. . . .

Supplementary Tables

Table S1. Targeted regions of the NGS scheme for continuous sequencing of thalassemia relevant regions

Targeted regions	Coordinates(GRCh38)	Length(bp)
HBA gene cluster[(1)]	chr16:82,745-230,000	147,255
HBB gene cluster	chr11:5,200,000-5,300,000	100,000
KLF1	chr19:12,883,700-12,887,700	4,000
GATA1	chrX:48,785,600-48,794,800	9,200
Promoter region of UGT1A1	chr2:233,758,985-233,762,062	3,077
BCL11A coding regions	chr2:60,449,567-60,464,000 chr2:60,468,600-60,469,000 chr2:60,545,700-60,546,500 chr2:60,553,000-60,554,000	16,633
MYB coding regions	chr6:135,181,475-135,217,977	3,380
ATRX coding regions	chrX: 77,508,353-77,786,001	8,217
G6PD coding regions	chrX: 154,531,390-154,547,569	1,638
rSNPs[(2)]	chr2:60,481,361-60,481,561 chr2:60,493,000-60,493,211 chr2:60,490,808-60,490,999 chr2:60,492,735-60,492,935 chr6:135,097,780-135,097,980 chr6:135,111,314-135,111,514 chr20:508,265-508,465	1,402
Total	-	280,819

Notes to Table S1: (1) HBA gene cluster is 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); HBB gene cluster is composed of 5' - *HBE* - *HBG2*- *HBG1* - *HBD* - *HBB* - 3' along with a distal locus control regions (known as LCR). **(2)** Regulatory SNPs: rs6732518, rs11886868, rs1427407, rs766432, rs9399137, rs949414

Table S2 - part1 Simulated haemoglobin mutations (SVs)

INDEX	COMMON NAME	HBVAR_ID	RANGE(HG19)	SIZE	TYPE
1	$-\alpha^{2.7}$	HbVar.1074	chr16:225799-228500	2701	subtype_A Del
2	$--^{CANT}$	HbVar.1085	chr16:218499-231600	13101	Subtype_A Del
3	$--^{SEA}$	HbVar.1086	chr16:215399-234700	19301	Subtype_A Del
4	$--^{SA}$	HbVar.1090	chr16:208599-232200	23601	Subtype_A Del
5	$--^{MC}$	HbVar.1102	chr16:189299-232500	43201	Subtype_A Del
6	60.3 Kb deletion	HbVar.3190	chr16:147577-207873	60295	Subtype_A Del
7	Kabylian deletion-insertion	HbVar.2875	chr11:5254797-5256837	2040	Subtype_A Del
8	12,620 bp deletion	HbVar.987	chr11:5238163-5250775	12612	Subtype_A Del
9	French HPFH deletion	HbVar.2873	chr11:5246099-5265833	19734	Subtype_A Del
10	German $(A\gamma\delta\beta)^0$	HbVar.1039	chr11:5219205-5272527	53322	Subtype_A Del
11	HPFH-6	HbVar.1048	chr11:5193973-5273251	79278	Subtype_A Del
12	$-\alpha^{5.3}$	HbVar.1080	chr16:217819-223066	5247	Subtype_B Del
13	$-\alpha^{3.5}$	HbVar.1075	chr16:226599-230100	3501	Subtype_C Del
14	$--^{SPAN}$	HbVar.1084	chr16:219999-230550	10551	Subtype_C Del
15	$-\alpha^{3.7}$ (type I)	HbVar.1076	chr16:223299-227103	3804	Subtype_D Del
16	$-\alpha^{4.2}$	HbVar.1079	chr16:219816-224074	4257	Subtype_D Del
17	$\alpha\alpha/\alpha\alpha\alpha^{anti3.7}$	NA	NA	NA	Subtype_D Dup
18	$\alpha\alpha/\alpha\alpha\alpha^{anti4.2}$	NA	NA	NA	Subtype_D Dup
19	$\alpha\alpha/\alpha\alpha\alpha\alpha$	NA	NA	NA	Subtype_D Dup
20	$--^{SEA}/-\alpha^{3.7}$	NA	NA	NA	Compound Del
21	$--^{CANT}/-\alpha^{3.7}$	NA	NA	NA	Compound Del
22	$--^{SPAN}/-\alpha^{3.7}$	NA	NA	NA	Compound Del
23	$-\alpha^{5.3}/-\alpha^{3.7}$	NA	NA	NA	Compound Del
24	$--^{SEA}/\alpha\alpha\alpha^{anti3.7}$	NA	NA	NA	Compound Del Compound Dup
25	$--^{SEA}/\alpha\alpha\alpha\alpha$	NA	NA	NA	Compound Del Compound Dup

