

**ANALYSIS OF THALASSEMIA VARIANT
SPECTRUM AND DEVELOPMENTAL
VALIDATION OF EXOME SEQUENCING FOR β -
THALASSEMIA IN MALAYSIA**

SYAHZUWAN BIN HASSAN

UNIVERSITI SAINS MALAYSIA

2025

**ANALYSIS OF THALASSEMIA VARIANT
SPECTRUM AND DEVELOPMENTAL
VALIDATION OF EXOME SEQUENCING FOR β -
THALASSEMIA IN MALAYSIA**

by

SYAHZUWAN BIN HASSAN

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

December 2025

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful

Thank you to Allah, and Alhamdulillah, for His will, this project and dissertation are completed. This thesis is dedicated to my family, Irnie Yuslele Tamam, Iman Nur Jannah, Hassan Yunus, and Mariati Kanapi. I express my sincerest appreciation to my supervisors, Dr. Zefarina Zulkafli, Professor Dr. Wan Zaidah Abdullah, Associate Professor Dr. Rosnah Bahar, Dr. Ezzeddin Kamil Mohamed Hashim, Associate Professor Dr. Muhammad Farid Johan, and Dr. Ezalia Esa, for their supervision, suggestions, and advice during the study and dissertation writing.

I also thank Dr. Khairul Nizam and my colleagues at the MGM, Mrs. Faidatul Syazlin, Mrs. Syahira, Mrs. Velintina, Mrs. Raihan, Ms. Haifa, Mrs. Fyqa, Ms. Mira, Mrs. Fatin, Mrs. Mell, Mrs. Fika, Mr. Bakri, Mr. Muhammad, Mrs. Suffya, and Dr. Nur Aisyah at the Hematology Unit, Cancer Research Centre of the Institute for Medical Research, for their assistance during my study. I thank the examiners, AP. Dr. Syahrilnizam, AP. Dr. Noor Haslina, and AP. Dr. Edinur Hisham; I also thank Pn. Kamarulasikin, and everyone at HUSM for the administrative support.

I thank my parents, Hassan Yunus and Mariati Kanapi, my wife and child, for their support and understanding during the study period. Finally, thank you to Dr. Zubaidah, Dr. Rahimah, Dr. Yuslina, Dr. Norafiza, Dr. Ermi, Dr. Azian, and Dr. Nurul. May the blessings and peace of Allah be upon our Prophet Muhammad ﷺ the last of the Messengers, and his Companions. All praise is to Allah, the Lord of all worlds.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	ii
TABLE OF CONTENTS.....	iii
LIST OF TABLES	xiii
LIST OF FIGURES	xx
LIST OF SYMBOLS	xxix
LIST OF ABBREVIATIONS	xxx
LIST OF APPENDICES	xxxvii
ABSTRAK	xxxix
ABSTRACT	xli
CHAPTER 1 INTRODUCTION.....	1
1.1 Research Background.....	1
1.2 Problem Statement	3
1.2.1 Heterogeneity of Thalassemia Mutations, Ambiguous Differential Diagnoses, and Diagnostic Challenges.....	3
1.2.2 Labor-Intensive Testing, and DNA Test Limitations.....	4
1.2.3 Need for a Unified DNA Analysis Strategy.....	5
1.3 Study Motivation.....	5
1.3.1 Understanding the Genotype-Phenotype Relationship and Hematological Parameter Evaluation for β -thalassemia Carrier Prediction	5
1.3.2 Determining the Spectrum of β -thalassemia and Variants.....	6
1.3.3 Bioinformatics Pipeline Assembly for Unified SNV, Indels, and CNV Analyses of Thalassemia.....	7
1.4 Hypotheses and Research Questions.....	8
1.4.1 Hypothesis.....	8
1.4.2 Research Questions	9
1.5 Objectives.....	9

1.5.1	General Objective.....	9
1.5.2	Specific Objectives.....	9
CHAPTER 2 LITERATURE REVIEW		12
2.1	Red blood cell.....	12
2.2	Hemoglobin.....	13
2.2.1	Hemoglobin Structure and Switching	14
2.3	Thalassemia and hemoglobinopathies.....	16
2.4	<i>HBB</i> gene and cluster mutations	17
2.4.1	β -globin hemoglobinopathies.....	17
2.4.2	β -thalassemia.....	17
2.4.3	HPFH and $\delta\beta$ -thalassemia.....	18
2.4.4	δ -thalassemia and variants.....	19
2.5	<i>HBA</i> gene and cluster mutations	19
2.5.1	α -thalassemia and variants	19
2.6	Thalassemia Phenotypes	20
2.6.1	α -thalassemia silent carrier, trait, HbH, and Hb Bart's	20
2.6.2	β -thalassemia trait, intermedia, and major	21
2.6.3	Non-transfusion Dependent Thalassemia (NTDT) and Transfusion-dependent Thalassemia (TDT).....	22
2.6.4	Genetic Modifiers of thalassemia.....	23
	2.6.4(a) Genetic Modifier of β -thalassemia	23
	2.6.4(b) Genetic Modifier of α -thalassemia	25
2.7	Conventional DNA Analysis.....	26
2.7.1	Reverse Dot-Blot Analysis.....	27
2.7.2	Gap-PCR	27
2.7.3	Amplification Refractory Mutation System (ARMS) or Allele-Specific Polymerase Chain Reaction (ASPCR).....	27
2.7.4	Multiplex Ligation Probe-Dependent (MLPA) Analysis.....	28

2.7.5	Sanger Sequencing	29
2.8	Screening and prevention of thalassemia via DNA analysis in Malaysia	32
2.9	Routine Diagnosis of Thalassemia and Variants in Malaysia	33
2.9.1	Molecular diagnosis of thalassemia in Malaysia	34
2.10	Local DNA analysis programs	37
2.10.1	Cascade Screening	37
2.10.2	Sixteen-year-old student screening	37
2.10.3	β -thalassemia carrier screening	37
2.11	High-throughput Sequencing (HTS)	38
2.11.1	Next Generation Sequencing	38
2.11.2	Third-Generation Sequencing (TGS)	40
2.12	Bioinformatics	42
2.13	Bioinformatics analysis of NGS data	43
2.13.1	Datasets Pre-processing	44
2.13.2	Germline Variant Discovery	44
2.13.3	Copy Number Variant (CNV) Discovery	45
2.13.4	Web-based NGS Analysis	47
2.13.5	NGS Applications in thalassemia diagnosis	48
CHAPTER 3	METHODOLOGY	51
3.1	Materials	51
3.1.1	Research design	51
3.1.2	Sample and Data Collection	52
3.1.3	Study subjects	53
3.1.4	Sample collection	53
3.1.5	Sample size	53
3.1.6	Sample selection criteria	55
3.1.6(a)	Inclusion criteria:	55

3.1.6(b)	Exclusion criteria:	55
3.1.7	Sample selection criteria for Whole Exome Sequencing (WES) and bioinformatic analysis	56
3.1.8	Instrumentation and Laboratory	57
3.2	Methods	57
3.2.1	Conventional DNA Analysis of of <i>HBB</i> mutations	57
3.2.1(a)	β -Multiplex ARMS (β -MARMS).....	58
3.2.1(b)	β -globin sequencing.....	60
3.2.1(c)	<i>HBB</i> cluster GAP-PCR	61
3.2.1(c)(i)	β -Multiplex GAP (β -MGAP).....	61
3.2.1(c)(ii)	β -Multiplex GAP 2 (β -MGAP 2) and $\gamma\delta\beta$ Asian-Indian Deletion/Inversion β -GAP	64
3.2.1(c)(iii)	Hb Lepore GAP	68
3.2.1(d)	β -globin MLPA (β -MLPA)	71
3.2.1(e)	δ -globin sequencing.....	72
3.2.2	Conventional DNA Analysis of <i>HBA</i> mutations.....	73
3.2.2(a)	α -Multiplex GAP (α -MGAP)	73
3.2.2(b)	α -Multiplex ARMS (α -MARMS).....	76
3.2.2(c)	α -triplications.....	79
3.2.2(d)	α -globin sequencing.....	80
3.2.2(e)	α -globin MLPA (α -MLPA)	82
3.3	Statistical analysis of the hematological parameters.....	83
3.3.1	Kruskal-Wallis test.....	83
3.3.2	Logistic Regression (LR) Analysis	84
3.4	Whole Exome Sequencing library preparation, sequencing, and bioinformatics analysis.....	85
3.4.1	DNA Qualitative and Quantitative Assessment for the Whole Exome Sequencing (WES).....	85
3.4.2	DNA Library Preparation for the WES.....	86

3.4.2(a)	Genomic DNA Fragmentation.....	87
3.4.2(b)	Adaptor-tagged Library Purification	88
3.4.2(c)	Adaptor-tagged Library Amplification.....	88
3.4.2(d)	Amplified Adaptor-tagged Library Purification	89
3.4.2(e)	Library DNA Quality and Quantity Assessment	90
3.4.2(f)	Hybridization and Capture.....	91
3.4.2(f)(i)	Hybridize DNA samples to the Capture Library	91
3.4.2(f)(ii)	Capture the hybridized DNA using streptavidin-coated beads.....	91
3.4.2(g)	Indexing and Sample Processing for Multiplexed Sequencing.....	93
3.4.2(g)(i)	Amplify the captured libraries to add index tags	93
3.4.2(g)(ii)	Purify the amplified captured libraries using AMPure XP beads.....	94
3.4.2(g)(iii)	Assess indexed library DNA quantity and quality.....	95
3.4.2(h)	Sample pooling, clustering, and sequencing.....	96
3.5	Bioinformatics analysis of the WES	96
3.5.1	Pre-processing of the raw sequence (FASTQ) data	97
3.5.2	Germline variants calling	99
3.5.3	Callset/Genotype Refinement and VCF annotation and prioritization	101
3.5.4	Copy Number Variant (CNV) Analysis	103
3.5.4(a)	CNV analysis by CNVkit	104
3.5.4(b)	CNV analysis by Control-Freec.....	106
3.5.4(c)	CNV analysis by cn.mops.....	106
3.5.4(d)	Performance metrics of the CNV tools.....	106
3.5.5	Breakpoint analysis of the novel duplications and deletions on the α -globin cluster.....	107
3.5.5(a)	Identification of $\alpha\alpha^{102}$ duplication on sample 168_13	107

3.5.5(b)	Identification of $\alpha\alpha\alpha^{91}$ duplication on sample 1006_16 and 61_14.....	110
3.5.5(c)	Breakpoint analysis of the -- ^{GB} deletion.....	113
CHAPTER 4	RESULTS.....	115
4.1	Samples' data description.....	115
4.2	Exome sequencing exploratory data analysis.....	124
4.2.1	Raw and trimmed FASTQ quality scores and GC percentages ...	124
4.2.2	BAM QC	126
4.3	Family's at-risk status	128
4.4	Classification based on the thalassemia genotype.....	130
4.5	Frequency of α -thalassemia and α -globin cluster duplication.....	133
4.5.1	α - and β -thalassemia and variant mutation spectrum and frequency.....	136
4.5.1(a)	Frequency of α -thalassemia and α -globin cluster duplication among Peninsular Malaysians	136
4.5.1(b)	Frequency of heterozygous carrier of thalassemia in the <i>HBA</i> cluster in Sabah	139
4.6	Frequency of β -, δ -thalassemia and variants, HPFH, and $\delta\beta$ -thalassemia, and the carriers with concomitant α -thalassemia	139
4.6.1	Frequency of heterozygous β -thalassemia, δ -thalassemia and variants, HPFH, and $\delta\beta$ -thalassemia carriers in Peninsular Malaysia	144
4.6.1(a)	Frequency of heterozygous β -thalassemia, δ -thalassemia and variants, HPFH, and $\delta\beta$ -thalassemia carriers (without α -thalassemia) in Peninsular Malaysia.....	144
4.6.1(b)	Frequency of concomitant α -thalassemia among the β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia carriers in Peninsular Malaysia....	148
4.6.2	Frequency of heterozygous β -thalassemia, δ -thalassemia and variants, HPFH, and $\delta\beta$ -thalassemia carriers in Sabah	152
4.6.2(a)	Frequency of heterozygous β -thalassemia, δ -thalassemia and variants, HPFH, and $\delta\beta$ -thalassemia (without α -thalassemia) carriers in Sabah.....	152

4.6.2(b)	Frequency of concomitant α -thalassemia among heterozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia carriers in Sabah.....	154
4.6.3	Frequency of heterozygous β -thalassemia and variant carriers in Sarawak.....	155
4.6.3(a)	Frequency of heterozygous β -thalassemia and variant carriers (without α -thalassemia) in Sarawak ...	155
4.6.3(b)	Frequency of concomitant α -thalassemia in β -thalassemia carriers in Sarawak.....	156
4.7	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia	156
4.7.1	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia in Peninsular Malaysia	159
4.7.1(a)	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia (without α -thalassemia) in Peninsular Malaysia.....	159
4.7.1(b)	Frequency of concomitant α -thalassemia among compound heterozygous and homozygous β -thalassemia subjects in Peninsular Malaysia	163
4.7.2	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia in Sabah	165
4.7.2(a)	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, and δ -thalassemia (without α -thalassemia) in Sabah.....	165
4.7.2(b)	Frequency of concomitant α -thalassemia in homozygous and compound heterozygous β -thalassemia and variant subjects in Sabah	168
4.7.3	Frequency of compound heterozygous and homozygous β -thalassemia and variants in Sarawak.....	169
4.7.3(a)	Frequency of homozygous and compound heterozygous β -thalassemia and variants subjects in Sarawak.....	169
4.7.3(b)	Frequency of concomitant α -thalassemia in homozygous and compound heterozygous β -thalassemia and variants subjects in Sarawak	169

