NGS4THAL, a one-stop molecular diagnosis and carrier screening tool for thalassemia and other hemoglobinopathies by next-generation sequencing

Yujie Cao

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Shau Yin Ha

Department of Pathology, Hong Kong Children's Hospital, Hong Kong S.A.R., China

Chi-Chiu So

Department of Pathology, Hong Kong Children's Hospital, Hong Kong S.A.R., China

Tong Ming For

Clinical Genetic Service, Department of Health, The Government of the Hong Kong S.A.R., China

Clara Sze-Man Tang

Department of Surgery, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Huoru Zhang

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Rui Liang

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Jing Yang

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Brian Hon-Yin Chung

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Godfrey Chi-Fung Chan

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Yu-Lung Lau

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Maria-Mercè Garcia-Barcelo

Department of Surgery, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Edmond Shiu-Kwan Ma

Department of Pathology, Hong Kong Sanatorium & Hospital

Pranee Sucharitchan

Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Nattiya Hirankarn

Center of Excellence in Immunology and Immune Mediated Diseases, Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Wanling Yang (✉ yangwl@hku.hk)

Department of Paediatrics and Adolescent Medicine, Queen Mary Hospital, The University of Hong Kong, Hong Kong S.A.R., China

Method Article

Keywords: thalassemia, hemoglobinopathies, sickle cell disease, NGS, bioinformatics pipeline, molecular diagnosis, carrier screening

Posted Date: May 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-542196/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Thalassemia is one of the most common genetic diseases and a major health threat worldwide. Accurate, efficient and scalable genetic testing methodology is much needed for its molecular diagnosis and carrier screening.

Results We developed NGS4THAL, a bioinformatics analysis pipeline analyzing next generation sequencing (NGS) data to detect pathogenic variants for thalassemia and other hemoglobinopathies. NGS4THAL recovers and realigns ambiguously mapped NGS reads derived from the homologous hemoglobin gene clusters to achieve accurate detection of point mutations and small insertion/deletions (InDels). And it uses several structural variant (SV) detection tools with complementary algorithms, and an inhouse database with control data on a number of known SVs to achieve accurate detection of hemoglobin SVs. Detected variants are matched with those in HbVar, allowing recognition of known pathogenic variants, including disease modifiers. Tested on simulation data, NGS4THAL achieved high sensitivity and specificity. For targeted NGS sequencing data from samples with laboratory-confirmed pathogenic hemoglobin variants, it achieved 100% detection accuracy. Application of NGS4THAL on whole genome sequencing data from unrelated studies detected thalassemia mutation carrier rates for Hong Kong Chinese and Northern Vietnamese that were consistent with those from epidemiological studies.

Conclusions NGS4THAL is a highly accurate and efficient molecular diagnosis tool for thalassemia and other hemoglobinopathies based on tailored analysis of NGS data, and is potentially scalable for carrier screening purposes.

Background

Thalassemia is one of the most common genetic diseases, with very high prevalence and carrier rates in many tropical and subtropical regions(1, 2). It is a group of heterogeneous, mostly recessively inherited disorders with a broad spectrum of mutations that often differs among geographical regions and ethnical groups(3). More than 400 thalassemia pathogenic variants have been documented in HbVar(4) up to date, which is a comprehensive database on pathogenic and benign hemoglobin variants.

Deletions in the α -globin gene locus are the underlying causes of most of the α -thalassemia (5), while point mutations and small insertions/deletions (InDels) in the β -globin gene region explain most β -thalassemia(6). Although most of the pathological mutations involve the coding regions of the hemoglobin genes, there are pathogenic variants involving the regulatory elements in intergenic regions or introns, affecting expression or splicing of the hemoglobin genes(7). In addition, some disease modifiers(8-10), both in the hemoglobin regions and in other regions of the genome, impact on disease severity of β -thalassemia and sickle cell disease (MIM: 603903), mainly through regulation of the fetal hemoglobin levels.

Traditionally, molecular diagnosis of thalassemia (as well as other hemoglobinopathies) starts with hematological tests and hemoglobin electrophoresis, followed by genetic tests to identify the exact pathogenic variant(s)(11, 12). This requires cumbersome laboratory work and step-by-step decision-making based on the circumstances, which makes it difficult to scale up. Certain disease mutations also may not show positive hematological features and might be missed by hematological screening(12). Thus, an accurate, efficient, and universally applicable thalassemia and hemoglobin variant detection platform is greatly needed in order to screen relevant globin gene variants and disease modifiers for this group of diseases without having to consider specific circumstances.

Next generation sequencing (NGS) technology provides a possibility of detecting almost all the pathological variants and disease modifiers in a single experimental setting and has been used to various degrees in the molecular diagnosis of thalassemia and other hemoglobinopathies(13-15). However, none of these attempts have resulted in an accurate, efficient, and scalable solution for globin gene variant detection, and as a result, NGS technology has not been widely applied in clinical laboratories and population carrier screening for hemoglobin diseases up to date.

The α -globin gene clusters are located in 16p13.3, a region features with homologous sequences, interspersed repeats, and various types of genetic variants(16). Short NGS reads derived from this region are often ambiguously mapped to multiple positions due to sequence homology. This not only causes missed detections of point mutations and InDels, but also affects read depth estimation, which is one of the features required for detection of structural variants (SVs). Furthermore, compound forms of hemoglobin SVs are common in thalassemia epidemic regions. For example, $-\text{SEA}/-\alpha^{3.7}$, a compound heterozygous genotype that causes hemoglobin H (HbH) disease, is common in Hong Kong. Being able to detect complex forms of genotypes is essential for analysis tools aiming at comprehensive thalassemia molecular diagnosis and carrier screening.

Here, we present NGS4THAL, a one-stop bioinformatics analysis pipeline aiming at detecting hemoglobin variants through analyzing various forms of NGS data. NGS4THAL recovers and realigns NGS reads with multiple alignments in the hemoglobin regions for accurate detection of point mutations and small InDels. It uses SV callers with complementary algorithms, tailored parameters, and an in-house database of relevant controls of specific SVs to improve SV detection. Validation of NGS4THAL on simulation data and targeted sequencing data on samples with globin gene variants previously confirmed from clinical laboratories achieved high sensitivity and specificity. It can also detect thalassemia pathogenic variants located in regulatory regions, as well as known disease modifiers. NGS4THAL is implemented in python and can be accessed from https://github.com/JavenCao/Realign_BAM.

