Molecular Biology Reports (2025)52:128
https://doi.org/10.1007/s11033-024-10196-2

ORIGINAL ARTICLE

A next-generation sequencing-based universal target panel and algorithm for one-stop detection of copy number alterations and single-nucleotide variations in the *HBB* gene cluster for rapid diagnosis of β-thalassemia

Debashis Pal[1] · Prosanto Kumar Chowdhury[1,2] · Kaustav Nayek[3] · Nidhan K. Biswas[4] · Subrata Das[4] · Anupam Basu[1]

Received: 7 November 2024 / Accepted: 19 December 2024

Abstract

Background This study aimed to develop and validate a targeted next-generation sequencing (NGS) panel along with a data analysis algorithm capable of detecting single-nucleotide variants (SNVs) and copy number variations (CNVs) within the beta-globin gene cluster. The aim was to reduce the turnaround time in conventional genotyping methods and provide a rapid and comprehensive solution for prenatal diagnosis, carrier screening, and genotyping of β-thalassemia patients.

Methods and results We devised a targeted NGS panel spanning an 80.4 kb region on chromosome 11, encompassing the beta-globin gene cluster and 5' locus control region. We also developed an advanced data analysis algorithm consisting of variant calling and depth plot analysis that facilitates simultaneous detection of SNVs and CNVs in a single run. The test panel and algorithm were validated with 14 in-house β-thalassemia carrier/patient samples and cross-checked against the HbVar database. We identified seven pathogenic SNVs and five CNVs within the beta-globin gene cluster in various combinations, such as heterozygous, homozygous, and compound heterozygous conditions. Additionally, the coordinates of 169 rare deletions and 11 fusion mutations reported in the HbVar database were checked to verify the theoretical ability of our developed gene panel to detect all the CNVs within the target region.

Conclusion The developed panel and NGS technology can detect both SNVs and CNVs in a single run and can also be utilized for prenatal diagnosis and carrier screening for hemoglobinopathies, underscoring its versatility and clinical utility.

Keywords β-thalassemia · Target NGS · SNVs · CNVs · CNVKit · Depth plot

Abbreviations

NGS	Next-generation sequencing
SNVs	Single-nucleotide variants
CNVs	Copy number variations
LCR	Locus control region

Introduction

β-Thalassemia and hemoglobinopathy are the most prevalent genetic disorders in India, South Asian countries, America, Italy, Middle Eastern countries, and other parts of the world [1, 2]. The worldwide carrier frequency of hemoglobinopathy is approximately 7%, whereas 1–5% of the global population are healthy carriers of different thalassemia mutations [3]. According to a report by the World Health Organization (WHO), approximately 40,000 infants are born annually with thalassemia, with the majority of them having β-thalassemia [4]. Consanguineous marriage, a lack of carrier testing, or the use of improper testing methods are the deep-rooted norms of the large percentage

✉ Anupam Basu
abasu@zoo.buruniv.ac.in

1 Department of Zoology, The University of Burdwan, Bardhaman, West Bengal 713104, India

2 Peerless Hospital and Research Centre, Kolkata, West Bengal, India

3 Department of Pediatric Medicine, Burdwan Medical College and Hospital, Burdwan, West Bengal, India

4 National Institute of Biomedical Genomics, Kalyani, West Bengal, India

Published online: 16 January 2025

of active thalassemia patients globally [5, 6]. High-performance liquid chromatography (HPLC) and capillary electrophoresis (CE) are the primary biochemical methods for β-thalassemia diagnosis and carrier screening. Elevated HbA2 levels associated with hypochromic, microcytic red blood cells are the key indicators of the heterozygous β-thalassemia trait [7]. Although HPLC and CE have been widely used for carrier screening, they have several limitations. In some β-thalassemia carriers, normal or borderline HbA2 can arise from mild β-thalassemia mutations, coinheritance of β-thalassemia with δ-thalassemia or iron deficiency anemia [8, 9]. In such cases, mutation screening is inevitable. Thalassemia is a genetic disorder caused primarily by mutations in the beta-globin (*HBB*) and alpha globin (*HBA*) genes. The *HBB* gene is a part of the large beta-globin gene cluster; it consists of a 5' regulatory LCR, followed by 5 beta-like globin genes, and a pseudo-beta gene (5′-epsilon– – gamma-G– – gamma-A— beta pseudogene 1– – delta– – beta–-−3′) [10]. The entire region is approximately 80 kb long. Pathogenic single-nucleotide alterations and large gross gene deletions are responsible for different types of thalassemia, which is evident from the entries in the HbVar database [11]. Ongoing research is consistently revealing many more novel rearrangements in this genomic region [12–14]. ARMS–PCR, Sanger sequencing, gap PCR, MLPA, and other genetic tests are required either alone or sequentially to understand the complete globin genotype of patients with thalassemia in a country with a diverse population lineage [15, 16]. This may increase TAT as well as expenditure and also poses a particular challenge for pregnant women, who have limited time for detection and intervention. The advancement of NGS technology, coupled with data analysis workflows and packages, enables the ease of complex genetic testing to detect SNVs and CNVs with diverse loci in a single run [17, 18]. NGS-based gene panels are extensively used in clinical settings for the identification of genetic mutations in many complex polygenic disorders, such as Alzheimer's disease (Invitae Hereditary Alzheimer's Disease Panel [https://www.ncbi.nlm.nih.gov/gtr/tests/553718/howtoorder/#]), autism spectrum disorder (Autism NGS Panel launched by Fulgent [https://www.fulgentgenetics.com/Autism]), and many more. However, globally, such a comprehensive thalassemia genomic-based solution is in the preliminary stage, and only one commercial NGS-based thalassemia kit is available (https://devyser.com/kits-and-reagents/devyser-thalassemia-ngs). Our developed NGS-based targeted gene panel and diagnostic data analysis algorithm offers a comprehensive solution, facilitating the detection of complex thalassemia cases as well as early prenatal genetic diagnosis, thereby enhancing the effectiveness of intervention strategies and contributing to improved global health outcomes.

Materials and methods

Study setting: The initial screening and filtering of 600 cases were performed using HPLC, ARMS–PCR, Sanger sequencing, and GAP-PCR. This study was approved by the ethics committee of The University of Burdwan, Purba Bardhaman, India, approval No. *IEC/BU/2017/01*. Written informed consent was obtained from all participants prior to their enrollment in the study. For minor participants, written consent was provided by their parents or legal guardians.

Subject inclusion criteria: In the cohort of 600 patients, subjects were selected based on the following criteria. (A) Mismatch pedigree through Sanger sequencing [S7–S11], (B) Failed PCR amplification for Sanger sequencing of the HBB gene [S12], (C) Ambiguous HPLC of the case/carrier state; we included β-thalassemia carrier subjects with normal or borderline HbA2 and slightly elevated HbF [S5, S6]. (D) HPFH carrier [S13, S14]. (E) β-thalassemia carrier [S1–S4]. The details of the hematological parameters and recruitment criteria are given in Table 1 and Supplementary Table S1. Additionally, one control sample was recruited for comparative analysis. Fourteen representative subjects with diverse diagnostic genotype conditions were subsequently selected for targeted next-generation sequencing (NGS) analysis (S1–S14) (Fig. 1).

