Supplementary Figures

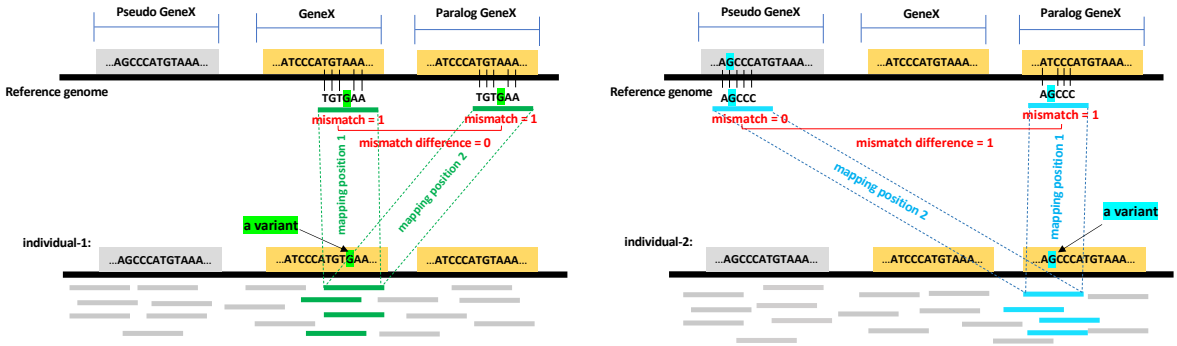
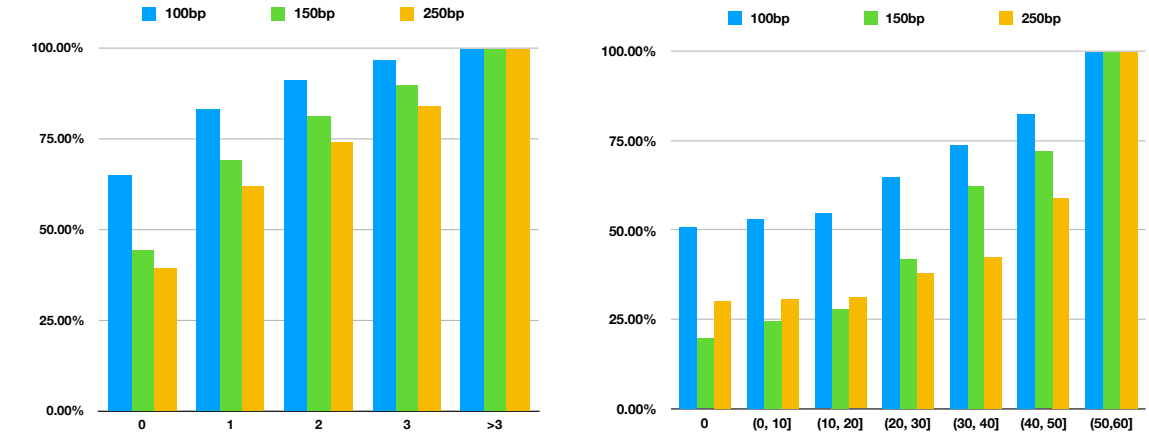


Figure S1. Scheme showing reads with multiple alignments in homologous gene locus

Notes to Figure S1: (A) a variant happening in a gene (GeneX) with variants-informative NGS reads mapping to multiple positions in its homologous paralog gene (Paralog GeneX) with mismatch difference equals to zero; **(B)** a variant happening in a gene (GeneX) with variants-informative NGS reads mapping to multiple positions in its homologous pseudogene with mismatch difference equals to 1.



(A) Mismatch differences between the primary hits and alternative hits for RMAs

(B) Mapping quality score(MAQ) for RMAs

Figure S2. Features for NGS reads with multiple alignments (RMAs) from hemoglobin gene regions

Notes to Figure S2: (A) Mismatch differences between the primary hits and alternative hits for RMAs. Showed on X-axis are the mismatch differences in base pair unit. Y-axis is the cumulative percentage, colored by different NGS read length. **(B)** Mapping quality score (MAQ) for RMAs. MAQ intervals are showed on X-axis. Y-axis is the cumulative percentage, colored by different NGS read length.