Implementation

Detection of point mutations and small InDels

A schematic illustration of NGS4THAL is presented in **Figure 1**. NGS4THAL tries to improve the detection of SNVs and small InDels through recovery and realignment of NGS sequencing reads with multiple alignments (RMAs), which are common in the two hemoglobin genomic regions, namely 16p13.3 and 11p15.4.

NGS4THAL firstly recovers informative RMAs in the hemoglobin regions from the alignment file in BAM format. The following criteria are applied for recovering a RMA: (1) it maps to one of the homologous regions, namely *HBZ-HBZP1*, *HBA2-HBA1-HBAP1* or *HBG1-HBG2*; and (2) it has a bwa(17)-based mapping quality score (MAQ) of zero due to having an alternative mapping position with an identical mapping profile as that of the primary mapping position, or (3) it has a MAQ < 30 and the difference of mismatches between the primary mapping hit and the alternative mapping hit is less than three base pairs under the short-read settings (*e.g.* 100bps - 300bps).

After recovering these reads, one read of the read pair, designated as R1, is realigned to either the primary mapping position or the alternative mapping position at random. The mate read, designated as R2, is mapped nearby, within five standard deviation of the mean insert size of the NGS library from R1. The MAQs of these realigned RAMs are adjusted to 60 in a new BAM file for downstream analysis. SNVs/InDels are detected from the new BAM file by GATK-HaplotypeCaller(18).

NGS4THAL was designed to detect point mutations and small InDels in globin gene clusters, regulatory elements, and certain thalassemia disease modifiers known to date (**Additional file 3: Table S1**) and this is accomplished by looking the detected variants up in the HbVar database. The detected variants are classified into known pathogenic SNVs/InDels, benign Hb variants, or novel variants, with the latter also being reported for manual annotation. Variants located in equivalent positions within the homologous genes, such as between *HBA2* and *HBA1* (*e.g.* *HBA2:c.64G>C* or *HBA1:c.64G>C*), cannot be distinguished based on available short reads NGS data. They will be reported as such for annotation and further test using traditional laboratory methods. Additional technical details can be found in **Note 1 in additional file 1**.

Detection of hemoglobin structural variants

Depending on the relative positions between the breakpoints of the SVs and the homologous sequences, we classified hemoglobin SVs into four subtypes: Subtype-A SVs, whose breakpoints are both located in non-homologous regions; Subtype-B and Subtype-C SVs, for which one of the two breakpoints is located in a homologous region; and Subtype-D SVs, whose breakpoints are both located in homologous regions. NGS4THAL uses complementary tools analyzing the mapped NGS reads for SV detection, based on features of insert size estimate, mapping positions of fragments upon splitting a read, and read-depth, respectively (**Additional file 2: Figure S3**).

BreakDancer⁽¹⁹⁾ relies on mapping information of respective reads from a read pair, while Pindel⁽²⁰⁾ uses the respective mapping positions upon splitting a single NGS read into fragments. CoNIFER⁽²¹⁾ makes

use of read depth information to infer CNVs (Copy Number Variants), initially developed for whole exome sequencing (WES) data. Some modifications were adopted to tailor the analysis process to better fit our purpose. Briefly, requirement on minimum alternative MAQ was lowered to 0 ($-q\ 0$), allowing BreakDancer to take RMAs into consideration to better detect SVs located in homologous sequences. An SV may cause insert size deviation from the norm of the library as well as mapping location aberrations when reads are split and respectively mapped. BreakDancer and Pindel are complementary to each other and combined use of the two may increase sensitivity⁽²⁰⁾. Thus, we included results from BreakDancer as a source of input for Pindel, which is a practice that has been shown to increase SV detection accuracy⁽²⁰⁾.

To make CoNIFER suited for analyzing sequencing data with continuous coverage, a set of probes covering the two hemoglobin loci were generated *in-silico*. Each probe is designed at 80 bps in length, with 25 bp overlap between adjacent probes. This modification allows the software to have more detailed and fine-tuned SV detection. To increase sensitivity, the detection threshold was also lowered from 1.0 to 0.5, effectively allowing detection of CNVs with 50% read-depth change relative to the baseline average, instead of the default threshold of 100% change. Finally, variants detected from a comprehensive use of BreakDancer, Pindel and CoNIFER were matched with those in the HbVar database. SVs matching the known hemoglobin SVs in the database were reported as such, while candidate novel SVs were reported for further manual annotation for their functional implications. More technical details of the algorithm can be found in **Note 1 in additional file 1**.

In addition, a stage-wise process was implemented in NGS4THAL in order to detect compound SVs (such as $-^{SEA}/-\alpha^{3.7}$), which is a difficult task for any single algorithm. The principle adopted was to fine tune the background baseline to make it in line with one type of the compound SVs, so that the coexistence of the SV from the other haplotype would be easier to detect. Specifically, in the screening stage, all the samples are analyzed using BreakDancer, Pindel and CoNIFER to determine the presence of SVs/CNVs. Then in the fine-profiling stage, samples with suspected large SVs covering both the *HBA2* and *HBA1* genes (such as $-^{SEA}$) are further tested for the possibility of compound heterozygous SVs (such as co-existence of $-\alpha^{3.7}$ in the form of $-^{SEA}/-\alpha^{3.7}$).

Briefly, in the fine-profiling stage, samples are grouped by the type of large SVs they are suspected to carry, for example, samples suspected of carrying $-^{SEA}$ are grouped altogether. The read-depth for HBA genes is re-examined by CoNIFER for each group of samples, with reference to the baselines estimated by the controls carrying the large SVs (such as $-^{SEA}$). NGS4THAL incorporated an in-house sample database with NGS sequencing data on wildtype and pre-defined heterozygous SVs of various types, from which control samples of a particular heterozygous SV can be drawn to be analyzed together with the samples in question. The sample size is set at 30 in minimum in both the screening-stage and the fine-profiling stage, ensuring accurate estimate of read depth for SV detection.

Datasets used to evaluate NGS4THAL

1. Simulated data

Point mutations and small InDels were simulated to cover *HBA2*, *HBA1*, *HBB*, *HBD* genes, on both exons and introns. The simulated SVs included the aforementioned hemoglobin SV subtypes, as well as some complex combinations - compound hemoglobin SVs. A total of 69 point mutations (including 66 thalassemia pathogenic variants, three α -globin variants causing other hemoglobinopathies), 19 small InDels, and 25 SVs spanning *HBA2* and *HBA1* genes (including 16 deletions, 3 duplications, and 6 compound hemoglobin SVs) were simulated (**Additional file 3: Table S2**).

