EBioMedicine 23 (2017) 150–159

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Rapid Targeted Next-Generation Sequencing Platform for Molecular Screening and Clinical Genotyping in Subjects with Hemoglobinopathies

Xuan Shang [a,z,aa,1], Zhiyu Peng [b,1], Yuhua Ye [a,z,aa,1], Asan [s,t,1], Xinhua Zhang [c], Yan Chen [d], Baosheng Zhu [e], Wangwei Cai [f], Shaoke Chen [g], Ren Cai [h], Xiaoling Guo [i], Chonglin Zhang [j], Yuqiu Zhou [k], Shuodan Huang [l], Yanhui Liu [m], Biyan Chen [n], Shanhuo Yan [o], Yajun Chen [p], Hongmei Ding [q], Xiaolin Yin [c], Liusong Wu [d], Jing He [e], Dongai Huang [f], Sheng He [g], Tizhen Yan [h], Xin Fan [g], Yuehong Zhou [r], Xiaofeng Wei [a,z,aa], Sumin Zhao [s,t], Decheng Cai [a,z,aa], Fengyu Guo [s,t], Qianqian Zhang [a,z,aa], Yun Li [u], Xuelian Zhang [a,z,aa], Haorong Lu [u], Huajie Huang [a,z,aa], Junfu Guo [s,t], Fei Zhu [a,z,aa], Yuan Yuan [s,t], Li Zhang [a,z,aa], Na Liu [u], Zhiming Li [a,z,aa], Hui Jiang [b], Qiang Zhang [a,z,aa], Yijia Zhang [a,z,aa], Wan Khairunnisa Wan Juhari [v], Sarifah Hanafi [v], Wanjun Zhou [a,z,aa], Fu Xiong [a,z,aa], Huanming Yang [b,y], Jian Wang [b,y], Bin Alwi Zilfalil [v], Ming Qi [w,ab], Yaping Yang [x], Ye Yin [b], Mao Mao [b,*], Xiangmin Xu [a,z,aa,**]

[a] Department of Medical Genetics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, Guangdong, China
[b] BGI-Shenzhen, Bei Shan Industrial Zone, Yantian District, Shenzhen, Guangdong, China
[c] Department of Hematology, 303rd Hospital of the People's Liberation Army, Nanning, Guangxi, China
[d] The Second Department of Pediatrics, Affiliated Hospital of Zunyi Medical College, Zunyi, Guizhou, China
[e] Genetic Diagnosis Center, First People's Hospital of Yunnan Province, Medical School of Kunming University of Science and Technology, Kunming, Yunnan, China
[f] Department of Biochemistry and Molecular Biology, Hainan Medical College, Haikou, Hainan, China
[g] Department of Genetic and Metabolic Laboratory, Guangxi Zhuang Autonomous Region Women and Children Health Care Hospital, Nanning, Guangxi, China
[h] Department of Medical Genetics, Liuzhou Municipal Maternity and Child Healthcare Hospital, Liuzhou, Guangxi, China
[i] Maternity and Child Health Care Hospital of Foshan City, Foshan, Guangdong, China
[j] Guilin Women and Children health care hospital, Guilin, Guangxi, China
[k] Department of Clinical Laboratory, Zhuhai Municipal Maternal and Child Healthcare Hospital, Zhuhai Institute of Medical Genetics, Zhuhai, Guangdong, China
[l] Maternal and Child Health Hospital in Meizhou, Meizhou, Guangdong, China
[m] Department of Prenatal Diagnosis Center, Dong Guan Maternal and Child Health Hospital, Dongguan, Guangdong, China
[n] Baise Women and Children Care Hospital, Baise, Guangxi, China
[o] Genetic Laboratory, Qinzhou Maternaland Child Health Hospital, Qingzhou, Guangxi, China
[p] Women and Children's Health Hospital of Shaoguan, Shaoguan, Guangdong, China
[q] Department of Gynecology and Obstetrics, The People's Hospital of Yunfu City, Yunfu, Guangdong, China
[r] Pingguo Women and Children Care Hospital, Baise, Guangxi, China
[s] Tianjin Medical Laboratory, BGI-Tianjin, BGI-Shenzhen, Tianjin, China
[t] Binhai Genomics Institute, BGI-Tianjin, BGI-Shenzhen, Tianjin, China
[u] BGI Clinical Laboratories-Shenzhen, BGI-Shenzhen, Shenzhen, China
[v] Department of Paediatric, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia
[w] School of Basic Medical Sciences, Zhejiang University, Hangzhou, Zhejiang, China
[x] Departments of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA
[y] James D. Watson Institute of Genome Sciences, Hangzhou, Zhejiang, China
[z] Guangdong Technology and Engineering Research Center for Molecular Diagnostics of Human Genetic Diseases, Guangzhou, Guangdong, China
[aa] Guangdong Key Laboratory of Biological Chip, Guangzhou, Guangdong, China
[ab] Center for Genetic & Genomic Medicine, Zhejiang University Medical School 1st Affiliated Hospital, James Watson Institute of Genome Sciences, Hangzhou, Zhejiang, China

ARTICLE INFO

Article history:
Received 29 June 2017
Received in revised form 15 August 2017
Accepted 15 August 2017
Available online 17 August 2017

ABSTRACT

Hemoglobinopathies are among the most common autosomal-recessive disorders worldwide. A comprehensive next-generation sequencing (NGS) test would greatly facilitate screening and diagnosis of these disorders. An NGS panel targeting the coding regions of hemoglobin genes and four modifier genes was designed. We validated the assay by using 2522 subjects affected with hemoglobinopathies and applied it to carrier testing in a cohort of 10,111 couples who were also screened through traditional methods. In the clinical genotyping analysis of 1182

* Correspondence to: Mao Mao, BGI-Shenzhen, Bei Shan Industrial Zone, Yantian District, Shenzhen, Guangdong 518000, China.
** Correspondence to: Xiangmin Xu, Department of Medical Genetics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, Guangdong 510515, China.
E-mail addresses: maomao@genomics.cn (M. Mao), xixm@smu.edu.cn, gzxuxm@pub.guangzhou.gd.cn (X. Xu).
[1] These authors contributed equally to this study.

http://dx.doi.org/10.1016/j.ebiom.2017.08.015

Keywords:
Hemoglobinopathy
Next-generation sequencing
Molecular screening
Clinical genotyping

β-thalassemia subjects, we identified a group of additional variants that can be used for accurate diagnosis. In the molecular screening analysis of the 10,111 couples, we detected 4180 individuals in total who carried 4840 mutant alleles, and identified 186 couples at risk of having affected offspring. 12.1% of the pathogenic or likely pathogenic variants identified by our NGS assay, which were undetectable by traditional methods. Compared with the traditional methods, our assay identified an additional at-risk 35 couples. We describe a comprehensive NGS-based test that offers advantages over the traditional screening/molecular testing methods. To our knowledge, this is among the first large-scale population study to systematically evaluate the application of an NGS technique in carrier screening and molecular diagnosis of hemoglobinopathies.

1. Introduction

Hemoglobinopathies, including sickle cell anemia and α-/β-thalassemia, are the most common monogenic diseases worldwide (Higgs et al., 2012). Approximately 340,000 children with significant hemoglobin (Hb) disorders are born worldwide each year, 90% of whom are in developing and low-income countries, where they represent a massive public health burden (Williams and Weatherall, 2012). Prevention programs based on carrier screening and prenatal diagnosis have resulted in a continuous decline in rates of thalassemia major at birth in the Mediterranean region (Cao and Kan, 2013). Such prospective screening has recently been introduced in at-risk countries in Asia (Ahmed et al., 2002; Cao and Kan, 2013). Because the identification of high-risk couples by carrier testing is necessary for effective prevention, prevention programs must determine how to perform comprehensive carrier screening in large populations.

