

LONDON  
SCHOOL *of*  
HYGIENE  
& TROPICAL  
MEDICINE



**Population genetic structure and genomic  
divergence in *Plasmodium knowlesi***

**Paul Cliff Simon Divis**

Thesis submitted in accordance with the requirements for the degree  
Doctor of Philosophy (PhD) of the University of London

**March 2017**

**Department of Pathogens and Molecular Biology  
Faculty of Infectious and Tropical Diseases  
LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE**

Funded by Ministry of Higher Education, Malaysia

## Declaration

I, Paul Cliff Simon Divis, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: 

Date: March 2017

## Abstract

*Plasmodium knowlesi* infections in humans have been increasingly seen in many countries across Southeast Asia, with cases mainly concentrated in Malaysia, since a major focus of infections was first described in Malaysian Borneo over 10 years ago. Clinical presentations show a wide spectrum of illness from mild to fatal, with the possible occurrence of asymptomatic infections. Two monkey species have been identified as the chief reservoir hosts; long-tailed macaque (*Macaca fascicularis*) and pig-tailed macaque (*M. nemestrina*). In order to explore the transmission of *P. knowlesi* infections, it is important to study the population genetic structure of this parasite.

To address this, a microsatellite genotyping toolkit consisting of 10 loci specific for *P. knowlesi* was developed and validated. Using these highly polymorphic markers, analysis of more than 500 *P. knowlesi* infections from human and wild macaque hosts across Malaysian Borneo and humans of peninsular Malaysia showed remarkable population genetic structure. Human clinical isolates were shown to comprise highly divergent subpopulations, respectively associated with forest-dwelling long-tailed macaque (Cluster 1) and pig-tailed macaque (Cluster 2) reservoir hosts. After analysis of initial whole genome sequence data, re-assessment of population genetic structure was undertaken by microsatellite analysis of more samples from wild macaques and humans from peninsular Malaysia, showing profound geographical divergence between Borneo (sympatric Cluster 1 and Cluster 2) and mainland peninsular Malaysia (Cluster 3). The overall three major subpopulations demonstrated by microsatellite

analyses matched the analysis inferred by the genome-wide sequence analysis of clinical isolates.

To allow further investigation of variation in genome-wide divergence between the sympatric subpopulations in Borneo, a simple laboratory kit consisting of allele-specific PCR assays was developed to distinguish the two subpopulations. This eased in identifying to which subpopulations *P. knowlesi* infections belonged, and subsequently generating more genome-wide sequences for comparative study. Further analyses revealed remarkable heterogeneity in the level of divergence between the sympatric subpopulation across the genome. Genomic architectures showed 20 high-divergence regions scattered in different chromosomes. These findings suggest independent adaptation of parasites in different macaque hosts that persist sympatrically.

## Acknowledgements

Completing this thesis has been the most challenging part in my life. I would not be able to make it without the support, prayer and guidance from a few special people. I would like to express my greatest gratitude to the following:

- My supervisor, Professor David Conway, for giving me the support and guidance as well as training me to be a good scientist.
- Professor Balbir Singh at the Malaria Research Centre, UNIMAS for being a wonderful mentor throughout my research career in malaria.
- Samuel Assefa and Craig Duffy for providing assistance and guidance pertaining to R and perl scripts; Lindsay Stewart for making sure that the laboratory was inhabitable; and the rest of the Conway group - Lee, Harvey and Sarah.
- Ozan Gundogdu and Eloise Thompson for assisting in specialised laboratory equipment at LSHTM; Dayang Shuaisyah and Khamisah Kadir at the Malaria Research Centre, UNIMAS for laboratory assistance in Malaysia.
- Clemens Kocken and his team at the Biomedical Primate Research Centre, Rijswijk, The Netherlands for providing DNA controls of all macaque *Plasmodium* species.
- Colin Sutherland for providing DNA controls of human *Plasmodium* species, and also his team, especially Khalid Beshir and Mary Oguike for giving permission to use their specialised laboratory equipment.
- Inke Lubis and Grace Dacuma for a great companionship during coffee break and conferences – I am going to miss our sharing of ups and downs during our PhD studies.
- Mirza, Ayu, Azmin, Rekaya, Ajis, Edris, Saloma, Sylvester, Cliff and Isabel: thank you for the friendship and encouragement to push me in completing this PhD thesis.
- Finally, to my family, especially my parents Simon Divis and Mune Robert, for the prayers, support and blessings throughout my life.

*And God is able to bless you abundantly, so that in all things at all times, having all that you need, you will abound in every good work – 2 Corinthians 9:8*

## Table of Contents

Declaration .....	2
Abstract .....	3
Acknowledgements .....	5
Table of Contents .....	6
List of Figures .....	10
List of Tables .....	13
Abbreviations .....	15
Publications .....	18
<b>Chapter One: Introduction</b> .....	<b>20</b>
1.1 Malaria burden worldwide .....	21
1.2 Malaria parasites and life cycles .....	22
1.2.1 <i>Plasmodium</i> species .....	22
1.2.2 Life cycles .....	23
1.3 Epidemiology of <i>Plasmodium knowlesi</i> infections .....	26
1.4 The genome of <i>Plasmodium knowlesi</i> .....	30
1.5 Population genetic structure of <i>Plasmodium</i> species in Southeast Asia .....	32
1.5.1 <i>P. falciparum</i> and <i>P. vivax</i> .....	32
1.5.2 <i>P. knowlesi</i> .....	33
1.6 Genetic markers for studying the population substructure .....	34
1.6.1 Microsatellites .....	35
1.6.2 Single nucleotide polymorphisms from whole genome sequencing data .....	36
1.7 Hypothesis and objectives .....	38
1.7.1 Hypothesis .....	38
1.7.2 Specific objectives .....	38
<b>Chapter Two: Divergent <i>Plasmodium knowlesi</i> in human populations associated with different macaque host species</b> .....	<b>40</b>
2.1 Introduction .....	42
2.2 Materials and methods .....	43
2.2.1 <i>P. knowlesi</i> samples from humans and macaques .....	43

2.2.2	Development of <i>P. knowlesi</i> microsatellite genotyping markers .....	44
2.2.3	PCR and genotyping protocols .....	46
2.2.4	Multiplicity of infection .....	47
2.2.5	Genetic diversity and population divergence .....	47
2.2.6	Haplotype relatedness and linkage disequilibrium .....	47
2.2.7	Assessment of population genetic substructure .....	48
2.3	Results .....	49
2.3.1	<i>P. knowlesi</i> microsatellites as genetic markers for population studies	49
2.3.2	Host-dependent genetic structure of <i>P. knowlesi</i> .....	54
2.3.2.1	Inter-host and intra-host diversity .....	54
2.3.2.2	Population genetic substructure .....	56
2.3.3	Geographical population genetic structure of <i>P. knowlesi</i> .....	58
2.3.3.1	Diversity among geographical sites .....	58
2.3.3.2	Population genetic substructure and population divergence ..	58
2.3.4	Evaluation of cluster assignment indices .....	66
2.4	Discussion .....	70

### **Chapter Three: Development of a simple genotyping assays for discriminating**

	<b>sympatric <i>Plasmodium knowlesi</i> subpopulations .....</b>	<b>74</b>
3.1	Introduction .....	76
3.2	Materials and methods .....	77
3.2.1	DNA samples .....	77
3.2.2	Single nucleotide polymorphism (SNP) of <i>P. knowlesi</i> .....	78
3.2.3	Development of allele-specific markers .....	78
3.2.3.1	Designing of PCR primers .....	78
3.2.3.2	Conventional PCR assay .....	79
3.2.3.3	Touchdown PCR assay .....	80
3.3	Results .....	81
3.3.1	Allele-specific markers to discriminate <i>P. knowlesi</i> subpopulations ....	81
3.3.2	Genotyping of <i>P. knowlesi</i> infections .....	84
3.4	Discussion .....	87

## Chapter Four: Three major divergent subpopulations of the malaria parasite

<i>Plasmodium knowlesi</i> .....	89
4.1 Introduction .....	91
4.2 Materials and methods .....	92
4.2.1 Study sites and DNA samples .....	92
4.2.2 Microsatellite genotyping of new samples .....	94
4.2.3 Analysis of microsatellite genotypes from previous data .....	94
4.2.4 Analyses of <i>P. knowlesi</i> population genetic substructure .....	95
4.3 Results .....	97
4.3.1 Genotypic diversity within <i>P. knowlesi</i> infections .....	97
4.3.2 Analysis of <i>P. knowlesi</i> population genetic structure with new samples .....	99
4.3.3 Analysis of population genetic structure incorporating new and previously acquired microsatellite data .....	101
4.3.4 Robustness and divergence of subpopulation clusters .....	103
4.4 Discussion .....	108

## Chapter Five: Genomic divergence between two sympatric *Plasmodium knowlesi* subpopulations .....

<i>Plasmodium knowlesi</i> subpopulations .....	112
5.1 Introduction .....	114
5.2 Materials and methods .....	115
5.2.1 DNA samples and genotyping assays .....	115
5.2.2 DNA library preparation .....	116
5.2.3 Whole genome sequencing and mapping on reference genome .....	117
5.2.4 Single nucleotide polymorphism (SNP) calling and filtration .....	118
5.2.5 Characterisation of nuclear genome-wide genetic patterns .....	119
5.2.5.1 Genomic diversity and population structure .....	119
5.2.5.2 Defining genomic regions .....	121
5.2.5.3 Diversity and signature of selection in genomic regions .....	122
5.2.6 Extra-chromosomal genomes .....	122
5.3 Results .....	123
5.3.1 Generation of new whole-genome sequence data, assembly and SNP calling .....	124



5.3.2 Population structure .....	126
5.3.3 Genome-wide patterns of variation .....	129
5.3.4 Genomic profiles of differentiation between subpopulations .....	132
5.3.5 Genomic regions of high and low divergence .....	134
5.3.6 Diversity of genomic regions between subpopulations .....	138
5.3.7 Phylogeny by extra-chromosomal genomes .....	143
5.3.7.1 Apicoplast genome .....	144
5.3.7.2 Mitochondrial genome .....	144
5.4 Discussion .....	148
<b>Chapter Six: General discussion and perspectives for the future</b> .....	151
<b>References</b> .....	159
<b>Appendices</b> .....	180
<b>Appendix 2.1</b> Map illustrates the macaque sampling sites from locations within a 30 km radius of Kapit town, Sarawak .....	180
<b>Appendix 2.2</b> Dataset – Genotypes at <i>P. knowlesi</i> microsatellite loci in infections from macaques (n = 47) and humans (n = 552) .....	181
<b>Appendix 2.3</b> Genotypic data of 10 pairs and one triplet of identical haplotypes detected in six geographical locations .....	182
<b>Appendix 2.4</b> Plots of delta <i>K</i> ( $\Delta K$ ) based on Evanno’s method for the determination of hypothetical ancestral population cluster ( <i>K</i> ) from the STRUCTURE analysis extracted using the STRUCTURE Harvester .....	183
<b>Appendix 4.1</b> Dataset - Additional datasets regarding genotypes in humans and macaques studied in Malaysian Borneo and peninsular Malaysia .	186
<b>Appendix 4.2</b> STRUCTURE analysis on 166 <i>P. knowlesi</i> infections across Malaysia and seven laboratory isolates obtained by 10 microsatellite loci .....	187
<b>Appendix 4.3</b> STRUCTURE analysis on 758 <i>P. knowlesi</i> genotypes obtained by 10 microsatellite loci .....	188
<b>Appendix 5.1</b> Summary of remapping the previously generated short read sequences of <i>P. knowlesi</i> against the version 2.0 reference genome .....	189

## List of Figures

<b>Figure 1.1</b>	Map showing malaria burden worldwide in 2000 and 2016 .....	22
<b>Figure 1.2</b>	Life cycle of mammalian <i>Plasmodium</i> species .....	24
<b>Figure 1.3</b>	Genetic recombination of malaria parasites in mosquito hosts ....	26
<b>Figure 1.4</b>	Geographical range of <i>P. knowlesi</i> infections in humans, mosquito vectors and macaques across Southeast Asia regions ..	29
<b>Figure 1.5</b>	Distribution of multi-gene families ( <i>SICAvar</i> and <i>KIR</i> ) and telomere-like repeats in 14 chromosomes of the <i>P. knowlesi</i> strain H version 1.0 genome .....	31
<b>Figure 2.1</b>	Map of sampling locations of 599 <i>P. knowlesi</i> infections genotyped in this study .....	45
<b>Figure 2.2</b>	Multiple genotype <i>P. knowlesi</i> infections and diversity among infections in three host species from Kapit .....	56
<b>Figure 2.3</b>	Population genetic structure of <i>P. knowlesi</i> from infections in three host species in Kapit .....	58
<b>Figure 2.4</b>	Diversity and genetic structure of <i>P. knowlesi</i> in human infections from nine different geographical locations .....	60
<b>Figure 2.5</b>	Correlation between degree of cluster admixture and multi- locus linkage disequilibrium (standardised index of association) .	63
<b>Figure 2.6</b>	Isolation-by-distance model and principal component analysis (PCA) of the human <i>P. knowlesi</i> isolates .....	65
<b>Figure 2.7</b>	Population genetic structure of <i>P. knowlesi</i> infections from all 512 humans, 34 long-tailed macaques and 10 pig-tailed macaques in Malaysia .....	67
<b>Figure 2.8</b>	Allele frequency distributions and genetic differentiations of 10 microsatellite loci between two <i>P. knowlesi</i> subpopulation clusters .....	68
<b>Figure 2.9</b>	Intermediate cluster assignment indices in <i>P. knowlesi</i> infections in humans and macaques .....	69
<b>Figure 3.1</b>	Example of a region in <i>P. knowlesi</i> genome for developing an allele-specific marker .....	80

<b>Figure 3.2</b>	Gel electrophoresis shows allele specificity of PCR primer sets C1A and C2J for discriminating <i>P. knowlesi</i> infections of Cluster 1 and Cluster 2 subpopulations, respectively .....	84
<b>Figure 3.3</b>	Proportions of isolates positive for the new PCR discrimination of Cluster 1 and Cluster 2 subpopulations for <i>P. knowlesi</i> infections in 355 infections from Malaysian Borneo .....	85
<b>Figure 3.4</b>	Proportions of isolates positive for the new PCR discrimination of <i>P. knowlesi</i> considered to be 'Cluster 1' and 'Cluster 2' subpopulations in 62 infections from Peninsular Malaysia .....	86
<b>Figure 4.1</b>	DNA samples of <i>P. knowlesi</i> infections derived from 134 humans and 48 macaques across Malaysia .....	93
<b>Figure 4.2</b>	Multiplicity of infections (MOI) in 134 human and 48 macaque hosts across Malaysia .....	99
<b>Figure 4.3</b>	Subpopulation cluster assignments of 166 individual <i>P. knowlesi</i> infections in human and macaque hosts across Malaysia and seven laboratory isolates .....	100
<b>Figure 4.4</b>	STRUCTURE analysis on 758 <i>P. knowlesi</i> genotypes using 10 microsatellite loci .....	102
<b>Figure 4.5</b>	Population genetic structure of combined 751 <i>P. knowlesi</i> infections across Malaysia and seven laboratory isolates .....	104
<b>Figure 4.6</b>	Principal coordinate analysis deduced from genetic distance matrix of 10 microsatellite loci in 751 <i>P. knowlesi</i> infections across Malaysia and seven laboratory isolates .....	106
<b>Figure 5.1</b>	Types of SNP alleles and distribution of minor allele frequency of biallelic SNPs across 14 chromosomes derived from 80 <i>P. knowlesi</i> isolate whole genome sequences .....	126
<b>Figure 5.2</b>	Population structure of <i>P. knowlesi</i> infections of 80 whole genome sequences .....	127
<b>Figure 5.3</b>	Genome-wide nucleotide diversity ( $\pi$ ) of two <i>P. knowlesi</i> subpopulation clusters in Malaysian Borneo .....	131
<b>Figure 5.4</b>	Genome-wide Tajima's <i>D</i> values of two <i>P. knowlesi</i> subpopulation clusters in Malaysia Borneo .....	131

<b>Figure 5.5</b>	Comparison of genome-wide Tajima's $D$ distributions between two <i>P. knowlesi</i> subpopulation clusters in Malaysian Borneo .....	132
<b>Figure 5.6</b>	Genome-wide plots of $F_{ST}$ divergence between the sympatric Cluster 1 and Cluster 2 subpopulations of <i>P. knowlesi</i> in Malaysian Borneo. ....	133
<b>Figure 5.7</b>	Frequency distribution of $F_{ST}$ values for SNPs at different levels of minor allele frequencies (MAF) .....	133
<b>Figure 5.8</b>	Distribution of genetic differentiation ( $F_{ST}$ ) in 14 chromosomes between Cluster 1 and Cluster 2 <i>P. knowlesi</i> subpopulations in Malaysian Borneo .....	135
<b>Figure 5.9</b>	Determination of genomic regions of divergence between two subpopulations of <i>P. knowlesi</i> in Malaysian Borneo .....	137
<b>Figure 5.10</b>	Genomic landscape between two divergent subpopulations of <i>P. knowlesi</i> in Malaysian Borneo .....	141
<b>Figure 5.11</b>	Spectrum of nucleotide diversity in genomic regions between Cluster 1 and Cluster 2 subpopulations .....	142
<b>Figure 5.12</b>	Distribution of 10-kb windows $\pi$ in high divergence regions (HDR) and low divergence regions (LDR) between Cluster 1 and Cluster 2 subpopulations .....	143
<b>Figure 5.13</b>	Maximum likelihood phylogenies inferred by 30.6-kb apicoplast genomes derived from 65 <i>P. knowlesi</i> infections .....	145
<b>Figure 5.14</b>	Maximum likelihood phylogenies inferred by 5.9-kb mitochondrial genomes of 129 <i>P. knowlesi</i> haplotypes .....	146
<b>Figure 5.15</b>	Genealogical network based on 129 <i>P. knowlesi</i> mitochondrial DNA genomes from Malaysian Borneo and mainland Southeast Asia showing 56 major haplotypes .....	147

## List of Tables

<b>Table 2.1</b>	Primers for genotyping of <i>P. knowlesi</i> microsatellites and location of loci in <i>P. knowlesi</i> version 1.0 genome sequence .....	50
<b>Table 2.2</b>	Species-specificity of primers for 19 microsatellite loci .....	52
<b>Table 2.3</b>	Numbers of isolates genotyped in 10 human and 2 macaque populations across Malaysia using 10 microsatellite loci .....	53
<b>Table 2.4</b>	Allelic diversity, measured as expected heterozygosity ( $H_E$ ), of 10 microsatellite loci of <i>P. knowlesi</i> in 12 populations across Malaysia .....	60
<b>Table 2.5</b>	Proportion of isolates designated as Cluster 1 and Cluster 2 in 10 geographical populations of human <i>P. knowlesi</i> isolates .....	62
<b>Table 2.6</b>	Test of multi-locus linkage disequilibrium of <i>P. knowlesi</i> isolated in humans by measuring the standardised index of association ( $I_A^S$ ) using only unique haplotypes in each geographical sites .....	62
<b>Table 2.7</b>	Pairwise measures of fixation indices ( $F_{ST}$ values above diagonal) and geographical distance (in kilometres below diagonal) across 10 populations of <i>P. knowlesi</i> from human isolates .....	64
<b>Table 3.1</b>	Unique primer pairs designed at each locus to discriminate two <i>P. knowlesi</i> subpopulation clusters .....	82
<b>Table 3.2</b>	Summary of PCR assay optimisations for discriminating two subpopulations of <i>P. knowlesi</i> infections .....	83
<b>Table 3.3</b>	Summary of PCR cycling condition of primer sets C1A and C2J for genotyping <i>P. knowlesi</i> parasites of Cluster 1 and Cluster 2 subpopulations, respectively .....	83
<b>Table 3.4</b>	Specificity of allele-specific PCR assays for primer sets C1A and C2J for <i>P.knowlesi</i> Cluster 1 and Cluster 2 subpopulations, respectively .....	85
<b>Table 4.1</b>	Summary of <i>P. knowlesi</i> mixed genotype infections in 134 human and 48 macaque hosts across Malaysia using 10 microsatellite loci .....	98

<b>Table 4.2</b>	Assignment of combined 753 <i>P. knowlesi</i> genotypes into three subpopulation clusters determined by minimum of two out of three assignment methods .....	106
<b>Table 4.3</b>	Summary of subpopulation cluster assignment on combined 758 <i>P. knowlesi</i> genotypes according to host and geographical origins .....	108
<b>Table 5.1</b>	Definition of the terminal genes before subtelomeric regions (version 2.0 of <i>P. knowlesi</i> H strain) .....	120
<b>Table 5.2</b>	Summary of mapping the short read sequences of 21 isolates generated using the MiSeq sequencing platform against the <i>P. knowlesi</i> H strain version 2.0 reference genome .....	125
<b>Table 5.3</b>	Subpopulation cluster assignments on 35 <i>P. knowlesi</i> infections in humans from Kapit, Malaysian Borneo .....	128
<b>Table 5.4</b>	Mean nucleotide diversity ( $\pi$ ) and pairwise differentiation ( $F_{ST}$ ) between Cluster 1 and Cluster 2 subpopulations .....	130
<b>Table 5.5</b>	Summary of mean value indices among low divergence regions (LDR), intermediate divergence regions (IDR) and high divergence regions (HDR) between two divergent of <i>P. knowlesi</i> subpopulations in Malaysian Borneo .....	139
<b>Table 5.6</b>	Locations and lengths of high divergence regions (HDRs) and low divergence regions (LDRs) in 14 chromosomes of <i>P. knowlesi</i> .....	140

## Abbreviations

%	percent
%poly	percentage of polyclonal infections
°C	degree Celsius
Δ	delta
∧	and
A+T	adenine and thymine
AIDS	acquired immune deficiency syndrome
BLAST	Basic Local Alignment Search Tool
bp	base pair
bq	base quality
Chr	chromosome
<i>D</i>	Tajima's <i>D</i> value
DAPC	discriminant analysis of principal component
DNA	deoxyribonucleic acid
dNTPs	deoxyribonucleotide triphosphates
<i>F<sub>ST</sub></i>	fixation indices
G+C	guanine and cytosine
<i>Hd</i>	haplotype diversity
HDR	highly divergence region
<i>H<sub>E</sub></i>	expected heterozygosity
HIV	human immunodeficiency virus
Hm	human
IAM	infinite alleles model
<i>I<sub>A</sub><sup>S</sup></i>	standardised index of association
ID	identity
IDR	intermediate divergence region
ITS	interstitial telomeric sequences
kb	kilobase pair
KIR	knowlesi interspersed repeats
km	kilometre
LD	linkage disequilibrium

LDR	low divergence region
LT	long-tailed macaque
MAF	minor allele frequency
Mb	megabase pair
MCMC	Markov chain Monte Carlo
MgCl <sub>2</sub>	magnesium chloride
MgSO <sub>4</sub>	magnesium sulphate
min	minute(s)
mM	millimolar
MOI	multiplicity of infection
mq	mapping quality
MS <sub>10</sub>	all ten microsatellite loci
MW	molecular weight marker
mya	million years ago
n	number
nd	not done
neg	negative
nM	nanomolar
P	probability
PCA	principal component analysis
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
Pct	<i>Plasmodium coatneyi</i>
Pcy	<i>Plasmodium cynomolgi</i>
Pfi	<i>Plasmodium fieldi</i>
Pfrg	<i>Plasmodium fragile</i>
Pin	<i>Plasmodium inui</i>
Pk	<i>Plasmodium knowlesi</i>
<i>Pkmsp-1</i>	<i>Plasmodium knowlesi</i> merozoite surface protein 1
<i>Pknbpxa</i>	<i>Plasmodium knowlesi</i> normocyte binding protein
PT	pig-tailed macaque
Q	Phred quality
QC	quality control



SD	standard deviation
SE	standard error
sec	second(s)
SICAvAr	schizont infected cell agglutination variants
SNP	single nucleotide polymorphisms
SSM	stepwise mutation model
SSRs	simple sequence repeats
STR	short tandem repeats
U	unit
UK	United Kingdom
UNIMAS	Universiti Malaysia Sarawak
USA	United States of America
WGS	whole genome sequencing
$\mu$ l	microliter
$\mu$ M	micro molar
$\pi$	nucleotide diversity ( $\rho i$ )

## Publications

### Papers:

Divis PC, Lin LC, Rovie-Ryan JJ, Kadir KA, Anderios F, Hisam S, Sharma RSK, Singh B, Conway DJ: **Three divergent subpopulations of the malaria parasite *Plasmodium knowlesi***. *Emerg Infect Dis* 2017, **23**.