To test the robustness of NGS4THAL, we simulated ten independent heterozygous carriers for each simulated variant. We also generated ten cases with homozygous $\alpha^{3.7}$. An additional 870 independent individuals without any polymorphisms were simulated as negative controls, summing up to a simulated cohort of 2,000 samples. VarSim(22) and ART(23) were used sequentially to generate NGS reads with continuous coverage of the HBA (hg19: chr16:132,744-279,999, size = ~147kbps) and the HBB regions (hg19: chr11:5,221,230-5,321,230, size = 100kbps). The simulated NGS reads were similar to those generated by Illumina HiSeq 2500 platform in terms of base pair quality scores. Paired-end reads with read length of 150bps were produced at around 100-fold coverage depth.

2. Targeted sequencing data on samples with clinically confirmed mutations

Seventy-one patient samples with hemoglobin variants causal to thalassemia or other hemoglobinopathies, confirmed in clinical laboratory settings, were kindly provided by the Department of Pathology in Queen Mary Hospital and Clinical Genetic Service, Department of Health, Hong Kong, and Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. The confirmed hemoglobin variants included Hb CS, Hb QS, Hb Luton, HbQ-Thailand, Hb E, -28 (A>G), CD 41/42 (-CTTT), IVS I-5(G>C), IVS II-654 C>T, IVS I-1(G>A), CD 17 (AAG>TAG), Hb Malay, Hb Hoshida, CD 35 TAC>TAA, CD71/72 (+A), $-\text{SEA}$, $-\alpha^{3.7}$, $-\alpha^{4.2}$, and compound hemoglobin SVs, such as $-\text{SEA}/-\alpha^{4.2}$, $-\text{SEA}/-\alpha^{3.7}$ and $-\text{SEA}/\alpha\alpha^{\text{anti}4.2}$.

Enrichment probes were designed by online tools from Roche NimbleDesign. To maximize coverage and continuous capture of sequences from the two hemoglobin gene loci, including the regulatory elements, the designed probes were allowed to have no more than three matching positions in human genome. Besides, a number of genes and polymorphic sites known to modify disease severity were also included (**Note 2 in additional file 1**). The probes were manufactured by Roche NimbleGen. NGS library preparation, including fragment enrichment, were prepared based on SeqCap EZ HyperCap Workflow Version 1.2. Paired-end sequencing of 100bps read length was generated on Illumina HiSeq 1500, performed by the Centre of Genomic Sciences in the University of Hong Kong. Around 1500-fold coverage depth was achieved for these samples.

3. Whole exome sequencing (WES) and whole genome sequencing (WGS) datasets

Whole exome sequencing (WES) data on 185 individuals were provided by the Asian Primary Immunodeficiency (PID) Network(24) (referred to as Hong Kong WES cohort below). Exomes were

captured by probes from various sources, including Agilent SureSelect, Illumina TruSeq and xGen Exome Research Panel from Integrated DNA Technologies. Paired-end sequencing with read length of 100bps or 150bps and sequencing depth of 50-150 folds were applicable to these data.

Whole genome sequencing (WGS) data on a total of 954 samples from a Hirschsprung's disease study (25, 26) conducted in The University of Hong Kong was used to test the performance of NGS4THAL, including 447 Hirschsprung's disease cases and 507 healthy controls. The ethnical origins of the samples include Hong Kong Chinese, Mainland Chinese, and Northern Vietnamese (**Additional file 3: Table S3**). The data were 150bps in NGS read length with paired-end sequencing and an average of 30-fold coverage. Different types of the detected variants from this dataset were rendered for orthogonal validation by traditional experimental methods, including Sanger sequencing after PCR amplification and Multiplex Ligation Dependent Probe Amplification (MLPA) (**Note 3 in Additional file 1**).

Haplotype analysis

To define haplotype(s) that might be in linkage disequilibrium (LD) with the $-^{SEA}$ deletion, we used 40 heterozygotes $-^{SEA}$ samples, confirmed by NGS4THAL from the aforementioned cohorts with targeted sequencing and WGS data, together with 3, 516 wildtype individuals on the $-^{SEA}$ locus with genome-wide SNP genotyping data and WGS data, all of whom were Hong Kong Chinese in ethnic origin(27). The wildtype genotype was confirmed either by report from NGS4THAL or by heterozygosity on a high quality SNP (rs2974771) located within the range of $-^{SEA}$ deletion. These data were used as the discovery cohort for LD analysis between $-^{SEA}$ and haplotypes in this region (**Note 4 in Additional file 1**).

A replication cohort was collected from the 1000 Genomes Project with phased genotypes(28), including 21 $-^{SEA}$ heterozygotes and 381 wildtype samples from matching ethnicities (**Additional file 3: Table S6**). Recombination rate was derived from the HapMap project(29), and the LD patterns surrounding $-^{SEA}$ were based on SNP genotyping data of Hong Kong individuals, analyzed by Haploview(30). PHASE (version 2.1)(31) was used to determine the haplotypes of the genotype data and their frequencies. χ^2 test was used to calculate the association *P* values and odds ratios of associations between $-^{SEA}$ and haplotype in linkage.

Results

Detection of point mutations and small InDels

RMAs (NGS reads with multiple alignment) occur within *HBZ-HBZP1* and within *HBAP1-HBA2-HBA1* (**Additional file 2: Figure S4**) in the α -globin gene locus, and between *HBG1* and *HBG2* in the β -globin gene locus (**Additional file 2: Figure S5**). NGS4THAL identifies these RMAs and realigns the variant-informative RMAs to equally likely positions in the reference genome at random. This process was shown to significantly improve variant detection located in the homologous regions (**Additional file 2: Figure S6**).

Testing of NGS4THAL on simulation data showed that, without the process of RMA recovery and realignment, hemoglobin SNVs located in homologous regions were often missed, including variants located in exon 1 and exon 2 of *HBA1* and *HBA2* genes (**Figure 2 - A and B**). This was because NGS reads derived from these regions often failed to meet mapping quality threshold and were not contributing to the variant detection process (**Additional file 2: Figure S7**). In contrast, upon recovery and realignment of the informative RMAs by NGS4THAL, detection of variants in these regions was significantly improved. For SNVs in the simulation data, NGS4THAL achieved a 97.87% sensitivity, compared to 68.09% in the absence of NGS reads recovery and realignment process. Considering SNVs with small InDels altogether, NGS4THAL achieved a sensitivity of 98.48%, compared to a 77.27% sensitivity without the RMA realignment process.