A traditional routine strategy for thalassemia carrier screening first identifies people with phenotypic traits associated with thalassemia by using hematological and biochemical tests and subsequent molecular genetic testing in this selected group to generate definitive diagnoses (Cao and Kan, 2013; Traeger-Synodinos et al., 2015). However, a phenotypic screening approach in carrier testing would not detect individuals with "silent" forms of thalassemia, most of whom may be missed because they have normal or borderline red cell indices and/or HbA_2 levels (Piel and Weatherall, 2014; Traeger-Synodinos et al., 2015). In addition, the characterization of disease-causing defects in samples from individuals suspected of having thalassemia may require various labor-intensive methodologies (Cao and Kan, 2013; Traeger-Synodinos et al., 2015). At least 1530 mutations causing thalassemia or abnormal hemoglobin variants, ranging from single-base changes to large rearrangements, have been characterized to date (HbVar database for human hemoglobin variants and thalassemia mutations, n.d). Moreover, several variants that modify β-thalassemia phenotypes have been identified at various loci (Giardine et al., 2011; Sankaran and Weiss, 2015; Thein, 2013). The application of high-throughput molecular approaches to direct mass screening and accurate molecular diagnosis of hemoglobinopathies raises several challenges.

Next-generation sequencing (NGS) has been shown to allow rapid, multiplex and high-throughput detection of genetic variants (Korf and Rehm, 2013). NGS technologies—applied to the whole genome, the exome, or targeted gene panels—have been effectively used in research settings, as well as in clinical testing and diagnosis of genetic disorders (Stark et al., 2016; Yang et al., 2013). Recently, target capture and NGS have been validated in patient-based and population-based carrier testing of Mendelian recessive diseases, thereby yielding high-quality genotype calls and acceptable false-positive and false-negative rates and cost-effectiveness (Chong et al., 2012; Haque et al., 2016).

Here, to evaluate whether NGS might be suitable for application in clinical management of monogenic diseases in a large-scale population, we retrospectively analyzed 2522 subjects with various hemoglobinopathies by using a targeted NGS approach and validated this assay for carrier testing of multiple genetic defects in a cohort of 20,222 individuals pre-typed by routine screening methods. A population-based study of hemoglobinopathies in southern China was also performed. Our results demonstrated that NGS, compared with traditional methods, can detect pathogenic or likely pathogenic variants in a more precise and general manner, thus providing an effective platform for molecular screening and clinical genotyping in subjects with hemoglobinopathies.

2. Materials and Methods

2.1. Study Participants

The study was approved by the Medical Ethics Committee in accordance with the Declaration of Helsinki. The study methodology was designed with guidance by STROBE (von Elm et al., 2007).

Samples from two groups of subjects were collected for this study. In Group I, 2522 subjects included 2087 samples used for study methodology of NGS assay and the sample data of 435 Thalassemia Intermedia (TI)/Thalassemia Major (TM) patients from our previous cohort (Fig. 1) (Liu et al., 2014). The 2087 subjects were from China and Malaysia and included 49 normal controls and 2038 clinical samples with various hemoglobinopathies (Table S1 in the Supplementary appendix). They were used for assessing the performance of the NGS-based assay in detecting hemoglobinopathy-causing mutations in a retrospective study. The 49 controls had been sequenced previously, and they did not carry any pathogenic or likely pathogenic variants. The 2038 clinical samples had been analyzed by traditional molecular genotyping techniques (See the "Genotypic Analysis Using Traditional Methods" section). NGS analysis was performed on these pre-typed results with blinding, and the results were subsequently compared with the pre-typed results to obtain the concordance rate. Sanger sequence analysis and multiplex ligation-dependent probe amplification (MLPA) were performed to verify samples with discrepancies. Because some modifier genes had been shown to be responsible for modifying β-thalassemia phenotypes, we also analyzed whether genotypes at these loci might assist in β-thalassemia diagnosis in 1182 β-thalassemia patients with thalassemia major (TM) or thalassemia intermedia (TI; 747 patients of the 2087 samples described above and 435 patients from the cohort described in a previous study by our group (Fig. 1; Liu et al., 2014). Among the 747 samples, 510 were β^0/β^0, and 237 were β^0/β^+. Among the 435 samples, 300 were β^0/β^0 and 135 were β^0/β^+. In total, the 1182 samples included 810 β^0/β^0 and 372 β^+/β^0.

Group II, which included a total of 10,111 couples (either before or during pregnancy) were randomly selected from five provinces (Guangxi, Guandong, Yunnan, Guizhou and Hainan) in southern China (Fig. 2). The age range of these samples was 18–48 years old. We applied the assay in a blind analysis for carrier testing on these couples, who were also enrolled in prevention programs for thalassemias using traditional screening methods (See the "Hematological Analysis" and "Genotypic Analysis Using Traditional Methods" sections). They were also included in an epidemiological survey of hemoglobinopathies.

2.2. Thalassemia Genotype Definition and Clinical Diagnosis in Thalassemia Patients

HBB genotype categories are defined as follows: (β^0): HBB:c.124_127delTTCT, HBB:c.52A>T, HBB:c.316-197C>T,

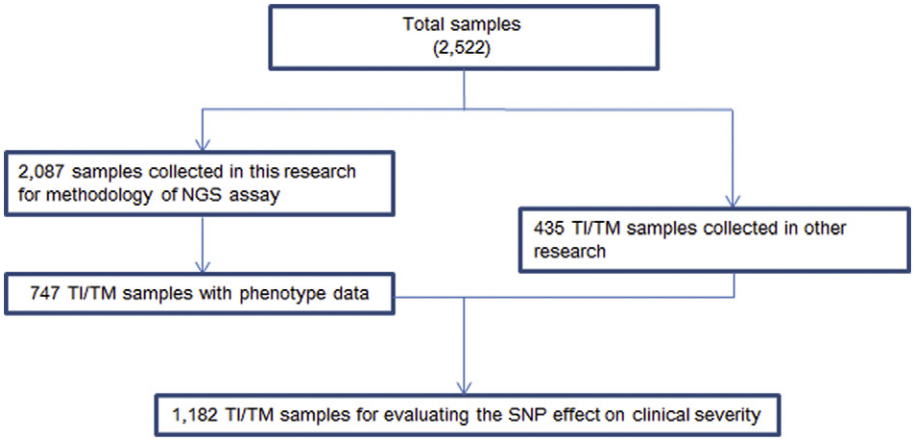

Fig. 1. The sample composition of Group I. A total of 2522 clinical samples with hemoglobinopathies were used in our study. The composition and application of these samples in this study are shown. Among them are included 1182 samples with TM or TI, which were included to evaluate SNP effects on clinical severity. The data from 435 TI/TM samples were collected from a cohort from our previous report, which had been thoroughly sequenced by NGS (Liu et al., 2014). Among the 747 samples, 510 β^0/β^0 patients were first used to identify the SNPs that might elevate HbF levels (see Fig. S6).

HBB:c.216_217insA, HBB:c.92 + 1G>T, HBB:c.130G>T, HBB:c.84_85insC, HBB:c.91A>G, HBB:c.45_46insC, HBB:c.165_177delTATGGGCAACCCT, HBB:c.315 + 1G>A, HBB:c.287_288insA, HBB:c.113G>A, HBB:c.93-1G>C, Chinese del(NC_000011.9:g.5191148_5270051del), Cantonese del(NC_000011.9:g.(5240000_5246696)_5271087del), Taiwanese del(NC_000011.9:g.5247493_5248849del), Filipino del(NC_000011.9:g.5134113_5252589del), β7.3k del(NC_000011.9:g.5247824_5255222del), β21k del (NC_000011.9:g.5246909_5268823del), β118kdel (NC_000011.9:g.5135464_5254173del); (β^+): HBB:c.-78A>G, HBB:c.79G>A, HBB:c.-79A>G, HBB:c.315 + 5G>C, HBB:c.-140C>T, HBB:c.-81 A>C, HBB:c.93–21G>A and SEA-HPFH(NC_000011.9:g.5222878_5250288del). *HBA* genotype categories are defined as: -α: $-\alpha^{3.7}$ (NC_000016.9:g.223300_227103del) and $-\alpha^{4.2}$(NC_000016.9:g.219817_(223755_224074)del); –: $-^{SEA}$(NC_000016.9:g.215400_234700del) and $-^{Thai}$(NC_000016.9:g.199800_233300del); $\alpha^T\alpha$: $\alpha^{CS}\alpha$ (Hb Constant Spring. HBA2:c.427T>C), $\alpha^{Westmead}\alpha$ (Hb Weastmead, HBA2:c.369C>G), $\alpha^{QS}\alpha$ (Hb Quong Sze, HBA2:c.377T>C), $\alpha^{Q\text{-}thailand}\alpha$ (HBA1:c.223G>C) and $\alpha^{Binyang}\alpha$ (HBA2:c.40G>T).

