LI LABORATORY INVESTIGATION

journal homepage: https://laboratoryinvestigation.org/

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Research Article

Haplotyping Using Long-Range PCR and Nanopore Sequencing to Phase Variants: Lessons Learned From the *ABCA4* Locus

Benjamin McClinton[a], Christopher M. Watson[a,b], Laura A. Crinnion[a,b], Martin McKibbin[a,c], Manir Ali[a], Chris F. Inglehearn[a], Carmel Toomes[a,*]

[a] Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK; [b] North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK; [c] Department of Ophthalmology, St. James's University Hospital, Leeds, UK

ARTICLE INFO

Article history:
Received 21 September 2022
Revised 11 January 2023
Accepted 17 April 2023

Keywords:
ABCA4
allele phasing
haplotype formation
long-range PCR
long-read sequencing
nanopore

ABSTRACT

Short-read next-generation sequencing has revolutionized our ability to identify variants underlying inherited diseases; however, it does not allow the phasing of variants to clarify their diagnostic interpretation. The advent of widespread, increasingly accurate long-read sequencing has opened up new applications not currently available through short-read next-generation sequencing. One such use is the ability to phase variants to clarify their diagnostic interpretation and to investigate the increasingly prevalent role of *cis*-acting variants in the pathogenesis of the inherited disease, so-called complex alleles. Complex alleles are becoming an increasingly prevalent part of the study of genes associated with inherited diseases, for example, in *ABCA4*-related diseases. We sought to establish a cost-effective method to phase contiguous segments of the 130-kb *ABCA4* locus by long-read sequencing of overlapping amplification products. Using the comprehensively characterized CEPH sample, NA12878, we verified the accuracy and robustness of our assay. However, in-field assessment of its utility using clinical test cases was hampered by the paucity and distribution of identified variants and by PCR chimerism, particularly where the number of PCR cycles was high. Despite this, we were able to construct robust phase blocks of up to 94.9 kb, representing 73% of the *ABCA4* locus. We conclude that, although haplotype analysis of variants located within discrete amplification products was robust and informative, the stitching together of larger phase blocks using overlapping single-molecule reads remained practically challenging.

Introduction

The ubiquitous adoption of short-read next-generation sequencing has significantly expanded the scope and availability of molecular diagnostic screening. When genetic testing is performed, typically a more targeted sequencing approach is performed initially using a custom hybridization reagent before exome or more latterly genome sequencing is undertaken.[1-3] Despite improved capabilities to detect sequence variants, the interpretation of the phase for recessive alleles remains challenging in many cases. Compounding the issue of detecting variants of unknown significance, a growing number of conditions display complex genetic architecture where the phase of variants alters their pathogenicity.

[*] Corresponding author.
E-mail address: c.toomes@leeds.ac.uk (C. Toomes).

ELSEVIER
https://doi.org/10.1016/j.labinv.2023.100160

When biallelic variants are detected in genes associated with autosomal recessive disease, cascade testing will be performed on available family members to establish their phase and to test if the detected variant segregates with the presenting phenotype. If no family members are available, as is often the case for late-onset disorders, progress is restricted. If the variants are close together and have been observed in other families, then predictions of phase can be made.[4] If variants are private (and previously unreported), it is not possible to confirm whether or not they are in *cis* or in *trans* and this limits therapeutic options for the subjects.

To date, laboratories have had limited capabilities to establish contiguous haplotypes from detected variants. Previous methods to phase variants include TA-cloning, but only variants that are close together can be phased using this technique, and it is time-consuming.[5] Several wet-laboratory strategies have been proposed that use molecular "barcodes" to tag DNA fragments before short-read sequencing.[6,7] Such approaches have typically been discontinued or do not enable interrogation of specifically targeted loci (ie, it is necessary to perform whole genome sequencing (WGS)). By contrast, the development of long-read single-molecule sequencers, such as those manufactured by Oxford Nanopore Technologies (ONT) and Pacific Biosciences, are enabling the direct assessment of physically separated variants in single reads.

In addition to phasing biallelic recessive alleles, long-read sequencing promises to allow the phasing of complex alleles, which are becoming increasingly apparent in many inherited diseases. A complex allele is one where multiple variants in a haplotype act as a single allele with altered (generally increased) pathogenicity compared with any of the constituent variants in isolation. This has become an increasingly important aspect of the study of genes associated with the recessive disease.[8] The limited capability to reliably establish the phase of the detected variants hinders the identification and study of these complex alleles.

An exemplar gene for haplotype analysis is *ABCA4* (NM_000350.3). *ABCA4*-related disease is the most common cause of inherited retinal disease.[9] Correctly phasing, and therefore diagnosing, cases is becoming increasingly important for this gene as clinical trials for therapies are currently underway.[10] In addition to its clinical importance, *ABCA4* is known to harbor complex alleles. For example, the common *ABCA4* variant c.5603A>T (NM_000350.3) p.(Asn1868Ile) (whose gnomAD reported minor allele frequency is 0.0665) has been associated with *ABCA4*-related disease, at reduced penetrance, only when in *trans* with a severely pathogenic variant. However, when the c.5603A>T allele is identified in *cis* with the *ABCA4* variant c.2588G>C (NM_000350.3) p.(Gly863Ala), disease penetrance is increased.[4,11] Given the clinical importance of this gene and the need to further understand the role complex alleles play in its pathogenesis, it is essential that a method exists to phase alleles in the absence of additional family members.