We didn't treat variants detected on *equivalent positions* within the homologous genes (such as *HBA2*:c.64G>C or *HBA1*:c.64G>C) as false positives, as under the short reads NGS sequencing scheme (e.g., 100-300bps read length and insert size around 300-500bps), the exact position of the point mutations between *HBA2* and *HBA1* cannot be determined with certainty, particularly for exon 1 - 2 of these two genes, which share nearly 2kb sequences of complete identity. Acknowledging this caveat, a 100% specificity was achieved as pathogenic variants were concerned with the realignment process.

Testing of NGS4THAL on the 71 samples sequenced by the targeted sequencing platform, 100% sensitivity was achieved for the pathogenic hemoglobin point mutations and small InDels (including variants causing thalassemia and other hemoglobinopathies), which were previously confirmed by traditional laboratory methods (**Additional file 2: Figure S8**). Specially, Hb Q-Thailand and Hb_Luton, located in exon 1 and exon 2 of *HBA2* and *HBA1*, would have been missed without the RMA realignment process.

Applying NGS4THAL on whole genome sequencing (WGS) data from the Hirschsprung's disease study(25, 26), a number of thalassemia pathogenic point mutations and small InDels common to local populations were detected, including Hb CS, Hb Westmead, CD41/42 (-CTTT) and -28 (A>G) for Hong Kong Chinese, and HbE, CD 17 (AAAG>TAG) and -50 G>A for Northern Vietnamese. The detected carrier rates for these variants were consistent with those from previous epidemiological reports(32)·(33) (**Additional file 3: Table S7** and **Additional file 3: Table S8**). Furthermore, all the unique mutations detected above were orthogonally confirmed by PCR amplification followed by Sanger sequencing (**Additional file 3: Table S9**).

Commonly used WES capture protocols in the market can achieve complete coverage of *HBA2*, *HBA1*, *HBM* genes, including introns through off-target capture and sequencing, and coverage on the exonic regions of *HBZ*, *HBB*, *HBD*, *HBG1* and *HBG2*. Hemoglobin variants located in these regions could be readily detected by NGS4THAL. A number of locally common thalassemia pathogenic point mutations and small InDels were detected from the Hong Kong WES cohort with high confidence, including Hb Westmead (carrier frequency was 1/185=0.5%) and Hb CS (1/185=0.5%) from exon 3 of *HBA2*, Codons 41/42 (-TTCT) (2/185=1.1%), -28 (A>G) (1/185=0.5%) and IVS-II-654 (C>T) (2/185=1.1%) from *HBB*

(**Additional file 3: Table S10**). The carrier rates for these pathogenic variants detected in the WES cohort were consistent with the epidemiology reports, although this is a small sample size for accurate estimation(32). On the other hand, intronic mutations in *HBB*, *HBG1* and *HBG2*, as well as intergenic variants that may affect expression of the hemoglobin genes might have been missed by WES, highlighting the shortcomings of the methodology on coverage.

Detection of structural variants

We tested NGS4THAL on simulation data composed of the four subtypes of hemoglobin structural variants (SVs) defined earlier, including heterozygous, homozygous, and compound heterozygous SV genotypes. Subtype-A deletions (such as $-\alpha^{2.7}$), presenting with definitive characteristics from paired end sequencing, could be detected without ambiguity. These SVs present characteristics such as aberrant insert size inferred from mapping positions of paired reads, breakpoints revealed by mapping of reads upon being split, and reduced/increased coverage depth, as detected by BreakDancer, Pindel and/or CoNIFER, respectively (**Figure 3 - A**).

Subtype-B (such as $-\alpha^{5.3}$) and Subtype-C (such as $-\alpha^{3.5}$) deletions, with one of the two breakpoints located in homologous sequences, can be detected by BreakDancer when analyzing changes in insert size from paired-end reads near the breakpoints, which works well especially with RMAs being recovered and used. Using $-\alpha^{5.3}$, a Subtype-B deletion, as an example (**Figure 3 - B**), tailored software BreakDancer detected two candidate deletions: one was consistent exactly with the simulated $-\alpha^{5.3}$ and the other detection was a false, longer SV with one of the ends incorrectly determined from ambiguous mapping. For NGS4THAL, this issue of breakpoints ambiguity can be resolved by matching the results with HbVar database and the detection matching with known variants in the database was reported as the most likely variant.

For algorithms based on mapping of split-reads, such as Pindel, the 5' breakpoint could be determined by analyzing NGS reads that span the breakpoint. However, the split-fragments of these reads could also be mapped to the paralogous *HBA1* gene, causing uncertainty in defining the 3' breakpoint. In this case, NGS4THAL firstly detects all the plausible SVs supported by the data, and makes the call by matching all the candidate results with the known hemoglobin SVs from the HbVar database. In addition, simultaneously applying CoNIFER also detected Subtype-B and Subtype-C deletions through changes in read depth.

The $-\alpha^{3.7}$ is a Subtype-D deletion, caused by crossing-over between misaligned homologous X,Y,Z boxes during meiosis(34). As both breakpoints are located in homologous sequences, Subtype-D variants cannot be detected by evaluating insert size aberrations that BreakDancer is based on (**Figure 3 - D**). Similarly, they cannot be detected by split-read mapping methods such as Pindel either. Nevertheless, Subtype-D deletions could still be detected by read depth-based algorithms, such as CoNIFER when parameters are adjusted. In the case of $-\alpha^{3.7}$ (hg19: chr16:223,299-227,103), NGS4THAL detected a

deletion spanning hg19:chr16:222,879-226,799. Matching the result with all variants in HbVar database reveals that this is likely the $-\alpha^{3.7}$ deletion (**Figure 3 - D**).

Detecting compound heterozygous forms of SVs, which are not uncommon in epidemic regions for thalassemia, can be more complicated. Taking $-\text{SEA}/-\alpha^{3.7}$ as an example, we found that in the screening stage, although the $-\text{SEA}$ deletion could be identified by both BreakDancer and Pindel (**Figure 4 - A**), the loss of another copy of *HBA2* on top of $-\text{SEA}$ (e.g., $-\alpha^{3.7}$) can't be correctly determined, even by CoNIFER with modified parameters (**Figure 4 - B**).