The definitions of TM and TI in this study were based on the following clinical indications as described previously (Steinberg et al., 2009; Thuret et al., 2010): (1) onset age < 24 months (TM), or onset age

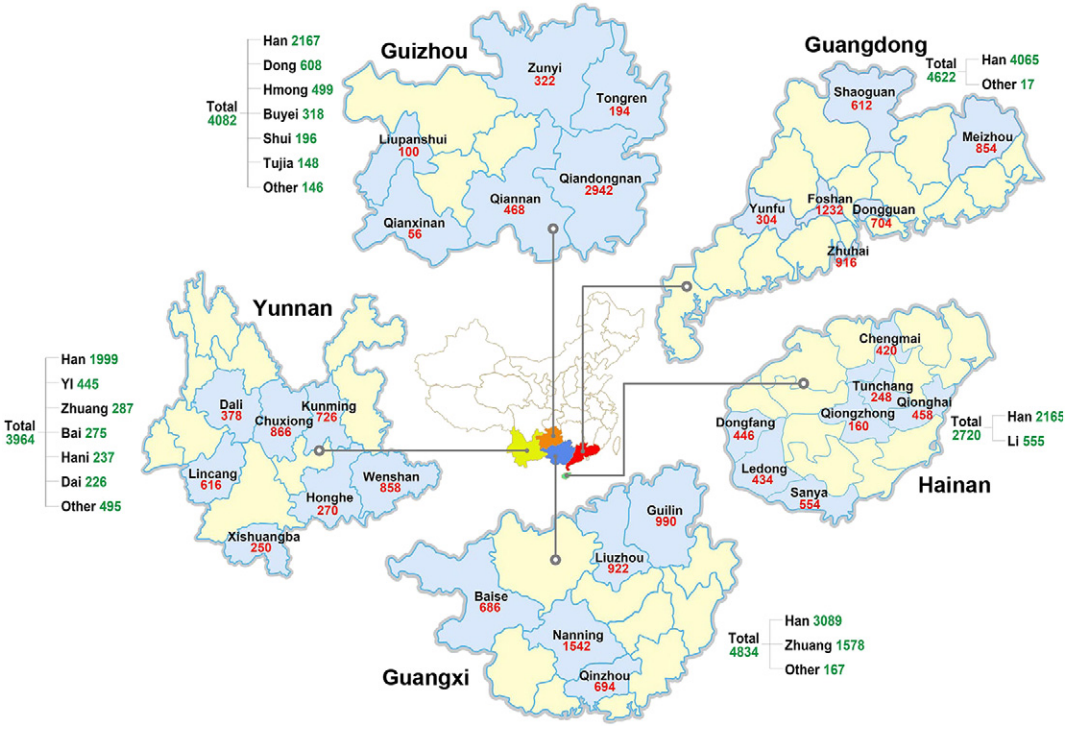

Fig. 2. The regional and ethnicity distribution of the 20,222 population samples from five provinces in southern China. Among the 20,222 individuals, 4622 were from Guangdong Province, 4834 were from Guangxi Province, 4082 were from Guizhou Province, 2720 were from Hainan Province and 3964 were from Yunnan Province. The ethnic groups of these samples are also listed.

≥ 24 months (TI); (2) transfusion times per year > 4 (TM), or transfusion times per year ≤ 4 (TI); (3) Hb F < 20 g/L (TM), or Hb F ≥ 20 g/L (TI). The Hb F level was measured before transfusion.

2.3. Hematological Analysis

Hematological parameters were determined with an automated hematology analyzer (Sysmex, Japan), and hemoglobin analysis was performed with either high-performance liquid chromatography (Bio-Rad, USA) or capillary electrophoresis (Sebia, France and Helena, USA).

2.4. Genotypic Analysis Using Traditional Methods

Genomic DNA was extracted from peripheral blood (PB) by using a standard phenol/chloroform method. The following genotyping techniques were used: Gap-PCR and multiplex ligation-dependent probe amplification (MLPA) for CNVs, reverse dot blot (RDB), high-resolution melting analysis (HRMA) and Sanger sequencing for SNVs. All techniques were performed according to standard procedures.

The genotypes of 2087 samples (49 normal controls and 2038 clinical samples with various hemoglobinopathies) were analyzed with Gap-PCR, MLPA, RDB and Sanger sequencing methods. The additional mutations detected by the NGS assay, which are listed in Supplementary Table S5, were subsequently validated by Gap-PCR, MLPA and Sanger sequencing methods.

The genotypes of selected samples from the 20,222 individuals were analyzed by using Gap-PCR, MLPA, HRMA and Sanger sequencing methods. The additional mutations detected by the NGS assay were subsequently confirmed by MLPA and Sanger sequencing.

2.5. Design of the NGS Assay

We targeted the entire protein-coding regions, key regulatory regions, known pathogenic copy number variants (CNVs) regions and single nucleotide variants (SNVs)/insertion and deletion variants (indels) in the non-coding regions of hemoglobin gene clusters (α- and β-globin gene clusters) and 4 modifier genes (*KLF1*, *BCL11A*, *HBS1L* and *MYB*). We used 21 autosomal SNP sites and 6 sex chromosomes genes as markers for identity and sex tracing of the samples. The total size of the targeted sequences was 275,234 bp. (Details are shown in Fig. S1–S3 and the "Protocol of NGS assay" section in the Supplementary appendix).

2.6. Genomic DNA Extraction and DNA Quality Characterization for NGS

Genomic DNA was extracted automatically from periphery whole blood with a MagPure Buffy Coat DNA Midi KF Kit (Magen, China). DNA sample quality and quantity were measured using gel electrophoresis and/or a NanoDrop spectrophotometer (Thermo Scientific, USA).

2.7. Library Preparation and Sequencing

We initially established a standard method based on mechanical fragmentation, which required multiple steps for the preparation of a targeted sequencing library (Adey et al., 2010), and used it to prepare a library for most of the archived clinical samples. Briefly, 1 μg genomic DNA per sample was sheared into fragments of 250 bp. The DNA was then purified after characterization of fragment size. End repair, A-tailing at the 3′ ends and adapter ligation was performed through pre-capture amplification. To further improve the library preparation in terms of input DNA and automation, we optimized a Tn5 transposase-based method for rapid and low-input library preparation. For each sample, approximately 50 ng DNA was fragmented into ~200-bp segments and barcoded using an in-house transposome system (Picelli et al., 2014). Then, pre-capture amplification and PCR product purification was performed. Targeted sequence capture was conducted by pooling each of the 47 indexed PCR products for the two methods for libraries and hybridization with custom capture probes. The yield libraries were validated and quantified before sequencing. DNA sequencing was performed on a HiSeq2000 or 4000 sequencer (Illumina, USA) in paired-end 100-bp reads (PE100). Each sample yielded an average 0.26–0.27 GB raw data, which provided a mean sequencing depth of 178–191 folds on the targeted regions, and 97.9–98.6% of nucleotides were covered by >20 uniquely mapped reads (Table S2 in the Supplementary appendix).