4.8	Mutation frequencies across Peninsular Malaysia, Sabah, and Sarawak	170
4.9	Descriptive and inferential statistical analysis of the genotype groups' hematological parameters	175
4.9.1	Removal of Outliers	175
4.9.2	Exploratory Data Analysis (EDA)	176
4.9.2(a)	Descriptive statistics	178
4.9.2(b)	Test for Normality	181
4.9.2(b)(i)	Red Blood Cell (RBC) ($10^{12}/L$)	181
4.9.2(b)(ii)	Hemoglobin (Hb) (g/dL)	187
4.9.2(b)(iii)	Mean Corpuscular Volume (MCV) (fL)	191
4.9.2(b)(iv)	Mean Corpuscular Hemoglobin (MCH) (pg)	195
4.9.2(b)(v)	Red Cell Distribution Width (RDW) (%)	199
4.9.2(b)(vi)	Hemoglobin A ₂ (HbA ₂) (CE) (%)	203
4.9.2(b)(vii)	Hemoglobin F (HbF) (CE) (%)	207
4.9.2(b)(viii)	Variants (CE) (%)	211
4.9.2(b)(ix)	Hemoglobin A ₂ (HbA ₂) (HPLC) (%)	214
4.9.2(b)(x)	Hemoglobin F (HbF) (HPLC) (%)	218
4.9.3	Kruskal-Wallis test	222
4.9.3(a)	Red Blood Cell Count ($10^{12}/L$)	222
4.9.3(b)	Hemoglobin (g/dL)	224
4.9.3(c)	Mean Corpuscular Volume (fL)	227
4.9.3(d)	Mean Corpuscular Hemoglobin (pg)	230
4.9.3(e)	Red Cell Distribution Width (%)	233
4.9.3(f)	Capillary Electrophoresis's HbA ₂ (%)	236
4.9.3(g)	Capillary Electrophoresis's HbF (%)	239
4.9.3(h)	Capillary Electrophoresis's Hb Variant (%)	242
4.9.3(i)	HPLC's HbA ₂ (%)	244
4.9.3(j)	HPLC's HbF (%)	247
4.9.3(k)	Summary statistic of Kruskal Wallis test and Post hoc analysis between Normal and other genotype group	250

4.9.4	Model evaluation by Logistic Regression (LR) analysis	253
4.9.4(a)	Exploratory Data Analysis (EDA) before Logistic Regression.....	253
4.9.4(a)(i)	Summary statistics	253
4.9.4(a)(ii)	Correlation between variables.....	254
4.9.4(a)(iii)	Distribution of variables	254
4.9.4(b)	Data partition	259
4.9.4(c)	Model evaluation	259
4.9.4(d)	Model performance.....	261
4.9.4(e)	Cross-validation of the models	266
4.10	Concordance of the SNV and indels detection.....	270
4.11	Validation of CNV	279
4.11.1	Concordance, validation, and breakpoint analysis of the CNV ...	279
4.11.1(a)	Novel $\alpha\alpha\alpha^{102}$ duplication breakpoint analysis of Case 168_13.....	279
4.11.1(b)	Novel $\alpha\alpha\alpha^{91}$ duplication breakpoint analysis of Case 1006_16 and 61_14.....	284
4.11.1(c)	Breakpoint analysis of the $--^{GB}$ deletion.....	291
4.11.1(c)(i)	Case 1437_17.....	291
4.11.1(c)(ii)	Case 1043_18.....	294
CHAPTER 5 CONCLUSION AND FUTURE RECOMMENDATIONS....		298
5.1	Spectrum of β -thalassemia mutations in Malaysia.....	298
5.1.1	Population estimates.....	298
5.1.2	β -thalassemia mutation spectrum in Peninsular Malaysia, Sabah, and Sarawak.....	298
5.2	Correlation of the hematological parameters among cases with β -globin gene mutations	304
5.2.1	The correlation of RBC, Hb, MCV, MCH, RDW, HbA2, HbF, and the variant value among the various genotype groups.....	304
5.3	The logistic regression comparing several models to predict the outcome of heterozygous β -thalassemia.....	307
5.3.1	Parameter selections and Models tested.....	307

5.3.2	Model of choice.....	309
5.4	Bioinformatics workflow, SNV, indels, and CNV calling validations	309
5.4.1	WES CNV performance assessment.....	309
5.4.1(a)	Performance of breakpoint estimations by CNVkit, Control-Freec, and cn.mops.....	315
5.4.2	SNV, indels, CNV concordance and NGS performance and advantage.....	317
5.5	Study strengths and limitations	321
5.5.1	Strengths.....	321
5.5.2	Limitations	322
CHAPTER 6 CONCLUSION AND FUTURE RECOMMENDATIONS....		325
6.1	Conclusion.....	325
6.2	Recommendations for Future Research	326
6.2.1	Epidemiological studies of β -thalassemia in Sarawak, Perlis, Terengganu, and Kelantan,.....	326
6.2.2	Genomic coverage for Targeted Sequencing (TS) of thalassemia	327
6.2.3	Tailored Enrichment for Thalassemia Diagnostics: Leveraging New References and Sequencing Technologies	328
REFERENCES.....		330
APPENDICES		
LIST OF PUBLICATIONS		

LIST OF TABLES

	Page
Table 2.1 Types of mutations and their effects, as well as the commonly genotyped β -thalassemia and variants among Malaysians	20
Table 3.1 Sample size for β -thalassemia spectrum analysis and WES	55
Table 3.2 Primer details in the β -MARMS for detection of the five <i>HBB</i> mutations	59
Table 3.3 PCR composition for β -MARMS	60
Table 3.4 Primer details in the β -globin sequencing.....	61
Table 3.5 PCR reaction composition of <i>HBB</i> before sequencing	61
Table 3.6 Primer details in the β -MGAP for detection of the eight <i>HBB</i> cluster deletions.....	63
Table 3.7 PCR composition for the β -MGAP	64
Table 3.8 Primer details in the β -MGAP 2 for detection of the HPFH-3, Chinese 100 kb, and Asian 49.3 kb deletions	66
Table 3.9 PCR composition for the β -MGAP 2	66
Table 3.10 Primer details in the $\gamma\delta\beta$ Asian-Indian Deletion/Inversion β -GAP ...	67
Table 3.11 Primer combinations and the amplicon sizes of the $\gamma\delta\beta$ Asian-Indian Deletion/Inversion.....	68
Table 3.12 PCR composition for the $\gamma\delta\beta$ Asian-Indian Deletion/Inversion β -GAP	68
Table 3.13 Primer details of Hb Lepore GAP	70
Table 3.14 PCR composition for the Hb Lepore GAP.....	71
Table 3.15 Thermal cycler settings for the β -MLPA	72
Table 3.16 Primer details in the δ -globin sequencing.....	73
Table 3.17 PCR reaction composition of <i>HBD</i> before sequencing.....	73

Table 3.18	Primer details in the α -MGAP for detection of the common deletional α -thalassemia.....	75
Table 3.19	PCR composition for the α -MGAP	76
Table 3.20	Primer details in the α -MARMS for detection of the common non-deletional α -thalassemia.....	78
Table 3.21	PCR composition for the α -MARMS.....	79
Table 3.22	Primer details in the α -triplications	80
Table 3.23	PCR composition for the α -triplications	80
Table 3.24	Primer details in the <i>HBA1</i> and <i>HBA2</i> sequencing	81
Table 3.25	PCR reaction composition of <i>HBA1</i> and <i>HBA2</i> before sequencing...	81
Table 3.26	Thermal cycler settings for the α -MLPA	83
Table 3.27	Preparation of pre-capture PCR Reaction mix.....	89
Table 3.28	Thermal cycler program for pre-capture PCR	89
Table 3.29	Thermal cycler program for hybridization.....	91
Table 3.30	Post-Capture PCR Cycling Program.....	94
Table 3.31	Primers used for the identification of $\alpha\alpha\alpha^{102}$ duplication	108
Table 3.32	PCR reaction composition for the identification of $\alpha\alpha\alpha^{102}$ duplication.....	108
Table 3.33	Primers used for the identification of $\alpha\alpha\alpha^{91}$ duplication.....	111
Table 3.34	PCR reaction composition for the identification of $\alpha\alpha\alpha^{91}$ duplication.....	111
Table 3.35	Primers used for the identification of $--^{GB}$ deletion.....	113
Table 3.36	PCR reaction composition for the identification of $--^{GB}$ deletion	114
Table 4.1	The number of hospitals, samples, and ethnicities in each state. The table describes the frequencies of a single ethnicity and dual ethnicity assigned to intermarriage children	117
Table 4.2	The frequency of samples with single ethnicity in each state	119

Table 4.3	The frequency of children of intermarriage bearing parental ethnicity.....	121
Table 4.4	Summary of available and missing FBC, Hb analysis, age, and ethnicity data in all samples according to the state	123
Table 4.5	Genotype frequency among children with both parents were carriers of β -thalassemia, variants, HPFH, or $\delta\beta$ -thalassemia	129
Table 4.6	Genotype frequency among children with low risk of having symptomatic thalassemia	129
Table 4.7	Genotype groups and the number of samples for each genotype group. Each group is described in the details column.....	131
Table 4.8	Frequency of heterozygous and homozygous α -thalassemia and variants, HbH, and segmental α -globin duplication.....	134
Table 4.9	Frequency of various α -thalassemia. Nomenclature follows the guidelines of the Human Genome Variant Society (HGVS)	134
Table 4.10	Frequency of α -thalassemia and variants, HbH, and <i>HBA</i> duplication among subjects of different ethnicities in Peninsular Malaysia	138
Table 4.11	Frequency of heterozygous and homozygous states among subjects of different ethnicities in Sabah	139
Table 4.12	Number of samples for the heterozygous carrier of β -, δ -thalassemia and variants, HPFH and $\delta\beta$ -thalassemia, and the carriers with concomitant α -thalassemia.....	141
Table 4.13	Frequency of heterozygous β -thalassemia, δ -thalassemia and variants, HPFH and $\delta\beta$ -thalassemia (without α -thalassemia). Nomenclature follows the guidelines of the Human Genome Variant Society (HGVS).	142
Table 4.14	Frequency of heterozygous state of β -thalassemia and variants, HPFH, $\gamma\delta\beta$ -thalassemia, $\delta\beta$ -thalassemia, and δ -thalassemia and variants among subjects of different ethnicities in Peninsular Malaysia	146

Table 4.15	Frequency of concomitant α -thalassemia and segmental α -globin duplication among various β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia carriers of different ethnicities in Peninsular Malaysia.....	150
Table 4.16	Frequency of β -thalassemia and variants, and Hemoglobin Lepore carriers among different ethnicity in Sabah	153
Table 4.17	Frequency of concomitant α -thalassemia and segmental α -globin duplication among the β -thalassemia and variants of different ethnicities in Sabah	155
Table 4.18	Frequency of heterozygous state of β -thalassemia and variants among subjects of different ethnicity in Sarawak	156
Table 4.19	Frequency of concomitant α -thalassemia among β -thalassemia carriers of different ethnicity in Sarawak.....	156
Table 4.20	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia	158
Table 4.21	Frequency of compound heterozygous and homozygous β -thalassemia, β -variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia among subjects of different ethnicities in Peninsular Malaysia.....	161
Table 4.22	Frequency of concomitant α -thalassemia and segmental α -globin duplication among homozygous and compound heterozygous β -thalassemia/variant of different ethnicities in Peninsular Malaysia.	164
Table 4.23	Frequency of compound heterozygous and homozygous β -thalassemia states among subjects of different ethnicity in Sabah ..	166
Table 4.24	Frequency of compound heterozygous and homozygous β -thalassemia states among subjects of mixed-ethnics in Sabah.....	167
Table 4.25	Frequency of concomitant α -thalassemia and segmental α -globin duplication among the homozygous and compound heterozygous β -thalassemia of different ethnicities in Sabah	168
Table 4.26	Frequency of heterozygous and homozygous β -thalassemia states among subjects of different ethnicities in Sarawak.....	169

Table 4.27	Frequency of concomitant α -thalassemia among β -globin gene and cluster mutation carriers and patients of different ethnicities in Sarawak.....	170
Table 4.28	Mutation frequency among children and parents in the Peninsular Malaysia, Sabah, and Sarawak.....	172
Table 4.29	List of variables, types, and missing and unique data.....	176
Table 4.30	Re-classified genotype for statistical analysis.....	176
Table 4.31	Frequency, ratio, and rank of each genotype group	178
Table 4.32	Summary statistics of the hematological parameters	180
Table 4.33	Summary statistics of each genotype group's Red Blood Cell (RBC) ($10^{12}/L$) parameters.....	183
Table 4.34	RBC ($10^{12}/L$) parameter normality statistics for each genotype group	184
Table 4.35	Summary statistics of each genotype group's Hemoglobin (Hb) (g/dL) parameters	188
Table 4.36	Hb (g/dL) parameter normality statistics for each genotype group .	189
Table 4.37	Summary statistics of each genotype group's Mean Corpuscular Volume (MCV) (fL) parameters	192
Table 4.38	MCV (fL) parameter normality statistics for each genotype group .	193
Table 4.39	Summary statistics of each genotype group's Mean Corpuscular Hemoglobin (MCH) (pg) parameters.....	196
Table 4.40	MCH (pg) parameter normality statistics for each genotype group .	197
Table 4.41	Summary statistics of each genotype group's Red Cell Distribution Width (RDW) (%) parameters.....	200
Table 4.42	RDW (%) parameter normality statistics for each genotype group .	201
Table 4.43	Summary statistics of each genotype group's Hemoglobin A ₂ (HbA ₂) (CE) (%) parameters.....	204
Table 4.44	HbA ₂ (CE) (%) parameter normality statistics for each genotype group	205

