biology

MDPI

Article

Targeted Double-Stranded cDNA Sequencing-Based Phase Analysis to Identify Compound Heterozygous Mutations and Differential Allelic Expression

Hiroki Ura [1,2,*], Sumihito Togi [1,2] and Yo Niida [1,2]

[1] Center for Clinical Genomics, Kanazawa Medical University Hospital, 1-1 Daigaku, Uchinada, Kahoku, Ishikawa 920-0923, Japan; togi@kanazawa-med.ac.jp (S.T.); niida@kanazawa-med.ac.jp (Y.N.)
[2] Division of Genomic Medicine, Department of Advanced Medicine, Medical Research Institute, Kanazawa Medical University, 1-1 Daigaku, Uchinada, Kahoku, Ishikawa 920-0923, Japan
* Correspondence: h-ura@kanazawa-med.ac.jp; Tel.: +81-076-286-2211 (ext. 8353)

Simple Summary: Phase analysis to distinguish between *in cis* and *in trans* heterozygous mutations is important for clinical diagnosis because *in trans* compound heterozygous mutations cause autosomal recessive diseases. However, conventional phase analysis is limited because of the large target size of genomic DNA. Here, we performed a targeted double-stranded cDNA sequencing-based phase analysis to resolve the limitation of distance using direct adapter ligation library preparation and paired-end sequencing; we elucidated that two heterozygous mutations on a patient with Wilson disease are *in trans* compound heterozygous mutations. Furthermore, we detected the differential allelic expression. Our results indicate that a targeted double-stranded cDNA sequencing-based phase analysis is useful for determining compound heterozygous mutations and confers information on allelic expression.

Abstract: There are two combinations of heterozygous mutation, i.e., *in trans*, which carries mutations on different alleles, and *in cis*, which carries mutations on the same allele. Because only *in trans* compound heterozygous mutations have been implicated in autosomal recessive diseases, it is important to distinguish them for clinical diagnosis. However, conventional phase analysis is limited because of the large target size of genomic DNA. Here, we performed a genetic analysis on a patient with Wilson disease, and we detected two heterozygous mutations chr13:51958362;G>GG (NM_000053.4:c.2304dup r.2304dup p.Met769HisfsTer26) and chr13:51964900;C>T (NM_000053.4:c.1841G>A r.1841g>a p.Gly614Asp) in the causative gene *ATP7B*. The distance between the two mutations was 6.5 kb in genomic DNA but 464 bp in mRNA. Targeted double-stranded cDNA sequencing-based phase analysis was performed using direct adapter ligation library preparation and paired-end sequencing, and we elucidated they are *in trans* compound heterozygous mutations. Trio analysis showed that the mutation (chr13:51964900;C>T) derived from the father and the other mutation from the mother, validating that the mutations are *in trans* composition. Furthermore, targeted double-stranded cDNA sequencing-based phase analysis detected the differential allelic expression, suggesting that the mutation (chr13:51958362;G>GG) caused downregulation of expression by nonsense-mediated mRNA decay. Our results indicate that targeted double-stranded cDNA sequencing-based phase analysis is useful for determining compound heterozygous mutations and confers information on allelic expression.

Keywords: phase analysis; compound heterozygous mutation; next-generation sequencing; targeted double-stranded cDNA sequencing; allelic expression

Citation: Ura, H.; Togi, S.; Niida, Y. Targeted Double-Stranded cDNA Sequencing-Based Phase Analysis to Identify Compound Heterozygous Mutations and Differential Allelic Expression. *Biology* **2021**, *10*, 256. https://doi.org/10.3390/biology10040256

Received: 5 March 2021
Accepted: 22 March 2021
Published: 24 March 2021

1. Introduction

Next-generation sequencing (NGS) is a powerful technology used in the clinical field for genetic diagnosis [1–3]. At present, the use of NGS in clinical diagnosis is largely for

comprehensive analysis, such as whole-genome sequencing, whole-exome sequencing, and gene targeting panel sequencing. Phase analysis, which detects specific compound heterozygous mutations, is not commonly performed using NGS technology.

During genetic diagnosis, multiple heterozygous mutations could be detected at specific loci. There are two combinations of heterozygous mutations. *In cis* heterozygous loss-of-function mutation still retains one functionally active allele as both mutations are located at the same allele (Figure 1A), whereas *in trans* compound heterozygous mutation does not retain any functionally active alleles as each mutation is found at a different allele. A patient who has *in trans* compound heterozygous loss-of-function mutation will be affected by an autosomal recessive disease similar to a patient who has loss-of-function homozygous mutation. Therefore, the clinical diagnoses must distinguish between *in trans* and *in cis* heterozygous mutations when there are more than two heterozygous mutations at a particular gene locus.