2.8. Data Analysis and Interpretation

The Illumina Pipeline software program (version 1.8) was used to process the raw image data. After image processing was complete, the raw fastq data was automatically delivered to an in-house integrated pipeline named Gaea (see Fig. S4A in the Supplementary appendix), which was based on Hadoop MapReduce Technology and has been designed to accelerate the speed of data analysis for subsequent steps of bioinformatic analysis (Marx, 2013). In detail, data filtering was performed using an in-house Perl script called GaeaFastqQC to filter out the reads with low quality or from sequencing adaptors. Then, read mapping to the human reference genome (UCSC build hg19) was performed using the BWA-based software GaeaBwaStreaming. In addition, duplicate marking, indel realignment and base quality score recalibration steps were used in post-processing of the mapped reads. Finally, SNVs (missense, splice, nonsense, synonymous variants, as well as variants in introns and promoter regions, etc.) and indels were detected using GATK-based software, GaeaGenotyper, with the UnifiedGenotyper algorithm and annotated using an in-house program called BGICGAnnotation. Structural variations were detected in parallel by two individual programs, BatCNV and Breakpointer (see Fig. S4 B in the Supplementary appendix). BatCNV was based on a hidden Markov model (HMM) algorithm and was used to detect large gene deletions and duplications by analyzing changes in read depth within the targeted region. Three QC criteria were used to filter out abnormal samples at first. Then, GC correction and Batch correction were used to minimize data fluctuations. Finally, an HMM was applied to predict the most likely copy number. Breakpointer used an alternative approach based on clustering split reads to pinpoint the exact break-points of CNVs (if possible). The properly mapped paired-end reads with soft-clipped ends were collected first. The clipped ends were remapped with the BWA-MEM algorithm. Then, reads with the same clipped position and remapped position were clustered into a group. The exact breakpoint was inferred from the group's clipped-remapped position pair. Finally, variants were interpreted mainly on the basis of the guidelines of the American College of Medical Genetics (ACMG) and the literature (MacArthur et al., 2014; Richards et al., 2008, 2015) and were classified as pathogenic, likely pathogenic, of unknown clinical significance, likely benign or benign, on the basis of the level of evidence supporting pathogenicity (Table S3 in the supplementary appendix). Only pathogenic and likely pathogenic variants were considered reportable and have been included in the final evaluation of each individual's mutation carrier status. The process is summarized in Fig. S5 in the Supplementary appendix. The pathogenicity of novel variants was analyzed by experts in hemoglobinopathy and listed in "Index of the variants" section in Supplementary. Notably, the pathogenicity of variants in modifier genes was not determined according to the criteria of Table S3 since these variants themselves were not direct disease cause. However, genotypes of these genes should be considered in accurate diagnosis of β-thalassemia patients and identification of at-risk couples because modifier genes could affect the clinical severity of thalassemia (Liu et al., 2014).

2.9. Statistical Analysis to Identify SNPs Associated with Clinical Severity

A total of 1182 β-thalassemia patients were enrolled to identify the modifying variants that might significantly affect the clinical severity of

Table 1
Performance evaluation of variant detection with an NGS-based gene panel.

Locus	NGS assay	Traditional molecular assay			Concordance rate (%)
		concordance (pre-typed)	concordance (re-typed)	Discordance	
α-globin locus	4174	4157	16	1[a]	99.98
β-globin locus	4174	4165	9	0	100

This analysis was based on the mutant allele counts in the 2087 samples tested, for a total allele count of 4174.
Concordance (pre-typed): number of alleles whose pre-typed results were consistent with the result of NGS.
Concordance (re-typed): number of alleles whose re-typed results were consistent with the result of NGS. (NGS detected additional variants not reported in the pre-typed results. These 25(16 + 9) variants were subsequently confirmed through traditional molecular methods.)
[a] This allele was a mutant allele with a deletion but was diagnosed as wild-type by the NGS assay.

β-thalassemia (Fig. 1). As shown in Fig. S6, a *t*-test of the 510 β^0/β^0 samples, was implemented in Perl (v5.16.3) to compare the mean HbF levels between the carrier and non-carrier group of each variant with frequencies over 0.01. An association study was conducted in Plink, with the HbF z-score as the dependent variable. The variants found to be significant by both *t*-test and the association study according to *p* values after Bonferroni correction were input into a Cox regression model, together with the *KLF1* functional mutations and *HBA* disease-causing mutations. The onset ages for each sample were introduced as the dependent variable in the Cox regression. Finally, the positive variants were considered to exert a significant effect on both the HbF levels and onset ages of β thalassemia patients. All statistical analyses were performed using SPSS 19.0 software. *P*-values < 0.05 were considered significant.

3. Results

3.1. Accurate and Comprehensive Identification of Genetic Variants Associated With Hemoglobinopathies

In a retrospective study, we developed a comprehensive NGS-based test for hemoglobinopathies and validated its genotyping performance in 2087 samples (2038 hemoglobinopathy patients and 49 controls) pre-typed by traditional molecular testing methods (Table S1 in the Supplementary appendix). The NGS assay identified 72 hemoglobinopathy-associated variants, accurately characterizing nearly all common and several rare thalassemia mutations known to exist in Chinese populations (Table S4 in the Supplementary appendix). No hemoglobinopathy-associated mutations were detected in the 49 controls in our study cohort. In summary, we validated our target capture and NGS assays by using a large set of reference samples, which yielded high-quality genotype calls, as well as excellent concordance rates at the α- and β-globin loci (Table 1). The assay identified 14 additional variants in 25 individuals (Table S5 in the Supplementary appendix). All the additional mutations were also subsequently re-typed by traditional molecular methods (Table 1). The assay failed to detect only one large deletion in the α-globin cluster due to aberrant sequencing depth distribution in the CNV analysis pipeline, possibly because of low DNA quality.

3.2. Wider Molecular Characterization of Patients with TM and TI

To assess the ability of our assay to accurately diagnose β-thalassemia patients in clinical settings, we systematically analyzed the NGS data provided by 1182 samples (810 β^0/β^0 TM or TI and 372 β^0/β^+ TM or TI; see detailed information of phenotype parameters including Hb F levels, onset of transfusion and number of transfusions per year in Table 2a). In addition to the main genotypes in the *HBA* and *HBB* genes, we identified a group of single-nucleotide polymorphisms

Table 2
Summary of the phenotypic data (a) and effect of variants on the distribution of TM and TI in 1182 β-thalassemia patients (b).

(a) The phenotypic data

	β^0/β^0 ($n = 810$)	β^0/β^+ ($n = 372$)	*p*-value[a]
Sex (n)			
Males:females	524:286	233:139	0.494
Clinical data			
TM:TI	687:123	230:142	<0.001
Hematological data[b]			
HbF level(g/L)	12.63 ± 14.75	18.39 ± 15.58	<0.001
Onset age(month)	13.17 ± 27.54	34.85 ± 54.02	<0.001
Transfusion times per year	12.22 ± 6.41	9.03 ± 6.57	<0.001

(b) Effect of variants on the distribution of TM and TI

Number of variants	Total	β^0/β^0 ($n = 810$)			β^0/β^+ ($n = 372$)		
		TM (%)	TI (%)	*P*-value	TM (%)	TI (%)	*P*-value
0	430	279 (93.3)	20 (6.7)	NA	97 (74.0)	34 (26.0)	NA
1	471	270 (85.7)	45 (14.3)	0.003423[c]	95 (60.9)	61 (39.1)	0.02563[d]
2	232	121 (76.1)	38 (23.9)	2.977E^{-07}[c]	35 (47.9)	38 (52.1)	0.000335[d]
3	42	17 (53.1)	15 (46.9)	1.771E^{-11}[c]	3 (30)	7 (70)	0.009455[d]
4	7	0 (0)	5 (100)	NA	0 (0)	2 (100)	NA
5	0	0	0	NA	0	0	NA
Total	1182	687	123	NA	230	142	NA

Results based on χ^2 analysis.
The total number of the samples included was 1182 (sample composition shown in Fig. 1). The five variants: *KLF1* mutations, rs368698783 (*Xmn*I), rs61749494 (*BCL11A*), rs11759553 (*HBS1L-MYB*) and *HBA* disease-causing mutations.
Abbreviations: NA, not applicable.
[a] *P*-value was determined by either a Kruskal-Wallis test or the χ^2 test between 2 genotypes.
[b] Hematological data are shown as the means ± standard deviation.
[c] β^0/β^0 patients carrying variants had a significantly higher chance of exhibiting a TI phenotype than patients carrying no variants ($p < 0.01$).
[d] β^0/β^+ patients carrying variants had a significantly higher chance of exhibiting a TI phenotype than did patients carrying no variants ($p < 0.01$).

