TaKoHigh enables accurate variant calling and phasing of PCR-based long-read sequencing data

Yuka Eura[1], Atsushi Takahashi[2,3], Sheng Ye[1], Kazuya Sakai[4], Masanori Matsumoto[4,5], and Koichi Kokame[1]*

[1]Department of Molecular Pathogenesis, National Cerebral and Cardiovascular Center, Suita, Osaka, Japan

[2]Department of Genomic Medicine, National Cerebral and Cardiovascular Center, Suita, Osaka, Japan

[3]Omics Research Center, National Cerebral and Cardiovascular Center, Suita, Osaka, Japan

[4]Department of Blood Transfusion Medicine, Nara Medical University, Kashihara, Nara, Japan

[5]Department of Hematology, Nara Medical University, Kashihara, Nara, Japan

***Correspondence:**

Koichi Kokame

Department of Molecular Pathogenesis

National Cerebral and Cardiovascular Center

6-1 Kishibe-Shimmachi, Suita, Osaka 564-8565, Japan

E-mail: kame@ncvc.go.jp, Phone: +81-6-6170-1070

Abstract

Long-read sequencing (LRS) is a powerful approach for analyzing causative variants in hereditary diseases. When the target gene is known, long-range PCR amplicons provide high coverage and efficiency. However, artifacts such as chimeric reads, allelic imbalance, and uneven coverage across overlapping amplicons can compromise the accuracy of variant calling and phasing. Yet no existing software is tailored to the properties of PCR-based LRS. We developed TaKoHigh, the first variant calling and phasing tool optimized for LRS of PCR amplicons. TaKoHigh analyzes each amplicon individually, applies thresholds based on allelic balance, and connects haplotypes through overlaps between adjacent amplicons. In *ADAMTS13*-associated cases, TaKoHigh achieved 98% variant calling accuracy, outperforming Clair3 (64%) and Longshot (50%). It also successfully resolved compound heterozygosity in cases where conventional tools failed. TaKoHigh enables robust interpretation of PCR-based LRS data without requiring specialized experimental protocols, making it broadly applicable in both clinical and research settings.

Introduction

The identification of pathogenic variants is important for elucidating the cause and pathogenesis of inherited diseases, and this information is crucial for determining treatment strategies. Targeted sequencing using PCR is a cost- and time-efficient approach, and is particularly useful for the analysis of known disease-associated genes[1] [2]. The Sanger method is a gold standard for exon-level analysis, whereas PCR-based long-read sequencing (LRS) is emerging as a powerful tool for analyzing entire genes, including intronic regions[3]. Compared to whole-genome or whole-exome sequencing, PCR-based LRS provides higher coverage in target regions and requires less patient-derived genomic DNA, which is advantageous when sample volume is limited.

For Sanger sequencing, user-friendly commercial software is widely available and accessible even to non-specialists. In contrast, LRS requires a multi-step bioinformatics workflow that can be more technically demanding[4]. Popular tools such as Clair3[5] and Longshot[6] have been developed for variant calling and phasing in long-read data, but these tools do not explicitly consider the specific characteristics of PCR amplicons, such as artificial chimeras and allelic amplification bias[7]. This limitation may lead to incorrect genotyping or phasing, particularly in clinical samples with subtle allelic imbalances.

In this study, we describe the specific technical challenges of PCR-based LRS and present a dedicated solution based on our analysis of patients with hereditary thrombotic thrombocytopenic purpura (hTTP). We observed discrepancies between variant calls and phasing results obtained using Clair3/Longshot and those validated by Sanger sequencing of paternal and maternal genomes. To investigate the underlying cause, we manually performed variant calling and phasing by analyzing reads from individual PCR amplicons, and obtained results consistent with Sanger-based phasing. Through this process, we identified that the inaccuracies in variant calling and phasing were due to PCR-specific artifacts including chimeric reads and unequal amplification between alleles, which are common even in well-optimized PCR reactions.

Based on these findings, we developed TaKoHigh, a new software tool that enables accurate variant calling and phasing in real-world PCR-based long-read sequencing data. A usage README is provided as Supplementary Note, and compiled executables for macOS (Apple silicon) and Windows are available from the corresponding author upon reasonable request. TaKoHigh is robust to typical PCR-derived artifacts and requires no special experimental protocols, making it compatible with standard laboratory procedures. This tool represents a significant step toward the reliable interpretation of long-read amplicon data in clinical and research settings.

Results

Challenges of compound heterozygosity analysis using PCR-based LRS

hTTP is an autosomal recessive disease caused by pathogenic variants in both *ADAMTS13* alleles. Identifying the causative variant of the *ADAMTS13* gene is crucial for a definitive diagnosis[8-14]. In the case of patient USS-S, two candidate pathogenic variants, c.821T>C p.(Leu274Pro) and c.3116G>A p.(Cys1039Tyr), were identified using the Sanger method; however, one of them, c.3116G>A, was not detected in either parent, indicating a potential *de novo* mutation (Fig. 1a). Therefore, it was challenging to prove compound heterozygosity. To address this issue, we applied PCR-based LRS to determine whether the two variants were located on the same or different alleles using phasing, which utilizes the ability of long reads to reconstruct haplotypes based on single molecules (Fig. 1b, c).

Long-range PCR amplicons were designed to span the two candidate variants, with 3–8 kb overlaps between adjacent amplicons (Supplementary Table 1). These overlapping regions are essential for phasing, enabling the construction of haplotypes by linking heterozygous single-nucleotide variants (SNVs), one of the strengths of LRS. PCR was performed using long-range DNA polymerases under optimized conditions, and Nanopore sequencing followed standard protocols. Each of the six PCR reactions produced clear bands of the expected size (Extended Data Fig. 1a). The PCR amplicons were pooled, and libraries were prepared using the standard protocol. Sequencing was performed on GridION platform with the R10.4.1 flow cell in high-accuracy mode (260 bp/s), generating high-quality raw data (Supplementary Table 2). The FASTQ reads were quality-filtered and size-selected using Filtlong, and mapped to the reference genome (GRCh38) using Minimap2[15,16] to create BAM files (Extended Data Fig. 1b). Variant calling and phasing were performed using Clair3 and Longshot (Extended Data Fig. 2). Haplotypes were then constructed using WhatsHap from the resulting phased VCF files (Fig. 1c, full image in Extended Data Fig. 1c)[17].

To validate these bioinformatic results, the phased variants were compared with Sanger sequencing results of the parental DNA. Some discrepancies were observed (Fig. 1d), prompting us to investigate whether the standard analysis pipeline failed to resolve the correct phasing and, if so, to identify the cause.

Manual phasing reveals causes of miscalling and misphasing

Because the existing software failed to accurately perform variant calling and phasing, we attempted to manually call variants and execute phasing using the same sequence data from patient USS-S (Fig. 2). The reads corresponding to each amplicon were extracted by searching

for amplicon-specific sequences (Step 1, Fig. 2a). Approximately 50 reads were extracted for each amplicon, numbered, and used as input (Step 2, Fig. 2b). Heterozygous variants were selected for analysis from each amplicon. The procedure was as follows: the BAM file was loaded in the Integrative Genomics Viewer (IGV)[18], and candidate heterozygous variants were selected such that multiple variants were included per amplicon (Step 3, Fig. 2c). A total of 26 variants were identified from manual inspection in IGV, with their positions, reference/alternative bases, and detection status by Clair3 and Longshot summarized in Extended Data Fig. 3. For each amplicon, the corresponding reads were displayed in a sequence analysis software (Sequencher, Gene Codes Corporation), and the nucleotides at each selected site from Step 3 were identified and recorded (Step 4, Fig. 2d; full-size image in Extended Data Fig. 4).

The nucleotide data from Step 4 were entered into a Microsoft Excel spreadsheet. Based on the combinations of bases obtained from manual inspection in IGV, each base was color-coded into two haplotype groups (cyan and magenta). The next step involved creating haplotypes (Step 5, Fig. 2e) by adjusting base assignments within each read to ensure that adjacent bases were consistently grouped by color. Bases that differed from those indicated by IGV (yellow) or deletions (white) were also documented. Using the constructed haplotypes for each amplicon, all amplicons were successfully merged into a unified haplotype map without any contradictions (Step 6, Fig. 2f). Finally, to confirm the accuracy of the haplotype assignment, Sanger sequencing was performed for all heterozygous variants using the paternal and maternal genomes (Step 7, Fig. 2g). This approach confirmed the consistency of our phasing results and demonstrated that the candidate variant, c.3116G>A, was inherited from the father, validating the diagnosis of compound heterozygosity (Fig. 2g).

The strength of this manual approach lies in the separate analysis of each amplicon, allowing interpretation based on the specific characteristics of individual amplicons. This approach also revealed the cause of this error observed in existing software. Existing tools do not explicitly account for factors inherent to PCR-based sequencing, such as (1) PCR-induced chimeras (chimeric amplicons) and (2) allele-specific amplification bias (Extended Data Fig. 5). Our findings emphasize the necessity of interpreting long-read sequencing data within the context of each amplicon's quality and behavior. This result clearly demonstrates that even software with high analytical performance on PCR-free data can produce misleading results when applied to PCR-amplified sequences.

Design and implementation of TaKoHigh

Therefore, we developed a software tool named TaKoHigh (Targeted Amplicon-based variant

Calling and phasing Optimized for High-throughput long-read sequencing) to automate the manual steps demonstrated in the previous analysis. Conventional software such as Clair3 and Longshot attempts to phase variants using generic algorithms that do not consider PCR-specific biases or differences between amplicons, such as variability in coverage, error profiles, or phasing context (Fig. 3c). In contrast, TaKoHigh is designed to address these limitations by utilizing user-defined PCR information, applying specialized criteria for each amplicon, and explicitly detecting unphased gaps. The conceptual differences between these approaches, along with the overall analysis workflow, are illustrated in Fig. 3. TaKoHigh is, to our knowledge, the first amplicon-aware software specifically designed for PCR-based LRS. The strategy behind TaKoHigh is illustrated in Figs. 3 and 4. Unlike general variant-calling tools for LRS, TaKoHigh begins by extracting only the reads corresponding to each amplicon, based on the start and end positions and the read length. To minimize erroneous data, which is critical for accurate analysis in TaKoHigh, read groups were organized and analyzed separately for each amplicon (Fig. 3c). The optimal criteria for variant detection were determined according to the amplification ratio between alleles in each amplicon. Subsequently, SNVs were identified and haplotypes were constructed based on the co-occurrence frequency of adjacent SNVs within each amplicon (Fig. 3c). Adjustments, including correction of phase errors and cross-amplicon rescue, were applied using information from the overlapping regions of neighboring amplicons. In this way, the longest possible and most accurate haplotypes were generated (Fig. 3c). These steps were specifically designed to overcome the common issues in PCR-based LRS, such as allelic imbalance, the presence of chimeric reads, and uneven coverage between different amplicons.

Workflow of TaKoHigh for phasing across amplicons

The TaKoHigh variant caller accepts a BAM file comprising one or more amplicons as input (Fig. 4). Based on the provided amplicon information, the software generates split BAM files for each amplicon, followed by the identification of variants using specific homozygous and heterozygous variant criteria for each amplicon. The variant list is then constructed through cross-amplicon rescue. Within these variants, phasing is performed for each amplicon using heterozygous variants that meet the necessary quality criteria for reliable identification. The final step involves combining these elements to achieve phasing across amplicons. This approach leverages the strengths and mitigates the weaknesses of PCR-based LRS.

It is also important to note that when no heterozygous SNVs exist to act as a bridge between two amplicons, the software should not forcibly connect the haplotypes. Instead, it should clearly indicate the lack of connection. This output serves as a cue for users to perform

additional PCR amplification to bridge the missing segments. For example, in patient USS-2M, no bridging heterozygous SNVs were found between PCR2 and PCR3. Consequently, the software generated two haplotypes separated by a clearly defined gap (phase sets), as expected (Extended Data Fig. 6), thereby highlighting the need for an additional PCR.

Benchmarking and validation of TaKoHigh performance

To evaluate the performance of TaKoHigh, we used sequence data from patient USS-S and obtained a phased variant list (provided as a worksheet within a single Excel file in the Supplementary Data, TaKoHigh output for USS-S). Consistent with manual phasing, TaKoHigh concluded that the candidate variant not present in the parents was located on the paternal allele, thereby confirming compound heterozygosity. All 41 heterozygous variants (HET-1 to HET-41) used for phasing were verified by Sanger sequencing of the parental genome. The variants inherited from the paternal and maternal alleles are shown as blue and red, respectively. Notably, HET-2 was present in both parents, and HET-41 was not found in either, rendering its origin undetermined; it is therefore displayed as a black bar (Fig. 5a).

Next, we compared the results obtained from four methods: TaKoHigh, Clair3, Longshot, and manual inspection in IGV (Fig. 5b). A complete list of SNVs, both heterozygous and homozygous, detected by each method was compiled. We identified 41 heterozygous variants (HET-1 to HET-41) and 17 homozygous variants (HOMO-1 to HOMO-17). The presence or absence of variant calls across methods is indicated. TaKoHigh accurately identified the genotypes of all variants except HOMO-5. In contrast, Clair3 and Longshot missed several variants. The numerical detection performance of each software is summarized in Table 1.

