MGI's SE400 Sequencing Product Facilitates the Identification of Thalassemia Gene Cluster Deletion

In 2021, Guangzhou AmCare Genomics Laboratory collaborated with multiple hospitals in Guangzhou and published a research article titled "Identification of thalassemia gene cluster deletion by long-read whole-genome sequencing (LR-WGS)" in the *International Journal of Laboratory Hematology*. This study utilized the DNBSEQ-G400 SE400 long-read sequencing mode from MGI to conduct whole-genome sequencing (WGS) analysis of seven patients with thalassemia. The results showed that the long-read SE400 WGS based on the DNBSEQ-G400 genetic sequencer is a powerful tool for detecting pathogenic variants and breakpoints in thalassemia. It can not only detect known deletions but can also identify new coding/non-coding deletions associated with the disease[1].

Recommended application: Genetic Disease Diagnosis

Recommended model: DNBSEQ-G400RS

• SE400 boosts a higher diagnostic rate

Long-read SE400 whole-genome sequencing has significant advantages in detecting highly repetitive sequences, structural variants, and *Indel* variants, significantly improving the detection rate of pathogenic variants in thalassemia.

• Efficient and high-quality sequencing data output

DNBSEQ sequencing technology exhibits many excellent features such as high accuracy, low repeat rate and low index hopping rate.

• Highly automated library preparation process

This library preparation process can be efficiently and rapidly carried out with the assistance of MGI's self-developed automation systems.

Background

At present, high-throughput targeted sequencing or whole exome sequencing (WES) has become common practices in the clinical diagnosis of genetic diseases, achieving a positive detection rate ranging from 20% to 50%. Nonetheless, WES may not provide reliable diagnoses for specific gene variants, primarily due to technical constraints, especially in cases involving highly homologous sequences like pseudogenes. The pseudogenes that were obtained by single-gene cloning were first defined as genetic sequences within the genome that exhibit similarity to the coding gene sequence but do not undergo typical expression in 1997. Now, the human genome contains an approximately 20,000 pseudogenes, and it has been established that 98 pathogenic genes are situated within pseudogenes or in sequences with high homology, such as the globin gene cluster.

High-throughput sequencing typically has read lengths of 100-150 bp and cannot differentiate genetic variants originating from genes and pseudogenes. On the other hand, long-read high-throughput sequencing can directly detect repeat sequences, allowing for the distinguishing between highly homologous genes and pseudogenes. Compared to WES, whole-genome sequencing (WGS) in long-read sequencing modes has shown better performance in detecting structural variants (SVs), Indel variants and sequencing depth, making it conductive to identifying pathogenic variants related to rare diseases[2-4].

Thalassemia, as an autosomal recessive genetic blood disorder, has complicated types as well as high morbidity and mortality[5,6]. Thalassemia still affects over 20,000 newborns each year, particularly in South China[7]. The continuous improvement of PCR technology has made molecular technique the gold standard for thalassemia diagnosis. However, relevant assay methods have more or less various limitations, such as the ability to detect only specific types of variants and the inability to spontaneously detect deletion or non-deletion variants. Therefore, there is an urgent need in this field for new methods to detect gene breakpoints and fill the gaps in these methods.

FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES.

Research Description

In this study, 7 patients with thalassemia gene cluster deletions were selected for whole-genome sequencing (WGS) using the DNBSEQ-G400 SE400 long-read sequencing mode. This approach helped determine the breakpoints of the gene deletions. The results showed that all the patients carried heterozygous deletions within the globin gene clusters. Specifically, 3 patients had variants within the α-globin gene cluster, three had variants within the β-globin gene cluster, and 1 patient had deletions in both globin gene clusters. This finding is expected to facilitate the diagnosis of genetic diseases involving pseudogenes or highly repetitive sequences and provide options for prenatal diagnosis.