(SNPs) in four major modifier genes and cis element variants in the β-globin cluster that contribute to β-thalassemia (Giardine et al., 2011; Sankaran and Weiss, 2015; Thein, 2013). The cis-elements included hypersensitive sites (HS) 1–5 of the locus control region (LCR), promoter or untranslated region (UTR) of β-globin or γ-globin, etc. We first used a panel of SNPs discovered in the targeted regions to identify SNPs that might affect levels of HbF in 510 consecutive TI/TM patients with β^0/β^0 genotypes (Fig. 1). *KLF1* mutations, rs368698783 (*Xmn*I), rs61749494 (*BCL11A*), rs11759553 (*HBS1L-MYB*) and *HBA* disease-causing mutations were identified as the five most important ameliorators, as previously reported in other studies (Danjou et al., 2015; Galanello et al., 2009; Liu et al., 2014) (Fig. S6 in the Supplementary appendix). We then evaluated their effect on clinical severity in 1182 β-thalassemia patients. NGS-based genotyping was used to perform accurate molecular characterization of β-thalassemia patients by identifying disease-causing mutations and modifier variants simultaneously. As shown in Table 2b, 93.3% of β^0/β^0 patients with none of the five variants described above would be diagnosed as having TM, whereas this probability dropped dramatically (85.7%) in carriers of any one of the variants ($p = 0.0034$). The same trend was observed for β^0/β^+ patients ($p = 0.0256$). Similarly, the more variants carried by β^0/β^0 or β^0/β^+ patients, the greater their chance of developing a TI phenotype. Those who carried four of the five variants would be diagnosed as having TI without exception.

3.3. Evaluation of Targeted NGS Assay for Its Application to Carrier Testing

A total of 10,111 pregestational or prenatal couples from multiple ethnic groups from five provinces (Guangxi, Guandong, Yunnan, Guizhou and Hainan) were enrolled for evaluation of carrier screening (Fig. 2). The couples were separately screened with the NGS-based test and traditional routine screening methods (details shown in Fig. S7 in the Supplementary appendix). We identified 186 at-risk couples by using the NGS method, compared with 151 at-risk couples by using the routine method (Fig. 3). The proportions of couples identified as being at risk for thalassemia by using the NGS method and the routine method were 1.84% (186/10,111) and 1.50% (151/10,111), respectively. We identified 35 additional couples of the 10,111 couples by using the NGS method compared with the traditional screening/molecular testing method, thus representing a 23.2% increase in the number of positive screens (Table 3). Of these 35 couples, 17 were at risk for TM or TI, 13 for Hb H disease, three for TM or TI co-inherited via KLF1 variants, one for Hb Bart's hydrops fetalis and one for an atypical thalassemia caused by compound heterozygosity of KLF1 mutations (Table 3). All mutations in the α- or β-globin cluster and *KLF1* gene were confirmed through routine molecular methods. These 35 at-risk couples were not detected by the traditional screening pipeline owing to inherent deficiencies in the procedure. These deficiencies originated from the traditional screening strategy itself, such as misleading due to phenotype analysis and limitations of routine molecular methods. Briefly, 12 couples had a negative phenotype in the first step of hematologic testing, mutations in seven couples were misdiagnosed and mutations in 16 couples were undetected (Fig. 3 and Table S6 in the Supplementary appendix). The individuals showed negative phenotype had been confirmed. Interestingly, approximately half of these individuals were from the same ethnic group and lived in an isolated region of Guizhou Province. Therefore, this pattern might be related to the genetic background of a specific population.

3.4. A Population-Based Molecular Epidemiological Investigation of Hemoglobinopathies

We thoroughly analyzed 20,222 subjects by using an NGS assay to detect pathogenic variants for hemoglobinopathies. At total of 4840

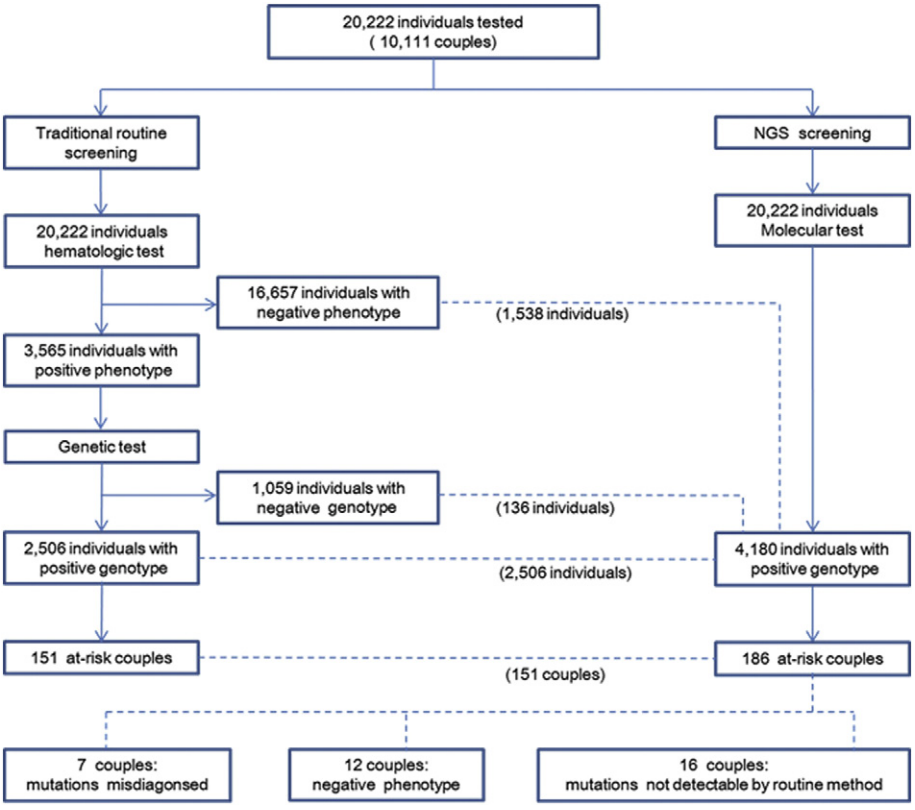

Fig. 3. Comparison of couples identified as being at risk for hemoglobinopathies by using the NGS method and the traditional routine method. A total of 4180 out of 20,222 individuals tested positive with the NGS method. Of these 4180 individuals, 2506 were also identified as positive by the traditional routine method. A total of 1538 individuals were not detected by the routine method because they showed negative phenotypes, and 136 individuals were not detected because their mutations were misdiagnosed or could not be detected by the routine method. Overall, an additional 35 couples were identified as being at risk for α- and β-thalassemia using our NGS assay.

Table 3
Details of the 151 at-risk couples identified by routine methods and the 186 at-risk couples identified by NGS.

Forms of hemoglobinopathy for which fetuses would be at risk		Number of couples testing positive	
Major classes	Genotypes	Routine	NGS
Hb Bart's	(- -/- -)	25	26
Hb Bart's or mild Hb Bart's	(- -/- -) ± (β^M/β^N)	4	4
Hb Bart's/Hb H	(- -/- - or - -/-α)	3	3
Hb Bart's/Hb H or mild Hb Bart's/Hb H	(- -/- - or - -/-α or --/$\alpha^T\alpha$) ± (β^0/β^N)	1	1
HbH	(- -/-α)	54	61
	(- -/$\alpha^T\alpha$)	16	19
	($\alpha^T\alpha/\alpha^T\alpha$)	1	2
	(- -/-α or - -/$\alpha^T\alpha$)	3	4
HbH or mild HbH	(- -/-α) ± (β^M/β^N)	8	9
	(- -/$\alpha^T\alpha$)± (β^M/β^N)	5	5
TM or severe TM	(β^0/β^0) ± (αα/αααα)	0	1
TM	(β^0/β^0)	16	18
	(β^+/β^0)	4	5
TM or TI	(β^0/β^0) ± (αα/-α)	4	4
	(β^0/β^0) ± (αα/$\alpha^T\alpha$)	0	1
	(β^0/β^0) ± (αα/-α or -α/-α)	2	2
	(β^+/β^0) ± (αα/-α)	3	3
	(β^+/β^0) ± (αα/$\alpha^T\alpha$)	1	1
	(β^0/β^0) ± ($KLF1^M/KLF1^N$)	0	2
TI	(β^+/β^+)	1	1
	(β^{++}/β^0)	0	1
	(β^0/β^N, αααα/αα)	0	7
	(β^0/β^N, αααα/αα or αααα/αααα)	0	1
	(β^+/β^N, αααα/αα)	0	1
	(β^{++}/β^N, αααα/αα)	0	1
TI or Mild TI	(β^+/β^+) ± (αα/-α or -α/-α)	0	1
	(β^+/β^N, αααα/αα) ± ($KLF1^M/KLF1^N$)	0	1
Atypical thalassemia	($KLF1^M/KLF1^M$)	0	1
Total		151	186