Table 4.45	Summary statistics of each genotype group's Hemoglobin F (HbF) (%) parameters of the CE	208
Table 4.46	HbF (CE) (%) parameter normality statistics for each genotype group	209
Table 4.47	Summary statistics of each genotype group's variants (%) of the CE.....	212
Table 4.48	Variants (CE) (%) parameter normality statistics for each genotype group.....	212
Table 4.49	Summary statistics of each genotype group's Hemoglobin A ₂ (HbA ₂) (%) parameters of the HPLC	215
Table 4.50	HbA ₂ (HPLC) (%) parameter normality statistics for each genotype group.....	216
Table 4.51	Summary statistics of each genotype group's Hemoglobin F (HbF) (%) parameters of the HPLC	219
Table 4.52	HbF (HPLC) (%) parameter normality statistics for each genotype group.....	220
Table 4.53	Post Hoc comparisons of hematological parameters and hemoglobin fractions between thalassemia genotype groups and the Normal group	252
Table 4.54	Summary statistics specifying the variables' types, missing data, and their unique count.....	253
Table 4.55	Summary statistics depicting the distributions of the variables	254
Table 4.56	Number of samples in train, test, and down-sampled datasets	259
Table 4.57	Summary statistics depicting the odds ratio (OR), 95% confidence interval (CI), and the p-value for each model.....	263
Table 4.58	Percentage of correct predictions from logistic regression model and Likelihood Ratio Test (LRT) between Null and Full models ...	263
Table 4.59	Likelihood Ratio Test (LRT) between models.....	263

Table 4.60	Accuracy of Model Predictions and Area Under the Curve (AUC) of each model	265
Table 4.61	Model rank based on R^2 , RMSE, AIC, and BIC weight.....	266
Table 4.62	Confusion matrix depicting the prediction of β -thalassemia and Normal genotype for the test and a new dataset	267
Table 4.63	Cross-validation statistics on the test and a new dataset for all models	268
Table 4.64	Cross-validation statistics on the new dataset for Model 1 and 4....	269
Table 4.65	Confusion matrix depicting prediction of β -thalassemia and Normal genotype for the new dataset using the repeated k-fold cross-validation	270
Table 4.66	Comparison of SNV, indels, and CNV mutations genotyped by conventional PCR and Whole Exome Sequencing	275
Table 4.67	FBC and Hb analysis of case ID 168_13 and his parents	280
Table 4.68	FBC and Hb analysis of case ID 1006_16 and his parents	285
Table 4.69	FBC and Hb analysis of case ID 1437_17 and his parents	291
Table 4.70	FBC and Hb analysis of case ID 1043_18 and her parents.....	294
Table 5.1	CNVkit, Control-Freec, and cn.mops false detection rates.....	311
Table 5.2	Key differences of each CNV caller	314
Table 5.3	Comparison of the actual breakpoint and the breakpoint estimations by CNVkit, Control-Freec, and cn.mops	316

LIST OF FIGURES

	Page
Figure 1.1 Study workflow	11
Figure 2.1 Molecular tetrameric structure of Hb with iron (heme) binding.....	13
Figure 2.2 Hemoglobin Switching. The graph shows changes in globin gene expression during prenatal and postnatal development (Venkatesan et al., 2020).....	16
Figure 2.3 The analog (earlier) version of Sanger Sequencing using chain termination method based on the enzymatic process with dideoxynucleotides (Hunkapiller et al., 1991)	30
Figure 2.4 Automated Sanger sequencing utilizes fluorescently labeled dNTPs allowing a single reaction sequencing (Adapted from “The Sanger (chain-termination) method for DNA sequencing” by Estevezj (2012) is licensed under CC BY-SA 3.0).....	32
Figure 2.5 DNA analysis test workflow for the detection of thalassemia.....	36
Figure 2.6 Illumina sequencing by synthesis (SBS). DNA templates are immobilized on a flowcell. When nucleotides are incorporated onto the DNA strands, they release light pulses that are captured by the sequencer and output the bases read from each cluster, along with the quality metrics for each base (Hassan et al., 2023).....	39
Figure 2.7 PacBio’s SMRTbell library base calling in SMRT cell containing millions of ZMW wells. Anchored polymerase incorporates the nucleotides, emitting light that is measured in real-time. A base-calling algorithm translates the light into DNA sequence (Hassan et al., 2023).....	41
Figure 2.8 ONT’s library base calling in nanopore protein. DNA bases passes through the nanopore, producing “squiggle” translated into nucleotides by neural network (Hassan et al., 2023)	42

Figure 3.1	Overall target-enrichment sequencing sample preparation workflow87
Figure 3.2	Representative sample electropherograms showing pre-capture analysis of amplified library DNA.....90
Figure 3.3	Representative sample electropherograms showing post-captured analysis of amplified index-tagged library DNA.....96
Figure 3.4	Preprocessing of the raw sequence reads into an analysis-ready binary alignment map (BAM) workflow99
Figure 3.5	Germline variant calling workflow 101
Figure 3.6	Callset/Genotype refinement and annotation workflow 103
Figure 3.7	CNV analysis workflow using CNVkit..... 105
Figure 3.8	a) Relative estimated <i>HBA</i> duplication of sample 168_13 (light gray area). b) Primer F181066_181085 and R82875_82856 amplifies 3.9 kb amplicon. F181747_181766 and R80456_80437 were used to sequence the amplicon. c) Gel electrophoresis showing the 3.9 kb amplicon size alongside 1 kb DNA ladder 109
Figure 3.9	a) Relative estimated <i>HBA</i> duplication of sample 1006_16 (light gray area). b) Primer F177870_177889 and R89153_89134 amplifies 2.5 kb amplicon. F178416_178435 and R87871_87850 were used to sequence the amplicon. c) Gel electrophoresis showing the 2.5 kb amplicon size of 1006_16 and 61_14 alongside 1 kb DNA ladder..... 112
Figure 4.1	Mean and per sequence quality score for a) Raw FASTQ reads; b) trimmed FASTQ reads..... 125
Figure 4.2	Percentage of reads duplicates and GC: a) raw paired-end FASTQ; b) trimmed paired-end FASTQ 126
Figure 4.3	a) Percentage of duplicate tagged reads across all samples' reads; b) mapped reads and mapped reads percentage across all samples . 127
Figure 4.4	Percentage of the various duplicated read types 128
Figure 4.5	RBC ($10^{12}/L$) Q-Q plot of each genotype group..... 185

Figure 4.6	Boxplot depicting the distribution of RBC ($10^{12}/L$) of each genotype group.....	186
Figure 4.7	Boxplot depicting the distribution of Hb (g/dL) of each genotype group	190
Figure 4.8	Boxplot depicting the distribution of MCV (fL) of each genotype group	194
Figure 4.9	Boxplot depicting the distribution of MCH (pg) of each genotype group	198
Figure 4.10	Boxplot depicting the distribution of RDW (%) of each genotype group	202
Figure 4.11	Boxplot depicting the distribution of HbA ₂ (CE) (%) of each genotype group.....	206
Figure 4.12	Boxplot depicting the distribution of HbF (CE) (%) of each genotype group.....	210
Figure 4.13	Boxplot depicting the distribution of Variant (CE) (%) of each genotype group.....	213
Figure 4.14	Boxplot depicting the distribution of HbA ₂ (HPLC) (%) of each genotype group.....	217
Figure 4.15	Boxplot depicting the distribution of HbF (HPLC) (%) of each genotype group.....	221
Figure 4.16	Kruskal Wallis test, violin, and boxplot depicting the distribution of RBC ($10^{12}/L$) among the genotype classes	223
Figure 4.17	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for RBC ($10^{12}/L$) among the genotype classes	224
Figure 4.18	Kruskal Wallis test and violin plot depicting the distribution of Hb (g/dL) among the genotype classes	226
Figure 4.19	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for Hb (g/dL) among the genotype classes	227

Figure 4.20	Kruskal Wallis test and violin plot depicting the distribution of MCV (fL) among the genotype classes.....	229
Figure 4.21	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for MCV (fL) among the genotype classes	230
Figure 4.22	Kruskal Wallis test and violin plot depicting the distribution of MCH (pg) among the genotype classes	232
Figure 4.23	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for MCH (pg) among the genotype classes	233
Figure 4.24	Kruskal Wallis test and violin plot depicting the distribution of RDW (%) among the genotype classes.....	235
Figure 4.25	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for RDW (%) among the genotype classes	236
Figure 4.26	Kruskal Wallis test and violin plot depicting the distribution of HbA ₂ (%) taken by CE among the genotype classes	238
Figure 4.27	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for HbA ₂ (%) taken by CE among the genotype classes	239
Figure 4.28	Kruskal Wallis test and violin plot depicting the distribution of HbF (%) taken by CE among the genotype classes	241
Figure 4.29	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for HbF (%) taken by CE among the genotype classes	242
Figure 4.30	Kruskal Wallis test and violin plot depicting the distribution of Hb Variants (%) taken by CE among the genotype classes.	243
Figure 4.31	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for Hb Variants (%) taken by CE among the genotype classes	244

Figure 4.32	Kruskal Wallis test and violin plot depicting the distribution of HbA ₂ (%) taken by HPLC among the genotype classes	246
Figure 4.33	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for HbA ₂ (%) taken by HPLC among the genotype classes	247
Figure 4.34	Kruskal Wallis test and violin plot depicting the distribution of HbF (%) taken by HPLC among the genotype classes	249
Figure 4.35	Heatmap depicting Dunn's multiple pairwise comparisons with Holm's method adjusted p-value for HbF (%) taken by HPLC among the genotype classes	250
Figure 4.36	Heatmap depicting the correlation between each variable.....	254
Figure 4.37	Boxplots showing the distribution of the hemoglobin indices among the heterozygous β -thalassemia and the Normal. a) MCV, b) MCH, c) RBC, and d) HbA ₂	255
Figure 4.38	Histogram showing the distribution of: a) MCV, b) MCH, c) RBC, and d) HbA ₂ variables among the heterozygous β -thalassemia carriers and the normal group.....	256
Figure 4.39	Scatterplot depicting the distribution of MCV and MCH for β -thalassemia carriers and Normal samples. The plot size indicates the HbA ₂ (%).....	257
Figure 4.40	Scatterplot depicting the distribution of RBC and MCH for β -thalassemia carriers and Normal samples. The plot size indicates the HbA ₂ (%).....	258
Figure 4.41	Visualization of collinearity for all models.....	260
Figure 4.42	Forest plot showing the odds ratio and confidence interval of each variable. a) Model 1; b) Model 2; c) Model 3; d) Model 4.....	264
Figure 4.43	Receiver operating characteristics (ROC) curve of each model	265
Figure 4.44	Spiderweb plot comparing the performance of each model.....	266
Figure 4.45	Repeated k-fold cross-validation accuracy for: a) Model 1, and b) Model 4	269

Figure 4.46	Prediction of the functional effects of <i>KLF1</i> Cd 322 (CGC>AGC) by: a) Polyphen-2; and b) MutationTaster2273
Figure 4.47	Prediction of the functional effects of <i>SUPT5H</i> IVS2-4146 (C>T) by Combined Annotation Dependent Depletion (CADD).....274
Figure 4.48	<i>HBA</i> MLPA analysis of the parents: a) MLPA analysis of the father showed multiple polymorphisms across the <i>HBA</i> cluster; b) MLPA analysis of the mother showed normal copy number of the <i>HBA</i> cluster.281
Figure 4.49	CNV analysis of the <i>HBA</i> cluster for case 168_13: a) MLPA analysis showed a duplication spanning from the HS40 (172 nt probe) until exon 3 of <i>HBQ1</i> gene (400 nt probe) and the multiple polymorphisms inherited from the father; b) CNV analysis of exome sequencing reads showed a copy number duplication spanning (blue dots) from the <i>MPG</i> gene until <i>HBQ1</i> gene.282
Figure 4.50	Both sequence reads were mapped to the downstream part of the reference. a) Sequence unmatching (red bar) the NC_000016:0-229460 reference sequence is the upstream sequence of $\alpha\alpha^{102}$ duplication; b) Sequence matching (green bar) the reference is the downstream sequence of $\alpha\alpha^{102}$ duplication.....283
Figure 4.51	Nucleotide blast (BLASTN) of the sequencing data. Forward and reverse sequencing reads are shown in A) and B), and C) and D), respectively: A) and D) amplicon matching the end segment of $\alpha\alpha^{102}$ duplication (NC_000016.10:181947); B) and C) amplicon matching the start segment of $\alpha\alpha^{102}$ duplication (NC_000016.10:79903)284
Figure 4.52	CNV analysis of the <i>HBA</i> cluster for case ID 1006_16: a) MLPA analysis showed a duplication spanning from the HS40 (172 nt probe) until downstream of <i>HBA1</i> gene (283 nt probe) and Asian polymorphism (292 nt probe) inherited from the father (ID 61_14); b) CNV analysis of exome sequencing reads showed a copy number duplication spanning (blue dots) from downstream of the <i>MPG</i> gene until <i>HBA1</i> gene.....286