Panel design: For comprehensive detection of SNVs as well as CNVs in the beta-globin gene cluster, a targeted genomic panel was designed. The beta-globin gene cluster is present in the short arm of chromosome 11, consisting of a 5′ regulatory LCR, followed by a 5 beta-like gene and a pseudo-beta gene [5′-epsilon– – gamma-G– – gamma-A— beta pseudogene 1– – delta– – beta–-−3′] and a 3′ regulatory region of the beta-globin gene (Fig. 2). For the panel design, we considered the human reference genome sequence version 38 (GRCh38). Genomic coordinates were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/) using the NCBI reference sequence NC_000011.10. The specific genomic coordinates for the different components of the beta-globin gene cluster are as follows: 5′ LCR: chr11:5,269,925–5,304,186; *HBE1*: 5268345–5269945; *HBG2*: 5253188–5254781; *HBG1*: 5248269–524985; *HBBP1*: 5241954–5243592; *HBD*: 5232838–5234483; *HBB*: 5225464–5227071; beta-globin gene 3′ regulatory region: 5,223,780–5,226,590. Based on these coordinates, we stitched the gene components together along with their intergenic regions to create a final panel (chr11: 5223780–5304186), covering a genomic region of 80.4 kb on the short arm of chromosome 11.

Library preparation: AmpliSeq for the Illumina custom library was designed using the DesignStudio™ tool (Illumina) (https://designstudio.illumina.com). A total of 336

Table 1 Hematological and clinical profiles of the recruited subjects for targeted sequencing

Sl. No.	Sample Id	Age*, sex (M/F)	CBC				HPLC			Reason for inclusion
			Hb (g/dl)	RBC (10^6/µl)	MCV (fl.)	MCH (pg)	HbA2 %	HbF %	HbE %	
1	S7	35, F	10.2	4.52	69	22.6	5.2	34.2	43	Sanger sequencing shows a mismatch pedigree; proband is homozygous for a point mutation, but one of the parents shows the absence of same variant
2	S8	22, F	7	3.68	91.6	26.4	3.5	74.7	0	
3	S9	20, F	5.9	2.62	74.8	22.5	3.0	59	0	
4	S10	31, M	10.1	4.7	–	–	10.2	21	0	
5	S11	3, F	6.6	3.26	84.4	20.2	1.1	98	0	
6	S12	6.5, F	10.7	5.18	66.8	20.7	0	96.5	0	Failed to PCR amplification of the HBB gene
7	S5	30, M	11.7	3.9	87.1	30.3	2.8	6.3	0	β-thalassemia carrier with low HbA 2%
8	S6	29, F	9.2	4.12	77.4	22.3	3.8	17.4	0	
9	S13	50, M	14.4	6.35	71.8	22.7	2.5	13.3	0	HPFH carrier; included to evaluate the efficacy of the developed panel and methodology for detecting CNVs
10	S14	30, M	14.1	5.9	72.5	23.9	2.5	25.6	0	
11	S1	10, M	8.3	4.25	77.4	19.5	4.7	7.4	0	β-thalassemia carrier; included to assess the efficacy of the developed panel in carrier screening
12	S2	52, F	7.0	3.3	65.9	20.8	5.5	10.8	0	
13	S3	31, F	9.2	4.8	59.8	18.9	4.8	10.4	0	
14	S4	7, F	8.0	4.07	72.5	19.7	8.3	1.4	0	
15	Control	27, F	12.2	4.88	66	24.9	2.4	0.3	0	To validate detection of the wild-type sequence

*Age in years

primer sets with average amplicon lengths of 352 bp were designed against the 80.4 kb targeted genomic region of chromosome 11, with a maximum coverage of 98.45% (Supplementary Figure S1).

Sample preparation and sequencing: The concentration of DNA was determined using a Qubit dsDNA HS Assay Kit on a Qubit 2.0 fluorometer. Approximately 10 ng of DNA was amplified with the AmpliSeq On-Demand Custom DNA Panel [19] using 5· AmpliSeq HiFi mix. After amplification, the primers were partially digested using FuPa reagent, and the amplified products of both primers of the same sample were pooled in a single tube. CD indices were ligated with each individual sample, after which the libraries were cleaned up with AMPure XP beads. The index-ligated library was then washed with 70% ethanol. The final amplification of the library was performed using 10· Library Amp primers. The final library was quantified using the Qubit method, and the size distribution of the library was determined using a TapeStation (Agilent Technology) according to the manufacturer's instructions, with an average fragment length between 400 and 550 bp. Approximately 10 pM of each library was pair-end sequenced on the Illumina MiSeq platform, and raw FastQ files were generated.

Data analysis algorithm for the identification of thalassemia causing both SNVs and CNVs: A step-by-step data analysis algorithm was developed to identify all the pathogenic SNVs and CNVs in the targeted region.

Step 1: Quality control of the raw data: Sequence quality analysis of the raw sequencing reads was performed using the FASTQC package. Low-quality sequencing reads were filtered out.

Step 2: Variant calling and annotation: High-quality reads were aligned against the human reference genome version hg38 using the DRAGEN DNA aligner with the Amplicon pipeline [20]. A targeted AmpliSeq bed file was used to specify the regions of interest, and a variant call file (VCF) was generated. Variants were annotated using the wANNOVAR tool [21].

Step 3: Variant filtering and identification of pathogenic SNVs: All the identified variants with minor allele frequencies $< 1\%$ were selected from 3 databases: (i) the 1000 Genomes Project Phase III, (ii) the Exome Aggregation Consortium (ExAC), and (iii) the Genome Aggregation Database (gnomAD) [22–24]. Combined Annotation Dependent Depletion (CADD) phred score > 20, was used to classify potential damaging variants, below which variants were often classified as benign [25]. To assess the impact of UTRs, upstream, downstream, intronic, and splice variants were further checked in the

