with reduced flow cell yields of up to 90% (S. Martin et al., 2022). This results in greater consumable costs, which would be additional to the greater computational demand of the approach. Similarly, when mapping sequence data to the reference genome from CRISPR targeting approaches, the off-target mapping of the sequences is several fold greater than PCR-based methods and target coverage is therefore often many factors lower (Hafford-Tear et al., 2019;

McDonald et al., 2021), with costs per target significantly higher. Innate challenges also exist with PCR approaches, including the presence of inhibitors, variable target length, optimisation time, amplification bias, and nucleotide errors (Potapov & Ong, 2017; Shagin et al., 2017). However, despite these challenges with long-range PCR enrichment, the approach is arguably more effective for scaled-up targeted phasing at present.

This study aims to identify the DNM parent-of-origin and zygosity using a targeted long-range and standard PCR approach for phasing 109 distinct DNMs previously identified in infertile men by patient–parent trio exome sequencing (Oud et al., 2022). Targeted amplification of regions encompassing each unique DNM is performed using an optimized PCR workflow designed to quickly increase PCR success rates and reduce presequencing base error from standard and long-range PCR. The combination of exome patient–parent trio data, targeted ONT sequencing, and validated DNMs are used to improve phasing and allele frequency confidence. Critical aspects of the process are assessed to ascertain practical application for large-scale use of the approach, including amplification length, long-read sequencing error rates, and overall phasing performance.

2 | METHODS

2.1 | Ethical considerations and study participants' consent

Seventy-seven infertile male patients with unexplained (idiopathic) nonobstructive azoospermia or severe to extreme oligozoospermia and their unaffected parents were enrolled from the Radboudumc outpatient clinic and Newcastle upon Tyne Hospitals NHS Foundation Trust (Newcastle, UK). This study was approved by the ethics Committees/Institutional Review Boards (Nijmegen: NL50495.091.14 version 5.0, Newcastle: REC Ref: 18/NE/0089). Written informed consent from all patients and their parents was obtained.

2.2 | WES

Genomic DNA was extracted from the whole blood for all patients at the time of evaluation and treatment at the fertility center. Parental DNA was obtained from saliva using the Oragene OG-500 kit (DNA Genotek). Exomes of patient–parent trios were prepared for sequencing using 1 µg of high-quality genomic DNA, which was quantified with the Qubit dsDNA HS kit (Thermo Fisher Scientific). Whole exome enrichment was performed using Illumina's Nextera DNA Exome Capture kit (Illumina) or Twist Bioscience's Twist Human Core Exome Kit, as per manufacturer protocols. Individual sample libraries were then indexed with Illumina Nextera DNA UD indexes (Illumina) before pooling libraries together. The final pooled library was sequenced on the NovaSeq 6000 platform (Illumina) to an average coverage depth of 72X (Illumina Nextera kit) or 99X (Twist Bioscience's kit).

2.3 | DNM/iSNP calling and DNM validation

Sequence reads were aligned to the Human Genome Reference Assembly GCRh37 using Burrows-Wheeler Alignment (BWA) version 0.7.17 (Li, 2013) and indexed using SAMtools (Li et al., 2009) and Picard version 1.6 (PicardToolkit, 2019). SNVs and indels were subsequently called by the Genome Analysis Toolkit HaplotypeCaller v4.1.4.1 (McKenna et al., 2010). De novo analysis was performed using custom-designed in-house analysis pipelines and validation of DNMs in the proband and parents was performed using standard Sanger Sequencing (Oud et al., 2022). iSNP calling was also performed using an in-house pipeline, which called proband and parent variants using Clair v2 (Luo et al., 2020), proband variants were selected based on unique parental inheritance with a minimum parental coverage of 5×. A single iSNP is selected for parent-of-origin assignment, see Section 2.7.

2.4 | Target amplification

In 77 infertile males analyzed, 109 rare protein-coding DNMs were identified and PCR amplified targeted phasing was performed, for a basic schematic of the methods see Supporting Information: Figure S1. Irrespective of known iSNP presence, primers were designed for the target regions (determined as per Figure 1b) of all 109 DNMs (Supporting Information: Table S1) using our in-house pipeline which designs primers using Primer3 version 2.3.6 (Untergasser et al., 2012) and Human Genome Reference Assembly GCRh37. Automated rounds of primer creation were run for each target, selecting the most optimal primer pair based on predicted fragment size, self-binding, hairpin factors, reverse primer to forward primer binding, and nonspecific genome binding. All expected fragment sizes were limited to a maximum length of 10 kb for greater enrichment success and speed,

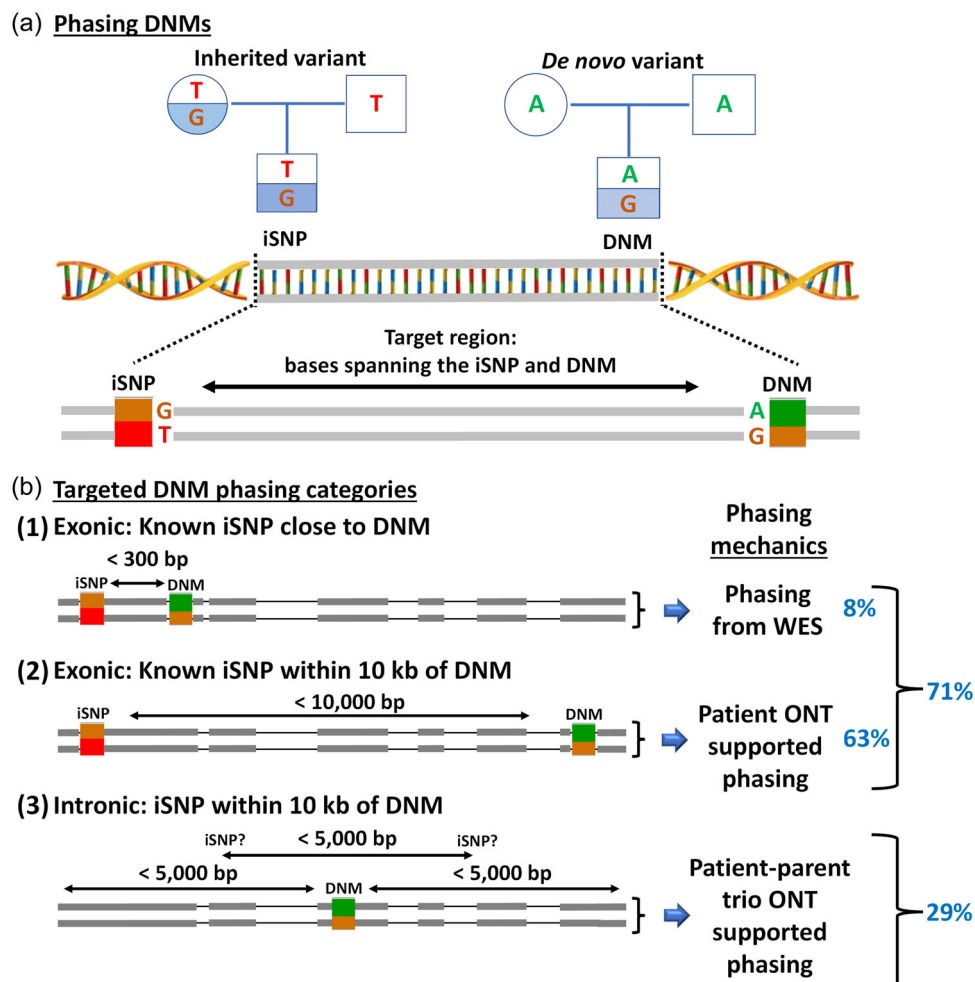


FIGURE 1 (a) Pedigree chart illustrating the bases of an iSNP and a DNM. To phase a DNM and determine the parent-of-origin, single reads must span from the DNM to the iSNP without fragmentation. The distance between these two points of interest dictates the feasibility of phasing the target region with short or long-read sequencers. PCR amplification followed by long-read sequencing allows target read lengths of up to 10,000 bp. (b) Illustration of the three categories in which targets are grouped and phased. The gray boxes upon the thin line in each category (1, 2, and 3) represent existing exome read coverage. Each category is based on the distance between the iSNP and DNM, which determines whether it can be phased with only the existing short-read WES trio data (1), WES and long-read proband data (2), or WES and long-read trio data (3). The percentage of samples that fit each category is highlighted in blue. DNM, de novo mutation; iSNP, informative single-nucleotide polymorphism; WES, whole exome sequencing.

though optimisation steps were required for 50% of the targets (Supporting Information: Figure S2).