- Figure 4.53 CNV analysis of the *HBA* cluster for the father of case ID 1006_16: a) MLPA analysis showed a duplication spanning from the HS40 (172 nt probe) until downstream of *HBA1* gene (283 nt probe) and Asian polymorphism (292 nt probe); b) CNV analysis of exome sequencing reads showed a copy number duplication spanning (blue dots) from downstream of the *MPG* gene until *HBA1* gene287
- Figure 4.54 Both sequence reads were mapped to the upstream part of the reference. a) Sequence matching (green bar) the NC_000016:0-229460 reference sequence is the downstream sequence of $\alpha\alpha\alpha^{91}$ duplication; b) Sequence unmatching (red bar) the reference is the upstream sequence of $\alpha\alpha\alpha^{91}$ duplication288
- Figure 4.55 Nucleotide blast (BLASTN) of the sequencing data. Forward sequencing reads are shown in A) and B), and reverse in C), respectively: A) amplicon matching the end segment of $\alpha\alpha\alpha^{91}$ duplication (NC_000016.10:179080); B) amplicon matching the start segment of $\alpha\alpha\alpha^{91}$ duplication (NC_000016.10:87820); C) reverse sequence read matching the end segment of $\alpha\alpha\alpha^{91}$ duplication.....289
- Figure 4.56 The start and end of $\alpha\alpha\alpha^{91}$ duplication: a) The start of the $\alpha\alpha\alpha^{91}$ duplication (NC_000016.10:87820) is at the intron 10 position 878 of the *NPRL3* gene; b) end (NC_000016.10:179080) at 1559 bp upstream of the *HBA1* gene.....290
- Figure 4.57 CNV analysis of the *HBA* cluster for the case ID 1437_17: a) MLPA analysis showed a heterozygous deletion spanning from upstream of the *HBM* (184 nt probe) until downstream of the *HBA1* gene (283 nt probe); b) CNV analysis of exome sequencing reads showed a heterozygous deletion (red dots) spanning from downstream of the *HBM* gene until *HBA1* gene292
- Figure 4.58 CNV analysis of the *HBA* cluster for the mother (case ID 1904_17) of the case ID 1437_17: a) MLPA analysis showed a heterozygous deletion spanning from upstream of the *HBM* (184

	nt probe) until downstream of the <i>HBAI</i> gene (283 nt probe); b) CNV analysis of exome sequencing reads showed a heterozygous deletion (red dots) spanning from downstream of the <i>HBM</i> gene until <i>HBAI</i> gene293
Figure 4.59	CNV analysis of the <i>HBA</i> cluster for the case ID 1043_14: a) MLPA analysis showed a compound heterozygous deletion spanning from upstream of the <i>HBM</i> (184 nt probe) until downstream of the <i>HBAI</i> gene (283 nt probe), and a deletion spanning from upstream of <i>HBA2</i> (391 nt probe) until upstream of <i>HBAI</i> (190 nt probe); b) CNV analysis of exome sequencing reads showed a heterozygous deletion (red dots) spanning from downstream of the <i>HBM</i> gene until <i>HBAI</i> gene295
Figure 4.60	CNV analysis of the <i>HBA</i> cluster of the father of the case ID 1043_14: a) MLPA analysis showed a heterozygous deletion spanning from upstream of the <i>HBM</i> (184 nt probe) until downstream of the <i>HBAI</i> gene (283 nt probe); b) CNV analysis of exome sequencing reads using the Control-Freec failed to call the deletion296
Figure 4.61	The 8 bp homologous sequences can be found at both 50 and 30 deletion breakpoints either at NG_000006.1:g.22766_22773 or NG_000006.1:g.39537_39544. Thus, the -- ^{GB} deletion can be mapped either from the 1164 bp upstream of <i>HBZPI</i> to 1151 bp downstream of the <i>HBAI</i> gene (NG_000006.1:g.22766_39536del) or from the 1156 bp upstream of <i>HBZPI</i> to 1159 bp downstream of <i>HBAI</i> gene (NG_000006.1:g.22774_39544del).297
Figure 5.1	The common β -thalassemia mutations among countries in SEA: Myanmar (Win et al., 2002), Northern Thailand (Sirichotiyakul et al., 2003), Cambodia (Munkongdee et al., 2016), Peninsular Malaysia (current study), Sabah (current study), Singapore (Tan et al., 1994), South Sumatra (Sari et al., 2022), Laos (Singha et al., 2021; Phengsavanh et al., 2022), Hanoi (Filon et al., 2000),

South Vietnam (Hao et al., 2001), Northeast Thailand (Yamsri et al., 2011), East Kalimantan (Susanto et al., 2020), West Java (Maskoen et al., 2017), and Central Java (Rujito et al., 2015). The n refers to heterozygous carrier cohorts while the n ^a refers to homozygous and compound heterozygous cohorts. Meanwhile, the *Filipinos refers to Filipinos residing in Taiwan (Ko et al., 1998)	301
--	-----

Figure 5.2	Comparison between conventional and NGS-based DNA analysis workflow for β -thalassemia	318
------------	--	-----

LIST OF SYMBOLS

α	Alpha
β	Beta
γ	Gamma
δ	Delta
ε	Epsilon
ζ	Zeta
ψ	pseudo
μ	micro

LIST OF ABBREVIATIONS

μL	microliter
AF	Amniotic fluid
AHSP	Alpha-Hemoglobin-Stabilizing Protein
AIC	Akaike's Information Criterion
AIC wt	Akaike's Information Criterion weight
ALFA	Allele Frequency Aggregator
ANOVA	Analysis of Variance
APOE	apolipoprotein E
ARMS	Amplification Refractory Mutation System
ASO	Allele-specific Oligonucleotide hybridization
ASPCR	Allele-specific oligonucleotide PCR
AUC	Area Under the Curve
BAF	B-Allele Frequency
BAM	Binary Alignment Map
<i>BCL11A</i>	BAF Chromatin Remodeling Complex Subunit
BFHS	Hb Bart's Hydrops Fetalis Syndrome
BIC	Bayesian Information Criterion
BIC wt	Bayesian Information Criterion weight
BLASTN	Nucleotide Basic Local Alignment Search Tool
BMD	bone marrow density
bp	base pair
BWA-MEM	Burrows-Wheeler Aligner-MEM
CADD	Combined Annotation Dependent Depletion
caret	Classification And REgression Training
CBS	circular binary segmentation
CCS	Circular Consensus Sequence
Cd	Codon
CE	Capillary Electrophoresis
CGE	Capillary Gel Electrophoresis
CGH	Comparative Genomic Hybridization
CI	Confidence interval

CLI	Command Line Interface
CMA	chromosome microarray
cn.mops	Copy Number estimation by a Mixture Of PoissonS
cnr	Copy number ratio
cns	Copy number segment
CNV	Copy Number Variant
CO ₂	Carbon dioxide
<i>COL1A1</i>	type 1 collagen
CS	Constant Spring
CV	Chorionic villus
dATP	Deoxyadenosine 5'-triphosphate
dbSNP	Single Nucleotide Polymorphism Database
dCTP	Deoxycytidine 5'-triphosphate
ddNTP	dideoxynucleotide triphosphates
df	Degree of freedom
DFO	Deferasirox
dGTP	Deoxyguanine 5'-triphosphate
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide triphosphate
dTTP	Deoxythymidine 5'-triphosphate
EDA	Exploratory Data Analysis
EDTA	ethylenediaminetetraacetic acid
EOF	Electro-osmotic Flow
FASTA	FAST-All File format
FASTQ	File format
FBC	Full Blood Count
FDR	False detection rate
fL	femtolitre
FN	False-negative
FNR	False negative rate
FP	False positive
FPR	False positive rate
FS	Fisher's Exact Test

GATK	Genome Analysis Toolkit
gDNA	genomic DNA
GE	Genome Explorer
GPU	Graphical processing units
GQ	Phred-scale Genotype Quality
GRCh38 or hg38	Genome Reference Consortium Human Build 38
GUI	Graphical User Interface
GVCF	Genome Variant Calling Format
GWAS	Genome-wide Association Studies
HAPMAP	Haplotype Map
Hb	Hemoglobin
HbA	Hemoglobin A or Hemoglobin A1
<i>HBA</i>	Hemoglobin subunit alpha
<i>HBA1</i>	Hemoglobin subunit alpha 1
<i>HBA2</i>	Hemoglobin subunit alpha 2
HbA ₂	Hemoglobin A ₂
<i>HBAP1</i>	Pseudo alpha 1
<i>HBAP2</i>	Pseudo alpha 2
<i>HBB</i>	Hemoglobin subunit beta
<i>HBBP1</i>	Hemoglobin subunit beta pseudogene 1
<i>HBD</i>	Hemoglobin subunit delta
Hb C	Hemoglobin C
Hb E	Hemoglobin E
<i>HBE1</i>	Hemoglobin subunit epsilon
HbF	Fetal Hemoglobin
<i>HBG1</i>	Hemoglobin subunit gamma 1
<i>HBG2</i>	Hemoglobin subunit gamma 2
HbH	Hemoglobin H disease
<i>HBM</i>	Hemoglobin subunit mu
<i>HBQ1</i>	Hemoglobin subunit theta 1
Hb S	Hemoglobin Sickle
<i>HBS1L</i>	HBS1 Like Translational GTPase
HbVar	Database of Human Hemoglobin Variants and Thalassemia Mutations
<i>HBZ</i>	Hemoglobin Subunit Zeta

<i>HBZP1</i>	Hemoglobin subunit zeta pseudogene 1
HCT	Hematocrit
HET	Count of heterozygous genotypes
HGMD	Human Gene Mutation Database
HGP	Human Genome Project
HMM	Hidden Markov model
HKL	Hospital Kuala Lumpur
HOMREF	Count of homozygous reference genotypes
HOMVAR	Count of homozygous variant genotypes
HPLC	High-pressure Liquid Chromatography
HPFH	Hereditary Persistence of Fetal Hemoglobin
HRM	High-Resolution Melting
HS3	Hypersensitive site 3
HS40	DNase I hypersensitive site 40
HSB	Hospital Sultanah Bahiyah
HTML	Hyper Text Markup Language
HTS	High-throughput Sequencing
HUSM	Hospital Universiti Sains Malaysia
IBM	International Business Machines
ICT	iron-chelating therapy
ID	Iron Deficiency
IDA	Iron Deficiency Anemia
IEF	Isoelectric Focusing
IGV	Integrative Genomics Viewer
IMR	Institute for Medical Research
indel	Insertion and deletion
INFO	Information
IQR	Interquartile range
Iso-Seq	isoform sequencing
IVS	Intervening Sequence
JePEM	Jawatankuasa Etika Penyelidikan Manusia
J-Sg	J-Singapore
KLF1	Kruppel-Like Factor 1
L	Liter

LCR	locus control region
LOVD	Leiden Open Variation Database
LRT	Likelihood test ratio
<i>LUC7L</i>	Putative RNA-binding protein Luc7-like 1 gene
MARMS	Multiplex Amplification Refractory Mutation System
Maser	Management and Analysis System for Enormous Reads
mb	megabase
MCC	Matthews Correlation Coefficient
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCP1	monocyte chemotactic protein-1
MCV	Mean Corpuscular Volume
Mdn	Median
MGM	Molecular Genetic Laboratory
mL	milliliter
MLPA	Multiplex Ligation Probe-dependent Amplification
MOH	Ministry of Health
MQ	Root Mean Square of the mapping quality of reads
MREC	Medical Research and Ethics Committee
mRNA	messenger Ribonucleic Acid
MTR	Malaysian Thalassemia Registry
ng	Nanogram
NGS	Next Generation Sequencing
NIR	No Information Rate
nm	Nanomole
NRC	normalized read count
NTC	No Template Control
nt	nucleotide
NTDT	Non-Transfusion Dependent Thalassemia
OA	Orang Asli
O ₂	Oxygen
ONT	Oxford Nanopore Technology
OMIM	Online Mendelian Inheritance in Man
PacBio	Pacific Bioscience

PCA	Principal component analysis
PCR	Polymerase Chain Reaction
PE	Pair end
pg	pictogram
PL	Phred-scaled genotype likelihoods
PND	Prenatal diagnosis
PoN	Panel of Normal
<i>POLR3K</i>	Polymerase III, RNA Subunit, K gene
QD	Quality by depth
Q-PCR	Quantitative Polymerase Chain Reaction
QSR	Quality Score Recalibration
QTL	Quantitative Trait Loci
QV	Phred Quality Score
RBC	Red Blood Cell
RD	Read Depth
RDW	Red cell distribution width
RFLP	Restriction Fragment Length Polymorphism
RMSE	Root Mean Square Error
ROC	Receiver operating characteristics
ROI	Region of Interest
RPKM	Repeats Per Kilobase Exon Model Per Million Mapped Reads
SAM	Sequence Alignment Map
SBS	sequencing by synthesis
SCA	Sickle Cell Anemia
SEA	Southeast Asia
<i>SOX6</i>	sex-determining region Y (SRY)- Box Transcription Factor 6
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SSCP	Single-strand Conformation Polymorphism
STR	Short Tandem Repeat
<i>SUPT5H</i>	SPT5 Homolog, DSIF Elongation Factor Subunit
SVD	singular value decomposition
TA	Ta repeat
TDT	Transfusion Dependent Thalassemia

<i>TGFB1</i>	Transforming growth factor β 1
TGS	Third Generation Sequencing
Tm	Melting temperature
TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
TS	Targeted sequencing
TW	Twin Peaks
T2T	Telomere-to telomere
T2T-CHM13	Telomere-to-Telomere assembly of the CHM13 cell line
T2T-CHM13+Y	T2T-CHM13 assembly and Y-chromosome
UCSC	University of California, Santa Cruz
<i>UGT1A</i>	Glucoronosyltransferase 1
UNIX	UNiplexed Information Computing System
USA	United States of America
UTR	Untranslated Region
UV-VIS	Ultraviolet-Visible Spectroscopy
VCF	Variant Calling Format
VDR	Vitamin D receptor
VIF	Variance inflation factor
VQSLOD	Variant Quality Score Log-Odds
VUS	Varaint of uncertain significance
WBC	White Blood Cell
WES	Whole Exome Sequencing
WHO	World Health Organisation
XHMM	eXome hidden Markov model
ZRPKM	z-scores reads per thousand bases per million reads sequenced