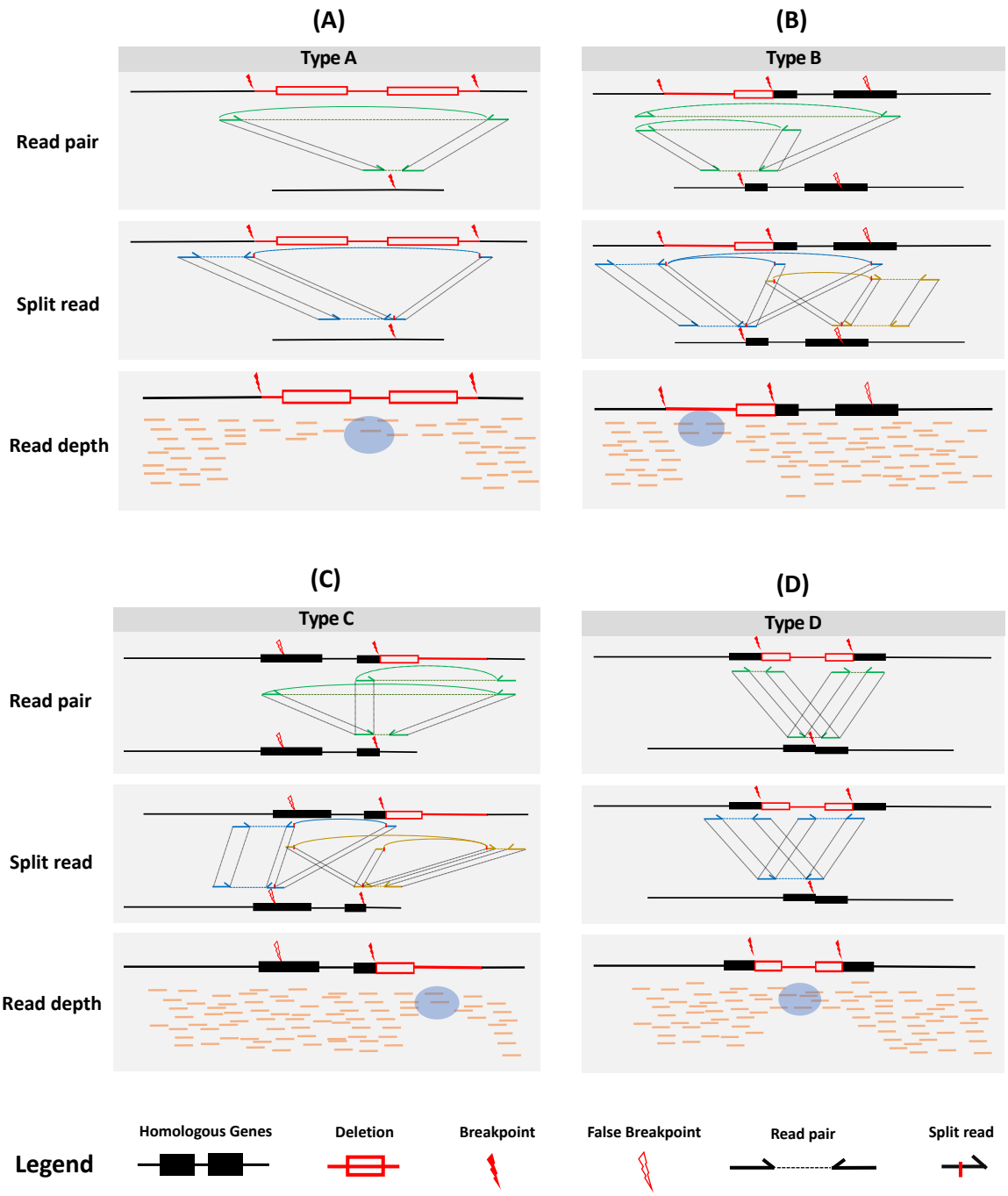
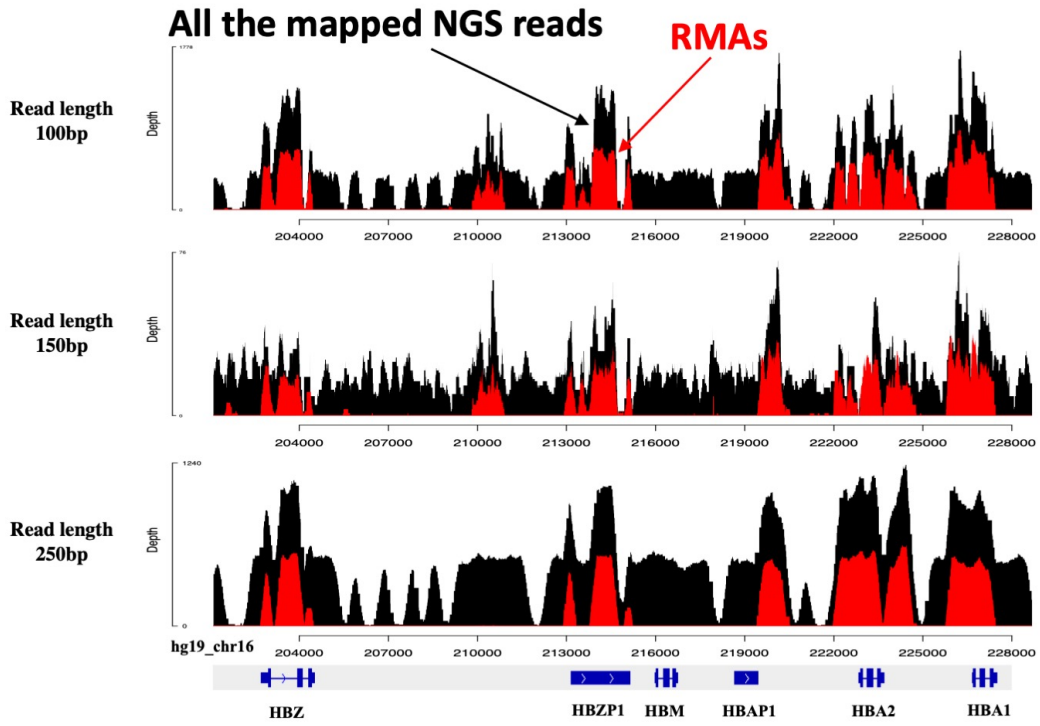
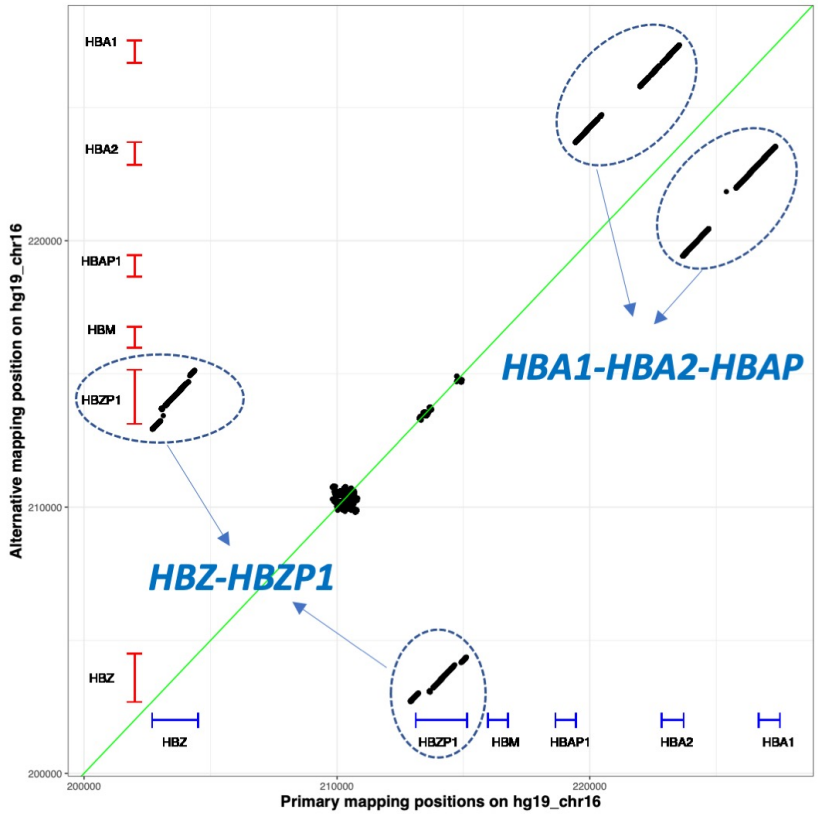


Figure S3. Illustration of characteristics of hemoglobin structural variants from NGS sequencing

Notes to Figure S3: Hemoglobin structural variants (SVs) (*e.g.*, deletions) classification are based on relative positions between the homologous regions and SVs, and their characteristics from paired-end short reads NGS sequencing, including read-pair, split-read and read-depth, are shown.