Here, we report our experience of sequencing *ABCA4* using a target enrichment strategy in which overlapping long-range PCR products are subjected to nanopore sequencing on a Flongle Flow Cell. While haplotype analysis of variants located within discrete amplification products was robust and informative, the stitching together of larger phase blocks using overlapping single-molecule reads remained practically challenging for this locus. Despite this, this assay promises to be a useful method for phasing physically distant variants in *ABCA4*-related diseases and beyond.

Materials and Methods

All patients were diagnosed and recruited to the study at St James's University Hospital, Leeds, UK. Blood samples were collected from patients and family members after obtaining informed consent and adhering to the Declaration of Helsinki. Ethical approval was provided by the Leeds East Teaching Hospitals NHS Trust Research Ethics Committee (project no. 17/YH/0032). Genomic DNA was isolated by Yorkshire Regional Genetics from whole blood using standard protocols. For the control subject, NA12878, genomic DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) from a lymphoblastoid cell line (Coriell Institute). Comprehensively characterized control genomic DNA from individual NA12878 was used in conjunction with publicly available data sets for this sample. NA12878 platinum Variant Call Files from the Genome in a Bottle Benchmarking study[12] were obtained from: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/GRCh37/.

PCR primers were designed using Primer3 (http://primer3.ut.ee/) and validated using the University of California Santa Cruz *in silico* PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr). All primers were purified using the desalted method and purchased from Sigma-Aldrich.

Following the manufacturer's guidelines, long-range PCR amplification was performed using the SequalPrep Long PCR Kit (Invitrogen). Each long-range PCR reaction consisted of approximately 50 ng of genomic DNA, 2 μL of 10× SequalPrep reaction buffer (Invitrogen), 0.3 μL of 5U/μL SequalPrep long polymerase (Invitrogen), 0.4 μL of dimethyl sulfoxide (Invitrogen), 2 μL of 10× SequalPrep Enhancer A (Invitrogen), and 0.5 μL each 10 pmol/μL primer, made up to 20 μL with distilled water. Thermocycling was performed according to the manufacturer's instructions, initially at 35 cycles and later at 25 cycles. Annealing temperatures between 52 °C and 58 °C were used, details are available upon request. Amplification products were purified using an Agencourt AMPure XP (Beckman Coulter) bead clean-up. Equimolar amounts of purified PCR products for each case were pooled.

Nanopore sequencing libraries were prepared using the LSK109 ligation sequencing kit (ONT). Briefly, this began with an end-repair and nickase treatment reaction consisting of 1.75 μL Ultra II end prep reaction buffer (New England Biolabs [NEB]), 1.75-μL FFPE DNA repair buffer (NEB), 1.5-μL Ultra II end prep enzyme mix (NEB), 1-μL FFPE DNA repair mix (NEB), 13.5-μL nuclease-free water, and 10 μL of equimolar PCR products at (50 ng/μL). The reaction was incubated at 20 °C for 5 minutes and then 65 °C for 5 minutes. A further bead-based clean-up reaction was performed using AMPure XP beads, after which sequencing adaptors were ligated to the end-repaired PCR products. This reaction comprised 30 μL of PCR products, 12.5 μL of Ligation Buffer (ONT), 5 μL of Quick Ligase (NEB), and 5 μL of AMX Adapter Mix (ONT). The reaction was incubated at room temperature for 10 minutes, after which a final AMPure XP bead clean-up was performed with washes that used Long Fragment Buffer (ONT). The final library was eluted in 6 μL of Elution Buffer (ONT) following a 10-minute incubation at 37 °C. Each library was sequenced using a Flongle Flowcell (R.9.4.1) on a MinION (ONT).

Base calling to convert the raw data from FAST5 to FASTQ format was performed using Guppy version 5.0.16 (http://nanoporetech.com). Adapter sequences were trimmed from read ends using Porechop version 0.2.4 (https://github.com/rrwick/Porechop) before NanoFilt version 2.8.0 (https://github.com/wdecoster/nanofilt) was used to remove low-quality reads (-q 10)[13] and select reads that were between 3 and 12 kb in length (this step also removed the first 75-bps from each read which are typical of lower quality). Reads were aligned to the human reference genome (build hg19) using minimap2 version 2.22 (https://github.com/lh3/minimap2).[14] File manipulation, including SAM to BAM conversion and read sorting by genomic coordinating, was performed using Samtools version 1.9 (http://www.htslib.org/).[15] Variant calling was performed using Nanopolish version 0.13.2. (https://github.com/jts/nanopolish) with an allele fraction of 0.33; only single nucleotide variants were considered due to the relatively higher error rate of indels in nanopore data. Haplotypes were constructed using WhatsHap version 1.1.[16] To inspect haplotypes manually, read groups were isolated by segregating reads based on a single variant using the Jvarkit tool biostar214299 (http://lindenb.github.io/jvarkit/Biostar214299.html).[5] This tool enabled the assessment of the degree of chimeric read formation. All reads with a nonreference allele in each of the indicated amplicons were isolated. These were then compared with the most distal *in trans* allele; the proportion of reads having the in *cis* variant was taken as the proportion of chimeric reads. Aligned reads were visualized using the Integrated Genome Viewer version 2.7.2.[17] Sequencing metrics were generated using NanoStat version 1.5.0 (https://github.com/wdecoster/nanostat) and Samtools Flagstat version 1.9.