Considering these scenarios, in the fine-profiling stage, as $-\text{SEA}$ deletion has already been detected, NGS4THAL groups these samples with a number of heterozygous $-\text{SEA}$ controls with similar sequencing depth for further evaluation by CoNIFER. In so doing, the loss of the second copy in HBA genes from the $-\text{SEA}/-\alpha^{3.7}$ genotype could be clearly detected (**Figure 4 - C**). The same principle holds true for other compound heterozygous SVs with position overlaps, such as $-\text{CANT}/-\alpha^{3.7}$ and $-\text{SPAN}/-\alpha^{3.7}$ (**Additional file 2: Figure S9**).

From simulated data, NGS4THAL achieved 98.40% sensitivity and 100% specificity for the detection of the hemoglobin SVs including deletions, duplications and compound rearrangements (**Additional file 3: Table S11**). NGS4THAL determines whether a detected SV is known in the HbVar database by checking the coordinates of the SV boundaries. Due to sequencing depth fluctuations and issues with RMAs, the boundaries of SVs determined by read-depth method are not always accurate. As a result, a few SVs were not accurately reported by NGS4THAL, accounting for the 1.6% of missed detections.

Testing NGS4THAL on the targeted sequencing data of the 71 samples with confirmed hemoglobin mutations led to fully accurate detection of the hemoglobin SVs, including $-\text{SEA}$, $-\alpha^{3.7}$ and $-\alpha^{4.2}$ (**Additional file 2: Figure S10**), as well as compound rearrangements, such as $-\text{SEA}/-\alpha^{4.2}$, $-\text{SEA}/-\alpha^{3.7}$ and $-\text{SEA}/\alpha\alpha^{\text{anti}4.2}$ (**Additional file 2: Figure S11**), without false positive detection. In addition, applying NGS4THAL on WGS data from the Hirschsprung's disease study led to detection of common hemoglobin SVs in samples from Hong Kong and North Vietnam (**Additional file 3: Table S7**), such as common deletions of $-\text{SEA}$, $-\alpha^{3.7}$ and $-\alpha^{4.2}$, (**Additional file 2: Figure S12**), and duplications of *HBA2* or *HBA1* (**Additional file 2: Figure S13**). The carrier rates of the thalassemia pathogenic deletions detected were consistent with those from epidemiology reports(32)·(35-37). A total of 19 samples with representative hemoglobin SVs, including $-\text{SEA}$, $-\alpha^{3.7}$, $-\alpha^{4.2}$ and α -region duplications were selected for verification by MLPA assay, and all were confirmed experimentally (**Additional file 3: Table S12**).

When NGS4THAL was applied to the WES data (**Additional file 3: Table S10**), a limited number of off-target NGS reads upstream of the *HBM* gene allowed the detection of the $-\text{SEA}$ deletion. These reads, which were near the $-\text{SEA}$ breakpoints, allowed BreakDancer and Pindel to make detection of three $-\text{SEA}$ heterozygotes, for a carrier frequency of 1.6%. In addition, NGS4THAL detected one $-\alpha^{3.7}$ heterozygote deletion (1/185=0.5%) with some uncertainty on the boundaries by CoNIFER. Notably, the detected rates

of $-^{SEA}$ and $-\alpha^{3.7}$ were both much lower than expected, which probably were caused by lack of adequate and continuous coverage by the WES data.

Population difference in thalassemia mutation carrier rates

Based on thalassemia screening on WGS cohorts, thalassemia mutation carrier rate was estimated to be 11.7% for Hong Kong Chinese population, with α -thalassemia mutation carrier rate estimated at 9.6% and β -thalassemia mutation carrier rate at 1.9% (**Additional file 3: Table S7**). The most common thalassemia mutations for Hong Kong Chinese population include $-^{SEA}$ deletion (4.53%), $-\alpha^{3.7}$ deletion (3.20%), $-\alpha^{4.2}$ deletion (0.53%) for α -thalassemia, and CD 41/42 (-CTTT) (1.60%) for β -thalassemia. The estimated thalassemia mutation carrier rate for Northern Vietnamese was 14.91% with high frequencies of Hb E (3.10%), $-^{SEA}$ deletion (3.70%) and $-\alpha^{3.7}$ (3.10%). Since the α -duplication may exacerbate disease severity in β -thalassemia, we also calculated carrier rates for duplications. In Hong Kong Chinese population, $\alpha\alpha/\alpha\alpha^{anti3.7}$ (2/375) and $\alpha\alpha/\alpha\alpha^{anti4.2}$ (1/375), accounting for 0.80% of the population. In the Northern Vietnamese population, the carrier rates for $\alpha\alpha/\alpha\alpha^{anti3.7}$ was estimated at 3.11% (5/161).

Thalassemia is not endemic in North China. The samples collected from North China have a detected thalassemia carrier rate of 2.7% (7/254) (**Additional file 3: Table S8**), of which 71.43% (5/7) were $-\alpha^{3.7}$. And the detected carrier rates for $\alpha\alpha/\alpha\alpha^{anti3.7}$ was 1.57% (4/254), suggesting that these mutations are likely recurrent. As expected, $-^{SEA}$ and other thalassemia mutations common to Hong Kong Chinese and Northern Vietnamese population were absent in the Northern Chinese samples. These mutations exert effect on disease severity of β -thalassemia and would have little functional impact working alone.

Impact of different sequencing platforms on thalassemia mutation detection

It was revealed that both targeted sequencing data of the hemoglobin regions with in-depth and continuous coverage and WGS with the widely adopted 30-fold coverage depth lead to mutation detections that are accurate or in line with expectations from epidemiology reports, for both point mutation/small InDels and SVs. WES with 50-150 folds sequencing depth did lead to detection of point mutations and small InDels located in well covered coding regions (**Additional file 3: Table S10**). However, WES has its intrinsic shortcomings in SV detection due to the coverage discontinuity and lack of coverage in introns and intergenic regions. Although off-targeted NGS reads might lead to identification of $-^{SEA}$ deletion, the detected rate is lower than expected. The carrier rate detected from Hong Kong WES dataset (5.9%, 11/185) is probably a reflection of missing detections of SVs and non-coding SNVs (**Additional file 3: Table S10**).