Additionally, we examined results using BAM files divided by amplicon and analyzed them using the same four methods. Except for the PCR2 BAM file, the variant calling results were identical to those obtained from the entire BAM file. For PCR2, however, analysis of the full BAM file with TaKoHigh successfully detected both SNVs and phasing, indicating that cross-amplicon rescue was effective (Fig. 5c). In contrast, Clair3 and Longshot failed to detect several variants in PCR2 and other regions (Extended Data Fig. 7). Manual inspection in IGV of the BAM files confirmed the missing variants in PCR2, while other amplicon regions were consistent with the results from TaKoHigh. The corresponding phased variant lists generated by TaKoHigh for USS-S, USS-2M, and USS-2M with gap rescue (PCR8 added) are provided as separate worksheets within a single Excel file in the Supplementary Data.

The accuracy of variant calling was highest for TaKoHigh (98%), followed by Clair3 (64%) and Longshot (50%) (Table 1). The reduced performance of existing software was attributed to the inability to rescue allele-specific amplification imbalance in PCR2.

Regarding phasing accuracy, only TaKoHigh produced correct phasing results (Fig. 5a), while Longshot and Clair3 generated incorrect results (Fig. 1d). The phasing errors in existing software were largely attributable to missed variant calls resulting from PCR chimeras and an amplification imbalance between alleles.

Discussion

TaKoHigh, developed in this study, is a software tool specialized in "variant calling and phasing" and is exclusively designed for LRS of amplicons. This represents the first haplotyping software designed specifically for PCR-based data. PCR amplicons pose unique challenges, including the presence of chimeric reads and allelic imbalance. However, these issues can be managed effectively when the analysis is designed with a clear understanding of PCR-specific features. In fact, the use of PCR amplification enables targeted enrichment of the region of interest, allowing for high coverage and subsequently more accurate variant calling and phasing.

Existing software tools are prone to generating errors in variant calls and phasing when applied to amplicon-derived data (Fig. 1d, Extended Data Fig. 8). When paternal and maternal genomes data are available, discrepancies can be identified by verifying with the Sanger method. However, in cases where paternal and maternal genomes are unavailable, such verification becomes impractical, and potentially erroneous results may be accepted without scrutiny, a critical issue in clinical or diagnostic settings. Moreover, if every variant required verification by Sanger sequencing, the core advantage of automated phasing would be nullified.

TaKoHigh specializes in amplicons and was developed based on a detailed understanding of their unique characteristics, allowing it to avoid common sources of error. Importantly, the software clearly indicates regions where haplotypes cannot be connected, rather than forcing a connection. This output serves as a practical cue for users to consider additional PCR amplification to bridge such gaps.

For example, when analyzing patient USS-2M with existing tools, multiple phasing errors were observed in regions where haplotypes should have formed (Extended Data Fig. 8, 9a, and 10). In contrast, TaKoHigh correctly output two separate haplotypes with a clearly defined gap (Extended Data Fig. 6). Based on this result, we added one additional amplicon that included the two closest heterozygous variants from each haplotype. Reanalysis successfully produced a single, continuous haplotype (Extended Data Fig. 6 and 9b). This allowed us to determine that the rare variants in the promoter region were located on the same allele as previously identified variants and therefore not likely to be causative. Additional IGV-based visualizations confirming phasing consistency in USS-2M are provided in Extended Data Figs. 8 and 10.

Currently, TaKoHigh focuses on calling SNVs from BAM files mapped using Minimap2 or other equivalent software, and performs high-accuracy phasing using only SNVs that meet stringent quality criteria. In future versions, we plan to extend its functionality to include insertions and deletions (InDels) detection following a similar strategy. Compared to SNVs, InDel calling is more challenging due to higher error rates in difficult regions, such as homopolymers. Nonetheless, a tool that can detect both SNVs and InDels and phase them reliably would be a valuable addition for PCR-based long-read sequencing applications.

Methods

Patients

Two families, USS-S and USS-2M, were analyzed in this study. Patient USS-S (male) was diagnosed with hTTP at age 4 and has since received weekly prophylactic FFP infusions. Although neither his childhood nor family history has been obtained, ADAMTS13 activity was severely deficient without detectable inhibitors. His father and mother had ADAMTS13 activity levels of 34.2% and 47.6%, respectively[14]. Patient USS-2M (male) experienced multiple episodes of thrombocytopenia during infancy and childhood. He was later found to have severely reduced ADAMTS13 activity (2.0–2.8%) without detectable inhibitors. His father and mother had ADAMTS13 activity levels of 8.6% and 26.8%, respectively. In the USS-S family, two candidate variants were identified by Sanger sequencing; however, one of them was not found in either parent, making it difficult to confirm compound heterozygosity. In contrast, only one causative variant was detected by Sanger sequencing in the USS-2M family. Written informed consent was obtained from all patients. This study was approved by the Research Ethics Committee of National Cerebral and Cardiovascular Center (approval number M23-017).

LRS using Nanopore GridION

Genomic DNA was extracted from peripheral blood using the Blood GenomicPrep Mini Spin Kit (Cytiva). Long-range PCR was performed with Platinum SuperFi II DNA Polymerase (Thermo Fisher Scientific) using appropriate primer sets (Supplementary Table 1).

For patient USS-S, six long PCR amplicons, each approximately 10 kb in length and designed with 3–8-kb overlaps, were used (Extended Data Fig. 1a). For patient USS-2M, amplicons ranging from 9–15 kb were prepared, with overlaps of 1–5 kb (Extended Data Fig. 9a). To bridge the gap between two haplotypes in USS-2M, an additional 6-kb amplicon was generated and incorporated into the analysis (Extended Data Fig. 6 and 9b; Supplementary Table 1).

Library preparation was carried out using the Ligation Sequencing Kit V14 (Oxford

Nanopore Technologies). Sequencing was performed on a GridION platform equipped with an R10.4.1 flow cell. The instrument and software were operated according to the manufacturer's instructions using the high-accuracy mode (260 bp/s). Sanger sequencing was performed as described previously[9].

Bioinformatic analysis of LRS data

Raw sequencing data were generated using the Nanopore GridION platform. Base calling quality was assessed using NanoPlot (https://github.com/wdecoster/NanoPlot)[19], and the results are summarized in Supplementary Table 2. Reads were filtered using Filtlong (https://github.com/rrwick/Filtlong) with a minimum mean quality of 97; for USS-S, minimum read length of 7,000 bp and maximum of 10,500 bp; for USS-2M, minimum of 8,000 bp and maximum of 16,000 bp. The selected reads were then aligned to the reference genome using Minimap2 (https://github.com/lh3/minimap2).

Variant calling and phasing were performed using Clair3 (https://github.com/HKU-BAL/Clair3) and Longshot (https://github.com/pjedge/longshot). Phased VCF files were generated, and phased haplotypes were constructed using WhatsHap (https://whatshap.readthedocs.io/en/latest/index.html)[17].

TaKoHigh

TaKoHigh is implemented in Python 3, with selected components in C++. The core steps are:

1. Read selection

Reads are selected by mapping coordinates corresponding to each amplicon, using the provided amplicon definitions.

2. Variant calling for each amplicon

Variant calling is performed per amplicon. Homozygous genotypes are assigned for allele frequencies ≥0.95, and heterozygous genotypes for allele frequencies between 0.4 and 0.6. Strand bias is also considered. If an amplicon shows extreme allelic imbalance (i.e., no sites fall within the expected heterozygous range), the amplicon is designated "PCR-U", and site-level genotypes within that amplicon are withheld pending cross-amplicon refinement.

3. Refinement of variant calls (cross-amplicon rescue)

Variant calls are refined using overlaps among amplicons. Confident calls on one amplicon are propagated to overlapping PCR-U amplicons covering the same site; other unknown genotypic sites on that PCR-U amplicon with similar allele ratios are assigned the corresponding genotype.

4. Haplotype construction for each amplicon

Two haplotypes are constructed per amplicon from the most frequent co-occurrence patterns of

nearby variants, integrating all variants detected on that amplicon.

5. Integration of haplotypes across amplicons

Using heterozygous variants in overlapping regions, per-amplicon haplotypes are linked across amplicons to yield global haplotypes.

To facilitate reuse, a usage README is provided as Supplementary Note, while platform-specific executables are available from the corresponding author upon reasonable request.

Validation by Sanger sequencing

To confirm the accuracy of both variant calling and phasing by TaKoHigh, we performed Sanger sequencing on selected SNVs. Targeted SNVs were chosen from each amplicon, including those used for phasing and representative variants called by TaKoHigh. PCR primers were designed to flank these sites, and amplification was performed using standard Taq polymerase. Sequencing reactions were carried out using the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific), followed by capillary electrophoresis on an ABI 3500xl Genetic Analyzer. Resulting chromatograms were analyzed with Sequencher (Gene Codes Corporation). The accuracy of variant calls and their parental origin were confirmed.

Acknowledgments

We thank Dr. Keiko Sonoda for her advice on Nanopore sequencing. This work was supported in part by grants from the Ministry of Health, Labour, and Welfare of Japan (Grant number JPMH23FC1022 to K.K.); the Japan Society for the Promotion of Science (Grant numbers 23K07567 to Y.E. and 23K27372 to K.K.); the Japanese Society of Thrombosis and Haemostasis (2024 to Y.E.); and the Takeda Japan Medical Office Funded Research Grant (2022 to Y.E.).

Author contributions

Y.E., A.T., and K.K. designed the study and developed the core algorithms. Y.E. conducted all experiments. A.T. wrote and optimized the software code. Y.E. drafted the manuscript. S.Y. and K.K. contributed to data analysis and revised the manuscript. K.S. and M.M. recruited the patients. All authors discussed the results and approved the final version of the manuscript.

Code/Software availability

Executable binaries of TaKoHigh for macOS (Apple silicon) and Windows are available from the corresponding author upon reasonable request for academic use. A detailed usage README is provided as Supplementary Note. The source code is not publicly released at this time.

Competing interests

The authors declare no competing interests.

Correspondence and requests for materials and software should be addressed to Koichi Kokame.

Table 1. Comparison of variant calling accuracy for patient USS-S using three software tools

Type	Sanger	TaKoHigh	Clair3	Longshot
Heterozygous	41	41	21	20
Homozygous	17	16	16	9
Total	58	57	37	29

The number of variants (41 heterozygous and 17 homozygous) detected by each method is shown. Sanger sequencing was used as the reference standard. Phasing accuracy is not considered in this comparison. TaKoHigh detected 57 out of 58 variants (98% accuracy), outperforming Clair3 (37/58, 64%) and Longshot (29/58, 50%). The homozygous variant HOMO-5 was not detected by TaKoHigh, which accounts for the one discrepancy (see Extended Data Fig. 10).

Figure Legends

Figure 1. PCR-based long-read sequencing (LRS) to determine compound heterozygosity in patient USS-S. a, Two candidate variants in *ADAMTS13* (c.821T>C and c.3116G>A) identified by Sanger sequencing. **b**, Purpose of PCR-based LRS: to determine whether the two variants are located on the same or different alleles. **c**, Schematic overview of the LRS approach and IGV snapshot of BAM alignments mapped to the reference genome (GRCh38). Phased haplotypes (phase sets) generated by Clair3 and Longshot are shown as blue lines, with start and end positions indicated in green (Clair3) or red (Longshot). **d**, Comparison of phasing results between Sanger-based parental analysis and those obtained using Clair3 and Longshot. Misphasing was observed with both standard tools.

Figure 2. Step-by-step manual phasing of long-read data from patient USS-S. Manual construction and validation of haplotypes from FASTQ data using seven steps: **a**, Step 1 – Selection of reads corresponding to each amplicon based on amplicon-specific sequences. **b**, Step 2 – Collection of approximately 50 reads per amplicon. **c**, Step 3 – Identification of heterozygous SNVs using BAM files viewed in IGV (see Extended Data Fig. 3). **d**, Step 4 – Verification of the selected SNVs on each read using Sequencher (see Extended Data Fig. 4). **e**, Step 5 – Construction of haplotypes within each amplicon by grouping adjacent SNVs based on co-occurrence. **f**, Step 6 – Integration of all amplicon-specific haplotypes into a single, contiguous haplotype (see Extended Data Fig. 5). **g**, Step 7 – Validation of the phasing result by trio Sanger sequencing (Patient, P [father], M [mother]).

Figure 3. Overview of the variant calling and phasing strategy implemented in TaKoHigh, a tool optimized for PCR-based long-read sequencing. a, Schematic of overlapping long-range PCR design targeting a disease-causing gene. Overlaps between adjacent amplicons are essential for connecting per-amplicon haplotypes (cross-amplicon phasing). **b**, Representative reads obtained from PCR-based LRS. Such reads may contain PCR-specific artifacts including chimeric reads, allelic imbalance, and unintended junctions. **c**, TaKoHigh bioinformatic pipeline. BAM files corresponding to each amplicon are extracted, and SNVs are called using amplicon-specific amplification profiles. Haplotypes are then constructed based on co-occurrence patterns of adjacent SNVs. Haplotypes from poorly performing amplicons are rescued using data from neighboring amplicons. Finally, per-amplicon haplotypes are connected across amplicons to generate a complete phased region.

