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

LIST OF TABLES

Page

LIST OF FIGURES

Page

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	apoliprotein 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
BCL11A	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
COLIA1	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
HBA	Hemoglobin subunit alpha
HBA1	Hemoglobin subunit alpha 1
HBA2	Hemoglobin subunit alpha 2
HbA_2	Hemoglobin A_2
HBAP1	Pseudo alpha 1
HBAP2	Pseudo alpha 2
HBB	Hemoglobin subunit beta
HBBP1	Hemoglobin subunit beta pseudogene 1
HBD	Hemoglobin subunit delta
Hb C	Hemoglobin C
Hb E	Hemoglobin E
HBE1	Hemoglobin subunit epsilon
HbF	Fetal Hemoglobin
HBG1	Hemoglobin subunit gamma 1
HBG2	Hemoglobin subunit gamma 2
HbH	Hemoglobin H disease
HBM	Hemoglobin subunit mu
HBQ1	Hemoglobin subunit theta 1
Hb S	Hemoglobin Sickle
HBS1L	HBS1 Like Translational GTPase
HbVar	Database of Human Hemoglobin Variants and Thalassemia Mutations
HBZ	Hemoglobin Subunit Zeta

HBZP1	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
LUC7L	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
POLR3K	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
SOX6	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
SUPT5H	SPT5 Homolog, DSIF Elongation Factor Subunit
SVD	singular value decomposition
TA	Ta repeat
TDT	Transfusion Dependent Thalassemia

TGFB1	Transforming growth factor β1
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
UGT1A	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

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 HbA_2 (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 HbA_2), Model 2 (RBC, MCH, dan HbA_2), dan Model 3 (MCV, MCH, dan HbA_2) tidak meningkatkan ramalan pembawa β-talasemia secara signifikan. Untuk memudahkan proses genotip, saluran HaplotypeCaller dan CalculateGenotypePosteriors daripada The Genome Analysis Toolkit (GATK) digunapakai 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 $\alpha\alpha\alpha^{91}$ dan *de novo* $\alpha\alpha\alpha^{102}$. 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 HbA_2 (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 HbA_2), Model 2 (RBC, MCH, and HbA_2), and Model 3 (MCV, MCH, and HbA_2) 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 HbA_2 (Model 4) is compared to more comprehensive models including RBC, MCV, MCH, and HbA_2 (Model 1), MCV, MCH, and HbA_2 (Model 2); and RBC, MCH, and HbA_2 (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 HbA_2 levels. Although not pathogenically relevant, the co-occurrence of δ-thalassemia or variants in β-thalassemia carriers reduces the HbA_2 levels, potentially masking the diagnosis of β-

thalassemia. Locally, the most common δ-thalassemia and variant was HbA2-Indonesia (HBD:c.208G>C) [δcd 69 (GGT>CGT)] and HbA2-Deventer (HBD:c.202G>A) [δcd 67 GTG>ATG)] (Hassan et al., 2019). Since δ-thalassemia modifies HbA_2 levels, diagnosing borderline HbA_2 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* (α2 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 HbA_2. 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 HbA_2, (Model 1) and three simpler models with MCV, MCH, and HbA_2 (Model 2), RBC, MCH, and HbA_2 (Model 3), and MCH and HbA_2 (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 HbA_2 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 HbA_2) 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 HbA_2 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, HbA_2, 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 HbA_2 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.