Materials and Methods

Sample collection

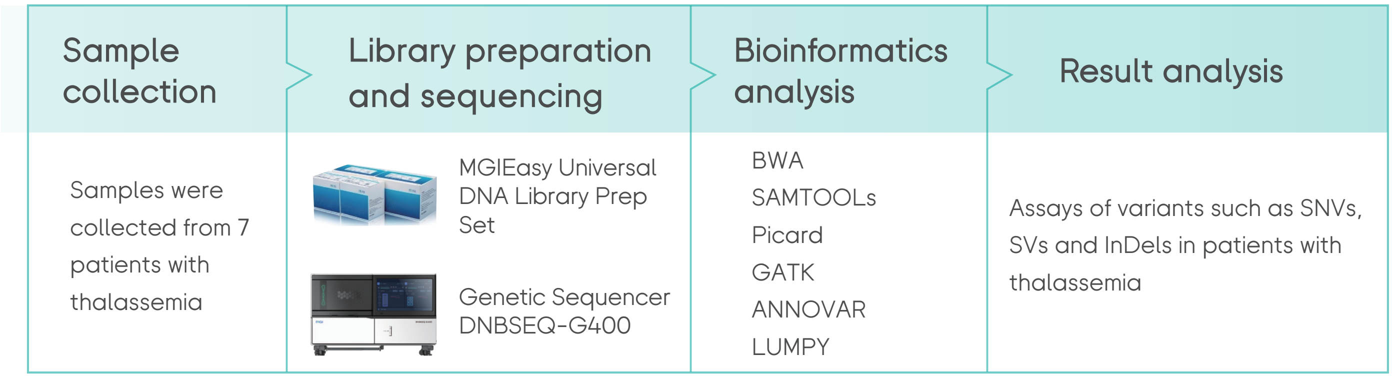
This study collected samples from seven patients with thalassemia whose hematological characteristics suggested the presence of α or β gene deletions. Informed consent was obtained from those patients. Sanger sequencing was performed to exclude the possibility of point mutations in the *HBA1*, *HBA2* , or *HBB* genes. Multiplex ligation-dependent probe amplification (MLPA) analysis indicated that all seven subjects had heterozygous large segment deletions within the α-globin gene cluster or β-globin gene cluster, but the deletion breakpoints had not been determined.

DNA extraction, library preparation, and WGS

A SolPure Blood DNA Extraction Kit was used to extract genomic DNA from peripheral blood samples of patients, and relevant products from MGI were used for subsequent library preparation and sequencing. The steps for library preparation are as follows: genomic DNA was first fragmented, followed by end repair and adapter ligation. PCR amplification was then conducted, the purified PCR products (fragment lengths of 400-600 bp) were subjected to single-strand circularization and preparation of DNA nanoballs. Finally, single-end 400-bp long-read sequencing (SE400) was carried out on the DNBSEQ-G400 genetic sequencer. It is noteworthy that when dealing with a large number of samples, MGI's automated extraction and library preparation systems can significantly save labor and improve efficiency.

Bioinformatics analysis

The raw sequencing data from the DNBSEQ-G400 platform were stored in FASTQ file (Fq). The sequencing data was aligned to the human reference genome GRCh37/hg19 using the BWA software (version: 0.7.5) with the MEM algorithm. SAMTools (version: 0.1.18) and Picard tool (version: 1.93, https://broadinstitute.github.io/picard/) were used for sorting, indexing, and de-duplicating BAM files. The GATK software (version: 3.1-1) and the HaplotypeCaller tool were employed to detect SNVs and INDELs, and ANNOVAR was used to annotate SNVs and INDELs (reference information was derived from databases such as SNP, 1000 Genomes, and other published databases). Additionally, LUMPY was used to calculate structural variants (SVs). The potential and reliable breakpoints were identified by counting the total number of softClip reads that could be aligned to the human reference genome. The sequence cut by softClip was retrieved to find the other end of the breakpoint.



Sample collection

Library preparation and sequencing

Bioinformatics analysis

Result analysis

Samples were collected from 7 patients with thalassemia

MGIEasy Universal DNA Library Prep Set

Genetic Sequencer DNBSEQ-G400

BWA
SAMTOOLs
Picard
GATK
ANNOVAR
LUMPY

Assays of variants such as SNVs, SVs and InDels in patients with thalassemia

Results

SE400 WGS analyses of patients with thalassemia

The study selected 7 patients coded by Guangzhou Women and Children Medical Center and sent their peripheral blood samples to the independent laboratory (the AMCARE Genomic Laboratory in Guangzhou, China) for WGS and data analysis. The DNBSEQ-G400 genetic sequencer generated a total of 118.89 GB of WGS data. All 7 patients were found to carry heterozygous deletions. Among these patients, 3 had deletions in the α-globin gene cluster, 3 had deletions in the β-globin gene cluster, and 1 had deletions in both globin gene clusters. Specifically, two patients had the same DNA breakpoint location that led to a 19 kb gene deletion, and two other patients shared identical breakpoint location with the gene deletion of 27 kb in length (Table 1 and Figure 1).