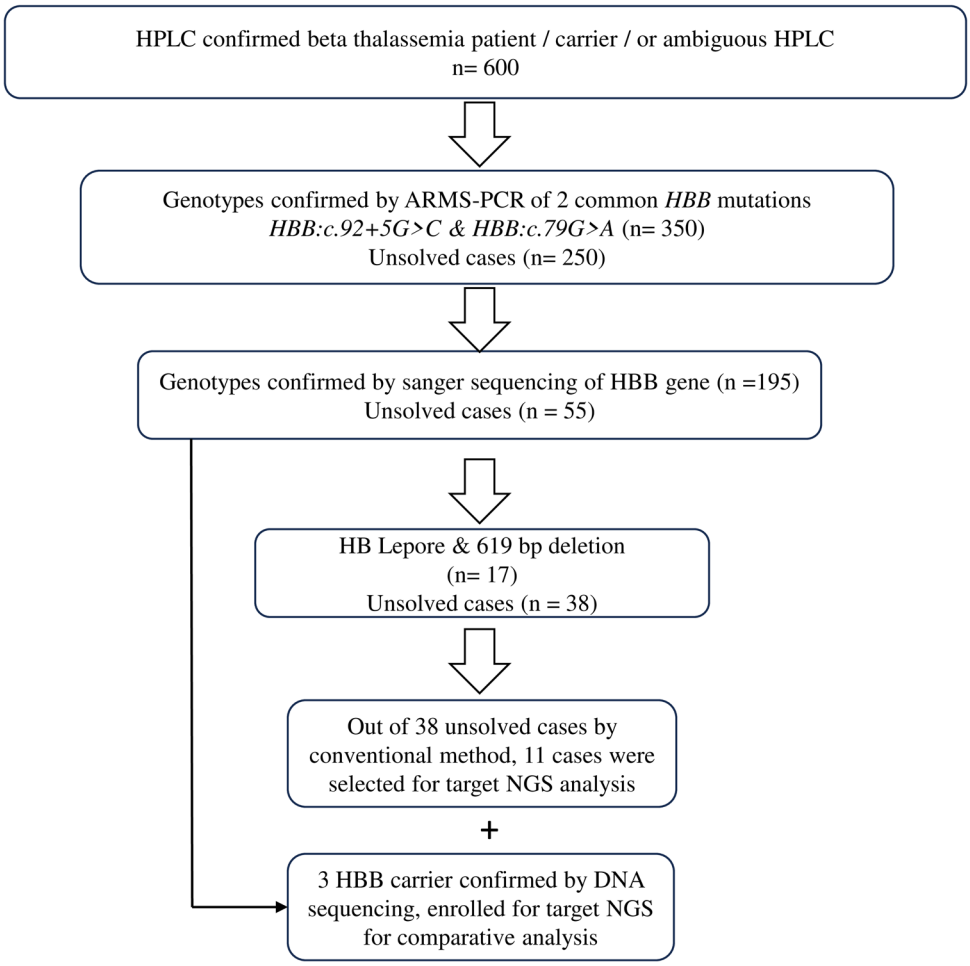

Fig. 1 Schematic representation of conventional *HBB* genotyping methods used for identifying thalassemia patients and carriers, along with the case selection process for targeted NGS analysis

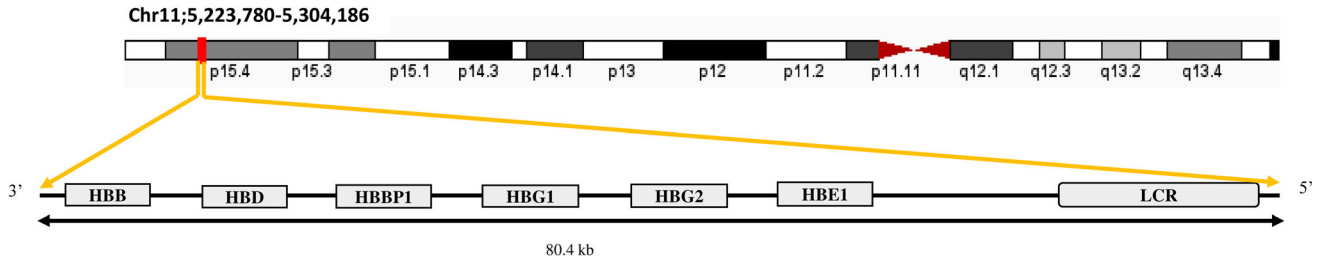

Fig. 2 Genomic region for target sequencing encompassing 80.4 kb on the short arm of chromosome 11. The diagram depicts the organization of beta-like globin genes, along with the 5′ LCR and 3′ regulatory region

ClinVar [26] and HbVAR databases to ascertain the mutation type and associated hemoglobin variants.

Step 4: Clinical correlations with identified variants: Homozygous, heterozygous, or compound heterozygous variants were clinically correlated with patient data and confirmed with parental gene sequencing. In unidentified cases or instances of mismatch mutation inheritance with parental sequences or long stretches of homozygous variants identified in the annotated VCF, we conducted further CNV analysis.

Step 5: Detection of CNVs in the target region: To investigate the presence of CNVs in the target region, we followed a multistep approach. We applied the CNVKit tool to the BAM file for CNV detection. Furthermore, a depth-based approach was adopted for validation of the CNV results. CNVkit takes the alignment and bed files as input and outputs chromosomal segments where there is considerable deviation in depth compared with

adjacent regions. In the depth-based approach, all reads aligned to the targeted regions were extracted from the alignment file (BAM) and arranged in pileup format. The pileup file was subsequently scanned from beginning to end in a per-base manner, the depth was estimated, and the depth information was stored in a depth file. The depth file is plotted and overlapped with the CNVkit results for the final list of CNV regions. Furthermore, the run of heterozygosity was estimated and validated by the genotype information of SNP data from the same region. Additionally, the Integrated Genome Viewer (IGV) tool [27] was used to visualize the coverage information from the BAM file (Figure 3).

Sanger sequencing: To validate the identified SNVs and INDELs from target sequencing, we performed Sanger sequencing of the *HBB* and *HBD* genes. The primer set used for amplification of the *HBB* gene was as follows: *HBB* forward primer 5′-GCCAAGAGATATATCTTA-GAG-3′ and reverse primer 5′-CCATTCTAAACTG-TACCCTGT-3′. The primer set used for amplification of the *HBD* gene was as follows: forward primer: 5′-CTGAGTCAAGACACACATGACAG-3′. reverse primer: 5′-TGGTATGCATAATTTGAGTTGTTG-3′.

Validation for various deletion and fusion mutations in the global population: To test the utility of the developed panel for detecting different mutations in the global population, we accessed the HbVar database. Accordingly, we mapped the submitted variants with our panel coordinates. We curated entries for deletion and fusion mutations involved in the beta-globin gene cluster. This mapping involved cross-referencing the genomic positions of submitted mutations from diverse ethnic groups in the HbVar database with the coverage provided by our developed target panel [11].

Results

Identification of SNVs in the beta-globin cluster: In accordance with the variant filtering criteria mentioned in the Methods, we successfully identified seven rare pathogenic SNVs in beta-globin gene clusters with the developed NGS method (Table 2): (1) Splice variant; *HBB: c.92 + 5G > C*, which was observed in six subjects; out of them, two subjects exhibited heterozygous conditions, whereas the remaining four subjects presented homozygous conditions (Fig. 4A). (2) *HBB: c.2T > A* was found in a single subject (Fig. 4B). (3) A homozygous substitution mutation, *HBB: c.79G > A*, was found in one patient (Fig. 4C). (4) A homozygous single-nucleotide deletion (-T) mutation in codon 15 of the *HBB* gene, *HBB: c.46delT*, (Fig. 4D) (5) A homozygous mutation in the *HBB* promoter, *HBB: c.−140 C > T*, (Fig. 4E). (6) A heterozygous delta globin gene mutation, *HBD: c.61G > A* (Fig. 4F). (7) A heterozygous upstream promoter mutation in the HBD gene, *HBD: c.−118 C > T* (rs549964658) (Fig. 4G). All seven identified SNVs were confirmed by Sanger sequencing (Supplementary Figure S2).