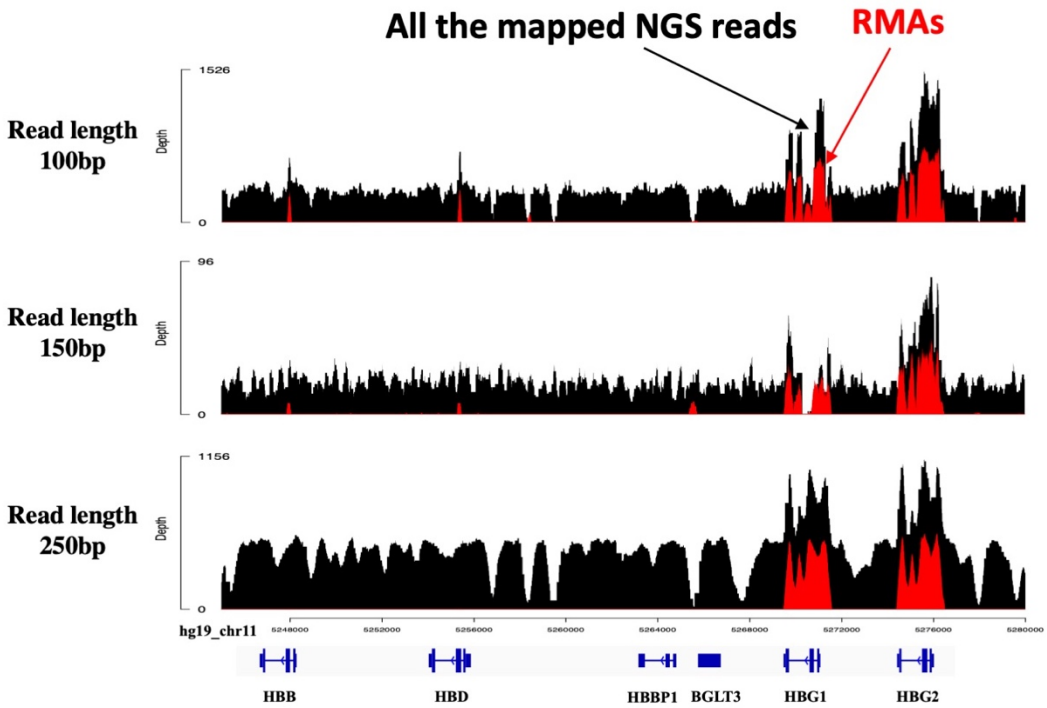
(A)



(B)

Figure S4. Positioning of reads with multiple alignments (RMAs) in HBA gene clusters

Notes to Figure S4: (A) Distribution of RMAs in HBA gene clusters. Black curves show all the alignments; red curves show RMAs. Y-axis is the depth of coverage from NGS sequencing with different read length. **(B)** RMAs happen between homologous genes and pseudogenes. Primary mapping positions and alternative mapping positions for RMAs are shown on X-axis and Y-axis, respectively.



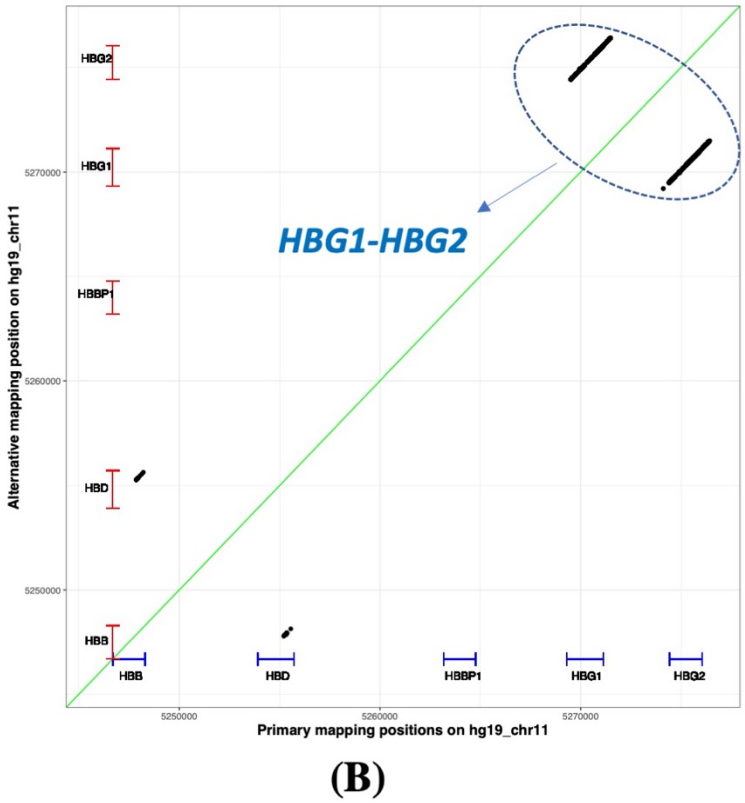
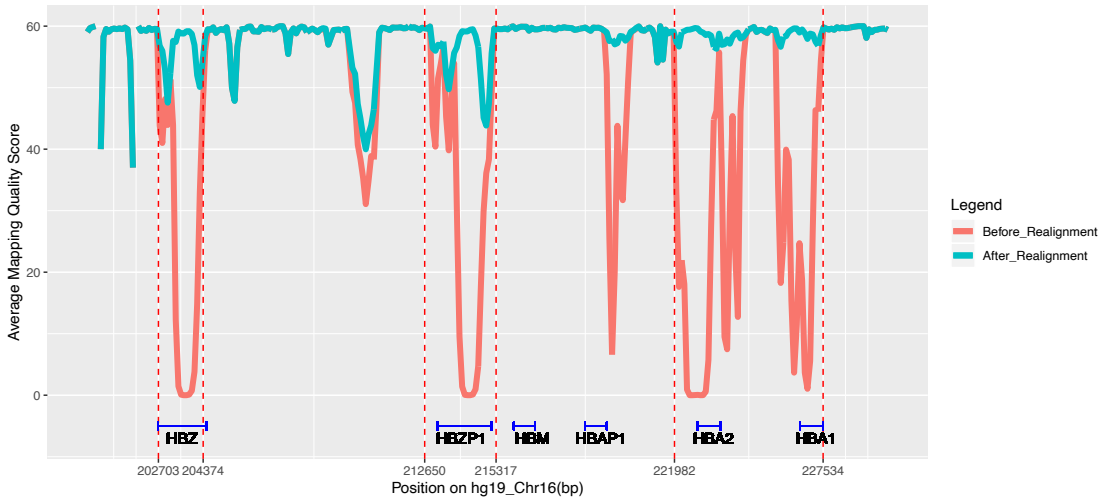
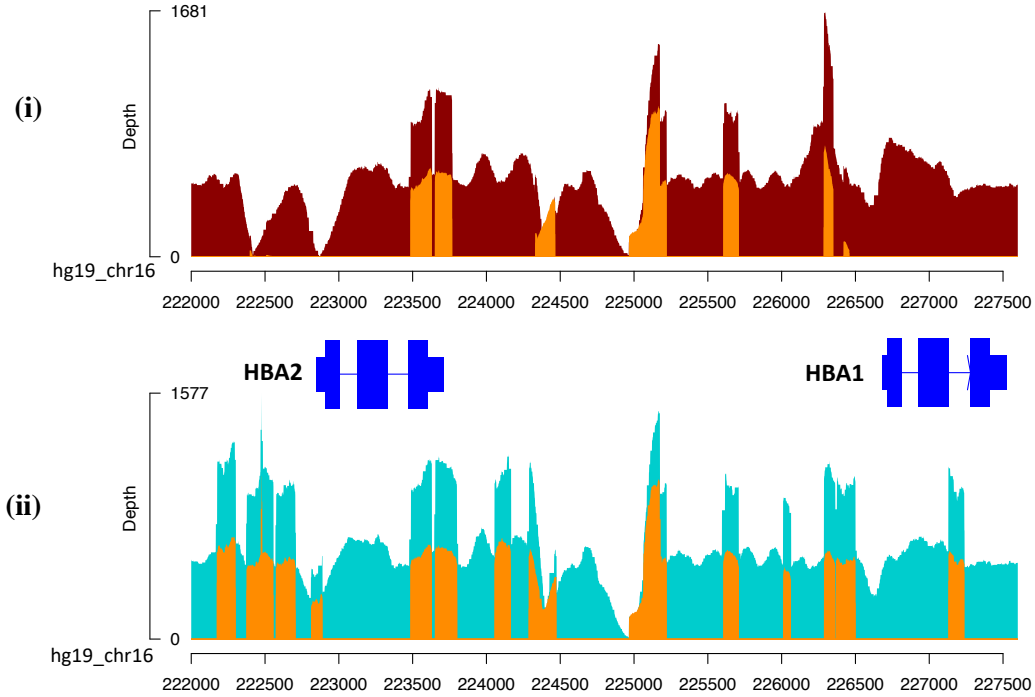