The phase blocks created by WhatsHap were curated by visually inspecting the overlaps between amplicons to ensure that they contained heterozygous variants, acknowledging that variant-lacking overlaps would initiate a new phase block.

Exemplar data sets were obtained from internal cases that had undergone WGS and from the Public Genomes Project (https://www.personalgenomes.org.uk/data).

Results

To facilitate the design of amplicons spanning the *ABCA4* locus, we assessed the density of polymorphic variants at the *ABCA4* locus using gnomAD version 2.1.1; 385 variants with a minor allele frequency of >0.10 were identified, with a theoretical distribution of 2.96 variants per kb. These data informed our design strategy, which used long-range PCR products with a minimum 2-kb overlap between amplicons and allowed us to position the amplicons so they captured the maximum variation in the overlapping region. We next pilot-tested our theoretical amplicon design *in silico* on a series of 10 data sets (2 in-house WGS and 8 from the Public Genomes Project). We observed that all cases would have at least one heterozygous variant per amplicon overlap. Furthermore, we observed that across these 10 pilot cases, the average number of variants per amplicon overlap ranged from 1 to 9.44. We also observed that when using the gnomAD version 2.1.1. data set, there was an average of 8.9 (minimum, 1; maximum, 18) variants per overlap.

The assay was established to enable variant detection across the entire 130.2-kb *ABCA4* locus (chr1: 94587651-94457435 hg19). Variant-containing overlaps and the uniformity of coverage across specific long-range amplification products for the control NA12878 is visible in Figure 1. We immediately observed that variant calling and the creation of parental haplotypes were hampered by the chimeric read formation within amplification products, measured as reads that are inconsistent with any established parental transmission of a given variant. To investigate this in detail, reads were segregated into respective read groups using the most proximal heterozygous variant in an amplification product. The proportion of chimeric products in the most distal heterozygous variant was assayed (Fig. 2). This allowed the proportion of chimeric reads (per individual amplicon) to be assayed. The observed chimeric reads (range, 12%-42%; mean 26%) were thought to be linked to chimeric PCR product formation, as has been previously observed.[18] To assess the effect of the number of rounds of PCR on chimeric read formation, we reduced the thermocycling count from 35- to 25-cycles and then again assayed minor allele fractions at the outermost heterozygous variants within each amplification product. The decrease in cycle count reduced the effect of chimerism by an average of 18% (range 2%-33%), an observation that was verified using the well-characterized GIAB sample (NA12878) (Table 1). At 25 cycles, the proportion of chimeric reads was low (range, 1%-13%; mean 7%). Despite the improvement in data quality, the reduction in cycle count significantly decreased the molarity of available amplification products from approximately 9 nM to approximately 1.3 nM per amplicon. However, the reduction in chimeric reads observed at a lower cycle was deemed sufficient to prevent a high number of chimeric reads confounding phase block formation.

Assayed single nucleotide variants were verified using the GIAB data set (case 1, NA12878). We compared our experimentally derived phase blocks with publicly available phased genotypes derived from parental inheritance data for this control. Our data set indicated that the *ABCA4* locus in case 1 could be resolved into