Haplotype(s) in linkage disequilibrium with the $-^{SEA}$ deletion

All the detected Δ ^{SEA} deletions from our datasets shared the same breakpoints with the coordinates of hg19:chr16: 215395-234699, consistent with its nature as a founder mutation(38). Population substructure analysis of the WGS cohorts by principal component analysis showed that Δ ^{SEA} deletions were only detected from samples genetically similar to and clustered with Southeast Asian populations, and in cases often defying self-reported geographical origins(**Additional file 2: Figure S14**).

Local LD structure showed that Δ ^{SEA} deletion located in a LD-block with recombination hotspots located immediately upstream (the telomere end of 16p13.3) and some distance downstream (**Additional file 2: Figure S15**). We selected six SNPs from this interval to define haplotypes that might be associated with the Δ ^{SEA} deletion. From the discovery cohort, a haplotype with G(rs2685118) - G(rs1203957) - A(rs3918352) - T(rs11642609) - G(rs1203974) - A(rs8045291) was found in high LD with Δ ^{SEA} (**Additional file 3: Table S13**). This haplotype was detected in all but one (39 of 40, 97.5%) Δ ^{SEA} carriers and in only 25.63% of the non Δ ^{SEA} individuals, with an estimated OR of 113.19 and a *P-value* < 2.2e-16. Using this haplotype as a surrogate for Δ ^{SEA} deletion achieves a sensitivity of 97.50% and specificity of 74.37% (**Additional file 3: Table S14**). In the replication cohort, this risk haplotype (GGATGA) was associated with Δ ^{SEA} with similar strength (OR of 114.61, *P-value* < 2.2e-16), which would achieve a sensitivity of 95.45% and a specificity of 84.51%.

Discussion

Thalassemia is a major public health burden in many countries, especially in tropical and subtropical regions. Global population migration, however, has led to the spread of the disease to historically non-epidemic regions, such as Northern Europe and the Americas(39). It has been shown that establishing nation-wide or territory-wide thalassemia screening programs can significantly reduce the incidence rates of thalassemia or other hemoglobinopathies, moving forward towards the goal of controlling these ancient diseases(40). It is commonly believed that pre-conception knowledge of carrier status, which can be achieved by population screening or screening for couples who are planning to have children, could offer initiatives to at-risk subjects to take necessary measures in avoiding having children with thalassemia. This is especially valuable in regions where termination of pregnancies is not acceptable for religious, cultural or legal reasons.

The current laboratory diagnosis methods for thalassemia starts with detection of reduced levels of mean corpuscular volume/mean corpuscular hemoglobin (MCV/MCH), which could miss certain mutations due to various reasons. For example, silent heterozygous α -thalassemia and β -thalassemia carriers, compound heterozygotes with both an α -globin gene deletion and a duplication (α -/ $\alpha\alpha$), heterozygous Hb E carriers, and double heterozygotes of β -thalassemia and α -thalassemia mutations, may not show reduction in MCV/MCH levels. Diagnosis methods based on hemoglobin electrophoresis could also be problematic, especially for some silent heterozygous mutation carriers. For example, coexistence of β -thalassemia mutations may significantly reduce HbH levels. Furthermore, downstream genetic tests, such as Gap-PCR and MLPA, which are the gold standard for the diagnosis of certain

mutations, are often cumbersome and are targeted towards certain genetic mutations. Many of these techniques are only available to some tertiary hospitals, preventing their wide-spread applications.

Although NGS sequencing of hemoglobin regions has the potential to reveal nearly the complete spectrum of genetic mutations, data analysis poses a big challenge due to sequence complexity of the hemoglobin regions. This is aggravated by the fact that bioinformatic expertise is still not widely available in clinical laboratories. Thus, it is necessary to develop an accurate, comprehensive, and efficient platform to aid the molecular diagnosis of thalassemia and other hemoglobinopathies. NGS4THAL, a NGS data analysis pipeline, aims at serving this purpose. NGS4THAL particularly deals with issues caused by homologous sequences among the hemoglobin genes and pseudogenes for accurate detection of point mutations and small InDels. Using a combination of complementary SV callers with tailored parameters, and aided by specific in-house controls on one of the two types of heterozygous mutations, NGS4THAL achieved excellent performance in hemoglobin SV detections as well. Tested by both simulated and real-world thalassemia patient data, NGS4THAL achieved high sensitivity and specificity.

Homologous DNA sequences in human genome cause challenges for mutation detection by NGS featuring short reads and small insert sizes. Different approaches have been proposed to deal with the alignment issues, namely 'unique' strategy, 'best match' strategy and 'all matches' strategy(41). They refer to keeping only the uniquely mapped reads, reads with the best matches, or all reads including those with lower mapping quality scores, respectively. To our experience, most of the analysis tools by default removes RMAs before variants calling, akin to the 'unique' strategy. In a genome-wide perspective, this approach balances sensitivity and specificity. However, when it comes to a given region, such as the HBA gene locus, this approach leads to the removal of most of the NGS reads derived from *HBA1* and *HBA2*, resulting in failure of detection of mutations in or across these regions.

NGS4THAL makes use of a combination of three approaches: (1) the reads with MAQ > 30 were used as they were, akin to the 'unique' strategy; (2) the reads with MAQ = 0 were reassigned to one of the equally likely mapping positions at random, which follows the 'best match' strategy; (3) for the reads with $0 < \text{MAQ} < 30$ and with differences of mismatches < 3bp between the primary and the alternative hits, they are realigned with the consideration of them being variants-informative, which is akin to the 'all matches' strategy. Tested by both simulated data and real-world sequencing data, the use of a combination of different strategies was shown to handle the multiple alignments issue well from homologous hemoglobin genes and pseudogenes, significantly increasing sensitivity for point mutation detection. Many 'false' detections are caused by the simultaneous reporting of mutations from *HBA1* and *HBA2*, an issue cannot be solved by short read sequencing. Some other false positive callings caused by RMA realignments are from upstream of *HBA2* and between *HBA2* and *HBA1*. As they are not involved in pathogenesis for thalassemia and other hemoglobinopathies, they were treated as likely false detections and were not reported.

Currently, there is no widely accepted *best practice* for SV detection. Most SV detection tools are specialized for a certain type of SVs by making use of limited sequencing signatures. Comparing mapping positions/orientations of read pair and split read, evaluation of read depth and *de novo* assembly of relevant reads are among the most used approaches. Integrating results from complementary SV callers has been suggested to improve SV detection(42). Classifying the hemoglobin SVs into four subtypes helped us select relevant and complementary tools to increase sensitivity.