Assefa S, Lim C, Preston MD, Duffy CW, Nair MB, Adroub SA, Kadir KA, Goldberg JM, Neafsey DE, Divis P, et al: **Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi***. *Proc Natl Acad Sci U S A* 2015.

Divis PC, Singh B, Anderios F, Hisam S, Matusop A, Kocken CH, Assefa SA, Duffy CW, Conway DJ: **Admixture in Humans of Two Divergent *Plasmodium knowlesi* Populations Associated with Different Macaque Host Species**. *PLoS Pathog* 2015, **11**:e1004888.

### Posters:

**Genomic divergence and loci under selection in *Plasmodium knowlesi*** – Divis PCS, Sharma RSK, Kadir KA, Anderios F, Hisam S, Matusop A, Assefa SA, Duffy CW, Pain A, Singh B and Conway DJ. Presented at the Molecular Approaches to Malaria (MAM) 2016 – awarded best poster presentation.

**Admixture in humans of two divergent *Plasmodium knowlesi* populations associated with different macaque host species** – Divis PCS, Singh B, Anderios F, Hisam S, Matusop A, Kocken CH, Assefa SA, Duffy CW and Conway DJ. Presented at the EMBL Conference BioMalPar XI: Biology and Pathology of the Malaria Parasites 2015.

**Population genetic structure of the zoonotic malaria parasite *Plasmodium knowlesi* in Malaysia** – Divis PCS, Singh B, Anderios F, Hisam S, Matusop A, Assefa SA, Duffy CW and Conway DJ – Presented at the 63<sup>rd</sup> Annual Meeting of the American Society of Tropical Medicine and Hygiene 2014.

**Orals:**

**Admixture in humans of two divergent *Plasmodium knowlesi* populations associated with different macaque host species** – Divis PCS, Singh B, Anderios F, Hisam S, Matusop A, Kocken CH, Assefa SA, Duffy CW and Conway DJ. Presented at the British Society of Parasitology 2015

**Population genetic structure of the zoonotic malaria parasite *Plasmodium knowlesi* in Malaysia** – Divis PCS, Singh B, Assefa SA, Duffy CW and Conway DJ. Presented at the Genomic Epidemiology of Malaria Conference 2014.

# Chapter One

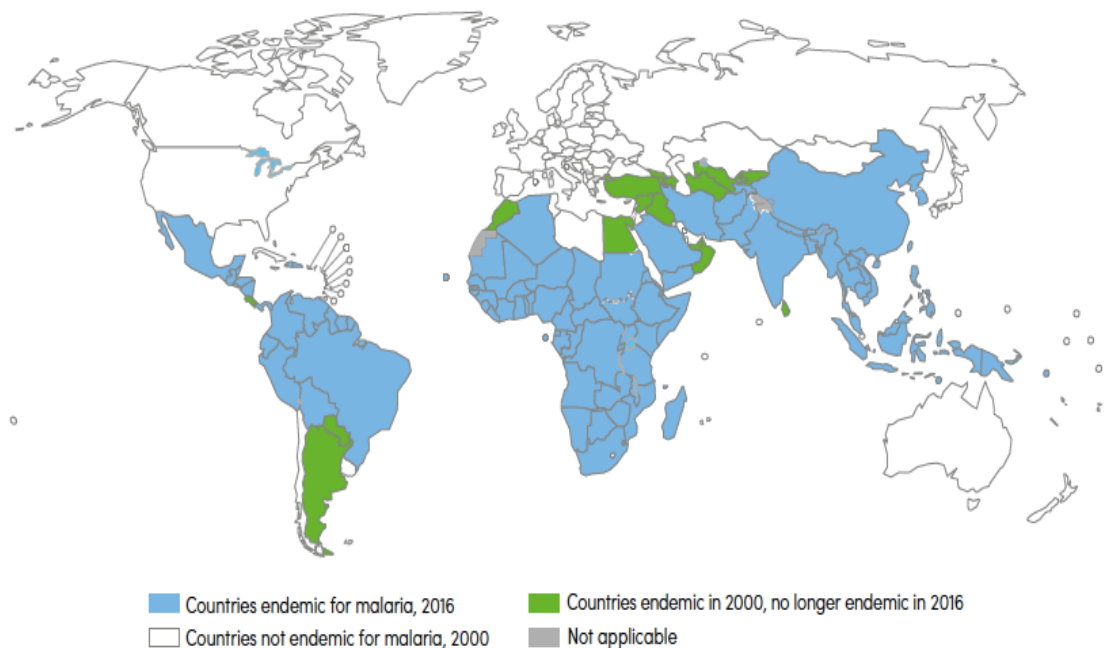
## Introduction

### 1.1 Malaria burden worldwide

Malaria is a noticeable health problem affecting 91 countries and territories worldwide (World Health Organization, 2016). In 2015, 212 million cases of malaria have been estimated, with infection rates higher among children aged 2 to 10 years and widely spread in the tropical and subtropical regions. While most malaria cases occur in the WHO African Region (90%), followed by WHO Southeast Asia Region (7%) and WHO East Mediterranean Region (2%), some countries that had endemic malaria in the year 2000 are no longer been considered as endemic after no cases were reported in three consecutive years (Figure 1.1). It was estimated that 438, 000 deaths were reported in 2015, making it the fifth infection to cause death worldwide after respiratory infections, HIV/AIDS, diarrhoeal diseases and tuberculosis.

The trademark presentation of malaria is fever, which starts with irregular patterns and further becoming periodic after a few days (Bartoloni and Zammarchi, 2012).

Other common clinical presentations in uncomplicated malaria include headache, chill, vomiting, fatigue, myalgia and nausea. In cases that are untreated or with delayed treatment, these can further progress to severe malaria with complications that may involve the central nervous system, respiratory system, renal systems and potentially haematopoietic systems. Nonetheless, asymptomatic carriage of malaria parasites in humans is widely reported, and infections may also associate with non-specific symptoms (Sifft et al., 2016, Fornace et al., 2015).



**Figure 1.1.** Map showing malaria burden worldwide in 2000 and 2016. Countries where no cases of malaria in three consecutive years are considered no longer endemic in 2016 (shown in green). Map adapted from Malaria World Report 2016 (World Health Organization, 2016).

Malaria is generally linked with poverty. Because of illness imposed by this disease, it is a public-health problem that contributes substantially to the social and economic development especially in endemic countries (Sachs and Malaney, 2002, Jimoh et al., 2007). Nonetheless, some endemic countries are now experiencing reduced malaria mortality rates, with some progress possibly relating to urbanisation and increased economic development (World Health Organization, 2016).

## 1.2 Malaria parasites and life cycles

### 1.2.1 *Plasmodium* species

The causative agent of malaria is the protozoan parasite of the genus *Plasmodium*.

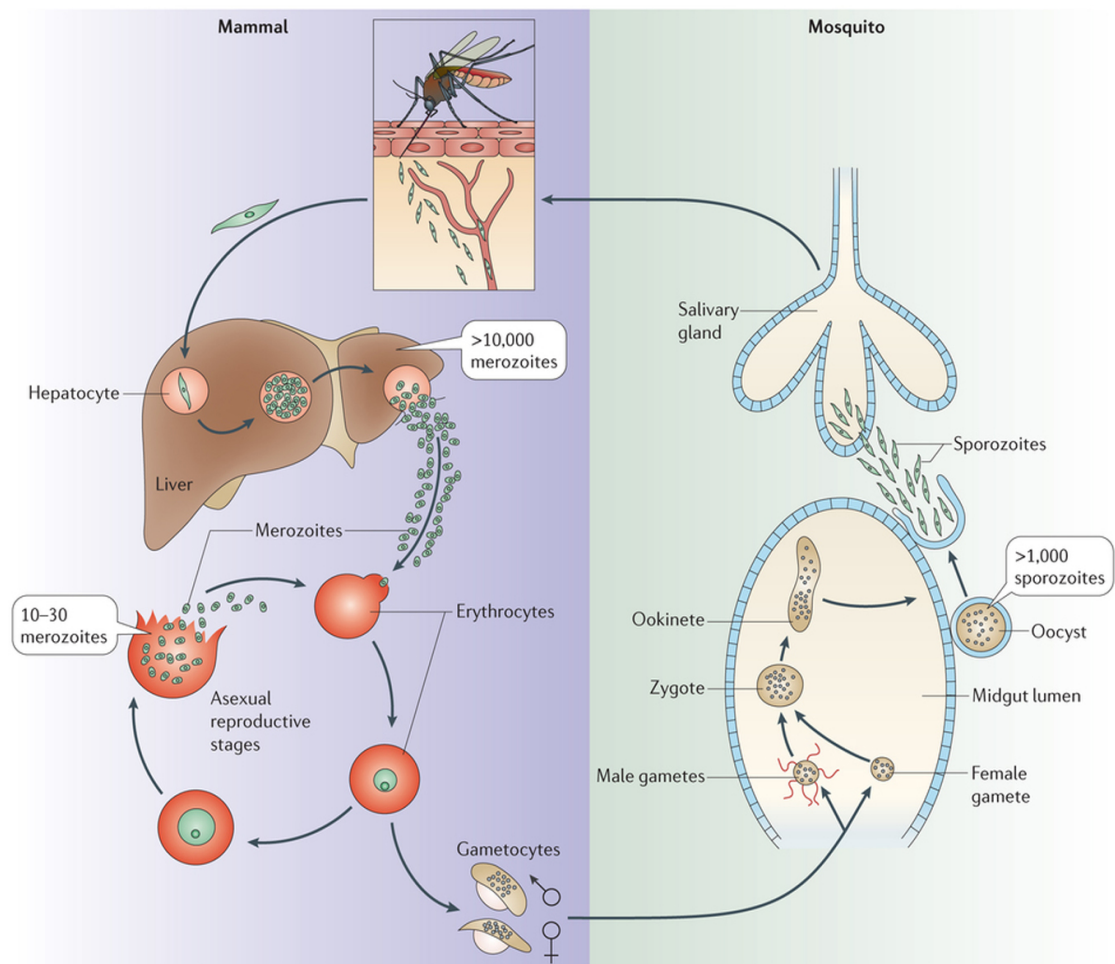
Globally, malaria is caused by five non-zoonotic human *Plasmodium* species (*P.*

*falciparum*, *P. vivax*, *P. malariae*, *P. ovale curtisi*, and *P. ovale wallikeri*), with *P. falciparum* responsible for the most deaths especially in Africa (Sutherland et al., 2010, World Health Organization, 2016). Outside the African continent, *P. vivax* is the major cause of malaria, especially in Latin America. In Southeast Asia as a whole, most malaria cases are due to *P. falciparum* or *P. vivax*, with highest incidence in remote forested areas by either or both species (World Health Organization, 2013).

Since 2004, *P. knowlesi* has become well recognised as a monkey malaria parasite that causes malaria in humans in Southeast Asia, and that can potentially cause severe and fatal infections (Singh et al., 2004, Cox-Singh et al., 2010, Rajahram et al., 2012). Other malaria parasites (*P. cynomolgi*, *P. coatneyi*, *P. fieldi* and *P. inui*) are also commonly found in monkeys of Southeast Asia, and two of these (*P. cynomolgi* and *P. inui*) have the capability to infect humans under experimental conditions (Garnham, 1966, Coatney et al., 1971). To date, only one case of natural *P. cynomolgi* infection in human has been reported in peninsular Malaysia (Ta et al., 2014). The epidemiology of *P. knowlesi* infections will be discussed further in this chapter, as it is the species focused on in this thesis.

### **1.2.2 Life cycles**

The life cycle of malaria parasites is complex, as it requires two hosts: vertebrates (including humans, non-human primates, and birds) and invertebrates (female *Anopheles* mosquitoes in the case of malaria parasites transmitted to mammals) (Figure 1.2). In general, during blood feeding of an infected mosquito on a human, haploid sporozoites are injected into the bloodstream and travel to the liver. Many sporozoites are destroyed by Kupffer cells and only fractions of the initial inocula



**Figure 1.2.** Life cycle of mammalian *Plasmodium* species.

A complete life of a malaria parasite involves sexual reproduction within the female anopheline mosquito host and asexual reproduction throughout the rest of the cycle. Image from published source (Menard et al., 2013).

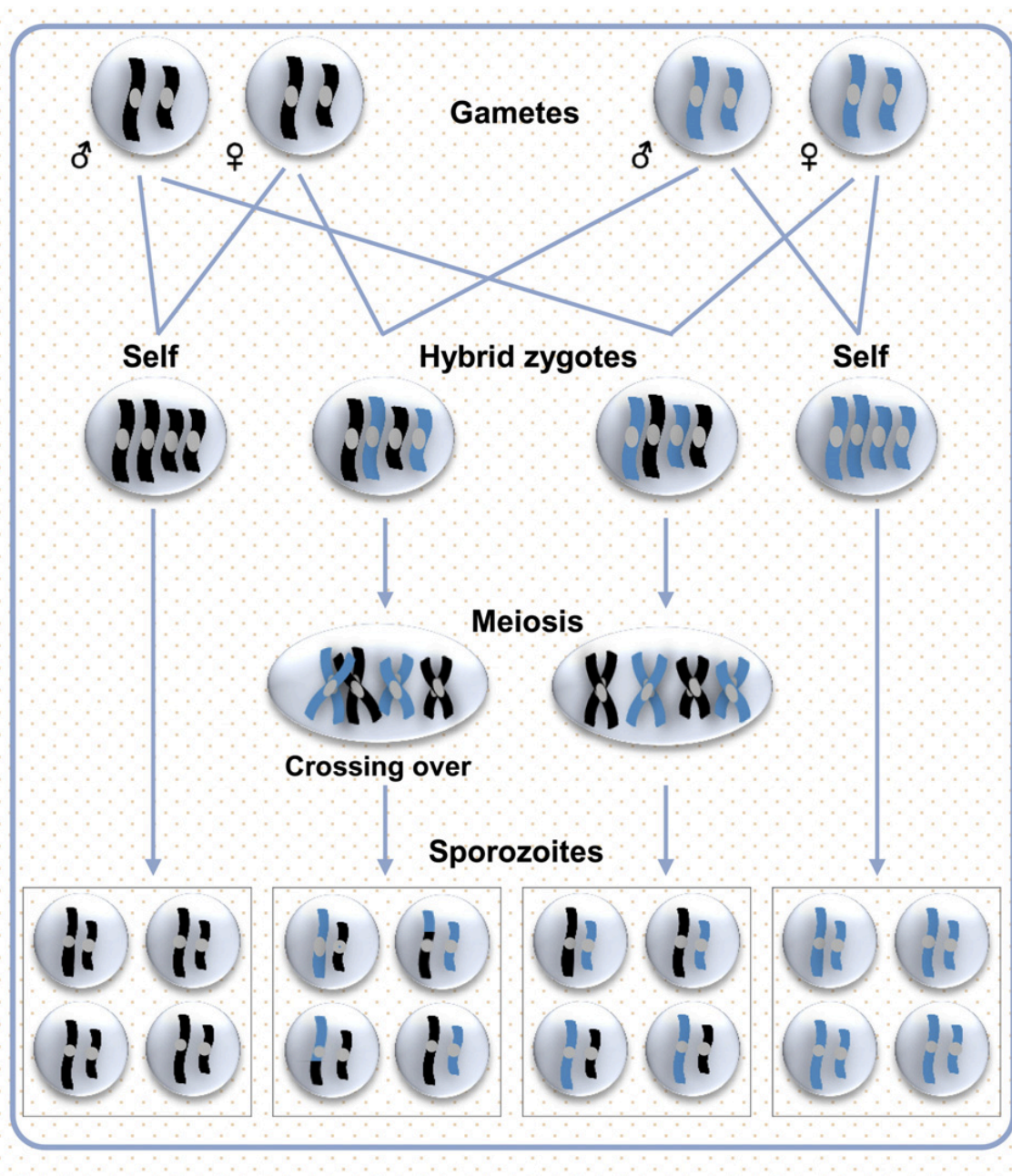
successfully invade individual hepatocytes (Pradel et al., 2004). Here, the parasite undergoes pre-erythrocytic stage development into schizonts containing thousands of merozoites that are the clonal products of mitosis. Once a schizont reaches maturity, it bursts and releases infective merozoites into the bloodstream to initiate the erythrocytic schizogony.

The asexual stages occur in the erythrocytes following merozoite invasion (Figure 1.2). They develop into morphological ring forms or early trophozoites, then into larger late trophozoites, and mature into schizonts containing 10 to 30 merozoites (Menard et al.,



2013). Matured schizonts rupture to release merozoites from erythrocytes, repeating the asexual cycles in humans. The multiple occurrences of asexual cycles in the human host results in increase of the parasite's population size in the host from  $10 - 10^2$  at the time of infection to  $10^8 - 10^{13}$  within a few weeks (Chang et al., 2013). Depending on *Plasmodium* species, the timing to complete an erythrocytic cycle varies, being either 24 hours (for *P. knowlesi*), 48 hours (for *P. falciparum*, *P. vivax* and *P. ovale*) or 72 hours (for *P. malariae*) (Coatney et al., 1971).

A small proportion of trophozoites enter into an alternative pathway to develop the sexual forms of haploid male or female gametocytes. When another female mosquito feeds blood from the infected human,  $10 - 10^3$  gametocytes are simultaneously taken up to the mosquito midgut (Chang et al., 2013). Here, gametocytes develop into male and female gametes, then fuse to form diploid zygotes. Meiosis occurs within a few hours following zygote formation, initiating genetic recombination between parental chromosomes (Figure 1.3) (Culleton and Abkallo, 2015). Then, the zygote matures into a motile ookinete, penetrates the midgut epithelium and finally develops into an oocyst. More than a thousand haploid sporozoites containing recombinant chromosomes are produced in each oocyst, which are then released into the mosquito haemolymph upon oocyst rupture, and finally reach the salivary glands of the mosquito. The life cycle continues when this mosquito probes the human skin before obtaining its blood meal (Figure 1.2).



**Figure 1.3.** Genetic recombination of malaria parasites in mosquito hosts. Fusion of haploid male and female gametocytes occurs in the midgut, producing diploid zygotes. Meiosis occurs within few hours, with recombination of genetic information between parental chromosomes, leading to development of haploid sporozoites containing recombinant chromosomes. Image from published source (Culleton and Abkallo, 2015).

### 1.3 Epidemiology of *Plasmodium knowlesi* infections

The first natural human infection of *P. knowlesi* was reported in 1965 when an American surveyor became ill upon returning to the United States from Pahang, peninsular Malaysia (Chin et al., 1965). This was followed by a case of *P. knowlesi*

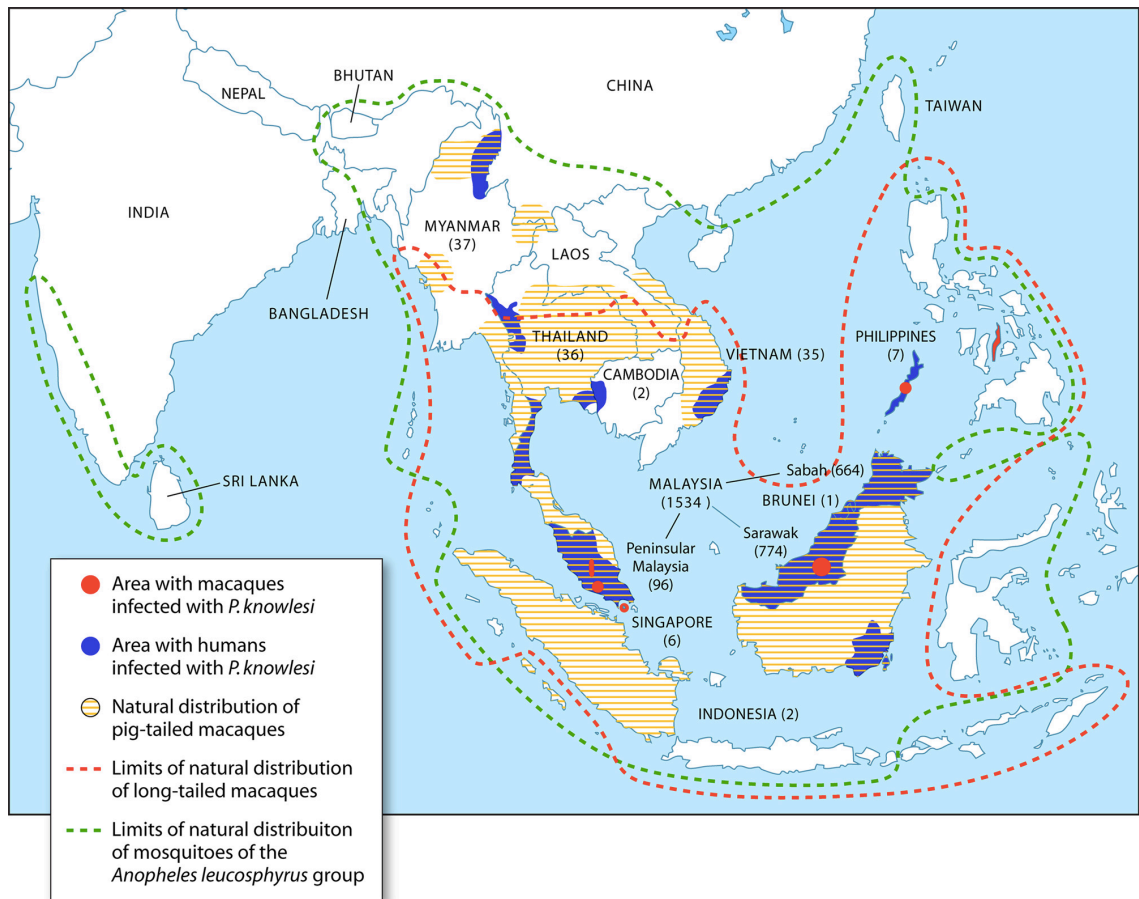
infection presumptively acquired in Johore, peninsular Malaysia (Fong et al., 1971). No cases were reported for more than 30 years, but then a study revealed that more than 50% of the malaria cases in Kapit Division, Malaysian Borneo were caused by *P. knowlesi* after characterisation by molecular methods (Singh et al., 2004). The finding of a high proportion of *P. knowlesi* infections in humans has led to an increase in the use of highly specific and sensitive polymerase chain reaction (PCR) methods elsewhere in the region (Singh and Daneshvar, 2013).