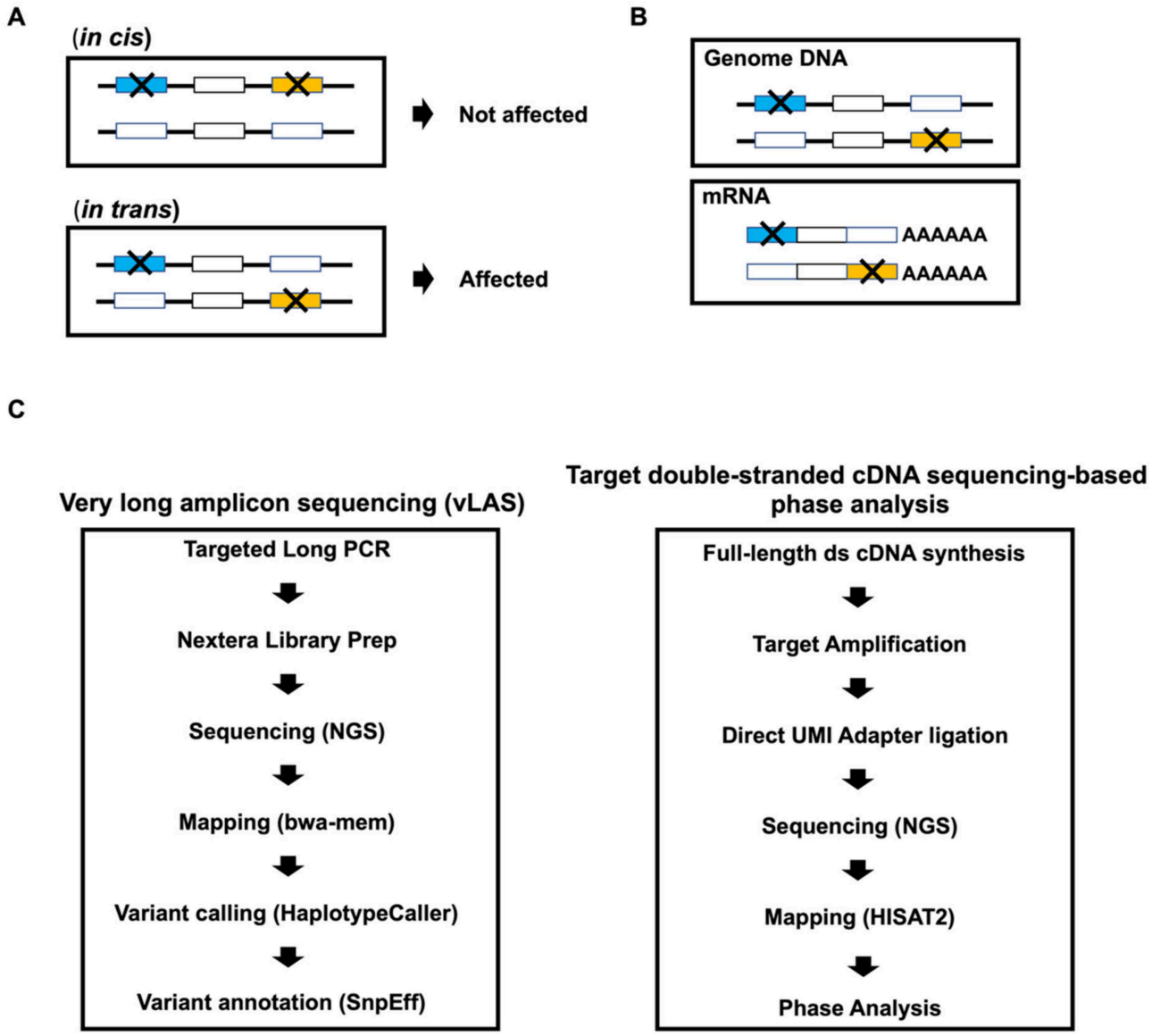


Figure 1. (**A**) Workflow for targeted double-stranded cDNA sequencing-based phase analysis. Types of compound heterozygous mutations (*in cis* and *in trans*); (**B**) Scheme of *in trans* compound heterozygous mutations on genomic DNA and mRNA; (**C**) Workflow for the detection of variants by very long amplicon sequencing (vLAS) and targeted double-stranded cDNA sequencing-based phase analysis.

However, a conventional phase analysis uses Sanger sequencing, which is limited by the distance between the two mutations and the required number of sequencing reads. The clone into the plasmid or fosmid must include the sequence of both mutations. Recently, an alternative approach for genetic diagnosis has been available with NGS technology of short read DNA sequencing (DNA-seq) [4–6]; however, it is also limited by the relatively short distances spanned by the reads. It is difficult to use except in cases of two heterozygous mutations in the same exon. To resolve the distance limitation, we used a targeted double-stranded cDNA sequencing-based phase analysis to detect two mutations in a single or paired-end read by removing large intron sequences by splicing (Figure 1B). In traditional RNA-seq, RNA is fragmented before cDNA synthesis [7]. To save the read including both mutations, we used the SMARTer RNA-seq method to first synthesize the full-length double-stranded cDNA once, and then fragment the double-stranded cDNA for NGS. Furthermore, the specific *ATP7B* double-stranded cDNA spanning two mutations was amplified from SMARTer cDNA that was, then, directly ligated to the adapter without fragmentation to ensure all paired-end reads contained both mutations.

Allele-specific expression of two alleles in a diploid individual may be potentially imbalanced, thereby contributing to phenotypic variation and disease pathophysiology among individuals [8,9]. Nonsense or frameshift mutations that induce nonsense-mediated mRNA decay (NMD) strongly affect imbalanced allelic expression due to targeted degradation [10]. There are approximately 30 million people with a genetic disorder worldwide, and it is estimated that about 30% of them have mutations with NMD-mediated differential allele expression [11]. Importantly, our targeted double-stranded cDNA sequencing-based phase analysis approach also provides information on allele-specific expression and may help for clinical diagnosis and provide a better understanding of the underlying molecular pathology.

Recently, third-generation sequencer such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) can be facilitated for phase analysis due to the production of long sequencing reads (>10 kb) [12–14]. The error rates of ONT and PacBio are relatively higher than Illumina NGS [15,16]. The long-read phase test software such as HapCUT and WhatsHap are required for read number and large number of reads including mutations for calculations using statistical algorithms [5,17]. Because these software calculate based on diploid genome, they are not suitable for mRNA; therefore, these software are suitable for comprehensive genome analysis but not for specific genome locus or mRNA. In this study, we verified our targeted double-stranded cDNA sequencing-based phase analysis in a patient with Wilson disease. First, we detected several variants in *ATP7B*, the responsible gene of Wilson disease, using a long PCR-based variant calling method (Figure 1C). Next, we compared the detection ability of compound heterozygous mutations between Nextera tagmentation and ThruPLEX direct adaptor ligation methods for library preparation. We also verified the detection ability of differential allelic expression by targeted double-stranded cDNA sequencing-based phase analysis.

2. Materials and Methods

2.1. Patient and Sample

A 40-year-old male was suspected of having Wilson disease because of low serum ceruloplasmin value, Kayser–Fleischer ring, and neuropsychiatric symptoms. After genetic counselling, peripheral blood was collected from the patient, and *ATP7B* testing was performed for a definitive diagnosis. Peripheral blood was also collected from the parents of the patient to confirm the compound heterozygous mutations.