Compound SVs (such as $-\text{SEA}/-\alpha^{3.7}$) are difficult to handle for any available tools based on a single algorithm. Making use of heterozygous controls on one type of the SVs allows the software to re-estimate read depth baselines and to facilitate the detection of the co-existing overlapping SV. Compound SVs are common in thalassemia epidemic regions, but fortunately, types of compound SVs are predictable according to local epidemiology data, allowing us to prepare this “specific control” dataset to aid the detection process, such as $-\text{SEA}$ in Southeast Asia and $-\text{THAI}$ in Thailand.

For targeted sequencing data, it is likely that certain deletions are even larger than the entire targeted region, which applies to some known thalassemia deletions. In this case, we suggest examination of run of homozygosity (ROH) on high quality SNVs in the region and violation of Mendelian inheritance if parental data are also available. NGS4THAL not only detects known thalassemia mutations, it is also capable of detecting novel hemoglobin variants, disease modifiers, and variants causal to other hemoglobinopathies. Looking up the detected variants in the HbVar database allows easy reporting of known causal mutations. We categorized all the variants into “Known Thalassemia Causal”, “Known Hemoglobinopathy Causal”, “Known non-Causal Hb variants” and “Novel variants” for downstream follow ups.

NGS4THAL is implemented in python and can be easily accessed in https://github.com/JavenCao/Realign_BAM, with detailed specifications. The RMA realignment steps implemented in NGS4THAL can be completed in minutes. The major time-consuming steps for NGS4THAL are variant callings by GATK-HaplotypeCaller and Pindel. Future improvements are needed to ramp up the efficiency when processing large amount of sequencing data, such as in a scenario of population screening.

Conclusions

In summary, taking the DNA sequence specificity of hemoglobin into consideration, we developed a tailored molecular diagnosis and carrier screening pipeline named NGS4THAL, which achieved high sensitivity and specificity for thalassaemia (and other haemoglobinopathies) pathogenic variants detection.

Availability And Requirements

Project name: NGS4THAL

Project home page: https://github.com/JavenCao/Realign_BAM

Operating system(s): Linux, MacOS

Programming language: python, shell

Other requirements: python module dependence(pysam, random, numpy), BreakDancer, Pindel, Conifer

License: GNU GPL v3

Any restrictions to use by non-academics: authorization from the author

List Of Abbreviations

NGS	Next generation sequencing
InDel	Insertion/deletion
SV	Structural variant
HbH	Hemoglobin H
RMA	reads with multiple alignments
MAQ	Mapping quality score
CNV	Copy number variants
WES	Whole exome sequencing
WGS	Whole genome sequencing
PID	Primary Immunodeficiency
MLPA	Multiplex Ligation Dependent Probe Amplification
LD	Linkage disequilibrium
OR	Odds ratio
MCV	Mean corpuscular volume
MCH	Mean corpuscular hemoglobin

Declarations

Ethics approval and consent to participate This study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster, with reference number UW 12-211

Consent for publication Not applicable

Availability of data and materials Not applicable

Competing interests The authors declare that they have no competing interests

Funding This work was supported by CHILDREN'S THALASSAEMIA FOUNDATION of Hong Kong (www.thalassaemia.org.hk, No. 2016/02 and No. 2018/05).

Authors' contributions WY conceived and supervised the project; YC developed and tested NGS4THAL; CCS, TMF, PS, and NH provided Thalassemia patient samples used in this study; CSMT and MMG-B provided the WGS data from the Hirschsprung's disease study; YLL and BHYC provided the WES samples; YC, JY, HZ and RL carried out bench experiments validating findings from NGS4THAL; YLL, GCFC, CCS, SKM and SYH gave suggestions on the design and execution of the project; YC and WY wrote and revised the manuscripts.

Acknowledgements We thank all the study participants from Queen Mary Hospital, Clinical Genetic Service of Hong Kong and Chulalongkorn University, Thailand. YC is supported by HKU Postgraduate Scholarship and the Edward & Yolanda Wong Fund for postgraduate students in the Department of Paediatrics and Adolescent Medicine, University of Hong Kong.

References

1. Piel FB, Weatherall DJ. The α -Thalassemias. *New England Journal of Medicine*. 2014;371(20):1908-16.
2. Rund D, Rachmilewitz E. β -Thalassemia. *New England Journal of Medicine*. 2005;353(11):1135-46.
3. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ*. 2008;86(6):480-7.
4. Giardine B, Borg J, Viennas E, Pavlidis C, Moradkhani K, Joly P, et al. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Research*. 2014;42(D1):D1063-D9.
5. Higgs DR. The molecular basis of α -thalassemia. *Cold Spring Harbor perspectives in medicine*. 2013;3(1):a011718.
6. Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, Maglott D, et al. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*. 2011;43:295.

7. Kalle Kwaifa I, Lai MI, Md Noor S. Non-deletional alpha thalassaemia: a review. *Orphanet Journal of Rare Diseases*. 2020;15(1):166.
8. Martyn GE, Wienert B, Yang L, Shah M, Norton LJ, Burdach J, et al. Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nature Genetics*. 2018;50(4):498-503.
9. Borg J, Papadopoulos P, Georgitsi M, Gutiérrez L, Grech G, Fanis P, et al. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nature Genetics*. 2010;42:801.
10. Wahlberg K, Jiang J, Rooks H, Jawaid K, Matsuda F, Yamaguchi M, et al. The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. *Blood*. 2009;114(6):1254-62.
11. Brancaleoni V, Di Pierro E, Motta I, Cappellini M. Laboratory diagnosis of thalassemia. *International journal of laboratory hematology*. 2016;38(S1):32-40.
12. Ip H-W, So C-C. Diagnosis and prevention of thalassemia. *Critical reviews in clinical laboratory sciences*. 2013;50(6):125-41.
13. Papasavva T, Van Ijcken WF, Kockx CE, Van Den Hout MC, Kountouris P, Kythreotis L, et al. Next generation sequencing of SNPs for non-invasive prenatal diagnosis: challenges and feasibility as illustrated by an application to β -thalassaemia. *European Journal of Human Genetics*. 2013;21(12):1403.
14. He J, Song W, Yang J, Lu S, Yuan Y, Guo J, et al. Next-generation sequencing improves thalassemia carrier screening among premarital adults in a high prevalence population: the Dai nationality, China. *Genetics in Medicine*. 2017.
15. Shang X, Peng Z, Ye Y, Zhang X, Chen Y, Zhu B, et al. Rapid Targeted Next-Generation Sequencing Platform for Molecular Screening and Clinical Genotyping in Subjects with Hemoglobinopathies. *EBioMedicine*. 2017;23:150-9.
16. Lam K-WG, Jeffreys AJ. Processes of copy-number change in human DNA: the dynamics of α -globin gene deletion. *Proceedings of the National Academy of Sciences*. 2006;103(24):8921-7.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 2009;25(14):1754-60.
18. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-8.
19. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*. 2009;6(9):677-81.
20. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-71.
21. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. *Genome research*. 2012;22(8):1525-32.