DNMs without exome-supported iSNPs within a 10 kb region were selected for intronic iSNP searching within 5 kb (Figure 1b-(3)). For this group primers were designed in 5 kb windows around the DNM. iSNPs from the ~5 kb region were identified in the introns of parental ONT data in the downstream bioinformatics. Standard and long-range PCR target enrichment were carried out using optimized running conditions of two separate supermixes/enzymes, BioRad iproof and TakaraBio PrimeSTAR GLX (Supporting Information: Table S2). Primary steps use dilution to reduce contaminants and inhibitors, where dilution was not enough a hotstart supermix (Quantabio RepliQa HiFi ToughMix) was used which reduces the need for annealing temperature specificity (Supporting Information: Table S2). The number of primer pairs that required bespoke optimisation was reduced to just 12% of the samples by using broad and basic steps as per the Supporting Information: Figure S2b. Only 5% of the primer pairs required redesigning, the rest were successfully optimized across the first 5 stages in Supporting Information: Figure S1b. Supporting Information: Figure S1b steps were chosen based on simple and effective PCR optimisation that could be easily applied on a broad scale for a large number of samples. These measures gave a PCR success rate of >80%, reducing the time required for primer-specific optimisation and/or redesign. Sample fragment sizes were confirmed using gel electrophoresis, and quantities were measured with the Qubit dsDNA HS kit (Thermo Fisher Scientific), with the best quality supermix enrichment for each given sample/target selected for sequencing.

2.5 | Optimising scalability

On average 30 targets per MinION run were amplified, pooled, and sequenced. Initially, a limited number of 11 samples were barcoded and sequenced, but a reduction in usable data was noted (Supporting Information: Table S3). As such, this approach was replaced with careful sample pooling instead. The 11 samples were not resequenced, as the small number meant there was still ample data per target (mean of 19 million bases per target). Targets were checked to ensure no overlap in genome location, and demultiplexing was performed based on region extraction of the mapped BAM files (see Section 2.7). Using this approach, a loss of 17.8% (SEM 7.2) of the total basecalled yield was avoided by bypassing barcode ligations, demultiplexing binning issues, and barcode trimming (Supporting Information: Table S3).

2.6 | Long-read sequencing

The PCR target enrichments of >20 ng were prepared for sequencing with the ONT ligation sequencing kit (SQK-LSK109) following the manufacturer's protocol, with adjustments for sample type and yield.

Individual sample libraries with <20 ng were concentrated at given bead clean-up steps. For sample pooling steps, samples were pooled in groups based on fragment size, to allow for more accurate normalisation before final pooling and flow cell loading. Prepared samples were sequenced on the MinION using the FLO-MIN106 version 9.4.1 flow cell platform. Flow cells were run until complete pore exhaustion, on average this was 72 h, with minimal refuel of flow cells performed whenever active pore percentages dropped below 70%, achieving up to 30 billion basecall yields per flow cell and an average coverage depth per sample of >35,000×.

2.7 | Bioinformatics: Basecalling, fastq clean up, and demultiplexing

The sequence signal data in multi-fast5 format were basecalled using Guppy version 3.4.4. The resulting fastq outputs were adapter trimmed, low-quality reads ends trimmed (-q 10), and short reads of <30 base pairs removed using cutadapt version 2.5 (M. Martin, 2011). Cleaned fastq files were mapped against Human Genome Reference Assembly GCRh37 using BWA-MEM (Li, 2013) (version 0.7.17), and sample targets were extracted from the resulting BAM file using SAMtools (Li et al., 2009) (version 0.1.19).

2.8 | Bioinformatics: Variant calling

Variants were called using Clair v2 (Luo et al., 2020) and filter parameters for variants required >20× coverage, allele frequency 20%–80%, 96% confidence, and allele agreement with the validated DNM. The remaining variants were error-polished using the high base accurate WES data, removing variants that disagree with any overlapping WES data. WES data for patients and parents were used in filtering ONT called variants illustrated in Figure 2 and Supporting Information: Tables S4 and S5. Intronic variants identified were further error-polished where the inheritance information was available from long-read parental data.

2.9 | Bioinformatics: Phasing

To identify reliable variants, reads were split based on DNM base type, and the remaining variants were checked for agreement with the split reads using bamql (Masella et al., 2016). The iSNP with the greatest combinations of split read agreement, coverage, and additional supporting sequencing data was selected for final phasing, see Supporting Information: Table S6 for iSNP information. This preliminary variant sorting and iSNP selection approach provided a more reliable variant list. The next step split the raw reads by base type at DNM and selected iSNP positions using bamql. After this step, the allele frequencies, parent-of-origin, and timing of the DNM event (pre/postzygosity) were determined. The parent-of-origin conclusion was re-affirmed manually in Integrative Genomics Viewer