Patient	Variant type	Hr	Deletion region	Deletion length	Copy number[a]
31941-ZZC	*HBA1, HBA2* deletions	16	157 009-330 001	172 993	1
31942-HZX	*HBA1, HBA2* deletions	16	215 396-234 699	19 304	1
31943-PHY	*HBB* deletions	11	5 222 878-5 250 289	27 412	1
31944-GXM	*HBB* deletions	11	5 236 361-5 257 771	21 411	1
31945-ZJL	*HBB* deletions	11	5 191 121-5 270 050	78 930	1
31946-LY	*HBA1, HBA2, HBB* deletions	11	5 222 878-5 250 288	27 411	1
		16	215 391-234 699	19 309	1
48197-YCT	*HBA1 HBA2* deletions	16	220 862-231 981	11 120	1

Table 1 WGS Analysis for identification of HBA1/2 and HBB variants

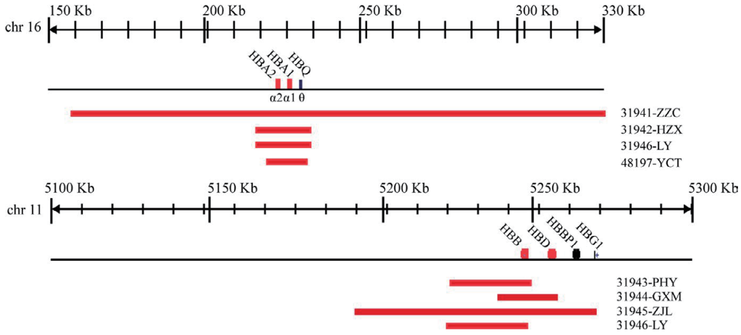


Figure 1. A summary of globin gene deletions in seven patients. The black line represents the normal region sequence, while the red filled box represents the deleted sequence in chr 11 and chr 16.

DNA structural variants of the 7 patients on chromosomes 11 and 16 (GRCh37/hg19)

1)Patient 31941-ZZC, a 32-year-old male, exhibited typical α-thalassemia characteristics. WGS analysis detected a rare 172 kb deletions of the α-globin gene cluster at chr16:57009-330001, which encompassed all coding exons of HBA1 and HBA2 (Figure 1, Table 1, and Table 2).

2)Patients 31942-HZX and 31946-LY had the same DNA breakpoints at chr16:215396-234699, resulting in a 50% reduction of HBA1 and HBA2 compared to the normal control group (Figure 1 and Table 2). By designing a pair of primers that covered the deleted region, we amplified a 1344 bp fragment (Figure 3D) and identified the gene deletion range of chr16:215256-235908. Additionally, we found that patient 31946-LY had a deletion of the β-globin gene at chr:115222878-5250288, which included all coding exons of HBB (Figure 1, Table 1, and Table 2), and was verified by PCR (Figure 2D). Both patients 31943-PHY and 31946-LY had high hemoglobin F (Hb F). MLPA analysis revealed deletions in the β-globin cluster.

3)Patient 31943-PHY, a 28-year-old female, exhibited persistent microcytic and hypochromic anemia with elevated Hb F (19.3%), which is a typical feature of hereditary persistence of fetal hemoglobin (HPFH) (Table 1 and Table 2). WGS detected a heterozygous deletion of the β-globin gene (Figure 1), which was later confirmed by PCR (Figure 2D).

4)Patient 31944-GXM had a microcytic anemia with Hb F-type (21%) HPFH (Table 1 and Table 2). WGS detected a 21.4 kb heterozygous deletion of the β-globin gene at chr11:5236361-5257771 (Table 2 and Figure 1), and this deletion was confirmed by PCR (Figure 2D).

5)Patient 31945-ZJL had a microcytic anemia and showed characteristic of hyperlipidemia (Table 1 and Table 2). WGS detected a 78.9 kb heterozygous deletion of the β-globin gene at chr11: 5 191 121- 5 270 050 (Table 2 and Figure 1), and this deletion was confirmed by PCR (Figure 2D).