Figure 4. Workflow of TaKoHigh for variant calling and haplotype phasing from PCR-based long-read sequencing data. FASTQ files generated from PCR-based LRS are first filtered for quality using Filtlong and aligned to the reference genome with Minimap2. The resulting BAM file is used as input for TaKoHigh, which runs in a Python 3 environment. In addition to the BAM files, TaKoHigh requires two inputs: a tab-delimited text file (.txt) specifying per-amplicon genomic coordinates and IDs; a reference genome sequence file (.rse and .cti). TaKoHigh extracts reads corresponding to each amplicon, performs SNV calling using amplicon-specific criteria, and ranks variants by confidence. High-confidence variants are then used to construct haplotypes within each amplicon. Haplotypes are subsequently connected across amplicons using overlapping regions. If no heterozygous SNVs are present in the overlapping region, a "gap" is reported. Additional amplicons may then be designed to bridge the gap and complete phasing.

Figure 5. Performance comparison of variant calls and haplotype phasing methods. a, Verification of 41 heterozygous variants phased by TaKoHigh using parental Sanger sequencing. Variants inherited from the paternal and maternal alleles are shown in blue and red, respectively. HET-2 was present in both parents, and HET-41 was not detected in either; both are displayed in black. **b**, Comparison of variant calling performance across four approaches: TaKoHigh, Clair3, Longshot, and manual inspection in IGV. All 41 heterozygous and 17 homozygous variants (HET-1 to HET-41, HOMO-1 to HOMO-17) were assessed. TaKoHigh correctly identified all variants except HOMO-5. Clair3 and Longshot missed several variants. Quantitative results are summarized in Table 1. **c**, Analysis of BAM files divided by amplicon. While no variants were detected by TaKoHigh in PCR2 alone, the complete BAM file allowed successful detection and phasing, indicating that cross-amplicon rescue was effective. The results from other PCR regions were consistent across both approaches. The results for individual PCR amplicons using Clair3, Longshot, and manual inspection in IGV are shown in Extended Data Fig. 7.

References

1 Chan, K. W. et al. Targeted Gene Sanger Sequencing Should Remain the First-Tier Genetic Test for Children Suspected to Have the Five Common X-Linked Inborn Errors of Immunity. *Front. Immunol.* **13**, 883446 (2022).

2 Marx, V. Method of the year: long-read sequencing. *Nat Methods* **20**, 6-11 (2023).

3 van Dijk, E. L. et al. Genomics in the long-read sequencing era. *Trends Genet.* **39**, 649-671 (2023).

4 Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

5 Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* **2**, 797-803 (2022).

6 Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* **10**, 4660 (2019).

7 Laver, T. W. et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.* **6**, 21746 (2016).

8 Levy, G. G. et al. Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura. *Nature* **413**, 488-494 (2001).

9 Kokame, K. et al. Mutations and common polymorphisms in ADAMTS13 gene responsible for von Willebrand factor-cleaving protease activity. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11902-11907 (2002).

10 Matsumoto, M. et al. Molecular characterization of ADAMTS13 gene mutations in Japanese patients with Upshaw-Schulman syndrome. *Blood* **103**, 1305-1310 (2004).

11 Moake, J. L. Thrombotic thrombocytopenic purpura: survival by "giving a dam". *Trans. Am. Clin. Climatol. Assoc.* **115**, 201-219 (2004).

12 Sadler, J. E. Von Willebrand factor, ADAMTS13, and thrombotic thrombocytopenic purpura. *Blood* **112**, 11-18 (2008).

13 Lotta, L. A., Garagiola, I., Palla, R., Cairo, A. & Peyvandi, F. ADAMTS13 mutations and polymorphisms in congenital thrombotic thrombocytopenic purpura. *Hum. Mutat.* **31**, 11-19 (2010).

14 Fujimura, Y. et al. Natural history of Upshaw-Schulman syndrome based on ADAMTS13 gene analysis in Japan. *J. Thromb. Haemost.* **9 Suppl 1**, 283-301 (2011).

15 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

16 Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572-4574 (2021).

17 Martin, M. et al. WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 085050 (2016).

18 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178-192 (2013).

19 De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39** (2023).

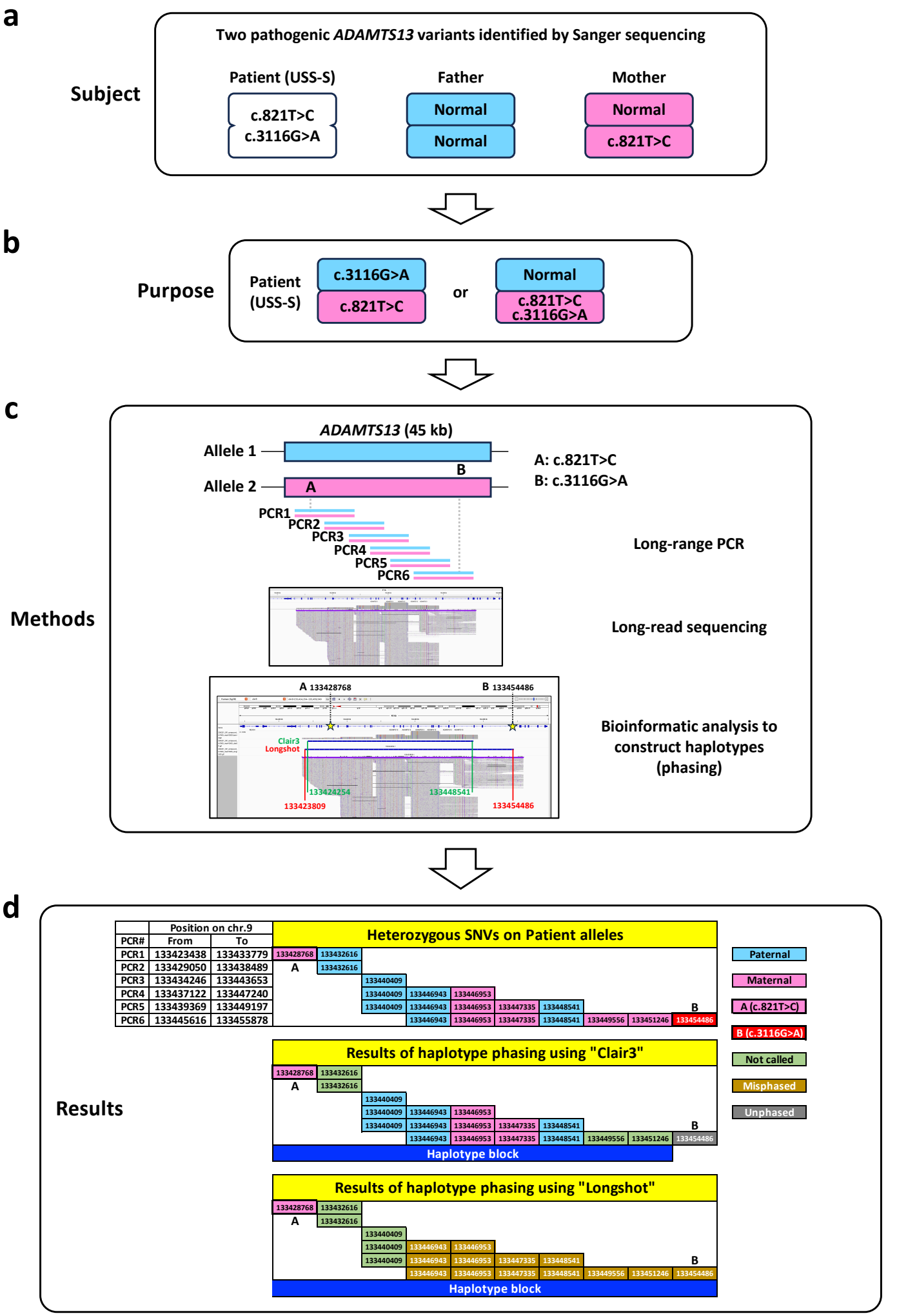

a
Subject
Two pathogenic ADAMTS13 variants identified by Sanger sequencing
Patient (USS-S)
c.821T>C
c.3116G>A
Father
Normal
Normal
Mother
Normal
c.821T>C
b
Purpose
Patient (USS-S)
c.3116G>A
c.821T>C
or
Normal
c.821T>C
c.3116G>A
c
Methods
ADAMTS13 (45 kb)
Allele 1
Allele 2
A
B
A: c.821T>C
B: c.3116G>A
PCR1
PCR2
PCR3
PCR4
PCR5
PCR6
Long-range PCR
Long-read sequencing
A 133428768
B 133454486
Clair3
Longshot
133424254
133448541
133423809
133454486
Bioinformatic analysis to construct haplotypes (phasing)
d
Results
Position on chr.9
PCR#
From
To
PCR1 133423438 133433779
PCR2 133429050 133438489
PCR3 133434246 133443653
PCR4 133437122 133447240
PCR5 133439369 133449197
PCR6 133445616 133455878
Heterozygous SNVs on Patient alleles
133428768
133432616
133440409
133446943
133446953
133447335
133448541
133449556
133451246
133454486
A
B
Paternal
Maternal
A (c.821T>C)
B (c.3116G>A)
Not called
Misphased
Unphased
Results of haplotype phasing using "Clair3"
Haplotype block
Results of haplotype phasing using "Longshot"
Haplotype block

Fig. 1

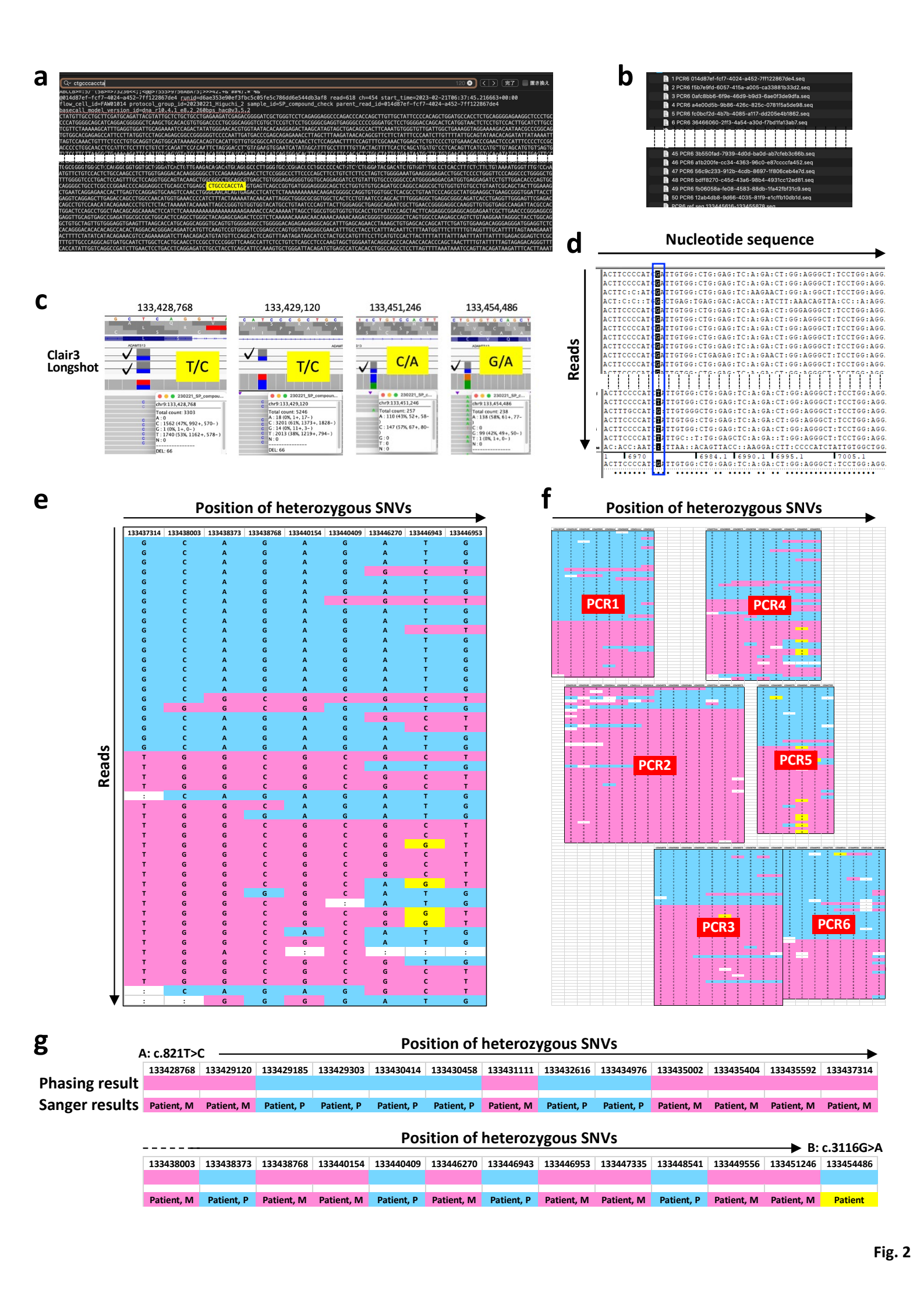

a
b
1 PCR6 014d87ef-fcf7-4024-a452-7ff122867de4.seq
2 PCR6 f5b7e9fd-6057-415a-a005-ca33881b33d2.seq
3 PCR6 0afc8bb6-6f9e-46d9-b9d3-6ae0f3de9dfa.seq
4 PCR6 a4e00d5b-9b86-426c-825c-0781f5a5de98.seq
5 PCR6 fc0bcf2d-4b7b-4085-a117-dd205e4b1862.seq
6 PCR6 36466060-2ff3-4a54-a30d-f7bd1fa13ab7.seq
45 PCR6 3b550fad-7939-4d0d-ba0d-ab7cfeb3c66b.seq
46 PCR6 a1b200fe-cc34-4363-96c0-e87ccccfa452.seq
47 PCR6 56c9c233-912b-4cdb-8697-1f806ceb4e7d.seq
48 PCR6 bdff8270-c45d-43a6-98b4-4931cc12ed81.seq
49 PCR6 fb06058a-fe08-4583-88db-1fa42fbf31c9.seq
50 PCR6 12ab4db8-9d66-4035-81f9-e1cffb10db1d.seq
c
Clair3
Longshot
133,428,768
133,429,120
133,451,246
133,454,486
T/C
T/C
C/A
G/A
d
Nucleotide sequence
Reads
e
Position of heterozygous SNVs
Reads
133437314 133438003 133438373 133438768 133440154 133440409 133446270 133446943 133446953
f
Position of heterozygous SNVs
PCR1
PCR4
PCR2
PCR5
PCR3
PCR6
g
A: c.821T>C
Position of heterozygous SNVs
133428768 133429120 133429185 133429303 133430414 133430458 133431111 133432616 133434976 133435002 133435404 133435592 133437314
Phasing result
Sanger results
Patient, M Patient, M Patient, P Patient, P Patient, P Patient, P Patient, M Patient, P Patient, P Patient, M Patient, M Patient, M Patient, M
Position of heterozygous SNVs
B: c.3116G>A
133438003 133438373 133438768 133440154 133440409 133446270 133446943 133446953 133447335 133448541 133449556 133451246 133454486
Patient, M Patient, P Patient, M Patient, M Patient, P Patient, M Patient, P Patient, M Patient, M Patient, P Patient, M Patient, M Patient