Subsequent studies showed that knowlesi malaria presents a spectrum of clinical manifestation from mild to fatal (Cox-Singh et al., 2008, Daneshvar et al., 2009, Cox-Singh et al., 2010, William et al., 2011, Rajahram et al., 2012, Rajahram et al., 2013). Following this, more knowlesi malaria cases have been described across Southeast Asia, including more regions in Malaysia (Cox-Singh et al., 2008, Vythilingam et al., 2008, William et al., 2011, Barber et al., 2012, Yusof et al., 2014), Thailand (Jongwutiwes et al., 2011, Jongwutiwes et al., 2004, Putaporntip et al., 2009, Sermwittayawong et al., 2012), Myanmar (Jiang et al., 2010, Ghinai et al., 2017), the Philippines (Luchavez et al., 2008), Singapore (Ng et al., 2008, Jeslyn et al., 2011), Vietnam (Van den Eede et al., 2009, Marchand et al., 2011), Indonesia (Figtree et al., 2010, Sulistyaningsih et al., 2010, Setiadi et al., 2016, Lubis et al., 2017) and Cambodia (Khim et al., 2011), as well as the Nicobar Islands of India (Tyagi et al., 2013). Knowlesi malaria is not only a major health concern in Southeast Asia, but imported cases into non-endemic countries in travelers returned from this region have alerted many healthcare providers worldwide (Kantele et al., 2008, Bronner et al., 2009, Ta et al., 2010, Berry et al., 2011, Link et al., 2012, Tanizaki et al., 2013, Cordina et al., 2014, Seilmaier et al., 2014).

There are perspectives that *P. knowlesi* infections is an emerging zoonotic pathogen (Cox-Singh, 2012, Antinori et al., 2013), although the “emerging” may be largely due to the advancement of molecular methods used for specific detection. Because the morphological resemblance to *P. malariae* during the erythrocytic stage (Lee et al., 2009b), the development of simple *P. knowlesi*-specific PCR assays was vital (Singh et al., 2004). Following this, study on “*P. malariae*” archival slides of patients obtained from Sarawak in 1990s reveals that most of them were indeed infected with *P. knowlesi* (Lee et al., 2009a), indicating that this parasite is not a very newly emerging pathogen.

It is known that long-tailed (*Macaca fascicularis*) and pig-tailed (*M. nemestrina*) macaque monkeys harbour *P. knowlesi* in nature (Coatney et al., 1971). The restricted distribution of these macaques overlaps with knowlesi malaria cases and the vector of the *Anopheles leucosphyrus* group in Southeast Asia (Figure 1.4). This suggests that local *P. knowlesi* transmission is strongly restricted to Southeast Asia, as it would only occur when suitable reservoir hosts and mosquito vector co-exist.

*P. knowlesi* infections in humans are likely to result from an ongoing zoonosis, as initially supported by sequence analyses of three polymorphic loci in the parasite (18S rRNA, circumsporozoite protein gene and mitochondrial genome) (Lee et al., 2011). Although high genetic heterogeneity was observed in *P. knowlesi* isolated from humans and wild macaques in Kapit Division of Sarawak, in Malaysian Borneo, identical haplotypes of these loci were evident in both humans and macaques. Entomological study identifies forest-dwelling mosquito *Anopheles latens* (of the *An.*



**Figure 1.4.** Geographical range of *P. knowlesi* infections in humans, mosquito vectors and macaques across Southeast Asia regions. Numbers in the parentheses represent numbers of *P. knowlesi* infections reported from 2004 to 2013. Image from published source (Singh and Daneshvar, 2013).

*leucosphyrus* group) as a local vector, which is attracted to both humans and macaques (Tan et al., 2008). In addition to this finding, other *Anopheles* species have been incriminated as vectors for *P. knowlesi*, including *An. cracens* and *An. introlatus* in peninsular Malaysia (Vythilingam et al., 2008, Jiram et al., 2012, Vythilingam et al., 2014), *An. dirus* in Vietnam (Marchand et al., 2011) and *An. balabacensis* in Sabah, Malaysian Borneo (Wong et al., 2015).

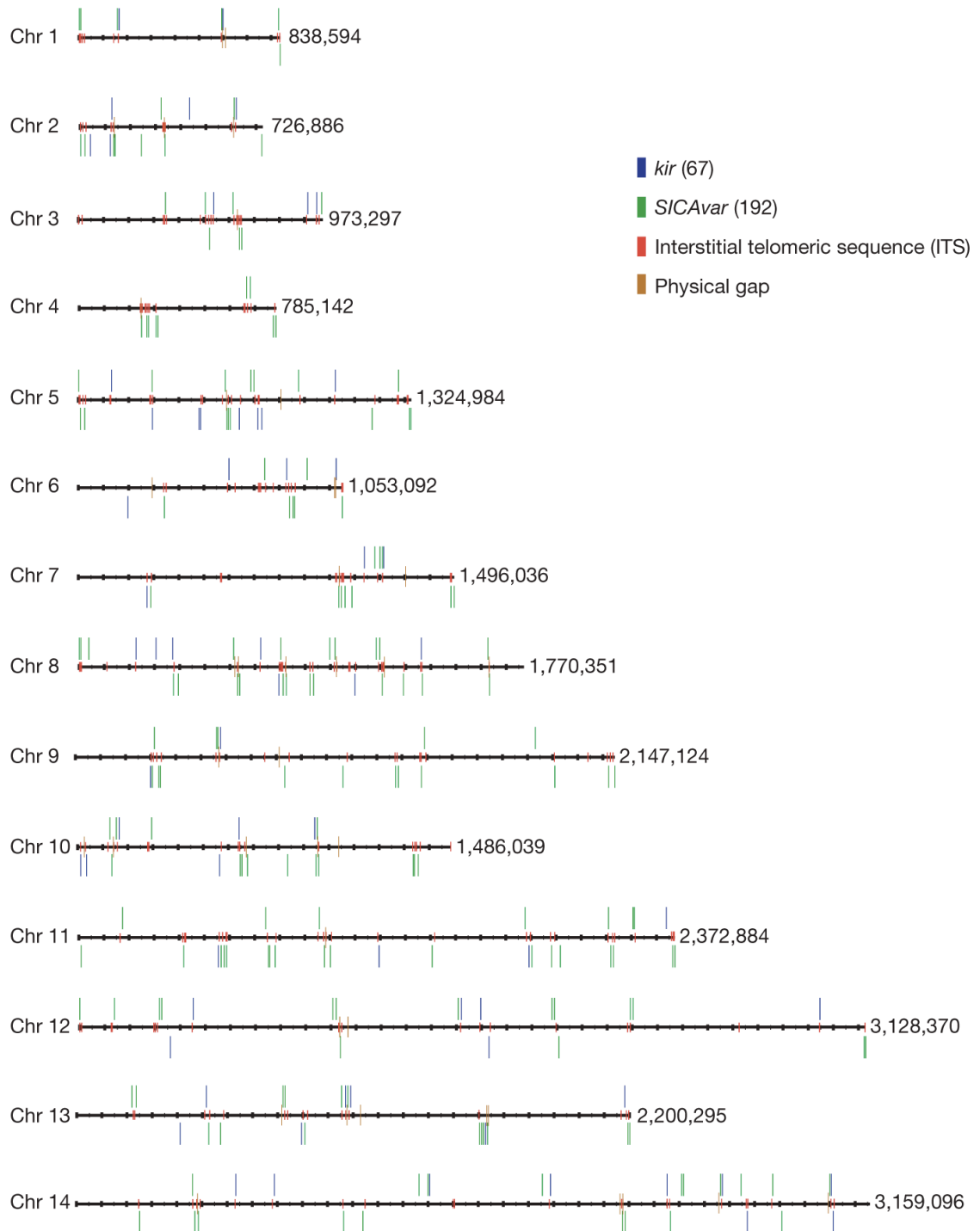
The research of this thesis is focused particularly on *P. knowlesi* infections in Malaysia, the country with most knowlesi malaria cases, which is also in malaria pre-elimination phase as there has been a progressive decrease of malaria cases over the past few

decades (World Health Organization, 2013). While cases of other human *Plasmodium* species decrease, the incidence of knowlesi malaria has apparently increased over time since it was first reported in 2004 (World Health Organization, 2015, William et al., 2013). As knowlesi malaria is zoonotic, it is not considered as part of an elimination campaign, but the incidence needs to be closely monitored (World Health Organization, 2015).

#### **1.4 The genome of *Plasmodium knowlesi***

The genome sequence of *P. knowlesi* was first published in 2008 (Pain et al., 2008). The parasite, referred to as strain H, was originally isolated from the patient infected in Malaysia in 1965. The nuclear genome sequence is approximately 24.1 Mb in length with chromosome size ranging from 0.84 Mb (chromosome 1) to 3.16 Mb (chromosome 14). Most *Plasmodium* species are A+T biased, and the G+C content for *P. knowlesi* is 38%, slightly less than in the closely related species *P. vivax* and *P. cynomolgi*. However, compared to *P. falciparum* (Gardner et al., 2002), the percentage of G+C in *P. knowlesi* is relatively high (Carlton et al., 2008, Pain et al., 2008).

At least 5,197 genes have been described in the *P. knowlesi* genome, with exons covering 49% of the total genome sequence (Tachibana et al., 2012). The multi-gene families, schizont infected cell agglutination variants (*SICAvar*) and knowlesi interspersed repeats (*KIR*), form the largest group and are randomly distributed throughout the 14 chromosomes (Pain et al., 2008). They are mostly surrounded by the interstitial telomeric sequences (ITS;GGGTT(T/C)A) and arranged in a tandem manner or as a large component of repeat units (Figure 1.5). A direct comparison of



**Figure 1.5.** Distribution of multi-gene families (*SICAvar* and *KIR*) and telomere-like repeats in 14 chromosomes of the *P. knowlesi* strain H version 1.0 genome. Multi-gene families are scattered across chromosome 1 to 14, with interstitial telomeric sequences (ITS) are found surrounding these gene. Image from published source (Pain et al., 2008).

the genome together with *P. vivax* and *P. cynomolgi* showed that the gene synteny is highly conserved across 14 chromosomes. Nonetheless, a number of genes are not syntenic and these mostly are of the multi-gene *SICAvar* and *KIR* families.

Comparative study among three major *Plasmodium* genome sequences reveals that the number and types of distribution of simple sequence repeats (SSRs) of both *P. knowlesi* and *P. vivax* are found to be lower than that of *P. falciparum* (Tyagi et al., 2011). This composition of A+T content in *P. knowlesi* is slightly higher than *P. vivax*, but lower than *P. falciparum*.

In this thesis, analyses comparing to the *P. knowlesi* reference H strain genome sequence were conducted based on reference version 1.0 (Pain et al., 2008), except for analyses presented in Chapter Five. During the progression of this thesis, the version 2.0 of the genome has been released in GeneDB in March 2014 (<http://www.genedb.org>) and this was used in the genome-wide analysis of *P. knowlesi* in Chapter Five.

## **1.5 Population genetic structure of *Plasmodium* species in Southeast Asia**

### **1.5.1 *P. falciparum* and *P. vivax***

Studies done on two human *Plasmodium* species (*P. falciparum* and *P. vivax*) of Southeast Asia have shown various features of population genetic structures. Genome-wide analyses of single nucleotide polymorphisms (SNP) in *P. falciparum* infections from Cambodia, Thailand and Vietnam indicate clear population genetic sub-structure of the parasite compared to those from West Africa (Miotto et al., 2013). Furthermore, the existence of multiple subpopulations in Cambodia alone reveals that the



population structure is highly complex, and this might lead to the existence of multiple forms of drug resistant *P. falciparum* (Miotto et al., 2013).

In different geographical foci of *P. falciparum* isolates in Malaysian Borneo, significant population structure with profound genetic differentiation indicated that foci were largely independent and relatedness declined with geographical distance (Anthony et al., 2005). Recently, reduced genetic diversity and microsatellite haplotypes over time in *falciparum* malaria have been reported in areas in Sabah where malaria is declining (Mohd Abd Razak et al., 2016).

Similar to *P. falciparum*, studies of *P. vivax* genetic structures in Southeast Asia reveal extensive heterogeneity. Significant genetic differentiation with a different level of linkage disequilibrium was observed from different subpopulations in Southeast Asia. Distinct genotypes of *P. vivax* have been circulating in many endemic areas, and multiple clone infections are common even in malaria-hypoendemic regions (Zhong et al., 2011, Imwong et al., 2007). Additionally, analyses of complete mitochondrial genomes isolated from *P. vivax* infections in Myanmar, Korea and China suggest a demographic history with ancient population expansion around 50,000 years ago (Miao et al., 2012). In a focus of *P. vivax* in Sabah, Malaysian Borneo, analysis using microsatellite markers indicates that malaria epidemic expansion may occur locally (Abdullah et al., 2013).

### **1.5.2 *P. knowlesi***

There has been very limited previous study regarding the population genetics of *P. knowlesi*. Analysis of the SNPs and haplotype tree structure of the mitochondrial DNA

of *P. knowlesi* isolated from humans and macaques of Sarawak, Malaysian Borneo, indicate that the parasite derived from an ancestral parasite approximately 98,000 – 478,000 years ago, predating human settlement in Southeast Asia (Lee et al., 2011). Moreover, the most recent common ancestor of *P. knowlesi* is likely to be older than that of *P. falciparum* (50,000 – 330,000 years ago) or *P. vivax* (53,000 – 265,000 years ago), and Asian macaques were likely the natural hosts throughout this time. Following this, *P. knowlesi* has undergone significant population expansion approximately 30,000-40,000 years ago, at a time when Borneo was part of mainland Southeast Asia and when the human population was growing in the region (Lee et al., 2011).

A study of the complete nucleotide sequence of merozoite surface protein 1 (*Pkmsp-1*) genes in a small number of *P. knowlesi* isolates from humans and macaques of Thailand suggested higher haplotype diversity in *P. knowlesi* isolated from humans than those of pig-tailed macaques (Putaporntip et al., 2013). The sample sizes were very small in this study, and as *Pkmsp-1* is a gene under strong selection its use as a single marker is inappropriate. More sampling of different isolates with multiple genetic markers is needed to investigate this systematically. In recent studies pertaining to *P. knowlesi* population substructure in Malaysia, analyses using different genes such as the Duffy binding protein (Fong et al., 2014, Fong et al., 2015), normocyte binding protein (Ahmed et al., 2016) and mitochondrial cytochrome oxidase subunit 1 (Yusof et al., 2016) showed some divergence between parasites from Borneo and mainland Southeast Asia.

### **1.6 Genetic markers for studying the population substructure**

Choosing appropriate genetic markers is crucial in order to study the biodiversity and

genetic structure of any species (Chenuil, 2006). There are many types of genetic markers, but appropriate ones may depend on the biological issues to be studied. It is emphasised that good genetic markers should not be under selection, and the use of multiple independent loci throughout the genome is recommended. In this research, two types of genetic markers are focused on.

### **1.6.1 Microsatellites**

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are loci containing tandemly repeated DNA motifs that are abundantly found in most eukaryotes. By definition, the composition of microsatellites is between 1 – 6 bp per repeat copy (mono- to hexa-nucleotide), whereas motifs with more than six nucleotides are referred to minisatellites (Guichoux et al., 2011). In general, microsatellites can be classified into three major families – pure (uninterrupted), compound and interrupted repeats (Jarne and Lagoda, 1996).

Microsatellites have been used for genotyping many organisms. In the human genome, a microsatellites locus is found every 2 – 30 kb on average (Guichoux et al., 2011). In order to identify recent population expansions, this genetic marker has some advantages because of the accumulation of new mutations in a short period of time, with mutation rate ranging from  $10^{-3}$  to  $10^{-6}$  per locus per generation (Guichoux et al., 2011, Oliveira et al., 2006).

The occurrence of mutation in microsatellites is indicated by changes of the repeat units, whether they involve one-step changes, multi-step changes, or directionally in favour of increased or decreased repeats (Ellegren, 2004). Generally, there are three

major models of mutation (Jarne and Lagoda, 1996, Oliveira et al., 2006). Firstly, the most favoured mechanism to estimate relations between individuals and population structure (Oliveira et al., 2006, Guichoux et al., 2011) is the stepwise mutation model (SSM), in which when microsatellites mutate, they only gain or lose one repeat. This may imply two alleles that differ by one repeat are most likely to be closely related and have a more recent common ancestor. Secondly, the infinite alleles model (IAM) considers that each mutation creates new allele randomly at rate  $\mu$ . For this model, for instance, a 15-repeat allele could be as closely related to a 10-repeat allele as to a 14-repeat allele. Finally, the  $K$ -allele model considers that a microsatellite can mutate into any one of  $K$  alleles randomly. The SSM model is likely to match the data best; as much of the mutation in microsatellites is due to slipped strand and imperfect DNA repair mechanisms during replication.

Multilocus microsatellite genotypes have been widely used to study the population genetics of *P. falciparum* and *P. vivax* (Abdullah et al., 2013, Anderson et al., 2000, Mobegi et al., 2012, Van den Eede et al., 2010), but prior to this thesis work no such studies were conducted for *P. knowlesi*.

### **1.6.2 Single nucleotide polymorphisms from whole genome sequencing data**

In general, SNPs are found to be more abundant than microsatellites in the genome of most species. In the human genome, SNPs are found in every 100 – 300 bp (Webster et al., 2002, Guichoux et al., 2011). Compared to microsatellites, SNPs have slower mutational rates, roughly  $10^{-9}$  per locus per generation (Oliveira et al., 2006).

In general, SNP analyses have been performed in genes of interest, or to broadly study the genetic diversity or mutation rates in populations. Examples of individual genes commonly investigated in malaria parasites include the circumsporozoite protein gene (Vargas-Serrato et al., 2003, Lee et al., 2011), merozoite surface protein gene (Atroosh et al., 2011, Putaporntip et al., 2013), and thrombospondin-related adhesive protein gene (Ohashi et al., 2014, Kosuwin et al., 2014).

The advancement of technology has given birth to high-throughput sequencing that has replaced or supplemented Sanger sequencing for genome-wide analysis. Whole genome sequencing is able to detect million of SNPs from the entire genome, enabling the study of many genes simultaneously. This technology offers lower cost per nucleotide than the standard Sanger capillary sequencing (Sboner et al., 2011).

The use of SNPs derived from whole genome sequencing has been applied to many organisms. Information from SNPs datasets have enabled study of the genomic divergence or genomic architecture in closely related organisms from different ecological environments or geographical locations. Considering studies on wild fish for example, this includes the cichlid *Astatotilapia* (Malinsky et al., 2015), three-spined stickleback *Gasterosteus aculeatus* (Guo et al., 2015) and Atlantic cod *Gadus morhua* L. (Berg et al., 2015) that show remarkable genomic heterogeneities among populations associated with different ecological environments.

In relation to malaria, a number of genome-wide studies have been conducted on *P. falciparum*. For example, analysis of SNPs from whole genome datasets of West African populations reveals evidence of loci under recent positive directional or

balancing selection (Mobegi et al., 2014, Duffy et al., 2015). Similarly for *P. vivax*, SNP analysis of population genomic data from different continents identified high genomic diversity relative to *P. falciparum* and revealed signatures of selection on loci likely to be related to drug resistance (Hupalo et al., 2016).

## **1.7 Hypothesis and objectives**

### **1.7.1 Hypothesis**

At the outset of this study, three hypotheses were considered:

1. If the *P. knowlesi* population is primarily dependent on macaques, with little transmission between humans, there should be strong differentiation between peninsular Malaysia and Malaysian Borneo, as macaques in these areas have been separated since the last glacial period.
2. If *P. knowlesi* is an ongoing zoonosis, there should be no major barrier of gene flow in *P. knowlesi* from macaques to humans.
3. There should be evidence of recent selection on particular loci in the *P. knowlesi* genome, due to encroachment of humans on the wild macaque habitats, leading to novel adaptive opportunities for the parasite.

### **1.7.2 Specific objectives**

Since there has been no analysis to determine the population genetic structure of *P. knowlesi* from different locations at a broad scale, this study would be the first to explore this. This project aims to increase our understanding of the population genetic patterns and genetic diversity of *P. knowlesi* throughout its distribution across Malaysia. The specific objectives are:

1. To develop multilocus microsatellite assays as genetic markers for large-scale genotyping analysis of *P. knowlesi* (Chapter Two).
2. To identify population substructure of *P. knowlesi* infections from humans and macaques from different geographical locations (Chapter Two and Chapter Four).
3. To develop simple allele-specific genotyping assays to distinguish different *P. knowlesi* subpopulations (Chapter Three).
4. To generate new genome-wide single nucleotide polymorphism (SNP) data from *P. knowlesi* isolates in humans using high throughput sequencing (Chapter Five).
5. To study the genomic divergence between *P. knowlesi* subpopulations using whole genome sequencing data (Chapter Five).

# Chapter Two



**Registry**  
 T: +44(0)20 7299 4646  
 F: +44(0)20 7299 4656  
 E: registry@lshtm.ac.uk

**RESEARCH PAPER COVER SHEET**

**PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.**

**SECTION A – Student Details**

<b>Student</b>	Paul Cliff Simon Divis
<b>Principal Supervisor</b>	Professor David Conway
<b>Thesis title</b>	Population Genetic Structure and Genomic Divergence in <i>Plasmodium knowlesi</i>

***If the Research Paper has previously been published please complete Section B, if not please move to Section C***

**SECTION B – Paper already published**

Where was the work published?	PLoS Pathogens		
When was the work published?	28 <sup>th</sup> May 2015		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

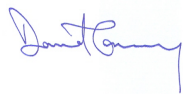
*\*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

**SECTION D – Multi-authored work**

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper (Attach a further sheet if necessary)	I designed the study, performed laboratory experiments, conducted statistical analyses and wrote the manuscript.
---	--

**Student Signature:** 

**Date:** 13<sup>th</sup> March 2017

**Supervisor Signature:** 

**Date:** 20<sup>th</sup> March 2017

## **Divergent *Plasmodium knowlesi* in human populations associated with different macaque host species**

### **2.1 Introduction**

The epidemiological emergence of infections can be traced by genotypic analyses, with a high level of resolution when pathogens have a high mutation rate, as illustrated by recently emerged viruses that now have a massive impact on global public health (Faria et al., 2014, Gire et al., 2014). Such analysis is more challenging for eukaryote pathogens with low mutation rate, although it is now clear that the major human malaria parasites *Plasmodium falciparum* and *P. vivax* have been endemic for many thousands of years after having been acquired as zoonotic infections from African apes (Liu et al., 2010, Liu et al., 2014). In contrast, natural human infections by *P. knowlesi* were almost unknown (Coatney et al., 1971) until a large focus of cases in Malaysian Borneo was described a decade ago (Singh et al., 2004). Infections have since been reported from throughout southeast Asia, within the geographical range of the long-tailed and pig-tailed macaque reservoir hosts (*Macaca fascicularis* and *M. nemestrina*) and mosquito vectors (of the *Anopheles leucosphyrus* group) (Singh and Daneshvar, 2013).

It is vital to determine the causes of this apparent emergence, as *P. knowlesi* can cause severe clinical malaria with a potentially fatal outcome (Cox-Singh et al., 2008, Rajahram et al., 2012, Barber et al., 2013). Molecular tools to discriminate *P. knowlesi* from other malaria parasite species were not widely applied until the zoonosis became

known, but analysis of DNA in archived blood samples from Malaysia and Thailand shows that it was already widespread twenty years ago (Lee et al., 2009a, Jongwutiwes et al., 2011). Sequences of parasite mitochondrial genomes and a few nuclear gene loci indicate ongoing zoonotic infection, as human *P. knowlesi* genotypes share most alleles identified in parasites sampled from wild macaques (Lee et al., 2011, Fong et al., 2014, Putaporntip et al., 2013).

To understand this zoonosis, and to identify whether human-to-human mosquito transmission is occurring, analyses of parasite population genetic structure in humans and macaques should be performed by extensive population sampling and characterisation of multiple putatively neutral loci. This study presents a *P. knowlesi* microsatellite genotyping toolkit and its application to the analysis of a large sample of isolates from human cases at ten different sites, as well as from both species of wild macaque reservoir hosts. Results reveal a profound host-associated sympatric subdivision within this parasite species, as well as geographical differentiation indicating genetic isolation by distance. The existence of two divergent parasite subpopulations, and their admixture in human infections provides unparalleled opportunity for parasite hybridisation and adaptation. Observations of some clinical infections with parasite types that appear intermediate between the two subpopulations may reflect this process, and are a possible result of human-to-human mosquito transmission.