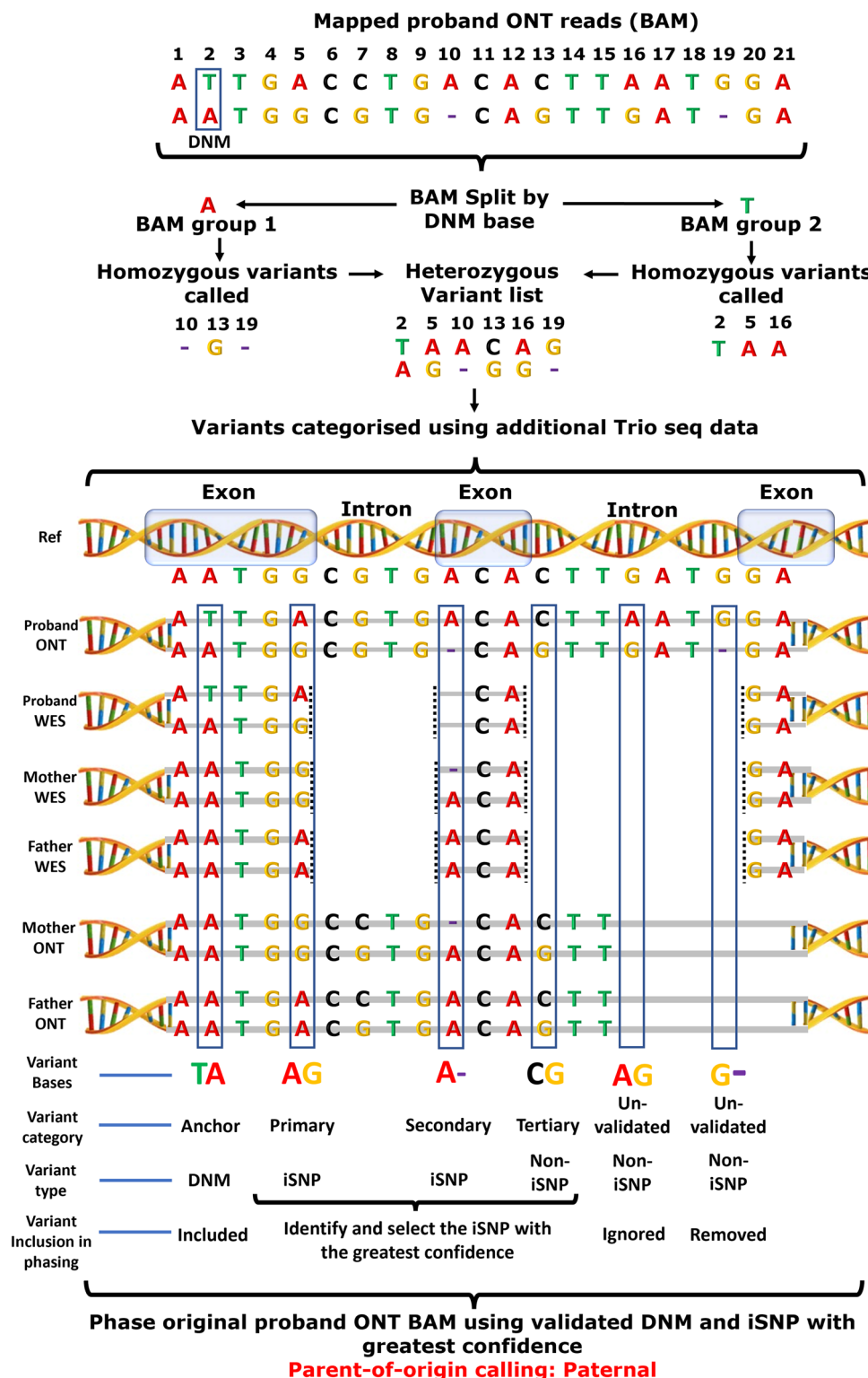


FIGURE 2 Using validated DNM for an anchored phasing pipeline. Pipeline overview for phasing, including illustration of variant strength and filtering. Reads are split by their Sanger-validated DNM nucleotide, then homozygous variants are identified in the two groups. These variants are compared to additional supporting sequencing data and prioritized for phasing based on their primary, secondary, tertiary, and unvalidated categories. These categories are based on the extent of the variant validation from WES and ONT Trio data. Small deletions that have no alternative supporting data are removed due to their high false call rate. An iSNP is selected based on the sequencing support category and used to phase the original proband ONT mapped reads, after which the allele frequencies, zygosity, and parent-of-origin are determined. DNM, de novo mutation; iSNP, informative single-nucleotide polymorphism.

(IGV) by visualizing the reads containing the DNM and the chosen iSNP from the pipeline.

Additional to manual analysis in IGV, postzygotic and prezygotic DNMs were determined by three category assessments, those that appeared postzygotic are detailed in Supporting Information: Table S7. The categories are; WES DNM base information (coverage/frequency), ONT DNM base information (coverage/frequency), and ONT allele information (coverage and allele frequencies). This information also helped determine background allele error, by presenting reads in the data that represent third and fourth allelic forms.

Allele and base error were calculated in two ways, one approach represented total error, which was calculated for a target DNM by combining all known false base frequencies for total base position error, and for total allele error, combining all known false allele frequencies. The other approach used the most prevalent base or allele frequency that could be determined as false. Assessment of error and noise in data helped support prezygotic and postzygotic DNM calls.

To provide an assessment of basecalling background error that is allele relevant, we ignored checking every base within every target as that would be unnecessarily extensive and time-consuming. Instead, we used the highest false base percentage of each iSNP used in phasing each target and calculated the mean false base percentage (Supporting Information: Table S8). Furthermore, the false base could be qualified as it had parental data support. A quality assessment of

the bases for each target within the BAM files was performed using Picard "QualityScoreDistributions" (PicardToolkit, 2019).

3 | RESULTS

WES of 77 infertile males and their unaffected parents identified 109 rare DNMs, all of which were independently validated by Sanger sequencing. Accurate phasing and parent-of-origin calling of the DNMs requires DNA molecules spanning a parentally iSNP and the DNM (Figure 1a). As such, the ability to call the parent-of-origin is primarily reliant on read lengths. This is notable in the WES data, where only 8% of DNMs could be phased as most iSNP were located >300 bp away from the DNM (Figure 1b-(1) and Table 1).

3.1 | Target amplification groups

For phasing, all DNM regions were targeted with either standard or long-range PCR, depending on required fragment size, and sequenced using ONT long-read sequencing. Primer pairs were designed for targeted long-read phasing of the 109 targets (Supporting Information: Table S1). The fragment size was capped at ~10 kb to simplify PCR optimisation. Unfortunately, amplification success was still impacted by