Figure S5. Positioning of reads with multiple alignments (RMAs) in HBB gene clusters

Legends are the same as those in Figure S4.



(A)



(B)

Figure S6. Comparison of the alignments before and after the realignment of informative reads with multiple alignments (RMAs)

Notes to Figure S6: **(A)** Comparison of mapping quality score (MAQ) of the alignments before (coloured in red) and after (coloured in cyan) the realignment process. Y-axis is average MAQ, calculated with a window size of 100bps and step size of 50bps from 1500-fold 100bps paired-end targeted sequencing with continuous coverage of HBA gene locus. **(B)** Comparison of candidate variants loci before and after the realignment process. (1) Red curves in track (i) show all the aligned NGS reads for *HBA2* and *HBA1* regions before the realignment process, and the orange curves are the candidate variants loci. Cyan curves in track (ii) show all the aligned NGS reads for *HBA2* and *HBA1* regions after the realignment process, and the orange curves are the candidate variants loci. (2) Candidate variants loci were defined by the re-assembled haplotypes from GATK-HaplotypeCaller(--*bamOutput* option).

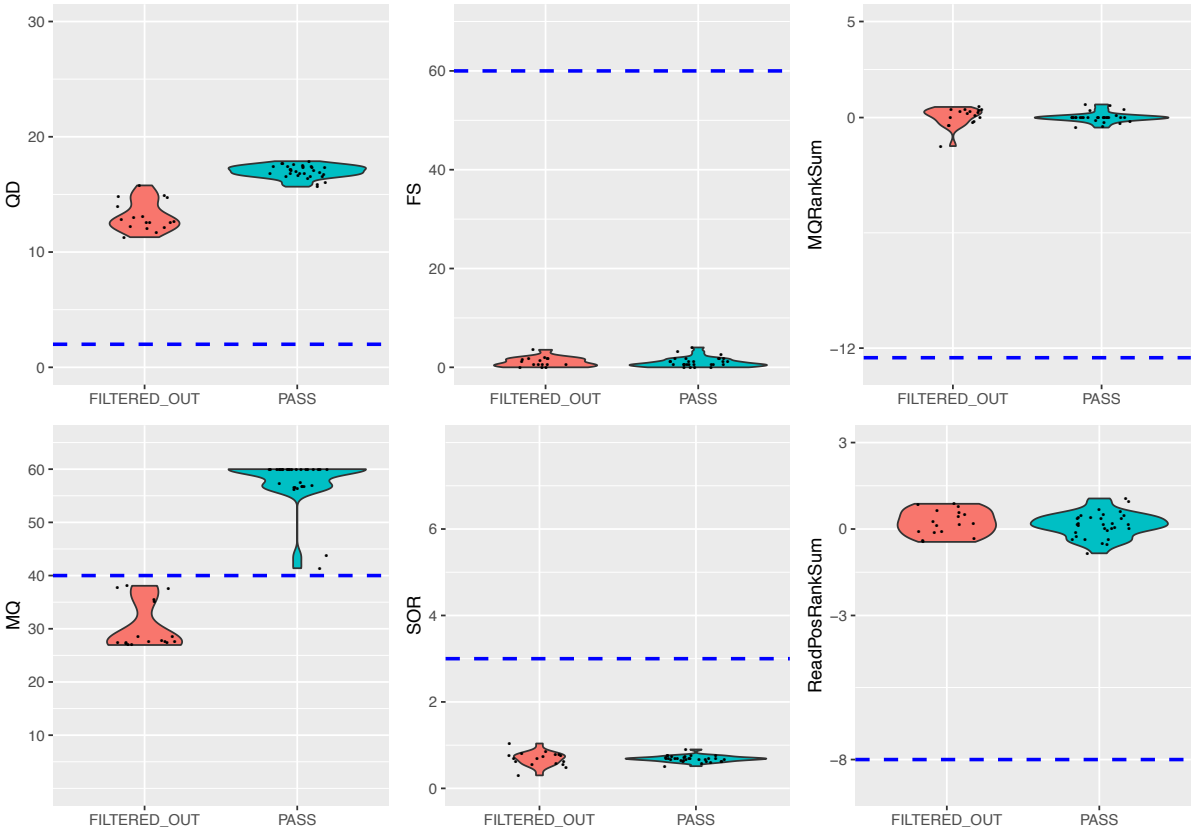


Figure S7. Annotation distribution for the simulated SNVs before realignment of reads with multiple alignments