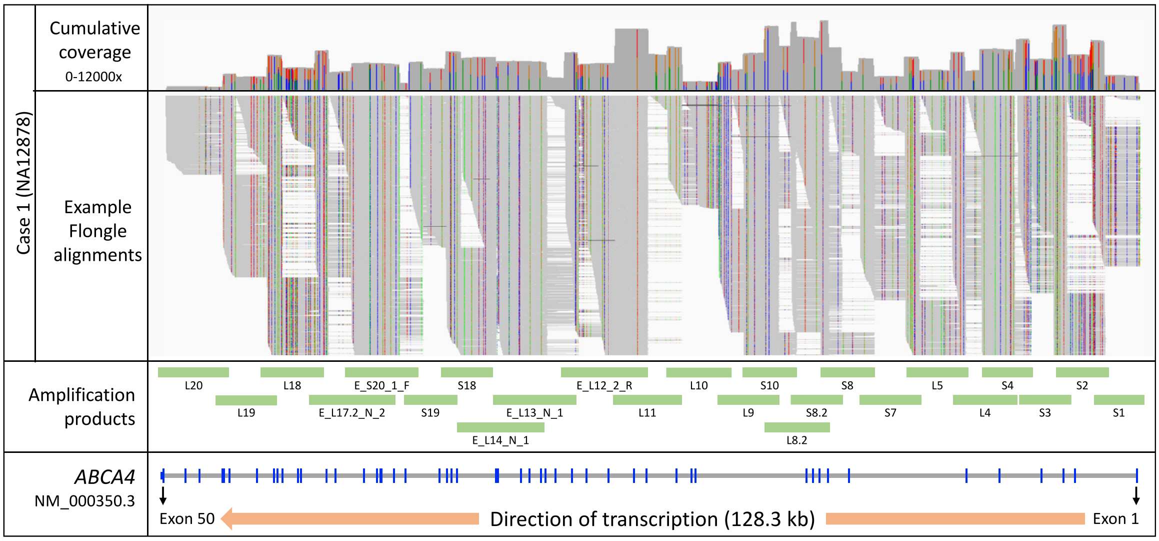

Figure 1.

Long-read alignments for case 1 (NA12878) showing the configuration of long-range amplification products (denoted in green) with respect to the *ABCA4* locus following 35 rounds of PCR amplification. Cumulative coverage is shown; the y-axis scale is 0-12,000×. Nonreference bases with an alternative allele fraction $\geq$ 0.3 are indicated.

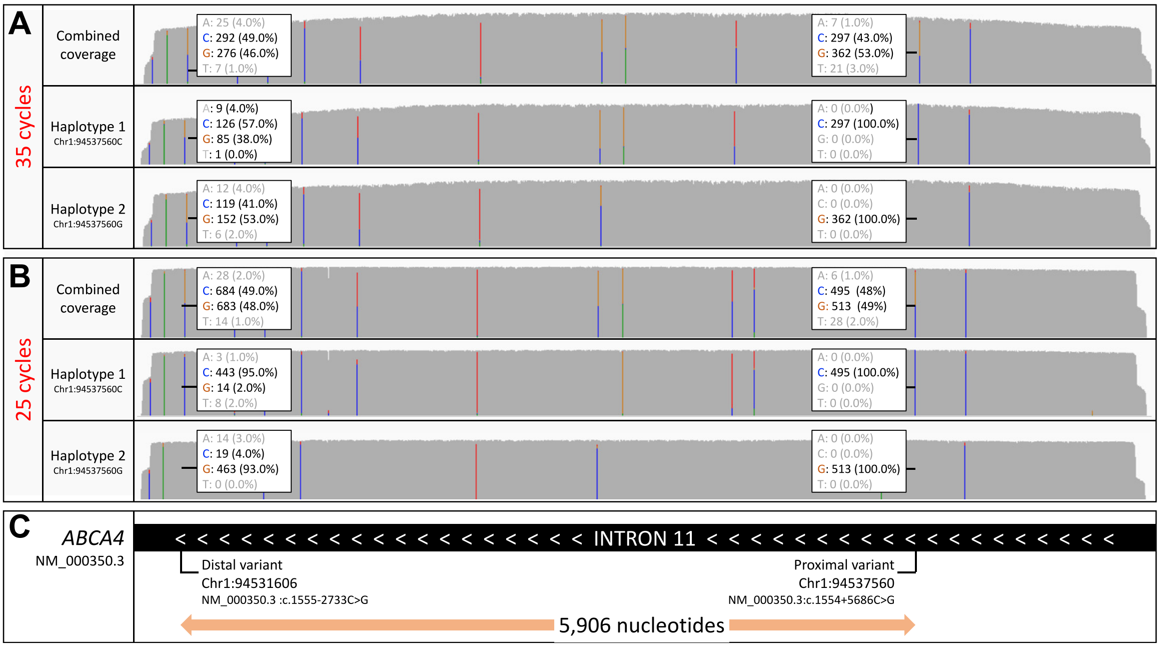

Figure 2.
Exemplar amplification products demonstrate the effect of reducing the chimeric read formation when the thermocycling count is reduced from (A) 35 cycles to (B) 25 cycles. Measurements of chimerism are calculated by selecting reference and nonreference reads at the most proximally located heterozygous variant, then interrogating the most distally located heterozygous variant. Note the difference between chimeric reads (an observed incorrect inheritance of parental genotypes) and the presence of sequencing errors, the latter occurring at a significantly lower rate. (C) The location of the amplification product with respect to *ABCA4*. Genomic coordinates are reported according to the human reference genome, build hg19.

a minimum of 2-phase blocks due to a lack of variants at one amplicon overlap. One of the 2 constructed phase blocks was 94.9 kb, spanning exons 1-30 and representing approximately 73% of the assayed locus. Within the constructed phase blocks, all phased genotypes were concordant with those present in the GIAB parental-phased data set (Supplementary Table S1).

Table 1
Monitoring the effect of cycle count on chimerism at the outermost heterozygous variants for each amplification product.