TABLE 1 Comparison of WES and targeted long-read phasing

Patient	Gene	WT > DNM	iSNP distance (bp)	DNM position - iSNP position (Based on ONT allele data)			WES				Long-read approach		Parent of Origin
				Paternal	Maternal	Proband	Proband DNM position bases	Proband iSNP position bases	Proband allele Coverage	WES Quality High*/Low	Proband DNM position bases	Proband allele Coverage	
02164	TBC1D7	C > T	97	C-C C-C	C-T C-C	T-C C-T	T (95/193 49%) C (98/193 51%)	C (68/162 42%) T (94/162 58%)	T-C 29 (41%) C-T 42 (59%)	High	T (57687/122202 47%) C (64515/122202 53%)	T-C 50516 (52%) C-T 46883 (48%)	Paternal
00052	TP63	G > A	82	G-A G-G	G-G G-G	A-A G-G	A (68/128 53%) G (60/128 47%)	A (25/61 41%) G (35/61 57%)	A-A 11 (50%) G-G 11 (50%)	High	A (39893/81843 49%) G (41950/81843 51%)	A-A 29200 (48%) G-G 31331 (52%)	Paternal
01247	c10orf71	C > T	54	C-A C-A	C-A C-G	T-G C-A	T (32/208 15%) C (176/208 85%)	A (46/114 40%) G (68/114 60%)	T-A 4 (9%) C-A 18 (41%) C-G 22 (50%)	High	T (6856/33394 21%) C (26538/33394 79%)	T-A 5135 (17%) C-A 10168 (33%) C-G 15604 (50%)	Maternal
01974	PLEKHA1	C > G	2	C-G C-G	C-A C-A	G-G C-A	G (73/162 45%) C (89/162 55%)	G (74/164 45%) A (90/164 55%)	G-G 73 (45%) C-A 88 (55%)	High	G (53781/113106 48%) C (59325/113106 52%)	G-G 50562 (48%) C-A 55876 (52%)	Paternal
01743	STXB2	C > T	23	C-T C-T	C-C C-T	T-T C-C	T (90/146 62%) C (56/146 38%)	T (101/175 58%) C (74/175 42%)	T-T 88 (62%) C-C 55 (38%)	High	T (3288/6267 52%) C (2979/6267 48%)	T-T 4680 (49%) C-C 4815 (51%)	Paternal
01889	USH2A	C > A	252	C-A C-A	C-G C-A	A-A C-G	A (57/107 53%) C (50/107 47%)	A (135/236 57%) G (101/236 43%)	A-A 17 (59%) C-G 12 (41%)	High	A (1499/2989 50%) C (1490/2989 50%)	A-A 2621 (58%) C-G 1935 (42%)	Paternal
01988	GRIP1	G > A	41	G-A G-A	G-G G-G	A-A G-G	A (24/51 47%) G (27/51 53%)	A (18/38 47%) G (20/38 53%)	A-A 11 (41%) G-G 16 (59%)	High	A (36021/78771 46%) G (42750/78771 54%)	A-A 32195 (48%) G-G 34902 (52%)	Paternal
02166	MKL1	G > A	46	G-T G-T	G-C G-T	A-T G-C	A (89/170 52%) G (81/170 48%)	T (128/240 53%) C (110/240 46%)	A-T 88 (52%) G-C 81 (48%)	High	A (62327/113736 55%) G (51409/113736 45%)	A-T 47252 (56%) G-C 36579 (44%)	Paternal
01691	SIGLEC10	C > T	274	C-A C-G	C-G C-G	T-A C-G	T (82/178,46%) C (96/178,54%)	A (61/132 46%) G (71/132 54%)	T-A 28 (38%) C-G 46 (62%)	High	T (1388/3359 41%) C (1971/3359 59%)	T-A 1786 (35%) C-G 3317 (65%)	Paternal
01862	HEL22	G > A	140	G-T G-C	G-C G-C	A-T G-C	A (73/146,50%) G (73/146,50%)	T (19/27 70%) C (8/27 30%)	A-T 3 (75%) G-C 1 (25%)	Low	A (6282/11514 55%) G (5174/11514 45%)	A-T 8684 (50%) G-C 8698 (50%)	Paternal
01511	RPA1	T > C	73	T-G T-A	T-A T-A	C-G T-A	C (33/84 39%) T (51/84 61%)	G (9/25 36%) A (16/25 64%)	0 (0%) T-A 8 (100%)	Low	C (2517/6357 40%) T (3791/6357 60%)	C-G 4339 (40%) T-A 6448 (60%)	Paternal
01209	SORCS2	G > A	100	G-G G-G	G-A G-G	A-G G-A	A (27/75 36%) G (48/75 64%)	G (3/7 43%) A (4/7 57%)	0 (0%) G-A 1 (100%)	Low	A (1587/13237 12%) G (11650/13237 88%)	A-G 966 (8%) G-G 2166 (18%) G-A 9084 (74%)	Paternal
01137	GPAT4	C > T	61	C-C C-C	C-T C-C	T-C C-T	T (54/110 49%) C (56/110 51%)	C (15/34 44%) T (19/34 56%)	T-C 7 (58%) C-T 5 (42%)	Low	T (21383/45575 47%) C (24192/45575 53%)	T-C 18824 (49%) C-T 19505 (51%)	Paternal
01982	MYH3	C > T	95	C-A C-G	C-G C-G	T-A C-G	T (54/127 43%) C (73/127 57%)	A (93/181 51%) G (88/181 49%)	T-A 1 (20%) C-G 4 (80%)	Low	T (25176/50630 50%) C (25454/50630 50%)	T-A 19644 (49%) C-G 20340 (51%)	Paternal

Note: All targets have at least 1 read covering the DNM and iSNP in the WES data. Quality of WES phasing is determined by the coverage cut-off of >10× coverage per allele. Blue bases and alleles represent wt, and red bases and alleles depict the mutant. The given iSNP base pair distance is in relation to the DNM position.

Abbreviations: DNM, de novo mutation; iSNP, informative single-nucleotide polymorphism; WES, whole exome sequencing; wt, wild-type.

*High: Individual allele coverage >10×.

target length within the 10 kb range, with ideal lengths found to be <4 kb (Supporting Information: Figure S2a). PCR optimisation steps were required for 50% of all targets (Supporting Information: Figure S2b). The amplification size had no impact on error or quality in base calls or allele assignment (Supporting Information: Figure S3 and Table S8). Importantly, for 71% of cases, an iSNP was identified in the available trio-based WES data within 10 kb of the DNM (Figure 1b-(1) and (2)). For this group of DNMs, phasing can be done by targeted long-read sequencing of the proband only, since the iSNP is already typed in patients and their parents.

Cases, where an iSNP could not be found in the coding region, had primers designed to cover 5 kb regions around the DNM position for parent and proband samples (Figure 1b-(3)). We chose the 5 kb region based on the analysis of 4344 DNMs identified in 53 whole genome sequenced individuals/children (Smits et al., 2022). This revealed the presence of at least 1 iSNP within 5 kb of any given DNM in 81% of cases (Supporting Information: Tables S9 and S10). In the end, we obtained long-read sequencing data with iSNPs for 77 out of 109 DNMs selected (71%, Supporting Information: Tables S9 and S11).

3.2 | Variants, filtering, error polishing, and phasing

PCR targeted long-read sequencing of the successfully phased 77 DNMs produced an average allele coverage of 35,430X, with a mean background noise of 4% at iSNP positions (see Supporting Information: Table S8 and Section 2.7 for background noise calculation). This increased coverage compared to alternative targeting methods, WES, or WGS is expected to help with error reduction and mosaicism detection (Wright et al., 2019). The DNMs, which were initially identified with short-read WES and validated with Sanger sequencing, were used to anchor the preliminary phasing of long reads. This anchoring approach groups long-reads by the base information at the DNM position. After variant calling was performed (Section 2.7), homozygous variants were removed and heterozygous variants were checked and filtered based on agreement with the DNM grouped reads. All remaining variants from the long-read sequencing approach were error-polished and filtered using WES and parental ONT data (Figure 2 and Supporting Information: Tables S4 and S5). Following this, iSNPs were identified from the remaining variants. The iSNP with the greatest confidence (coverage, supporting data, DNM allele agreement) was selected for phasing. When phasing the reads based on the DNM and selected iSNP, additional alleles were allowed for the DNM in case of a postzygotic event, but these were screened for credible biological relevance, that is, the DNM wild-type (wt) would have to match the iSNP of the DNM alt. Importantly, this anchored approach filtered out 10% more presumably falsely called variants in comparison to the standard filtering of variants based on quality criteria, sequencing coverage, and consensus (see Supporting Information: Table S4). Our data contain a high percentage of false variants, this is largely due to handling raw data rather than consensus polished data. We decided to use this raw data approach to increase sensitivity to detect mosaic mutations. Small indels made

up on average 25% of the false positives, and on average 94% of all indels detected in the long-read sequencing data were likely false positives (see Supporting Information: Table S5, and for an example of a false indel see Supporting Information: Figure S4). Because of this high error rate for indel calling, we decided to remove all indels without supporting data available, which is noted in the illustration of our approach (Figure 2). For simplicity, these indels include those that occur over homopolymer regions, which may exaggerate the error of nonhomopolymer indel calls, though recent “Guppy” basecaller updates have been suggested to reduce homopolymer error.