Fig. 2

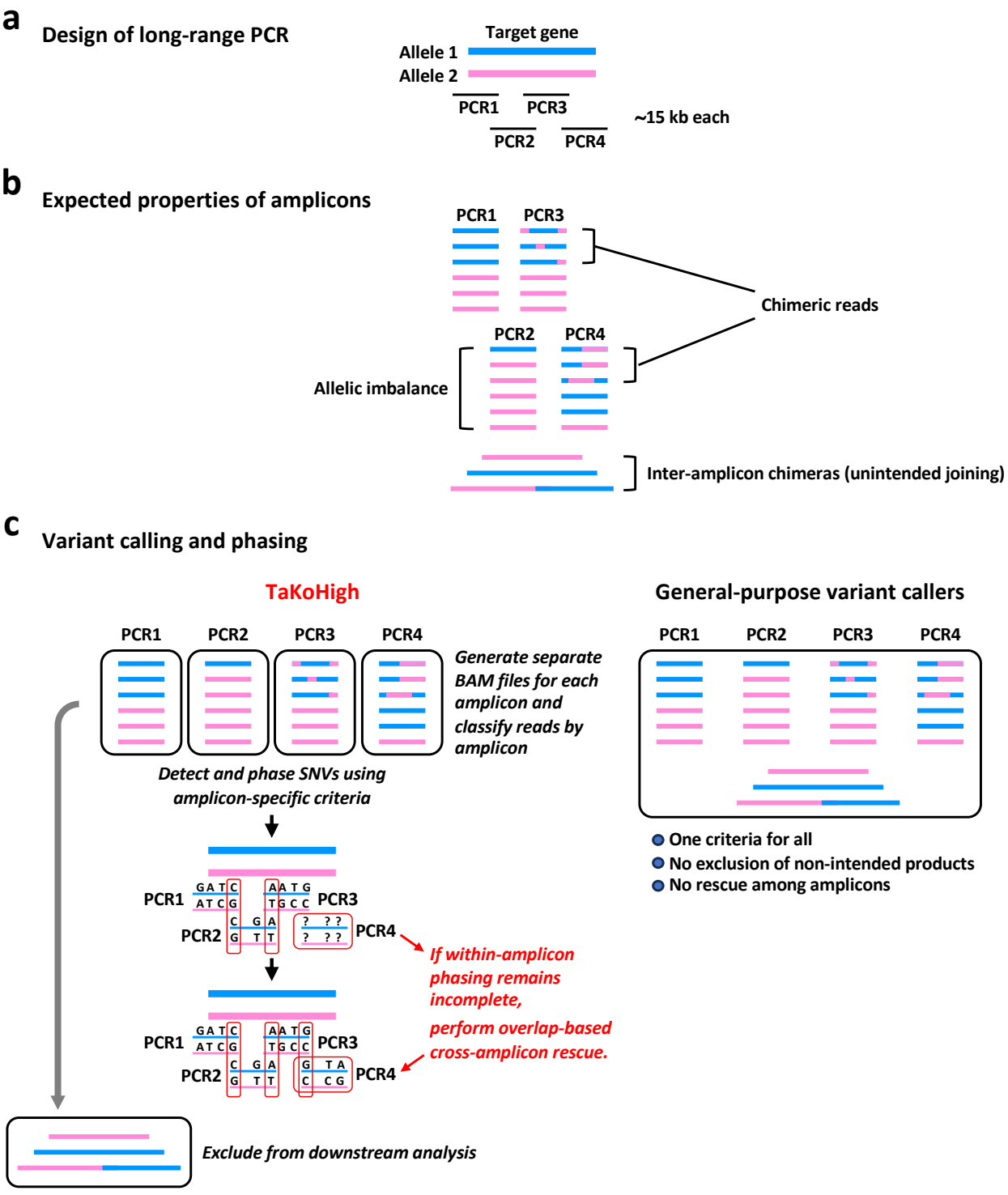

a
Design of long-range PCR
Target gene
Allele 1
Allele 2
PCR1
PCR3
PCR2
PCR4
~15 kb each
b
Expected properties of amplicons
PCR1
PCR3
Chimeric reads
PCR2
PCR4
Allelic imbalance
Inter-amplicon chimeras (unintended joining)
c
Variant calling and phasing
TaKoHigh
PCR1
PCR2
PCR3
PCR4
Generate separate BAM files for each amplicon and classify reads by amplicon
Detect and phase SNVs using amplicon-specific criteria
If within-amplicon phasing remains incomplete,
perform overlap-based cross-amplicon rescue.
Exclude from downstream analysis
General-purpose variant callers
One criteria for all
No exclusion of non-intended products
No rescue among amplicons

Fig. 3

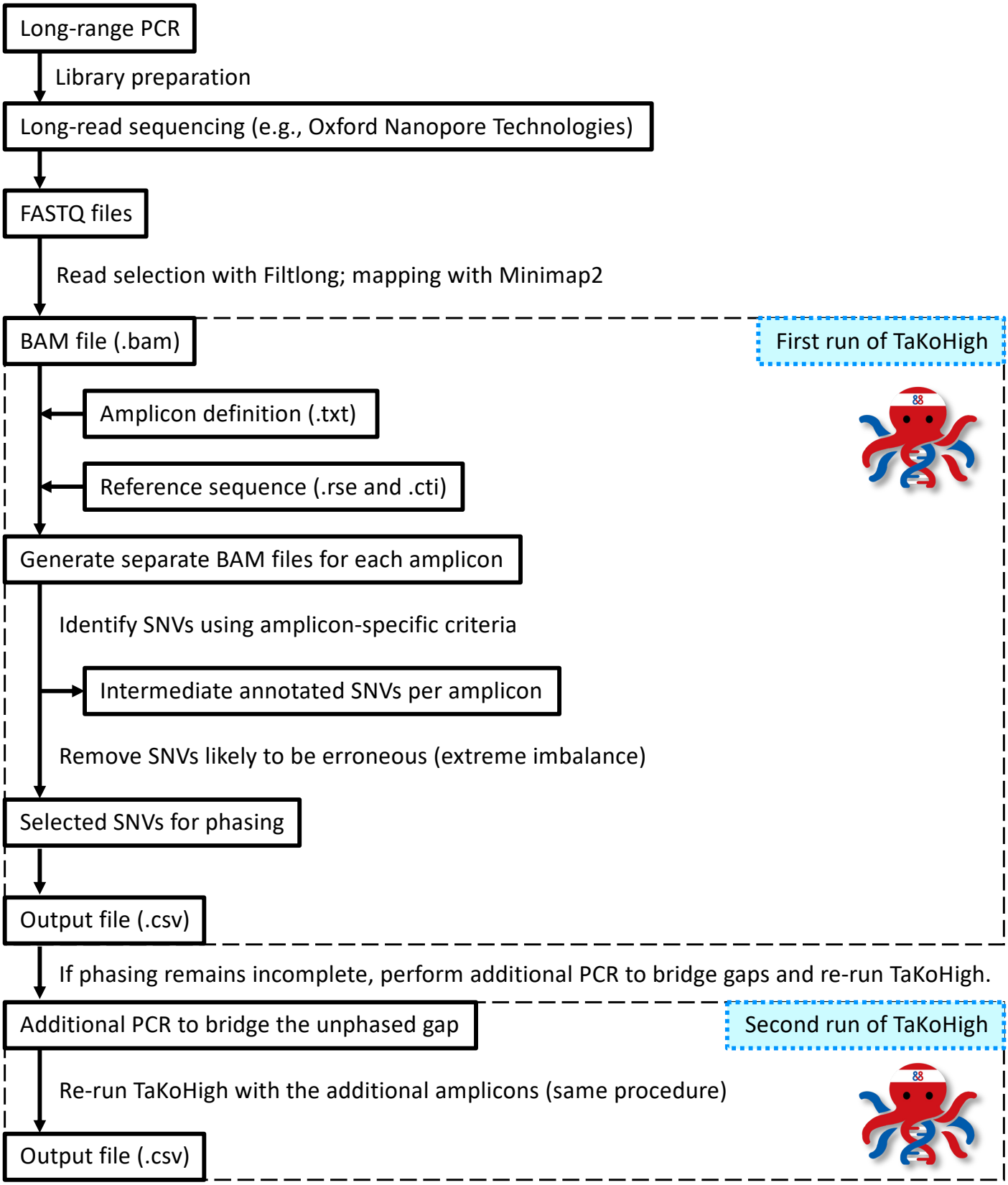
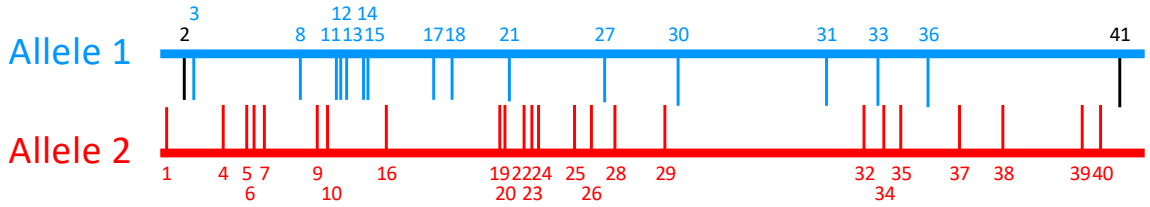

Long-range PCR
Library preparation
Long-read sequencing (e.g., Oxford Nanopore Technologies)
FASTQ files
Read selection with Filtlong; mapping with Minimap2
BAM file (.bam)
First run of TaKoHigh
Amplicon definition (.txt)
Reference sequence (.rse and .cti)
Generate separate BAM files for each amplicon
Identify SNVs using amplicon-specific criteria
Intermediate annotated SNVs per amplicon
Remove SNVs likely to be erroneous (extreme imbalance)
Selected SNVs for phasing
Output file (.csv)
If phasing remains incomplete, perform additional PCR to bridge gaps and re-run TaKoHigh.
Additional PCR to bridge the unphased gap
Second run of TaKoHigh
Re-run TaKoHigh with the additional amplicons (same procedure)
Output file (.csv)