LIST OF APPENDICES

APPENDIX A	The MOH's ethical approval from the MREC (identification number NMRR-20-2382-56195)
APPENDIX B	The university's ethical approval from the JEPeM (study protocol code of USM/JEPeM/21060437)
APPENDIX C	The hematological parameters and genotypes of families selected for WES
APPENDIX D	List of gene of interest comprising 570 <i>HBA</i> and <i>HBB</i> cluster and modifier genes
APPENDIX E	The dataset information and tool parameters for CNV analysis by Control-Freec
APPENDIX F	Control-Freec copy number plot
APPENDIX G	cn.mops workflow
APPENDIX H	Number of families (n) and hospital/health clinic for each state
APPENDIX I	Frequency of heterozygous and homozygous states of α -thalassemia and variants, HbH, and segmental α -globin duplication among subjects of different ethnicities in Peninsular Malaysia
APPENDIX J	Frequency of heterozygous state of β -thalassemia and variants, HPFH and $\delta\beta$ -thalassemia, and δ -thalassemia and variants among subjects of different ethnicities in Peninsular Malaysia
APPENDIX K	Frequency of concomitant α -thalassemia and segmental α -globin duplication among β -globin gene and cluster mutation carriers of different ethnicities in Peninsular Malaysia
APPENDIX L	Frequency of compound heterozygous and homozygous β -thalassemia, variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia among different ethnicities in Peninsular Malaysia
APPENDIX M	Frequency of concomitant α -thalassemia and segmental α -globin duplication in individuals with homozygous and compound heterozygous β -thalassemia, variants, δ -thalassemia, HPFH, and $\delta\beta$ -thalassemia across different ethnicities in Peninsular Malaysia
APPENDIX N	Hb (g/dL) Q-Q plot of each genotype group
APPENDIX O	MCV (fL) Q-Q plot of each genotype group
APPENDIX P	MCH (pg) Q-Q plot of each genotype group
APPENDIX Q	RDW (%) Q-Q plot of each genotype group
APPENDIX R	CE's HbA ₂ (%) Q-Q plot of each genotype group
APPENDIX S	CE's HbF (%) Q-Q plot of each genotype group

APPENDIX T	CE's variant (%) values Q-Q plot of each genotype group
APPENDIX U	HPLC's HbA ₂ (%) values Q-Q plot of each genotype group
APPENDIX V	HPLC's HbF (%) values Q-Q plot of each genotype group
APPENDIX W	CNV calling FDR and the breakpoint estimations of the CNV by MLPA, CNVkit, Control-Freec, and cn.mops

**ANALISIS SPEKTRA VARIAN TALASEMIA DAN VALIDASI
PEMBANGUNAN PENJUJUKAN EKSOM UNTUK TALASEMIA β DI
MALAYSIA**

ABSTRAK

β -talasemia merupakan masalah kesihatan yang serius di Malaysia, terutamanya dalam kalangan etnik tertentu. Kajian ini mengenal pasti mutasi β -talasemia dalam 736 keluarga yang menyertai saringan trio. Penjujukan Eksom Lengkap (WES) digunakan untuk mengesahkan Varian Bilangan Salinan (CNV) yang tidak diketahui, menentukan varian dalam kes genotip yang tidak informatif, dan menghimpun saluran bioinformatik untuk pengesanan serentak Varian Nukleotida Tunggal (SNV), indel, dan CNV dalam talasemia. Parameter hematologi dibandingkan di antara kumpulan β -talasemia, dan gabungan pembolehubah yang sesuai untuk β -talasemia heterozigot diterokai. Empat puluh enam, 16, dan 8 mutasi telah direkodkan dalam *HBB*, *HBA*, dan *HBD*. Lebih separuh daripada 571 kanak-kanak dari Semenanjung Malaysia adalah pesakit, 163 (28.5%) adalah pesakit Hb E/ β -talasemia diikuti oleh 139 (24.3%) β -talasemia homozigot atau heterozigot kompaun. Hb E, IVS 1-5 (G>C), IVS 1-1 (G>T), dan Hb Malay adalah mutasi yang paling kerap ditemui dalam kalangan Melayu. IVS 2-654 (C>T), Cd 41/42 (-TTCT), dan -28 (A>G) adalah mutasi utama dalam kalangan Cina. Insiden β -talasemia di Sarawak adalah rendah. Daripada 19 keluarga, lapan kanak-kanak mempunyai β -talasemia homozigot atau heterozigot kompaun, manakala empat mempunyai Hb E/ β -talasemia. β -Filipino adalah mutasi utama dalam kalangan etnik di Sabah. Daripada 146 kanak-kanak, 90 (61.6%) adalah β -Filipino homozigot yang bergantung kepada transfusi. Berdasarkan heterogeniti ini, analisa DNA yang

komprehensif melangkaui ujian spesifik mengikut populasi sangat diperlukan. Berbanding dengan pembawa dan pesakit β -talasemia yang tidak mempunyai α -talasemia, mikrositosis ($p < .001$) dan hipokromia ($p < .001$) berkurangan apabila pembawa β -talasemia mempunyai α -talasemia, memperburuk mikrositosis ($p = .004$) dan hipokromia ($p < .001$) dalam pesakit Hb E/ β -talasemia, dan mengurangkan peratusan varian pada pembawa Hb E ($p < .001$). Namun, tiada perubahan yang ketara pada parameter hematologi bagi kumpulan homozigot atau heterozigot kompaun yang mempunyai α -talasemia. Penurunan MCH (OR=0.52, $p < 0.001$) dan peningkatan HbA₂ (OR=18.1, $p < 0.001$) dalam Model 4 adalah mencukupi untuk meramal pembawa β -talasemia. Walaupun semua model menunjukkan prestasi ramalan yang baik (93.5-93.8%), Model 1 (MCV, RBC, MCH, dan HbA₂), Model 2 (RBC, MCH, dan HbA₂), dan Model 3 (MCV, MCH, dan HbA₂) tidak meningkatkan ramalan pembawa β -talasemia secara signifikan. Untuk memudahkan proses genotip, saluran HaplotypeCaller dan CalculateGenotypePosteriors daripada The Genome Analysis Toolkit (GATK) digunakan dan ianya 100% selaras dengan kaedah konvensional untuk pengesanan varian. Tiga perisian CNV telah dibandingkan (CNVkit, cn.mops, dan Control-Freec). Control-Freec menunjukkan prestasi terbaik, membantu dalam mengenal pasti dua CNV yang novel iaitu $\alpha\alpha$ ⁹¹ dan *de novo* $\alpha\alpha$ ¹⁰². Kesimpulannya, kajian pelbagai aspek ini memberikan pemahaman yang komprehensif tentang spektrum β -talasemia di Malaysia. Saluran bioinformatik yang dihimpunkan menawarkan strategi yang lebih mudah dan bernilai untuk aliran kerja yang lebih cekap dan menyumbang kepada pembangunan strategi saringan yang lebih berkesan.

**ANALYSIS OF THALASSEMIA VARIANT SPECTRUM AND
DEVELOPMENTAL VALIDATION OF EXOME SEQUENCING FOR β -
THALASSEMIA IN MALAYSIA**

ABSTRACT

β -thalassemia poses a significant health problem in Malaysia, particularly among specific ethnicities. This study identified β -thalassemia mutations in 736 families participating in trio screening. Whole Exome Sequencing (WES) was used to validate the unknown Copy Number Variant (CNV), determine variants in the uninformative genotyping cases, and assemble a bioinformatics pipeline for simultaneous detections of thalassemia SNVs, indels, and CNVs. Hematological parameters were compared among β -thalassemia groups, and suitable combinations of predictors for heterozygous β -thalassemia were explored. Forty-six, 16, and 8 mutations were recorded in *HBB*, *HBA*, and *HBD*, respectively. More than half of the 571 Peninsular Malaysian children were patients, 163 (28.5%) were Hb E/ β -thalassemia patients followed by 139 (24.3%) homozygous or compound heterozygous β -thalassemia. Hb E, IVS 1-5 (G>C), IVS 1-1 (G>T), and Hb Malay were the most common mutations among the Malays. IVS 2-654 (C>T), Cd 41/42 (-TTCT), and -28 (A>G) were common in Chinese. Sarawak had a low incidence of β -thalassemia. Of the 19 families, eight children were homozygous or compound heterozygous β -thalassemia, while four were Hb E/ β -thalassemia. β -Filipino was prevalent in ethnicities of Sabah. Of 146 children, 90 (61.6%) were transfusion-dependent homozygous β -Filipino. The heterogeneity highlights the need for comprehensive genotyping, beyond population-specific variants. Compared to those without α -thalassemia, concomitant α -thalassemia mitigated microcytosis ($p < .001$)

and hypochromia ($p < .001$) in β -thalassemia carriers, exacerbated microcytosis ($p = .004$) and hypochromia ($p < .001$) in Hb E/ β -thalassemia patients, and reduced Hb E carrier variant level ($p < .001$). However, co-inheritance of α -thalassemia did not significantly alter hematological parameters in homozygous or compound heterozygous β -thalassemia. A parsimonious combination of decreased MCH (OR=0.52, $p < 0.001$) and increased HbA₂ (OR=18.1, $p < 0.001$) in Model 4 were sufficient for heterozygous β -thalassemia prediction. All models demonstrated good predictive performance (93.5-93.8%), however, Model 1 (MCV, RBC, MCH, and HbA₂), Model 2 (RBC, MCH, and HbA₂), and Model 3 (MCV, MCH, and HbA₂) did not significantly improve heterozygous β -thalassemia prediction. To simplify the genotyping process, this study used The Genome Analysis Toolkit (GATK) HaplotypeCaller and CalculateGenotypePosteriors pipelines and demonstrated 100% concordance with conventional methods for variant detections. Three CNV callers were compared (CNVkit, cn.mops, and Control-Freec). Control-Freec exhibited superior performance, assisting in the identification of two novel CNVs of $\alpha\alpha\alpha$ ⁹¹ and *de novo* $\alpha\alpha\alpha$ ¹⁰². In conclusion, this multifaceted study provides a comprehensive understanding of the β -thalassemia spectrum in Malaysia. The assembled bioinformatics pipeline offers simpler and valuable strategies for a more efficient workflow and contributes to developing more effective molecular diagnostic strategies.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Cascade screening is a family-based approach to identifying individuals at risk for genetic conditions. In Malaysia's National Thalassemia Prevention and Control Program, parents of children under 12 seeking deoxyribonucleic acid (DNA) testing for β -thalassemia must undergo cascade screening as a family unit. This trio study involves the DNA testing of a proband and both parents and is often used by clinical laboratories to help with the classification of genetic variants. It aims to prevent incorrect diagnoses, particularly in children with Non-Transfusion Dependent Thalassemia (NTDT) and Transfusion Dependent Thalassemia (TDT). Since the data derived from Full Blood Count (FBC) and Hemoglobin (Hb) analysis in transfusion patients can be ambiguous and inconsistent, which may lead to a wide range of potential diagnoses, the inclusion of the parents' FBC and Hb analysis can expedite an accurate diagnostic process.

The research is divided into two parts. The initial part investigates the spectrum of mutations present in archived β -thalassemia DNA analysis data from families who have utilized this service. The study includes archived β -thalassemia DNA data for mutation patterns in carriers and patients, and identifying novel segmental deletions and duplications in the hemoglobin subunit beta (*HBB*) and hemoglobin subunit alpha (*HBA*) gene clusters. The DNA testing uses a variety of Polymerase Chain Reaction (PCR) techniques, including Gap-PCR, Amplification Refractory Mutation System (ARMS)-PCR, Sanger Sequencing, and Multiplex Ligation-dependent Probe Amplification (MLPA), focusing on the α - and β -globin

genes and their clusters. To address the complexity of identifying concurrent α -thalassemia in carriers and β -thalassemia patients, common α -thalassemia mutations in all the samples are genotyped by ARMS and Gap-PCR.

The second part involved genotyping by Whole Exome Sequencing (WES). The Next Generation Sequencing (NGS) based trio exome analysis provides a powerful approach to identify causal mutations for inherited diseases, in recessive, dominant disease, or *de novo* variants. Eighty archived DNA samples, comprising 26 family trios and two unrelated individuals, underwent WES. They comprised children with an unmatched genotype-phenotype diagnosis and families with unidentified α - or β -globin gene cluster deletions or duplications confirmation of preliminary results obtained by the conventional DNA analysis. The bioinformatics analysis involved a sequential evaluation of appropriate tools capable of genotyping SNVs, indels, and CNVs. These tools aim to streamline and replace the complex conventional DNA testing procedures used for thalassemia diagnosis.

To investigate the genotype-phenotype correlation of β -thalassemia, samples are classified into their respective genotype groups. This study uses non-parametric Kruskal-Wallis tests to compare RBC and Hb parameters across these groups. Additionally, the predictive performance of different models in identifying heterozygous β -thalassemia using logistic regression analysis. A parsimonious model incorporating MCH and HbA₂ (Model 4) is compared to more comprehensive models including RBC, MCV, MCH, and HbA₂ (Model 1), MCV, MCH, and HbA₂ (Model 2); and RBC, MCH, and HbA₂ (Model 3). The goal is to determine if the parsimonious model could effectively predict the presence of heterozygous β -thalassemia without the need for additional hematological parameters.

1.2 Problem Statement

1.2.1 Heterogeneity of Thalassemia Mutations, Ambiguous Differential Diagnoses, and Diagnostic Challenges

Thalassemia is characterized by its diverse mutation spectrum, with over a thousand variants recorded in various databases (Giardine et al., 2021; Halim-Fikri et al., 2024; Stephanou et al., 2019). It is a significant health concern in Malaysia, with a prevalence of approximately 4.5% among the Malay and Chinese populations (George, 1998), and has a substantial impact on the country's healthcare system. The diverse ethnic composition of Malaysia, combined with the prevalence of consanguinity and intermarriages, may have contributed to a complex genetic landscape. This diversity can lead to a wide range of thalassemia mutations, making accurate diagnosis and management challenging. The heterogeneous nature of thalassemia mutations, coupled with the diverse genetic makeup of the Malaysian population, presents a significant challenge for accurate diagnosis and effective management of the disorder.