Amplicon name	Amplicon coordinates (chromosome: start-stop)	Amplicon size (bp)	Proximal variant (g.)	Distal variant (g.)	Proportion of chimeric reads			Yield of the amplification product	
					35× (%)	25× (%)	Reduction (%)	35× (nM)	25× (nM)
S1	chr1:94581079-94587651	6572	chr1:94587362G>T	chr1:94581258T>A	31	2	29	47.41	0.87
S2	chr1:94576013-94583016	7003	chr1:94582249T>G	chr1:94576360A>G	19	5	14	10.27	1.70
S3	chr1:94571091-94577994	6903	chr1:94577423C>T	chr1:94572434C>G	17	12	5	13.24	1.55
S4	chr1:94566272-94573027	6755	chr1:94572434C>T	chr1:94567223T>C	16	3	13	11.69	1.96
L4	chr1:94562394-94570893	8499	chr1:94570625G>A	chr1:94562924C>T	31	9	22	8.11	0.57
L5	chr1:94556161-94564368	8207	chr1:94562924C>T	chr1:94556894A>G	27	8	19	9.94	0.85
S7	chr1:94550053-94558218	8165	chr1:94557434T>C	chr1:94550715G>A	24	7	17	10.62	0.38
S8	chr1:94544899-94552183	7284	chr1:94550715G>A	chr1:94545160T>C	42	11	31	3.31	1.11
S8.2	chr1:94540974-94547825	6851						25.74	2.60
L8.2	chr1:94537456-94546134	8678	chr1:94545160C>T	chr1:94537560G>C	28	8	20	8.63	4.36
S10	chr1:94534599-94541790	7191	chr1:94540069G>C	chr1:94534980G>C	24	1	23	24.53	2.21
L9	chr1:94531257-94539449	8192	chr1:94537560G>C	chr1:94531606G>C	25	3	22	6.28	0.41
L10	chr1:94524472-94533111	8639	chr1:94532013C>T	chr1:94524784A>G	19	4	15	11.50	0.20
L11	chr1:94517468-94526499	9031	chr1:94526044A>G	chr1:94520451G>T	15	13	2	2.79	Too low
E_L12_2	chr1:94510646-94522068	11,422	chr1:94520451G>T	chr1:94512360G>A	23	11	12	3.77	5.15
E_L13	chr1:94501620-94512558	10,938	chr1:94512360G>A	chr1:94504545C>T	29	10	19	16.72	4.08
E_L14	chr1:94496875-94508372	11,497	chr1:94505971G>A	chr1:94498133G>A	33	12	21	5.70	Too low
S18	chr1:94494792-94501639	6847	chr1:94501594A>C	chr1:94495487G>A	28	8	20	16.85	4.66
S19	chr1:94489852-94496894	7042	chr1:94496253G>A	chr1:94492773C>T	19	4	15	30.10	0.50
E_S20	chr1:94482132-94491845	9713	chr1:94488326T>A	chr1:94486355A>G	12	1	11	12.96	4.86
E_L17.2	chr1:94477294-94488723	11,429	chr1:94488326T>A	chr1:94480037C>T	40	7	33	4.29	7.25
L18	chr1:94470930-94479249	8319	chr1:94478847G>A	chr1:94471154C>T	31	10	21	9.72	2.76
L19	chr1:94464973-94472982	8009	chr1:94472909G>A	chr1:94465132G>T	37	11	26	59.19	1.80
L20	chr1:94457435-94466789	9354	chr1:94466659A>G	chr1:94464553C>T	26	7	19		0.18

No variants were available within overlapping regions for amplicon S8.2. Genomic coordinates are according to the human reference genome hg19.
(g.), genomic nomenclature.

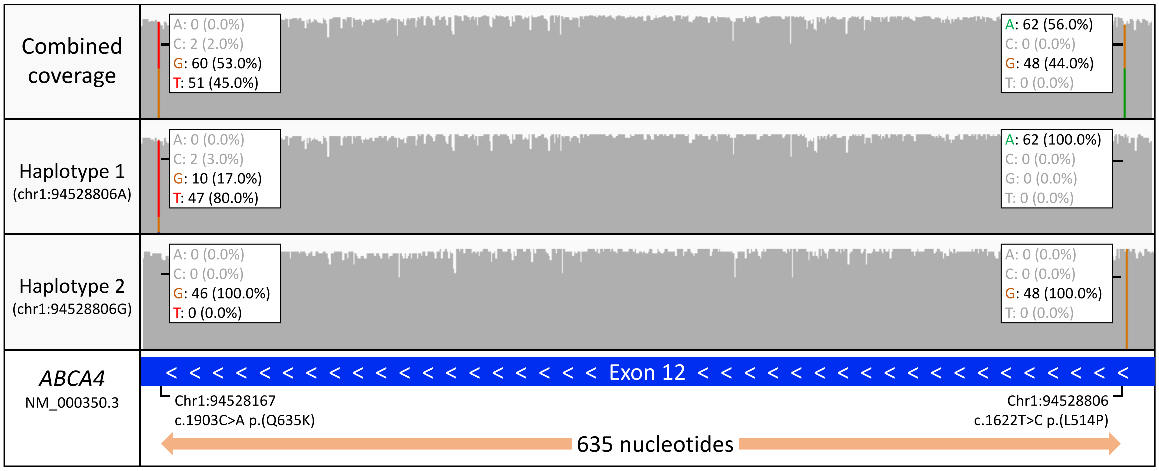

Figure 3.
Target variants c.1903C>A p.(Q635K) and c.1622T>C p.(L514P) (case 3, sample ID: 3515) have been phased into their corresponding haplotypes. Reads were separated according to the nucleotide sequence at the most proximally located target variant (chr1:94,528,806). Genomic coordinates are reported according to the human reference genome build hg19.