## **2.2 Materials and methods**

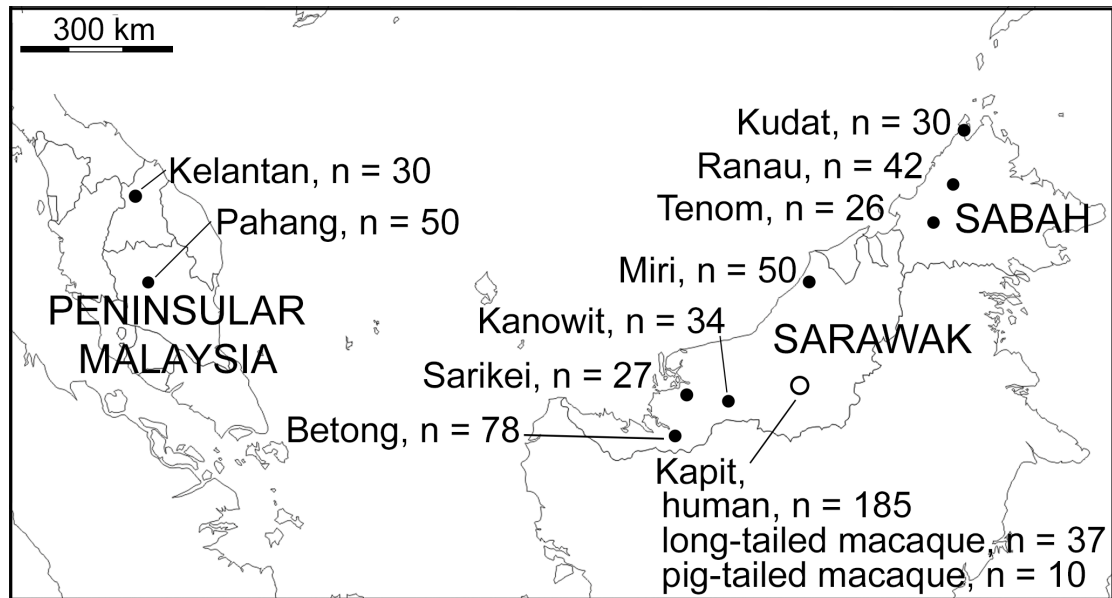
### **2.2.1 *P. knowlesi* samples from humans and macaques**

A total of 599 DNA samples from different *P. knowlesi* infections of humans and macaques were analysed from collections performed at 10 different geographical sites (Figure 2.1). 552 samples were from human *P. knowlesi* malaria patients from all of the sites, eight in Malaysian Borneo (Sarawak and Sabah states) and two in Peninsular Malaysia (Kelantan and Pahang states).

For samples from Sarawak, DNA was extracted at the University Malaysia Sarawak (UNIMAS) in Kuching from previously reported blood samples collected between 2000 and 2011 (Singh et al., 2004, Cox-Singh et al., 2008, Daneshvar et al., 2009, Foster et al., 2014) as well as new samples collected in 2012 and 2013, allowing analysis of five sites: Kapit (n = 185), Betong (n = 78), Kanowit (n = 34), Sarikei (n = 27) and Miri (n = 50). Samples from Sabah were collected in 2013, and DNA was extracted by the Sabah Public Health Reference Laboratory, allowing analysis of three sites: Kudat (n = 30), Ranau (n = 42) and Tenom (n = 26). For Peninsular Malaysia, blood samples collected from Kelantan (n = 30) and Pahang (n = 50) underwent DNA extraction at the Institute for Medical Research in Kuala Lumpur. DNA from blood samples of a total of 47 wild macaques (long-tailed macaque, *Macaca fascicularis* n = 37; pig-tailed macaque, *M. nemestrina* n = 10) previously collected within 30 km radius of Kapit town in Sarawak (Lee et al., 2011) were also included in the analyses (Appendix 2.1). The presence of *P. knowlesi* DNA was confirmed in all samples at UNIMAS by nested PCR assays (Singh et al., 2004, Lee et al., 2011).

### **2.2.2 Development of *P. knowlesi* microsatellite genotyping markers**

A combination of three microsatellite mining tools, i) iMEX (Mudunuri and Nagarajaram, 2007), ii) mreps (Kolpakov et al., 2003), and iii) MSATCOMMANDER



**Figure 2.1.** Map of sampling locations of 599 *P. knowlesi* infections genotyped in this study.

A total of 552 samples were from *P. knowlesi* malaria patients from 10 geographical locations: Peninsular Malaysia (Kelantan and Pahang), Sarawak (Kapit, Betong, Miri, Kanowit and Sarikei), and Sabah (Kudat, Ranau and Tenom). Additionally, 47 samples were from wild macaques (37 long-tailed and 10 pig-tailed macaques) with *P. knowlesi* infections in Kapit.

(Faircloth, 2008), were used to identify simple sequence repeat loci from the *P. knowlesi* version 1.0 reference genome (Pain et al., 2008). Loci with perfect trinucleotide simple repeat sequences were carefully selected using customised perl-script commands based on narrow criteria to maximise their likely utility for genotyping: i) a minimum of 7 repeat copies in each microsatellite in the reference sequence, ii) located at non-telomeric chromosomal regions as defined by regions syntenic with the *P. vivax* reference genome (Carlton et al., 2008), and iii) absence of any homopolymeric tracts adjacent to the microsatellite sequence that could give rise to additional size polymorphism.

As a result, 19 trinucleotide repeat loci widely spaced in the genome were shortlisted and PCR primers were designed using PrimerSelect software (DNASTAR, USA) for hemi-nested PCR assays. The specificity of PCR was tested using DNA controls of all human *Plasmodium* species, common malaria parasites of the Southeast Asian macaques (*P. knowlesi*, *P. coatneyi*, *P. inui*, *P. cynomolgi* and *P. fieldi*), as well as human, long-tailed and pig-tailed macaque DNA. Loci for which primers showed complete specificity of amplification from *P. knowlesi* were tested further for genotyping performance.

### **2.2.3 PCR and genotyping protocols**

Genotyping of each microsatellite locus was performed using a hemi-nested protocol with a fluorescent dye-labelled inner primer during the second round PCR amplification. Both first and second round PCR amplifications were conducted in individual tubes or wells for each locus, in 11 µl reaction volume containing 0.2 mM each dNTP (Bioline, UK), 2 mM MgSO<sub>4</sub>, 1X ThermoPol II reaction buffer (NEB, UK), 0.275 U *Taq* DNA polymerase (NEB, UK), 0.1 µM of each forward and reverse primer, and 1 µl sample DNA template. The PCR cycling conditions were as follows: initial denaturation at 94°C for 2 min, followed by 28 cycles of 94°C for 30 sec, annealing at 56°C for 30 sec and elongation at 68°C for 30 sec, with a final elongation step at 68°C for 1 min. Final PCR products were pooled into three groups of loci with different product size and dye profiles together with Genescan 500 LIZ molecular size standards (Applied Biosystems, UK) and run on a Genetic Analyzer 3730 capillary electrophoretic system (Applied Biosystems, UK). GENEMAPPER version 4.0 software (Applied Biosystems, UK) was used for scoring of allele electrophoretic size, and quantification of peak heights.

#### **2.2.4 Multiplicity of infection**

Infections containing multiple haploid parasite genotypes were apparent as multiple electrophoretic peaks for a locus corresponding to different alleles. The apparent genotypic multiplicity of infection (MOI) was determined by the locus with the most alleles detected in the infection, considering peaks with height of at least 25% relative to the predominant allele within each isolate. The predominant allele per locus within each infection was counted for subsequent population genetic analyses.

#### **2.2.5 Genetic diversity and population divergence**

Allelic diversity at each locus was measured as the virtual heterozygosity ( $H_E$ ) using FSTAT software version 2.9.3.2 (<http://www2.unil.ch/popgen/softwares/fstat.htm>), and allele frequency distributions were also inspected using GenAlEx version 6 (Peakall and Smouse, 2006) within the Microsoft Excel platform. Genetic differentiation between each population was measured by pairwise fixation indices ( $F_{ST}$ ) using FSTAT, with Bonferroni correction on a nominal significance level of 0.05 applied for multiple comparisons across the population pairs. To test for correlation between genetic differentiation and geographical distance, a Mantel test for isolation by distance was performed with Rousset's linearised  $F_{ST}/(1-F_{ST})$  plotted against the natural log of geographic distance using Genepop version 4.2 (Rousset, 2008, Raymond and Rousset, 1995).

#### **2.2.6 Haplotype relatedness and linkage disequilibrium**

The relatedness of haplotypes between individual isolates was assessed by measuring the pairwise proportion of shared alleles, excluding samples with missing data at any locus. A matrix of pairwise similarity among isolates was calculated based on the

identical or mismatched alleles from a complete set of loci and the distribution of shared alleles between sample pairs for each population was visualised using a customised perl-script command. To test for non-random allele assortment, multi-locus linkage disequilibrium (LD) was assessed by the standardised index of association ( $I_A^S$ ) using LIAN version 3.6 (Haubold and Hudson, 2000), with significance of the  $I_A^S$  values tested by Monte-Carlo simulation with 10,000 data permutations to generate the null distribution under linkage equilibrium.

### **2.2.7 Assessment of population genetic substructure**

To explore evidence of population substructure in the entire population, a Bayesian analysis was performed using the STRUCTURE version 2.3.4 software (Pritchard et al., 2000) using samples with no missing data at any locus. Individuals in the population pool were clustered to the most likely population ( $K$ ) by measuring the probability of ancestry using the multi-locus genotype data. The program parameters were set to admixture model with correlated allele frequency, with 50,000 burn-in period and 100,000 Markov chain (MCMC) iterations. To run the simulation,  $K$  value was predefined from 1 – 10 and the run was performed in 20 replicates for each  $K$ . The most probable  $K$  value was then calculated according to Evanno's method (Evanno et al., 2005) using the webpage interface STRUCTURE Harvester (Earl and vonHoldt, 2012). The assignment of a sample to a subpopulation cluster was based on the inferred cluster scores by STRUCTURE analysis, where samples with inferred cluster scores within a range in relation to the  $K$ -value were assigned together as one subpopulation cluster. The intermediate cluster assignment indices were calculated based on the proportion of shared cluster ancestries per individual isolate inferred by the cluster scores from the STRUCTURE analysis.



Principal component analysis (PCA) using the GenAlEx package was also performed independently for the same purpose. Samples with missing data at any locus were excluded, and the genetic distance matrix was generated based on the allelic mismatches between pairs of isolates. A two-dimensional PCA plot was generated considering the first two highest eigenvalues, and genetic clusters were determined based on the eigenvector coordinates along the axes of variation.

## **2.3 Results**

### **2.3.1 *P. knowlesi* microsatellites as genetic markers for population studies**

Hemi-nested PCR assays were developed for amplification of 19 tri-nucleotide simple sequence repeat loci from throughout the genome of *P. knowlesi* and tested for species-specificity using control DNA from all 10 known parasite species of humans, long-tailed or pig-tailed macaques, as well as human and macaque DNA to identify those suitable for genotyping samples from all hosts (Table 2.1). Assays for 11 loci were entirely species-specific for *P. knowlesi*, and 10 of these gave a clear single electrophoretic peak for each allele without any stutter bands (Table 2.2). These were used to genotype *P. knowlesi* infections in a total of 599 humans and wild macaques with a high rate of success, 556 (92.8%) scoring clearly for all 10 loci (Table 2.3; Appendix 2.2). Numbers of alleles at each locus ranged from 7 (for locus NC03\_2) to 21 (for locus CD05\_06) (Appendix 2.2).

**Table 2.1.** Primers for genotyping of *P. knowlesi* microsatellites and location of loci in *P. knowlesi* version 1.0 genome sequence.

For each locus, the first two primers are used in the Nest 1 reaction, while the second and third primers are used in the Nest 2 reaction. In each primer ID, the F and R labels denote forward and reverse primers, respectively. The third primer for each locus is internal to the other two and labelled with a specific fluorescent dye for genotyping on the capillary electrophoresis. Only loci 1 – 10 were *P. knowlesi*-species specific and used for subsequent genotyping experiments.

No	Locus	Chromosome (location of locus)*	Primer ID	Fluorescent dye label	Sequence (5' → 3')
1	NC03_2	3 (762,413 – 762,566)	N03_2R5	6FAM	AGACTCATGTGCGGCGTTCCTT
			N03_2mF1		GCGGGGAGGACGATAAACCCATA
			N03_2mR1		CGTCAAATGAAGAGAGCATTGCTC
2	CD05_06	5 (110,541 – 110,790)	C5R2	VIC	GCTACAATGTTTGAATCAGAAGG
			C5F2		GCCCATTGCAGCTATGCAC
			C5R1		GTTTTCGCTCCATGTTCCAGCC
3	CD08_61	8 (943,277 – 943,507)	C8R7	VIC	CTTGAACGTGCGTTTACATTCC
			C8F3		GATCAGTGGACTGGTATACACAGATA
			C8R5		CCATGTAGGATGTATATTTCTTCG
4	NC09_1	9 (217,751 – 218,039)	N09_1mR2	PET	TTCTCCACTTGACTTAAGGATTAAGC
			N09_1F2		TTGAAGCGGAATAGGGAAGGAT
			N09_mR1		CACACAGGTACGTGCATACATATAAG
5	NC10_1	10 (760,484 – 760,758)	N10_1mR2	NED	ATGTAGTAATGTTGGGGCTGTTGGTG
			N10_1F1		CATGGCTGGTATCCCCCTGTTCC
			N10_1R2		GGAAGGAGAGGACTAGTGCTGAAAGAG
6	CD11_157	11 (2,266,650 – 2,266,897)	C11bR3	NED	AATACATTTTCGGGAATATTCTCG
			C11bF2		TAGGGATCGTCAGAGGGAGG
			C11R2		GTAGCAAATATGCTCGTTCAGG
7	NC12_2	12 (930,911 – 931,257)	N12_2F4	VIC	TTCGTTGTTCTGTTTTCTTTGTTACT
			N12_2mR1		ATGGATCACTACCTTGTGGTG
			N12_2mF1		GTGGAAAGGAGGCAAACACACAG
8	NC12_4	12 (1,577,676 – 1,577,910)	N12_4mF2	VIC	GTCCTGATGAAATTGAATTTGTGC
			N12_4mR2		AGGGGTCGTATCGTCTCGTACG
			N12_4mF1		GGAGTTCCTCCGTTCCGAATG
9	CD13_61	13 (774,903 – 775,077)	C13aF2	6FAM	CCACTGATTGACAAGAAGAAGTTG
			C13aR1		GTTCCAATTGTTGGCCCCATTG
			C13aF1		CCAACAACCATTTCAGTTGACAAAG
10	CD13_107	13 (1,220,220 – 1,220,405)	C13bR6	NED	CGTGTGAAGAACGTAGTAGTACTG
			C13bF1		GATGACCACGTTATGGATAATGTTG
			C13bR1		CGTAAAACTCTGCACCTCCTTTGC
11	NC02_2	2 (180,599 – 180,839)	N02_2F5	6FAM	TGGCCAACAAGGGTAGCATCA
			N02_2R1		AATGCAGTTCCTTTCTTTACACGAG
			N02_2F1		AAGAGGCGAAGGAAGATAATCACATAGG
12	NC02_4	2 (76,039 – 76,381)	N02_4mR3	VIC	GGTATAACATTTTCATAACTCAGAAGAC
			N02_4F1		ACACTGTTACGACTTTTTCTTTCCATT
			N02_4mR2		TGGTGTGGAGGATCCATAGACA
13	CD03_40	3 (147,219 – 147,454)	C3F2	VIC	ATAGAGCGTGAAAAGAACAAGAG
			C3R1		ACCTCTCAATAGGGTAGGGTTAG
			C3F1		CGATTCTGAAAAAAGAAGAAGAG
14	NC04_1	4 (60,354 – 60,595)	N04_1R7	VIC	GCGCGGCACCACCTTTTAT
			N04_1mF2		GATTTCGTTATTCTCGTGCACACAAG
			N04_1R4		CCCCAATTTGATGTATAAGGTAGCAGAG
15	NC05_2	5 (885,398 – 885,603)	N05_2R1	VIC	ATTATTACAACCTCATGGATGGAAGATTT
			N05_2mF2		TCCCCTCACCTTGGGAGAGTG
			N05_2mR1		GCAGTCGTTACGCGGTTACAT

No	Locus	Chromosome (location of locus)*	Primer ID	Fluorescent dye label	Sequence (5' → 3')
16	NC08_3	8 (69,052 – 69,207)	N08_3mR1 N08_3mF1 N08_3R1	VIC	GATGAAGAATTTGTAGAGGCCG CTTCGGTAGAAGGAAAAAAGTGTG ACCATGATGTTTATTGTAGGGCTGAA
17	NC10_5	10 (1,126,914 – 1,127,165)	N10_5mF3 N10_5mR1 N10_5mF2		TGACCAGCTAGCCAATCTGTCA TGCATGCACACGGTACCAATTA CGGTCTGCATTGTCCCG
18	CD11_86	11 (1,339,944 – 1,340,119)	C11aF2 C11aR1 C11aF1		TGGGTGATGAGGTCGTGAATAGG GGTTGACGAGCAGGGTAAAACTGAG TATGCCTGCGGGAAGGGTGAG
19	NC12_7	12 (2,483,130 – 2,483,440)	N12_7R1 N12_7F2 N12_7mR1		TTCCCCTCGTGCGACTCTTCT AGGTAAGAGCCACGCAAGAATAACAAC GGCTTGGCAGCTTTACTTAAGTTCC

\*location of locus is based on primers used in second nest PCR

**Table 2.2.** Species-specificity of primers for 19 microsatellite loci.

The specificity of primers were initially visualised under agarose gel, and those with single band were further confirmed using the GeneMapper software. All primers for each locus did not cross-react to *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, human and macaque DNA.

Locus ID	Chr	Repeat motifs in reference genome	Visualisation of species-specificity	
			Agarose gel	GeneMapper
NC02_2	2	(TTA)9	Pk	Pk, with 1 bp stutter peaks
NC02_4	2	(ATA)7	Pk, Pin, Pfi	Not done
NC03_2	3	(AAG)11	Pk	Pk
CD03_40	3	(AAG)11	Pk	Pk, Pct, Pfi
NC04_1	4	(GAA)10	Pk, Pct	Not done
NC05_2	5	(ATT)9	Pk, Pct	Not done
CD05_06	5	(TAA)7	Pk	Pk
NC08_3	8	(TCA)7	Pk	Pk, Pcy, Pfi, Pct
CD08_61	8	(TAC)11	Pk	Pk
NC09_1	9	(GAA)9	Pk	Pk
NC10_1	10	(TTA)9	Pk	Pk
NC10_5	10	(AAT)7	Pk, Pfrg	Not done
CD11_86	11	(CAA)7	Pk	Pk, Pcy, Pct, Pin
CD11_157	11	(GAG)8	Pk	Pk
NC12_2	12	(AAT)16	Pk	Pk
NC12_4	12	(GAA)7	Pk	Pk
NC12_7	12	(TAT)7	Pk, Pct	Not done
CD13_61	13	(AAC)8	Pk	Pk
CD13_107	13	(AGG)7	Pk	Pk

Abbreviation: Chr – chromosome, Pk – *P. knowlesi*, Pct – *P. coatneyi*, Pfi – *P. fieldi*, Pcy – *P. cynomolgi*, Pin – *P. inui*, Pfrg – *P. fragile*

**Table 2.3.** Numbers of isolates genotyped in 10 human and 2 macaque populations across Malaysia using 10 microsatellite loci.

Locus	Sarawak										Sabah				Peninsular Malaysia				Total
	LT	PT	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	
	Kapit (n=37)	Kapit (n=10)	Kapit (n=185)	Betong (n=78)	Kanowit (n=34)	Sarikei (n=27)	Miri (n=50)	Kudat (n=30)	Ranau (n=42)	Tenom (n=26)	Kelantan (n=30)	Pahang							
NC12_2	36	10	171	72	34	24	50	28	39	26	28	50	39	26	28	50	568	94.8	
NC03_2	36	10	184	77	34	25	50	30	39	26	28	50	39	26	28	50	589	98.3	
NC09_1	36	10	182	76	34	26	50	29	40	26	27	50	40	26	27	50	586	97.8	
NC12_4	36	10	183	76	34	26	50	30	41	26	29	50	41	26	29	50	591	98.7	
NC10_1	37	10	184	77	34	26	50	30	41	26	27	50	41	26	27	50	592	98.8	
CD08_61	36	10	184	76	34	27	50	30	42	26	29	50	42	26	29	50	594	99.2	
CD11_157	34	10	183	78	34	26	50	30	40	26	28	50	40	26	28	50	589	98.3	
CD13_61	36	10	181	77	33	26	50	30	40	26	28	50	40	26	28	50	587	98.0	
CD05_06	36	10	183	77	34	26	50	30	40	26	29	50	40	26	29	50	591	98.7	
CD13_107	36	10	183	77	34	26	50	30	39	26	27	50	39	26	27	50	588	98.2	
All loci	34	10	167	71	33	24	50	28	38	26	25	50	38	26	25	50	556	92.8	

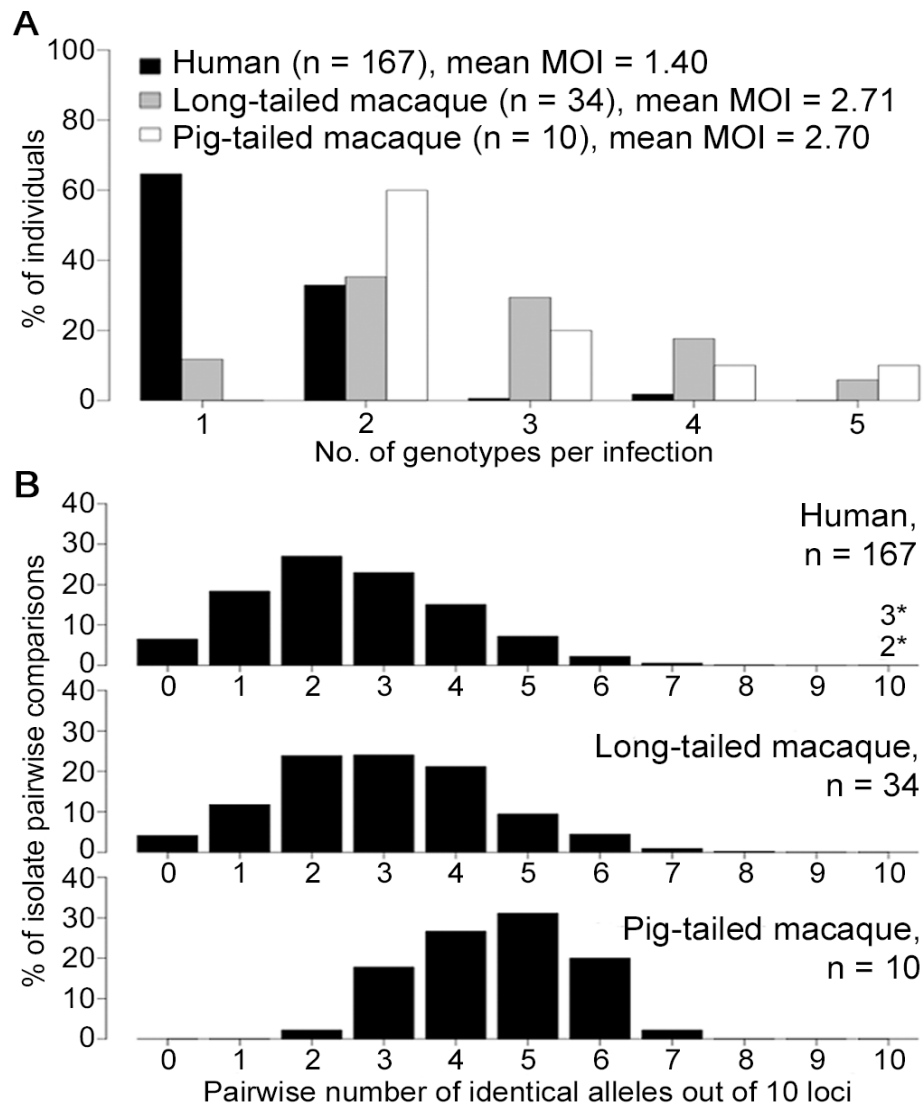
Abbreviation: LT – long-tailed macaque, PT – pig-tailed macaque, Hm – human

### 2.3.2 Host-dependent genetic structure of *P. knowlesi*

#### 2.3.2.1 Inter-host and inter-host diversity

Parasites from different host species sampled from Kapit where high numbers of clinical cases are seen were first analysed using complete 10-locus genotype data. Almost all *P. knowlesi* infections in macaques contained multiple genotypes, with no significant difference between long-tailed macaques (88% of 34 were mixed, mean MOI = 2.71) and pig-tailed macaques (100% of 10 were mixed, mean MOI = 2.70;  $P = 0.65$  for comparison between macaque host species), whereas only a minority of human *P. knowlesi* infections had multiple genotypes (35% of 167, mean MOI = 1.40;  $P < 10^{-15}$  for comparison between humans and macaques) (Figure 2.2A). To allow equally weighted sampling per host, the predominant allele at each locus within each infection was counted for subsequent analysis (Appendix 2.2).