Differential diagnoses for β -thalassemia can be challenging due to concomitant α -thalassemia, phenotypic variability, disease complexity, and molecular diagnostic test limitations. The simultaneous occurrence of α - and β -thalassemia, coupled with the diverse range of mutations (single nucleotide variants (SNV), small insertions and deletions (indels), deletions, duplications, insertions, nucleotide substitutions, and rearrangements), can significantly alter the hematological parameters, complicate diagnosis, and management. β -thalassemia carriers depend significantly on the increase in HbA₂ levels. Although not pathogenically relevant, the co-occurrence of δ -thalassemia or variants in β -thalassemia carriers reduces the HbA₂ levels, potentially masking the diagnosis of β -

thalassemia. Locally, the most common δ -thalassemia and variant was HbA2-Indonesia (HBD:c.208G>C) [δ cd 69 (GGT>CGT)] and HbA2-Deventer (HBD:c.202G>A) [δ cd 67 GTG>ATG)] (Hassan et al., 2019). Since δ -thalassemia modifies HbA₂ levels, diagnosing borderline HbA₂ β -thalassemia is vital. Each DNA test is limited to detecting several types of mutations. For instance, HbH disease ($-\alpha^{3.7}/--^{SEA}$) patients showed the existence of *HBA2* ($\alpha 2$ gene) when diagnosed by multiplex GAP-PCR. This could be due to the conversion of the *HBA1* into the *HBA2* gene, which must be ruled out by DNA sequencing and MLPA. The complex inheritance patterns of thalassemia mutations can lead to phenotypic variability, making accurate diagnosis challenging. The limitations of current DNA analysis methods, which are often designed to detect specific mutation types, can hinder the comprehensive assessment of genetic status, resulting in misdiagnoses or incomplete information. This complexity necessitates multiple laboratory methods, increasing the testing cost.

1.2.2 Labor-Intensive Testing, and DNA Test Limitations

For many years, DNA analyses of thalassemia have relied on two main steps: identifying common mutations using targeted DNA analysis methods such as ARMS-PCR, reverse dot-blot analysis, and Gap-PCR, while rare mutations are detected via Sanger sequencing and MLPA analysis, respectively. The MLPA enables the identification of large indels within the α - and β -globin clusters and complements the other methods meant to genotype known mutations. Unknown deletions and duplications discovered by MLPA require validation. For deletions and duplications spanning across the *HBA* and *HBB* clusters, validations by long-range PCR are possible because of the many probes available within these clusters.

However, larger deletions and duplications are harder to characterize due to the limited number of MLPA probes available beyond *HBA* and *HBB* clusters.

Although Sanger sequencing enables detections of SNV and indels in each globin gene, homozygous mutations must rule out the possibility of accompanying deletion on the other allele using MLPA. Indels are common in *HBB*, thus extra reads are required to accommodate the frameshifted heterozygous mixed Sanger sequencing reads. The combination of multiple methods is required to comprehensively assess the genetic status of β -thalassemia patients. This labor-intensive approach extends the laboratory turnaround time (TAT) and hinders efficient diagnosis.

1.2.3 Need for a Unified DNA Analysis Strategy

Detections of the frequently occurring thalassemia mutations are well established. These require multiple test methods for precise DNA analysis. They are mostly mutation-specific, which is laborious, resource-intensive, and time-consuming. A unified diagnostic approach that can simultaneously detect various mutation types is necessary to streamline the process. NGS technologies, such as WES and targeted sequencing, offer promising solutions for developing a comprehensive and efficient diagnostic approach for thalassemia.

1.3 Study Motivation

1.3.1 Understanding the Genotype-Phenotype Relationship and Hematological Parameter Evaluation for β -thalassemia Carrier Prediction

Thalassemia exhibits variable clinical presentations, even among individuals with identical genotypes. This inconsistency could be attributed to the coexistence of thalassemia with other diseases or the influence of genetic modifiers. Although the

genotype-phenotype relationship between carriers and patients is generally straightforward, the influence of α -thalassemia on the phenotype, RBC, and Hb parameters is unclear. To analyze the genotype-phenotype relationship, the non-parametric Kruskal-Wallis test is employed to identify differences in RBC and Hb parameters across different thalassemia genotypes.

β -thalassemia carriers are characterized by reduced MCV and MCH; and elevated RBC and HbA₂. To assess the association between binary outcome variables (heterozygous β -thalassemia and normal haplotype), four regression models were compared. These models include a full model incorporating RBC, MCV, MCH, and HbA₂, (Model 1) and three simpler models with MCV, MCH, and HbA₂ (Model 2), RBC, MCH, and HbA₂ (Model 3), and MCH and HbA₂ (Model 4), respectively.

1.3.2 Determining the Spectrum of β -thalassemia and Variants

Mutation spectrum analysis reveals a higher prevalence of mild β -thalassemia mutations compared to classical β -thalassemia mutations (Hassan, 2015). This skew likely arises from prioritizing the genotyping of β -thalassemia carriers with borderline HbA₂ levels over classical β -thalassemia carriers. Since the classical β -thalassemia mutations are not systematically diagnosed, the true spectrum of mutations remains unknown. Since cascade screening encompasses both patients and carriers, a more accurate mutation spectrum can be expected. Previous DNA analysis studies have primarily focused on specific regions or populations, limiting the understanding of the nationwide distribution of β -thalassemia mutations in Malaysia. The cascade screening data, encompassing trios with β -thalassemia and variants, provides a more comprehensive representation of the mutation landscape in Malaysia and shows actual mutation prevalence in Malaysia.

In Malaysia, the prevalence of α -thalassemia is lower than β -thalassemia (Ahmad et al., 2012; George, 1998). Cases of α -thalassemia due to single-gene deletion might be overlooked due to unremarkable red cell indices in routine blood counts and hence, are unlikely to be investigated. While cascade screening focuses on systematically identifying β -thalassemia and its variants, α -thalassemia status is also investigated within families exhibiting β -thalassemia and variant carrier phenotypes. Specifically, individuals with normal or α -thalassemia red cell indices are further assessed for α -thalassemia.

1.3.3 Bioinformatics Pipeline Assembly for Unified SNV, Indels, and CNV Analyses of Thalassemia

Whole Exome Sequencing (WES) targets only protein-coding regions of the human genome, comprising less than 2% of the total genome. WES offers coverage of over 95% of exons, reducing cost and sequencing time. Additionally, 85% of disease-associated variants reside within exons, making WES a preferred choice for diagnostic applications (Choi et al., 2009).

This study utilized WES as a reference to design a customized Targeted Sequencing (TS) panel for thalassemia, hemoglobinopathy, and their genetic modifiers. In the bioinformatics analysis, suitable tools that can genotype Single Nucleotide Variant (SNV), insertion and deletion (indel), and Copy Number Variant (CNV) sequentially were assessed. Both studies established the mutation heterogeneity, estimated the relative size of the CNVs, and facilitated the development of a customized TS and a bioinformatics pipeline compatible with SNV, indel, and CNV detection. CNV analysis using data derived from WES and targeted sequencing is a rapidly evolving field with the introduction of more open-source tools. To establish a customized bioinformatics pipeline for thalassemia, the

accuracy of multiple CNV detection tools was evaluated. This involved sequencing several cascade screening samples with unknown CNVs detected by MLPA. False negative and positive results were identified by comparing the index case and their parents' results. The study workflow is depicted in Figure 1.1.

1.4 Hypotheses and Research Questions

1.4.1 Hypothesis

- 1) The spectrum of β -thalassemia and hemoglobinopathy mutations among families participating in cascade screening will reveal a diverse range of genetic variants.
- 2) There will be significant differences in hematological parameters among the different β -thalassemia and hemoglobinopathies groups, and a parsimonious model (MCH and HbA₂) will perform equivalently to more complex models in predicting the presence of β -thalassemia trait.
- 3) A customized bioinformatics pipeline can be developed for the accurate identification of SNVs, indels, and segmental deletions and duplications in the *HBA* and *HBB* clusters using whole-exome sequencing (WES) data.
- 4) WES can be used to accurately validate novel and previously unidentified segmental deletions and duplications in the *HBA* cluster that were initially detected by MLPA.

1.4.2 Research Questions

- 1) What is the diversity of β -thalassemia and hemoglobinopathy mutations observed in families undergoing cascade screening?
- 2) Are there significant differences in hematological parameters among different types of β -thalassemia and hemoglobinopathies, and can a parsimonious model using only MCH and HbA₂ effectively predict the presence of β -thalassemia trait compared to more complex models?
- 3) What are the key components and steps involved in developing a customized bioinformatics pipeline for WES analysis of *HBA* and *HBB* clusters to accurately identify SNVs, indels, and segmental deletions and duplications?
- 4) What is the concordance rate between WES and MLPA in validating novel and previously unidentified segmental deletions and duplications in the *HBA* cluster?

1.5 Objectives

1.5.1 General Objective

To determine the spectrum of β -mutation and establish a bioinformatics pipeline to simplify DNA analysis of β -thalassemia.

1.5.2 Specific Objectives

- 1) To determine the spectrum of β -thalassemia and hemoglobinopathies mutations among families subscribing to cascade screening of β -thalassemia and hemoglobinopathies.

- 2) To analyze the correlation between RBC, Hb, MCV, MCH, RDW, HbA₂, HbF, and variant parameters in the various groups of β -thalassemia and variants and evaluate the performance of predictive models using RBC, MCV, MCH, and HbA₂ to identify β -thalassemia carriers.
- 3) To customize a bioinformatics pipeline for sequencing analysis of SNV, indels, and segmental deletions and duplications in *HBA* and *HBB* clusters.
- 4) To validate novel and previously unidentified segmental deletions and duplications in the *HBA* cluster detected by MLPA using WES.

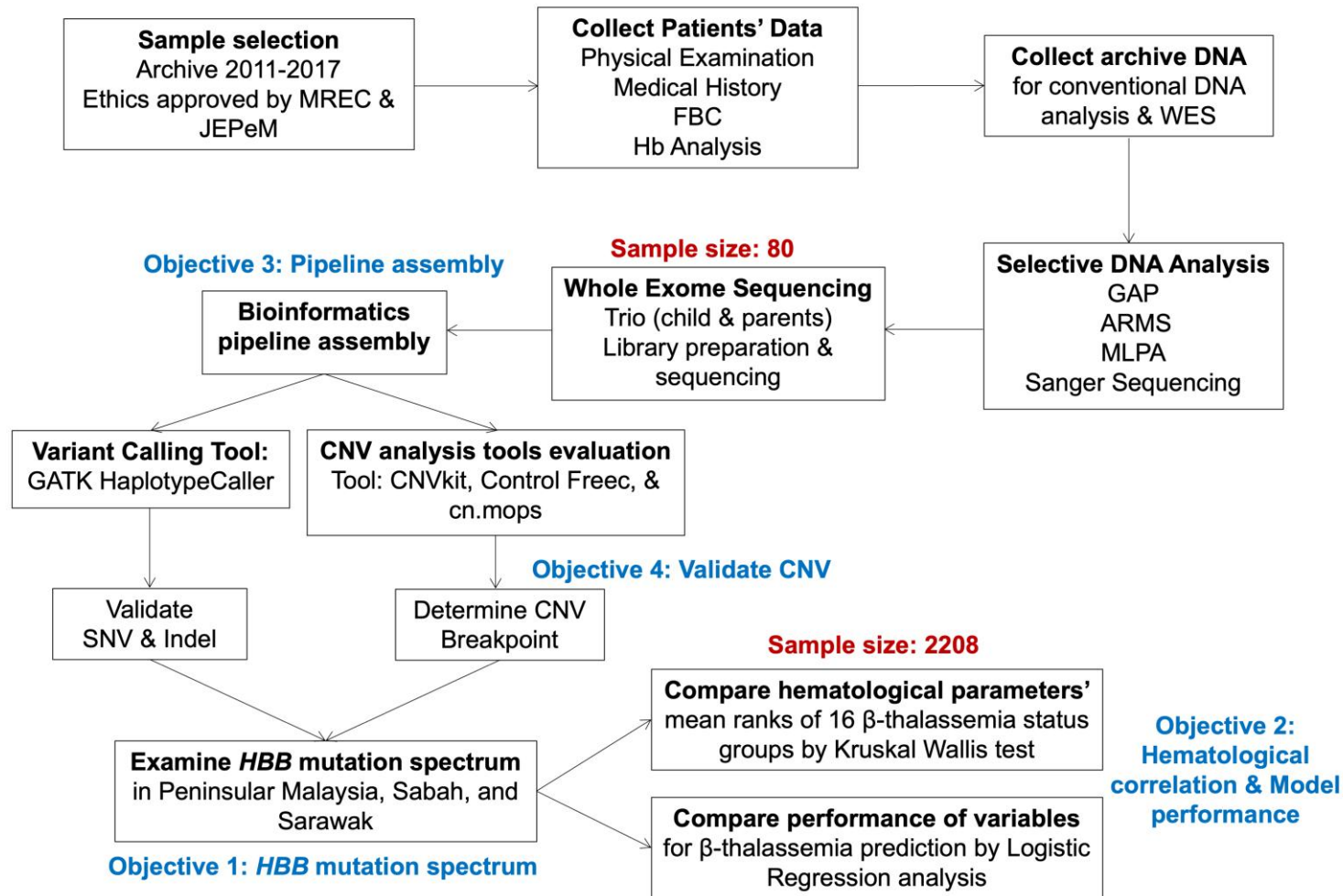


Figure 1.1 Study workflow

CHAPTER 2

LITERATURE REVIEW

2.1 Red blood cell

Blood cells are formed in the bone marrow by the process of erythropoiesis and arise from the same bone marrow stem cells. Stem cells are pluripotent and capable of developing into any blood cell type (Thomson et al., 1998). These immortal, undifferentiated, pluripotent stem cells give rise to erythrocytes, leukocytes (white blood cells, WBC), and platelets. Stem cells, under the influence of erythropoietin, develop into erythroblasts and start to synthesize Hb.