Abbreviations: Hb Bart's, Hb Bart's' hydrops fetalis; HbH, Hb H disease; TM, thalassemia major; TI, thalassemia intermedia; β^M, mutant β-globin allele; β^N, normal β-globin allele; $KLF1^M$, mutant *KLF1* allele; $KLF1^N$, normal *KLF1* allele. ±: with or without. Atypical thalassemia: mutant *KLF1* homozygotes or compound heterozygotes will show microcytic hypochromic anemia, a phenotype similar to thalassemia.

mutant alleles from 4180 individuals were detected, thus yielding an overall variant carrier frequency of 23.93%, including 16.77% for α-thalassemia, 4.70% for β-thalassemia, 1.70% for CNV variants (1.69% associated with the α-globin cluster and 0.01% associated with the β-globin cluster), 0.43% for KLF1 variants and 0.33% for structural hemoglobin variants (Table 4). Compared with a previous screening approach that was partly based on molecular testing of a few common α-thalassemia defects (-$\alpha^{3.7}$/, -$\alpha^{4.2}$/and HBA2:c.369C>G in all samples (Xiong et al., 2010)), the NGS approach enabled us to identify an additional 586 (12.1%) mutant alleles, including 343 pathogenic copy-number variations (CNV), 114 disease-causing mutations with no phenotypic red cell changes, 87 alleles with zinc-finger mutations in KLF1, four alleles with new variants (- -$^{30.8}$, -$\alpha^{1.2}$, HBA1: c.96-1G>C and HBB: c.41C>T) and 38 carriers co-inherited with α- and/or β-globin gene defects. We characterized 79 genetic variants of various hemoglobinopathies, including 14 types of pathogenic KLF1 mutations in a single test (see "Index of the variants" section in Supplementary appendix). KLF1 mutations and CNVs frequencies were included in this epidemiological data because of their pathogenicity or modification of the β-thalassemia phenotype (Perkins et al., 2016; Thein, 2013). The pathogenicity of new variants in α-/β-globin gene cluster (see "Index of the variants" section in Supplementary appendix) was evaluated mainly according to the criteria of "K" type in Table S3.

4. Discussion

In this study, we developed and validated a target-based NGS assay for molecular screening and clinical genotyping in hemoglobinopathies with a single test. Our target regions included all eight globin genes (*HBZ*, *HBA1*, *HBA2*, *HBE1*, *HBG1*, *HBG2*, *HBD* and *HBB*) and validated genetic modifiers (*KLF1*, *BCL11A* and *MYB*). Notably, although our target region was much larger than a previously reported NGS region that covers only three globin genes (*HBA1*, *HBA2* and *HBB*; He et al., 2017), the panel we designed did not include all potentially relevant genes that might affect the phenotype. However, the aim of this study was to demonstrate a practical and accurate methodology to detect variants in known genes related to hemoglobinopathies in clinical diagnosis, and not to search for novel modifier genes or sites.

Regarding genetic epidemiological investigation, we found a relatively high carrier frequency of 23.93% among individuals from the five provinces sampled (Table 4), which was higher than had been previously reported (Xiong et al., 2010). The increased carrier frequency was because we performed a thorough molecular analysis of each individual recruited for epidemiological investigation, whereas in previous studies, molecular analysis was performed only in selected individuals with positive phenotypes. The coverage of disease-causing loci was also more comprehensive in the NGS assay than the traditional genotyping assay. In addition, this is the first systematic report of molecular epidemiological data of hemoglobinopathies in Guizhou Province. Our updated epidemiological data may make local governments more aware of the severity of this disorder and more public funds might be allocated to prevention programs.

In the aspect of public prevention the birth of affected fetus, as described in Fig. 3 and Table 3, the traditional screening/molecular testing methods, which failed to detect 35 at-risk couples but identified by our NGS method due to additional disease-causing mutations and/or modifier mutations, may confer small risk for thalassemias in mutation-positive couples. Our assay was designed to capture key regions, such as α- and β-globin gene clusters, associated with hemoglobinopathies used for deep sequencing, which ensures the detection of any disease-conferring sequence changes in carrier screening. Unlike traditional carrier screening assays, which are designed to search for only the most common mutations within a gene, owing to cost considerations (Hallam et al., 2014; Xiong et al., 2010), the NGS-based approach identified both common and rare, annotated and novel variants in carriers with and without thalassemic trait phenotypes, significantly improving the detection of carrier status and therefore improving the detection rate of at-risk couples.

α-Thalassemia mutations affect up to 5% of the global population and Hb H disease, caused by the loss or inactivation of three of the four functional α-globin genes, is the most common form of symptomatic α-thalassemia (Piel and Weatherall, 2014; Vichinsky, 2013; Williams and Weatherall, 2012). Of the 10,111 couples screened, 104 (1%) were found to be at risk of having a baby with Hb H disease, and nearly 1/3 of these (31/104) had non-deletional Hb H disease that would require prenatal diagnosis. One of the many aims of this study was to further evaluate the outcomes in couples who were identified as being at-risk for having children with Hb H disease, thus potentially revealing critical clinical information for their potential offspring and helping clinicians provide precise genetic counseling after careful evaluation of all the known disease-causing mutations and modifier variants. The broad phenotypic diversity and high genetic heterogeneity of patients with Hb H disease (Lal et al., 2011;Piel and Weatherall, 2014; Vichinsky, 2013) may make such an assay necessary. In addition, our results showed that the frequency of alpha triplication (ααα$^{anti3.7}$ and ααα$^{anti4.2}$) was 1.67% (Table 4). Alpha triplication is not a pathogenic mutation, but the heterozygote of β thalassemia is asymptomatic, whereas the heterozygote of β thalassemia combined with alpha triplication results in a thalassemia intermedia phenotype. However, considering many couples' eagerness to have offspring, ethical challenges remain regarding fetuses with Hb H disease or thalassemia intermedia, which is associated with late disease onset and possibly improved survival.

Table 4
The results of a survey of prevalence of hemoglobinopathies among 20,222 people in five provinces in southern China.