Pairwise comparisons of each of the complete 10-locus profiles revealed that all infections in Kapit were genotypically distinct, except for one identical pair and one identical triplet of human infections (Figure 2.2B; Appendix 2.3). There was a much higher average proportion of shared alleles among pig-tailed macaque infections than among those in long-tailed macaques or humans (medians of 5, 3 and 2 identical alleles out of 10 loci respectively). Analysis of allele frequencies revealed that *P. knowlesi* parasites from pig-tailed macaques are very highly divergent from those in long-tailed macaques ( $F_{ST} = 0.217$ ,  $P < 0.001$ ), whereas those in humans have an intermediate level of relatedness ( $F_{ST} = 0.067$  versus long-tailed macaques,  $F_{ST} = 0.104$  versus pig-tailed macaques;  $P < 0.001$  for both).



**Figure 2.2.** Multiple genotype *P. knowlesi* infections and diversity among infections in three host species from Kapit.

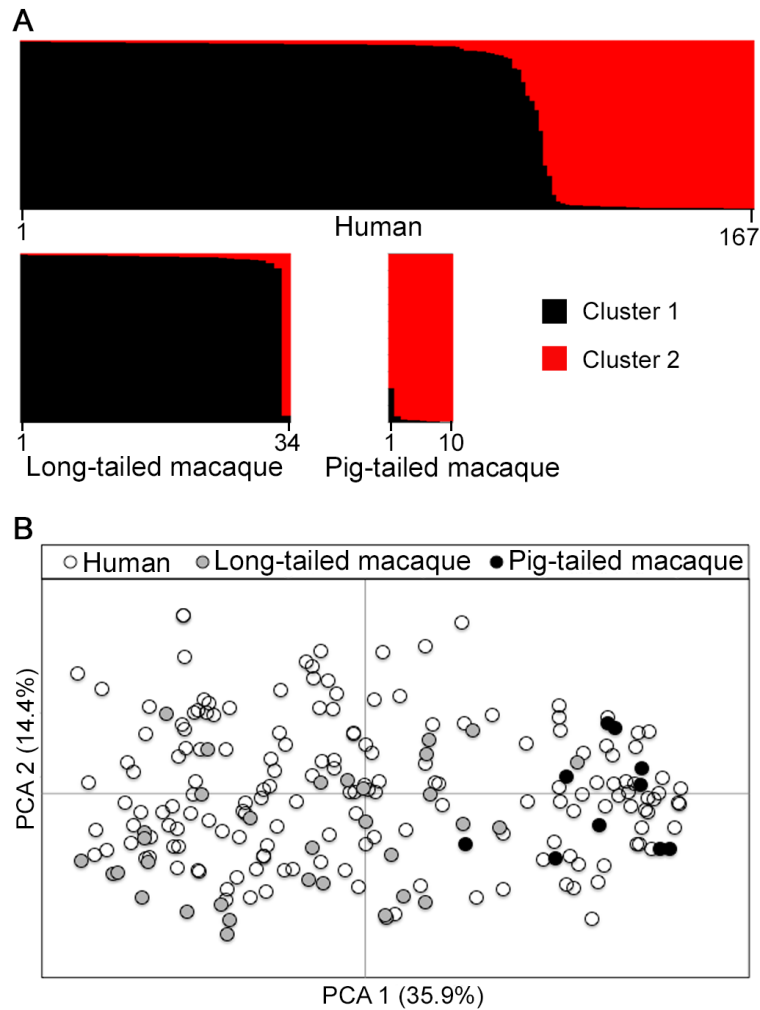
All data for the multiple genotype infections and distribution of identical alleles between infections derived from complete genotyping of 10 microsatellite loci. (A) Numbers of different *P. knowlesi* genotypes per infection (multiplicity of infection, MOI) showing significant difference between human and macaque infections (Fisher's Exact  $P < 1 \times 10^{-15}$ ), but not between long-tailed and pig-tailed macaques (Fisher's Exact  $P = 0.65$ ). (B) Numbers of identical alleles out of 10 loci in pairwise comparisons of infections, showing a similar diversity among infections in humans and long-tailed macaques, but a higher average identity among infections from pig-tailed macaques. All infections had a different 10-locus genotype, except for 5 of the 167 human infections (there was one pair sharing an identical 10-locus genotype, and a triplet of infections sharing another 10-locus genotype, indicated with asterisks).

### 2.3.2.2 Population genetic substructure

A Bayesian model-based STRUCTURE analysis of multi-locus genotype data from all hosts sampled in Kapit indicated the existence of two sub-population clusters of *P. knowlesi* ( $K = 2$ ;  $\Delta K = 936.75$  based on Evanno's estimation of  $K$ -population) (Figure 2.3A, Appendix 2.2 and Appendix 2.4). An individual infection genotype was assigned to be predominantly of a particular cluster if the STRUCTURE analysis score exceeded 0.5 for that cluster. All except one of the long-tailed macaque infections were assigned to the Cluster 1 subpopulation, whereas all pig-tailed macaque infections were assigned to the Cluster 2 subpopulation, while 71% of human infections were assigned to Cluster 1 and 29% to Cluster 2 (Figure 2.3A). A small minority of those which were primarily assigned to either cluster appeared to have a degree of mixed assignment, with scores nearer 0.5 than either zero or 1.0 for the alternative clusters (Appendix 2.2), which is analysed in a separate section below.

An independent scan by principal component analysis (PCA) showed an almost complete separation between parasites from long-tailed macaques and pig-tailed macaques along the first principal component, while parasites from humans covered the whole distribution and overlapped with all of the samples from both of the macaque hosts (Figure 2.3B).





**Figure 2.3.** Population genetic structure of *P. knowlesi* from infections in three host species in Kapit.

Both STRUCTURE and principal component analyses were conducted based on the complete 10-locus genotype dataset (167 from humans, 34 from long-tailed macaques, 10 from pig-tailed macaques). (A) Bayesian model-based STRUCTURE analysis indicates two subpopulation clusters throughout the whole dataset ( $K = 2$ ,  $\Delta K = 936.75$ ), with almost complete partitioning between the two macaque host species. Cluster 1 is shown in black while Cluster 2 is in red. (B) Principal component analysis (PCA) of the genetic divergence among all infections. The percentage of variation captured by each of the first two principal components is shown in brackets. Infections from the different macaque host species are almost completely separated by the first principal component, while human infections are distributed throughout the full range on both axes.

### **2.3.3 Geographical population genetic structure of *P. knowlesi***

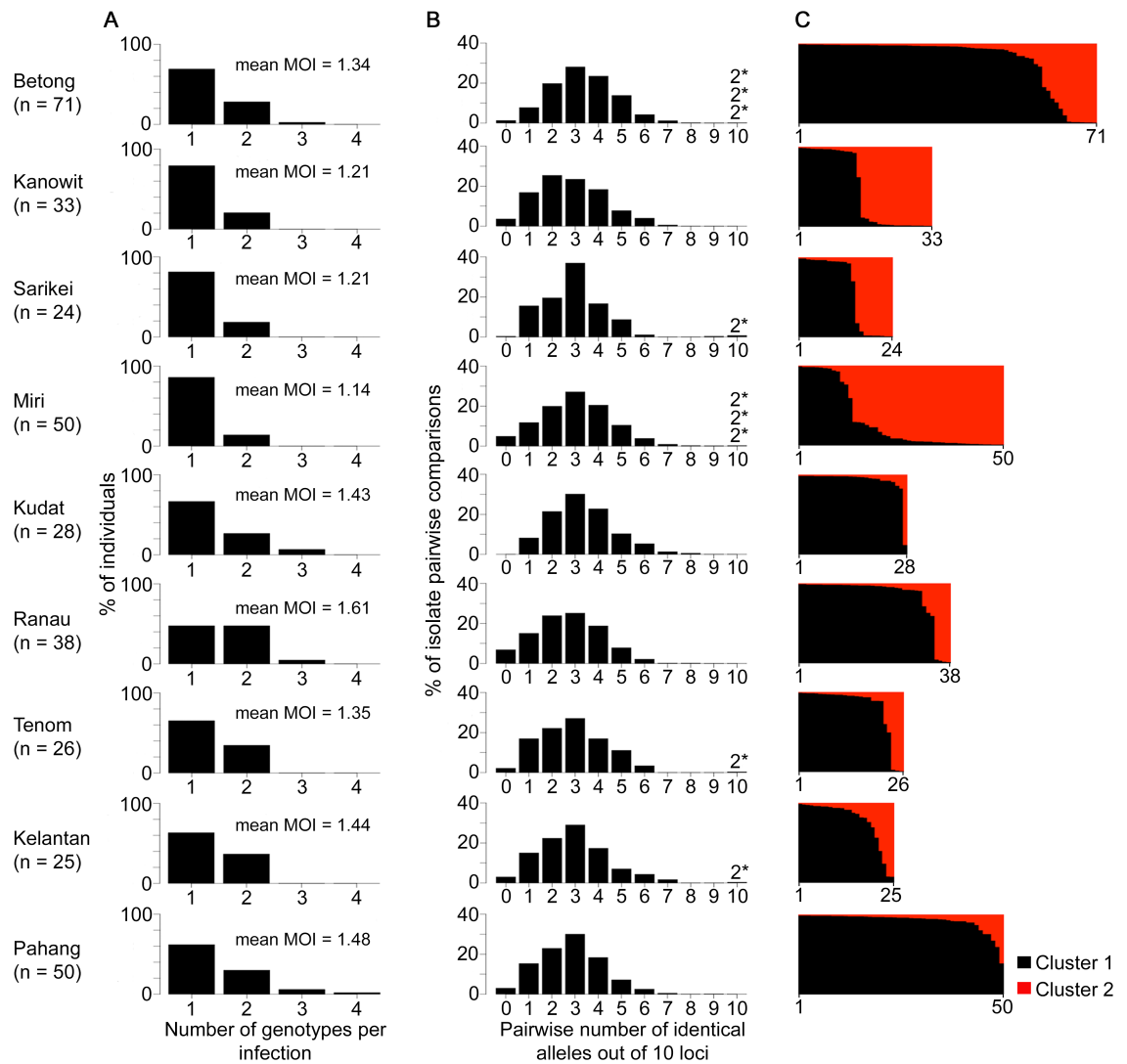
#### **2.3.3.1 Diversity among geographical sites**

Next, a further 367 human *P. knowlesi* infections from nine other geographical sites (Figure 2.1) were analysed. Most human infections had single *P. knowlesi* genotypes (Figure 2.4A; Appendix 2.2), and there were no differences in the proportions of mixed genotype infections across all sites (Comparison across 10 sites including Kapit: Pearson's  $\chi^2$ ,  $P = 0.096$ ; 32% of infections having  $> 1$  genotype overall). There were no differences in allelic diversity among the different sites ( $H_E$  estimates between 0.67 and 0.75,  $P > 0.1$  for all pairwise Wilcoxon Signed Rank tests across all 10 loci, Table 2.4).

Pairwise comparisons among genotypes from different infections showed a similar level of diversity at each site, with a median of 2 or 3 identical alleles out of 10 loci in each site (Figure 2.4B). Every infection had a different multi-locus genotype, and there were virtually none that shared alleles at more than 7 loci, except for nine pairs of identical haplotypes (three pairs in Betong, three in Miri, and one in each of Sarikei, Tenom and Kelantan) (Figure 2.4B). Each identical haplotype pair was shared by infections from different individuals sampled at the same site within the same year, except for two of the identical haplotype pairs in Miri, shared by individual infections sampled one and two years apart (Appendix 2.3).

#### **2.3.3.2 Population genetic substructure and population divergence**

There were two subpopulation clusters ( $K = 2$ ,  $\Delta K = 174.94$ , Appendix 2.4) throughout all of these sites, as had been seen in Kapit, but the relative frequency of the clusters varied geographically ( $P < 0.0001$ , Figure 2.4C). The Cluster 1 subpopulation was more



**Figure 2.4.** Diversity and genetic structure of *P. knowlesi* in human infections from nine different geographical locations.

All *P. knowlesi* data for analysing multiple genotype infections, distribution of identical alleles between isolates and STRUCTURE analysis were derived from complete genotyping of 10 microsatellite loci. (A) Proportions of infections containing different numbers of genotypes (multiplicity of infection, MOI). Comparison across all 10 geographical locations including Kapit (data in Figure 2.3) showed no significant differences (Pearson's  $\chi^2$  with 10000 replicates,  $P = 0.096$ ). (B) Distribution of numbers of identical alleles out of 10 loci in pairwise comparisons of infections. In six of the populations, a small number of pairs of identical multi-locus genotypes were seen as indicated with a label here (2\*) and tabulated in detail in Appendix 2.2. (C) Subpopulation clusters inferred by the Bayesian model-based STRUCTURE analysis with Cluster 1 (black) and Cluster 2 (red) corresponding to those identified in Figure 3 ( $K = 2$ ,  $\Delta K = 174.94$ ). Proportions of isolates assigned as Cluster 2 are highest at the sites in Sarawak (top four panels in the figure; locations of all sites are shown in Figure 2.1).

**Table 2.4.** Allelic diversity, measured as expected heterozygosity ( $H_E$ ), of 10 microsatellite loci of *P. knowlesi* in 12 populations across Malaysia.

Locus	Sarawak										Sabah				Peninsular Malaysia		
	LT	PT	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	Hm	
	Kapit (n=37)	Kapit (n=10)	Kapit (n=185)	Betong (n=78)	Kanowit (n=34)	Sarikei (n=27)	Miri (n=50)	Kudat (n=30)	Ranau (n=42)	Tenom (n=26)	Kelantan (n=30)	Pahang (n=50)					
NC12_2	0.85	0.82	0.86	0.85	0.86	0.84	0.84	0.84	0.90	0.86	0.80	0.92					
NC03_2	0.57	0.00	0.51	0.55	0.43	0.63	0.51	0.54	0.63	0.58	0.48	0.60					
NC09_1	0.78	0.78	0.89	0.85	0.89	0.92	0.85	0.90	0.77	0.87	0.54	0.63					
NC12_4	0.86	0.51	0.84	0.85	0.74	0.80	0.77	0.81	0.81	0.77	0.63	0.70					
NC10_1	0.65	0.78	0.73	0.72	0.57	0.67	0.65	0.53	0.61	0.66	0.76	0.83					
CD08_61	0.82	0.00	0.67	0.15	0.47	0.14	0.42	0.83	0.82	0.85	0.83	0.80					
CD11_157	0.77	0.84	0.80	0.79	0.81	0.88	0.77	0.76	0.84	0.80	0.67	0.75					
CD13_61	0.52	0.36	0.76	0.73	0.82	0.72	0.66	0.69	0.73	0.67	0.77	0.72					
CD05_06	0.81	0.60	0.78	0.80	0.80	0.82	0.74	0.73	0.79	0.55	0.78	0.70					
CD13_107	0.47	0.76	0.63	0.41	0.82	0.72	0.76	0.13	0.43	0.50	0.73	0.56					
Mean $H_E$	0.71	0.54	0.75	0.67	0.72	0.71	0.70	0.67	0.73	0.71	0.70	0.72					
SE	0.05	0.10	0.04	0.07	0.05	0.07	0.04	0.07	0.04	0.04	0.04	0.04					

The data was analysed based on the single clone samples and only predominant alleles in the multiple clone infections. Abbreviation: LT – long-tailed macaque, PT – pig-tailed macaque, Hm – human.

frequent overall, but Cluster 2 was also common at each of the sites in Sarawak, particularly in Miri and Kanowit where it was more frequent than Cluster 1 (Table 2.5; Appendix 2.2).

Over all human infections, there was a similarly high level of divergence in allele frequencies between the two subpopulation clusters as was seen between parasites from the two different macaque host species ( $F_{ST} = 0.194$ ,  $P < 0.001$ ). As expected, the degree of cluster admixture at each sampling site ( $p1 * p2$ , where  $p1$  and  $p2$  are the local frequencies of Cluster 1 and Cluster 2 respectively) correlated positively with the ( $I_A^S$ ) index of multi-locus linkage disequilibrium (Table 2.6; Figure 2.5, Spearman's Rho = 0.678,  $P = 0.015$ ).

Analysis of geographical divergence on the basis of  $F_{ST}$  indices derived from population allele frequencies (Table 2.7) identified a pattern strongly consistent with isolation by distance (Mantel test of matrix correlation  $P < 0.0001$ , Figure 2.6A). The greatest level of divergence was seen between peninsular Malaysia and Borneo as expected, although isolation by distance was also apparent within Borneo (Mantel test  $P = 0.0016$ ). The overall pattern consistent with isolation by distance remained when only infections with Cluster 1 genotypes were analysed ( $P = 0.0016$ , Figure 2.6A). There was a similar trend for the smaller number of samples with Cluster 2 genotypes, although this was not significant ( $P = 0.0922$ ), indicating that the majority of the geographical differentiation is independent of the Cluster subpopulation structure. A principal components analysis of all individual infection genotypes showed that most of the overall diversity is among those defined as Cluster 1 by the STRUCTURE analysis (Cluster 2 infections covered only part of the first principal component distribution),

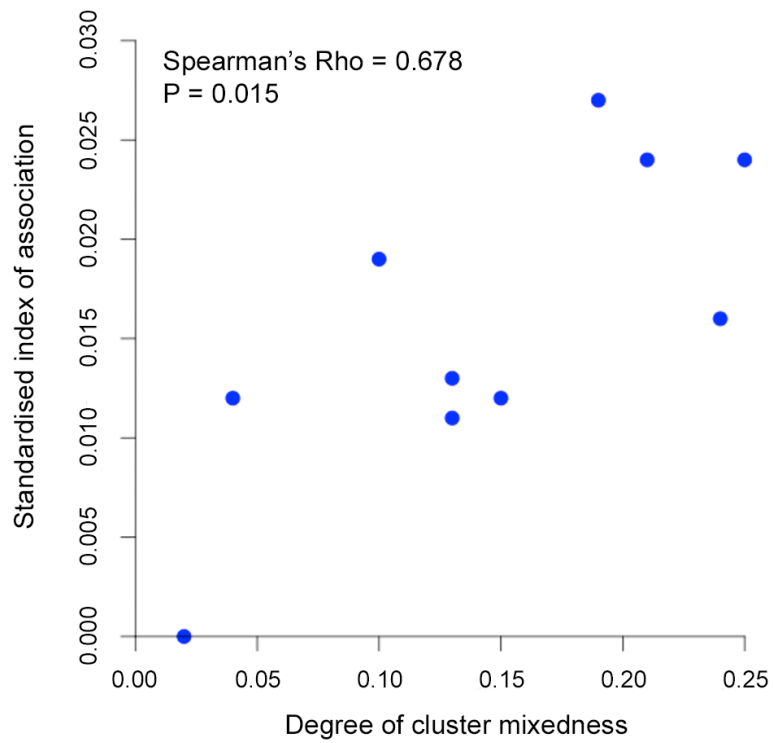
**Table 2.5.** Proportion of isolates designated as Cluster 1 and Cluster 2 in 10 geographical populations of human *P. knowlesi* isolates.

Regions	Sites	Frequency of Cluster 1 ( $p_1$ )	Frequency of Cluster 2 ( $p_2$ )	Degree of cluster mixedness ( $p_1 * p_2$ )
Sarawak	Kapit	0.71	0.29	0.21
Sarawak	Betong	0.82	0.18	0.15
Sarawak	Kanowit	0.45	0.55	0.25
Sarawak	Sarikei	0.58	0.42	0.24
Sarawak	Miri	0.26	0.74	0.19
Sabah	Kudat	0.96	0.04	0.04
Sabah	Ranau	0.89	0.11	0.10
Sabah	Tenom	0.85	0.15	0.13
Peninsular	Kelantan	0.84	0.16	0.13
Peninsular	Pahang	0.98	0.02	0.02

Pearson's  $\chi^2$  for test of homogeneity of proportions across populations,  $P < 0.0001$

**Table 2.6.** Test of multi-locus linkage disequilibrium of *P. knowlesi* isolated in humans by measuring the standardised index of association ( $I_A^S$ ) using only unique haplotypes in each geographical sites.

Regions	Sites	n	$I_A^S$	P value
Sarawak	Kapit	164	0.024	<0.001
Sarawak	Betong	68	0.012	0.035
Sarawak	Kanowit	33	0.024	0.005
Sarawak	Sarikei	23	0.016	0.090
Sarawak	Miri	47	0.027	<0.001
Sabah	Kudat	28	0.012	0.133
Sabah	Ranau`	38	0.019	0.008
Sabah	Tenom	25	0.011	0.185
Peninsular	Kelantan	24	0.013	0.167
Peninsular	Pahang	50	-0.001	0.564



**Figure 2.5.** Correlation between degree of cluster admixture and multi-locus linkage disequilibrium (standardised index of association).

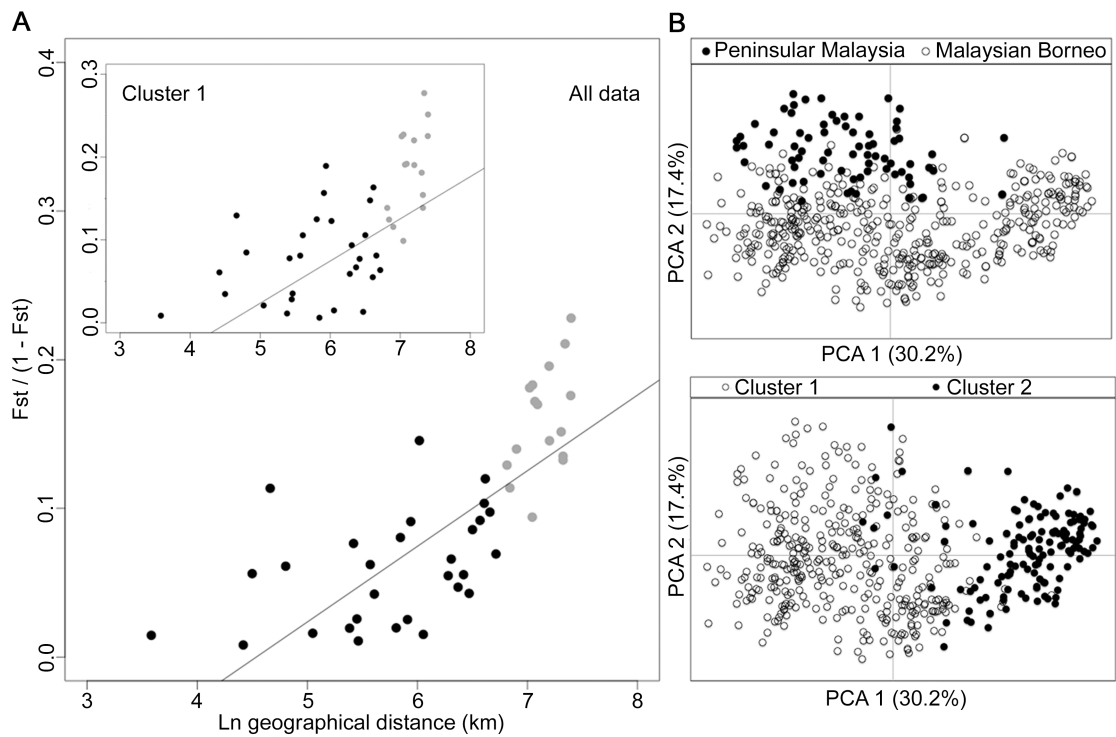
The degree of *P. knowlesi* cluster admixture was estimated as the local cluster mixedness ( $p1*p2$ ) based on the proportions of infections designated as Cluster 1 ( $p1$ ) and Cluster 2 ( $p2$ ) at each of 10 sampling sites for human infections across Malaysia (Spearman's Rho = 0.678, P = 0.015).