Notes to Figure S7: (1) Blue dashed lines are cut-off values to determine PASS/FILTERED_OUT status. PASS status is defined as: QD > 2.0 and FS < 60.0 and MQ > 40.0 and MQRankSum > -12.5 and Read-PosRankSum > -8.0. **(2)** QD: variant confidence/quality by depth; MQ: RMS mapping quality; FS: Phred-scaled *P-value* using Fisher's exact test to detect strand bias; SOR: symmetric odds ratio of 2x2 contingency table to detect strand bias; MQRankSum: Z-score from Wilcoxon rank sum test of alternative to reference read mapping qualities; ReadPosRankSum: Z-score from Wilcoxon rank sum test of alternative to reference read position bias.

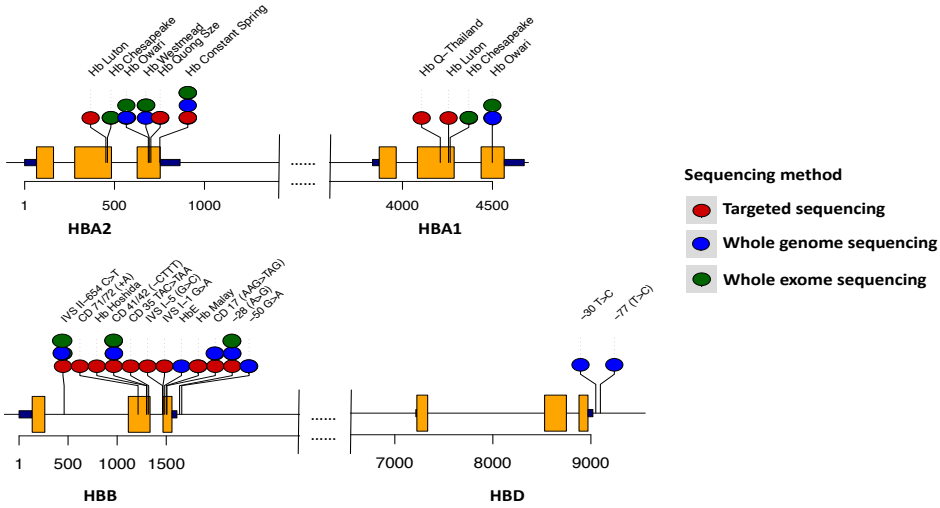


Figure S8. Detected mutation spectrum by NGS4THAL from real-word NGS sequencing cohorts

Notes to Figure S8: (1) HGVS names for the listed mutations are as follows: Hb Luton: HBA2:c.269A>T(or HBA1), Hb Chesapeake: HBA2:c.278G>T(or HBA1), Hb Owari: HBA2: c.364G>A (or HBA1), Hb Westmead: HBA2: c.369C>G, Hb Quong Sze(Hb QS): HBA2:c.377T>C, Hb Constant Spring(Hb CS): HBA2: c.427T>C, Hb Q-Thailand: HBA1:c.223G>C, IVS II-654 C>T: HBB:c.316-197C>T, CD 71/72 (+A): HBB:c.217dupA, Hb Hoshida: HBB:c.130G>C, CD 41/42 (-CTTT): HBB:c.126_129delCTTT, CD 35 TAC>TAA: HBB:c.108C>A, IVS I-5 (G>C): HBB:c.92+5G>C, IVS I-1 G>A: HBB:c.92+1G>A, HbE: HBB:c.79G>A, Hb Malay: HBB:c.59A>G, CD 17 (AAG>TAG): HBB:c.52A>T, -28 (A>G): HBB: c.-78A>G, -50 G>A: HBB:c.-100G>A, -30 T>C: HBD: c.-80T>C, -77 (T>C): HBD:c.-127T>C. **(2)** Hb Luton, Hb Chesapeake, Hb Owari, Hb Q -Thailand, Hb Hoshida are hemoglobin variants.