3.3 | Validation of the phasing approach using WES data

The parent-of-origin for the DNMs was initially assessed from the short-read WES data. Short-reads, 150 bp in length, encompassing DNM and iSNPs in the WES data were available for 14 of the 109 DNMs, but only 9 (8%) (Figure 1b-(1)) had acceptable coverage (>10× per allele, Table 1). These DNMs were also target amplified and sequenced using ONT long-read sequencing to provide a control cohort for our long-read phasing method. For all 9 DNMs with acceptable coverage in the WES data, the parent-of-origin assignment agreed between the two approaches and the DNM allele frequencies obtained were comparable (Table 1). For both short and long-read sequencing, 11 out of 14 DNMs showed allele frequencies around 50% ($\pm 10\%$), unlike confirmed mosaic mutations (Supporting Information: Table S7) no significant third allelic form was observed, indicating that these were germline DNM events. Of the three DNMs deviating from prezygotic allele frequencies, two were determined as postzygotic (SORCS2 and C10orf71), and one (SIGLEC10) may result from allelic sequencing bias (Supporting Information: Table S12). Alternatively, this could be a result of a copy number variant, however, we did not identify the expected change in coverage. ONT long-read sequencing yield for SIGLEC10 was significantly reduced compared to average ONT sequencing, and while both WES and ONT allele data showed DNM allele frequency deviations from 50%, no significant third allelic form was observed. In addition, the DNM base frequencies of both WES and ONT data were within the prezygotic range of $50\% \pm 10\%$.

For the DNM affecting the C10orf71 gene in patient 01247, the de novo mutated allele (T-A) had a much lower allele frequency of 9% and 17% when detected with both WES and ONT approaches, respectively, with similar percentages observed in DNM base frequencies. Apart from the wt allele and the DNM allele, we clearly observed a third allelic form that represented the wt version of the DNM allele. The presence of this third allelic form suggests that the DNM likely occurred as a postzygotic event. The postzygotic mutation is observed in both the WES and ONT data in this case. Postzygotic mutations can, however, be missed when WES has low coverage, as seen for the DNM in SORCS2 in patient 01209. The postzygotic DNM in SORCS2 presents an average discrepancy of 15% from prezygotic norms in the WES base frequencies. A third

allelic form is not shown in the WES allele data due to only having 1× coverage. Greater base discrepancies are observed in the ONT base frequencies, with an average deviation from prezygotic norms of 38%, and with several thousand times more coverage in the ONT data, a wt of the DNM allele is observed at 18%.

3.3.1 | WES versus ONT DNM allele frequency comparison and parent-of-origin

In total, 77 DNMs of 109 DNMs (71%) were phased and the parent-of-origin determined, with 64 of these (83%) being of paternal origin. Across the 77 phased DNMs there was a DNM allele frequency range of 6%–78% in the ONT data and 15%–71% in the WES DNM base data, clearly showing the presence of postzygotic and prezygotic mutations. From these 77 DNMs, our targeted long-read sequencing approach identified 61 clearly prezygotic DNMs, with a possible third allelic form found in the remaining 16 DNMs. The separation of pre and postzygotic DNMs can be established more clearly through a combined analysis of their base frequencies, allele frequencies, DNM wt allele frequency, and background sequencing error (see Supporting Information: Table S7). By doing this we confirmed 69 prezygotic DNMs and 8 postzygotic mutations, as depicted in Supporting Information: Tables S7 and S12, and Figures S5 and S6. On average a third of allelic data is identified as false in our ONT approach

(Supporting Information: Table S8), and that impacted the allele frequency significantly. After removal of these false allelic data, we noticed that the average ONT allele frequency for prezygotic DNMs was around 50% with an average of 49.6%, and a standard error margin (SEM) of 0.8% (Supporting Information: Table S13 and Figure S5). For germline DNMs, we expect the mutant allele frequency to be approximately 50%. While raw allele frequencies were expectedly much worse than DNM base frequencies, the polished allele frequencies are slightly more accurate compared to the WES DNM and ONT DNM prezygotic nucleotide base frequencies of 48.6% (SEM 1.0%) and 48.8% (SEM 0.8%), respectively (Supporting Information: Table S13). As expected, postzygotic frequencies of mutant alleles deviated significantly from this, with an average of 15.9% and a standard error margin of 2.4% in ONT allele data. Comparatively the frequencies of postzygotic mutations are less consistent in WES data, where the postzygotic frequency average was 28.8% (SEM 4.9%) (Supporting Information: Table S13).

The parent-of-origin was also investigated within the pre and postzygotic fractions, where we observe the prezygotic paternal preference of 83% and the postzygotic paternal origin of 62% (five out of eight), see Supporting Information: Figure S6. Classification of postzygotic DNMs was determined as per Supporting Information: Table S7 and a detailed IGV illustration is depicted in Figure 3. All postzygotic calls were corroborated by the DNM base frequencies in the short-read and long-read data, and further supported by all

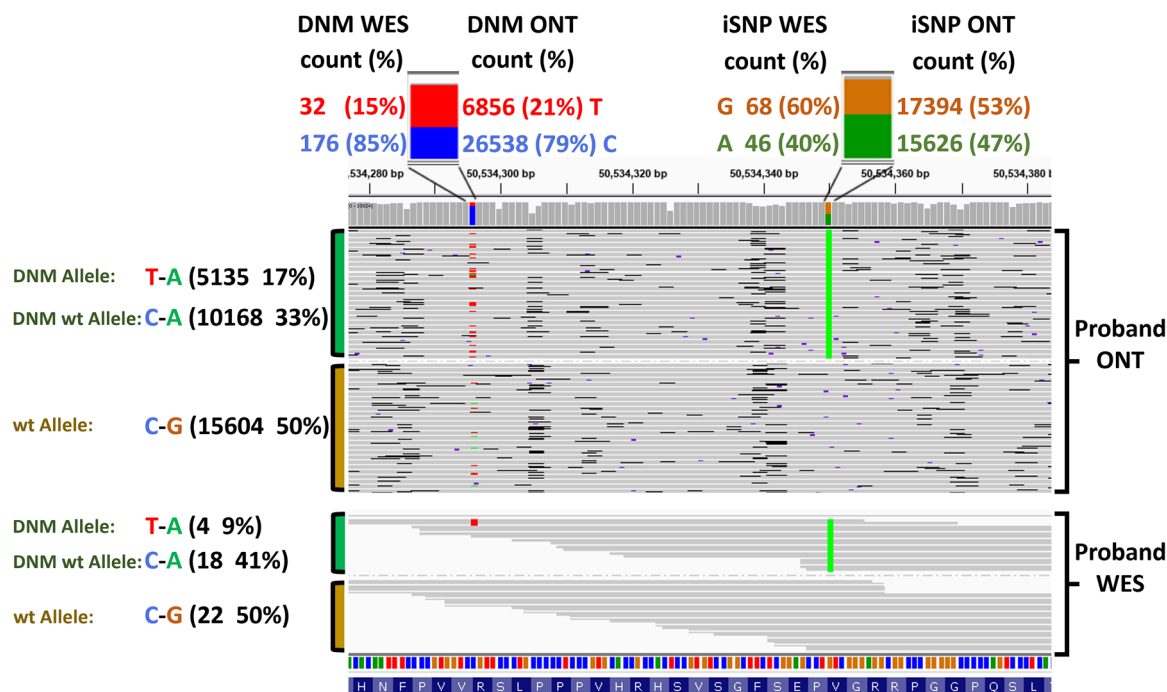


FIGURE 3 Example of the postzygotic DNM in C10orf71 visualized in IGV. The iSNP and DNM base associated (colored) counts and percentages are displayed at the top of the plot as a descriptive extension of the coverage bar. This postzygotic DNM is observed in the ONT and WES allele data and is supported by the discrepancies in the DNM positional base frequencies, while the iSNP base frequencies remain close to the expected 50:50 ratio. The IGV visual of the reads presents them grouped by iSNP base type. This read view illustrates a third allelic form in the reads that matches the base and iSNP combination expected, had the DNM not occurred (C-A 33% in ONT, 41% in WES). DNM, de novo mutation; iSNP, informative single-nucleotide polymorphism; WES, whole exome sequencing.

identified allele frequencies and the degree of error in target data (see Supporting Information: Tables S8 and S7).