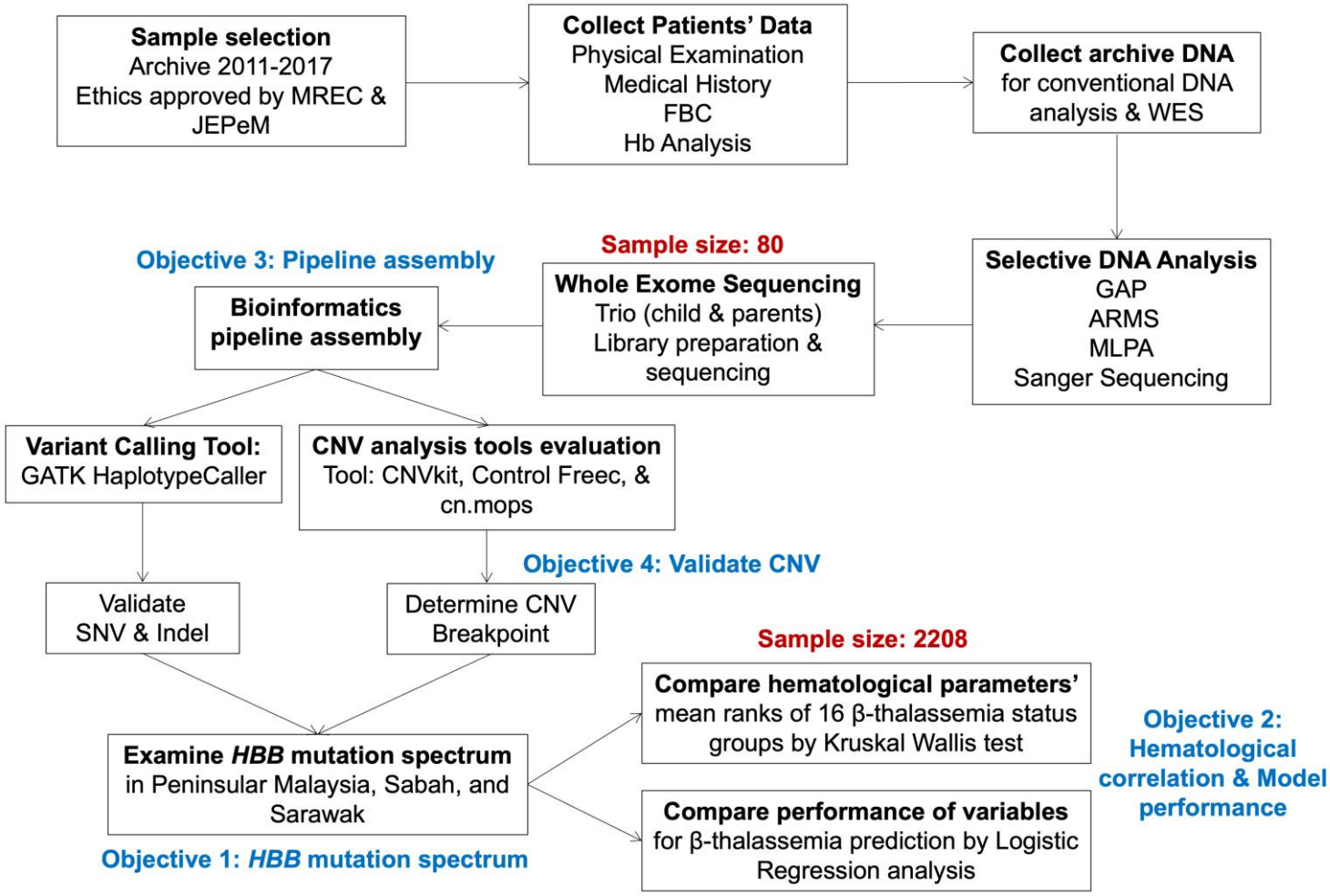

Figure 1.1 Study workflow

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 O_2 transport capacity in the RBC. The O_2 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 (O_2) to the periphery while transporting carbon dioxide (CO_2) 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 CO_2 and returns to the lungs, where the CO_2 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.