Fig. 4

a



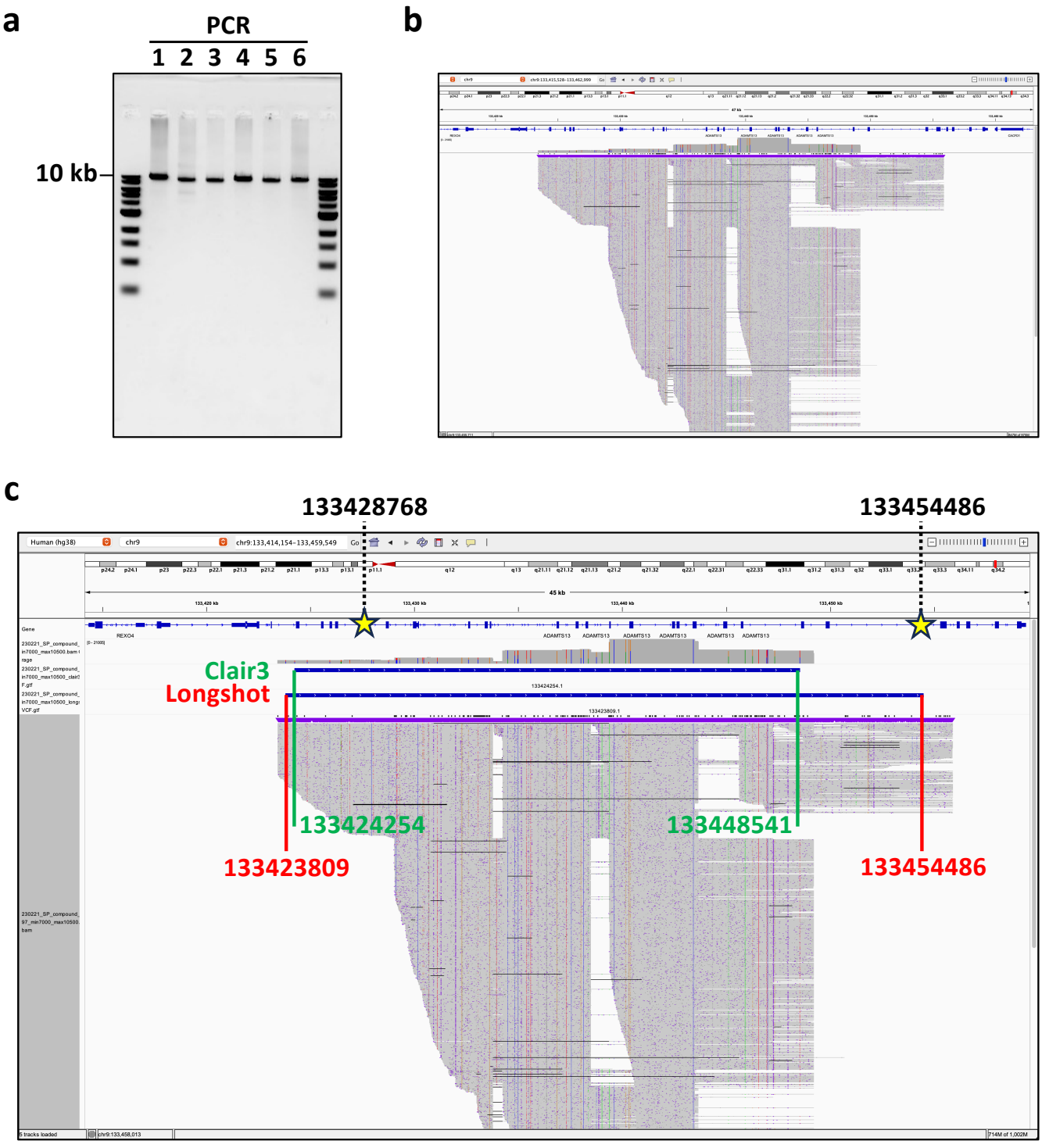
b

	Phased haplotype		TaKoHigh	Clair3	Longshot	Manual
Position of Chr.9	hap1	hap2	genotype	genotype	genotype	genotype
133423809	C	T	HET-1		HET-1	HET-1
133424254	C	T	HET-2	HET-2	HET-2	HET-2
133424860	T	C	HET-3	HET-3	HET-3	HET-3
133425943	T	C	HET-4	HET-4	HET-4	HET-4
133426474	G	A	HET-5		HET-5	HET-5
133426501	A	T	HET-6		HET-6	HET-6
133426901	G	T	HET-7		HET-7	HET-7
133427946	C	C	HOMO-1	HOMO-1	HOMO-1	HOMO-1
133428039	T	C	HET-8	HET-8	HET-8	HET-8
133428768	T	C	HET-9	HET-9	HET-9	HET-9
133429120	T	C	HET-10	HET-10		HET-10
133429155	T	C	HET-11			HET-11
133429185	T	C	HET-12			HET-12
133429303	G	C	HET-13			HET-13
133430170	C	C	HOMO-2	HOMO-2	HOMO-2	HOMO-2
133430414	G	T	HET-14			HET-14
133430458	T	C	HET-15			HET-15
133431111	C	T	HET-16			HET-16
133432286	G	G	HOMO-3	HOMO-3	HOMO-3	HOMO-3
133432616	G	C	HET-17			HET-17
133433061	T	C	HET-18			HET-18
133433609	G	G	HOMO-4	HOMO-4	HOMO-4	HOMO-4
133433956	T	T		HOMO-5		
133434495	A	C	HET-19	HET-19		HET-19
133434817	T	C	HET-20	HET-20		HET-20
133434976	C	G	HET-21			HET-21
133435002	T	C	HET-22	HET-22		HET-22
133435404	G	T	HET-23	HET-23		HET-23
133435592	C	T	HET-24	HET-24		HET-24
133436862	G	G	HOMO-6	HOMO-6		HOMO-6
133437314	G	T	HET-25	HET-25		HET-25
133437563	C	C	HOMO-7	HOMO-7		HOMO-7
133438003	C	G	HET-26			HET-26
133438133	C	C	HOMO-8	HOMO-8		HOMO-8
133438373	A	G	HET-27			HET-27
133438768	G	C	HET-28	HET-28	HET-28	HET-28
133439016	A	A	HOMO-9	HOMO-9	HOMO-9	HOMO-9
133439376	A	A	HOMO-10			HOMO-10
133439607	C	C	HOMO-11	HOMO-11		HOMO-11
133440154	A	G	HET-29	HET-29		HET-29
133440318	G	G	HOMO-12	HOMO-12		HOMO-12
133440409	G	C	HET-30	HET-30		HET-30
133443421	C	C	HOMO-13	HOMO-13		HOMO-13
133445114	A	G	HET-31	HET-31		HET-31
133445896	A	A	HOMO-14	HOMO-14	HOMO-14	HOMO-14
133446270	A	G	HET-32	HET-32	HET-32	HET-32
133446558	T	T	HOMO-15	HOMO-15	HOMO-15	HOMO-15
133446943	T	C	HET-33	HET-33	HET-33	HET-33
133446953	G	T	HET-34	HET-34	HET-34	HET-34
133446998	C	C	HOMO-16	HOMO-16	HOMO-16	HOMO-16
133447335	A	G	HET-35	HET-35	HET-35	HET-35
133448541	T	C	HET-36	HET-36	HET-36	HET-36
133449556	C	G	HET-37		HET-37	HET-37
133451246	C	A	HET-38		HET-38	HET-38
133451821	G	G	HOMO-17	HOMO-17	HOMO-17	HOMO-17
133453946	T	C	HET-39		HET-39	HET-39
133454166	G	C	HET-40		HET-40	HET-40
133454486	A	G	HET-41		HET-41	HET-41

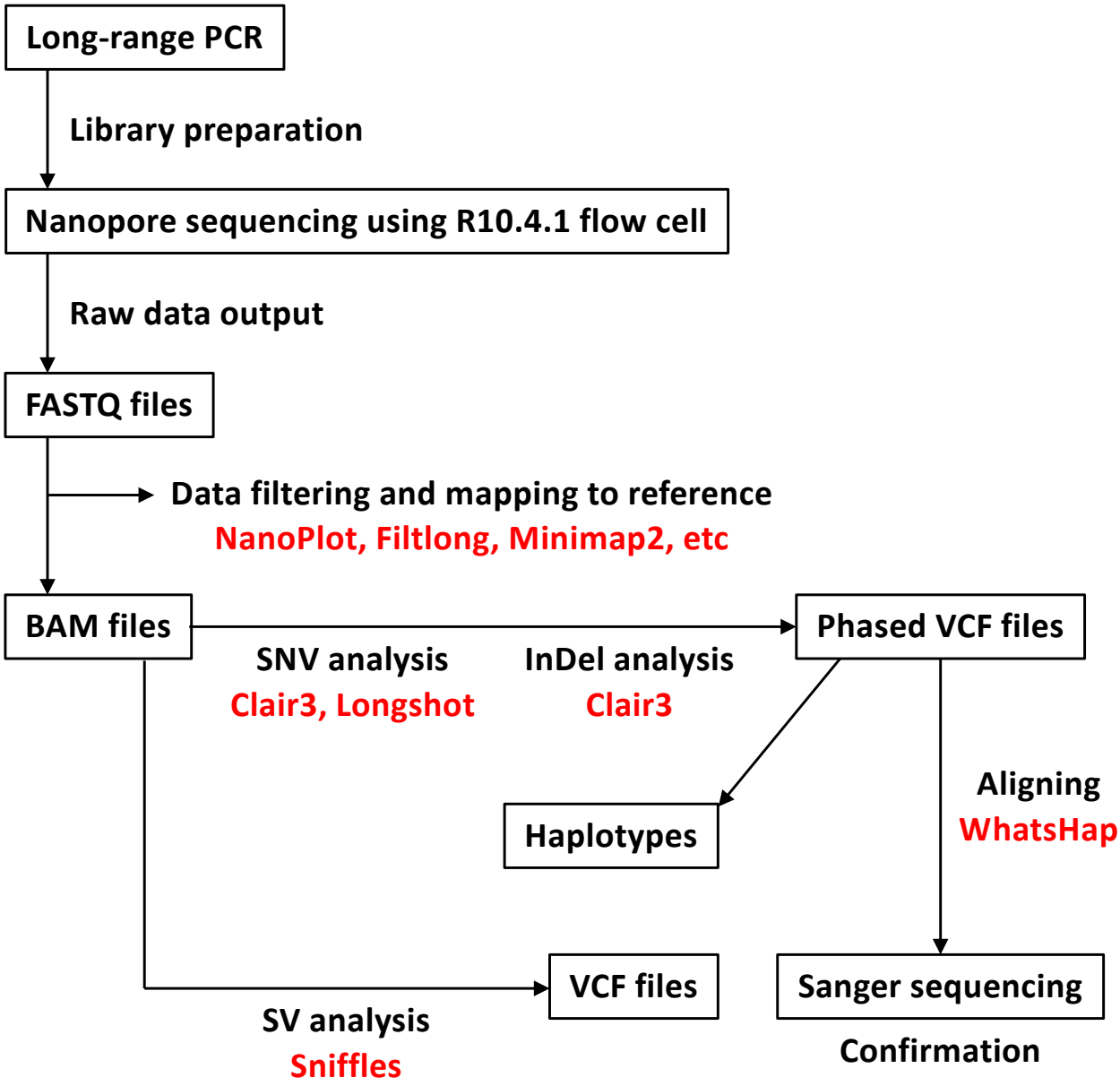
c

	Phased haplotype		TaKoHigh						
Position of Chr.9	hap1	hap2	All	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6
133423809	C	T	HET-1	HET-1					
133424254	C	T	HET-2	HET-2					
133424860	T	C	HET-3	HET-3					
133425943	T	C	HET-4	HET-4					
133426474	G	A	HET-5	HET-5					
133426501	A	T	HET-6	HET-6					
133426901	G	T	HET-7	HET-7					
133427946	C	C	HOMO-1	HOMO-1					
133428039	T	C	HET-8	HET-8					
133428768	T	C	HET-9	HET-9					
133429120	T	C	HET-10	HET-10					
133429155	T	C	HET-11	HET-11					
133429185	T	C	HET-12	HET-12					
133429303	G	C	HET-13	HET-13					
133430170	C	C	HOMO-2	HOMO-2					
133430414	G	T	HET-14	HET-14					
133430458	T	C	HET-15	HET-15					
133431111	C	T	HET-16	HET-16					
133432286	G	G	HOMO-3	HOMO-3					
133432616	G	C	HET-17	HET-17					
133433061	T	C	HET-18	HET-18					
133433609	G	G	HOMO-4	HOMO-4					
133433956	T	T							
133434495	A	C	HET-19			HET-19			
133434817	T	C	HET-20			HET-20			
133434976	C	G	HET-21			HET-21			
133435002	T	C	HET-22			HET-22			
133435404	G	T	HET-23			HET-23			
133435592	C	T	HET-24			HET-24			
133436862	G	G	HOMO-6			HOMO-6			
133437314	G	T	HET-25			HET-25	HET-25		
133437563	C	C	HOMO-7			HOMO-7	HOMO-7		
133438003	C	G	HET-26			HET-26	HET-26		
133438133	C	C	HOMO-8			HOMO-8	HOMO-8		
133438373	A	G	HET-27			HET-27	HET-27		
133438768	G	C	HET-28			HET-28	HET-28		
133439016	A	A	HOMO-9			HOMO-9	HOMO-9		
133439376	A	A	HOMO-10			HOMO-10	HOMO-10		
133439607	C	C	HOMO-11			HOMO-11	HOMO-11	HOMO-11	
133440154	A	G	HET-29			HET-29	HET-29	HET-29	
133440318	G	G	HOMO-12			HOMO-12	HOMO-12	HOMO-12	
133440409	G	C	HET-30			HET-30	HET-30	HET-30	
133443421	C	C	HOMO-13			HOMO-13	HOMO-13	HOMO-13	
133445114	A	G	HET-31				HET-31	HET-31	
133445896	A	A	HOMO-14				HOMO-14	HOMO-14	HOMO-14
133446270	A	G	HET-32				HET-32	HET-32	HET-32
133446558	T	T	HOMO-15				HOMO-15	HOMO-15	HOMO-15
133446943	T	C	HET-33				HET-33	HET-33	HET-33
133446953	G	T	HET-34				HET-34	HET-34	HET-34
133446998	C	C	HOMO-16				HOMO-16	HOMO-16	HOMO-16
133447335	A	G	HET-35					HET-35	HET-35
133448541	T	C	HET-36					HET-36	HET-36
133449556	C	G	HET-37						HET-37
133451246	C	A	HET-38						HET-38
133451821	G	G	HOMO-17						HOMO-17
133453946	T	C	HET-39						HET-39
133454166	G	C	HET-40						HET-40
133454486	A	G	HET-41						HET-41