2.2. Genomic DNA Extraction

We extracted all genomic DNA samples used in this study from peripheral whole blood using a rapid extraction method [18]. The DNA amount and optical density (A260/280 ratio) were measured using a Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA).

2.3. Very Long Amplicon Sequencing (vLAS)

Long-range PCR-based NGS, also known as vary long amplicon sequencing (vLAS), was performed at the *ATP7B* genomic region, as previously described [19]. Briefly, a set of very long-range PCR products (approximately 20 kb each) covering the entire gene locus was produced by KOD One (TOYOBO, Osaka, Japan) touchdown PCR. The long PCR primer sequences used in this study are shown in Supplementary Material, Table S1. An NGS library was prepared from purified PCR products using a Nextera Flex DNA kit (Illumina, San Diego, CA, USA), according to the manufacturer's protocol.

2.4. Total RNA Extraction and Full-Length Double-Stranded cDNA Synthesis

We extracted total RNA from peripheral blood mononuclear cells with TRIzol reagent (Thermo Fisher Scientific), according to the manufacturer's instructions. RNA concentration and purity were measured spectrophotometrically (Nanodrop, Thermo Fisher Scientific). The RNA integrity number was determined using a TapeStation 4200 with High Sensitivity RNA ScreenTape (Agilent Technologies, Santa Clara, CA, USA). Full-length double-stranded cDNA was synthesized from 50 ng of total RNA using a SMART-Seq® HT kit (Takara Bio USA, Mountain View, CA, USA), according to the manufacturer's standard protocol.

2.5. Library Preparation for Targeted Double-Stranded cDNA Based Sequencing

We amplified double-stranded cDNA of the *ATP7B* locus (621 bp) by harboring two pathogenic mutations with two specific primers targeting *ATP7B* exon 4/5 and exon 9 (exon 4/5 forward primer, 5′-acattgagctgacaatcacagg-3′ and exon 9 reverse primer, 5′-gagagacatgagtttagccagg-3′). The PCR product was purified using a PCR purification kit (Roche Diagnostics, Mannheim, Germany), and NGS libraries were prepared using either a Nextera XT DNA Library Prep kit (Illumina) or ThruPLEX® Tag-Seq kit (Takara Bio USA), according to the manufacturer's respective protocol.

2.6. Next Generation Sequencing

The libraries were quantified using the HS Qubit dsDNA assay (Thermo Fisher Scientific) and KAPA Library Quantification kit (KAPA Biosystems, Wilmington, MA, USA). According to the standard Illumina protocol, targeted double-stranded cDNA sequencing-based libraries were sequenced (2 × 250 bp) on an Illumina MiSeq. FASTQ files were generated using bcl2fastq (Illumina).

2.7. Data Analysis

The FASTQ files in vLAS were aligned to the reference human genome (hg38) using a Burrows-Wheeler Aligner MEM algorithm (BWA-MEM version 0.7.17-r1188) [20]. Haplotype variants were identified using GATK HaplotypeCaller (version 4.0.6.0) [21]. For analysis and interpretation, we used the following software packages: SAMtools (version 1.9), BEDTools (version v2.27.1), vcftools (version 0.1.16), and Integrative Genomic Viewer (IGV 2.4.13), and analysis approach as described previously [22–27]. For variant annotation, we used the following databases: SnpEff (version SnpEff 4.3t), dbSNP (version 151), TOPMED, ClinVar, Human Genetic Variation Database (HGVD), and ToMMo (version 3.5) [28–31]. For in silico analysis, we used dbNSFP (v3.2) that complies a prediction score from 29 prediction algorithms [32]. The original vLAS data presented in this study are available on request from the corresponding author. The vLAS data are not publicly available due to the personal information protection law in Japanese. The raw data of targeted double-stranded cDNA sequencing-based phase analysis in this study cannot be identified by individual information due to short sequence size and have been deposited in the Sequence Read Archive database of NCBI under the BioProject accession number PRJNA699678.

2.8. Phase Analysis

The FASTQ files in SMARTer Nextera and SMARTer ThruPLEX were added to unique molecular identifier (UMI, also known as a molecular barcode) information using UMI-tools and were aligned to the reference human genome (hg38) using HISAT2 (version 2.1.0), as described elsewhere [33,34]. The obtained Sequence Alignment/Map (SAM) format files were converted to the Binary Alignment/Map (BAM) file format using SAMtools. Duplicate reads in BAM files were removed using UMI-tools according to UMI information. The BAM files provide the read name and sequence information. We extracted the read name from BAM files with about 10 base-specific sequences around the mutation site. For one mutation (chr13:51958362;G or G>GG), we extracted the read name from BAM files with ATGGGGGGCG and ATGGGGGGGCG, whereas CCCGTGGACC and CCCGTGGATC were used for the other mutation (chr13:51964900;C or C>T). We detected compound heterozygous mutations by comparing the read name, which is common or specific to each mutation.

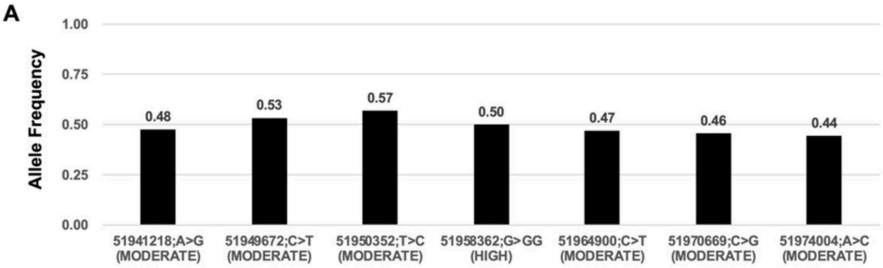
2.9. CHIPS and Sanger Sequencing

To verify the detected haplotype variants, CEL nuclease-mediated heteroduplex incision with polyacrylamide gel electrophoresis and performed silver staining (CHIPS) analysis and Sanger sequencing were performed and direct DNA sequencing, as described previously [26,35].

