Note 1. Technique details on development of NGS4THAL

(1) Realignment of informative reads with multiple alignments (RMAs) from hemoglobin regions for point mutation/small InDel detection

RMA positioning To increase detection sensitivity for point mutation and small InDel, informative RMAs were examined, recovered, and realigned. In the alignment file in BAM format, XA tag is a field with information on the alternative mapping for each read (if the read is a RMA). Firstly, to determine where RMAs happened, we extracted and plotted reads with XA tags using trackViewer packages(1). Then to further understand the relationship between homologous hemoglobin local elements (such as genes, pseudogenes) and the RMAs, we plotted the primary mapping positions against alternative mapping positions for these RMAs, and found that the RMAs mainly happened within three clusters, namely *HBZ-HBZP1*, *HBA2-HBA1-HBAP1* and *HBG1-HBG2*.

Defining features of RMAs from hemoglobin regions After RMA positioning, we studied the features of these RMAs, including difference of mismatch between the primary hit and alternative hit (mismatch difference for short), as well as mapping quality scores (MAQ). Mismatch difference (in base pair unit) between the primary hit and alternative hit was calculated based on the edit-distance from

NM tags in the BAM file. Mismatch difference = 0 means that the mapping profiles are identical at the two mapping positions (**Additional file 2:** Error! Reference source not found. **- A**) and nearly 60% of RMAs fell into this category (**Additional file 2:** Error! Reference source not found. **- A**). Besides, these RMAs are usually with MAQ equals zero. The remaining ~40% RMAs were with the mismatch difference $\neq$ 0, which may carry information on variants between homologous genes and pseudogenes (**Additional file 2:** Error! Reference source not found. **- B**). In total, around 95% of RMAs had the mismatch difference <= 3 . MAQ is used to quantify the probability that a NGS read is misplaced. We calculated the average MAQ (scored by bwa mem algorithm) with a window size of 100bps and step size of 50bps for the hemoglobin regions. Alignments with MAQ less than 30 account for 45%-65% of all the RMAs (**Additional file 2:** Error! Reference source not found. **- B**). Regions with low MAQ were located in *HBZ*, *HBZP1*, *HBAP1*, *HBA2* and *HBA1* for α-globin locus, and in *HBG1* and *HBG2* for β-globin locus.

Realignment of RMAs in hemoglobin regions To balance between specificity while keeping enough informative reads, RMAs meeting the following criteria were recovered and realigned: (1) being mapped to the homologous regions, namely *HBZ-HBZP1*, *HBA2-HBA1-HBAP1* and *HBG1-HBG2*; and (2) having bwa(2)-based mapping quality score (MAQ) of zero due to having an alternative

mapping position with identical mapping profile as the primary mapping position; and (3) MAQ < 30 and the difference of mismatches between the primary mapping hit and the alternative hit is less than 3 bps under the short reads settings (*e.g.* 100bps - 300bps). Identical mapping profile means that a NGS read can be mapped either to position 1 or to position 2 on the reference genome with the same number of mismatches. The mismatch can be 0 or not 0 (*e.g.*, 1, 2, >=3). We didn't include RMAs with MAQ > 30, since they could be treated as unique mapping and realignment of them would increase downstream burdens due to false positive detection. Once we recovered the candidate RMAs, R1 read of a pair of RAM was randomly placed into either primary mapping position or alternative mapping position, and then mate R2 read was placed at its mapping position that was no farther than [mean insert size of the NGS library ± 5 * standard deviation]. Also, MAQ of these realigned RAMs were fixed to 60 before generating the realigned BAM file. Here, the mean insert size of the NGS library was estimated by uniquely aligned read pairs with MAQ = 60. To exclude read pairs from two ends of large deletions, we restricted the selected read pairs having insert sizes of less than 1000 bps. Python module pysam(3) was used to complete the above process.