Identification of CNVs in the target region: CNV calling revealed different lengths of deletions in the beta-globin gene cluster. The long deletions included the following: (1) A heterozygous *HBD–HBB* deletion (Fig. 5A). (2) A compound heterozygous deletion involving *HBD–HBB* and *HBE1–HBB* genomic regions (Fig. 5B) (3) A heterozygous *HBBP1–HBB* deletion (Fig. 5C) (4) A heterozygous *HBG1–HBB* deletion (Fig. 5D) and (5) Deletion of the entire target region encompassing 5 beta-globin genes, including the pseudo-beta gene (*HBE1*, *HBG2*, *HBG1*, *HBBP1*, *HBD*, and *HBB*) and the entire LCR in the heterozygous condition, where we found all the variant sites to be in the homozygous condition (Fig. 5E). When CNV analysis of the control subjects was performed, equal depths across the target region and homozygous and heterozygous variants distributed throughout the target region were observed (Fig. 5F). All the identified deletions within the target region and the ClinVar accession numbers are listed in Supplementary Table S2.

Among the subjects recruited for testing the efficacy of the genomic panel, the distribution of genetic mutations associated with thalassemia was as follows (Table 3): (1) Compound heterozygous mutations involving both SNVs in the *HBB* gene and CNVs in the beta-globin gene cluster (S7–S11). (2) Compound heterozygous deletions of the *HBD–HBB* and *HBE1–HBB* genomic regions (S12). (3) Two cases of coinheritance of point mutations in the *HBB* and *HBD* genes (S5, S6). (4) Two subjects with different lengths of deletions, *HBD–HBB* and *HBG1–HBB*, in the target region in the heterozygous state (S13, S14). (5) Heterozygous SNV in the *HBB* gene (S1–S3). (6) Entire target region deletion in the heterozygous state (S4).

Coverage of different deletion and fusion mutations in the global population: An extensive examination of the HbVar database identified 169 deletions and 11 fusion mutations in the beta-globin gene cluster across global populations. The coordinates of these mutations are listed in Supplementary Tables S3 and S4. All the reported deletions and fusion mutations were either encompassed within the target region or extended beyond this region. Therefore, if a deletion is present in any sample, our developed panel and methodology can detect the deleted region of the genome within the beta-globin gene cluster.

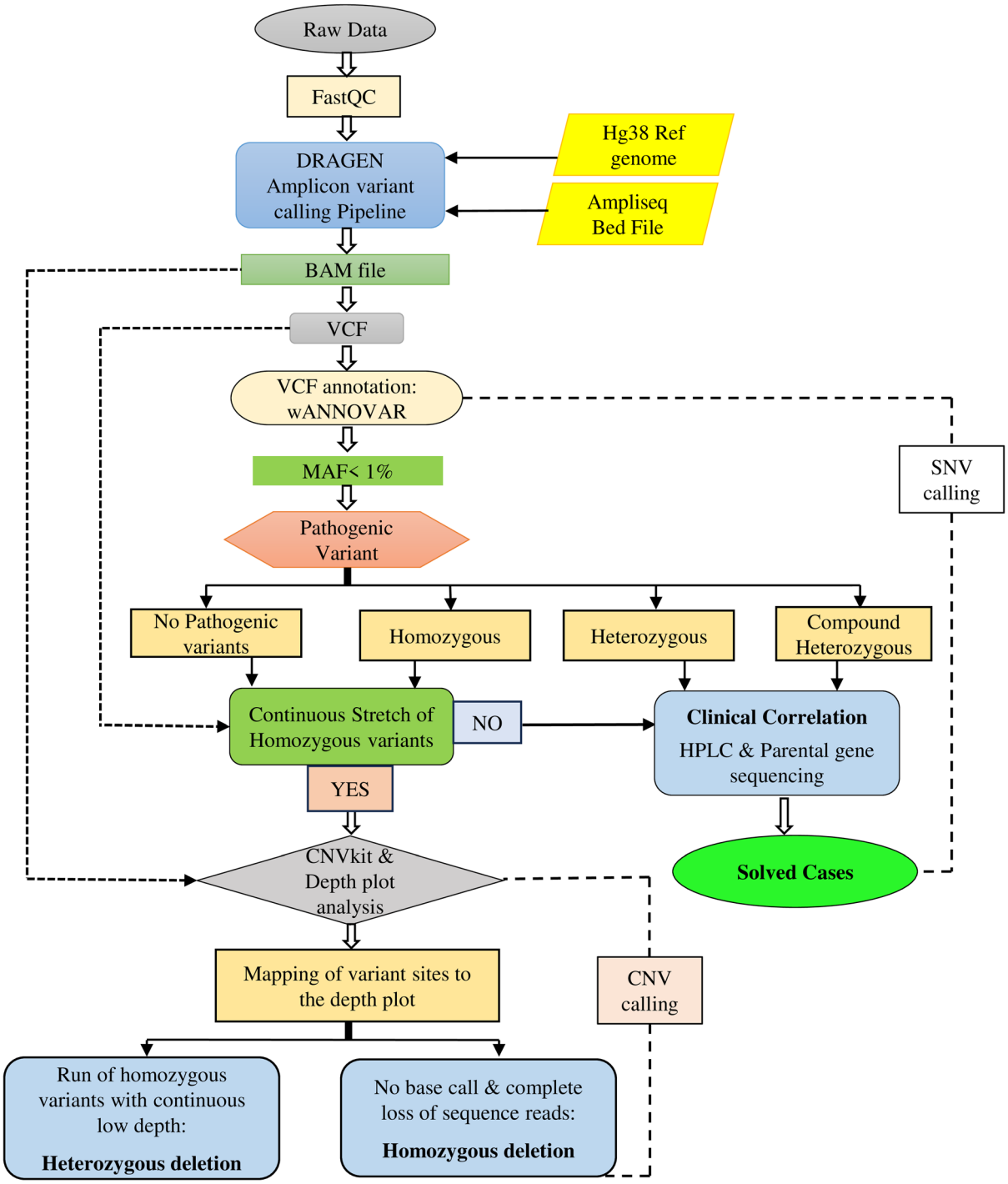

Fig. 3 Stepwise representation of raw sequence data processing, variant calling and detection of rare causative SNVs and CNVs in the targeted region

Discussion

β-Thalassemia or hemoglobinopathy can occur due to various types of mutations in the beta-globin gene cluster [11]. ARMS–PCR is a commonly used technique for identifying known mutations in a homogenous population. Reports of different investigations and our unpublished data revealed that a wide spectrum of different SNVs as well as CNVs exists, with very low frequencies responsible for β-thalassemia in India as well as in other countries [28, 29]. In such cases, the application of ARMS–PCR is constrained due to its limited success [30, 31]. Sanger

Table 2 Pathogenic variants identified by SNV calling through the filtering method of the developed algorithm of target next-generation sequencing (NGS)