4 | DISCUSSION

In this study, we optimized and applied a multiplexed long-read sequencing approach which makes use of high-quality short-read exome data to perform routine phasing of DNMs. We report the phasing results of 77 DNMs from 64 of our 77 patient–parent trios linked to male infertility, achieving successful phasing for 71% of the 109 DNMs investigated using this long-read targeted approach. In contrast, only 9 of these DNMs (8%) could be reliably phased based on short-read WES data alone.

Short-read exome sequencing has become an increasingly common tool in research and diagnostic of genetic disease, with the patient–parent trio-based sequencing routine for the detection of DNMs. With only 8% of DNMs being phasable in our cohort of 77 patients when using short-read WES alone, it is clear that an alternative approach is needed to determine the parent-of-origin and timing of DNMs. Our method uses standard and long-range PCR with standard optimisation steps to achieve the simplest and quickest large-scale success. PCR is a simple and common wet lab practice, providing greater enrichment and target specificity than any alternative target-based approach. The sequencing of target enriched long DNA strands with ONT allowed us, in most cases, to acquire 10s of thousands of times coverage per target, with many targets run per flow cell, supporting the projects scale demands. To overcome challenges with error and postzygotic mutations we used the WES data and Sanger-validated DNMs to polish the variant analysis, which limited computational demand as no complex algorithms were required, and processing could be quick.

DNMs are known to arise from mutational events occurring during gametogenesis, predominantly during spermatogenesis rather than during oogenesis. This is assumed to be associated to the scale of male gamete production and failure of DNA repair mechanisms which lead to the increased opportunity for mutational events to occur (Aitken & Baker, 2020; Evenson et al., 2020; Grégoire et al., 2013; Haldane, 1947; Kong et al., 2012). Previous literature has shown that DNMs occur on the paternal allele approximately 80% of the time (Goldmann et al., 2016; Kong et al., 2012; Yuen et al., 2016). In agreement with this literature, 83% of all phased DNMs in this study were determined to be of paternal origin.

Parent-of-origin and zygosity information adds another layer to our understanding of potential disease-causing variants. This is important when investigating genetic diseases, especially those that likely have complex and varied mechanisms. In our cohort of 77 patients, 51 patients were confirmed to suffer from nonobstructive severe oligospermic or azospermic phenotypes. In the original publication related to this study (Oud et al., 2022), we showed that six out of the eight likely causative DNMs identified in these patients were of paternal origin (Supporting Information: Tables S9 and S14, Figure S7). As DNMs predominantly cause a dominant effect, it

suggests that DNMs with a deleterious effect on the health of an individual can escape negative selection in spermatids from the paternal germline.

Accurate detection of the DNM allele frequency is critical to differentiate prezygotic from postzygotic mutational events, important in clinical settings for estimating the recurrence risk (Almobarak et al., 2021; Scanga et al., 2021). Our approach yielded a highly accurate allele frequency average of 49.6% in the prezygotic mutations, with an SEM of 0.84% (Supporting Information: Table S13). Though similar accuracy may be achievable with more computationally demanding methods, the strength of our method lies in utilizing the WES data and DNM validation practices commonly available. This shows that bioinformatic cleaning and more complex haplotype processing steps are unnecessary, with accurate results achievable through simple DNM and DNM-anchored iSNP selection. In total, 8 of the 77 phased DNMs were classified as postzygotic events (10%), largely in agreement with current literature results of 6.5%–10% (Acuna-Hidalgo et al., 2015; Sasani et al., 2019; Ye et al., 2018), supporting the validity of our method. Interestingly, while there was a significant correlation between WES and ONT postzygotic base/allele frequencies, 25% of the postzygotic DNMs could not be determined from WES DNM base frequencies. This demonstrates the importance of combining phasing analysis with deep coverage long-read sequencing to further characterize the timing of DNMs. As can be expected for postzygotic DNMs (Girard et al., 2016), we see less paternal bias even though our numbers are small (five out of eight postzygotic DNM are paternal, 62%).

We here use a standard PCR amplicon targeting approach with long-read sequencing, rather than CRISPR-Cas targeting or adaptive sampling. Despite CRISPR-Cas recently becoming a choice method for long-read targeted sequencing (Gilpatrick et al., 2020; Hafford-Tear et al., 2019; Liu et al., 2019; McDonald et al., 2021), and adaptive sampling being the latest technology, a large number of targets and small target sizes in our cohort would make either alternative complex and costly. Standard PCR targeting is optimal for the routine application that does not require methylation data, read lengths greater than 10–20 kb, or directly representative read counts (Aird et al., 2011). While the CRISPR-Cas approach can have a 10–100-fold enrichment of the target region compared to standard low coverage long-read WGS, it still results in 95.4% off-target sequencing (Gilpatrick et al., 2020). This off-target sequencing issue significantly limits the number of samples that can be run per flow cell, and only a single sample can be run if demultiplexing is based on the genomic position of the target. Similarly, adaptive sampling coverage and multiplexing is also restrictive comparatively to PCR, with only up to fourfold enrichment of the target compared to standard sequencing. The reverse is seen when comparing this to the standard amplicon approach used herein, where dozens of samples were run per flow cell and no off-target mapping was identified, and enrichment of the target was in fold factors >100. Based on using the optimal CRISPR-Cas approach of 2–3 gRNAs, and taking into account the reduced sample number per flow cell, CRISPR-Cas methods could have >40-fold increase in cost per target compared to PCR.

Nonetheless, CRISPR-Cas target enrichment shows great promise, and will likely be the best approach for targets larger than 10–20 kb. Comparative to PCR, adaptive sampling would also significantly increase costs due to the drop in flow cell yield and the specific computational infrastructure required. Despite this, adaptive sampling is an intriguing technology, with advantages over challenging genomic regions and swift targeted diagnostics.

No correlation was observed in basecalling error from PCR extension in targets of greater sizes. However, it is worth considering the potential increases in base error and bias from PCR approaches which would compound the lower accuracy inherent to long-read sequencing. Our data support the importance of minimizing target region sizes when performing PCR-based amplification for targeted sequencing, especially when performing primer optimisation for >100 bespoke primer pairs. Limiting target sizes will reduce labor-intensive PCR optimisation and though not observed in our study it may also reduce base error from PCR fidelity issues. We should, however, be mindful that for 11% of the DNMs studied no iSNPs were found within the 5 kb window, so minimizing the target region can also negatively impact phasing. For another 18% of DNMs, however, the sequencing data was of insufficient quality for phasing purposes, so clearly a balance must be found between sequencing quality and target size.

Since ONT released the MinION platform in 2014, there have been extensive leaps in advancing both the chemistry and the bioinformatic tools. This has resulted in raw base accuracy moving from as low as ~60% (Loman & Watson, 2015) to the current 92%–97% in the 9.4.1 flow cell chemistry used in this investigation. It should be noted that further increases in accuracy have also been suggested in recent flow cell chemistry, such as the release this year of the R10.4.1 flow cell. Bioinformatic tools that include the variant caller “Clair,” used here, have also shown increased confidence in variant calls but are thought to be reaching their limit, with greater confidence requiring significant alternative algorithms or improvements in chemistry (Luo et al., 2020). Despite the bioinformatic improvements in base calling and variant calling, we observe that the accuracy of long-read data on long-range and standard PCR products still causes far greater false positives than WES short-read data. After filtering ONT variants by read depth and quality scores, our anchored approach filtered an additional 50% of the remaining variants on average. If false variants that were missed before our anchored filtering approach were included in the phasing process it is likely some targets would be phased incorrectly or not phased at all. Many phasing tools such as “whatshap” carry out phasing with the understanding that variants within the vcf file are correct, so the removal of false variants is important.