Variants calling After realignment of the informative RMAs, SNVs/InDels were called from the realigned BAM file by GATK-HaplotypeCaller(4). To make sure that we didn't miss known variants (including known pathogenic SNVs/InDels

for thalassemia and other hemoglobinopathies, known neutral Hb variants) from HbVar(5) database and novel Hb variants, SNVs/InDels that could be called on *equivalent positions* on the homologous genes, such as HBA2:c.64G>C and HBA1:c.64G>C, were both reported for further examination, such as using laboratory methods. Besides, to trace the influence of RMA realignment on the variants calling process, candidate variants loci, which is suggested by the re-assembled haplotypes from --*bamOutput* option of GATK-HaplotpCaller(4), were outputted before and after the realignment process.

(2) Integrative usage of various tailored software in a stage-wise process to detect hemoglobin structural variants (SVs)

Usage of the algorithm based on read-pair analysis BreakDancer first tallies anomalous read pairs with discordant separation distance and/or alignment orientation, then regions harbouring more anomalous read pairs than expected are recognized as different structural variants[6], amongst regions with the anomalous read pairs mapped too far apart showing increased insert sizes are defined as deletions. Subtype-A deletions, with unique aligned NGS read pairs from two ends of them can present with typical increased insert size (**Additional file 2:** Error! Reference source not found. **- A**). Due to the fusion-gene nature of Subtype-D deletions, read pairs physically from two ends of them can be mapped at positions with normal insert sizes, so no anomalous read pairs with increased

insert size could be used by BreakDancer to detect Subtype-D deletions (**Additional file 2:** Error! Reference source not found. **- D**).

When considering read pairs near the breakpoints of Subtype-B and Subtype-C deletions, only one read of the read pair could be uniquely mapped, and the other mate read would be mapped ambiguously to homologous regions, resulting in at least three insert size changes (**Additional file 2:** Error! Reference source not found. **- B,C**). To increase detection sensitivity, we included these RMAs by lowing the MAQ cutoff to zero (*-q 0*), allowing BreakDancer to report all these candidate events; then by matching with the known hemoglobin SVs from HbVar database, we reported the known Hb variants and novel Hb variants.

Use of the algorithm based on split-read analysis Pindel first finds an anchor point near the SV breakpoint by the uniquely mapped read of a read-pair, and then the exact breakpoint was determined by splitting and mapping the unmapped mate read[7]. NGS reads overlapping with the breakpoint of Subtype-A deletion can be split to determine the exact breakpoint (**Additional file 2:** Error! Reference source not found. **- A**), however, for Subtype-D deletions, NGS reads physically overlapping with the breakpoints are expected to map properly without clipping certain bases (although they are RMAs) (**Additional file 2:** Error! Reference source not found. **- D**), so these reads cannot be split to determine the breakpoint by Pindel.

For Subtype-B and Subtype-C deletions, after finding anchor points by the uniquely mapped reads of read pairs, the split-fragments of the mate reads could be ambiguously mapped to homologous regions, causing false SV callings (**Additional file 2:** Error! Reference source not found. **- B,C**). We filtered the Pindel calling results by matching the detected breakpoints with the known Hb SVs in the database. Besides, considering that change in insert size and split reads at the breakpoints always happen together near a SV, to further increase SV detection sensitivity and specificity, we included the BreakDancer results as an input for Pindel[7].

Use of the algorithm based on read-depth analysis CoNIFER was originally designed for detecting CNV(Copy Number Variant) from exome sequencing data by analysis of coverage depth(8). The read-depth change for these 4 SV subtypes should be obvious and be easily used (**Additional file 2:** Error! Reference source not found.). And we made CoNIFER suited for analyzing continuous sequencing data by inputting the *in-silico* probes with overlapping baits. Also, to increase the detection sensitivity and maintaining high specificity, we lowered the detection cut-off from default 1.0 to 0.5, which let the tailored CoNIFER be able to detect candidate CNVs with more than 50% read depth change comparing with the baseline average, instead of the default parameter requiring 100% change in depth.