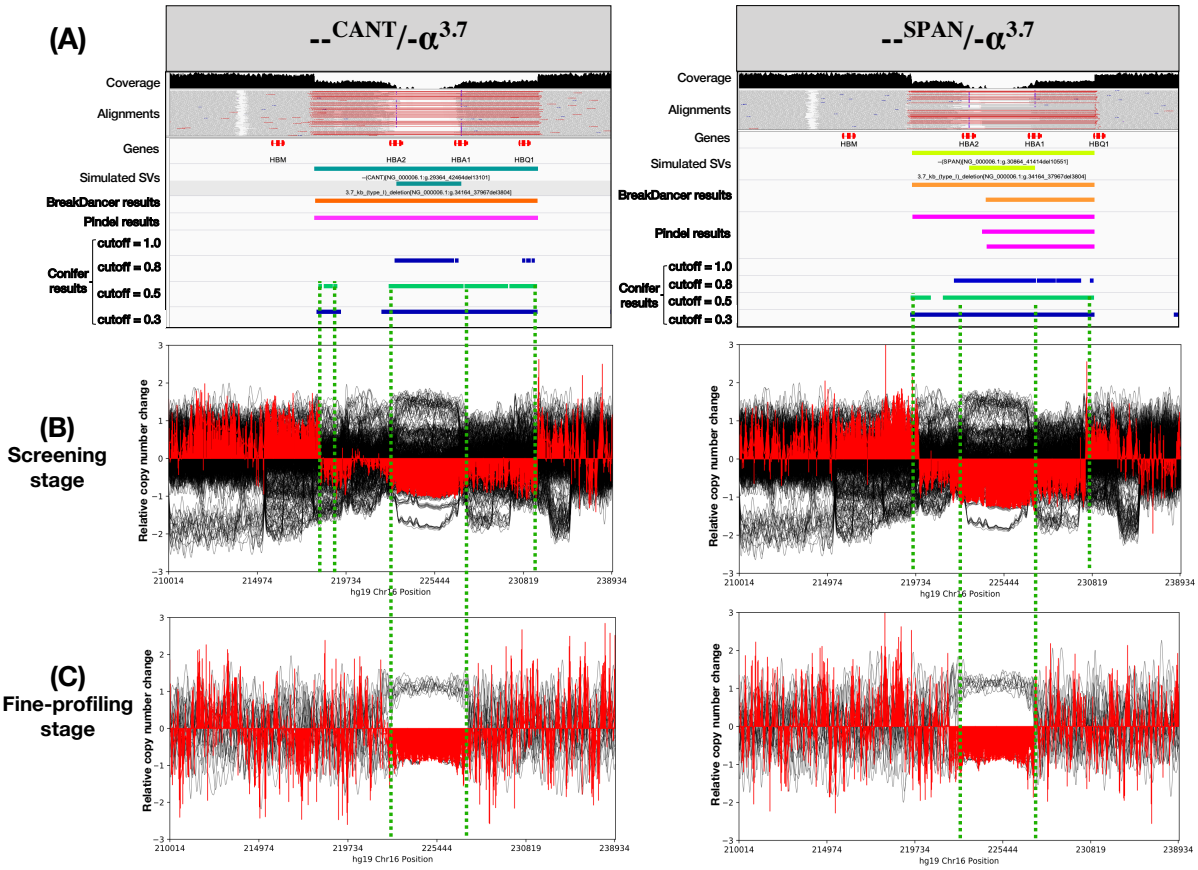


Figure S9. Detection of two compound rearrangements by NGS4THAL through stage-wise process in simulation

Notes to Figure S9: (A) Alignment landscape by IGV screenshot; (B) read-depth change analysis in the screening stage; (C) read-depth change analysis in the fine-profiling stage. Other legends are the same as those in Figure 4.

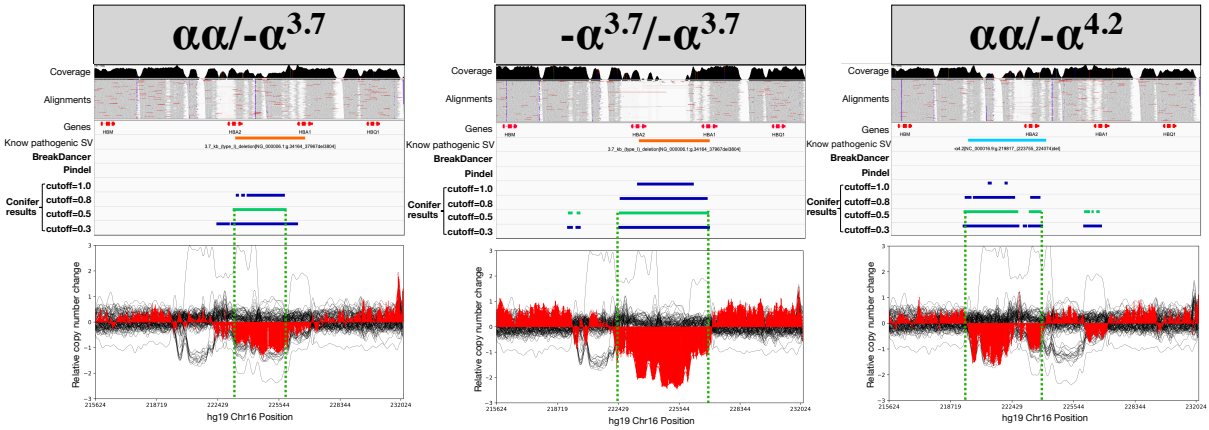


Figure S10. Detection of deletions from the targeted sequencing data by NGS4THAL

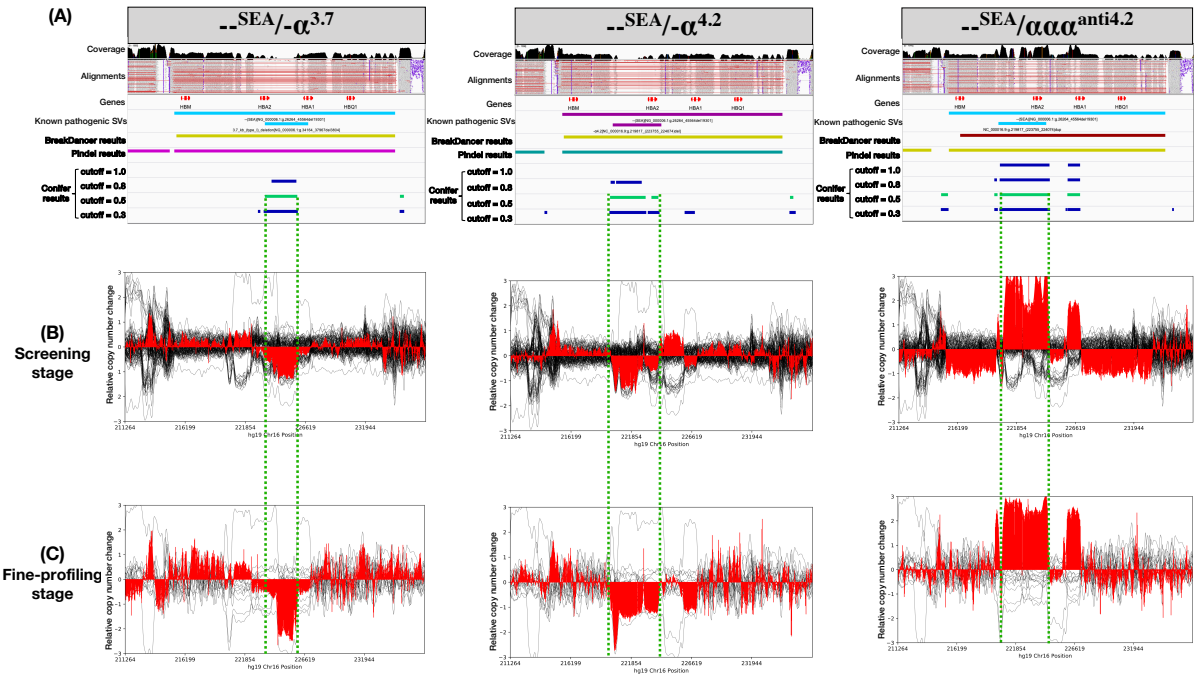


Figure S11. Detection of compound rearrangements from the targeted sequencing data by NGS4THAL

Notes to Figure S11: (A) Alignment landscape by IGV screenshot; (B) read-depth change analysis in the screening stage; (C) read-depth change analysis in the fine-profiling stage. Other legends are the same as those in Figure S9.