Having established the assay, we next sought to apply our method to a series of clinically relevant but hitherto unphased variants. We first tested this approach by selecting a case (case 2, 3515) whose known biallelic *ABCA4* (NM_000350.3) pathogenic variants (c.1622T>C, p.[L541P] and c.1903C>A, p.[Q635K]) were located within a single long-range PCR amplicon of 7.2 kb. Using the American College of Medical Genetics guidelines, these variants are considered to be pathogenic and likely pathogenic, respectively (ClinVar accession numbers VCV000099067.36 and VCV000099095.2). The sequence reads were segregated based on their nucleotide sequence at position c.1622 (chr1:94528806) (Fig. 3). This confirmed that the variants were arranged in *trans*, and the proportion of chimeric reads was determined to be low (17%). In the absence of parental samples, this confirmed a molecular diagnosis of Stargardt disease in this patient.

We then expanded our test to 4 cases requiring variants to be phased using multiple amplicons to confirm or refute their clinical significance (mutations detailed in Table 2). This targeted approach required the analysis of between 3 and 9 amplification products per case. Summary sequencing metrics and the mean read depth per amplicon are detailed in Table 2. All target variants were identifiable from their nanopore sequence traces. However, for each case, we observed at least one overlap that did not harbor any heterozygous variants (minimum: 1 and maximum: 2) that prevented the phasing of the target variants. Nevertheless, the majority (16/21) of the amplicon overlaps from all these cases containing variants and enabled phase blocks of up to 17.6 kb to be established (Supplementary Table 2). Although a lack of variation at the *ABCA4* locus in these cases limited the functionality of this assay, our data show that this assay can work well but is dependent on the level of heterozygosity in the case. It is, therefore, effective for phasing small loci, particularly if WGS data are available to help inform amplicon design or assay viability.

Discussion

We have described our experience using a long-range PCR target enrichment method to phase variants at the *ABCA4* locus. Although this approach is technically robust, several challenges, most notably a lack of heterozygous sequence variants among the overlapping segments of amplification products, have limited its operational utility. Despite this, phase blocks of up to 94.9 kb were constructed and verified using the GIAB data set, showing the potential of this method.

Our approach was successfully applied to a previously investigated case in which compound heterozygous *ABCA4* variants had been identified. This case was selected because the target variants were sufficiently close to enable haplotype formation from a single amplicon; we established that both variants were in *trans*, confirming a diagnosis of Stargardt disease. Previously, to phase this case, standard methods would have involved PCR amplification followed by TA-cloning. Not only is nanopore sequencing quicker, less labor intensive, and negates the need for bacterial facilities, it enables chimeric amplicons to be identified. This work is an extension of previous studies in which we established the phase of pathogenic variants in *TMEM231* (a cause of Joubert syndrome),[19] and, more recently, resolved the parent-of-origin of a *de novo UBE3A* variant causing Angelman syndrome at the imprinted 15q11-13 locus.[20]

When we assessed a larger cohort of cases with variants separated by multiple intervening amplicons, we found it was impossible to resolve the phase. This was primarily due to a lack of heterozygous variants in the overlapping intervals. This was unexpected based on our examination of gnomAD, and 10 pilot cases that indicated that all overlaps contained at least one variant. Although we anticipated our design strategy to be successful, limitations related to the variant distribution across our selected cases limited the practical utility of the assay.

A previous report described the presence of chimeric amplification products and reference alignment bias and concluded that long-range amplification was a major pitfall preventing robust phasing using nanopore sequencing.[18] Our presented workflow, which benefits from the significantly increased read depth per sample offered by Flongle flowcells (a flowcell that offers a reduced output at a lower cost), has enabled us to better determine the proportion of chimeric products. This gives greater confidence in the construction of intra-amplicon haplotypes. Furthermore, since 2016, there has been a general increase in the viability of nanopore-based sequencing. This is the result of greater throughput and increased accuracy from subsequent editions of pores, together with advances in base calling in subsequent versions of Guppy and variant calling using Nanopolish. We conclude that, by comparison with the publicly available

NA12878 case, our constructed phase blocks were accurate, and we were able to robustly phase, particularly within single amplicons. Notwithstanding the lack of polymorphic variation within the overlapping fragments, there is no reason why even larger phase blocks than 94.9 kb could not be constructed.