Note 2. Continuous targeted capture followed by next generation sequencing for the detection of thalassemia mutations and disease modifiers

Targeted regions inclusion We developed a targeted capture enrichment scheme in order to sequence the regions relevant for detecting hemoglobin variants and reported thalassemia disease modifiers. Firstly, HBA gene clusters composed of 5'- *HBZ* - *HBZP* - *HBAP1* - *HBA2* - *HBA1* - *HBQ* - 3' along with the upstream regulatory elements (named Multispecies Conserved Sequences or MCS-R1 to R4), and HBB gene clusters composed of 5' - *HBE* - *HBG2*- *HBG1* - *HBD* - *HBB* - 3' along with a distal locus control regions (known as LCR) were continuously captured. This covers most of the known pathogenic variants for thalassemia and other hemoglobinopathies, including single nucleotide variants, small InDels, large deletions and duplications. Secondly, to enable prediction of the disease severity, exonic regions of the master erythroid transcription factor genes, including *BCL11A*(9-12), *KLF1*(13), *GATA1*(14), *MYB*(15) which control fetal to adult hemoglobin switch, and regulatory SNPs(16, 17) (rs6732518, rs11886868, rs1427407, rs766432, rs9399137, rs9494145) which are associated with fetal hemoglobin levels were also targeted. In addition, considering the high frequency carriers rates of Glucose-6-phosphate dehydrogenase deficiency (G6PD) variants in malaria endemic regions, *G6PD* gene was targeted as well. In total, 280kbps

sequences are targeted in the capture enrichment scheme (**Additional file 3:** Error! Reference source not found.).

Design probes to achieve continuous NGS coverage Probe hybridization was used to enrich DNA fragments from the targeted regions. Considering the wide existence of repeats in hemoglobin regions, to get more coverage, we allowed the designed probes to have one or more than one, but no more than three targeting positions on human genome (1 <= close match <= 3), which achieved 85.8% coverage of the proposed regions. Further checking with RepeatMasker(18) showed that the uncovered regions (13.7%) were mostly repeat elements. Also, we checked that the probe-free gaps were predominantly smaller than 500bps and were mostly repeats elements. With average fragment size of 500bps or even shorter during the NGS library construction, this capture scheme meant to ensure that with off-target NGS reads sequenced from the fragments near the probe-free gaps, continuous coverage of the two hemoglobin regions could be achieved. The DNA probes were designed and synthesised by Roche Nimblegene(19).

NGS Sequencing Library Preparation All of the NGS libraries, including the capture reactions, were prepared based on the protocols of SeqCap EZ HyperCap Workflow Version 1.2. One hundred nanograms of genomic DNA was fragmented by Bioruptor pico. The fragmented DNA was subjected to end-repair, A-tailing at the 3’ end, and adaptor (with indexes, such as NEXTflex-96 DNA

Barcode can be used as the index) ligation at the terminal ends. The adapter ligated library with size selected in the range of 250-450bp by the dual-SPRI method. PCR was performed and the products were validated on an Agilent Bioanalyzer.

Before hybridization, libraries were normalized and combined with different indices into a single pool of one microgram in total prior to enrichment. The pooled DNA libraries were mixed with our designed capture probes of targeted regions (**Additional file 3:** Error! Reference source not found.). The hybridization was performed at 47°C for 16-20 hours to ensure targeted regions bound to the capture probes thoroughly. Streptavidin beads were used to capture probes containing the targeted regions of interest. Four wash steps with different wash buffers and temperatures were performed to remove nonspecific binding from the beads. The enriched libraries on the beads were then amplified by PCR. The enriched libraries were validated by Agilent Bioanalyzer, Qubit and qPCR for quality control analysis. The libraries were denatured and diluted to optimal concentration and applied in the cluster generation steps. HiSeq PE Cluster Kit v4 with cbot was used for cluster generation on the flow cell. Illumina HiSeq SBS Kit v4 was used for PE101 sequencing. The NGS sequencing was performed in Centre for Genomic Sciences, The University of Hong Kong.