22. Mu JC, Mohiyuddin M, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*. 2014;31(9):1469-71.
23. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28(4):593-4.
24. Lee PP-W, Lau Y-L. Improving care, education, and research: the Asian primary immunodeficiency network. *Annals of the New York Academy of Sciences*. 2011;1238(1):33-41.
25. Tang CS-m, Dattani S, So M-t, Cherny SS, Tam PK, Sham PC, et al. Actionable secondary findings from whole-genome sequencing of 954 East Asians. 2018;137(1):31-7.
26. Tang CS-m, Li P, Lai FP-L, Fu AX, Lau S-T, So MT, et al. Identification of Genes Associated With Hirschsprung Disease, Based on Whole-Genome Sequence Analysis, and Potential Effects on Enteric Nervous System Development. *Gastroenterology*. 2018;155(6):1908-22.e5.
27. Wang Y-F, Zhang Y, Lin Z, Zhang H, Wang T-Y, Cao Y, et al. Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nature Communications*. 2021;12(1):772.
28. The Genomes Project C. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
29. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The International HapMap Project. *Nature*. 2003;426(6968):789-96.
30. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2004;21(2):263-5.
31. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. 2003;165(4):2213-33.
32. Lau Y-L, Chan L-C, Chan Y-YA, Ha S-Y, Yeung C-Y, Waye JS, et al. Prevalence and genotypes of α - and β -thalassemia carriers in Hong Kong—implications for population screening. *New England Journal of Medicine*. 1997;336(18):1298-301.
33. O’Riordan S, Hien TT, Miles K, Allen A, Quyen NN, Hung NQ, et al. Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. *British Journal of Haematology*. 2010;150(3):359-64.
34. Lam K-WG, Jeffreys AJ. Processes of de novo duplication of human α -globin genes. *Proceedings of the National Academy of Sciences*. 2007;104(26):10950-5.
35. Nguyen NT, Sanchaisuriya K, Sanchaisuriya P, Van Nguyen H, Phan HTT, Fucharoen G, et al. Thalassemia and hemoglobinopathies in an ethnic minority group in Central Vietnam: implications to health burden and relationship between two ethnic minority groups. 2017;8(3):221-8.
36. Nguyen HV, Sanchaisuriya K, Nguyen D, Phan HTT, Siridamrongvattana S, Sanchaisuriya P, et al. Thalassemia and hemoglobinopathies in Thua Thien Hue province, central Vietnam. *Hemoglobin*. 2013;37(4):333-42.

37. O’Riordan S, Hien TT, Miles K, Allen A, Quyen NN, Hung NQ, et al. Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. 2010;150(3):359-64.
38. Qiu Q-W, Wu D-D, Yu L-H, Yan T-Z, Zhang W, Li Z-T, et al. Evidence of recent natural selection on the Southeast Asian deletion (–SEA) causing α -thalassemia in South China. BMC evolutionary biology. 2013;13(1):1.
39. Lorey F, Cunningham G, Vichinsky EP, Lubin BH, Witkowska HE, Matsunaga A, et al. Universal newborn screening for Hb H disease in California. Genetic testing. 2001;5(2):93-100.
40. Cousens NE, Gaff CL, Metcalfe SA, Delatycki MB. Carrier screening for Beta-thalassaemia: a review of international practice. European Journal of Human Genetics. 2010;18(10):1077-83.
41. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012;13(1):36-46.
42. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.

Figures

Figure 1

Workflow of NGS4THAL: RMA: NGS read with multiple alignment; SNV: single nucleotide variants; InDel: small insertion and deletion; SV: structural variant; CNV: copy number variant.

(A) HBA2 gene

(B) HBA1 gene

(C) HBB gene

(D) HBD gene

Figure 2

Detection of point mutations and small indels by NGS4THAL on simulated data.

Figure 3

Detection of hemoglobin structural variants by NGS4THAL using simulated data: (1) Hemoglobin structural variants (SVs) classification are based on relative positions between the homologous regions and SV breakpoints. (2) For each SV subtype, the top panel shows the alignment landscape in the format of IGV screenshot, followed by the simulated SVs and detection results from BreakDancer, Pindel and CoNIFER (with cutoffs of 1.0, 0.8, 0.5 and 0.3), respectively. CoNIFER detections with cutoff of 0.5 are used in NGS4THAL and are shown in green. (3) Lower panels in each section show read-depth changes detected by CoNIFER. Black lines are read-depth changes for all the simulated individuals. Red curves

show the read depth changes for the SV carriers in question. Y-axis is the relative copy number with reference to the baseline defined by all input samples; vertical dashed green lines illustrate the start and end of the CoNIFER variant calls using a cutoff of 0.5. (3) The simulated SVs are 2.7 kb deletion[NG_000006.1:g.36664_39364del2701] representing SV Subtype A, 5.3 kb deletion[NG_000006.1:g.28684_33930del5246] representing Subtype B, 3.5 kb deletion[NG_000006.1:g.37464_40964del3501] representing Subtype C, and 3.7 kb (type I) deletion[NG_000006.1:g.34164_37967del3804] representing Subtype D.

Figure 4

Detection of a compound structural variant (–SEA/–α^{3.7}) by NGS4THAL through a stage-wise process: (A) shows the alignment landscape in the format of IGV screenshot, as well as detection results from BreakDancer, Pindel and CoNIFER. (B) shows the read-depth change in the initial screening stage when all the samples were pooled together for analysis by CoNIFER. (C) shows the read-depth changes in the fine-profiling stage when pooling all the –SEA deletion carriers together, from in-house control database if necessary. Detection of homozygous –α^{3.7} shown on the right do not need the fine profiling stage.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.pdf](#)
- [Additionalfile2.pdf](#)
- [Additionalfile3.pdf](#)