Our study provides an approach for accurately phasing and parent-of-origin calling DNMs in a set of 77 patients. To our knowledge, this is the first time that phasing of DNMs has been investigated on this scale using long and standard range PCR targeted ONT sequencing, where each sample has a uniquely specific target. We optimized the method for efficiency and streamlined the laboratory and computational pipelines for processing large numbers

of DNMs for detailed phasing analysis. We incorporate additional short-read sequencing patient–parent trio data and Sanger-validated DNMs that are commonly available from DNM discovery pipelines like ours. This approach enabled us to improve DNM phasing and postzygotic calling. This data-supported and anchored phasing approach can be of great use in both research and diagnostic settings where DNMs are routinely studied and interpreted.

ACKNOWLEDGMENTS

We appreciate the participation of all patients and their parents in this study. We thank Arron Scott and Bryan Hepworth (Newcastle University) for technical support, the genomics core facility for whole exome sequencing (WES), the Bioinformatic support unit (BSU) for initial WES processing. This project was funded by an Investigator Award in Science from the Wellcome Trust (209451) and a Royal Society and Wolfson Foundation award (WM160091) to J. A. V.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Giles S. Holt  <http://orcid.org/0000-0002-7494-3736>

Manon S. Oud  <http://orcid.org/0000-0001-9513-3030>

Miguel J. Xavier  <http://orcid.org/0000-0003-0709-7223>

Joris A. Veltman  <http://orcid.org/0000-0002-3218-8250>

REFERENCES

- Acuna-Hidalgo, R., Bo, T., Kwint, M. P., van de Vorst, M., Pinelli, M., Veltman, J. A., Hoischen, A., Vissers, L. E., & Gilissen, C. (2015). Post-zygotic point mutations are an underrecognized source of de novo genomic variation. *The American Journal of Human Genetics*, 97(1), 67–74. <https://doi.org/10.1016/j.ajhg.2015.05.008>
- Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1), 241. <https://doi.org/10.1186/s13059-016-1110-1>
- Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2), R18. <https://doi.org/10.1186/gb-2011-12-2-r18>
- Aitken, R. J., & Baker, M. A. (2020). The role of genetics and oxidative stress in the etiology of male infertility—A unifying hypothesis? *Frontiers in Endocrinology*, 11, 581838. <https://doi.org/10.3389/fendo.2020.581838>
- Almobarak, S., Hu, J., Langdon, K., Ang, L., & Campbell, C. (2021). α -tropomyosin gene (TPM3) in an infant with Nemaline myopathy. *Clinical Case Reports*, 9(3), 1672–1676. <https://doi.org/10.1002/ccr3.3866>
- Awadalla, P., Gauthier, J., Myers, R. A., Casals, F., Hamdan, F. F., Griffing, A. R., Côté, M., Henrion, E., Spiegelman, D., Tarabeux, J., Piton, A., Yang, Y., Boyko, A., Bustamante, C., Xiong, L., Rapoport, J. L., Addington, A. M., DeLisi, J. L., Krebs, M. O., ... Rouleau, G. A. (2010). Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *American Journal of Human Genetics*, 87(3), 316–324. <https://doi.org/10.1016/j.ajhg.2010.07.019>
- Campbell, I. M., Stewart, J. R., James, R. A., Lupski, J. R., Stankiewicz, P., Olofsson, P., & Shaw, C. A. (2014). Parent of origin, mosaicism, and recurrence risk: Probabilistic modeling explains the broken

- symmetry of transmission genetics. *American Journal of Human Genetics*, 95(4), 345–359. <https://doi.org/10.1016/j.ajhg.2014.08.010>
- Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., ... Cartwright, R. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073. <https://doi.org/10.1038/nature09534>
- Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., ... Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), eabf7117. <https://doi.org/10.1126/science.abf7117>
- Evenson, D. P., Djira, G., Kasperon, K., & Christianson, J. (2020). Relationships between the age of 25,445 men attending infertility clinics and sperm chromatin structure assay (SCSA®) defined sperm DNA and chromatin integrity. *Fertility and Sterility*, 114(2), 311–320. <https://doi.org/10.1016/j.fertnstert.2020.03.028>
- Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, 49(12), 1684–1692. <https://doi.org/10.1038/ng.3991>
- Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B., Kleefstra, T., Brunner, H. G., ... Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509), 344–347. <https://doi.org/10.1038/nature13394>
- Gilpatrick, T., Lee, I., Graham, J. E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F. J., & Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology*, 38(4), 433–438. <https://doi.org/10.1038/s41587-020-0407-5>
- Girard, S. L., Bourassa, C. V., Lemieux Perreault, L.-P., Legault, M.-A., Barhdadi, A., Ambalavanan, A., Brendgen, M., Vitaro, F., Noreau, A., Dionne, G., Tremblay, R. E., Dion, P. A., Boivin, M., Dubé, M. P., & Rouleau, G. A. (2016). Paternal age explains a major portion of de novo germline mutation rate variability in healthy individuals. *PLoS One*, 11(10), e0164212.
- Goldmann, J. M., Wong, W. S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E., Hoischen, A., Roach, J. C., Vockley, J. G., Veltman, J. A., Solomon, B. D., Gilissen, C., & Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48(8), 935–939. <https://doi.org/10.1038/ng.3597>
- Grégoire, M.-C., Massonneau, J., Simard, O., Gouraud, A., Brazeau, M.-A., Arguin, M., Leduc, F., & Boissonneault, G. (2013). Male-driven de novo mutations in haploid germ cells. *Molecular Human Reproduction*, 19(8), 495–499. <https://doi.org/10.1093/molehr/gat022>
- Hafford-Tear, N. J., Tsai, Y.-C., Sadan, A. N., Sanchez-Pintado, B., Zarouchlioti, C., Maher, G. J., Liskova, P., Tuft, S. J., Hardcastle, A. J., Clark, T. A., & Davidson, A. E. (2019). CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. *Genetics in Medicine*, 21(9), 2092–2102. <https://doi.org/10.1038/s41436-019-0453-x>
- Haldane, J. B. S. (1947). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of Eugenics*, 13(4), 262–271. <https://doi.org/10.1111/j.1469-1809.1946.tb02367.x>
- Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., ... Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412), 471–475. <https://doi.org/10.1038/nature11396>
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics*, 1303, 3997.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Liu, G., Zhang, Y., & Zhang, T. (2019). Computational approaches for effective CRISPR guide RNA design and evaluation. *Computational and Structural Biotechnology Journal*, 18, 35–44. <https://doi.org/10.1016/j.csbj.2019.11.006>
- Loman, N. J., & Watson, M. (2015). Successful test launch for nanopore sequencing. *Nature Methods*, 12(4), 303–304. <https://doi.org/10.1038/nmeth.3327>
- Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nature Methods*, 13(9), 751–754. <https://doi.org/10.1038/nmeth.3930>
- Luo, R., Wong, C.-L., Wong, Y.-S., Tang, C.-I., Liu, C.-M., Leung, C.-M., & Lam, T.-W. (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*, 2, 1–8. <https://doi.org/10.1038/s42256-020-0167-4>
- Magi, A., Semeraro, R., Mingrino, A., Giusti, B., & D'Aurizio, R. (2018). Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in Bioinformatics*, 19(6), 1256–1272. <https://doi.org/10.1093/bib/bbx062>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17, 10–12.
- Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., & Leggett, R. M. (2022). Nanopore adaptive sampling: A tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*, 23(1), 11. <https://doi.org/10.1186/s13059-021-02582-x>
- Masella, A. P., Lalansingh, C. M., Sivasundaram, P., Fraser, M., Bristow, R. G., & Boutros, P. C. (2016). BAMQL: A query language for extracting reads from BAM files. *BMC Bioinformatics*, 17(1), 305. <https://doi.org/10.1186/s12859-016-1162-y>
- McDonald, T. L., Zhou, W., Castro, C. P., Mumm, C., Switzenberg, J. A., Mills, R. E., & Boyle, A. P. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nature Communications*, 12(1), 3586. <https://doi.org/10.1038/s41467-021-23918-y>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- O'roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., Karakoc, E., Mackenzie, A. P., Ng, S. B., Baker, C., Rieder, M. J., Nickerson, D. A., Bernier, R., Fisher, S. E., Shendure, J., & Eichler, E. E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), 585–589. <https://doi.org/10.1038/ng.835>
- Oud, M. S., Smits, R. M., Smith, H. E., Mastrorosa, F. K., Holt, G. S., Houston, B. J., de Vries, P. F., Alobaidi, B., Batty, L. E., Ismail, H., Greenwood, J., Sheth, H., Mikulasova, A., Astuti, G., Gilissen, C., McElenny, K., Turner, H., Coxhead, J., Cockell, S., ... Veltman, J. A. (2022). A de novo paradigm for male infertility. *Nature Communications*, 13(1), 154. <https://doi.org/10.1038/s41467-021-27132-8>

- Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*, 39(4), 442–450. <https://doi.org/10.1038/s41587-020-00746-x>
- PicardToolkit. (2019). Picard Toolkit. *Broad Institute, GitHub Repository*. Retrieved March 13, 2020, from <http://broadinstitute.github.io/picard/>
- Player, R., Verratti, K., Staab, A., Bradburne, C., Grady, S., Goodwin, B., & Sozhamannan, S. (2020). Comparison of the performance of an amplicon sequencing assay based on Oxford Nanopore technology to real-time PCR assays for detecting bacterial biodefense pathogens. *BMC Genomics*, 21(1), 166. <https://doi.org/10.1186/s12864-020-6557-5>
- Potapov, V., & Ong, J. L. (2017). Examining sources of error in PCR by single-molecule sequencing. *PLoS One*, 12(1), e0169774. <https://doi.org/10.1371/journal.pone.0169774>
- Sasani, T. A., Pedersen, B. S., Gao, Z., Baird, L., Przeworski, M., Jorde, L. B., & Quinlan, A. R. (2019). Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife*, 8, e46922. <https://doi.org/10.1101/552117>
- Scanga, H. I., Liasis, A., Pihlblad, M. S., & Nischal, K. K. (2021). NYX-related congenital stationary night blindness in two siblings due to probable maternal germline mosaicism. *Ophthalmic Genetics*, 42(5), 588–592. <https://doi.org/10.1080/13816810.2021.1941129>
- Shagin, D. A., Shagina, I. A., Zaretsky, A. R., Barsova, E. V., Kelmanson, I. V., Lukyanov, S., Chudakov, D. M., & Shugay, M. (2017). A high-throughput assay for quantitative measurement of PCR errors. *Scientific Reports*, 7(1), 2718. <https://doi.org/10.1038/s41598-017-02727-8>
- Smits, R. M., Xavier, M. J., Oud, M. S., Astuti, G. D. N., Meijerink, A. M., de Vries, P. F., Holt, G. S., Alobaidi, B. K. S., Batty, L. E., Khazeeva, G., Sablauskas, K., Vissers, L. E. L. M., Gilissen, C., Fleischer, K., Braat, D. D. M., Ramos, L., & Veltman, J. A. (2022). De novo mutations in children born after medical assisted reproduction. *Human Reproduction*, 37, 1360–1369. <https://doi.org/10.1093/humrep/deac068>
- Soifer, L., Fong, N. L., Yi, N., Ireland, A. T., Lam, I., Sooknah, M., Paw, J. S., Peluso, P., Concepcion, G. T., Rank, D., Hastie, A. R., Jojic, V., Ruby, J. G., Botstein, D., & Roy, M. A. (2020). Fully phased sequence of a diploid human genome determined de novo from the DNA of a single individual. *G3 Genes/Genomes/Genetics*, 10(9), 2911–2925. <https://doi.org/10.1534/g3.119.400995>
- Taylor, J. L., Debost, J., Morton, S. U., Wigdor, E. M., Heyne, H. O., Lal, D., Howrigan, D. P., Bloemendal, A., Larsen, J. T., Kosmicki, J. A., Weiner, D. J., Homsy, J., Seidman, J. G., Seidman, C. E., Agerbo, E., McGrath, J. J., Mortensen, P. B., Petersen, L., Daly, M. J., & Robinson, E. B. (2019). Paternal-age-related de novo mutations and risk for five disorders. *Nature Communications*, 10(1), 3043. <https://doi.org/10.1038/s41467-019-11039-6>
- Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., & Schork, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3), 215–223. <https://doi.org/10.1038/nrg2950>
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, 40(15), e115. <https://doi.org/10.1093/nar/gks596>
- Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8), 565–575. <https://doi.org/10.1038/nrg3241>
- Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology*, 37(2), 124–126. <https://doi.org/10.1038/s41587-018-0004-z>
- Wright, C. F., Prigmore, E., Rajan, D., Handsaker, J., McRae, J., Kaplanis, J., Fitzgerald, T. W., FitzPatrick, D. R., Firth, H. V., & Hurles, M. E. (2019). Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nature Communications*, 10(1), 2985. <https://doi.org/10.1038/s41467-019-11059-2>
- Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J. A., & Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genetics*, 43(9), 864–868. <https://doi.org/10.1038/ng.902>
- Ye, A. Y., Dou, Y., Yang, X., Wang, S., Huang, A. Y., & Wei, L. (2018). A model for postzygotic mosaicism quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. *Genome Research*, 28(7), 943–951. <https://doi.org/10.1101/gr.230003.117>
- Yuen, R. K., Merico, D., Cao, H., Pellicchia, G., Alipanahi, B., Thiruvahindrapuram, B., Tong, X., Sun, Y., Cao, D., Zhang, T., Wu, X., Jin, X., Zhou, Z., Liu, X., Nalpathamkalam, T., Walker, S., Howe, J. L., Wang, Z., MacDonald, J. R., ... Scherer, S. W. (2016). Genome-wide characteristics of de novo mutations in autism. *NPJ Genomic Medicine*, 1, 160271–1602710. <https://doi.org/10.1038/npgenmed.2016.27>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Holt, G. S., Batty, L. E., Alobaidi, B. K. S., Smith, H. E., Oud, M. S., Ramos, L., Xavier, M. J., & Veltman, J. A. (2022). Phasing of de novo mutations using a scaled-up multiple amplicon long-read sequencing approach. *Human Mutation*, 43, 1545–1556. <https://doi.org/10.1002/humu.24450>