Sl. No	Identified mutation	Functional position	Consequence	MAF (1000genome)	ExAC	GenomeAD	CADD phread score	Clinvar	HbVar (β + or β0)
1	*HBB: c.92 + 5G > C*	Intronic	Splicing	0.0012	0.0007	0.00006	NA	Pathogenic	β+
2	*HBB: c.2T > A*	Exonic	Start loss	NA	NA	0.00004	25.2	Pathogenic	β + or β0 unclear
3	*HBB: c.79G > A*	Exonic	Nonsynonymous	0.0028	0.0003	0.00003	23.2	Pathogenic	β+
4	*HBB: c.46delT*	Exonic	frameshift deletion	NA	NA	NA	NA	Pathogenic	β0
5	*HBB: c.−140 C > T*	upstream		NA	NA	0.00006	NA	Pathogenic	β+
6	*HBD: c.−118 C > T*	UTR5		0.0044	NA	0.00003	NA	Pathogenic	NA
7	*HBD: c.61G > A*	Exonic	Nonsynonymous	0.0016	0.0009	0.00006	23.8	NA	NA

NA not available

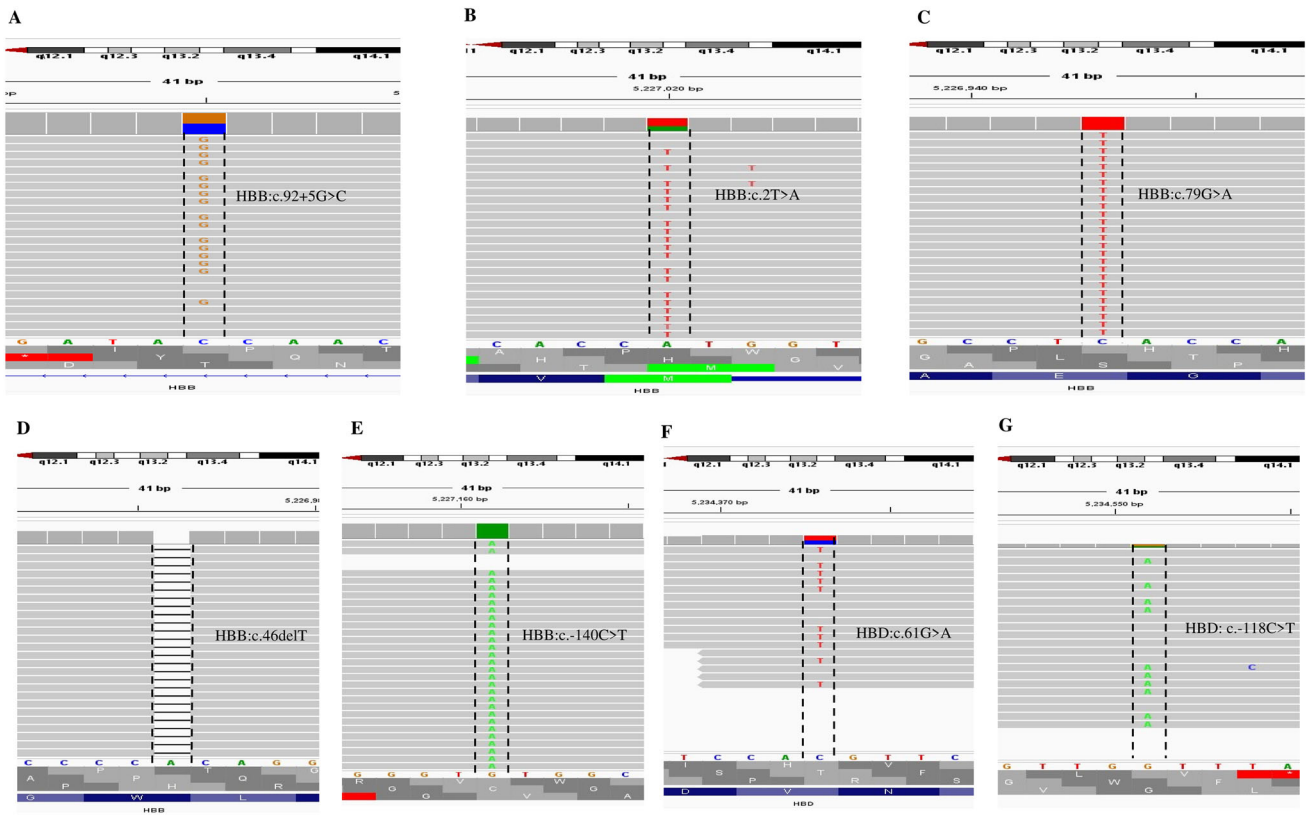

Fig. 4 IGV views of identified rare SNVs in the target region. **A**. *HBB*:c.92 + 5G > C **B**. *HBB*:c.2T > A. **C**. *HBB*:c.79G > A **D**. *HBB*:c.46delT **E**. *HBB*:c.−140 C > T **F**. HBD: c.61G > A **G**. *HBD*: c.−118 C > T

sequencing also faces limitations due to the size of the beta-globin gene family and its controlling region, which spans approximately 100 kb [32–34]. Pathogenic mutations responsible for these conditions are dispersed among different globin genes, including beta, gamma, and delta globin genes, and control the LCR [11]. Moreover, CNVs are very common in thalassemia, and Sanger techniques cannot detect CNVs. Another major drawback of the Sanger method is that the coinheritance of heterozygous point mutations with large deletions in the same genomic