Table S2 - part2 Simulated hemoglobin mutations (SNVs and small InDels)

INDEX	CHR	POS	HBVAR_ID	REF	ALT	TYPE
1	chr16	222912	HbVar.1059	A	G	Causal SNV
2	chr16	222954	HbVar.16	T	C	Causal SNV
3	chr16	222980	HbVar.2514	C	T	Causal SNV
4	chr16	223006	HbVar.2734	G	A	Causal SNV
5	chr16	223061	HbVar.2600	G	A	Causal SNV
6	chr16	223127	HbVar.1235	G	A	Causal SNV
7	chr16	223299	HbVar.2532	A	T	Causal SNV
8	chr16	223484	HbVar.167	G	A	Causal SNV
9	chr16	223502	HbVar.169	C	A	Causal SNV
10	chr16	223547	HbVar.187	T	C	Causal SNV
11	chr16	223597	HbVar.703	T	C	Causal SNV
12	chr16	223691	HbVar.1071	A	G	Causal SNV
13	chr16	226700	HbVar.1178	T	C	Causal SNV
14	chr16	226717	HbVar.3083	T	A	Causal SNV
15	chr16	226758	HbVar.16	T	C	Causal SNV
16	chr16	226785	HbVar.1176	G	T	Causal SNV
17	chr16	226855	HbVar.2778	G	C	Causal SNV
18	chr16	226931	eqvtHbVar.1235	G	A	Same as HbVar.1235, but on HBA1
19	chr16	227011	HbVar.87	G	A	Causal SNV
20	chr16	227020	HbVar.2908	T	C	Causal SNV
21	chr16	227103	eqvtHbVar.2532	A	T	Same as HbVar.2532, but on HBA1
22	chr16	227130	HbVar.2901	A	T	Causal SNV
23	chr16	227295	eqvtHbVar.167	G	A	Same as HbVar.167, but on HBA1
24	chr16	227313	HbVar.169	C	A	Causal SNV
25	chr16	227358	eqvtHbVar.187	T	C	Same as HbVar.187, but on HBA1
26	chr16	227375	HbVar.199	T	C	Causal SNV
27	chr16	227410	eqvtHbVar.707	A	T	Same as HbVar.707, but on HBA1

Continue - Table S*2* - part2 Simulated haemoglobin mutations (SNVs and small InDels)

INDEX	CHR	POS	HBVAR_ID	REF	ALT	TYPE
28	chr16	227506	HbVar.2722	G	A	Causal SNV
29	chr11	5246838	HbVar.3164	T	C	Causal SNV
30	chr11	5246889	HbVar.960	T	C	Causal SNV
31	chr11	5246925	HbVar.495	G	T	Causal SNV
32	chr11	5246958	HbVar.940	T	C	Causal SNV
33	chr11	5247153	HbVar.889	G	A	Causal SNV
34	chr11	5247851	HbVar.880	C	A	Causal SNV
35	chr11	5247992	HbVar.853	C	A	Causal SNV
36	chr11	5248031	HbVar.2946	T	G	Causal SNV
37	chr11	5248173	HbVar.808	C	A	Causal SNV
38	chr11	5248219	HbVar.788	G	T	Causal SNV
39	chr11	5248249	HbVar.779	C	G	Causal SNV
40	chr11	5248280	HbVar.772	C	T	Causal SNV
41	chr11	5253995	HbVar.1004	C	T	Causal SNV
42	chr11	5254289	HbVar.1002	G	A	Causal SNV
43	chr11	5255289	HbVar.2777	T	A	Causal SNV
44	chr11	5255423	HbVar.998	C	T	Causal SNV
45	chr11	5255650	HbVar.995	G	A	Causal SNV
46	chr11	5255749	HbVar.994	G	T	Causal SNV
47	chr11	5255778	HbVar.992	T	C	Causal SNV
48	chr16	222912	HbVar.1062	AT	A	Causal InDel
49	chr16	222970	HbVar.2889	CG	C	Causal InDel
50	chr16	223004	HbVar.2871	GAG	G	Causal InDel
51	chr16	223141	HbVar.2726	CC	C	Causal InDel
52	chr16	223176	HbVar.2575	AGC	A	Causal InDel
53	chr16	223217	HbVar.2730	GGCCGACGCGCTGACCAACGC CGTGGCGCACGTGGACGACATG	G	Causal InDel
54	chr16	223518	HbVar.2867	CGAGTTCACCCC	C	Causal InDel
55	chr16	226777	HbVar.2800	A	AT	Causal InDel
56	chr16	226808	HbVar.2871	GAG	G	Causal InDel
57	chr16	226986	HbVar.1068	GGCTCTGCCCAGGT	G	Causal InDel
58	chr16	227078	HbVar.2504	CGCCCTGAGC	C	Causal InDel