6)In the case of patient 48197-YCT, a deletion in the α-globin gene cluster was detected by MLPA (This deletion was identified using 18 probes, which spanned the region between 16p13.3:160, 313-196, 305, which encompassed sequences upstream of the HBA2-, HBA2-, HBA1-, and HBQ-related regions.). Nevertheless, the precise location of the breakpoint remained unclear. An 11 kb deletion in the α-globin gene cluster between 220,861 and 231,981 in chr 16 was detected by WGS (Table 2, Figure 2A, B). Using this WGS data, PCR amplification was performed, and the breakpoint fragments were verified by Sanger sequencing (Figure 2C).

Sample ID	Gender	Age	Hb (g/L)	MCV (fL)	MCH (pg)	HbA (%)	HbA2 (%)	HbF (%)
31941-ZZC	Male	32	142	66.6	21.3	97.5	2.5	0
31942-HZX	Male	30	139	70.2	23.6	97.6	2.4	0
31943-PHY	Female	28	120	71.7	23.3	76	4.7	19.3
31944-GXM	Female	31	123	76.5	24.9	77.2	1.8	21
31945-ZJL	Male	36	136	68.4	21	84.3	2.4	13.3
31946-LY	Female	25	123	70	22.2	80.5	4	15.5
48197-YCT	Female	28	103	67	22	97.4	2.6	0