3. Results

3.1. Screening for Pathogenic Mutations of ATP7B

Since not only coding region mutations but also deep intronic mutations and large intragenic deletion are known as *ATP7B* pathogenic mutations, very long amplicon sequencing (vLAS) was used for gene mutation screening [36,37]. First, we performed genetic mutation screening of *ATP7B*, the responsible gene of Wilson disease by very long amplicon sequencing (vLAS). The average depth on the *ATP7B* locus was 645, and we detected 168 single nucleotide variants (SNVs) and insertion/deletions (INDELs) variants by haplotype variant calling. These variants were assigned based on functional class as one high, six moderate, four low, and 157 modifiers by SnpEff database. Of those variants assigned as high, we identified a frameshift mutation; we also identified six missense mutations from the moderate category. The allele frequency of all detected variants is approximately 0.5, indicating that all detected variants are heterozygous (Figure 2A). We performed in silico analysis using dbNSFP database which provides the normalized score based on value in the prediction algorithms (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster, PROVEAN, VEST3, MetaSVM, MetaLR, CADD, DANN, and fathmm-MKL). The scores of the missense variant chr13:51964900;C>T (NM_000053.4:c.1841G>A r.1841g>a p.Gly614Asp) were higher than 0.7 and higher than other variants, indicating that this variant is predicted to cause damage at the protein level (Figure 2B). ClinVar database provides information that the variant (NM_000053.4:c.2304dup r.2304dup p.Met769HisfsTer26) is pathogenic and the variant (NM_000053.4:c.1841G>A r.1841g>a p.Gly614Asp) is likely pathogenic. According to the ACMG guideline of interpretation of sequence variants, 2015, the variant (NM_000053.4:c.2304dup r.2304dup p.Met769HisfsTer26) fulfilled the criteria of pathogenic, and the variant (NM_000053.4:c.1841G>A r.1841g>a p.Gly614Asp) was annotated likely pathogenic [38]. In addition, our analysis using the Trans-Omics for Precision Medicine (TOPMed) which provides allele worldwide database of allele frequency and Japanese population databases such as HGVD and ToMMo indicated that five of these seven variants are polymorphisms, except for two variants chr13:51958362;G>GG (NM_000053.4:c.2304dup r.2304dup p.Met769HisfsTer26) and chr13:51964900;C>T (NM_000053.4:c.1841G>A r.1841g>a p.Gly614Asp) (Figure 2C). Then, on the basis of on these findings, we focused on these two variants, both of which were validated by CHIPS and Sanger sequencing (Figure 2D,E).



B

Position	dbSNP146	SIFT	Polyphen2_HDIV	Polyphen2_HVAR	LRT	MutationTaster	PROVEAN	VEST3	MetaSVM	MetaLR	CADD	DANN	fathmm-MKL
51941218	rs1801249	0.02571	0.06737	0.22023	0.2247	0.08979	0.0167	0.00483	0.40226	0.00011	0.04135	0.0104	0.02244
51949672	rs732774	0.00963	0.19838	0.20453	0.84324	0.81033	0.01126	0.26259	0.38266	0.00011	0.31598	0.1149	0.24202
51950352	rs1061472	0.36263	0.52422	0.55836	0.84324	0.81033	0.45551	0.58595	0.43974	0.00011	0.41857	0.8685	0.65634
51964900	.	0.78421	0.89865	0.97092	0.84324	0.81033	0.91545	0.93321	0.89984	0.88591	0.87253	0.906	0.78813
51970669	rs1801244	0.08518	0.20066	0.22257	0.09612	0.08979	0.13822	0.08601	0.283	0.00011	0.16656	0.1624	0.09122
51974004	rs1801243	0.11351	0.2477	0.27121	0.1038	0.08979	0.19093	0.0361	0.30166	0.00011	0.05255	0.4105	0.13535

C

#CHROM	POS	REF	ALT	dbSNP	ClinVar	TOPMED	HGVD	ToMMo (MALE)	ToMMo (FEMALE)
chr13	51941218	A	G	rs1801249	Benign	0.4355	0.404	0.408	0.412
chr13	51949672	C	T	rs732774	Benign	0.4442	0.401	0.408	0.412
chr13	51950352	T	C	rs1061472	Benign	0.469	0.398	0.408	0.412
chr13	51958362	G	GG	rs137853287	Pathogenic	0.0001	-	-	-
chr13	51964900	C	T	-	Likely Pathogenic	-	-	-	-
chr13	51970669	C	G	rs1801244	Benign	0.4	0.489	0.496	0.492
chr13	51974004	A	C	rs1801243	Benign	0.394	0.54	0.541	0.54

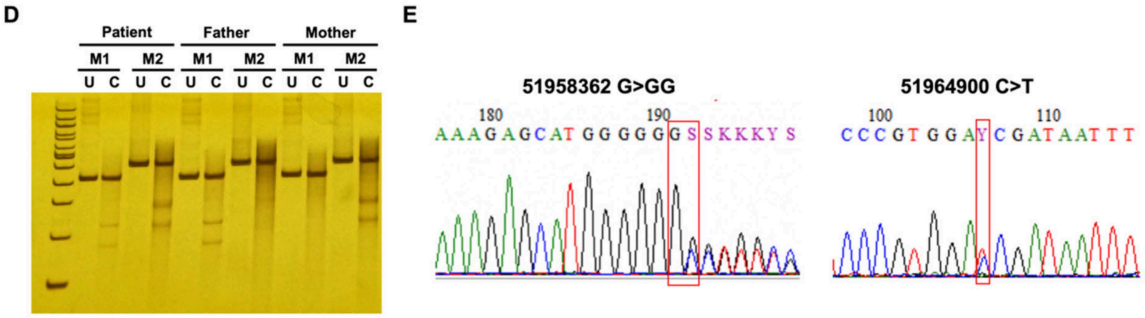


Figure 2. (**A**) Detection of haplotype mutations in the *ATP7B* locus. Allele frequency of variants; (**B**) In silico analysis of variants; (**C**) A summary of variants from the dbSNP, ClinVar, TOPMED, HGVD, and ToMMo databases; (**D**) Trio analysis by CHIPS technology assay (M1: chr13:51964900;C>T, M2: chr13:51958362;G>GG); (**E**) Patient electropherograms of the *ATP7B* mutations loci by Sanger sequencing.