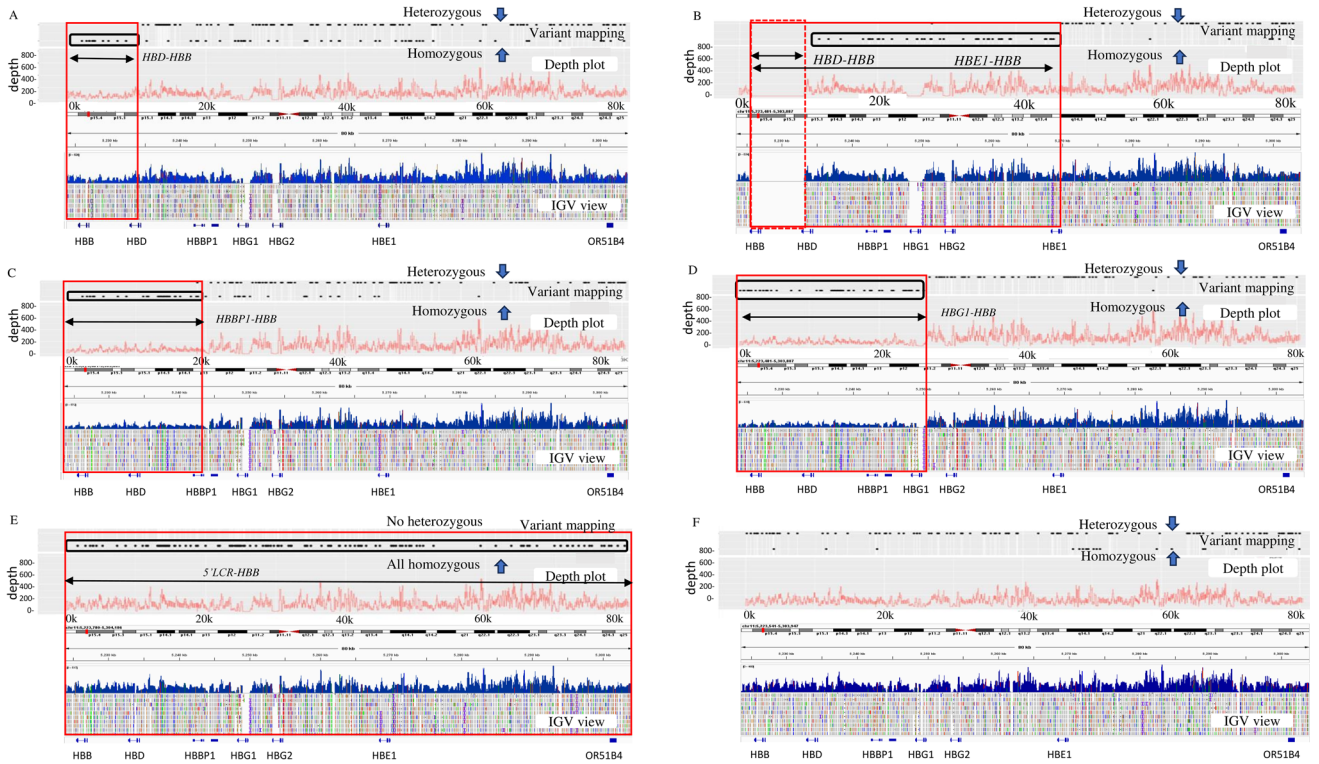

Fig. 5 Depth plot analysis, variant mapping and IGV validation of identified deletions in the beta-globin gene cluster. Each upper panel represents depth plot analysis, and each lower panel represents the IGV view of the BAM file with coverage. The region of deletions is marked in the red box, whereas stretches of homozygous variants are indicated with black outlines. **A** Representative plot of heterozygous deletions encompassing the *HBD–HBB* genomic region (S7, S13). **B** Compound heterozygous deletion of *HBD–HBB* and *HBE1–HBB*. The common region of deletion with complete loss of sequence reads is indicated in the red dotted box, (S12). **C** Heterozygous deletion from *HBBP1–HBB*, (S8). **D** Genomic deletion from *HBG1–HBB* (S9, S14). **E** Representative plot of the entire target region deletion (S4, S10, S11). **F** Representative plot of a control subject; homozygous and heterozygous variants are distributed throughout the entire target region

region appears as a homozygous mutation for that point mutation. Such scenarios can lead to false interpretations and may not reflect the actual genotype. Gap-PCR can be applied for long deletions where the exact breakpoint and the length of deletion are well known [35]. In heterogeneous populations, unknown deletions cannot be detected using this method, which can lead to false-negative results. Although MLPA is used in a few cases for long deletions, known deletions mixed with point mutations in trans patients cannot be detected by MLPA alone [36]. All the abovementioned conventional techniques alone are unable to identify different causal deletions and point mutations simultaneously for thalassemia.

To date, 291 deletion mutations have been reported in the IthaGenes database [37]. Research continues to identify many more novel deletions in this particular region. A recent study by Zhu et al. (2019) revealed a novel large deletion of 223 kb extending from the 3′UTR of the *HBD* gene to 215 kb downstream of the *HBB* gene [38]. Another study by Zhang et al. (2024) revealed a long deletion of 755 kb covering the *HBB* gene locus [39]. Accurately resolving so many large deletions requires a more extensive library panel or whole-genome sequencing to pinpoint breakpoints at base resolution. However, this approach generates a vast amount of data, which can be challenging to handle in smaller clinical settings. To address these challenges, we designed a targeted NGS panel focusing on the beta-globin gene cluster and the LCR, which are crucial for hemoglobin synthesis. This targeted NGS panel and data analysis method is optimized to detect both SNVs and CNVs effectively within the target region as well as part of a large deletion segment within the target region. Our approach is simple and comprehensive. To identify rare pathogenic SNVs, we first focused on variants with MAFs $< 1\%$ and assessed their pathogenicity based on the CADD Phred score. We consider CADD scores greater than 20 as potentially damaging variants. UTRs and upstream, downstream, intronic, and splice variants were verified for their pathogenicity by cross-referencing data from the HbVar and ClinVar databases. However, if there was evidence of a long stretch of homozygous variants from the VCF, we then executed CNV calling. CNV

Table 3 Genotypes of the recruited subjects as determined using our developed gene panel and methodology

Sl No.	Sample Id	Allele 1	Allele 2	Interpretation	Detection method
1	S7	*HBB: c.79G > A*	*HBD–HBB del*	Compound heterozygous	Allele 1: SNV calling Allele 2: CNV calling
2	S8	*HBB: c.46delT*	*HBBP1–HBB del*	Compound heterozygous	Allele 1: SNV calling Allele 2: CNV calling
3	S9	*HBB: c.92 + 5G > C*	*HBG1–HBB* del	Compound heterozygous	Allele 1: SNV calling Allele 2: CNV calling
4	S10	*HBB: c.−140 C > T*	Entire target region del	Compound heterozygous	Allele 1: SNV calling Allele 2: CNV calling
5	S11	*HBB: c.92 + 5G > C*	Entire target region del	Compound heterozygous	Allele 1: SNV calling Allele 2: CNV calling
6	S12	*HBD–HBB*	*HBE1–HBB*	Compound heterozygous	Allele 1: CNV calling Allele 2: CNV calling
7	S5	*HBB: c.92 + 5G > C*	*HBD: c.− 118 C > T*	Coinheritance of *HBB* & *HBD* heterozygous mutations	Allele 1: SNV calling Allele 2: SNV calling
8	S6	*HBB: c.92 + 5G > C*	*HBD: c.61G > A*	Coinheritance of *HBB* & *HBD* heterozygous mutations	Allele 1: SNV calling Allele 2: SNV calling
9	S13	*HBD–HBB*	Normal	Heterozygous deletion	CNV calling
10	S14	*HBG1–HBB*	Normal	Heterozygous deletion	CNV calling
11	S1	*HBB: c.92 + 5G > C*	Normal	Heterozygous	SNV calling
12	S2	*HBB: c.2T > A*	Normal	Heterozygous	SNV calling
13	S3	*HBB: c.92 + 5G > C*	Normal	Heterozygous	SNV calling
14	S4	Entire target region del	Normal	Heterozygous deletion	CNV calling

del deletion

calling was performed using CNVkit and depth plot analysis and was further confirmed by IGV analysis (Fig. 5). Thus, the developed gene panel and workflow can be used to identify SNVs and CNVs in a single assay (Fig. 3). Importantly, this method can reduce the TAT and overall cost in clinical settings.

Initially, we started with 600 β-thalassemia patient/carrier samples. Primary *HBB* genotyping was performed using conventional genotyping methods, of which 14 were selected for targeted NGS. The detailed genotype and clinical parameters are depicted in Supplementary Tables S5 and S6. In this cohort, we successfully identified 5 such cases in which both the SNV and CNV were

coinherited (S7–S11). The conventional DNA sequencing method yields false-positive results that are homozygous for a particular SNV. Specifically, Sanger sequencing relies on PCR amplification of the target region. In the presence of large deletions that encompass *HBB* and other genes in the beta globin gene cluster, the deleted allele fails to produce an amplification product. As a result, only the remaining intact allele is amplified, leading to a false homozygous appearance for any SNV present on the intact allele. This misrepresentation masks the true genotype and prevents accurate detection of compound heterozygosity involving SNVs and CNVs.