RBCs are deformable, allowing them to pass through capillaries that are a lot smaller in diameter than they are (Nikinmaa, 1990). They are characterized by a specific biconcave shape, high Hb content, and a lack of intracellular organelles such as the nucleus, mitochondria, or endoplasmic reticulum (Mazeron et al., 1997). They are densely packed with Hb molecules, which secures an O₂ transport capacity in the RBC. The O₂ binding and delivery properties of RBCs are guided by the allosteric properties of the Hb molecule.

The main role of Hb is to deliver oxygen (O₂) to the periphery while transporting carbon dioxide (CO₂) and other wastes generated by metabolizing tissues to the lungs. The oxygen is delivered in the form of oxygen-saturated Hb molecules (oxyhemoglobin) to the mitochondria, where it is used for aerobic respiration, while the oxygen-desaturated Hb (deoxyhemoglobin) collects CO₂ and returns to the lungs, where the CO₂ is released through exhalation (Perutz, 1978).

2.2 Hemoglobin

Globin is a globular protein molecule in several forms. Globin molecules are synthesized in the cytoplasm in parallel with the synthesis of heme. Each globin chain has a heme group attached. A heme group is a non-protein prosthetic group constituting an iron atom contained in the center of a large heterocyclic organic ring called a porphyrin (Casiday & Frey, 2007). A combination of heme and globin occurs at the histidine residue of the globin protein to form Hb. Both the heme group and the Hb protein undergo conformational changes upon the binding and unbinding of oxygen. Each Hb molecule is made up of two α -like globin chains and two β -like globin chains (Figure 2.1). The combination of tetra-globin chains (α , β , δ , γ , ζ , and ε -chains) determines the Hb type. Each chain has a different structure, electrical charge, electrophoretic mobility, and oxygen affinity.

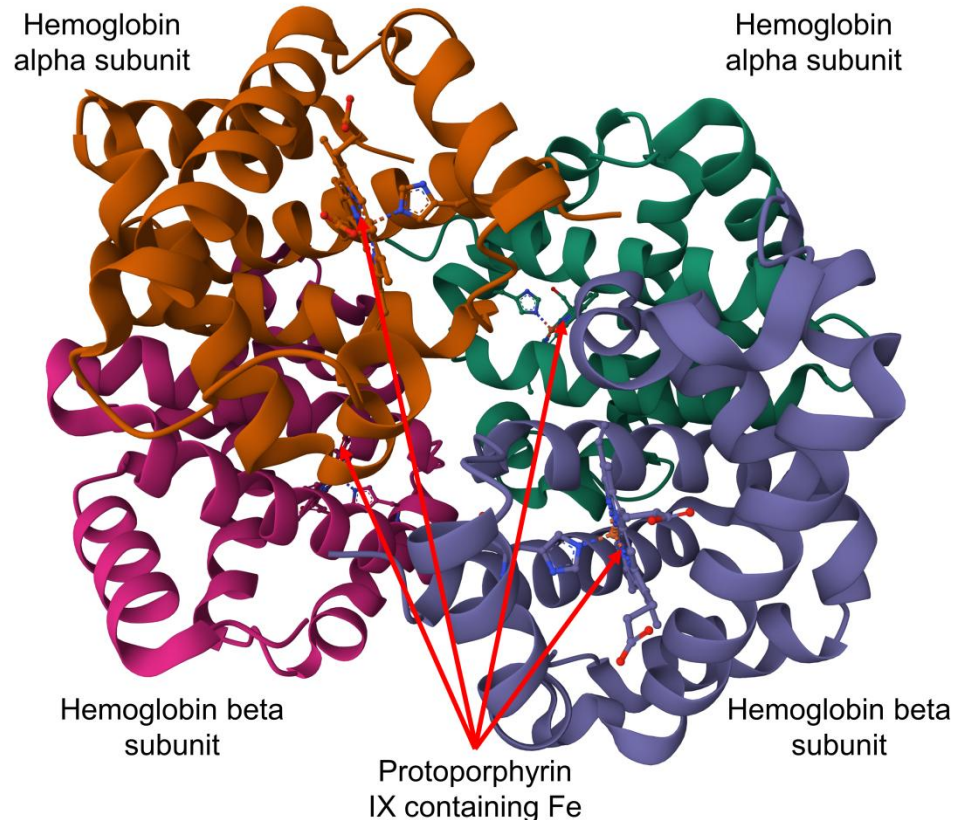


Figure 2.1 Molecular tetrameric structure of Hb with iron (heme) binding

2.2.1 Hemoglobin Structure and Switching

The human α -globin gene cluster is located near the telomeric of chromosome 16 spans and 26 kb and includes seven hemoglobin subunits: three functional α -like protein-coding genes of *HBZ* (zeta, ζ), *HBA2* (alpha 2, $\alpha 2$) and *HBA1* (alpha 1, $\alpha 1$), two expressed genes with unknown function *HBM* (mu, μ) and *HBQ1* (theta 1, $\theta 1$), and three pseudogenes of *HBAP1* (pseudo alpha 1, $\psi\alpha 1$), *HBAP2* (pseudo alpha 2, $\psi\alpha 2$), and *HBZP1* (zeta pseudogene 1, $\psi\zeta 1$). The three pseudogenes carry inactivating mutations and are not expressed. Previously classed as a pseudogene, microarray analysis of adult reticulocytes and cord blood samples of the *HBM* gene have demonstrated transcription of this gene (Goh et al., 2005).

The *HBA2* and *HBA1* have identical coding sequences. They differ slightly over the 5' untranslated regions and the introns and differ significantly over the 3' untranslated regions. The *HBA2* gene functions as the major α -globin gene in humans, encoding two to three times more protein than the *HBA1* gene (Liebhaber et al., 1986). Inclusive of the transcript and UTRs, *HBA2* is located at position chr16:172876-173710 with 835 sequence bases, while *HBA1* is located at position chr16:176680-177522 with 843 sequence bases (GRCh38.p13). Both genes have identical protein lengths of 142 amino acids.

The human β -globin gene cluster is located in the short arm of chromosome 11 (11p15.5q) and spans 81.7 kb. It contains the *HBB* (beta, β), the *HBD* (delta, δ), the *HBE1* (epsilon1, ϵ), the *HBG1* (A-gamma, $\text{A}\gamma$), the *HBG2* (G-gamma, $\text{G}\gamma$), and the *HBBP1* (pseudo beta, $\psi\beta$). The five functional globin genes are arranged in the order of their developmental expression. The expression of these genes is controlled by the locus control region (LCR) centromeric of the globin gene cluster (Hill et al., 1984). The *HBB* gene, which spans 1.6 kb and is located at position chr11: 5225464-

5227071, contains three exons, two intervening sequences (introns), and 5' and 3' untranslated regions (UTRs). The two *HBB* introns are 130 and 850 base pairs (bp), located between codons 30 and 31, and 104 and 105, respectively. Inclusive of the transcript and UTRs, *HBB* has 1608 sequence bases. It is regulated by an adjacent 5' promoter comprising TATA, CAAT, and duplicated CACCC boxes. Unlike the α -globin genes that are in duplicate forms and have four α -globin genes, each diploid cell contains only two β -globin genes. Inclusive of the start codon, the α - and β -globin chains contain 142 and 147 codons, respectively. The term codon refers to the sequence of three bases that determine the specificity of an amino acid to be included in a polypeptide chain.

The structure of Hb changes during the human development (Figure 2.2). Normal hemoglobin at all stages of life is a tetramer of two pairs of globin chains. Human embryonic hemoglobin comprises four embryonic hemoglobin: Hb-Gower I, Hb-Gower II, Hb-Portland I, and Hb-Portland II. Hb-Gower II consists of four polypeptide chains of two α - and ϵ -chains ($\alpha_2\epsilon_2$) (Huehns et al., 1964), while Hb-Gower I is composed of two ζ - and ϵ -chains ($\zeta_2\epsilon_2$) (Gale et al., 1979). The Hb-Portland I is composed of two ζ - and γ -chains ($\zeta_2\gamma_2$) (Sinhaseni et al., 1954), while the Hb Portland-II is composed of two ζ - and β -chains ($\zeta_2\beta_2$) (Randhawa et al., 1984). Embryonic Hb is confined to the yolk-sac stage of development and is replaced by fetal hemoglobin (HbF) until shortly before term. Fetal hemoglobin comprising two α - and γ -chains ($\alpha_2\gamma_2$) is also produced early in pregnancy and gradually replaces the embryogenic hemoglobin and becomes the major hemoglobin from about eight weeks post-conception until six weeks after birth. Then, it is replaced by adult hemoglobin. In adult Hb, α -chains are combined with β -chains (HbA, $\alpha_2\beta_2$) and with δ -chains (HbA₂, $\alpha_2\delta_2$). The HbF is replaced by HbA and HbA₂ during the first year of life, and

traces of the HbF are retained in normal adults. Ninety-five to 99% of adult Hb is HbA, while the rest is HbA₂ (less than 3.3%), and 1% or less is HbF (Nathan & Oski, 1993).

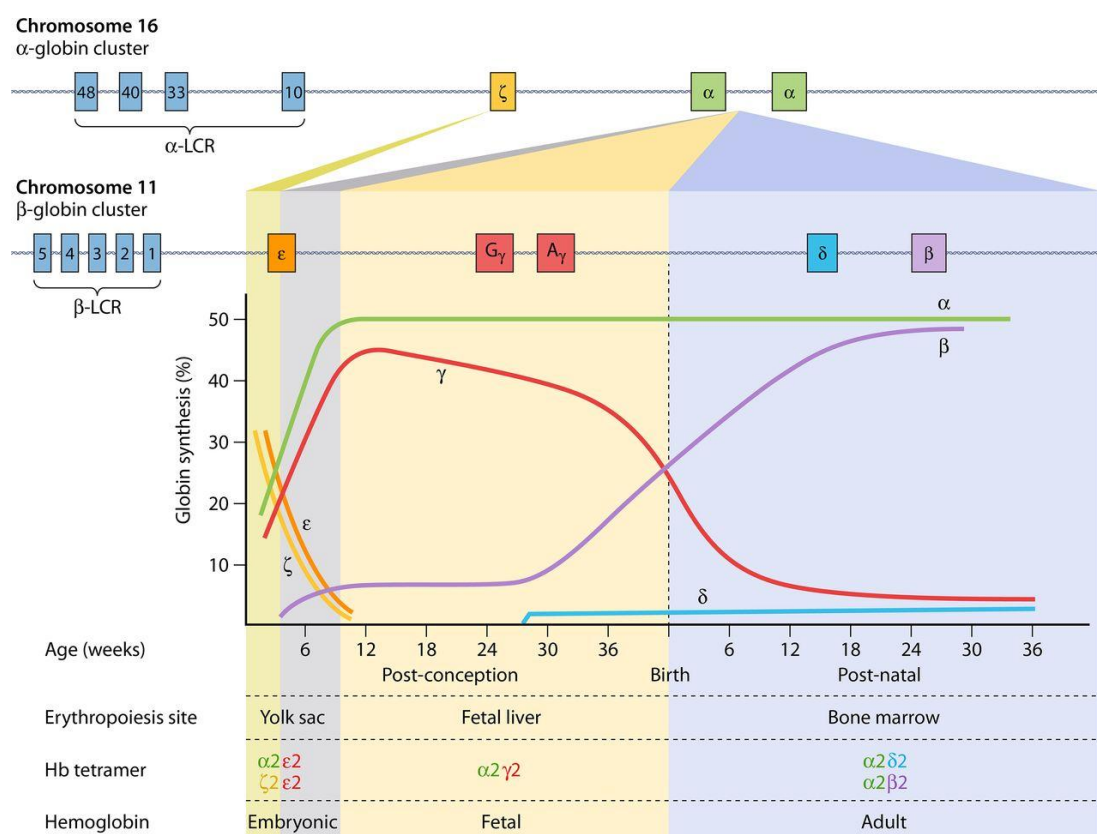


Figure 2.2 Hemoglobin Switching. The graph shows changes in globin gene expression during prenatal and postnatal development (Venkatesan et al., 2020)

2.3 Thalassemia and hemoglobinopathies

Thalassemia and hemoglobinopathies are genetic disorders caused by gene defects that hinder the normal production of hemoglobin, the major protein found in RBCs responsible for binding and carrying oxygen-rich RBCs from the lungs to the various body tissues. Normal adult hemoglobin consists of two pairs of α and β -chains, respectively. The synthesis of these proteins is coordinated to ensure equal production levels in erythropoietic cells (Ribeil et al., 2013). When one or both copies fail to produce normal β -globin, the α -globin gene continues its normal α -globin production. The consequence of these mutations is an imbalance of α/β -globin chain

synthesis, evidently in the homozygous forms, leading to the accumulation of free α -globin chains forming highly toxic aggregates (Khandros et al., 2012). Mutations in the *HBA* and the *HBB* are associated with α - and β -thalassemia, or variants, respectively.

2.4 *HBB* gene and cluster mutations

Mutations in the *HBB* clusters are extremely heterogeneous and are classified according to the phenotypes of thalassemia, variants, HPFH, and $\delta\beta$ -thalassemia. The most common type is β -thalassemia. β -thalassemia is caused by nucleotide substitutions, insertions, deletions, and segmental deletions of the *HBB* cluster. Nucleotide substitutions accounted for the majority. Most of the base changes are in the coding regions of the *HBB* gene.