3.2. Detection of Compound Heterozygous Mutations by Targeted Double-Stranded cDNA Sequencing-Based Phase Analysis

The genomic distance between chr13:51958362;G>GG and chr13:51964900;C>T is 6.5 kb (Figure 3A), whereas the distance in mRNA is only 464 bp, indicating that it is possible to determine whether compound heterozygous mutation is *in trans* or *in cis* mutation using short read targeted double-stranded cDNA sequencing-based phase analysis. First, we amplified the *ATP7B* specific locus double-stranded cDNA (product size: 621 bp) from full-length double-stranded cDNA. Next, we constructed NGS libraries from *ATP7B*-specific double-stranded cDNA using two different methods, i.e., Nextera and ThruPLEX. Nextera is a conventional approach that performs tagmentation and adapter insertion simultaneously using transposon technology. In contrast, we performed direct adapter ligation on both ends of *ATP7B*-specific double-stranded cDNA with the ThruPLEX approach. To evaluate targeted double-stranded cDNA sequencing-based phase analysis performance of these two approaches, we compared mapping rate, coverage, and detection efficiency of mutation and compound mutation. The ThruPLEX mapping rate was higher than that of Nextera (Figure 3B). Furthermore, the ThruPLEX mapping rate without UMI information was almost the same as that of Nextera, indicating that UMI information improves the

ThruPLEX mapping rate. Although the coverage of Nextera was low at both ends, we found that the coverage of ThruPLEX was uniform (Figure 3C). The read number, including that of each variant by Nextera was higher than that found by ThruPLEX, reflecting the observation that the mapping read number by Nextera was higher than that by ThruPLEX (Figure 3D,E). We also found that the allele frequencies at both positions (chr13:51958362 and chr13:51964900) were almost identical between Nextera and ThruPLEX (Figure 3F). We also detected paired-end reads including sequences of both positions in ThruPLEX but not Nextera (Figure 3G). The results from ThruPLEX show that chr13:5198362;GG and chr13:51964900;C (chr13:5198362;G and chr13:51964900;T) are within the same allele, indicating that the compound heterozygous mutation is *in trans* mutation. These results indicate that the targeted double-stranded cDNA sequencing-based phase analysis by ThruPLEX approach is better than Nextera as it detects paired-end reads that include the sequences of both positions.

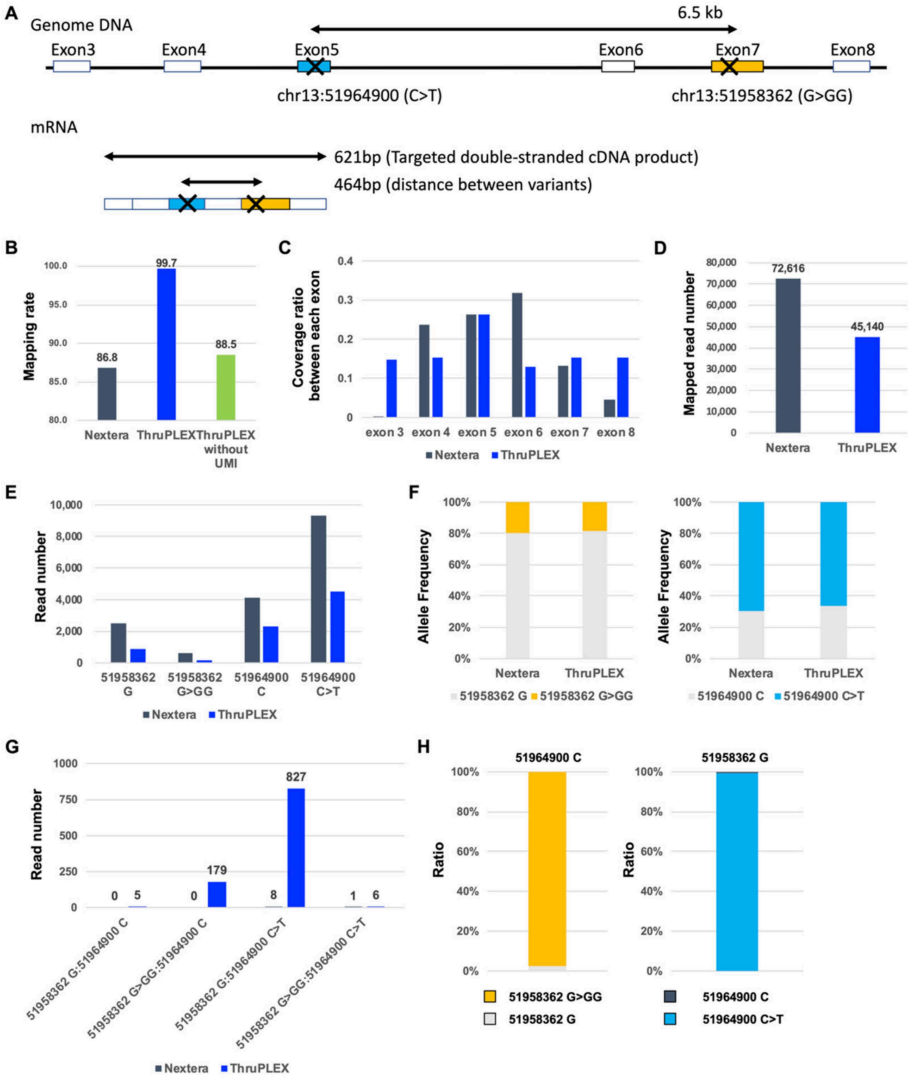


Figure 3. (**A**) Targeted double-stranded cDNA sequencing-based phase analysis. Genomic DNA and mRNA of human *ATP7B*; (**B**) The mapping rate for each sample using different approaches (Nextera versus ThruPLEX); (**C**) The coverage ratio between each exon; (**D**) The mapped read number for each sample; (**E**) The read number harboring each mutation for each sample; (**F**) Allele frequency for each position; (**G**) The read number harboring two mutations for each sample; (**H**) The ratio of reads harboring two mutations.