**Table 2.7.** Pairwise measures of fixation indices ( $F_{ST}$  values above diagonal) and geographical distance (in kilometres below diagonal) across 10 populations of *P. knowlesi* from human isolates.

Regions	Hosts	Sampling sites	Sarawak					Sabah				Peninsular Malaysia			
			Human	Human	Human	Human	Human	Human	Human	Human	Human	Human	Human	Human	Human
			Kapit	Betong	Kanowit	Sarikei	Miri	Kudat	Ranau	Tenom	Kelantan	Pahang			
Sarawak	Human	Kapit	-	0.025***	0.016*	0.011	0.041***	0.041***	0.062***	0.015	0.127***	0.086***			
Sarawak	Human	Betong	233	-	0.053***	0.014	0.084***	0.065***	0.107***	0.053**	0.155***	0.102***			
Sarawak	Human	Kanowit	156	90	-	0.008	0.019*	0.094***	0.079***	0.052**	0.145***	0.123***			
Sarawak	Human	Sarikei	236	36	83	-	0.025	0.089***	0.084***	0.045	0.153***	0.114***			
Sarawak	Human	Miri	273	380	333	369	-	0.127***	0.074***	0.071**	0.164***	0.147***			
Sabah	Human	Kudat	646	824	741	780	411	-	0.102***	0.019	0.186***	0.119***			
Sabah	Human	Ranau	594	748	666	713	346	106	-	0.058*	0.150***	0.132***			
Sabah	Human	Tenom	426	614	534	584	226	218	122	-	0.174***	0.117***			
Peninsular	Human	Kelantan	1340	1149	1202	1119	1336	1631	1625	1544	-	0.059***			
Peninsular	Human	Pahang	1146	935	993	911	1174	1516	1490	1516	263	-			

Indicative adjusted nominal level (5%) for multiple comparisons was 0.000758 after Bonferroni corrections. P-values were obtained after 66,000 permutations.  $F_{ST}$  values in (\*\*\*) indicate highly significant, (\*\*) indicate moderately significant, (\*) indicate weakly significant and white indicate not significant.





**Figure 2.6.** Isolation-by-distance model and principal component analysis (PCA) of the human *P. knowlesi* isolates.

(A) Relationship between transformed genetic differentiation and natural log of geographical distance (Euclidean distances of population pairs ranged from 36 km to 1631 km) for all pairs of sites across Malaysia (Mantel test of matrix correlation  $P < 0.0001$ ), with a similar relationship when analysing only isolates from the Cluster 1 subpopulation (Mantel test  $P = 0.0016$ ). Black dots denote pairs of populations within Malaysian Borneo and Peninsular Malaysia, while grey dots denote pairs of populations between Malaysian Borneo and Peninsular Malaysia. Evidence of isolation by distance remained when only sites within Borneo were considered (Mantel test  $P = 0.0016$ ). (B) PCA of the whole infection haplotype dataset indicated differentiation of Peninsular Malaysia isolates from Malaysian Borneo isolates by the second principal component axis, whereas isolates defined as Cluster 1 and Cluster 2 by STRUCTURE analysis were almost completely differentiated along the first principal component axis.

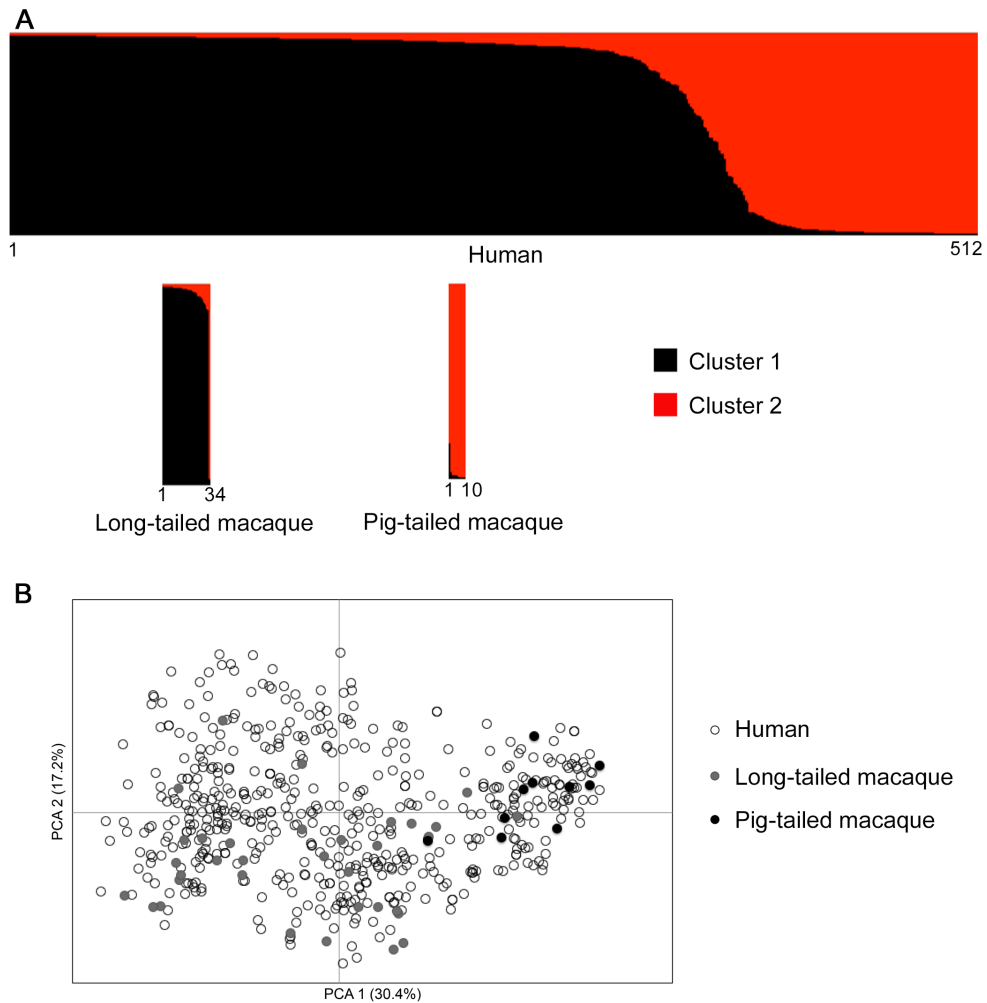
and infections from peninsular Malaysia are restricted to part of the second principal component distribution (Figure 2.6B).

Combination of the macaque samples together with all of the human samples across the 10 geographical locations confirmed the definition of the two *P. knowlesi* subpopulation clusters (Figure 2.7), which correspond to those shown above. Allele frequency distributions showed that some loci were particularly differentiated between the subpopulation clusters, with  $F_{ST} > 0.3$  for loci NC03\_2 and CD13\_61 (Figure 2.8).

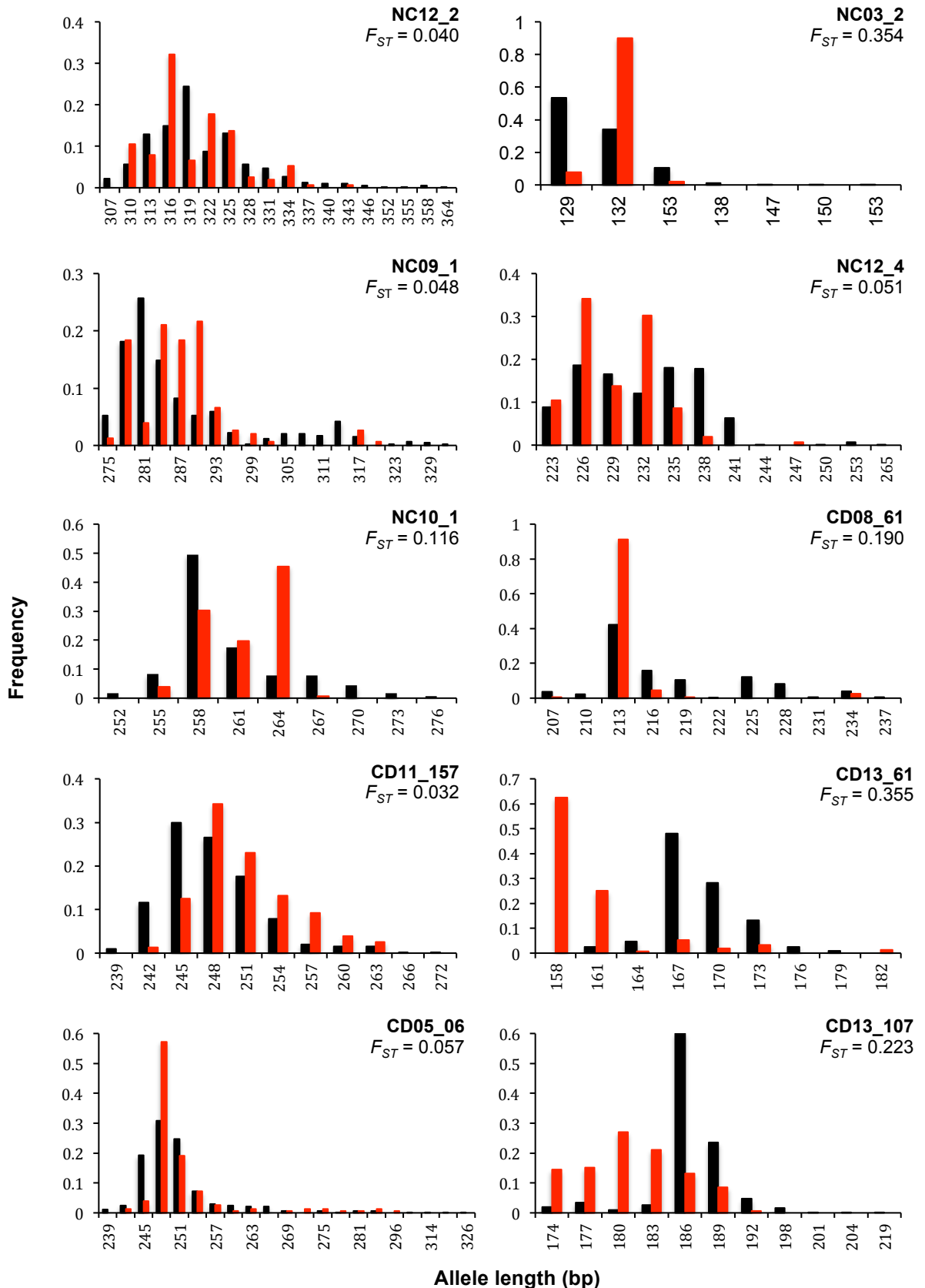
#### **2.3.4 Evaluation of cluster assignment indices**

Most individual infection genotypes had a clear majority of putative ancestry assignment to either Cluster 1 or Cluster 2, but a small minority of infections had a more intermediate profile (Figure 2.3A and Figure 2.4C). Quantitative analysis of the proportional Cluster 1 and Cluster 2 ancestry assignments for each infection genotype based on the STRUCTURE analysis yielded an index of the degree of intermediate cluster assignment for each infection. This has a maximum possible value of 0.5, although most infections had values closer to zero.

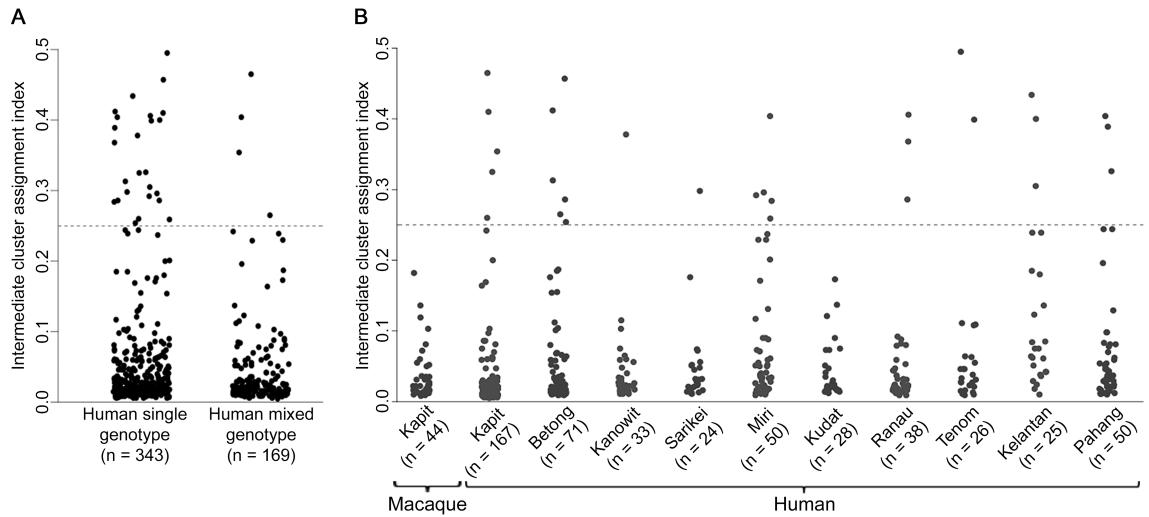
The intermediate cluster assignment indices showed no difference between single and mixed genotype human infections (Mann-Whitney test  $P = 0.20$ , Figure 2.9A), whereas both of these independently had higher indices than the macaque infections ( $P < 0.001$  for both comparisons). When analysis was focused on Kapit alone, the distribution of intermediate cluster indices were not significantly different between human and macaque infections ( $P = 0.25$ , Figure 2.9B). However, human infections from Kelantan



**Figure 2.7.** Population genetic structure of *P. knowlesi* infections from all 512 humans, 34 long-tailed macaques and 10 pig-tailed macaques in Malaysia. (A) Bayesian model-based STRUCTURE analysis corresponds to those mentioned in Figure 2.3A where two subpopulation clusters were observed throughout the whole dataset ( $K = 2$ ,  $\Delta K = 136.39$ ). When all of the human and macaque samples were analysed together, little evidence of admixture was seen between parasites sampled in the two macaque populations. (B) The principal component analysis (PCA) of all *P. knowlesi* isolates also indicates the infections from different macaque host species were almost completely separated by the first principal component while human infections were widely distributed throughout the full range on both axes.



**Figure 2.8.** Allele frequency distributions and genetic differentiations of 10 microsatellite loci between two *P. knowlesi* subpopulation clusters. Major subpopulation cluster (Cluster 1,  $n = 404$ ) is colour-coded in black while minor subpopulation cluster (Cluster 2,  $n = 152$ ) is in red.



**Figure 2.9.** Intermediate cluster assignment indices in *P. knowlesi* infections in humans and macaques.

The index for each infection was based on the proportion of shared ancestry between Cluster 1 and Cluster 2 inferred by the STRUCTURE analysis on all samples genotyped at 10 microsatellite loci. An index below 0.25 signifies the individual parasites were predominantly either Cluster 1 or Cluster 2 while those above 0.25 had more intermediate assignment (up to a maximum index value of 0.5). (A) There was no significant difference between single or multiple genotype human infections in the distribution of the indices (Mann-Whitney U test,  $P = 0.20$ ). (B) Intermediate cluster assignment indices were lower in infections from macaques in Kapit than in human infections overall (Mann-Whitney U test,  $P < 0.001$ ), but were not significantly different from human infections in Kapit ( $P = 0.25$ ). Among geographical sites, the indices for human infections were significantly higher in Kelantan than in Kapit, Betong, Kanowit, Kudat or Ranau ( $P < 0.05$  for each pairwise comparison after Bonferroni correction).

had a significantly higher distribution of values compared to five of the sites in Borneo (Mann-Whitney test  $P < 0.05$  for each comparison after Bonferroni correction) (Figure 2.9B). Across the different sites, there was no significant correlation between the local population admixture of both clusters ( $p1 * p2$ , Table 2.5) and the mean or variance of intermediate cluster assignment indices ( $P = 0.33$  and  $P = 0.59$  respectively).

## 2.4 Discussion

This study shows that human *P. knowlesi* is an admixture of two divergent parasite populations associated with different forest-dwelling macaque reservoir hosts. In human infections, the long-tailed macaque-associated *P. knowlesi* type (Cluster 1) is most common overall and at most of the geographical sites, while the pig-tailed macaque-associated type (Cluster 2) is also common at sites in Sarawak. The estimate of divergence between these two sympatric parasite subpopulations ( $F_{ST}$  index of  $\sim 0.22$  averaged over 10 microsatellite loci) may be conservative, due to high allelic diversity of the microsatellite loci which restricts the potential upper range of fixation indices (Balloux and Lugon-Moulin, 2002, Meirmans and Hedrick, 2011). The differentiation varied among the loci, with two of the microsatellite loci being particularly highly differentiated between the clusters ( $F_{ST} \sim 0.35$ ), so the robustness of the two assigned clusters was confirmed by repeat analyses which excluded these. Previous analysis of *P. knowlesi* mitochondrial DNA sequences from a relatively small number of human and long-tailed macaque infections in Kapit did not indicate two divergent lineages (Lee et al., 2011), although analysis of samples from Sabah suggests that sequences from pig-tailed macaque infections are differentiated from sequences from long-tailed macaque infections (Muehlenbein et al., 2015).

The results confirm that humans have mostly single genotype *P. knowlesi* infections whereas macaques have polyclonal infections, supporting the expectation that there is a higher rate of transmission among macaques (Lee et al., 2011, Tan et al., 2008). The number of *P. knowlesi* genotypes per infection in humans is lower than was previously seen in microsatellite analyses of the endemic human malaria parasites *P. falciparum* and *P. vivax* in some of the same areas in Malaysia (Anthony et al., 2005, Abdullah et al., 2013), whereas the number of *P. knowlesi* genotypes per infection in macaques is much higher. Levels of multi-locus linkage disequilibrium in *P. knowlesi* here are lower than reported in *P. vivax* or *P. falciparum* in these areas (Abdullah et al., 2013, Anthony et al., 2005), indicating that recombination in *P. knowlesi* probably commonly occurs in mosquitoes containing a macaque blood meal with multiple parasite genotypes.

It is unknown how the two sympatric *P. knowlesi* subpopulations are genetically isolated. The observation of a single long-tailed macaque with a *P. knowlesi* Cluster 2 type infection (otherwise only seen in pig-tailed macaques and humans) suggests there is not an absolute barrier in terms of primate host susceptibility, although there are differences in ecology. Additional sampling of both long-tailed and pig-tailed macaques will be important to confirm the host associations of different parasites (Muehlenbein et al., 2015). Both macaque species are widespread, but long-tailed macaques prefer secondary forest near human settlements where they have access to farms for food, whereas pig-tailed macaques spend more time in ground foraging in primary forests, generally having less frequent contact with humans (Fa and Lindburg, 1996). There may be differential susceptibility of mosquito species to the respective parasite types, as suggested for subpopulations of another malaria parasite elsewhere (Joy et al., 2008), or different mosquitoes may feed on the respective macaque host species.

Genetic differentiation in *P. knowlesi* was also strongly correlated with geographical distance, overall and for the Cluster 1 parasites. The observation of highest  $F_{ST}$  values between populations from Malaysian Borneo and Peninsular Malaysia was expected, as the South China Sea has separated macaques in these areas since the last glacial period (Ziegler et al., 2007), but a test for isolation by distance remained significant when analysing only sites within Borneo.

A small minority of human infections had intermediate cluster assignment indices, which could potentially result from occasional crossbreeding between the two genotypic clusters, although this cannot be concluded from these data alone.

Hybridisation between species or sub-species can offer opportunities for adaptation, and has been associated with emergence of novel host-specificity or pathogenicity in other parasitic protozoa (Goodhead et al., 2013) and fungi (Stukenbrock et al., 2012). Switching of host species has occurred repeatedly in malaria parasites of birds (Ricklefs et al., 2014) and small mammals (Schaer et al., 2013), as well as apes and humans (Liu et al., 2010, Liu et al., 2014), but the occurrence of parasite hybridisation and introgression has not been investigated. The potential occurrence of inter-cluster hybridisation in even a minority of human *P. knowlesi* infections, combined with the possibility of human-mosquito-human transmission, may increase the potential for *P. knowlesi* adaptation to the human host or to mosquito species that are more abundant than the currently known forest-associated vectors.

Genome-wide analysis of *P. knowlesi* populations would enable further evaluation of the genetic structure of this zoonotic parasite species, and scanning for loci under selection within each of the two subpopulations. Human clinical isolates containing



single species infections would be relatively straightforward to analyse, as *P. knowlesi* sequences would be unmixed with those of other human malaria species. In contrast, as natural macaque infections usually contain a mixture of different malaria parasite species (Lee et al., 2011), to obtain unambiguous genome sequences it may be necessary to sequence from individual parasites isolated from these hosts (Nair et al., 2014). Although experimental studies on *P. knowlesi* are usually conducted *in vivo* in non-human primates (Lapp et al., 2013, Hamid et al., 2011, Murphy et al., 2014), new approaches to adapt the parasites to *in vitro* growth using human erythrocytes have been successful (Lim et al., 2013, Moon et al., 2013). Analysis of phenotypic differences between the different host-associated types may be investigated using both *in vivo* and *in vitro* experimental systems, while continued epidemiological and clinical surveillance for increasing incidence or disease severity is of the highest priority.

# Chapter Three

**Registry**

T: +44(0)20 7299 4646  
 F: +44(0)20 7299 4656  
 E: [registry@lshtm.ac.uk](mailto:registry@lshtm.ac.uk)

**RESEARCH PAPER COVER SHEET**

**PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.**

**SECTION A – Student Details**

<b>Student</b>	Paul Cliff Simon Divis
<b>Principal Supervisor</b>	Professor David Conway
<b>Thesis title</b>	Population Genetic Structure and Genomic Divergence in <i>Plasmodium knowlesi</i>

***If the Research Paper has previously been published please complete Section B, if not please move to Section C***

**SECTION C – Prepared for publication, but not yet published**

Where is the work intended to be published?	Malaria Journal
Please list the paper's authors in intended authorship order:	Paul CS Divis, Ting H Hu, Samuel A Assefa, Balbir Singh, David J Conway
Stage of publication	Draft in preparation

**SECTION D – Multi-authored work**

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper (Attach a further sheet if necessary)	I designed the study, performed laboratory experiments, conducted statistical analyses and draft the manuscript.
---	--

**Student Signature:** 

**Date:** 13<sup>th</sup> March 2017

**Supervisor Signature:** 

**Date:** 20<sup>th</sup> March 2017

## Development of a simple genotyping assay for discriminating sympatric *Plasmodium knowlesi* subpopulations

### 3.1 Introduction

In Malaysian Borneo, two macaque species (*Macaca fascicularis* and *M. nemestrina*) have been incriminated as the reservoir hosts of *P. knowlesi* (Lee et al., 2011) while *Anopheles latens* and *An. balabacensis* have been shown to be vectors (Tan et al., 2008, Wong et al., 2015). Analyses in Chapter Two indicate that these two macaque species are reservoirs for *P. knowlesi* in humans, and that the parasites are divided into two genetically divergent subpopulations (Divis et al., 2015). This population substructure has been subsequently further confirmed by whole genome sequence analysis of parasites in human infections sampled from Sarawak state in Borneo where the two subpopulations sympatrically co-exist (Assefa et al., 2015). However, the sequence analysis showed that previously collected laboratory macaque-passaged isolates of *P. knowlesi*, which had mostly originated from Peninsular Malaysia, formed a third divergent subpopulation (Assefa et al. 2015).