Variants	Guangdong		Guangxi		Guizhou		Hainan		Yunnan		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
α-thalassemia	**587**	**12.70**	**924**	**19.11**	**378**	**9.26**	**1225**	**45.04**	**277**	**6.99**	**3391**	**16.77**
(- -^{SEA}/) deletion	290	6.27	347	7.18	159	3.90	98	3.60	80	2.02	974	4.82
(- -^{THAI}/) deletion	4	0.09	5	0.10	1	0.02	0	0.00	1	0.03	11	0.05
(- -$^{30.8}$/) deletion*	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
(-α$^{3.7}$/) deletion	151	3.27	247	5.11	120	2.94	420	1.44	138	3.48	1076	5.32
(-α$^{4.2}$/) deletion	64	1.38	108	2.23	41	1.00	419	15.40	21	0.53	653	3.23
(-α$^{27.6}$/) deletion	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
(-α$^{21.9}$/) deletion	0	0.00	2	0.04	0	0.00	0	0.00	0	0.00	2	0.01
(-α$^{2.7}$/) deletion	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
(-α$^{2.4}$/) deletion	0	0.00	3	0.06	0	0.00	0	0.00	0	0.00	3	0.01
(-α$^{1.2}$/) deletion*	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
HBA2:c.369C> (Hb Westmead)	31	0.67	125	2.59	12	0.29	239	8.79	18	0.45	425	2.10
HBA2:c.427T>C (Hb CS)	23	0.50	65	1.34	35	0.86	6	0.22	16	0.40	145	0.72
HBA2:c.377T>C (Hb QS)	12	0.26	13	0.27	7	0.17	40	1.47	0	0.00	72	0.36
HBA1:c.223G>C with -α4.2 (Hb Q-Thailand)	4	0.09	2	0.00	0	0.00	1	0.04	0	0.00	7	0.03
HBA2:c.40G>T (Hb Binyang)	0	0.00	0	0.00	2	0.05	0	0.00	1	0.03	3	0.01
HBA1:c.353_355dupTCA (Hb Phnom Penh)	3	0.06	1	0.02	0	0.00	1	0.04	0	0.00	5	0.02
HBA2:c.178G>C (Hb Zurich-Albisrieden)	2	0.04	0	0.00	0	0.00	0	0.00	0	0.00	2	0.01
HBA1:c.99G>A (Hb Amsterdam)	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
HBA1: c.1A>G	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
HBA1: c.2T>A	0	0.00	0	0.00	0	0.00	1	0.04	0	0.00	1	0.005
HBA1: c.2delT	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBA1: c.95 + 1G>A	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBA1: c.96-1G>C*	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
HBA1: c.96-2A>G	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	1	0.005
HBA1:c.223G>C	2	0.04	0	0.00	0	0.00	0	0.00	0	0.00	2	0.01
β-thalassemia	**194**	**4.20**	**322**	**6.66**	**189**	**4.63**	**139**	**5.11**	**107**	**2.70**	**951**	**4.70**
HBB: c.124_127delTTCT	68	1.47	135	2.79	95	2.33	96	3.53	27	0.68	421	2.08
HBB: c.52A>T	15	0.32	85	1.76	55	1.35	2	0.07	22	0.55	179	0.89
HBB: c.316−197C>T	46	1.00	19	0.39	15	0.7	5	0.18	5	0.13	90	0.45
HBB: c.−78A>G	31	0.67	28	0.58	9	0.22	12	0.44	7	0.18	87	0.43
HBB: c.216_217insA	3	0.06	6	0.12	0	0.00	2	0.07	3	0.08	14	0.07
HBB: c.79G>A (Hb E)	7	0.15	15	0.31	8	0.20	0	0.00	29	0.73	59	0.29
HBB: c.92 + 1G>T	1	0.02	7	0.14	0	0.00	0	0.00	2	0.05	10	0.05
HBB: c.−79A>G	0	0.00	2	0.04	1	0.02	2	0.07	0	0.00	5	0.02
HBB: c.130G>T	0	0.00	3	0.06	2	0.05	1	0.04	0	0.00	6	0.03
HBB: c.84_85insC	1	0.02	2	0.04	1	0.02	0	0.00	1	0.03	5	0.02
HBB: c.315 + 5G>C	0	0.00	5	0.10	0	0.00	0	0.00	0	0.00	5	0.2
HBB: c.−100G>A	13	0.28	14	0.29	2	0.05	19	0.70	6	0.15	54	0.27
HBB: c.−140C>T	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
HBB: c.113G>A	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	1	0.005
HBB: c.45_46insG	2	0.04	0	0.00	0	0.00	0	0.00	0	0.00	2	0.01
HBB: c.*110T>C	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBB: c.-11_-8delAAAC	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
HBB: c.304G>C (Hb Rush)	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBB: c.316-3C>T	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBB: c.316-90A>G	2	0.04	0	0.00	0	0.00	0	0.00	1	0.03	3	0.01
Chinese del	3	0.06	0	0.00	0	0.00	0	0.00	0	0.00	3	0.01
SEA-HPFH	1	0.02	0	0.00	0	0.00	0	0.00	1	0.03	2	0.01
β-thalassemia modifier	**11**	**0.24**	**42**	**0.87**	**3**	**0.07**	**19**	**0.70**	**12**	**0.30**	**87**	**0.43**
KLF1	11	0.24	42	0.87	3	0.07	19	0.70	12	0.30	87	0.43
Structural hemoglobin variants	**13**	**0.28**	**19**	**0.39**	**12**	**0.29**	**13**	**0.48**	**10**	**0.25**	**67**	**0.33**
HBA1: c.84G>T (Hb Hekinan)	4	0.09	6	0.12	4	0.10	8	0.29	2	0.05	24	0.12
HBB: c.68A>C (Hb G-Coushatta)	0	0.00	0	0.00	0	0.00	0	0.00	2	0.05	2	0.01
HBB: c.170G>A (Hb J-Bangkok)	2	0.04	0	0.00	1	0.02	1	0.04	0	0.00	4	0.02
HBB: c.341T>A (Hb New York)	5	0.11	12	0.25	6	0.15	1	0.04	5	0.13	29	0.14
HBB: c.34G>A (Hb Hamilton)	0	0.00	0	0.00	0	0.00	3	0.11	0	0.00	3	0.01
HBB: c.352C>T (Hb Tsukumi)	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	1	0.005
HBB: c.4G>T (Hb Niigata)	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
HBB: c.328G>A (Hb San Diego)	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
HBB:c.265C>G(Hb Oofuna)	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
HBB: c.41C>T (New abnormal Hb)*	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
CNV variants	**94**	**2.03**	**62**	**1.28**	**74**	**1.81**	**25**	**0.92**	**89**	**2.25**	**344**	**1.70**
ααα$^{anti3.7}$/	45	0.97	31	0.64	17	0.42	14	0.51	37	0.93	144	0.71
ααα$^{anti4.2}$/	47	1.02	30	0.62	56	1.37	10	0.37	51	1.29	194	0.96
αααα$^{69.4}$/*	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.005
αααα$^{20.9}$/*	0	0.00	0	0.00	0	0.00	1	0.04	0	0.00	1	0.005
αααα$^{121.2}$/*	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
HKαα/	0	0.00	1	0.02	0	0.00	0	0.00	0	0.00	1	0.005
β 67.8k dup*	1	0.02	0	0.00	0	0.00	0	0.00	0	0.00	1	0.005
β 204k dup*	0	0.00	0	0.00	1	0.02	0	0.00	0	0.00	1	0.005
Total											**4840**	**23.93**

20,222 samples: 4622 from Guangdong Province, 4834 from Guangxi Province, 4082 from Guizhou Province, 2720 from Hainan Province and 3964 from Yunnan Province N, number of chromosomes. *New variants identified in this study.
A total of 4840 α- and β- and KLF1 variant chromosomes (3757 α-variants, 996 β-variants and 87 KLF1 variants) were detected, corresponding to 4180 samples (3114 carried α-variants, 758 carried β-variants, 51 carried KLF1 variants, 222 carried both α- and β- variants, 22 carried both α- and *KLF1* variants, 8 carried both β- and KLF1 variants and 5 carried α-, β- and *KLF1* variants). Of the 4180 samples, 3565 carried only one type of variant chromosome (2758 α-, 757 β- and 50 KLF1), 570 carried two types of variant chromosome (356 α + α, 187 α + β, 1 β + β, 17 α + *KLF1*, 8 β + *KLF1* and 1 *KLF1* + *KLF1*), and 45 carried three types of variant chromosome (33 α + α + β, 2 α + β + β, 5 α + α + *KLF1* and 5 α + β + *KLF1*).

Considering the broad mutation spectrum and high prevalence of hemoglobinopathy-causing mutations across ethnic groups (Higgs et al., 2012; Weatherall, 2010), our technology may make it technically feasible to offer more effective preconception screening and diagnosis in large populations worldwide. Additional key challenges must be addressed to make this technique applicable to routine population screening, including introducing strict management policies, decreasing running costs, and improving quality control and protocol standardization such as the parameter settings and optimization of the platform. In China, a series of regulatory requirements for NGS that were recently established by the China Food and Drug Administration (CFDA) and the National Health and Family Planning Commission (NHFPC; Zhang and Li, 2017) must be followed. The major factors limiting the application of NGS are the expensive instruments, relatively high reagent cost and additional specific bioinformatics technicians. However, if applied to a large-scale population screen, the NGS strategy, compared with the traditional strategy, has a substantial advantage in time cost. In this study, the NGS had a throughput of 3000 samples per run. With the introduction of automated instruments for the entire workflow, the total run time for 3000 samples was 178 h, and the cost per sample was approximately $30. The cost of the test could be expected to decrease by half after approximately two years with the introduction of locally produced instruments and reagents. We believe that with continuous improvements in equipment and available methods, a comprehensive panel for hemoglobinopathies could be developed in the next few years. Furthermore, to make the NGS assay achievable for routine practice in developing countries, some conditions also need improvement, such as policy and financial support from local governments, training of skilled technicians and infrastructure construction.