Note 3. Experimental confirmation of variants detected by NGS4THAL

To confirm the thalassemia pathogenic variants detected by NGS4THAL from analyzing WGS data, representative variants were selected for orthogonal validation by experimental methods. The representative variants were selected according to the following criteria: (1) at least one sample was included for each kind of the detected pathogenic variants; (2) if a pathogenic variant was detected in more than one ethnic populations (Hong Kong Chinese, Northern Vietnamese and Mainland Chinese), then at least one sample from each population was included; (3) if both heterozygote and homozygote were detected for a pathogenic variant, then both the genotypes would be confirmed.

In details, Sanger sequencing was used for confirmation of point mutations/small InDels. PCR primers (same as sanger sequencing primer) were designed online by NCBI Primer-BLAST tool and synthesized by IDT company (**Additional file 3:** Error! Reference source not found.), and Sanger sequencing was done by Centre for Genomic Sciences, The University of Hong Kong. Structural variants were confirmed by MPLA methods with Holland SALSA MLPA P140 HBA probemix for α-thalassemia *in vitro* diagnosis. Despite the waiting-to-be-confirmed positive samples reported by NGS4THAL, a negative sample composed of TE0.1 (without DNA), a positive sample which was previously

verified by MLPA to carry -$\alpha^{3.7}$, and three reference baseline controls with NGS4THAL-tested-negative for CNV/SV were also included in the MLPA assay. Blood-derived DNA were first treated with RNase to eliminate the influence from RNA. We followed the instructions from the retailer and visualized the results by Coffalyser.Net software.

. … .. . .

Note 4. Haplotype analysis

The 3, 516 non_--^{SEA} individuals were samples collected from whole genome sequencing dataset and samples selected from our in-house SLE-GWAS cohort(20) (**Additional file 3:** Error! Reference source not found.). Absence of --^{SEA} was confirmed by NGS4THAL for samples from WGS dataset. For those with only SNP genotyping data, absence of_--^{SEA} status was inferred by the heterozygosity of SNP rs2974771, which is located in --^{SEA} region (hg19, chr16:221057). In the SLE GWAS dataset, this SNP has a genotyping rate of 98.4%, Hardy-Weinberg Equilibrium(P<1.0E-4) and the minor allele frequency of 0.43 in Hong Kong Chinese.

REFERENCES

1. Ou J, Zhu LJ. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nature Methods. 2019;16(6):453-4.

2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60.

3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. Proceedings of the National Academy of Sciences. 2008;105(5):1620-5.

10. Masuda T, Wang X, Maeda M, Canver MC, Sher F, Funnell AP, et al. Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. Science. 2016;351(6270):285-9.

11. 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.

12. Sankaran VG, Menne TF, Xu J, Akie TE, Lettre G, Van Handel B, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science. 2008;322(5909):1839-42.

13. 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.

14. 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.

15. Jiang J, Best S, Menzel S, Silver N, Lai MI, Surdulescu GL, et al. cMYB is involved in the regulation of fetal hemoglobin production in adults. Blood. 2006;108(3):1077-83.

16. Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nature genetics. 2007;39(10):1197-9.

17. Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA, Sripichai O, et al. A genome-wide association identified the common genetic variants influence disease severity in β 0-thalassemia/hemoglobin E. Human genetics. 2010;127(3):303-14.

18. Smit A, Hubley R, Green P. RepeatMasker at http://repeatmasker. org. Google Scholar. 2016.

19. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nature biotechnology. 2011;29(10):908.

20. Wang Y-F, Zhang Y, Zhu Z, Wang T-Y, Morris DL, Shen JJ, et al. Identification of ST3AGL4, MFHAS1, CSNK2A2 and CD226 as loci associated with systemic lupus erythematosus (SLE) and evaluation of SLE genetics in drug repositioning. Annals of the Rheumatic Diseases. 2018:annrheumdis-2018-213093.