Because this discovery of population structure has been very recent, it is not yet known whether there are differences in the clinical presentations between the major parasite genetic subpopulations. Nonetheless, a dimorphism of *P. knowlesi* normocyte binding protein (*Pknbpxa*), a gene encoding protein involved in red blood cell invasion, has been putatively associated with disease severity in a separate analysis (Ahmed et al., 2014, Pinheiro et al., 2015). In order to investigate further the biological and

clinical differences in *P. knowlesi* infections between the two major subpopulations, it is useful to design a rapid genotyping assay that is less demanding of laboratory time than multi-locus microsatellite analysis or whole genome sequencing.

In this chapter, a simple genotyping assay for distinguishing two divergent parasite subpopulations that exist sympatrically in Malaysian Borneo has been developed and validated for its reliability. However, this is shown to work only for *P. knowlesi* infections sampled in Borneo, but not for infections acquired in Peninsular Malaysia. This indicates that further analysis of the population genetic composition of the latter region is needed. This highly specific assay is potentially useful as the first step to identify the source of *P. knowlesi* infections in humans associated with different macaque host species in Borneo, enabling larger scale epidemiological surveillance and clinical studies.

## **3.2 Materials and methods**

### **3.2.1 DNA samples**

A total of 397 DNA samples of *P. knowlesi* from human clinical isolates were analysed in this chapter. These were sampled from eight sites in Malaysian Borneo: Betong (n = 29), Kanowit (n = 34), Miri (n = 46), Sarikei (n = 23) and Kapit (n = 52) from the Sarawak state, and Kudat (n = 46), Ranau (n = 62) and Tenom (n = 48) from the Sabah state, and two sites in Peninsular Malaysia: Kelantan (n = 15) and Pahang (n = 42). Subpopulation assignments (Cluster 1 or Cluster 2) for each individual infection were previously made by multi-locus microsatellite analysis in Chapter Two (Divis et al., 2015) or whole genome sequencing analysis (Assefa et al., 2015). DNA samples from long-tailed

macaques (from Kapit in Borneo, n = 10; and Selangor in Peninsular Malaysia, n = 15) and pig-tailed macaques (from Kapit, n = 5) were also included in the analysis here.

To demonstrate species-specificity of new PCR assays, DNA controls of human *Plasmodium* species (*P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae*) and Southeast Asian macaques (*P. knowlesi*, *P. coatneyi*, *P. inui*, *P. cynomolgi* and *P. inui*), as well as human, long-tailed and pig-tailed macaque DNA were also included in the analysis.

### **3.2.2 Single nucleotide polymorphism (SNP) of *P. knowlesi***

High-quality biallelic SNPs were obtained from a database of a separate study investigating the population genomics of *P. knowlesi* (Assefa et al., 2015). A SNP dataset was previously generated following mapping the short-read sequences against the version 1.0 of *P. knowlesi* strain H reference genome. Isolates that were used in the previous study, which were obtained from the Malaysian Borneo, were separated into two main groups: Cluster 1 (n = 38) was associated with long-tailed macaque host and Cluster 2 (n = 10) was associated with pig-tailed macaque host (Divis et al., 2015, Assefa et al., 2015). A total of 9,293 SNPs with complete fixation (pairwise fixation index,  $F_{ST} = 1$ ) of alternative alleles between these two subpopulation clusters were identified. Data from the laboratory-adapted lines that were seen to form a third subpopulation cluster based on the whole genome analysis (Assefa et al., 2015) were not used for the discriminatory SNP assay design here.

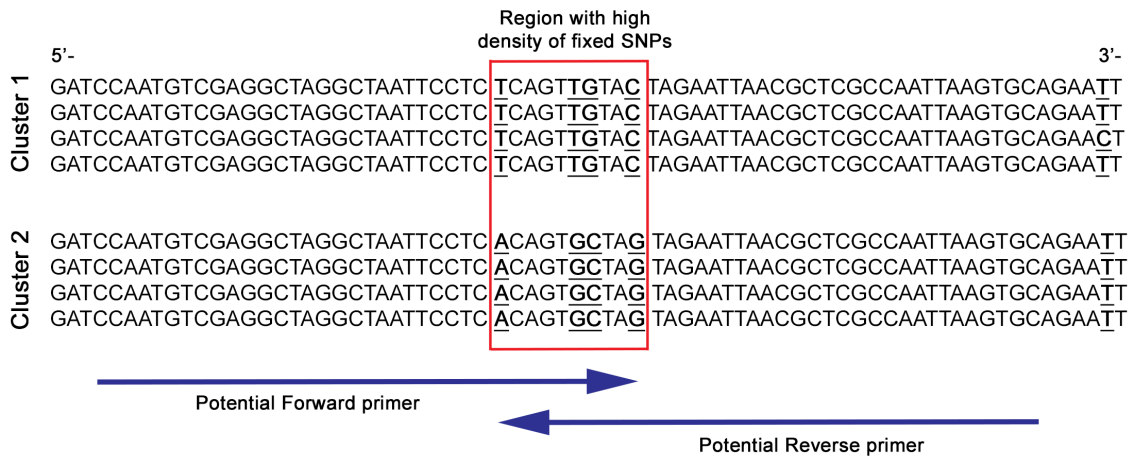
### **3.2.3 Development of allele-specific markers**

#### **3.2.3.1 Designing of PCR primers**

Genome regions containing a high density of fixed SNPs were identified, having multiple SNPs less than 5 nucleotides apart. This is useful to design high-specificity PCR primers to distinguish the two subpopulations, with discrimination particularly at the 3'-ends. Regions containing incompletely fixed SNPs ( $F_{ST} < 1$ ) within 30-nucleotide length were omitted, resulting in only regions with putative type-specific fixed SNPs together with non-polymorphic nucleotides that were suitable for PCR primer design (Figure 3.1). Oligonucleotides for allele-specific PCR assays for each subpopulation Cluster 1 and Cluster 2 were designed using the PrimerSelect software (DNASTAR, USA), based on conserved DNA sequence of *P. knowlesi* strain H reference genome ([www.genedb.org](http://www.genedb.org)). DNA sequences for each oligonucleotide were checked for any potential of cross-species matching with other *Plasmodium* species (human or macaque parasites) using BLAST search of available DNA sequences ([www.ncbi.nlm.nih/BLAST/](http://www.ncbi.nlm.nih/BLAST/)).

### **3.2.3.2 Conventional PCR assay**

PCR assays specific for each subpopulation cluster were optimised in 11- $\mu$ l reaction volume containing 1X Green GoTaq® Flexi buffer (Promega, Madison WI, USA), 5 mM MgCl<sub>2</sub>, 0.2 mM each dNTP (Bioline, UK), 0.275 U GoTaq® DNA polymerase (Promega, Madison WI, USA), 1 – 2  $\mu$ l DNA template and 0.25  $\mu$ M of each forward and reverse primers. Each primer pair underwent gradient PCR in order to determine the optimum annealing temperature with the following thermal cycling conditions: 94°C for 2 min, followed by 35 cycles of initial denaturation of 94°C for 30 sec, gradient annealing from 50 to 66°C for 1 min, extension at 72°C for 30 sec, and final elongation at 72°C for 2 min.



**Figure 3.1.** Example of a region in *P. knowlesi* genome for developing an allele-specific marker.

Each sequence in Cluster 1 and Cluster 2 subpopulations represents data from an individual isolate (only four from each cluster are shown but all isolate sequences were used for the design). SNPs are shown in bold and underlined while region consisting of high density of fixed SNPs ( $F_{ST} = 1$ ) are highlighted in red box. Potential forward and reverse PCR primers are designed surrounding this region where no sequence polymorphisms are found.

### 3.2.3.3 Touchdown PCR assay

Touchdown (TD) PCR was also tested in order to increase the specificity and sensitivity in PCR amplification (Korbie and Mattick, 2008). This PCR method was performed in two phases. Following the initial denaturation of 94°C for 2 min, DNA was subjected to 15 cycles of TD programme of 94°C for 30 sec, annealing temperature of 10°C above the primer melting temperature for 45 sec and 72°C for 30 sec, with a 1°C decrease of the annealing temperature every cycle. Following the TD programme, additional 25 cycles of amplification were performed with the following conditions: 94°C for 30 sec, annealing temperature of 5°C below the primer melting temperature for 45 sec, 72°C for 30 sec and final elongation at 72°C for 5 min. PCR products were separated by electrophoresis on agarose gel stained with ethidium bromide, and visualised under ultraviolet transillumination.



### 3.3 Results

#### 3.3.1 Allele-specific markers to discriminate *P. knowlesi* subpopulations

Five sets of PCR amplification assays were developed for each subpopulation Cluster 1 and Cluster 2, with oligonucleotides containing complete SNP fixation of alternative alleles accumulated at the 3'-end (Table 3.1). The strict criteria that were used to develop these PCR assays allowed one primer set located in Chromosome 1 while four were in Chromosome 6 for each subpopulation cluster. Because the SNPs were fixed between these two subpopulations, PCR primers were designed at the same loci for both subpopulation clusters.

All ten PCR assays were tested independently for determining the optimum PCR conditions. Of these, four assays did not perform well and were excluded from the *P. knowlesi*-specificity test. Six PCR assays showed to be highly specific for *P. knowlesi* when tested alongside four *Plasmodium* species of humans, four other *Plasmodium* species of long-tailed and pig-tailed macaques, as well as human and macaque host DNA (Table 3.2).

Further optimisation of PCR assays for the specificity against both *P. knowlesi* subpopulations, one assay each for Cluster 1 (primer set C1A) and Cluster 2 (primer set C2J) subpopulations showed complete specificity with no cross-reaction between parasites of the two subpopulations (Table 3.3). The amplification products ranged between 200 and 300 bp for primer set C1A and approximately 500 bp for primer set C2J (Figure 3.2).

**Table 3.1.** Unique primer pairs designed at each locus to discriminate two *P. knowlesi* subpopulation clusters.

Subpopulation	Set	Chromosome (location of locus)*	Primer ID	Sequence (5' → 3')
Cluster 1	C1A	1 (745,563 – 745,763) <sup>a</sup>	C101AF	GTTTGGTACGTTCAAGTGTGCGCTATGG
			CX01AR	CGTCTCCGCTTGTGTTTTCCATGTAC
	C1B	6 (35,925 – 36,162) <sup>b</sup>	C106AF	TCCATGTGCACCCTGGCATAACATGGTAC
			C106AR	TGTACAGAGTGTACAGGAGCTGGGAC
	C1C	6 (17,264 – 17,523) <sup>c</sup>	C106BF	GATATAACCACATGTTTGCTTCGAAGGAA
C106BR			GGAAAGGTACCTCTTCCTCATAGTCCC	
C1D	6 (1,014,117 – 1,014,286) <sup>d</sup>	C106CF	GGATGATTTAGGTAAGGATGAGGAGGGT	
		CX06CR	CGTCATCCTTATCCTTTTTACCTTATCC	
C1E	6 (1,039,470 – 1,039,954) <sup>e</sup>	C106DF	GATGATAATTATCTTAAAGAGCCGGATG	
		C106DR	CAAGACATTATGAACATTGGACCGATTA	
Cluster 2	C2F	1 (745,563 – 745,763) <sup>a</sup>	C201AF	GTTTGGTACGTTCAAGTGTGCTCTACAT
			CX01AR	CGTCTCCGCTTGTGTTTTCCATGTAC
	C2G	6 (17,264 – 17,523) <sup>c</sup>	C206AF	GATATAACCACATGTTTGCTTCGAAAGAG
			C206AR	GGAAAGGTACCTCTTCCTCATAGTCCA
	C2H	6 (35,925 – 36,162) <sup>b</sup>	C206BF	TCCATGTGCACCCTGGCATAACATGGCAT
			C106AR	TGTACAGAGTGTACAGGAGCTGGGAC
C2I	6 (1,014,117 – 1,014,286) <sup>d</sup>	C206CF	GGATGATTTAGGTAAGGATGAGGAGTGC	
		CX06CR	CGTCATCCTTATCCTTTTTACCTTATCC	
C2J	6 (1,039,470 – 1,039,954) <sup>e</sup>	C206DF	GATGATAATTATCTTAAAGAGCCGGAG	
		C206DR	CAAGACATTATGAACATTGGACCGACTG	

SNPs with complete fixed ( $F_{ST} = 1$ ) as shown at the 3'-end sequence in bold and underlined fonts. Superscripted letters (\*) at the end of locations indicate same loci for both subpopulation clusters.

**Table 3.2.** Summary of PCR assay optimisations for discriminating two subpopulations of *P. knowlesi* infections.

Primer set	Types of PCR	Optimum annealing temperature (°C)	Allele specificity	PCR summary
C1A	Conventional PCR	60	Cluster 1	Cluster 1 allele-specific
C1B	Conventional PCR	<i>nd</i>	<i>nd</i>	Weak amplification at gradient PCR
C1C	Conventional PCR	61	Cluster 1 and Cluster 2	Not specific
C1D	Conventional PCR	<i>nd</i>	<i>nd</i>	Weak amplification at gradient PCR
C1E	Touchdown PCR	68, 60	Cluster 1 and Cluster 2	Not specific; faint bands on Cluster 2 DNA
C2F	Conventional PCR	60	Cluster 2	Cluster 1 allele-specific but inconsistent amplifications
C2G	Touchdown PCR	68, 60	Cluster 1 and Cluster 2	Not specific
C2H	Conventional PCR	<i>nd</i>	<i>nd</i>	Weak amplification at gradient PCR
C2I	Conventional PCR	<i>nd</i>	<i>nd</i>	Weak amplification at gradient PCR
C2J	Touchdown PCR	68, 62	Cluster 2	Cluster 2 allele-specific

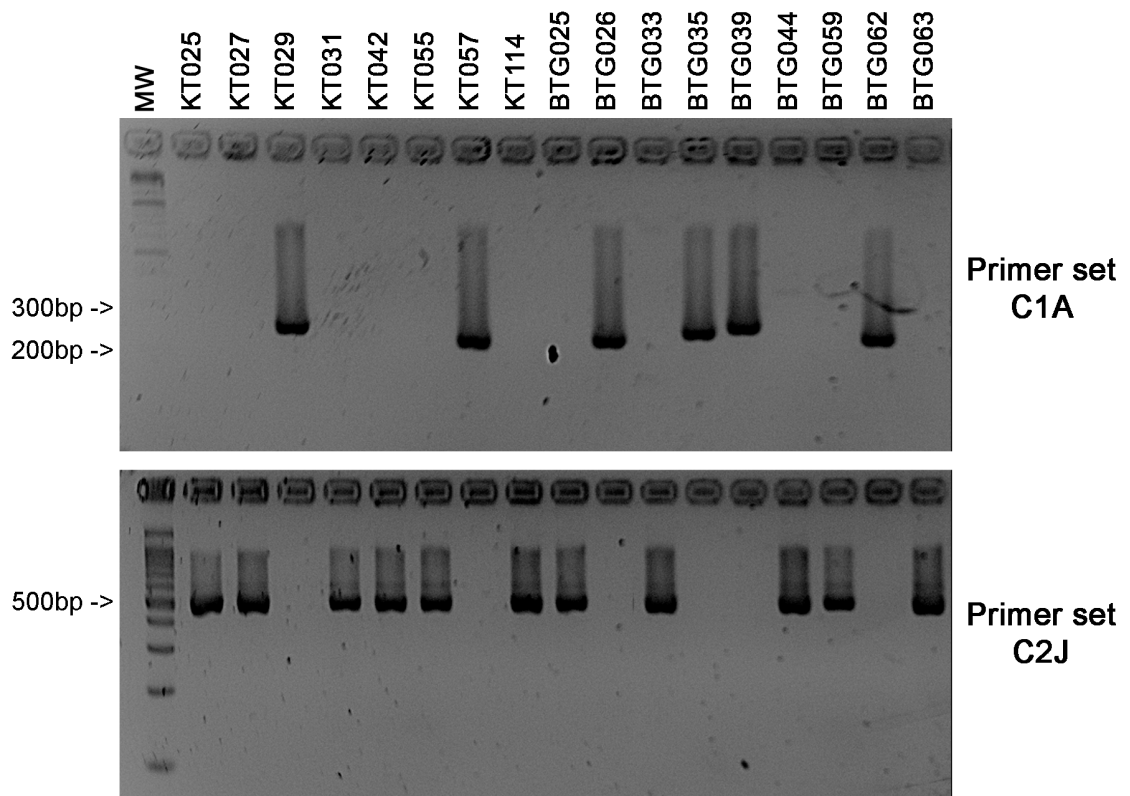
The allele specificity of primers tested on DNA of Cluster 1 and Cluster 2 clinical infections where identity of infections were confirmed previously by Assefa *et al.* (2015) and Divis *et al.* (2015). All primer sets, except for sets C1B, C1D, C2H and C2I, did not cross-react to *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, *P. coatneyi*, *P. fieldi*, *P. inui*, *P. cynomolgi*, human and macaque DNA. Abbreviation: *nd* – not done.

**Table 3.3.** Summary of PCR cycling condition of primer sets C1A and C2J for genotyping *P. knowlesi* parasites of Cluster 1 and Cluster 2 subpopulations, respectively.

Primer set C1A for Cluster 1 subpopulation	94°C	2 minutes	35 cycles
	94°C	30 seconds	
	60°C	1 minute	
	72°C	30 seconds	
	72°C	5 minutes	

Primer set C2J for Cluster 2 subpopulation	94°C	2 minutes	10 cycles
	94°C	30 seconds	
	68°C, decrease 1°C every cycle	45 seconds	
	72°C	30 seconds	
	94°C	30 seconds	25 cycles
	62°C	45 seconds	
	72°C	30 seconds	
	72°C	5 minutes	

PCR chemistry based on GoTaq® Flexi DNA Polymerase (Promega, USA)



**Figure 3.2.** Gel electrophoresis shows allele specificity of PCR primer sets C1A and C2J for discriminating *P. knowlesi* infections of Cluster 1 and Cluster 2 subpopulations, respectively.

Band sizes of primer set C1A may differ, ranging between 200 and 300 bp, while band sizes of primer set C2J are virtually consistent at approximately 500bp. Primer set sequences are shown in Table 3.1. Abbreviation: MW – molecular weight marker

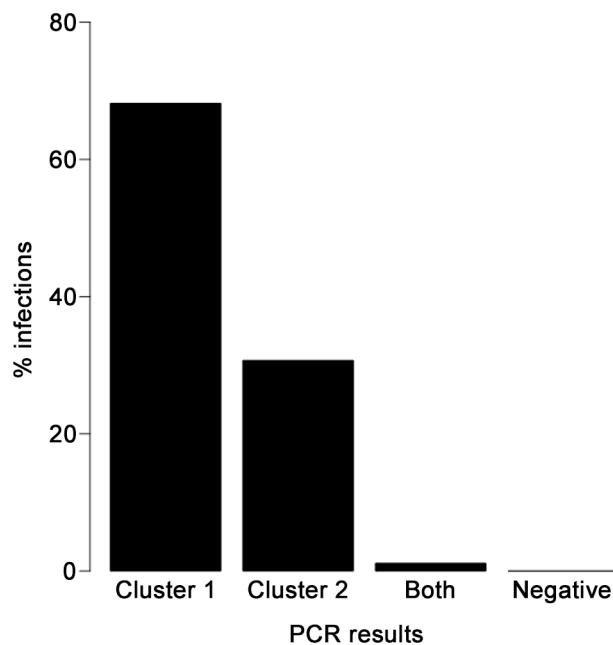
### 3.3.2 Genotyping of *P. knowlesi* infections

Analysis on 340 DNA samples of human cases and 15 from wild macaques sampled in Malaysian Borneo discriminated two subpopulations using the allele-specific primer sets C1A and C2J (Table 3.4). Only four samples were positive with both primer sets (1.1%, Figure 3.3), and these samples were each found to have mixed genotype infections as identified by microsatellite analysis in Chapter Two (Divis et al., 2015). Therefore, the concordance between the marker SNP discrimination and the previous subpopulation cluster assignment was 100% for samples from Malaysian Borneo.

**Table 3.4:** Specificity of allele-specific PCR assays for primer sets C1A and C2J for *P. knowlesi* Cluster 1 and Cluster 2 subpopulations, respectively.

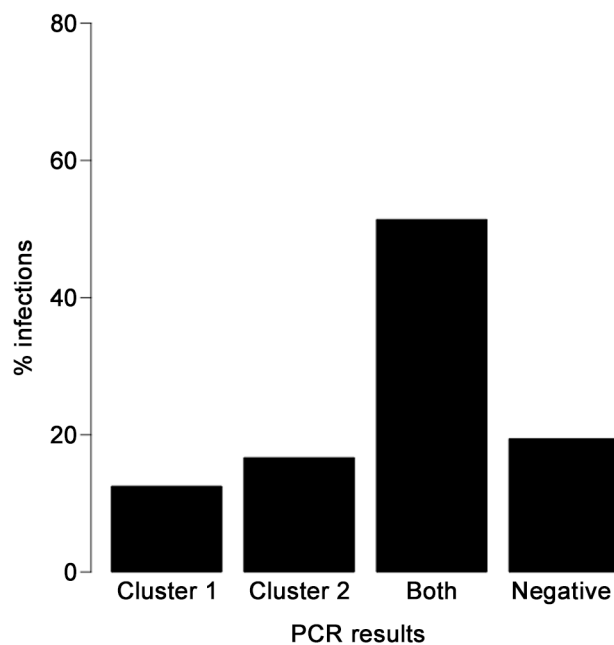
Region	Location	Host	n	Primer sets			Neg
				C1A	C2J	Both	
Malaysian Borneo	Betong	Human	29	21	7	1	0
Malaysian Borneo	Kanowit	Human	34	14	19	1	0
Malaysian Borneo	Miri	Human	46	15	31	0	0
Malaysian Borneo	Sarikei	Human	23	14	9	0	0
Malaysian Borneo	Kapit	Human	52	27	25	0	0
Malaysian Borneo	Kudat	Human	46	46	0	0	0
Malaysian Borneo	Ranau	Human	62	53	9	0	0
Malaysian Borneo	Tenom	Human	48	43	4	1	0
Malaysian Borneo	Kapit	Long-tailed macaque	10	9	0	1	0
Malaysian Borneo	Kapit	Pig-tailed macaque	5	0	5	0	0
Peninsular Malaysia	Kelantan	Human	15	1	4	3	7
Peninsular Malaysia	Pahang	Human	42	7	5	24	6
Peninsular Malaysia	Selangor	Long-tailed macaque	15	1	3	10	1

All DNA samples were blinded with regards to their cluster assignments deduced from microsatellite (Divis et al., 2015) and whole genome sequencing (Assefa et al., 2015) analyses. Abbreviation: Neg – PCR negative for both primer sets.



**Figure 3.3.** Proportions of isolates positive for the new PCR discrimination of Cluster 1 and Cluster 2 subpopulations for *P. knowlesi* infections in 355 infections from Malaysian Borneo.

In contrast, applying the assays to infections from 57 humans and 15 macaques from Peninsular Malaysia showed a high proportion of samples double positive for both alleles (51.4%; Figure 3.4; Table 3.4), and also a substantial proportion (19.4%) of samples that were negative. Therefore, this single-locus genotyping assay does not separate two discrete types among samples from Peninsular Malaysia.