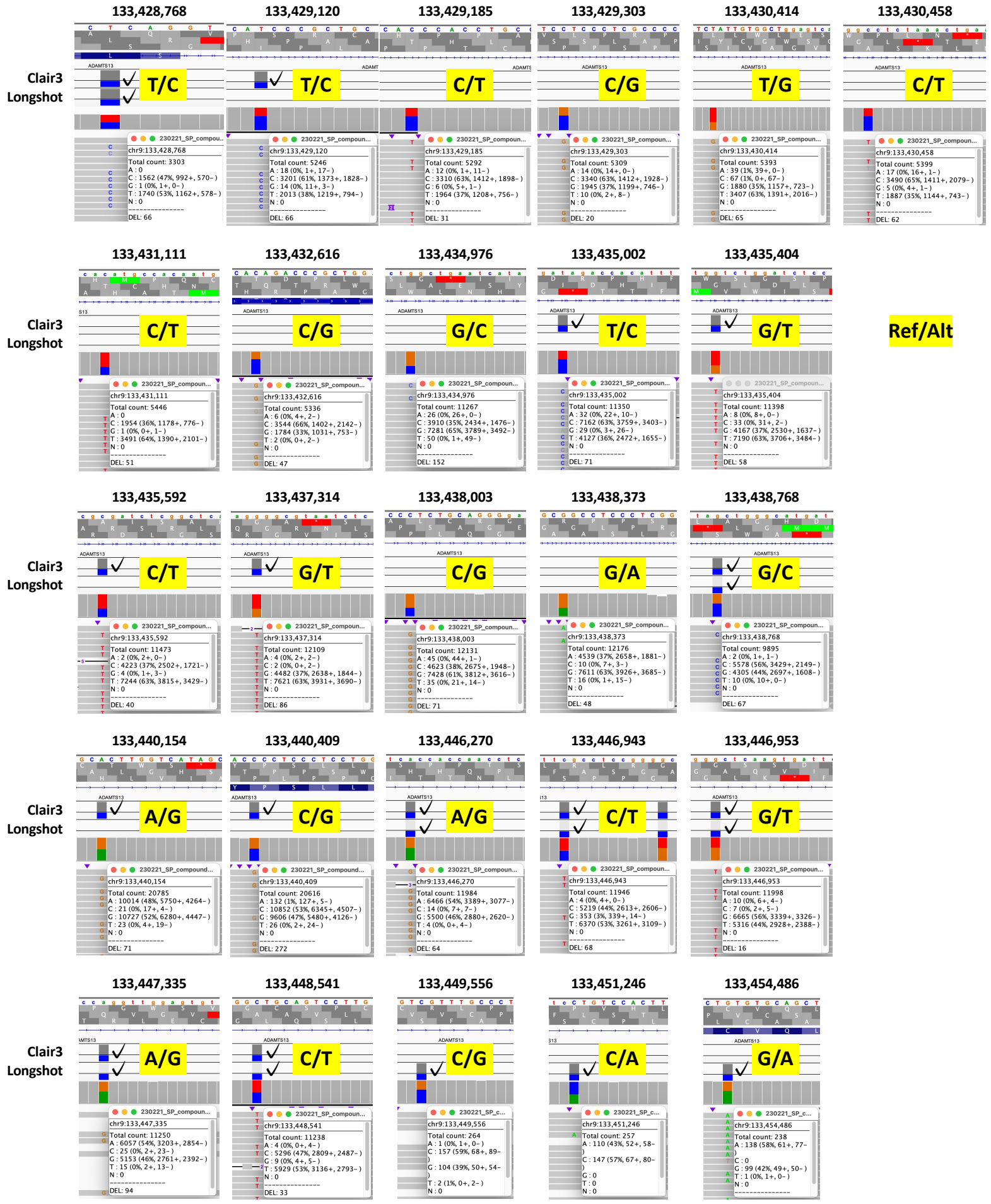
Fig. 5



Extended Data Figure 1. Validation of PCR amplicons, BAM files, and haplotypes for patient USS-S.
a, Electrophoresis of PCR amplicons (PCR1–PCR6) on a 1.2% agarose gel stained with SYBR Green. **b**, Visualization of BAM files mapped to the reference genome using IGV. **c**, Phased haplotypes (phase sets) displayed in IGV. Phased VCF files generated by Clair3 and Longshot were processed with WhatsHap to construct haplotypes. Haplotypes (phase sets) are shown in blue, with start and end positions marked in green (Clair3) or red (Longshot). Yellow stars indicate the two key variant positions used to assess compound heterozygosity. BAM files and haplotype tracks are overlaid in IGV for comparison.

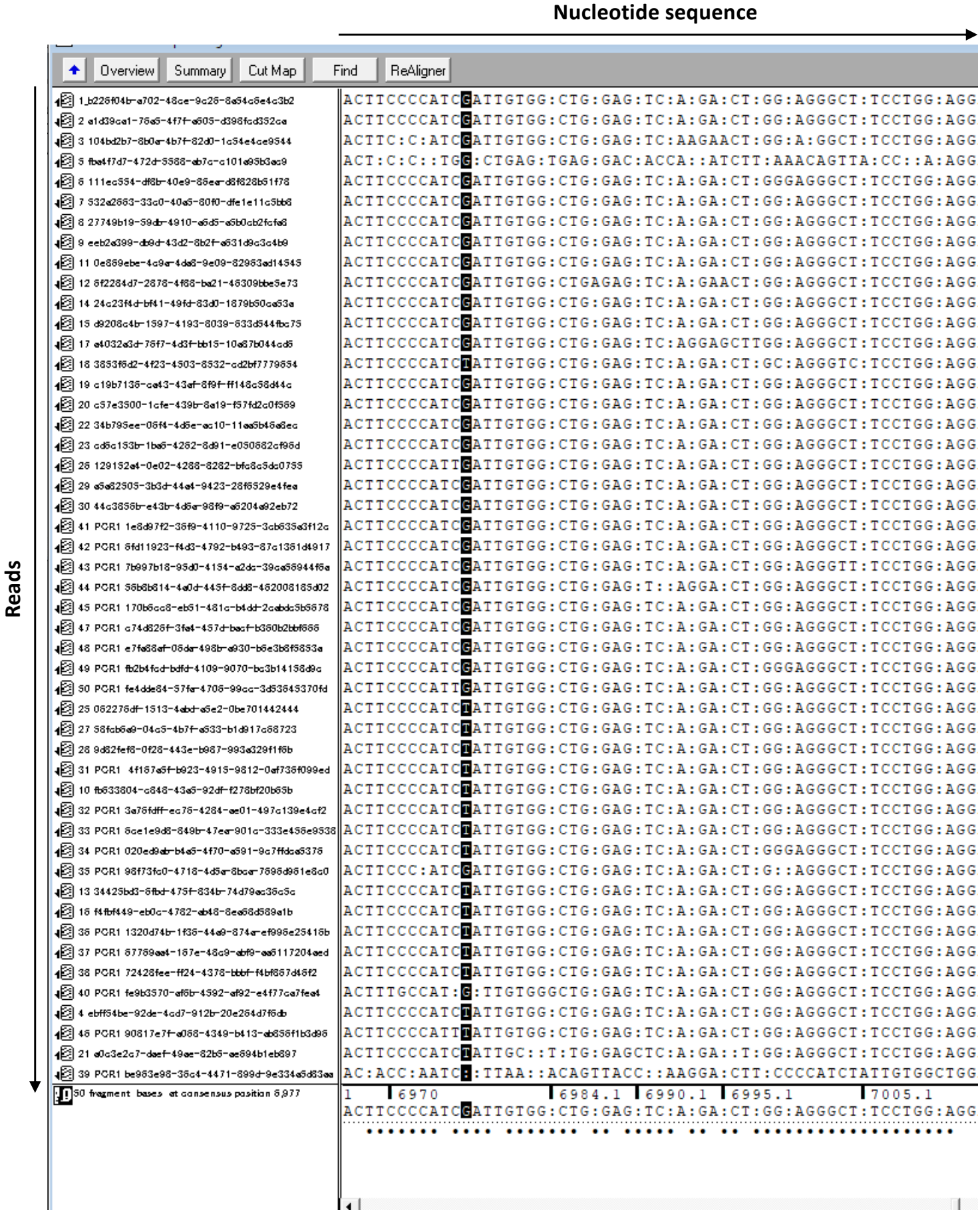


Extended Data Figure 2. Overview of the bioinformatic pipeline for PCR-based long-read sequencing.
Long-range PCR amplicons were used to generate sequencing libraries for nanopore-based long-read sequencing. The resulting FASTQ files were processed using a combination of tools: NanoPlot for quality assessment, Filtlong for read filtering, and Minimap2 for alignment to the reference genome. Variant calling and phasing were performed using Clair3 and Longshot, followed by making haplotypes with WhatsHap. Single-nucleotide variants (SNVs) and structural variants (SVs) were output as VCF files, and selected variants were validated by Sanger sequencing.

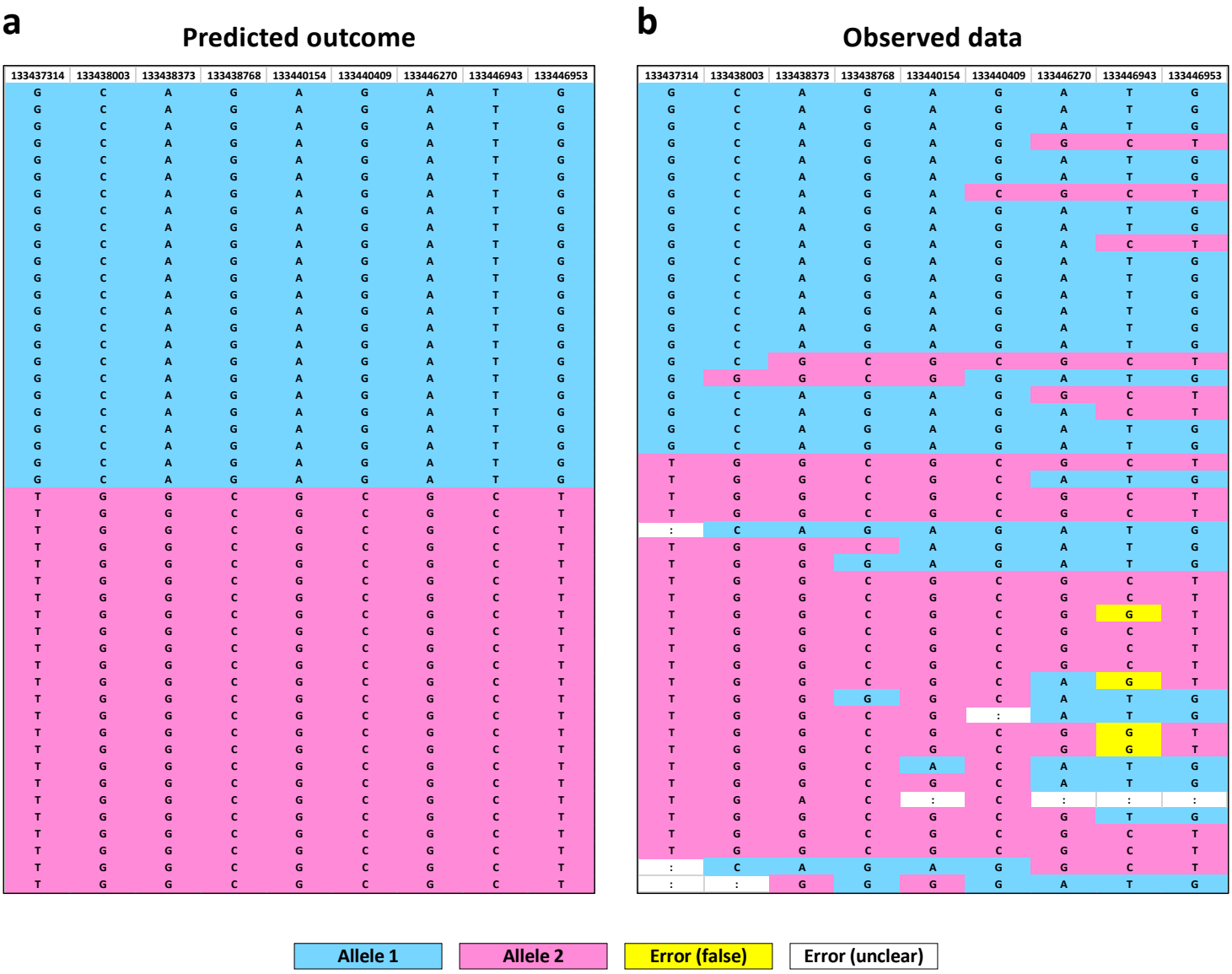


Extended Data Figure 3. Heterozygous SNVs used for manual phasing of USS-S data.

A total of 26 heterozygous SNVs were selected from BAM files using IGV with default settings. These SNVs were chosen to ensure that at least one overlapped with each amplicon. Reference and alternative alleles are highlighted in yellow. Checkmarks indicate successful variant calls by Clair3 and/or Longshot.