Our workflow presents a cost-effective strategy for the robust phasing of individual alleles. Alternative methods include native genomic sequencing without enrichment or CRISPR-based target enrichment before either Nanopore or PacBio SMRT sequencing.[21,22] Although amplification-free methods will not have the issue of chimeric read formation, as no chimeric PCR products will be created, far fewer copies of the target DNA are obtained. This presents several practical challenges. First, the amount of input DNA required is far higher for amplification-free methods than for PCR-based methods, and this quantity of DNA is often unavailable for many samples. Second, the resulting read depth is far lower for amplification-free enrichment and direct sequencing approaches, limiting variant calling at the single-base level. As the resulting read depth is so much lower than PCR-based methods, typically a full MinION flowcell per sample (or alternatively PacBio SMRT sequencing) is used to achieve sufficient read depth. As such, these methods are far more expensive than the method presented here. Sequencing a single case on the Flongle Flow Cell represents an approximately 10-fold cost reduction compared with sequencing a single case on the MinION flowcell. In addition to this, the increased cost of CRISPR reagents as compared with PCR reagents must be considered. Nevertheless, the analysis of native genomic DNA retains methylation status, which has allowed other investigators to create haplotype-phased methylation calls in primary tissue and corresponding paired tumors.[23] Similarly, amplification-free adaptive sampling using a real-time selection of target loci may allow the phasing of large genomic regions using bulk genomic DNA.[24]

Although amplification-free methods will not suffer from chimeric read formation, they will have the same issue with a lack of variation if a tiling approach is used. Methods leveraging ultralong reads can avoid the need to adopt a tiling approach; however, these methods rely on access to fresh blood or large amounts of high-molecular weight DNA, precluding their use on most samples. Although the tiled approach, which has been used here, proved to be practically challenging, it was possible to utilize this method on a broad range of samples, including previously extracted freeze-thawed DNA samples. Furthermore, many methods leveraging ultralong read lengths require expensive, specialist equipment, such as the SAGE-HLS machine. As such, although the ultralong read sequencing methods are extremely powerful, the genomic architecture of the region in question must be considered when selecting an enrichment method, as for many cases, an approach using long-range PCR, as described here, may be as effective.

An improved strategy could involve increasing the coverage of amplification products spanning the region of interest, so that multiple staggered amplicons cover each nucleotide. This would increase the likelihood of identifying sufficient variation to enable the stitching together of large phase blocks. However, this is impractical across large distances, and although it reduces the risk of a lack of variation preventing phasing, it does not eliminate it. While in the presented cases, it would have been possible to further redesign the amplicons to increase the number of overlaps with heterozygous variants, a more targeted approach in cases where the distribution of variants is already known through prior sequencing may be beneficial. A further issue with the current method was the decrease in amplified material generated with only 25 PCR cycles, making equimolar pooling practically

Table 2
Summary assay metrics for sequenced samples

Local sample ID	Case number	Target variants			Yield (MinKnow)	Unfiltered read count	Filtered read count	Reads on target	Reads on target (%)	Mean combined read depth across the target locus (×)	Target amplicons (N)	Variants per overlap (Mean)	Overlaps without variant (N)
		Allele 1	Allele 2	Interval (kb)									
GIAB (NA12878) 35×	1[a]	N/A	N/A	132,260	1.47 Gb	262,875	89,280	84,266	94.38	4808	23	5.1	1
GIAB (NA12878) 25×	1[b]	N/A	N/A	132,260	776.12 Mb	102,932	64,628	63,116	97.66	3557	23	5.1	1
3515	2	c.1622T>C	c.1903C>A	635	119.44 Mb	33,033	9064	120[c]	80.20	108	1	N/A	N/A
5863	3	(c.2588G>C, c.5603A>T)	c.4537del	40,787	674.08 Mb	236,349	18,386	7081	38.51	1295	8	4.1	1
5607	4	c.3259G>A	c.6089G>A	37,331	164.3 Mb	55,512	7437	2767	37.21	548	7	6.0	1
3536	5	c.1906C>T	c.3113C>T	19,195	199.87 Mb	93,411	7036	1,421	20.20	608	4	4.0	2
1337	6	c.5603A>T, c.5819T>C	c.6817-2A>C	17,667	683.13 Mb	108,826	39,608	26,012	65.67	11,115	3	4.3	1

N/A, not applicable.
[a] PCR performed with a total of 35 cycles.
[b] PCR performed with a total of 25 cycles.
[c] At the interrogated amplicon.

challenging. To overcome this, relatively good-quality DNA samples and robust PCR primers should be used.

PacBio SMRT sequencing is an alternative single-read sequencing platform with increased per-base accuracy compared to nanopore sequencing.[25] While this presents as an attractive alternative, SMRT sequencing has a lower maximum read length compared with nanopore sequencing. As such, if a large gene, such as *ABCA4*, was phased, a tiling approach similar to the one adopted here would be required. Furthermore, the choice of sequencing platform would not affect the practical challenges surrounding enrichment by long-range PCR, such as chimeric read formation. In addition, SMRT sequencing does not have a reduced-cost flow cell equivalent to the Flongle. Thus, if SMRT sequencing was used in conjunction with long-range PCR, as described here, large multisample libraries would need to be pooled to reduce the per-sample cost.