Table 2: Hematological data of 7 patients with thalassemia

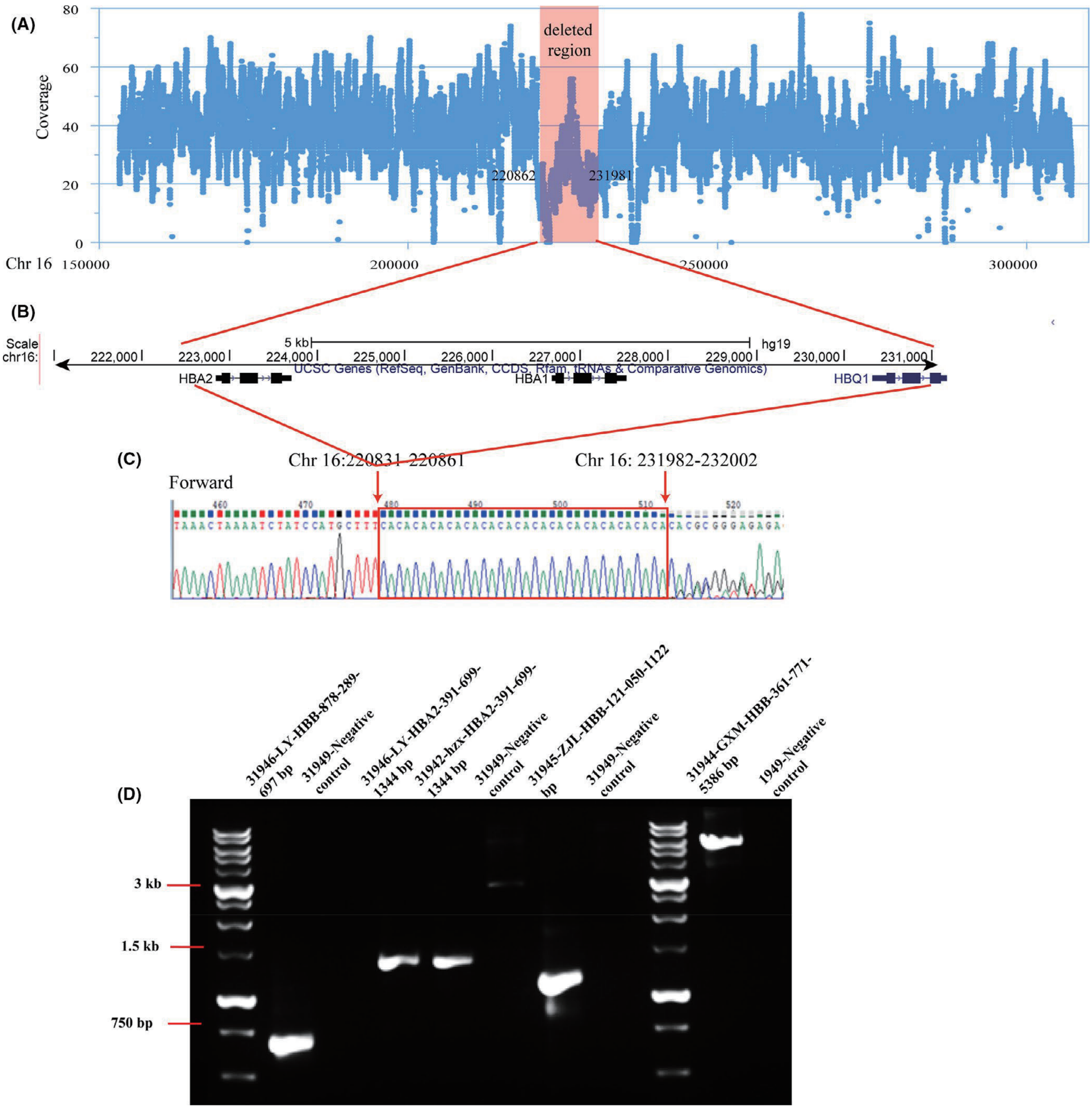


Figure 2. A summary of molecular analysis of hemoglobin deletions in 7 patients. A, Red rectangle: The WGS report successfully defined a structural variant in chromosome 16: 220,862-231,981 (GRCh37/hg19). The table shows that low coverage (<15 individuals' genomes) is the deleted region. B shows the normal region of the HBA1/2 cluster on chromosome 16. C, Sanger sequencing confirmed the breakpoint at chr16:220,862-231,981 for patient 48197-YCT. D, PCR confirmed the DNA breakpoints in the 7 patients.

Conclusions

This study selected 7 patients with thalassemia as the subjects and performed 400bp long-read whole-genome sequencing (WGS) to accurately detect thalassemia gene deletions.

The results show that the SE400 WGS independently developed by MGI is a powerful tool for detecting thalassemia breakpoints. It can not only detect known deletions but also identify new coding/non-coding deletions related to the disease, thereby expanding the research scope to the non-coding region and improving diagnostic efficiency. The findings of this study suggest that this is a highly promising diagnostic tool.

DNBSEQ-G400 Genetic Sequencer

References

1. Jiang F, Lyu GZ, Zhang VW, Li DZ. Identification of thalassemia gene cluster deletion by long-read whole-genome sequencing (LR-WGS). *Int J Lab Hematol*. 2021 Aug;43(4):859-865.
2. Verdura E, Schluter A, Fernandezeulate G, et al. A deep intronic splice variant advises reexamination of presumably dominant SPG7 cases. *Ann Clin Transl Neurol*. 2020;7(1):105-111.
3. Gilissen C, Hehirkwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*. 2014;511(7509):344-347.
4. Lindstrand A, Eisfeldt J, Pettersson M, et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Med*. 2019;11(1):68.
5. Taher AT, Weatherall DJ, Cappellini MD. Thalassaemia. *Lancet (London, England)*. 2018;391(10116):155-167.
6. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. *Bull World Health Organ*.2001;79(8):704-712.
7. Lai K, Huang G, Su L, He Y. The prevalence of thalassemia in mainland China: evidence from epidemiological surveys. *Sci Rep*.2017;7(1):920-930.

FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES.

Recommended Ordering Information

Category	Product	Cat. NO.
Instruments	Genetic Sequencer DNBSEQ-G400RS	900-000170-00
	MGISP-100RS Automated Sample Preparation System	900-000206-00
	MGISP-960RS Automated Sample Preparation System	900-000146-00
Software	MegaBOLT Bioinformatics analysis accelerator	900-000555-00
Library Prep	MGIEasy Universal DNA Library Prep Set（16 RXN）	1000006985
Sequencing Reagents	DNBSEQ-G400RS High-throughput Sequencing Set (FCL SE400)	1000016946

MGI Tech Co.,Ltd

Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083

Version: November 2023

☎ +86-4000-688-114

en.mgi-tech.com

MGI-service@mgi-tech.com

Authors: Huo Mingxing, Ma Jinglin

Editor-in-Charge: Wang Qiwei

Reviewer: Jiang Yao

1. For StandardMPS and CoolMPS: Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, Spain, UK, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania. Unless otherwise informed, StandardMPS sequencing reagents, and sequencers for use with such reagents are not available in Hong Kong. No purchase orders for StandardMPS products will be accepted in the USA until after January 1, 2023.
2. For HotMPS sequencers: This sequencer is only available in selected countries, and its software has been specially configured to be used in conjunction with MGI's HotMPS sequencing reagents exclusively.
3. For HotMPS reagents: This sequencing reagent is only available in selected countries.