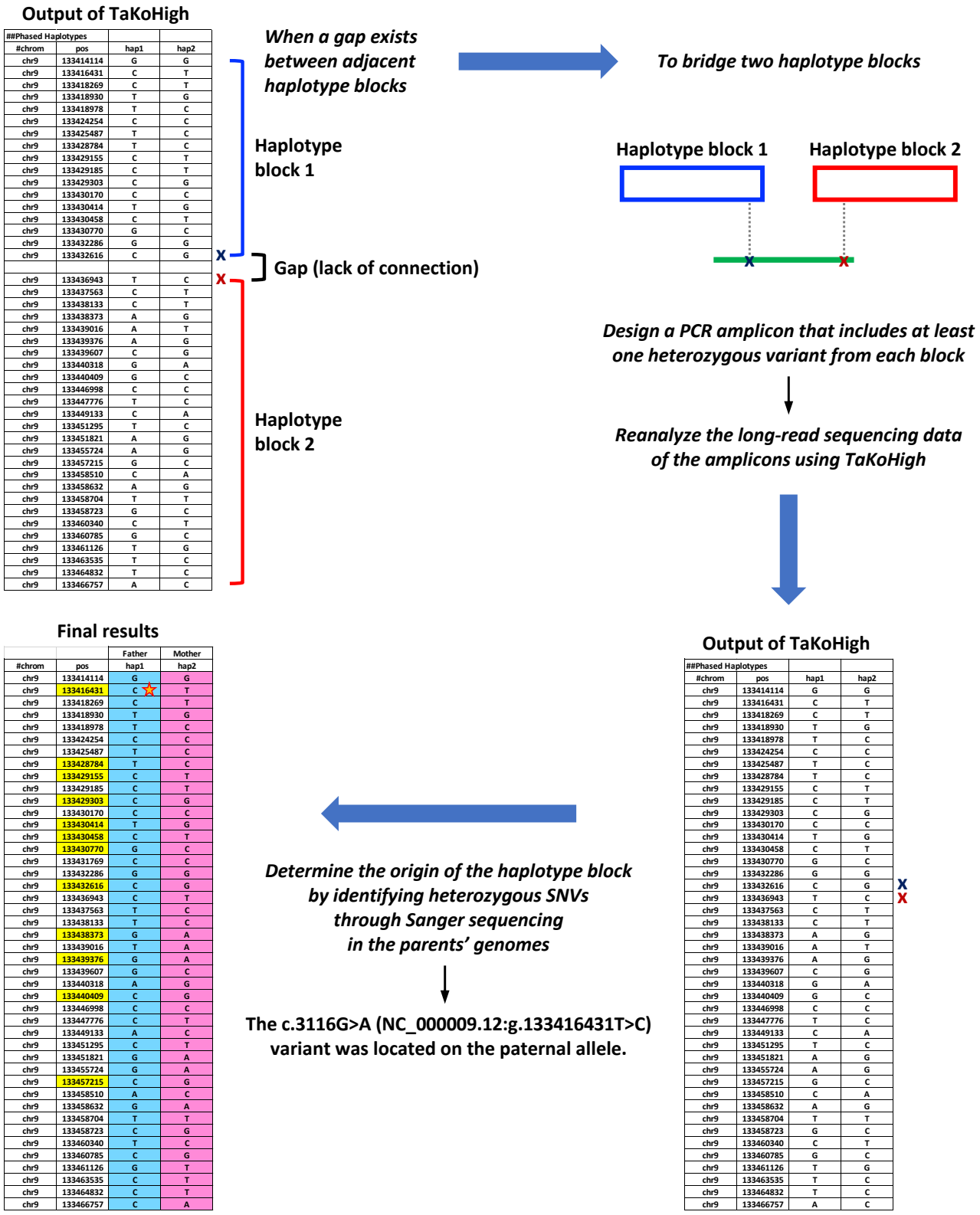


Extended Data Figure 4. Validation of selected heterozygous SNVs in individual reads using Sequencher.
Reads collected for each amplicon were imported into Sequencher and assembled against the corresponding reference sequence. The genotypes of the selected heterozygous SNVs were visually verified in each read. These results correspond to the SNVs shown in Fig. 2C and Extended Data Fig. 3.



Extended Data Figure 5. Illustration of phasing challenges in PCR amplicons during manual haplotype construction.

Manual phasing of the PCR4 amplicon from patient USS-S is shown as a representative case. **a**, Predicted outcome: Two alleles are cleanly separated based on heterozygous SNVs, with balanced amplification and no sequencing errors or PCR chimeras. **b**, Observed data: Chimeric reads, allele-specific amplification bias, and sequencing errors introduce inconsistencies, that prevent accurate haplotype reconstruction. Allele 1 is shown in blue, allele 2 in pink, false calls in yellow, and uncertain bases in white.



Extended Data Figure 6. Resolution of unconnected haplotypes in patient USS-2M by additional PCR and TaKoHigh analysis.

Initial analysis of the USS-2M data revealed two haplotypes (phase sets) due to the absence of overlapping heterozygous SNVs in one region. To bridge this gap, an additional amplicon was designed to include one heterozygous SNV from each haplotype. Sequencing and reanalysis with TaKoHigh successfully merged the haplotypes into a single contiguous phased region. Among the variants detected, one was extremely rare (NC_000009.12:g.133416431T>C; AF = 0.000013 in ToMMo60KJPN). Phasing showed that it resided on the paternal allele, which already carried a known causal variant (c.3099dupT, not detected by TaKoHigh because it is an InDel). Thus, no additional pathogenic variant was identified on the maternal allele. To validate the merged haplotype, 12 heterozygous SNVs—including the rare variant (highlighted in yellow)—were tested via Sanger sequencing of parental DNA. All results were concordant.

Clair3

	Phased haplotypes		Clair3						
Position in Chr.9	hap1	hap2	Total	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6
133423809	C	T							
133424254	C	T	HET-2	HET-2					
133424860	T	C	HET-3	HET-3					
133425943	T	C	HET-4	HET-4					
133426474	G	A							
133426501	A	T							
133426901	G	T							
133427946	C	C	HOMO-1	HOMO-1					
133428039	T	C	HET-8						
133428768	T	C	HET-9	HET-9					
133429120	T	C	HET-10	HET-10					
133429155	T	C							
133429185	T	C		HET-12					
133429303	G	C		HET-13					
133430170	C	C	HOMO-2	HOMO-2	HOMO-2				
133430414	G	T							
133430458	T	C							
133431111	C	T							
133432286	G	G	HOMO-3	HOMO-3	HOMO-3				
133432616	G	C							
133433061	T	C							
133433609	G	G	HOMO-4	HOMO-4	HOMO-4				
133433956	T	T	HOMO-5		HOMO-5				
133434495	A	C	HET-19			HET-19			
133434817	T	C	HET-20			HET-20			
133434976	C	G				HET-21			
133435002	T	C	HET-22			HET-22			
133435404	G	T	HET-23		HET-23	HET-23			
133435592	C	T	HET-24			HET-24			
133436862	G	G	HOMO-6		HOMO-6	HOMO-6			
133437314	G	T	HET-25			HET-25			
133437563	C	C	HOMO-7		HOMO-7	HOMO-7	HOMO7		
133438003	C	G			HET-26	HET-26			
133438133	C	C	HOMO-8		HOMO-8	HOMO-8	HOMO-8		
133438373	A	G				HET-27			
133438768	G	C	HET-28			HET-28			
133439016	A	A	HOMO-9			HOMO-9	HOMO-9		
133439376	A	A				HOMO-10	HOMO-10		
133439607	C	C	HOMO-11			HOMO-11	HOMO-11	HOMO-11	
133440154	A	G	HET-29			HET-29			
133440318	G	G	HOMO-12			HOMO-12	HOMO-12	HOMO-12	
133440409	G	C	HET-30				HET-30	HET-30	
133443421	C	C	HOMO-13			HOMO-13	HOMO-13	HOMO-13	
133445114	A	G	HET-31						
133445896	A	A	HOMO-14				HOMO-14	HOMO-14	HOMO-14
133446270	A	G	HET-32						
133446558	T	T	HOMO-15				HOMO-15	HOMO-15	HOMO-15
133446943	T	C	HET-33					HET-33	
133446953	G	T	HET-34						
133446998	C	C	HOMO-16				HOMO-16	HOMO-16	HOMO-16
133447335	A	G	HET-35						
133448541	T	C	HET-36					HET-36	
133449556	C	G							
133451246	C	A							
133451821	G	G	HOMO-17						HOMO-17
133453946	T	C							HET-39
133454166	G	C							
133454486	A	G							

Longshot

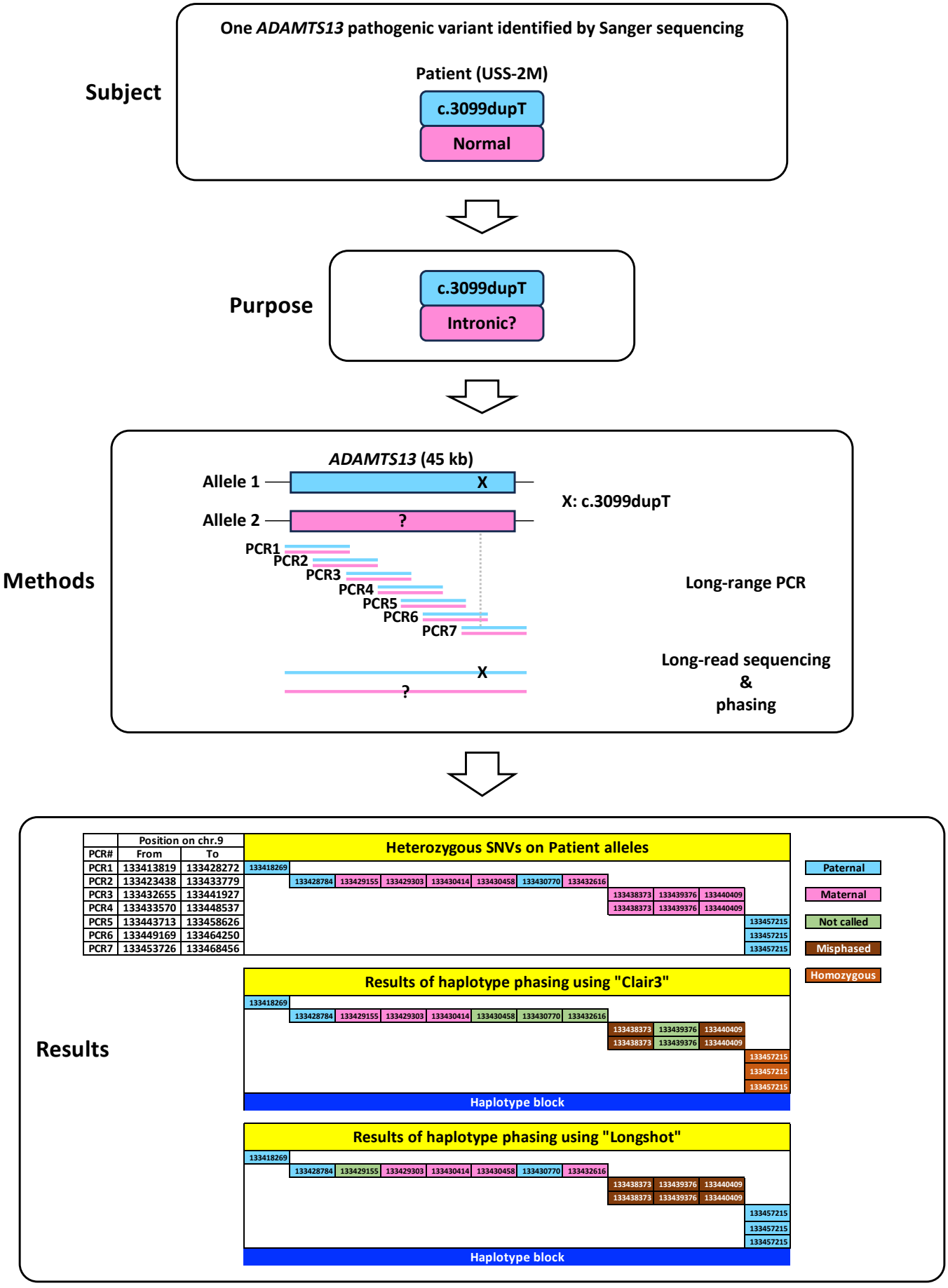
	Phased haplotypes		Longshot						
Position in Chr.9	hap1	hap2	Total	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6
133423809	C	T	HET-1	HET-1					
133424254	C	T	HET-2	HET-2					
133424860	T	C	HET-3	HET-3					
133425943	T	C	HET-4	HET-4					
133426474	G	A	HET-5	HET-5					
133426501	A	T	HET-6	HET-6					
133426901	G	T	HET-7	HET-7					
133427946	C	C	HOMO-1	HOMO-1					
133428039	T	C	HET-8	HET-8					
133428768	T	C	HET-9	HET-9					
133429120	T	C		HET-10					
133429155	T	C		HET-11					
133429185	T	C		HET-12					
133429303	G	C		HET-13					
133430170	C	C	HOMO-2	HOMO-2	HOMO-2				
133430414	G	T		HET-14					
133430458	T	C		HET-15					
133431111	C	T		HET-16					
133432286	G	G	HOMO-3	HOMO-3	HOMO-3				
133432616	G	C		HET-17					
133433061	T	C		HET-18					
133433609	G	G	HOMO-4		HOMO-4				
133433956	T	T							
133434495	A	C			HET-19				
133434817	T	C			HET-20				
133434976	C	G							
133435002	T	C							
133435404	G	T			HET-23				
133435592	C	T			HET-24				
133436862	G	G			HOMO-6				
133437314	G	T			HET-25		HET-25		
133437563	C	C			HOMO-7		HOMO-7		
133438003	C	G			HET-26		HET-26		
133438133	C	C			HOMO-8		HOMO-8		
133438373	A	G			HET-27		HET-27		
133438768	G	C	HET-28				HET-28		
133439016	A	A	HOMO-9				HOMO-9		
133439376	A	A					HOMO-10		
133439607	C	C					HOMO-11		
133440154	A	G					HET-29		
133440318	G	G					HOMO-12		
133440409	G	C					HET-30		
133443421	C	C					HOMO-13		
133445114	A	G					HET-31		
133445896	A	A	HOMO-14				HOMO-14		HOMO-14
133446270	A	G	HET-32				HET-32		HET-32
133446558	T	T	HOMO-15				HOMO-15		HOMO-15
133446943	T	C	HET-33				HET-33		HET-33
133446953	G	T	HET-34				HET-34		HET-34
133446998	C	C	HOMO-16				HOMO-16		HOMO-16
133447335	A	G	HET-35						HET-35
133448541	T	C	HET-36						HET-36
133449556	C	G	HET-37						HET-37
133451246	C	A	HET-38						HET-38
133451821	G	G	HOMO-17						HOMO-17
133453946	T	C	HET-39						HET-39
133454166	G	C	HET-40						HET-40
133454486	A	G	HET-41						HET-41