Continue - Table S2 - part2 Simulated haemoglobin mutations (SNVs and small InDels)

INDEX	CHR	POS	HBVAR_ID	REF	ALT	TYPE
59	chr16	227377	HbVar.1142	T	TT	Causal InDel
60	chr11	5246715	HbVar.2721	TTT	T	Causal InDel
61	chr11	5246853	HbVar.1237	TTAGCCA	T	Causal InDel
62	chr11	5247900	HbVar.871	ATCACT	AA	Causal InDel
63	chr11	5247992	HbVar.851	C	CA	Causal InDel
64	chr11	5248177	HbVar.803	ACCAACTT	A	Causal InDel
65	chr11	5248206	HbVar.2870	ACAGGGCAGTAACGGCAGACTTCTCCT	A	Causal InDel
66	chr11	5255265	HbVar.1000	C	CA	Causal InDel
67	chr16	223269	HbVar.124	C	G	Hb variants
68	chr16	222973	HbVar.28	A	G	Hb variants
69	chr16	223477	HbVar.165	A	C	Hb variants

Notes to **Table S2**: Hemoglobin structural variants (SVs) (*e.g.*, deletions) classification are based on relative positions between the homologous regions and SVs; *Causal* refer to known thalassemia pathogenic.

Table S2. Demographic information of whole genome sequencing cohorts

Groups	Collecting cities/Provinces	Count	Total
Hong Kong population	Hong Kong	375	375
Northern Vietnamese population	Hanoi	161	161
Mainland Chinese population	Beijing	35	418
	Changchun	2	
	Shenyang	31	
	Shandong	186	
	Wuhan	110	
	Zhejiang	5	
	Guiyang	21	
	Shenzhen	9	
	Guangzhou	19	

Table S3. Oligonucleotide sequences used for variants confirmation by Sanger sequencing

Group	Target variants	Oligonucleotide sequence	Annealing Temp (°C)	Product Size(bp)
1	Hb Westmead (HBA2:c.369C>G)[1]	F = CCTCTCAGGGCAGAGGATCAC R = TGCAGAGAGGTCCTTGGTCT	61.64 60.18	416, 2246 [2]
2	-28 (A>G) (HBB:c.-78A>G) Hb E (HBB:c.79G>A) CD 17 (AAG->TAG)(HBB:c.52A>T) -50 G>A (HBB:c.-100G>A) CD 41/42 (-CTTT) (HBB:c.126_129delCTTT)	F = TGCTTAGAACCGAGGTAGAGTTT R = AGGTGCCCTTGAGGTTGTC	59.17 59.54	915
3	IVS-II-654 (C->T) (HBB:c.316-197C>T)	F = GTGAGTCTATGGGACGCTTGA R = TTGGACAGCAAGAAAGCGAG	59.52 58.77	1000
4	-30 T>C (HBD:c.-80T>C) -77 (T->C) (HBD:c.-127T>C)	F = TTCCAACCTCTCAAAATTCCCTTG R = CTGAGGGTAGGAAAACAGCCCAA	59.41 62.40	711

Notes to Table S3: (1) Hb CS (HBA2: c.427T>C) was confirmed by MLPA (Holland SALSA MLPA P140 HBA probemix for α-thalassemia *in vitro* diagnostic assay) in which a probe was specifically designed for Hb CS; (2) extension time of the PCR amplification was limited to 40 seconds to avoid the unintended products longer than 500bps.