3.3. Validation of in trans Compound Heterozygous Mutation by Trio Analysis

To validate our identification of *in trans* compound heterozygous mutation, we investigated whether the patient's parents possess the two mutations, using CHIPS and Sanger sequencing (Figures 2D,E and 4A). We found that the father had one of the mutations (chr13:51964900;C>T) and the mother had the other mutation (chr13:51958362;G >GG). Moreover, we obtained the same results using RNA-seq analysis (Figure 4B). These results indicate that the patient inherited both mutations from his parents and has *in trans* compound heterozygous mutation (Figure 4C).

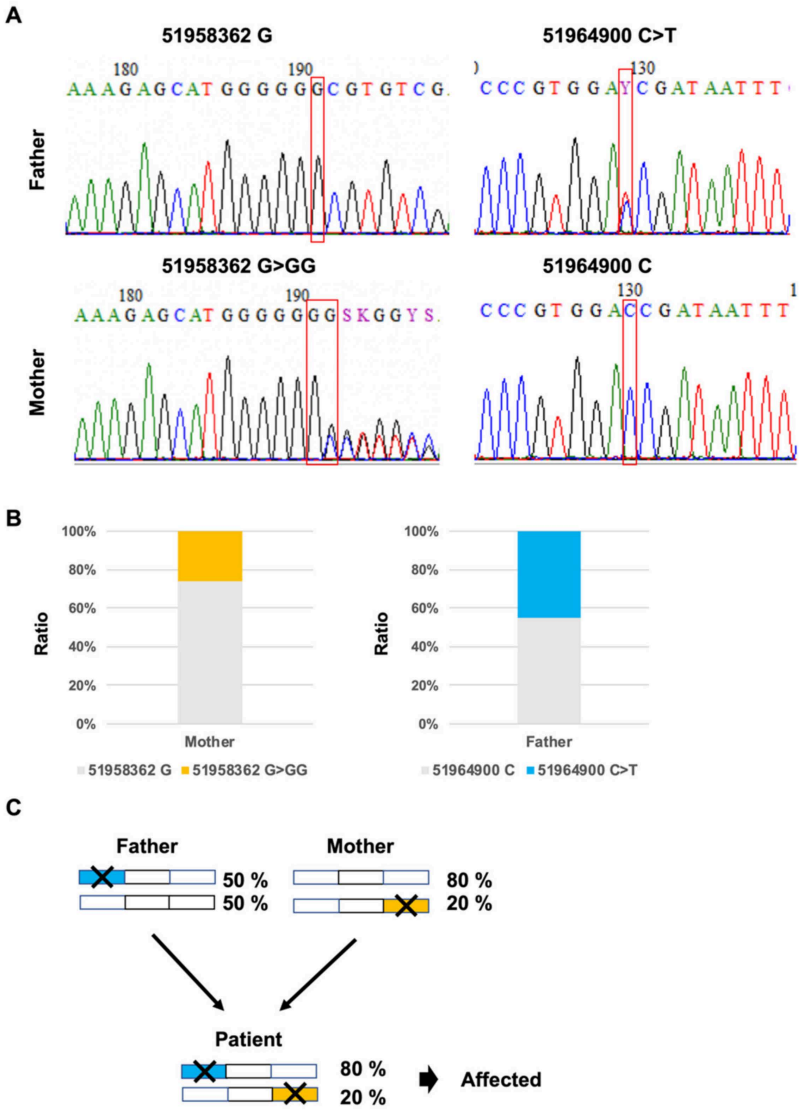


Figure 4. (**A**) Parent electropherograms of the *ATP7B* mutations loci by Sanger sequencing; (**B**) The expressed allele frequency of variants by targeted double-stranded cDNA-based sequencing; (**C**) Predicted inheritance and disease modelling based on targeted double-stranded cDNA-based sequencing.

3.4. Frameshift Mutation Causes Differential Allelic Expression

We also examined whether the two heterozygous mutations cause differential allelic expression. The results obtained from targeted double-stranded cDNA sequencing-based phase analysis indicated that expression of the mutated allele (chr13:51958362;G>GG) was lower than that of the wild type allele (Figure 3F). Furthermore, the expression of the mutated allele (chr13:51958362;G>GG) in the mother was also lower than the wild type

allele (Figure 4B). In contrast, expression of the mutated allele (chr13:51964900;C>T) showed the same expression level as the wild type allele. These results suggest that the mutation (chr13:51958362;G>GG) causes differential allelic expression. Therefore, it is possible to detect differential allelic expression, as well as compound heterozygous mutations simultaneously, by targeted double-stranded cDNA sequencing-based phase analysis.

4. Discussion

NGS-based applications for clinical diagnosis have advanced and are widely used in Mendelian-inherited diseases and cancer. The intended usage of NGS-based applications including long-read sequencing largely facilitates genome-wide comprehensive analysis such as whole-genome sequencing, whole-exome sequencing, and gene targeting panel sequencing. Several studies have reported phase analysis for genome-wide comprehensive analysis [4–6,12]. However, NGS-based phase analysis is rarely reported for a single gene despite often being required for clinical diagnosis.