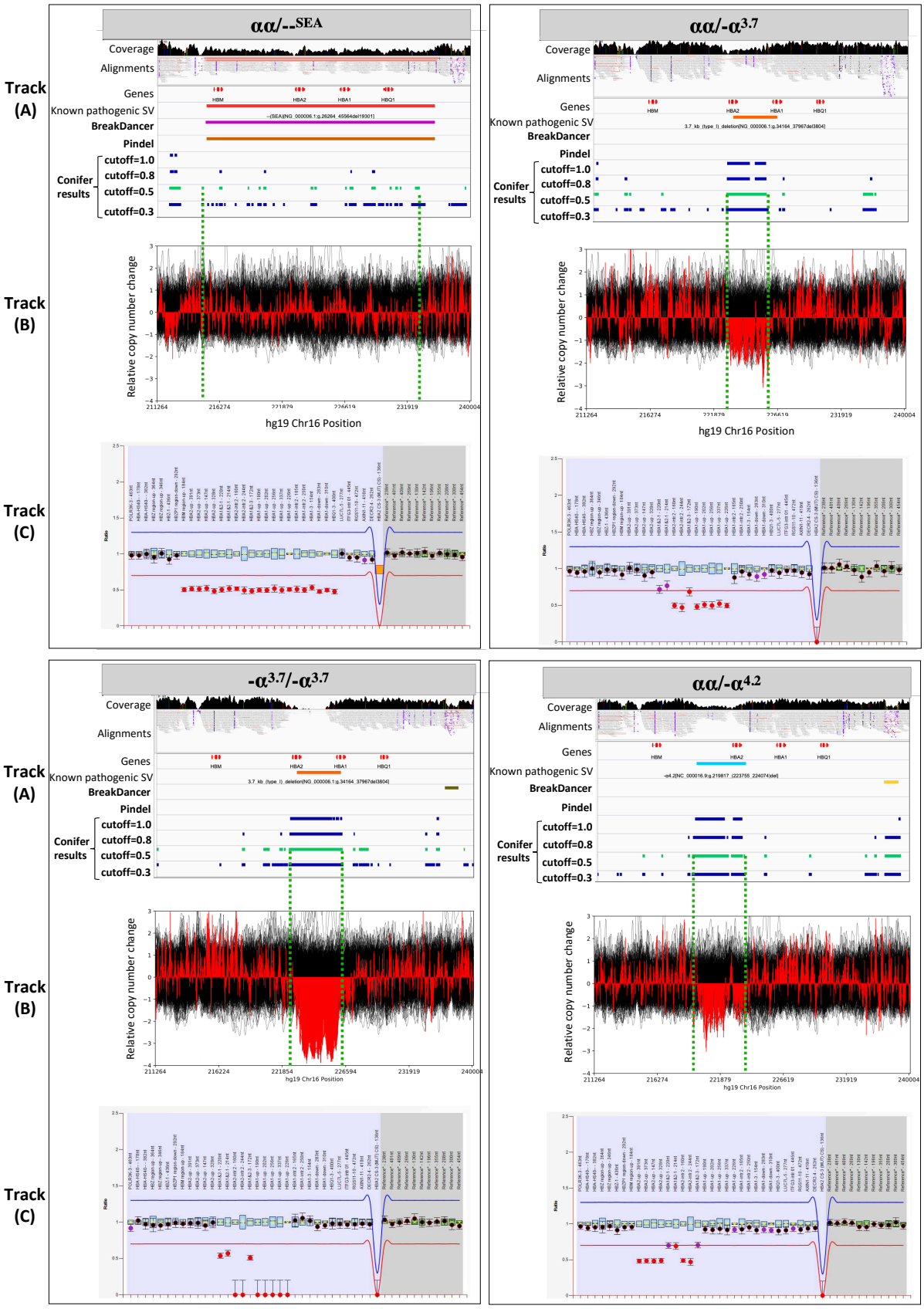


Figure S12. Detection of thalassemia pathogenic deletions from whole genome sequencing data by NGS4THAL

Notes to Figure S12: (1) Legends in track(A) and track(B) have the same meanings as those in Figure 3. **(2)** track(C) shows the confirmation of the SVs by assay of multiplex ligation-dependent probe amplification (MLPA). X-axis shows positions of the MLPA probes (different scale from Track(A) and Track(B)), and Y-axis shows the probe ratios compared to the non-deletion reference samples. The probe ratios had a normal range of 0.7 - 1.3 for non-deletion samples. Probe ratio of 0.5 indicated a single copy deletion, except for probes of 172 nt, 214 nt and 220 nt, for which a ratio of 0.75 indicates a single copy deletion, since these three probes hybridize to both *HBA1* and *HBA2*. Besides, the 135 nt probe (the one on the most right hand side of the purple regions) would only give a signal on samples containing the Hb CS point mutation.