A significant challenge in thalassemia carrier diagnosis is the potential for false-negative results associated with borderline HbA2 in HPLC. Low or borderline HbA2 levels in β-thalassemia carriers are often linked to the coinheritance of pathogenic mutations in the *HBB* and *HBD* genes in cis or trans conditions [40]. To assess the efficacy of our developed targeted NGS panel and workflow in identifying such carriers with borderline HbA2, we applied our method to Subjects S5 (HbA2: 2.8%, HbF: 6.3%) and S6 (HbA2: 3.8%, HbF: 17.4%). Our approach successfully detected the coinheritance of pathogenic point mutations in both the *HBB* and *HBD* genes.

We identified one unique sample (S12) where Sanger sequencing was unable to amplify the *HBB* gene. Further NGS analysis revealed that allele 1 had a deletion in *HBD–HBB* and that allele 2 had a deletion in the *HBE–HBB* genomic region. Both alleles shared a common deletion region, resulting in complete loss of sequence reads in the *HBD–HBB* genomic region. The complete deletion was clearly evident in CNVkit and depth plot analysis and was further confirmed through IGV (Fig. 5B). Conversely, S13 and S14 exhibited heterozygous deletions from the *HBD–HBB* and *HBG1–HBB* genomic regions, respectively, leading to increases in HbF levels (Fig. 5A–D).

Carrier screening for thalassemia was also validated through the developed targeted panel-based NGS methodology. In our sample cohort, we successfully identified 3 heterozygous diseases causing SNVs in S1–S3 by simple SNV calling, whereas a heterozygous deletion encompassing the entire target region was identified in the S4 sample by CNV calling (Fig. 5E). We also identified 5 different types of deletions under either heterozygous or compound heterozygous conditions (Fig. 5, Supplementary Figure S3).

We prepared the target gene panel comprises an 80.4 kb genomic region of the *HBB* gene cluster containing all the genes of the beta-globin family and their controlling regions. The aim of this gene panel is to detect any mutations in the global population. Accordingly, we evaluated the effectiveness of our target genomic panel through the coordinates of the different CNVs responsible for thalassemia, as reported in the HbVar. Our findings demonstrated that our targeted genomic panel was able to detect all the deletion fusion mutations reported in HbVar involving beta-globin gene clusters, which demonstrates its applicability in the detection of thalassemia-causing mutations in the world population.

In addition to β-thalassemia, alpha-thalassemia cases are rare. The loss of a single alpha allele is quite common, whereas the simultaneous loss of three alpha alleles, leading to HbH disease, is very rare in the Indian population. Individuals who are alpha carriers or have a loss of two alleles remain clinically unaffected owing to the presence of four alpha alleles. The frequency of alpha thalassemia carriers varies from 1 to 18% in different regions of India, whereas less than 2% of the population are carriers of alpha gene triplication [41, 42]. Alpha gene mutations are known to influence the severity of β-thalassemia. However, thalassemia caused by alpha gene mutations alone or in combination with beta mutations is rare. Therefore, detecting mutations in the *HBB* gene cluster is particularly important in the Indian context. In this study, we focused exclusively on mutations in the beta-globin gene cluster and did not consider mutations in the alpha-globin gene.

Our developed targeted NGS gene panel and combined algorithm for the detection of β-thalassemia genotypes are very unique. This single-assay-based, cost-effective method results in less TAT. The developed panel and NGS algorithm can detect both SNVs and CNVs responsible for β-thalassemia in a single run, whereas conventional genetic testing methods for thalassemia are not able to detect point mutations, indels, or CNVs in a single run. Thus, the developed method has considerable clinical relevance for thalassemia detection, not only for suspected carrier subjects but also for rapid prenatal diagnosis. The advent of third-generation sequencing (TGS), including long-read single-molecule real-time (SMRT) sequencing, offers significant potential for detecting large deletions that are challenging to identify with conventional short-read-based gene panels utilizing NGS. However, current TGS methods are limited by high error rates, are unsuitable for SNPs, have large data outputs and formats, and have high costs [43]. In thalassemia, the mutational landscape includes both SNVs and CNVs, further constraining the clinical applicability of TGS at this time. Our NGS-based, comprehensive one-stop solution effectively detects a wide range of mutations with robust coverage, making it a practical and reliable choice. As advancements in TGS chemistry and bioinformatics pipelines increase accuracy and reduce costs, the transition to a TGS-based platform could become a feasible and impactful option in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11033-024-10196-2.

Acknowledgements We thank National Institute of Biomedical Genomics (NIBMG) for sequencing service. The authors also extend their appreciation to the authorities of Burdwan Medical College and Hospital, and The University of Burdwan. Special thanks to the patients who generously agreed to participate in the present study.

Author contributions D. P. did the experiments, data analysis and wrote the manuscript; P. K. C. and K. N. did the clinical workup; S. D. under supervision of N. K. B. did a part of the data analysis, A. B. conceptualized, planned and supervised the entire work, data analysis and checked the manuscript.

Funding This work was supported by the Department of Biotechnology (DBT), Government of India, for partial funding [BT/PR26461/MED/12/821/2018].

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

Declarations

Competing interests The authors declare no competing interests.

Ethics approval This study was approved by the ethics committee of The University of Burdwan, Purba Bardhaman, India, approval No. *IEC/BU/2017/01*.