Our targeted double-stranded cDNA sequencing-based phasing method successfully demonstrated detection of *in trans* compound heterozygous mutations (chr13:51958362;G >GG and chr13:51964900;C>T) in *ATP7B*. This approach has five advantages. First, it overcomes distance limitations by using mRNA instead of genomic DNA. In fact, the average length of mRNA is approximately 2.7 kb as compared with approximately 55 kb for genes. For this reason, it is possible to detect compound heterozygous mutations using mRNA more readily. The distance between our identified compound heterozygous mutations in the *ATP7B* gene tested in this study is 464 bp in mRNA and 6.5 kb in genomic DNA. The second advantage is NGS library size. Although an RNA-seq library size is usually around 300 bp because of fragmentation, it can be extended to up to 1 kb (Illumina NGS library limitation) using our direct adapter ligation method. Thus, it is possible to detect almost any compound heterozygous mutation in our method as the distance limitation of library size is 1 kb as compared with an average mRNA length of 2.7 kb, as long as there is gene expression in peripheral blood. More recently, the PacBio HiFi sequencing method for long-read sequencing was developed [39]. The method yields highly accurate long-read sequencing which provides applications such as single nucleotide and structural variant detection. In addition, it is possible to improve distance limitation in our method by combination with PacBio HiFi sequencing method. Although *ATP7B* expression in blood is low, RT-PCR was still possible as SMARTer cDNA synthesis amplifies full-length cDNA. Importantly, our method detected compound heterozygous mutations in paired-end sequencing reads, which was in contrast to the Nextera method. The third advantage is simplicity. Our approach does not involve new techniques or computationally intense statistical analysis. In fact, our approach has the following three steps: (1) full-length double-stranded cDNA synthesis, (2) targeted amplification and, (3) direct unique molecular identifier (UMI) adapter ligation in library preparation. In addition, the kits required for the third step (SMARTer, KOD One, and ThruPLEX) are commercially available. Furthermore, it was not difficult to detect a compound heterozygous mutation, because paired-end sequencing reads, which harbor approximately 10 specific sequences around each mutation, were directly extracted without the need for installation of specific software. The fourth advantage is the ability to analyze differential allele expression. It has been reported that 9% to 30% of disease-causing mutations have an impact on RNA expression [40]. Therefore, the measurement of mutant allele expression and provision of expression data would confer critically needed information not otherwise readily available for clinical diagnosis. Loss-of-function mutations with the premature stop codon have been identified to cause differential allele expression patterns by NMD [41]. Our findings, in the current study, suggest that the frameshift mutation (chr13:51958362;G>GG) may cause differential allelic expression by decreasing expression of the same allele through NMD. In addition, analyzing differential allelic expression by conventional phase analysis using sanger sequencing is difficult because sanger sequencing of many clones is required. For example, sanger sequencing of at least 1000 clones must be performed for the same level as our targeted double-stranded cDNA sequencing-based

phase analysis. Furthermore, differential allele expression analysis cannot be affected by allelic different amplification bias, recombination, and duplication during PCR by UMI technology in targeted double-stranded cDNA sequencing-based phase analysis. Fifth, our method does not require trio analysis. Trio analysis can also detect compound heterozygous mutations; however, it inherently requires sample collection of the patient and both parents, a requirement often difficult to meet in actual clinical practice. In contrast, our method only requires a total RNA sample of the patient only.

A disadvantage of our targeted double-stranded cDNA sequencing-based phase analysis is that intron variants cannot be investigated due to dealing with mRNA. It is required for expression analysis to decide the pathological significance of putative intron variants because most intron variants affect expression and splicing machinery. In practice, targeted double-stranded cDNA sequencing-based phasing needs, for vLAS variant calling, to get information of putative compound heterozygous mutations due to targeted sequencing (Figure 1), therefore, intron variants can be detected during this preliminary step (vLAS). Moreover, targeted double-stranded cDNA sequencing-based phase analysis can investigate differential allelic expression if there are appropriate heterozygous mutations. It might also be possible to analyze low quality fragmented RNA such as FFPE sample because targeted double-stranded cDNA sequencing-based sequencing phase analysis target the short region of double-stranded cDNA.

This study demonstrates that our targeted double-stranded cDNA sequencing-based phase analysis detects compound heterozygous mutations accurately without the need for trio analysis and determines differential allelic expression through NMD by frameshift mutation. We conclude that this method is useful for the determination of compound heterozygous mutations in clinical diagnosis.

Supplementary Materials: The following are available online at https://www.mdpi.com/2079-7737/10/4/256/s1, Table S1: ATP7B Long PCR primers.

Author Contributions: H.U. designed and planned the experiments; Y.N. supervised the experiments; H.U., S.T., and Y.N. performed the experiments; H.U. analyzed data; H.U. and Y.N. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Kanazawa Medical University (no. 11181, 26699).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Kanazawa Medical University (no. G111, approved 10 November 2015).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original vLAS data presented in this study are available on request from the corresponding author. The vLAS data are not publicly available due to the personal information protection law in Japanese. The raw data of targeted double-stranded cDNA sequencing-based phase analysis in this study cannot be identified individual information due to short sequence size and have been deposited in the Sequence Read Archive database of NCBI under the BioProject accession number PRJNA699678.

Acknowledgments: We thank the members of the Center for Clinical Genomics at the Kanazawa Medical University Hospital for helpful discussion and feedback on this manuscript. This work is supported by the Kanazawa Medical University (no. 11181, 26699).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hartman, P.; Beckman, K.; Silverstein, K.; Yohe, S.; Schomaker, M.; Henzler, C.; Onsongo, G.; Lam, H.C.; Munro, S.; Daniel, J.; et al. Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. *Mol. Genet. Metab. Rep.* **2019**, *19*, 100464. [CrossRef] [PubMed]
2. Voelkerding, K.V.; Dames, S.; Durtschi, J.D. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: A paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *J. Mol. Diagn.* **2010**, *12*, 539–551. [CrossRef] [PubMed]