Table S4. Discovery samples for --^{SEA} associated haplotype analysis

Datasets	--^{SEA} carrier counts	non --^{SEA} control counts
SLE GWAS cohort	0	2, 587
Targeted sequencing cohort	15	0
Whole genome sequencing cohort	25	929
Total	40	3, 516

Table S5. --^{SEA} heterozygotes$^{(1)}$ from 1000 Genome project

Index	Sample name	Sex	Population$^{(2)}$	Sample count from the population
1	HG00684	female	CHS	64
2	HG00690	female	CHS	
3	HG00699	female	CHS	
4	HG00728	male	CHS	
5	HG00729	female	CHS	
6	HG00864	female	CDX	109
7	HG00978	female	CDX	
8	HG01794	female	CDX	
9	HG01805	female	CDX	
10	HG01807	female	CDX	
11	HG01812	female	CDX	
12	HG01816	male	CDX	
13	HG02151	female	CDX	
14	HG02179	female	CDX	
15	HG02181	female	CDX	
16	HG02182	female	CDX	
17	HG02353	male	CDX	
18	HG02374	male	CDX	
19	HG02397	male	CDX	
20	HG02410	male	CDX	
21	HG01846	male	KHV	124
22	HG02028	female	KHV	

Total: --^{SEA} carriers count 22, and 381 non_--^{SEA} individuals

Notes to Table S5: **(1)** They were all founders from 1000G phase 3 release; --^{SEA} = esv3637547 in the vcf file. **(2)** CHS: Southern Han Chinese; CDX: Dai Chinese; KHV: Kinh Vietnamese

Table S6 Summary table of detected thalassemia pathogenic variants by NGS4THAL from whole genome sequencing cohorts

Hong Kong Chinese population (N=375)			Northern Vietnamese population (N=161)		
Pathogenic variants	**Carrier counts**	**Percentage**	**Pathogenic variants**	**Carrier counts**	**Percentage**
Total thalassemia mutation[1]	44	11.7	**Total thalassemia mutation**	24	14.9
α-thalassemia mutation	36	9.6	**α-thalassemia mutation**[3]	12	7.5
(1) $--^{SEA}$	17[2]	4.5	(1) $--^{SEA}$	6	3.7
(2) $-\alpha^{3.7}$	12	3.2	(2) $-\alpha^{3.7}$	5[4]	3.1
(3) $-\alpha^{4.2}$	2	0.5	(3) Hb Westmead	1	0.6
(4) Hb CS	2[2]	0.5			
(5) Hb Westmead	3	0.8			
β-thalassemia mutation	7	1.9	**β-thalassemia mutation**	11	6.8
(1) CD 41/42 (-CTTT)	6	1.6	(1) HbE	5[4]	3.1
(2) -28 (A>G)	1	0.3	(2) CD 17 (AAG>TAG)	3	1.9
			(3) -50 G>A	3	1.9
δ-thalassemia mutation	1	0.3	**δ-thalassemia mutation**	1	0.6
(1) -30 T>C	1	0.3	(1) -77 (T>C)	1	0.6

Notes to Table S6: (1) --^{SEA}: NG_000006.1: g.26264_45564del19301, -$\alpha^{3.7}$: NG_000006.1: g.34164_37967del3804, -$\alpha^{4.2}$: NC_000016.9:g.219817_(223755_224074)del, HGVS names for other variants listed are the same as those in **Error! Reference source not found.**. **(2)** Include a sample from compound heterozygote of --^{SEA}/Hb CS, and each of the mutation haplotype was counted as 1. **(3)** Include a carrier with both α- and β-thalassemia mutations, and each of the mutation was counted. **(4)** A homozygote was counted as 2.