2.4.1 β -globin hemoglobinopathies

This thalassemia phenotype forms structural β -globin chain variants if they are synthesized at a reduced rate, known as β -thalassemic hemoglobinopathies, and is caused by base changes on the coding regions of the *HBB* gene. Although there are many structural Hb variants, most of them are rare, and only three, Hb S, C, and E, are common (Weatherall & Clegg, 2001a).

2.4.2 β -thalassemia

β -thalassemia is a heterogeneous group of autosomal recessive disorders resulting in a genetically deficient synthesis of the β -globin chain in Hb (Thein, 2005). The main pathophysiologic mechanism is the imbalance between α -chains and their tetramer, β -globin chains. The β -globin chain's deficit leads to excess α -globin chains, which precipitate in the form of inclusions that damage the erythroid precursors in the

bone marrow and spleen, causing ineffective erythropoiesis (Stamatoyannopoulos & Nienhuis, 1994; Steinberg et al., 2001).

β -thalassemia has the most heterogeneous mutations. In β -thalassemia, nucleotide substitutions can occur at 5' and 3' untranslated regions (UTR), coding sites, introns, and the poly-A tail of the gene. The most common coding site mutations are Cd 19 (AAC>AGC) Hb Malay, Cd 17 (AAG>TAG), and Cd 15 (TGG>TAG) (Hassan et al., 2013). Meanwhile, CAP +1 (A>C), -28 (A>G), and Poly A (AATAAA>AATAGA) are common UTR mutations. Mutations on the intronic site adjacent to the codons lead to abnormal mRNA splicing; however, intronic mutations away from the codons may also have the same effects as in the common deep intronic variant, IVS 2-654 (C>T). Common intronic mutations are IVS 1-5 (G>C) and IVS 1-1 (G>T) (Hassan et al., 2013). A small nucleotide insertion or deletion led to a reading frameshift mutation that changed the reading frame, resulting in a different translation from the original. They are termed indels. The most common indels are Cd 41/42 (-TTCT), Cd 35 (TAC>TA-), and Cd 8/9 (+G). Large deletions causing β -thalassemia are the least heterogeneous. β -Filipino is the most frequent deletional β -thalassemia in Malaysia (Teh et al., 2014; Thong et al., 1999).

2.4.3 HPFH and $\delta\beta$ -thalassemia

HPFH and $\delta\beta$ -thalassemia are characterized by the deletion of genomic sequences downstream of the γ -globin gene comprising $\psi\beta$ -, δ -, and β -globin genes, resulting in the persistence of fetal γ -globin expression into adulthood. The deletions vary in size. They are categorized into High Persistence Fetal Hemoglobin (HPFH) and $\delta\beta$ -thalassemia, and the latter is associated with more severe phenotypes. Amongst the most common HPFH and $\delta\beta$ -thalassemia are HPFH-6, $\delta\beta$ Thai 12.5 kb deletion, and $\gamma\delta\beta$ Siriraj I 118 kb deletion.

2.4.4 δ -thalassemia and variants

δ -thalassemia and variants occur when there is a mutation in the *HBD* gene. They are not pathogenically relevant. However, concomitant β - and δ -thalassemia reduce the HbA₂ percentage to a normal level, masking the diagnosis of the β -thalassemia state. A recent study found heterogeneous δ -thalassemia and variants where, the most common were variant HbA₂-Indonesia (HBD:c.208G>C) and δ -thalassemia HbA₂-Deventer (Hassan et al., 2019). Co-inheritance of δ -thalassemia lowers HbA₂ levels, a β -thalassemia carrier may present with a borderline HbA₂ concentration, leading to a misdiagnosis. Therefore, diagnosing β -thalassemia in individuals with borderline HbA₂ is essential.

2.5 *HBA* gene and cluster mutations

Mutations in the *HBA* clusters are heterogeneous. α -variants are the most heterogeneous; however, α -thalassemia is more common in many populations. α -thalassemia is caused by nucleotide substitutions, insertions, deletions, and segmental deletions of the *HBA* cluster. Unlike β -thalassemia, large deletions are the most common mutation forms.

2.5.1 α -thalassemia and variants

Normal individuals have two α -globin genes per haploid genome, and the α -globin genotype is represented as $\alpha\alpha/\alpha\alpha$. In α -thalassemia, deletional forms are more common; thus, mutations are commonly classified into deletional and non-deletional α -thalassemia. The deletional α -thalassemia of $-\alpha^{3.7}$, $--^{SEA}$, is the most common. Nucleotide substitutions can occur at 5' and 3' untranslated regions (UTR), coding sites, introns, and the poly-A tail of the gene. Coding regions' base substitutions led to most variants, while mutations on the UTRs, introns, and poly-A tail produced α -thalassemia phenotypes. Although variants are heterogeneous, only several types, such

as α^{CS} and α^{Adana} are common. Poly A (AATAAA>AATA--) is known to produce a severe phenotype. Table 2.1 depicts the types of β -thalassemia mutations and their effects on the gene.

Table 2.1 Types of mutations and their effects, as well as the commonly genotyped β -thalassemia and variants among Malaysians

Type	Cause	Effect	Common mutation
Point mutation (single nucleotide substitution)	Missense mutation	Results in a protein in which one amino acid is substituted for another	Cd 19 (AAC>AGC) Hb Malay & Cd 26 (GAG>AAG) Hb E
	Nonsense mutation	A stop codon replaces an amino acid codon, leading to premature termination of translation	Cd 17 (AAG>TAG)
Indel (Insertion or deletion)	Frameshift mutation	Causes a change in the reading frame, leading to the introduction of unrelated amino acids into the protein	Cd 41/42 (-TTCT) Cd 71/72 (+A)
Structural variation	Segmental deletion	Loss of large chunks of DNA leading to loss of gene function	β -Filipino deletion, $\delta\beta$ -Thai, HPFH-6
	Segmental duplication	Increase in the number of copies of a gene, change $\alpha:\beta$ ratio	<i>HBG</i> duplication (benign), $\alpha\alpha\alpha^{\text{anti3.7}}$, $\alpha\alpha\alpha^{\text{anti4.2}}$

2.6 Thalassemia Phenotypes

Both α - and β -thalassemia phenotypes are classified separately. α -thalassemia is classified into four phenotypes; silent carrier, trait, HbH, and Hb Bart's. Meanwhile, β -thalassemia is classified into four phenotypes; silent, minor, intermedia, and major.

2.6.1 α -thalassemia silent carrier, trait, HbH, and Hb Bart's

It is recognized that α -thalassemia status is the result of reduced α -globin gene expression. It is caused by either deletional or non-deletional mechanisms. Single gene deletion or variant results in α -thalassemia silent carrier status. The 4.2 kb

deletion of DNA (leftward type, $-\alpha^{4.2}$) and the deletion of 3.7 kb of DNA (rightward type, $-\alpha^{3.7}$) have been identified in our population (Ahmad et al., 2012, 2013).

α -thalassemia trait (minor) is characterized by having two residual functional α -gene, and two mutated α -globin genes either in cis (double mutations of α -gene at the same chromosome, $--/\alpha\alpha$) or in trans (one mutated α -gene in each chromosome, $-\alpha/-\alpha$). Carriers of unstable α -variants are classified in this phenotype. The deletions include the SEA, ($--^{SEA}$), the Thai ($--^{THAI}$), and the Filipino ($--^{FIL}$) deletions, homozygous $-\alpha^{3.7}$ ($-\alpha^{3.7}/-\alpha^{3.7}$), and variants such as α^{CS} and α^{Adana} . They are presented with mild hypochromic microcytic anemia.

A three-gene deletion results in significant production of hemoglobin H (HbH), which has four β -chains (β_4). HbH can be caused by a compound heterozygous gene deletion (three gene deletions), or two gene deletions with a point mutation such as α^{CS} and α^{Adana} (non-deletional HbH). However, the latter are generally more severe. While most HbH patients do not require transfusions, the clinical presentation can be heterogeneous. HbH individuals may typically have hepatosplenomegaly, variable forms of jaundice, gallstones, and hemolytic crisis. Inheritance of double two α -globin gene deletions such as homozygous $--^{SEA}$, homozygous $--^{THAI}$, or compound heterozygous $--^{SEA}/--^{FIL}$, result in a total absence of the α -globin gene production giving rise to the Hb Bart's ($--/--$) hydrops fetalis. This phenotype has also been reported in fetuses with homozygous α^{Adana} (Nainggolan et al., 2010).

2.6.2 β -thalassemia trait, intermedia, and major

β -thalassemia minor or trait refers to carriers of a β -thalassemia mutation with a normal β -globin on the other allele. Carriers of β -thalassemia are clinically

asymptomatic. β -thalassemia major is characterized by infancy-onset severe anemia that requires lifelong blood transfusion for survival (Galanello & Origa, 2010). In β -thalassemia major, the excess unpaired α -globin chains aggregate to form inclusion bodies that damage RBC membranes, leading to intravascular hemolysis, damage and apoptosis of erythroid precursors, causing ineffective erythropoiesis (Yuan et al., 1993).

β -thalassemia intermedia is defined as the condition of having an intermediate condition between minor and major (Nathan & Oski, 1993). Most patients with β -thalassemia intermedia are homozygotes or compound heterozygotes for β -thalassemia (Galanello & Cao, 1988). Rarely, simple carrier of β -thalassemia can be symptomatic, such as in the case of co-inheritance of segmental duplication of the α -chain, which increases the imbalance ratio of β - and α -tetramer resulting in a more severe phenotype (Goossens et al., 1980; Henni et al., 1985). Conversely, inheritance of two severe β -thalassemia alleles may also lead to β -thalassemia intermedia phenotype when they co-inherited genetic modifiers that reduce the imbalance β - and α -tetramer, thus phenotype severity (Weatherall & Clegg, 2001b). Occasionally β -thalassemia intermedia have been used to describe severe Hb H disease.

2.6.3 Non-transfusion Dependent Thalassemia (NTDT) and Transfusion-dependent Thalassemia (TDT)

The separate phenotype classification of α - and β -thalassemia can be perplexing, especially for patients with a complex genotype such as concomitant α - and β -thalassemia. Thus, new symptomatic thalassemia called Transfusion-dependent Thalassemia (TDT) and Non-transfusion Dependent Thalassemia (NTDT) was proposed primarily to better classify various α - and β -thalassemia and variants

(Weatherall, 2012). The TDT refers to patients requiring lifelong regular blood transfusions to survive (thalassemia major) (Rachmilewitz & Giardina, 2011).

NTDT is a term to describe patients who do not require such lifelong regular transfusions for survival, although they may require occasional or even frequent transfusions in certain clinical settings and for defined periods (Musallam et al., 2013). NTDT encompasses five clinically distinct forms: β -thalassemia intermedia, hemoglobin E/ β -thalassemia (mild and moderate forms), α -thalassemia intermedia (Hb H disease), hemoglobin S/ β -thalassemia, and hemoglobin C thalassemia (Weatherall, 2012).

2.6.4 Genetic Modifiers of thalassemia

2.6.4(a) Genetic Modifier of β -thalassemia

Thalassemia severity is extremely diverse, from the clinically asymptomatic β -thalassemia carriers to the severe ends of TDT (Thein, 2005). Modifier genes are defined as genetic variants that lead to differences in the disease phenotype. They are classified into three groups: primary, the mutations of the β -globin genes; secondary, loci that are directly involved in globin synthesis; and tertiary, loci that are not involved in globin synthesis but that might modify the disease severity. Primary modifiers are presented by β -thalassemia phenotypes that range from silent carriers to homozygous (recessive), heterozygous (dominant), or compound heterozygous inheritance of the principal forms of the disease (Weatherall, 2001). Mutations can either cause a reduction or absence of β -globin chain synthesis and production (Thein, 2012).

The secondary modifiers are those that change the degree of globin chain imbalance in β -thalassemia (Musallam et al., 2013). They are associated with

concomitant α -thalassemia among homozygous β -thalassemia that decreased α /non- α -chain imbalance and disease severity (Galanello & Cao, 1988), or co-inheritance of α -globin duplication that increased α /non- α -chain imbalance and its severity (Goossens et al., 1980), and elevated production of γ -globin chains (HbF) in adulthood (Cao & Galanello, 2010). The most common forms of copy number gains in the α -globin genes are the $\alpha\alpha\alpha^{\text{anti3.7}}$ and $\alpha\alpha\alpha^{\text{anti4.2}}$. However, homozygous triplicated α - ($\alpha\alpha\alpha/\alpha\alpha$) or heterozygous quadruplicated α -globin genes ($\alpha\alpha\alpha\alpha/\alpha\alpha$) in β -thalassemia carriers did not increase the phenotype beyond thalassemia intermedia (Galanello et al., 1983; Thompson et al., 1989).

HbF is one of the most important modifiers because it can compensate for the lack of β -globin chains (HbA). Persistence HbF is caused by non-deletional HPFH mutations or genetic polymorphism involved in HbF regulation. The persistence of HbF expression in adulthood is caused either by mutations that trigger the binding of transcriptional activators of γ -globin or mutations that disrupt the binding sites of γ -globin repressors (Venkatesan et al., 2020).

HbF level is regulated by three major loci: HBG2:g.-158C>T on 11p15.4, *HBSIL-MYB* intergenic region on 6q23.3, and BAF Chromatin Remodeling Complex Subunit (*BCL11A*) on 2p16.1. *BCL11A* interacts with sex-determining region Y(SRY)-Box Transcription Factor 6 (SOX6) to repress the transcriptional of γ -globin genes (Xu et al., 2010). Of the three quantitative trait loci (QTL), *BCL11A* had the strongest effect on HbF/F cell levels, with 15.1 % of the trait variance (Menzel et al., 2007), and up to 12% in Sickle Cell Anemia (SCA) patients (Lettre et al., 2008; Sedgewick et al., 2008).