3. Meldrum, C.; Doyle, M.A.; Tothill, R.W. Next-generation sequencing for cancer diagnostics: A practical perspective. *Clin. Biochem. Rev.* **2011**, *32*, 177–195. [PubMed]
4. Yang, W.Y.; Hormozdiari, F.; Wang, Z.; He, D.; Pasaniuc, B.; Eskin, E. Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics* **2013**, *29*, 2245–2252. [CrossRef]
5. Bansal, V.; Bafna, V. HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **2008**, *24*, i153–i159. [CrossRef]
6. Delaneau, O.; Howie, B.; Cox, A.J.; Zagury, J.F.; Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **2013**, *93*, 687–696. [CrossRef]
7. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
8. Fan, J.; Hu, J.; Xue, C.; Zhang, H.; Susztak, K.; Reilly, M.P.; Xiao, R.; Li, M. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.* **2020**, *16*, e1008786. [CrossRef]
9. Kukurba, K.R.; Zhang, R.; Li, X.; Smith, K.S.; Knowles, D.A.; How Tan, M.; Piskol, R.; Lek, M.; Snyder, M.; Macarthur, D.G.; et al. Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.* **2014**, *10*, e1004304. [CrossRef]
10. Miller, J.N.; Pearce, D.A. Nonsense-mediated decay in genetic disease: Friend or foe? *Mutat. Res. Rev. Mutat. Res.* **2014**, *762*, 52–64. [CrossRef]
11. Frischmeyer, P.A.; Dietz, H.C. Nonsense-mediated mRNA decay in health and disease. *Hum. Mol. Genet.* **1999**, *8*, 1893–1900. [CrossRef]
12. Snyder, M.W.; Adey, A.; Kitzman, J.O.; Shendure, J. Haplotype-resolved genome sequencing: Experimental methods and applications. *Nat. Rev. Genet.* **2015**, *16*, 344–358. [CrossRef]
13. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [CrossRef]
14. Porubsky, D.; Garg, S.; Sanders, A.D.; Korbel, J.O.; Guryev, V.; Lansdorp, P.M.; Marschall, T. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* **2017**, *8*, 1293. [CrossRef]
15. Laehnemann, D.; Borkhardt, A.; McHardy, A.C. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief. Bioinf.* **2016**, *17*, 154–179. [CrossRef]
16. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8. [CrossRef]
17. Patterson, M.; Marschall, T.; Pisanti, N.; van Iersel, L.; Stougie, L.; Klau, G.W.; Schönhuth, A. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **2015**, *22*, 498–509. [CrossRef]
18. Lahiri, D.K.; Schnabel, B. DNA isolation by a rapid method from human blood samples: Effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality. *Biochem. Genet.* **1993**, *31*, 321–328. [CrossRef]
19. Togi, S.; Ura, H.; Niida, Y. Optimization and Validation of Multi-modular Long-range PCR-based Next-Generation Sequencing Assays for Comprehensive Detection of Mutation in Tuberous Sclerosis Complex. *J. Mol. Diagn.* **2021**. [CrossRef]
20. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595. [CrossRef]
21. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
22. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
23. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
24. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]
25. Thorvaldsdottir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinf.* **2013**, *14*, 178–192. [CrossRef]
26. Ura, H.; Togi, S.; Niida, Y. Dual Deep Sequencing Improves the Accuracy of Low-Frequency Somatic Mutation Detection in Cancer Gene Panel Testing. *Int. J. Mol. Sci.* **2020**, *21*, 3530. [CrossRef]
27. Ura, H.; Togi, S.; Niida, Y. Target-capture full-length double-strand cDNA sequencing for alternative splicing analysis. *RNA Biol.* **2021**, 1–8. [CrossRef]
28. Cingolani, P.; Platts, A.; Wang le, L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef]
29. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]
30. Higasa, K.; Miyake, N.; Yoshimura, J.; Okamura, K.; Niihori, T.; Saitsu, H.; Doi, K.; Shimizu, M.; Nakabayashi, K.; Aoki, Y.; et al. Human genetic variation database, a reference database of genetic variations in the Japanese population. *J. Hum. Genet.* **2016**, *61*, 547–553. [CrossRef] [PubMed]

31. Tadaka, S.; Saigusa, D.; Motoike, I.N.; Inoue, J.; Aoki, Y.; Shirota, M.; Koshiba, S.; Yamamoto, M.; Kinoshita, K. jMorp: Japanese Multi Omics Reference Panel. *Nucleic Acids Res.* **2018**, *46*, D551–D557. [CrossRef] [PubMed]
32. Liu, X.; Wu, C.; Li, C.; Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **2016**, *37*, 235–241. [CrossRef] [PubMed]
33. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **2017**, *27*, 491–499. [CrossRef]
34. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef]
35. Niida, Y.; Ozaki, M.; Inoue, M.; Takase, E.; Kuroda, M.; Mitani, Y.; Okumura, A.; Yokoi, A.; Fujita, S.; Yamada, K. CHIPS for genetic testing to improve a regional clinical genetic service. *Clin. Genet.* **2015**, *88*, 155–160. [CrossRef]
36. Woimant, F.; Poujois, A.; Bloch, A.; Jordi, T.; Laplanche, J.L.; Morel, H.; Collet, C. A novel deep intronic variant in ATP7B in five unrelated families affected by Wilson disease. *Mol. Genet. Genom. Med.* **2020**, *8*, e1428.
37. Chen, Y.C.; Yu, H.; Wang, R.M.; Xie, J.J.; Ni, W.; Zhang, Y.; Dong, Y.; Wu, Z.Y. Contribution of intragenic deletions to mutation spectrum in Chinese patients with Wilson's disease and possible mechanism underlying ATP7B gross deletions. *Parkinson. Relat. Disord.* **2019**, *62*, 128–133. [CrossRef]
38. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2015**, *17*, 405–424. [CrossRef]
39. Hon, T.; Mars, K.; Young, G.; Tsai, Y.C.; Karalius, J.W.; Landolin, J.M.; Maurer, N.; Kudrna, D.; Hardigan, M.A.; Steiner, C.C.; et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **2020**, *7*, 399. [CrossRef]
40. Stenson, P.D.; Mort, M.; Ball, E.V.; Evans, K.; Hayden, M.; Heywood, S.; Hussain, M.; Phillips, A.D.; Cooper, D.N. The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **2017**, *136*, 665–677. [CrossRef]
41. MacArthur, D.G.; Balasubramanian, S.; Frankish, A.; Huang, N.; Morris, J.; Walter, K.; Jostins, L.; Habegger, L.; Pickrell, J.K.; Montgomery, S.B.; et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **2012**, *335*, 823–828. [CrossRef] [PubMed]