Table S7. Thalassemia pathogenic variants detected by NGS4THAL from a Mainland Chinese population with whole genome sequencing data (sample size = 418)

Index	Sample_name	--$^{SEA(1)}$	-$\alpha^{3.7}$_region	-$\alpha^{4.2}$_region	Point_Mutation	Small_InDel
1	NC104		het del			
2	SD028		het del			
3	SD098		het del			
4	SD086		het del			
5	SD202		het del			
6	SD111				-77 (T>C)	
7	SD191				IVS-II-654 (C>T)	
8	NC42				-50 G>A$^{(1)}$	
9	NC58				-28 (A>G)	
10	NC43	het$^{(2)}$				
11	NC51	het				
12	NC65		het del			
13	NC10		dup$^{(2)}$			
14	NC187		dup			
15	NC225		dup			
16	SD017		dup			
17	SD102		dup			

Notes to Table S8: (1) HGVS names for variants listed are the same as those in **Error! Reference source not found.** and Table S6; (2) het: heterozygote, del: deletion, dup: duplication; duplications are excluded when calculating thalassemia pathogenic variants carrier rate; (3) samples highlighted in grey were those collected from North China, including Beijing, Jilin Province(Changchun City), Liaoning Province(Shenyang City) and Shandong Province.

Table S8. Sanger sequencing confirmation for representative pathogenic point mutations detected by NGS4THAL

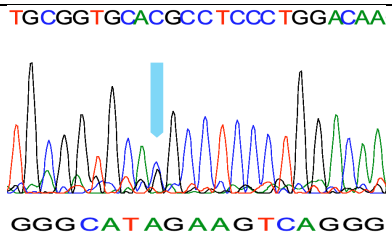
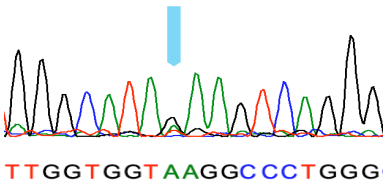
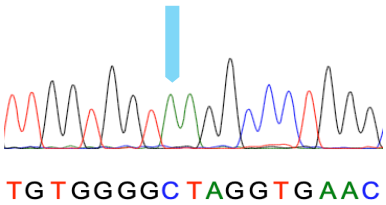
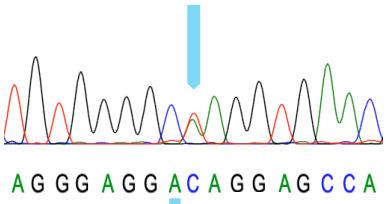
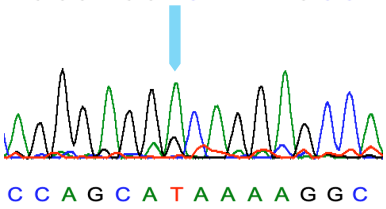
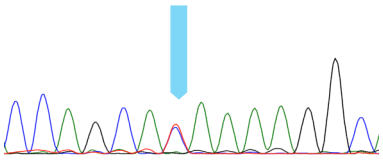
Index	Target variants	Sanger results
1	Hb Westmead (HBA2:c.369C>G)	
2	-28 (A>G) (HBB:c.-78A>G)	
3	Hb E (HBB:c.79G>A)	
4	CD 17 (AAG->TAG)(HBB:c.52A>T)	
5	-50 G>A (HBB:c.-100G>A)	
6	-30 T>C (HBD:c.-80T>C)	

Table S9. Thalassemia pathogenic variants detected by NGS4THAL from Hong Kong WES cohort (sample size = 185)[1]

Index	Sample_name	$--^{SEA}$	$-\alpha^{3.7}$_region	$-\alpha^{4.2}$_region	Point_Mutation	Small_InDel
1	A090512				Hb Westmead[2]	
2	A190007				Hb CS	
3	A190048		het del		-28 (A>G)	
4	A170537					CD 41/42 (-CTTT)
5	35A					CD 41/42 (-CTTT)
6	H12086				IVS-II-654 (C>T)	
7	A170405				IVS-II-654 (C>T)	
8	A180267	het				
9	A170693	het				
10	H12020	het				
11	PID12-092				Hb Owari[3]	
12	PID16-196				Hb Chesapeake[2]	

Notes to Table S10: (1) Whole exome sequencing data were primarily generated for primary immunodeficiency disorder; (2) HGVS names for variants listed are the same as those in **Error! Reference source not found.** and Table S6, het: heterozygote, del: deletion; (3) Hb Owari and Hb Chesapeake are α-chain variants and are excluded when calculating thalassemia pathogenic variants carrier rates.