References

1. Yadav SS, Panchal P, Menon KC (2022) Prevalence and management of β-Thalassemia in India. Hemoglobin 46(1):27–32. https://doi.org/10.1080/03630269.2021.2001346
2. Kattamis A, Forni GL, Aydinok Y et al (2020) Changing patterns in the epidemiology of β-thalassemia. Eur J Haematol 105(6):692–703. https://doi.org/10.1111/ejh.13512
3. Pinto VM, Russo R, Quintino S et al (2023) Coinheritance of PIEZO1 variants and multi-locus red blood cell defects account for the symptomatic phenotype in β-thalassemia carriers. Am J Hematol 98(6):E130–E133
4. Forni GL, Grazzini G, Boudreaux J et al (2023) Global burden and unmet needs in the treatment of transfusion-dependent β-thalassemia. Front Hematol 2:1187681
5. Mensah C, Sheth S (2021) Optimal strategies for carrier screening and prenatal diagnosis of α- and β-thalassemia. Hematol Am Soc Hematol Educ Program 2021(1):607–613. https://doi.org/10.1182/hematology.2021000296
6. Jameel T, Baig M, Murad MA et al (2024) Consanguineous marriages, premarital screening, and genetic testing: a survey among Saudi university students. Front Public Health. https://doi.org/10.3389/fpubh.2024.1328300
7. Hosseeini S, Kalantari E, Dorgalaleh A et al Thalassemia and Hemoglobinopathy Screening By HPLC Method and Comparison With Conventional Methods
8. Colaco S, Colah R, Nadkarni A (2022) Significance of borderline HbA2 levels in β thalassemia carrier screening. Sci Rep 12(1):5414
9. Sharifi A, Mahdieh N (2021) HBB mutations and HbA2 level: escaping the carrier screening programs. Clin Case Rep 9(2):973–977
10. Gallagher PG (2023) A novel β-Globin locus deletional syndrome: εγ-Thalassemia. Clin Chem 69(7):671–672. https://doi.org/10.1093/clinchem/hvad067
11. Giardine BM, Joly P, Pissard S et al (2021) Clinically relevant updates of the HbVar database of human hemoglobin variants and thalassemia mutations. Nucleic Acids Res 49(D1):D1192–D1196. https://doi.org/10.1093/nar/gkaa959
12. Bao XQ, Wang JC, Qin DQ et al (2022) A novel 5 kb deletion in the β-globin gene cluster identified in a Chinese patient. Hemoglobin 46(4):245–248
13. Yin ZZ, Yao J, Wei FX et al (2022) Targeted next-generation sequencing reveals a large novel β-Thalassemia deletion that removes the entire HBB gene. Hemoglobin 46(5):290–295
14. Li Y, Liang L, Guo W et al (2023) Identification of a novel 107 kb deletion in the alpha-globin gene cluster using third-generation sequencing. Clin Biochem 113:36–39
15. Hu S, Zhan W, Wang J et al (2020) Establishment and application of a novel method based on single nucleotide polymorphism analysis for detecting β-globin gene cluster deletions. Sci Rep 10(1):18298. https://doi.org/10.1038/s41598-020-75507-6
16. Munkongdee T, Chen P, Winichagoon P et al (2020) Update in laboratory diagnosis of Thalassemia. Front Mol Biosci 7:74
17. Pereira R, Oliveira J, Sousa M (2020) Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med 9(1):132
18. Singh AK, Olsen MF, Lavik LA et al (2021) Detecting copy number variation in next generation sequencing data from diagnostic gene panels. BMC Med Genom 14(1):214
19. AmpliSeq for Illumina On-Demand Custom and Community Panels. Reference Guide (Document # 1000000036408 v09). Illumina Proprietary (2020) Available online at: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/ampliseq-for-illumina/ampliseq-for-illumina-custom-and-community-panels-reference-guide-1000000036408-09.pdf (accessed January 30, 2021)
20. Illumina (2020) User Guide Illumina-DRAGEN-Bio-IT Platform 3.7 User-Guide-1000000141465-00.pdf
21. Rathinakannan VS, Schukov HP, Heron S et al (2020) ShAn: an easy-to-use tool for interactive and integrated variant annotation. PLoS ONE 15(7):e0235669. https://doi.org/10.1371/journal.pone.0235669
22. Fairley S, Lowy-Gallego E, Perry E et al (2020) The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res 48(D1):D941–D947
23. Ge F, Arif M, Yan Z, Worachartcheewan A, Shoombuatong W et al (2024) Review of computational methods and database sources for predicting the effects of coding frameshift small insertion and deletion variations. ACS Omega 9(2):2032–2047. https://doi.org/10.1021/acsomega.3c07662
24. Karczewski KJ, Francioli LC, Tiao G et al (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581(7809):434–443
25. Rentzsch P, Witten D, Cooper GM et al (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47(D1):D886–D894. https://doi.org/10.1093/nar/gky1016
26. Landrum MJ, Chitipiralla S, Brown GR et al (2020) ClinVar: improvements to accessing data. Nucleic Acids Res 48(D1):D835–D844. https://doi.org/10.1093/nar/gkz972
27. Robinson JT, Thorvaldsdottir H, Turner D et al (2023) igv.js: an embeddable JavaScript implementation of the Integrative Genomics viewer (IGV). Bioinformatics 39(1):btac830. https://doi.org/10.1093/bioinformatics/btac830

28. Huang TL, Zhang TY, Song CY et al (2020) Gene mutation spectrum of thalassemia among children in Yunnan Province. Front Pead 8:159
29. Singh P, Shaikh S, Parmar S et al (2023) Current status of β-Thalassemic burden in India. Hemoglobin 47(5):181–190. https://doi.org/10.1080/03630269.2023.2269837
30. Hassan S, Bahar R, Johan MF et al (2023) Next-generation sequencing (NGS) and third-generation sequencing (TGS) for the diagnosis of Thalassemia. Diagnostics 13(3):373. https://doi.org/10.3390/diagnostics13030373
31. Munkongdee T, Chen P, Winichagoon P et al (2024) Update in laboratory diagnosis of thalassemia. Front Molecularbiosci 27(7):74
32. Achour A, Koopmann TT, Baas F et al (2021) The evolving role of next-generation sequencing in screening and diagnosis of Hemoglobinopathies. Front Physiol 12:686689. https://doi.org/10.3389/fphys.2021.686689
33. Crossley BM, Bai J, Glaser A et al (2020) Guidelines for Sanger sequencing and molecular assay monitoring. J Vet Diagn Invest 32(6):767–775
34. Sabath DE (2023) The role of molecular diagnostic testing for hemoglobinopathies and thalassemias. Int J Lab Hematol 45:71–78
35. Yasin NM, Abdul Hamid FS, Hassan S et al (2022) Molecular and hematological studies in a cohort of beta zero South East Asia deletion (β°-thal SEA) from Malaysian perspective. Front Pediatr 10:974496. https://doi.org/10.3389/fped.2022.974496
36. Luo S, Chen X, Yuan D, Liu Y (2022) Detection of four rare thalassemia variants using single-molecule realtime sequencing. Front Genet 13:974999
37. Minaidou A, Tamana S, Stephanou C et al (2022) A novel tool for the analysis and detection of copy number variants associated with haemoglobinopathies. Int J Mol Sci 23(24):15920. https://doi.org/10.3390/ijms232415920
38. Zhu F, Wei X, Cai D et al (2019) A novel 223 kb deletion in the beta-globin gene cluster was identified in a Chinese thalassemia major patient. Int J Lab Hematol 41(4):456–460. https://doi.org/10.1111/ijlh.13021
39. Zhang R, Li R, Fang J et al (2024) Thalassemia caused by complex large fragment rearrangements. QJM 117(9):672–674. https://doi.org/10.1093/qjmed/hcae089
40. Colaco S, Nadkarni A (2021) Borderline HbA_2 levels: Dilemma in diagnosis of beta-thalassemia carriers. Mutat Res Rev Mutat Res 788:108387. https://doi.org/10.1016/j.mrrev.2021.108387
41. Thaker P, Mahajan N, Mukherjee MB, Colah RB (2022) Molecular heterogeneity of hb H disease in India. Thalassemia Rep 12(3):73–84. https://doi.org/10.3390/thalassrep12030012
42. Shaw J, Patra A, Khatun A et al (2024) Alpha globin gene alterations modifying the phenotype of homozygous beta thalassaemia. EJHaem 5(3):440–446. https://doi.org/10.1002/jha2.923
43. Scarano C, Veneruso I, De Simone RR, Di Bonito G, Secondino A, D'Argenio V (2024) The third-generation sequencing challenge: novel insights for the omic sciences. Biomolecules 14(5):568. https://doi.org/10.3390/biom14050568