We conclude that this is a valid and efficient method to phase variants, although it is ill-suited to do so at scale across a large region, such as the *ABCA4* locus. We propose that this method is valuable for phasing variants in smaller genes or those near larger genes. Additionally, this method may be well suited for phasing of variants that are spaced far apart in cases with available sequence data to allow for the design of bespoke target primers to ensure variants in the overlaps between amplicons, permitting more robust haplotype generation.

Author Contributions

B.M.C. performed the development of methodology, analysis, and writing of the paper, C.M.W. performed study concept and design, development of analysis, review, and revision of the paper, L.A.C. provided technical and material support, M.M.K. provided patient samples and clinical information, M.A. and C.I. provided analysis and interpretation of data and, review and revision of the paper, and C.T. performed the study concept and design, analysis and interpretation of data and, review and revision of the paper. All authors have read and approved the final paper.

Data Availability

The data that support the findings of this study are not publicly available due to the use of patient sequences but are available from the corresponding author upon request.

Funding

This work has been supported by a Horizon 2020, Marie Sklodowska-Curie Innovative Training Network entitled European Training Network to Diagnose, Understand, and Treat Stargardt Disease, a Frequent Inherited Blinding Disorder-StarT (813490).

Christopher Watson was supported by the NVIDIA Academic Hardware Grant Program.

Declaration of Competing Interest

C.M. Watson has received travel expenses to speak at Oxford Nanopore Technologies' organized conference. The other authors declare no conflict of interest.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.labinv.2023.100160

References

1. Kong X, Zhong X, Liu L, Cui S, Yang Y, Kong L. Genetic analysis of 1051 Chinese families with Duchenne/Becker muscular dystrophy. *BMC Med Genet*. 2019;20(1):139.
2. Chang YS, Chang CM, Lin CY, Chao DS, Huang HY, Chang JG. Pathway mutations in breast cancer using whole-exome sequencing. *Oncol Res*. 2020;28(2):107–116.
3. Crowley E, Warner N, Pan J, et al. Prevalence and clinical features of inflammatory bowel diseases associated with monogenic variants, identified by whole-exome sequencing in 1000 children at a single center. *Gastroenterology*. 2020;158(8):2208–2220.
4. Runhart EH, Sangermano R, Cornelis SS, et al. The common ABCA4 variant p. Asn1868ile shows nonpenetrance and variable expression of stargardt disease when present in trans with severe variants. *Invest Ophthalmol Vis Sci*. 2018;59(8):3220–3231.
5. Lindenbaum P. JVarkit: java-based utilities for Bioinformatics. Published online May 26, 2015. https://doi.org/10.6084/M9.FIGSHARE.1425030.V1
6. Marks P, Garcia S, Barrio AM, et al. Resolving the full spectrum of human genome variation using linked-reads. *Genome Res*. 2019;29(4):635–645.
7. Chen Z, Pham L, Wu TC, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res*. 2020;30(6):898–909.
8. Campbell P, Ellingford JM, Parry NRA, et al. Clinical and genetic variability in children with partial albinism. *Scientific Rep*. 2019;9(1):16576.
9. Bauwens M, Garanto A, Sangermano R, et al. ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. *Genet Med*. 2019;21(8):1761.
10. Cremers FPM, Lee W, Collin RWJ, Allikmets R. Clinical spectrum, genetic complexity, and therapeutic approaches for retinal disease caused by ABCA4 mutations. *Prog Retin Eye Res*. 2020;79:100861.
11. Schulz HL, Grassmann F, Kellner U, et al. Mutation spectrum of the ABCA4 gene in 335 Stargardt disease patients from a multicenter German cohort-impact of selected deep intronic variants and common SNPs. *Invest Ophthalmol Vis Sci*. 2017;58(1):394–403.
12. Zook JM, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource for benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32(3):246–251.
13. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666–2669.
14. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100.
15. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
16. Martin M, Patterson M, Garg S, et al. WhatsHap: fast and accurate read-based phasing. *bioRxiv. Published online November*. 2016;14:085050.
17. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26.
18. Laver TW, Caswell RC, Moore KA, et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci Rep*. 2016;6:21746.
19. Watson CM, Dean P, Camm N, et al. Long-read nanopore sequencing resolves a TMEM231 gene conversion event causing Meckel-Gruber syndrome. *Hum Mutat*. 2020;41(2):525–531.
20. Watson CM, Jackson L, Crinnion LA, Bonthron DT, Sheridan E. Long-read sequencing to resolve the parent of origin of a de novo pathogenic *UBE3A* variant. *J Med Genet*. 2022;59:1082–1086.
21. Nachmanson D, Lian S, Schmidt EK, et al. Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). *Genome Res*. 2018;28(10):1589–1599.
22. Watson CM, Crinnion LA, Hewitt S, et al. Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications. *Lab Invest*. 2020;100(1):135–146.
23. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38(4):433–438.
24. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Redfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnol*. 2021;39(4):442–450.
25. Aedui S, Ameur A, Vermeesch JR, Hestand MS. Single-molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*. 2018;46(5):2159–2168.