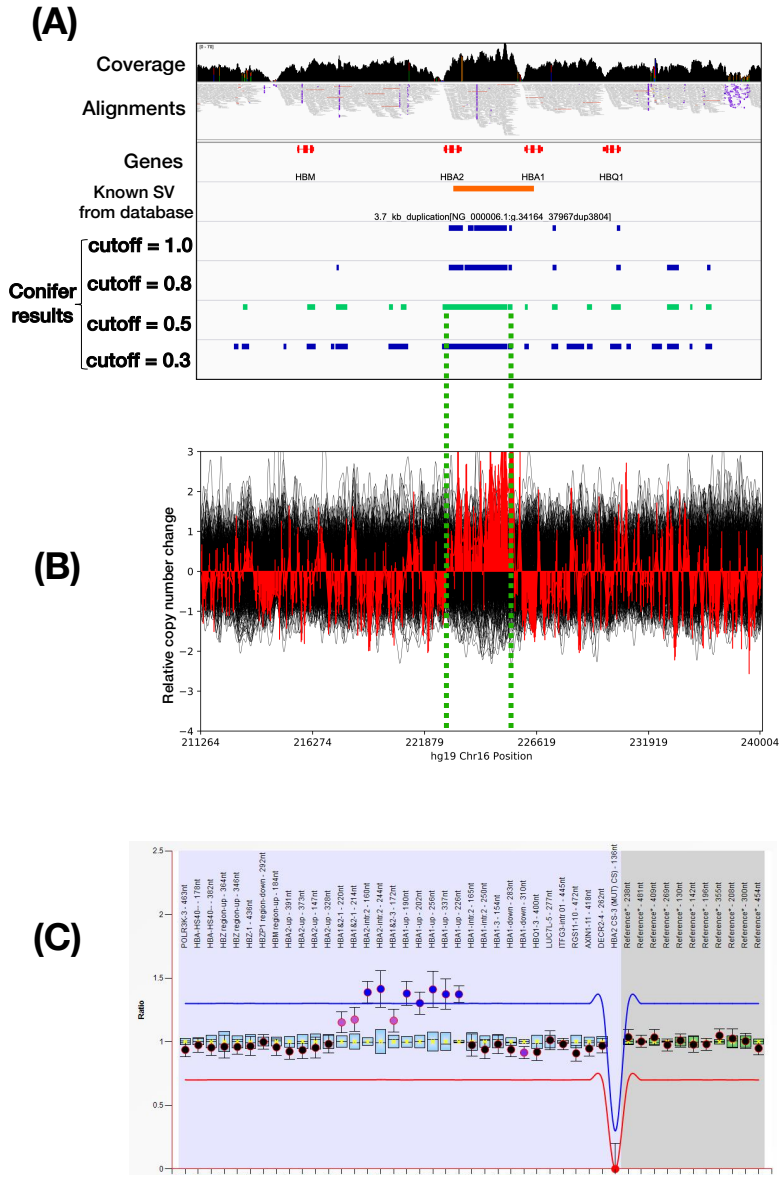


Figure S13. Detection of a heterozygous duplication from whole genome sequencing data by NGS4THAL

Notes to Figure S13: (A) Alignment landscape by IGV screenshot; (B) read-depth analysis by CoNIFER in the screening stage; (C) MLPA confirmation. Other legends are the same as those in Figure S12.

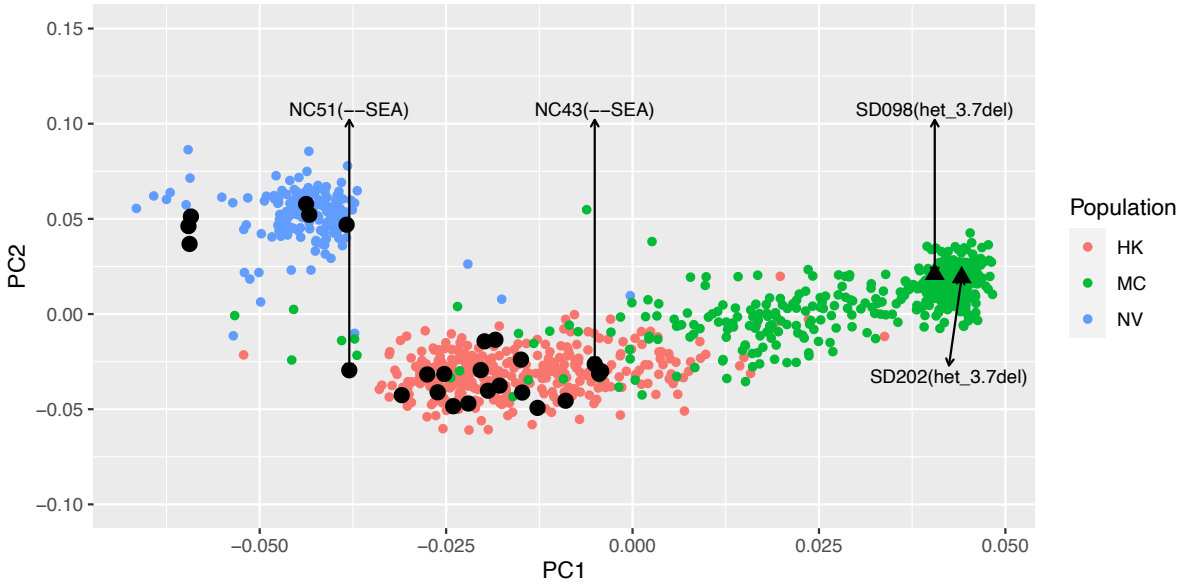
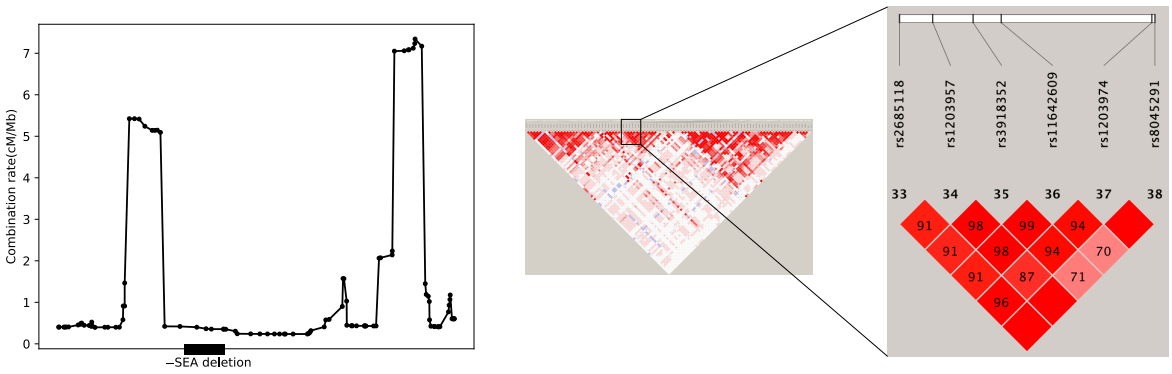


Figure S14. Genetic principal component analysis for populations collected from Hong Kong (HK), Northern North Vietnam (NV) and Mainland China (MC)

Notes to Figure S14: NC51 and NC43 are Mainland Chinese collected from the City of Guangzhou and were detected with heterozygous $--^{SEA}$ deletion; SD098 and SD202 are Northern Chinese collected from Shandong Province and were detected with heterozygous -$\alpha^{3.7}$ deletions.



(A) **(B)**

Figure S15. Local recombination rate and linkage disequilibrium patterns for the regions surrounding the --^{SEA} deletion

Notes to Figure S15: (A) Recombination rate for --^{SEA} deletion region. Recombination rate is from HapMap project; --^{SEA} deletion region is hg19:chr16:215398-234706. **(B)** LD (|D'|) landscape around --^{SEA} deletion. LD pattern determined by the selected SNPs downstream --^{SEA} is shown in the black box.