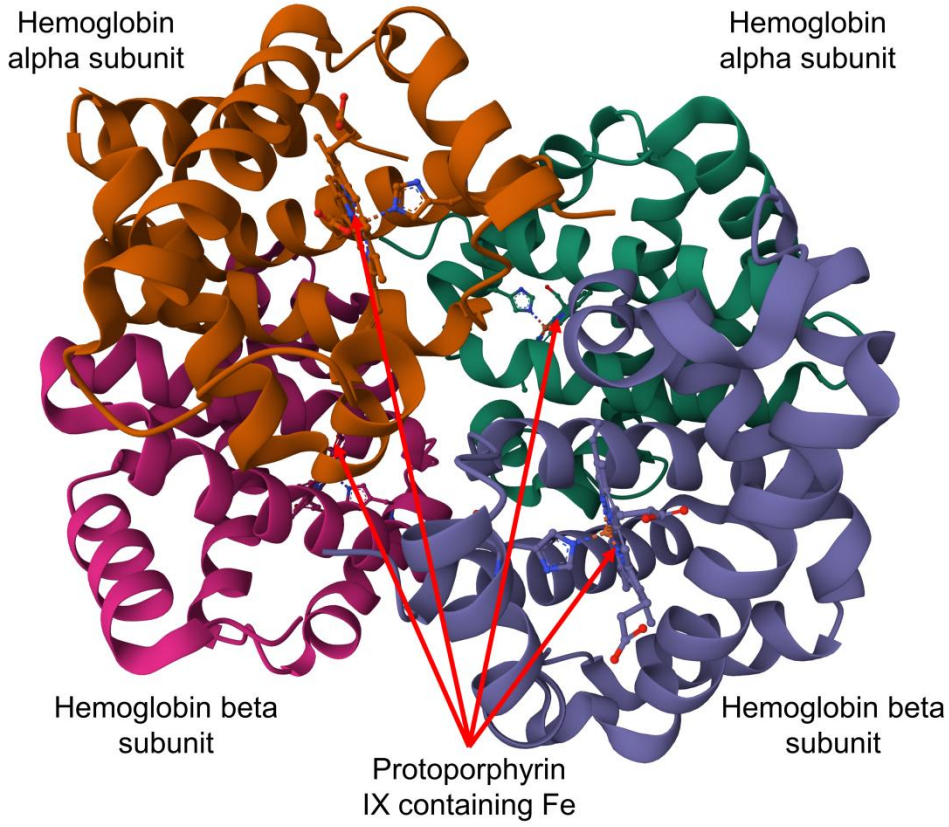

Figure 2.1 Molecular tetrameric structure of Hb with iron (heme) binding

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, α2) and *HBA1* (alpha 1, α1), two expressed genes with unknown function *HBM* (mu, μ) and *HBQ1* (theta 1, θ1), and three pseudogenes of *HBAP1* (pseudo alpha 1, ψα1), *HBAP2* (pseudo alpha 2, ψα2), and *HBZP1* (zeta pseudogene 1, ψζ1). 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 a-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 *HBE*1 (epsilon1, ε), the *HBG1* (A-gamma, $^{A}\gamma$), the *HBG2* (G-gamma, $^{G}\gamma$), and the *HBBP1* (pseudo beta, ψβ). 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 (α2ε2) (Huehns et al., 1964), while Hb-Gower I is composed of two ζ- and ε-chains (ζ2ε2) (Gale et al., 1979). The Hb-Portland I is composed of two ζ- and γ-chains (ζ2γ2) (Sinhaseni et al.,1954), while the Hb Portland-II is composed of two ζ- and β-chains (ζ2β2) (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 (α2γ2) 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, α2β2) and with δ-chains (HbA_2, α2δ2). The HbF is replaced by HbA and HbA_2 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 HbA_2 (less than 3.3%), and 1% or less is HbF (Nathan & Oski, 1993).