Manual inspection

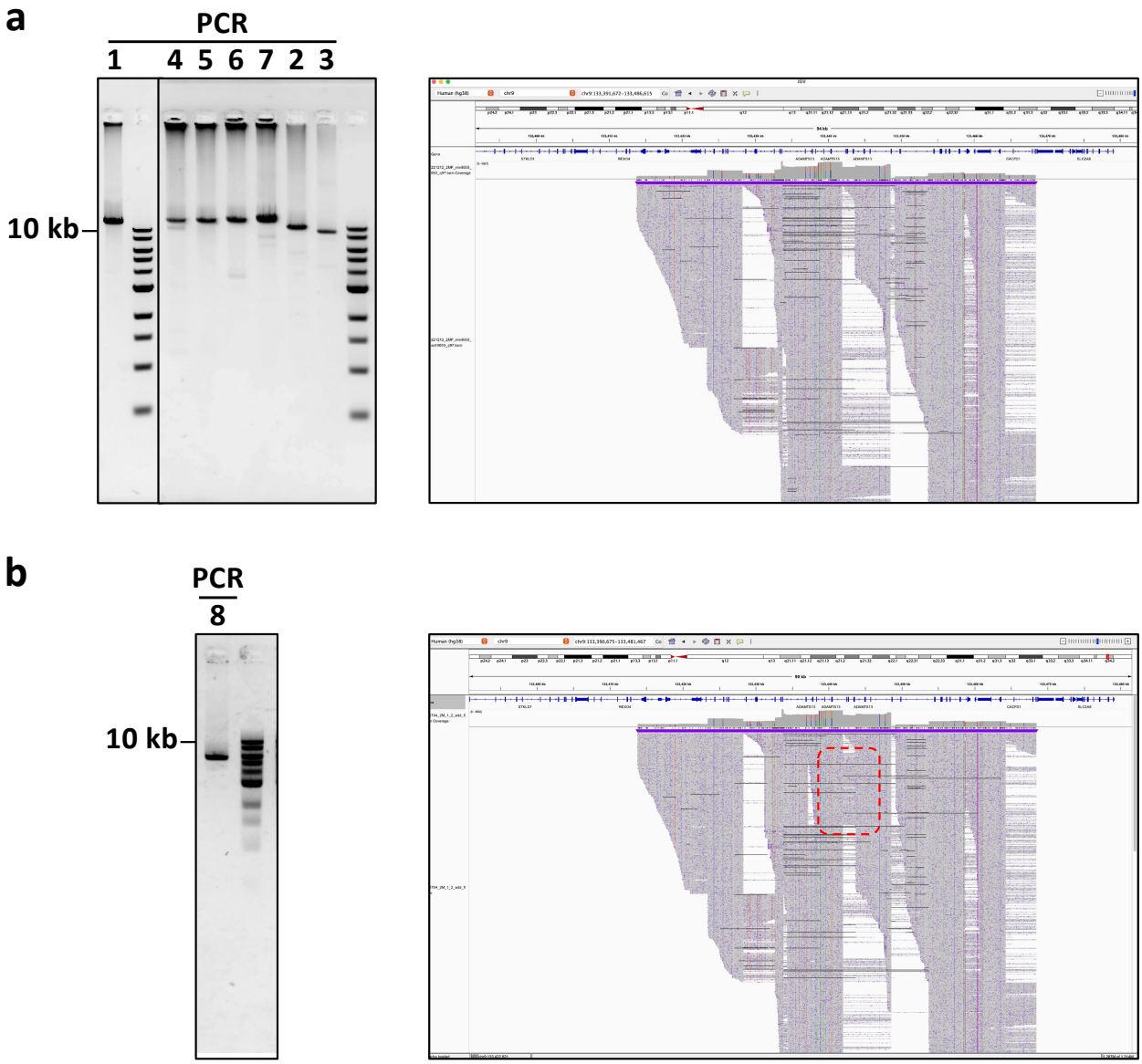
	Phased haplotypes		BAM data observed by IGV (default set)						
Position in Chr.9	hap1	hap2	Total	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6
133423809	C	T	HET-1	HET-1					
133424254	C	T	HET-2	HET-2					
133424860	T	C	HET-3	HET-3					
133425943	T	C	HET-4	HET-4					
133426474	G	A	HET-5	HET-5					
133426501	A	T	HET-6	HET-6					
133426901	G	T	HET-7	HET-7					
133427946	C	C	HOMO-1	HOMO-1					
133428039	T	C	HET-8	HET-8					
133428768	T	C	HET-9	HET-9					
133429120	T	C	HET-10	HET-10	HET-10				
133429155	T	C	HET-11	HET-11					
133429185	T	C	HET-12	HET-12					
133429303	G	C	HET-13	HET-13					
133430170	C	C	HOMO-2	HOMO-2	HOMO-2				
133430414	G	T	HET-14	HET-14					
133430458	T	C	HET-15	HET-15					
133431111	C	T	HET-16	HET-16	HET-16				
133432286	G	G	HOMO-3	HOMO-3	HOMO-3				
133432616	G	C	HET-17	HET-17					
133433061	T	C	HET-18	HET-18					
133433609	G	G	HOMO-4	HOMO-4	HOMO-4				
133433956	T	T							
133434495	A	C	HET-19		HET-19	HET-19			
133434817	T	C	HET-20		HET-20	HET-20			
133434976	C	G	HET-21			HET-21			
133435002	T	C	HET-22		HET-22	HET-22			
133435404	G	T	HET-23		HET-23	HET-23			
133435592	C	T	HET-24		HET-24	HET-24			
133436862	G	G	HOMO-6		HOMO-6	HOMO-6			
133437314	G	T	HET-25		HET-25	HET-25	HET-25		
133437563	C	C	HOMO-7		HOMO-7	HOMO-7	HOMO-7		
133438003	C	G	HET-26		HET-26	HET-26	HET-26		
133438133	C	C	HOMO-8		HOMO-8	HOMO-8	HOMO-8		
133438373	A	G	HET-27			HET-27	HET-27		
133438768	G	C	HET-28			HET-28	HET-28		
133439016	A	A	HOMO-9			HOMO-9	HOMO-9		
133439376	A	A	HOMO-10			HOMO-10	HOMO-10		
133439607	C	C	HOMO-11			HOMO-11	HOMO-11	HOMO-11	
133440154	A	G	HET-29			HET-29	HET-29	HET-29	
133440318	G	G	HOMO-12			HOMO-12	HOMO-12	HOMO-12	
133440409	G	C	HET-30			HET-30	HET-30	HET-30	
133443421	C	C	HOMO-13			HOMO-13	HOMO-13	HOMO-13	
133445114	A	G	HET-31				HET-31	HET-31	
133445896	A	A	HOMO-14				HOMO-14	HOMO-14	HOMO-14
133446270	A	G	HET-32				HET-32	HET-32	HET-32
133446558	T	T	HOMO-15				HOMO-15	HOMO-15	HOMO-15
133446943	T	C	HET-33				HET-33	HET-33	HET-33
133446953	G	T	HET-34				HET-34	HET-34	HET-34
133446998	C	C	HOMO-16				HOMO-16	HOMO-16	HOMO-16
133447335	A	G	HET-35					HET-35	HET-35
133448541	T	C	HET-36					HET-36	HET-36
133449556	C	G	HET-37						HET-37
133451246	C	A	HET-38						HET-38
133451821	G	G	HOMO-17						HOMO-17
133453946	T	C	HET-39						HET-39
133454166	G	C	HET-40						HET-40
133454486	A	G	HET-41						HET-41

Extended Data Figure 7. Variant calling results from Longshot, Clair3, and manual inspection in IGV.

For patient USS-S, BAM files were divided by amplicon and analyzed using Longshot, Clair3, and manual inspection in IGV (default settings). This figure shows the distribution and accuracy of variant calls for individual PCR amplicons. The results indicate that certain regions lacked sufficient variant information or showed inconsistencies among tools.

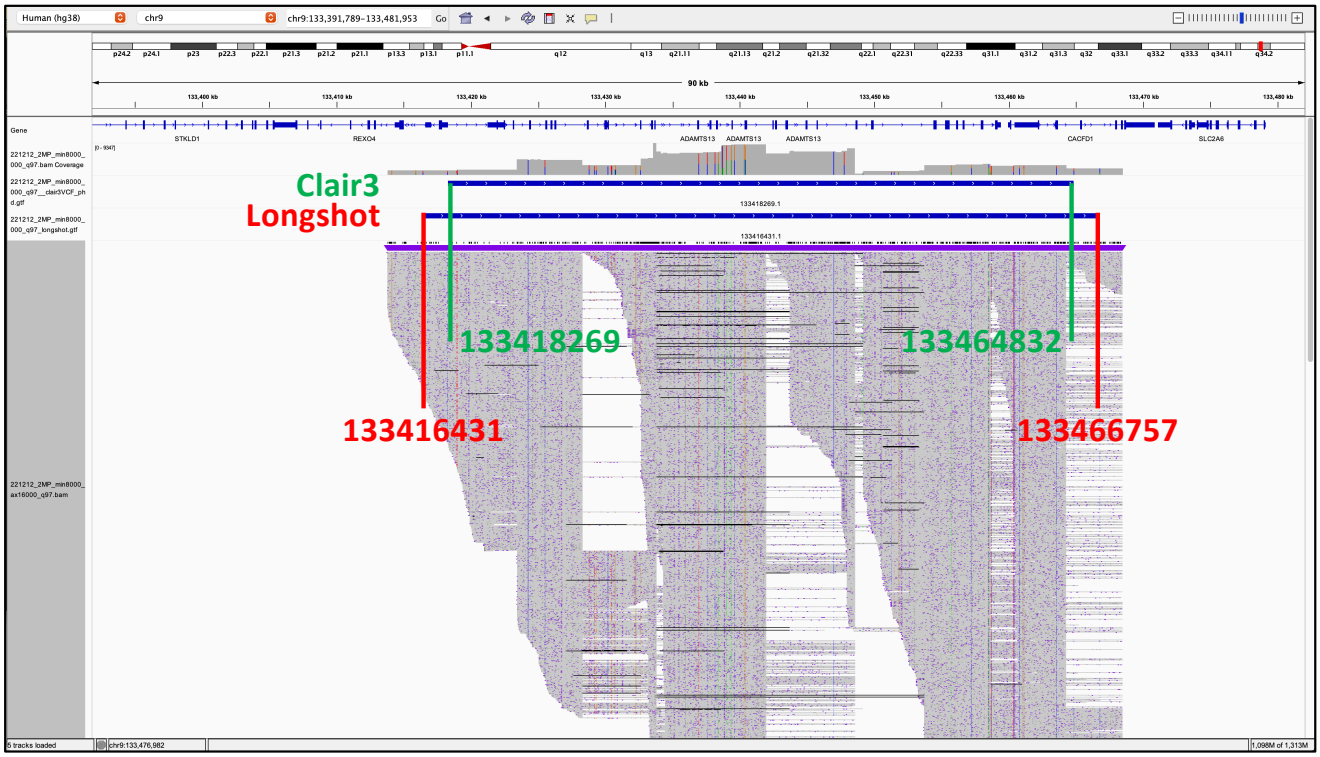


Extended Data Figure 8. PCR-based LRS to detect a causative variant on the maternal allele in patient USS-2M. While Sanger sequencing confirmed a known pathogenic variant on the paternal allele, no variant was found on the maternal allele. To investigate, we performed PCR-based long-read sequencing of *ADAMTS13*. Although variant calling and phasing were attempted with Clair3 and Longshot, results remained incomplete. Parental origin of selected heterozygous SNVs was determined by Sanger sequencing, revealing that both tools missed or miscalled variants, likely contributing to phasing failure.



Extended Data Figure 9. Validation of PCR amplicons and BAM file visualization for patient USS-2M.

a, PCR amplicons (PCR1–PCR7) were confirmed by 1.2% agarose gel electrophoresis with SYBR Green staining. **b**, An additional PCR (PCR3), designed to bridge the haplotype gap, was sequenced and visualized using IGV. Analysis confirmed successful integration of this region using TaKoHigh.



Extended Data Figure 10. Haplotypes of patient USS-2M visualized in IGV.

Phased VCF files generated by Clair3 and Longshot were used as input for WhatsHap to construct haplotypes. These haplotypes are shown in blue in IGV, with start and end points marked in green (Clair3) and red (Longshot). BAM files were displayed alongside for direct comparison and phasing verification.