Table S10. Detection of structural variants by NGS4THAL on simulation data

SV types	Deletion		Alpha_Duplication		non SV
Compound status	Non-compound	Compound[(1)]	Non-compound	Compound[(2)]	NA
Positive cases	160	60	30	20	1750
Report positive	160	59	27	20	0
Sensitivity for this category	100.0%	98.3%	90.0%	100.0%	NA
Total Sensitivity	(160+59+27+20) / (160+60+30+20) = 98.4%				
Total Specificity	1750 / 1750 = 100.0%				

Notes to Table S11: (1) Compound deletions include: $--^{SEA}/-\alpha^{3.7}$, $--^{CANT}/-\alpha^{3.7}$, $--^{SPAN}/-\alpha^{3.7}$, $-\alpha^{5.3}/-\alpha^{3.7}$, $--^{SEA}/\alpha\alpha\alpha^{anti3.7}$, $--^{SEA}/\alpha\alpha\alpha\alpha$; (2) Compound Alpha_Duplications include: $--^{SEA}/-\alpha\alpha\alpha^{anti3.7}$, $--^{SEA}/\alpha\alpha\alpha\alpha$; other simulated non-compound SVs can be found in **Error! Reference source not found.**.

Table S11. Confirmation of representative α-thalassemia SVs detected by NGS4THAL using MLPA

Groups	SVs	HKC[1]	NV	MC	Sub_Total
1	$--^{SEA}$ deletion	2[2]	1	1	4
2	$-\alpha^{3.7}$ deletion	1	2[3]	2	5
3	$-\alpha^{4.2}$ deletion	2	0	0	2
4	Duplications	3	0	5	8
5	Reference baseline controls	1	1	1	3
6	A positive control	A heterozygous $-\alpha^{3.7}$ carrier			1
7	Negative control	TE0.1 (no DNA sample)			1
					Total = 24

Notes to Table S12: (1) HKC: Hong Kong Chinese population, NV: Northern Vietnamese population, MC: Mainland Chinese population; (2) one is $\alpha\alpha/--^{SEA}$, and the other one is $\alpha\alpha/--^{SEA}$ with Hb CS point mutation; (3) one is heterozygote, and the other one is homozygote.

Table S12. Inferred haplotypes from the discovery samples for --^{SEA} associated haplotype analysis

Index	Haplotype	Freq_Case	S.E.(case)	Freq_Control	S.E.(control)
1	GGGCGA[(1)]	0.200	0.001[(3)]	0.345	0.001
2	GGATAA	0.170	0.009	0.313	0.001
3[(2)]	**GGATGA**	**0.530**	**0.009**	**0.129**	**0.001**
4	ATATAA	0.048	0.006	0.118	0.001
5	GGATAG	0.025	0.000	0.052	0.001
6	GTATAA	0.008	0.006	0.020	0.001
7	GTATGA	0.005	0.006	0.007	0.001

Notes to Table S12: **(1)** The selected SNPs are in the following order: rs2685118, rs1203957, rs3918352, rs11642609, rs1203974, rs8045291. **(2)** The haplotype highlighted in bold was carried by all the heterozygous --^{SEA} carriers. **(3)** Only haplotypes with frequencies greater than 0.005 in cases or controls are listed here.

Table S13 Contingency table of a risk haplotype(1) carriers and the presence of --^{SEA} deletion

Discovery cohort				**Replication cohort**			
	Carry --^{SEA}	non --^{SEA}	Total		Carry --^{SEA}	non --^{SEA}	Total
Carry risk haplotypes	39	901	940	Carry risk haplotypes	21	59	80
Not carry risk haplotypes	1	2,615	2,616	Not carry risk haplotypes	1	322	323
Totals	40	3,516	3,556	Totals	22	381	403
P-value	< 2.2e-16			*P-value*	3.20e-15		
Odds Ratio	113.19			Odds Ratio	114.61		
Sensitivity	97.50%			Sensitivity	95.45%		
Specificity	74.37%			Specificity	84.51%		

Notes to Table S14: The haplotype is defined as G (rs2685118) - G (rs1203957) - A (rs3918352) - T (rs11642609) - G (rs1203974) - A (rs8045291).