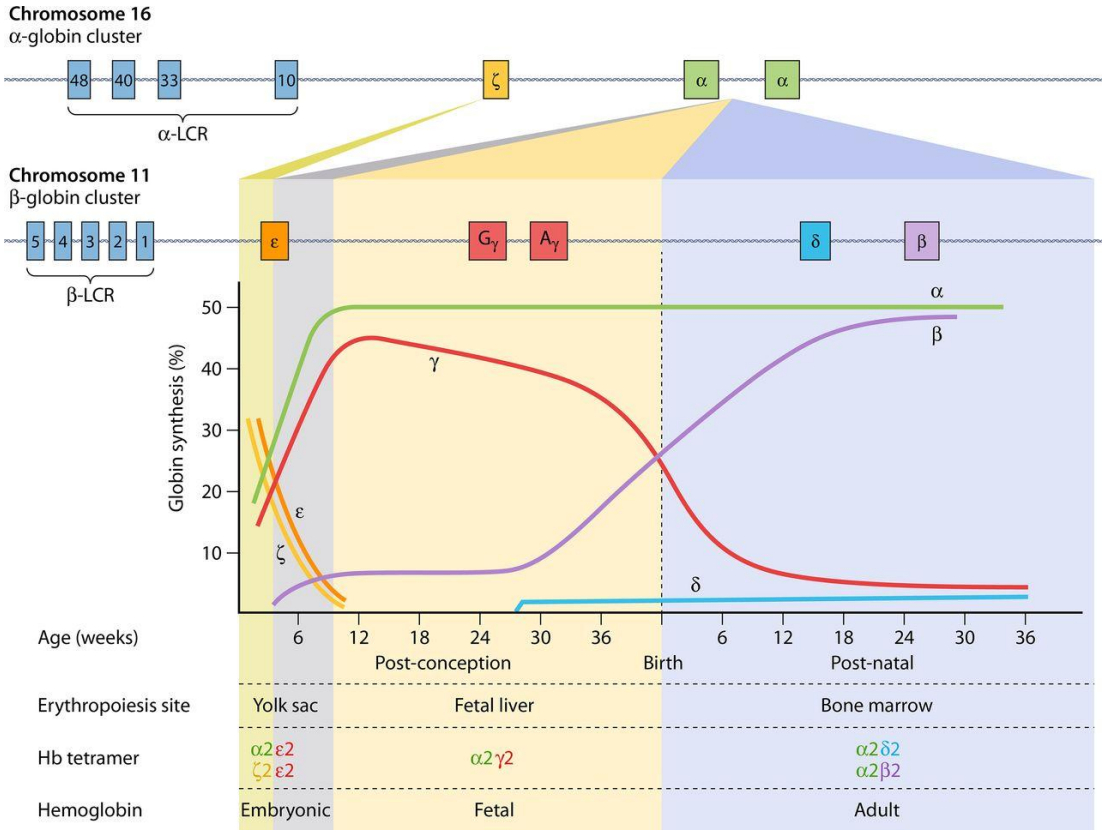

Figure 2.2 Hemoglobin Switching. The graph shows changes in globin gene expression during prenatal and postnatal development (Venkatesan et al., 2020)

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 δβ-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 δβ-thalassemia

HPFH and δβ-thalassemia are characterized by the deletion of genomic sequences downstream of the ^{A}γ-globin gene comprising ψβ-, 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 δβ-thalassemia, and the latter is associated with more severe phenotypes. Amongst the most common HPFH and δβ-thalassemia are HPFH-6, δβ Thai 12.5 kb deletion, and γδβ 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 HbA_2 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 HbA_2 levels, a β-thalassemia carrier may present with a borderline HbA_2 concentration, leading to a misdiagnosis. Therefore, diagnosing β-thalassemia in individuals with borderline HbA_2 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 αα/αα. 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, δβ-Thai, HPFH-6
	Segmental duplication	Increase in the number of copies of a gene, change α:β ratio	*HBG* duplication (benign), $\alpha\alpha\alpha^{anti3.7}$, $\alpha\alpha\alpha^{anti4.2}$

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, --/αα) or in trans (one mutated α-gene in each chromosome, -α/-α). 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 (β4). 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 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 $\alpha\alpha\alpha^{anti3.7}$ and $\alpha\alpha\alpha^{anti4.2}$. However, homozygous triplicated α- (ααα/ααα) or heterozygous quadruplicated α-globin genes (αααα/αα) 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 Hb F 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, *HBS1L-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).