Standards and guidelines for preconception population carrier screening with NGS technology have been established for several common monogenic disorders (Hallam et al., 2014; Richards et al., 2015). The comprehensive mutation spectrum of hemoglobinopathies and the well-established quality control procedures for variant detection by target-based NGS allowed us to develop an alternative to traditional molecular testing for carrier screening. Recently, efforts have been made to establish an NGS-based workflow for the identification of disease-causing mutations in Mendelian disorders (Umbarger et al., 2014). Our study shows that NGS-based diagnostic testing can achieve rapid, multiplex and high-throughput detection of genetic variants. Here, we used hemoglobinopathies as a model, because they are the most common human monogenic diseases and are associated with multiple mutations in disease-causing genes as well as modifier genes. Advances in NGS platforms and improved ease of operation should facilitate the mechanization and automatization of this approach, which might then serve as a solution for large-scale population-based carrier screening and an effective means of controlling severe hemoglobinopathies.

Funding Source

This study was supported by research funding from the National Natural Science Foundation of China (NSFC) (U1201222 and 31671314 to Dr. Xu), the NSFC (U1401221 to Dr. Shang), Science and Technology Program of Guangzhou (201604020045 to Dr. Xu) and Natural Science Foundation of Guangdong Province (2017A030313673 to Dr. Shang).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceived and designed the experiments: XS, ZP, HJ, HY, JW, YY, MM and XX. Performed the experiments: XS, XW, DC, FG, YL, XZ, HH, FZ, NL, ZL, QZ, YZ, WZ and FX. Analyzed the data: YY, A, SZ, QZ, HL, JG, YY, and LZ. Collected the samples: XZ, YC, BZ, WC, SC, RC, XG, CZ, YZ, SH, YL, BC, SY, YC, HD, XY, LW, JH, DH, SH, TY, BAZ, WKWJ, SH, XF, YZ, BAZ and MQ. Wrote the paper: XS, ZP, YY, SA, MQ, YY, MM and XX.

Acknowledgments

We thank the patients for their willingness to participate in this study; Jiajie Pu, Huilin Xu, Yi Cheng, Jianmei Zhong, Fei He, Jun Xiong, Yihong Li, Renhua Wu, Bo Wang, Hongyun Zhang, Chunna Fan, Huiqian Du, Yaling Wang, Ruoyu Wang, Yaoshen Wang, Kaixin Yang, Yaping Zhu and Wenwei Zhong et al. for assistance in collecting samples, doing experiment and data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ebiom.2017.08.015.

References

Adey, A., Morrison, H.G., Asan, Xun X., Kitzman, J.O., Turner, E.H., et al., 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11, R119.
Ahmed, S., Saleem, M., Modell, B., Petrou, M., 2002. Screening extended families for genetic hemoglobin disorders in Pakistan. N. Engl. J. Med. 347, 1162–1168.
Cao, A., Kan, Y.W., 2013. The prevention of thalassemia. Cold Spring Harb. Perspect. Med. 3, a011775.
Chong, J.X., Ouwenga, R., Anderson, R.L., Waggoner, D.J., Ober, C., 2012. A population-based study of autosomal-recessive disease-causing mutations in a founder population. Am. J. Hum. Genet. 91, 608–620.
Danjou, F., Francavilla, M., Anni, F., Satta, S., Demartis, F.R., Perseu, L., et al., 2015. A genetic score for the prediction of beta-thalassemia severity. Haematologica 100, 452–457.
von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., Vandenbroucke, J.P., Initiative, S.T.R.O.B.E., 2007. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370, 1453–1457.
Galanello, R., Sanna, S., Perseu, L., Sollaino, M.C., Satta, S., Lai, M.E., et al., 2009. Amelioration of Sardinian beta 0 thalassemia by genetic modifiers. Blood 114, 3935–3937.
Giardine, B., Borg, J., Higgs, D.R., Peterson, K.R., Philipsen, S., Maglott, D., et al., 2011. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat. Genet. 43, 295–301.
Hallam, S., Nelson, H., Greger, V., Perreault-Micale, C., Davie, J., Faulkner, N., et al., 2014. Validation for clinical use of, and initial clinical experience with, a novel approach to population-based carrier screening using high-throughput, next-generation DNA sequencing. J. Mol. Diagn. 16, 180–189.
Haque, I.S., Lazarin, G.A., Kang, H.P., Evans, E.A., Goldberg, J.D., Wapner, R.J., 2016. Modeled fetal risk of genetic diseases identified by expanded carrier screening. JAMA 316, 734–742.
HbVar database for human hemoglobin variants and thalassemia mutations. http://globin.bx.psu.edu/hbvar/menu.html bx.psu.edu.
He, J., Song, W., Yang, J., Lu, S., Yuan, Y., Guo, J., et al., 2017. Next-generation sequencing improves thalassemia carrier screening among premarital adults in a high prevalence population: the Dai nationality, China. Genet. Med. http://dx.doi.org/10.1038/gim.2016.218.
Higgs, D.R., Engel, J.D., Stamatoyannopoulos, G., 2012. Thalassaemia. Lancet 379, 373–383.
Korf, B.R., Rehm, H.L., 2013. New approaches to molecular diagnosis. JAMA 309, 1511–1521.
Lal, A., Goldrich, M.L., Haines, D.A., Azimi, M., Singer, S.T., Vichinsky, E.P., 2011. Heterogeneity of hemoglobin H disease in childhood. N. Engl. J. Med. 364, 710–718.
Liu, D., Zhang, X., Yu, L., Cai, R., Ma, X., Zheng, C., et al., 2014. KLF1 mutations are relatively more common in a thalassemia endemic region and ameliorate the severity of beta-thalassemia. Blood 124, 803–811.
MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., et al., 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469–476.
Marx, V., 2013. Biology: the big challenges of big data. Nature 498, 255–260.
Perkins, A., Xu, X., Higgs, D.R., Patrinos, G.P., Arnaud, L., Bieker, J.J., et al., 2016. Kruppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. Blood 127, 1856–1862.
Picelli, S., Bjorklund, A.K., Reinius, B., Sagasser, S., Winberg, G., Sandberg, R., 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 24, 2033–2040.
Piel, F.B., Weatherall, D.J., 2014. The alpha-thalassemias. N. Engl. J. Med. 371, 1908–1916.
Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., et al., 2008. ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. Genet. Med. 10, 294–300.
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al., 2015. 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. 17, 405–424.

Sankaran, V.G., Weiss, M.J., 2015. Anemia: progress in molecular mechanisms and therapies. Nat. Med. 21, 221–230.
Stark, Z., Tan, T.Y., Chong, B., Brett, G.R., Yap, P., Walsh, M., et al., 2016. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet. Med. 18, 1090–1096.
Steinberg, M.H., Forget, B.G., Higgs, D.R., Weatherall, D.J., 2009. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management. Second Edition. Cambridge University Press.
Thein, S.L., 2013. Genetic association studies in beta-hemoglobinopathies. Hematol. Am. Soc. Hematol. Educ. Program 2013, 354–361.
Thuret, I., Pondarre, C., Loundou, A., et al., 2010. Complications and treatment of patients with beta-thalassemia in France: results of the National Registry. Haematologica 95, 724–729.
Traeger-Synodinos, J., Harteveld, C.L., Old, J.M., Petrou, M., Galanello, R., Giordano, P., et al., 2015. EMQN Best Practice Guidelines for molecular and haematology methods for carrier identification and prenatal diagnosis of the haemoglobinopathies. Eur. J. Hum. Genet. 23, 560.
Umbarger, M.A., Kennedy, C.J., Saunders, P., Breton, B., Chennagiri, N., Emhoff, J., et al., 2014. Next-generation carrier screening. Genet. Med. 16, 132–140.
Vichinsky, E.P., 2013. Clinical manifestations of alpha-thalassemia. Cold Spring Harb. Perspect. Med. 3, a011742.
Weatherall, D.J., 2010. The importance of micromapping the gene frequencies for the common inherited disorders of haemoglobin. Br. J. Haematol. 149, 635–637.
Williams, T.N., Weatherall, D.J., 2012. World distribution, population genetics, and health burden of the hemoglobinopathies. Cold Spring Harb. Perspect. Med. 2, a011692.
Xiong, F., Sun, M., Zhang, X., Cai, R., Zhou, Y., Lou, J., et al., 2010. Molecular epidemiological survey of haemoglobinopathies in the Guangxi Zhuang Autonomous Region of southern China. Clin. Genet. 78, 139–148.
Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., et al., 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. 369, 1502–1511.
Zhang, R., Li, J.M., 2017. China's policies regarding next-generation sequencing diagnostic tests. Sci. Transl. Med. 9–11 *Suppl issue*.