**Figure 3.4.** Proportions of isolates positive for the new PCR discrimination of *P. knowlesi* considered to be 'Cluster 1' and 'Cluster 2' subpopulations in 62 infections from Peninsular Malaysia.

### 3.4 Discussion

This chapter reports development of a simple and rapid genotyping method for the discrimination of two sympatric *P. knowlesi* subpopulations in clinical isolates that are associated with different macaque host species in Borneo. The PCR genotyping assay has been validated for its specificity so that it can be used in future studies.

However, it did not reliably discriminate types in samples from Peninsular Malaysia, indicating that the population genetic structure of *P. knowlesi* needs further investigation in this area.

The specificity of PCR assays (primer sets C1A and C2J) for discriminating the two divergent Cluster 1 and Cluster 2 subpopulations appears complete for analysis of *P. knowlesi* infections in Malaysian Borneo. In order to discriminate parasites from Peninsular Malaysia more efficiently, further genotyping analysis is required to see whether they separate into clusters 1 and 2 (as in Borneo, as indicated by the results of Chapter 2), or if they comprise a separate cluster, such as the 'Cluster 3' identified in old laboratory isolates (Assefa et al., 2015).

While identification of *P. knowlesi* parasites in human and macaque samples by PCR has already been widely performed (Singh et al., 2004, Lee et al., 2011, Singh and Daneshvar, 2013), this relatively simple new genotyping tool should help in studying the prevalence of *P. knowlesi* clinical infections associated with different macaque host species. Using this tool, two divergent *P. knowlesi* parasite populations in Malaysian Borneo can be simply discriminated and further investigation on the clinical presentations and biological characterisation of the parasites can be performed without first needing multi-locus microsatellite genotyping or whole genome

sequencing on all samples. This also points to the need to further sample and characterise the parasite population genetic structure in Peninsular Malaysia, for which the separation into two types was not evident with the single-locus genotyping assay applied here.



# Chapter Four

**Registry**  
 T: +44(0)20 7299 4646  
 F: +44(0)20 7299 4656  
 E: registry@lshtm.ac.uk

**RESEARCH PAPER COVER SHEET**

**PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.**

**SECTION A – Student Details**

<b>Student</b>	Paul Cliff Simon Divis
<b>Principal Supervisor</b>	Professor David Conway
<b>Thesis title</b>	Population Genetic Structure and Genomic Divergence in <i>Plasmodium knowlesi</i>

***If the Research Paper has previously been published please complete Section B, if not please move to Section C***

**SECTION B – Paper already published**

Where was the work published?	Emerging Infectious Diseases		
When was the work published?	April 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*\*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

**SECTION D – Multi-authored work**

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper (Attach a further sheet if necessary)	I designed the study, performed laboratory experiments, conducted statistical analyses and wrote the manuscript.
---	--

**Student Signature:**  **Date:** 13<sup>th</sup> March 2017

**Supervisor Signature:**  **Date:** 20<sup>th</sup> March 2017

## Three major divergent subpopulations of the malaria parasite *Plasmodium knowlesi*

### 4.1 Introduction

*Plasmodium knowlesi* is transmitted by mosquitoes to humans from monkey reservoir hosts, with different *Anopheles* species of the *leucophyrus* group having been incriminated as potential vectors in different areas (Singh and Daneshvar, 2013, Vythilingam et al., 2016). Two macaque species (the long-tailed macaque *Macaca fascicularis*, and the pig-tailed macaque *M. nemestrina*) are the major reservoirs of infection (Lee et al., 2011, Vythilingam et al., 2008), and human infections in Malaysian Borneo have divergent genetic subpopulations that are respectively seen in the different-macaque species locally, indicating that there may be two independent zoonoses occurring sympatrically (Divis et al., 2015).

Significant geographical differentiation of parasites between Malaysian Borneo and Peninsular Malaysia was also evident in the microsatellite analysis, and separate studies have revealed divergence between the two regions at unlinked genes encoding the normocyte binding protein (Ahmed et al., 2016, Ahmed et al., 2014, Pinheiro et al., 2015), and the Duffy binding protein (Fong et al., 2014, Fong et al., 2015), as well as the 18S rRNA and mitochondrial cytochrome oxidase subunit 1 (Yusof et al., 2016). Whole genome sequencing has confirmed the presence of two divergent subpopulations of *P. knowlesi* in Malaysian Borneo and revealed a third divergent cluster of laboratory isolates maintained in laboratories since in the 1960s, most of

which were recorded to have originated originally from Peninsular Malaysia (Assefa et al., 2015).

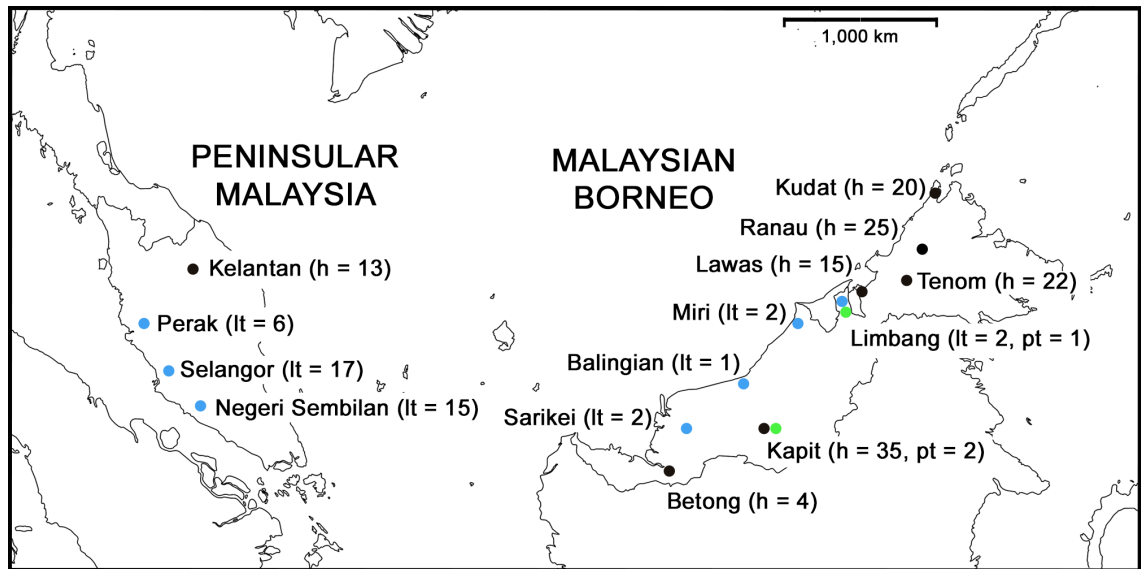
To resolve the population structure in relation to host species and geography, a new collection of 182 *P. knowlesi* infection samples from humans and wild macaques living in diverse areas of Malaysia was genotyped at ten microsatellite loci. Analysis was first conducted separately for the new data set, and then for a combined data set incorporating previous multi-locus microsatellite data (Chapter 2; (Divis et al., 2015), utilising several independent and complementary statistical approaches to identify genetic substructure.

All analyses revealed that two divergent genetic subpopulations of human cases occur sympatrically in Malaysian Borneo, which are respectively detected in long-tailed macaques and pig-tailed macaques in the same region, while a third divergent genetic subpopulation occurs in humans and most macaques in Peninsular Malaysia. This parasite species has undergone different sympatric and allopatric processes of divergence, which will affect its future adaptation to a changing environmental landscape, while current differences between the subpopulations need to be recognized in clinical and epidemiological studies.

## **4.2 Materials and methods**

### **4.2.1 Study sites and DNA samples**

*P. knowlesi* DNA samples were obtained from human clinical cases at seven sites, and macaque hosts at eight sites, across Malaysia (Figure 4.1). DNA was extracted from anticoagulated venous blood samples or dried blood spots, and tested for the



**Figure 4.1.** DNA samples of *P. knowlesi* infections derived from 134 humans and 48 macaques across Malaysia.

Samples obtained from humans are marked with black dots and labeled with 'h', long-tailed macaques with blue dots and labeled with 'lt', and pig-tailed macaques with green dots and labeled with 'pt'. These samples are independent of those studied previously in Chapter 2 (Divis et al., 2015).

presence of different malaria parasite species by species-specific PCR using methods previously described (Lee et al., 2011). Samples from 134 *P. knowlesi* positive human cases collected between 2012 and 2014, that had sufficient DNA for multi-locus genotyping, originated from Kapit (n = 35), Betong (n = 4) and Lawas (n = 15) in Sarawak state, from Kudat (n = 20), Ranau (n = 25) and Tenom (n = 22) in Sabah state, and from Kelantan (n = 13) in Peninsular Malaysia.

Samples from 48 *P. knowlesi*-positive macaques had sufficient DNA for multi-locus genotyping, which were collected between 2007 and 2014. Most were from long-tailed macaques, sampled from Selangor (n = 17), Perak (n = 6) and Negeri Sembilan (n = 15) in Peninsular Malaysia, and from Balingian (n = 1), Limbang (n = 2), Miri (n = 2) and

Sarikei (n = 2) in Sarawak, while pig-tailed macaque samples were from Limbang (n = 1) and Kapit (n = 2) in Sarawak. Sampling was performed according to the protocols of the Department of Wildlife and National Parks in Malaysia. DNA of *P. knowlesi* strain Nuri (kindly provided by Clemens Kocken at the Biomedical Primate Research Centre, The Netherlands) was also included in the genotyping as a control (Kocken et al., 2002).

#### **4.2.2 Microsatellite genotyping of new samples**

Each of the *P. knowlesi* positive DNA samples was genotyped at ten microsatellite loci (NC03\_2, CD05\_06, CD08\_61, NC09\_1, NC10\_1, CD11\_157, NC12\_2, NC12\_4, CD13\_61, CD13\_107) using hemi-nested PCR assays specific for *P. knowlesi* as previously described (Divis et al., 2015). Fluorescent dye-labelled PCR products were analysed by capillary electrophoresis on a Genetic Analyzer 3730 (Applied Biosystems, UK), using GeneScan 500 LIZ internal size standards, following which scoring of alleles and peak heights were done using GeneMapper version 4.0 software (Applied Biosystems, UK).

The genotypic multiplicity of each infection (MOI) was defined as the maximum number alleles detected at any individual locus. Electrophoretic peak heights above 200 fluorescent units of the expected molecular sizes were scored as alleles, and secondary peaks within an infection sample were scored if they have a height of at least 25% relative to the predominant allele. The multi-locus genotype profile of each infection, and allele frequency counts for population samples, were determined by counting the predominant allele at each locus within each infection.

#### **4.2.3 Analysis of microsatellite genotypes from previous data**

Whole genome sequence data of *P. knowlesi* samples were retrieved from the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>), and the reference genome sequence of strain H was obtained from GeneDB (<http://www.genedb.org/Homepage/Pknowlesi>). Most of the parasite genome short-read Illumina sequences available are from patients sampled in Malaysian Borneo (Assefa et al., 2015, Pinheiro et al., 2015), but a few are from older laboratory lines that originated from Peninsular Malaysia as well as one supposedly from the Philippines (Assefa et al., 2015). Although genome sequences indicate some historical mislabelling or contamination of the laboratory lines, meaning that individual identities are in question, it is clear that the majority are from Peninsular Malaysia (Assefa et al., 2015).

The raw short reads were aligned to *P. knowlesi* genome strain H using the BWA-MEM alignment tool with default parameters (Li, 2013). Lists of indels were identified using the SAMTools and VCFtools software (Li et al., 2009a, Danecek et al., 2011) with the following parameters: `mpileup -B -Q 23 -d 2000 -C 50 -ugf; varFilter -d 10 -D 2000` which was described elsewhere (Assefa et al., 2015). Using the ARTEMIS software (Rutherford et al., 2000), the putative microsatellite allele size was determined by inspecting the indels within the location of the PCR primers used for the second amplification PCR. The quality of the mapping within the microsatellite allele regions was assessed with the minimum depth of short read coverage at 30-fold.

#### **4.2.4 Analyses of *P. knowlesi* population genetic substructure**

Population genetic structure was evaluated by Bayesian clustering inference using the STRUCTURE version 2.3.4 software (Pritchard et al., 2000), on samples for which there

were no missing data at any locus. First, in order to allocate the probable ancestral assignment of a genotype into one or more of  $K$  clusters, the parameters were set for the admixture model on the basis of correlated allele frequency, without providing the sample source information. However, the sensitivity for population structure analysis can be improved by providing population information, in which an algorithm assumes that the probability of an individual being part of a population varies among locations or sources of origins (Hubisz et al., 2009). For this second test, the parameter was set to LOCPRIOR, which is informative for weak population structure signals, and it has the freedom to not detect any structure when there is none.

All STRUCTURE runs were performed with a burn-in period of 50,000 followed by 100,000 Markov chains (MCMC iterations). The simulations were replicated 20 times for  $K$  values ranging from 1 to 10. The optimal  $K$  value was calculated based on Evanno's method of  $\Delta K$  statistics implemented in the STRUCTURE HARVESTER webpage interface (Earl and vonHoldt, 2012, Evanno et al., 2005). For the optimum  $K$ , the 20-replicate runs were aligned at 10,000 permutations to determine the consensus of cluster scores using the CLUMPP version 1.1.2 (Jakobsson and Rosenberg, 2007).

To evaluate population structure independently, principal coordinate analysis (PCoA) was performed using the GenAlEx packages version 6 implemented in Microsoft Excel (Peakall and Smouse, 2006). A genetic distance matrix was first generated using the multi-locus microsatellite dataset, and a two-dimensional PCoA was plotted based on the first two highest eigenvalues. The  $K$ -means clusters were calculated using the first and second eigenvectors generated from the PCoA, and subsequently used to assign each individual infection to the most probable cluster. In addition, the discriminant



analysis of principal component (DAPC) from the *adegenet* 2.0.0 packages in R was applied to assess the population structure (Jombart et al., 2010). In this procedure, genotype data were first transformed into 40 uncorrelated principal components, and then variances were partitioned using the discriminant function into within-group and among-group components, while optimizing separations between groups.

Pairwise differentiation ( $F_{ST}$ ) between different subpopulations of *P. knowlesi* was calculated using the FSTAT software version 2.9.3.2 (Goudet, 1995).

### **4.3 Results**

#### **4.3.1 Genotypic diversity within *P. knowlesi* infections**

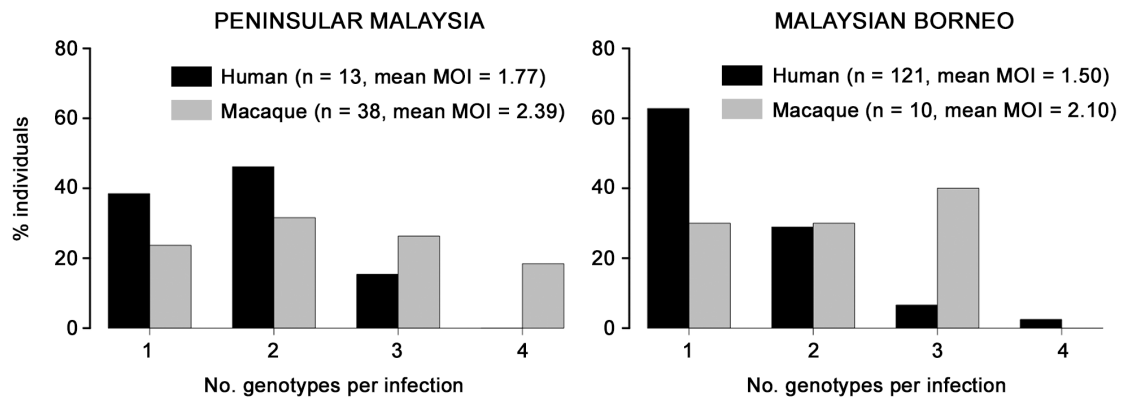
Out of 182 *P. knowlesi* infections genotyped for this study (134 from humans, 45 from long-tailed macaques, and 3 from pig-tailed macaques), 166 (91.2%) yielded complete genotype data for the panel of ten microsatellite loci, while the remainder were each genotyped for at least seven of the loci (Table 4.1; Appendix 4.1).

Among the human cases, single genotype infections were common, and the average number of genotypes per infection (multiplicity of infection, MOI) was less than two at all sites sampled. There was no significant difference in numbers of genotypes per infection in Malaysian Borneo and Peninsular Malaysia (mean MOI = 1.50 and MOI = 1.77 respectively, Fisher's Exact  $P = 0.14$ ). In contrast, multiple genotype infections were more common in macaques both in Malaysian Borneo (mean MOI = 2.10,  $P = 6.7 \times 10^{-3}$ ) and Peninsular Malaysia (mean MOI = 2.39,  $P = 9.8 \times 10^{-4}$ ) (Table 4.1; Figure 4.2). The predominant allele at each locus per infection was counted for subsequent statistical analyses on population structure.

**Table 4.1.** Summary of *P. knowlesi* mixed genotype infections in 134 human and 48 macaque hosts across Malaysia using 10 microsatellite loci.

Host and Site	Region	N	No. isolates with the following no. genotypes detected:				%poly	MOI	MS <sub>10</sub>
			1	2	3	4			
<b>HUMAN</b>									
Kapit	Sarawak	35	27	5	2	1	23	1.34	35
Betong	Sarawak	4	4	0	0	0	0	1.00	3
Lawas	Sarawak	15	7	7	0	1	53	1.67	14
Kudat	Sabah	20	13	6	1	0	35	1.40	20
Ranau	Sabah	25	13	10	2	0	48	1.56	25
Tenom	Sabah	22	11	7	3	1	50	1.73	22
Kelantan	Peninsular Malaysia	13	5	6	2	0	62	1.77	13
Total		134							132
<b>LONG-TAILED MACAQUE</b>									
Balingian	Sarawak	1	0	1	0	0	100	2.00	1
Limbang	Sarawak	2	0	1	1	0	100	2.50	1
Miri	Sarawak	2	1	1	0	0	50	1.50	1
Sarikei	Sarawak	2	1	0	1	0	50	2.00	2
Selangor	Peninsular Malaysia	17	8	6	2	1	53	1.76	15
Perak	Peninsular Malaysia	6	1	3	2	0	83	2.17	5
Negeri Sembilan	Peninsular Malaysia	15	0	3	6	6	100	3.20	6
Total		45							31
<b>PIG-TAILED MACAQUE</b>									
Limbang	Sarawak	1	0	0	1	0	100	3.00	1
Kapit	Sarawak	2	1	0	1	0	50	2.00	2
Total		3							3
Total all		182							166

All samples (N) were successfully genotyped at  $\geq 7$  loci, and 166 samples had complete genotypes for all 10 microsatellite loci (MS<sub>10</sub>). The percentage of polyclonal infections (%poly) and average genotypic multiplicity of infections (MOI) are shown.

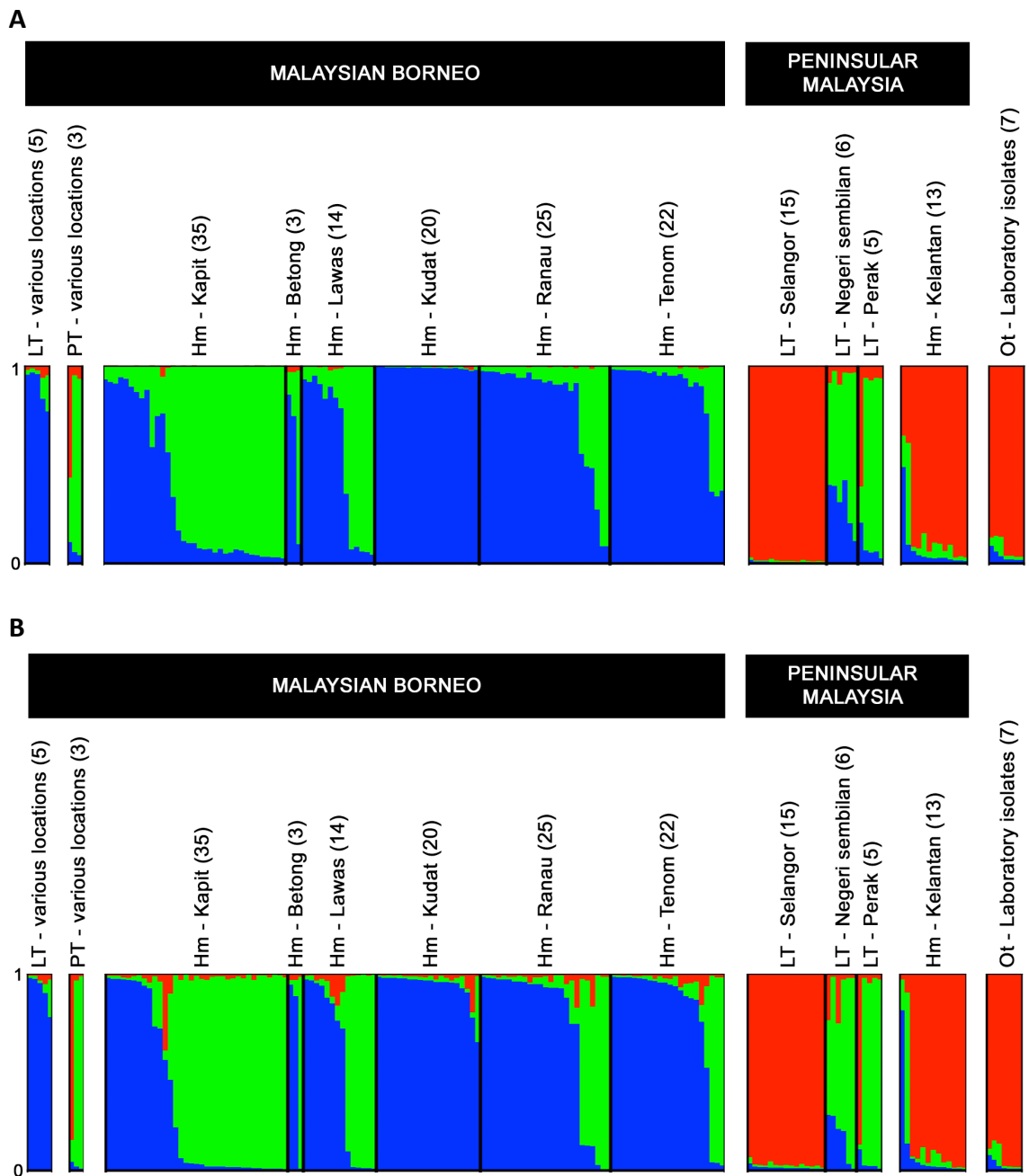


**Figure 4.2.** Multiplicity of infections (MOI) in 134 human and 48 macaque hosts across Malaysia.

Means of MOI were higher in macaque hosts than in human hosts for both regions, but the values were not significant for Peninsular Malaysia (Fisher's Exact  $P = 0.25$ ) compared to Malaysian Borneo ( $P = 0.01$ ).

#### 4.3.2 Analysis of *P. knowlesi* population genetic structure with new samples

Bayesian clustering analyses using two admixture models on the new sample of 166 infections with complete genotype data for the full panel of ten microsatellite loci identified three subpopulation clusters ( $K = 3$ ; Figure 4.3A with  $\Delta K = 37.72$  for the LOCPRIOR model; Figure 4.3B with  $\Delta K = 128.51$  for the non-LOCPRIOR model; Appendix 4.2), hereafter referred to Cluster 1 to 3. Human infections in Malaysian Borneo were assigned to Clusters 1 and 2, while long-tailed macaque infections were all Cluster 1 and pig-tailed macaque infections were mostly Cluster 2 (one pig-tailed macaque infection was assigned as intermediate between Clusters 2 and 3), confirming the existence of two major sympatric subpopulations in Malaysian Borneo, as reported previously (Divis et al., 2015, Pinheiro et al., 2015, Assefa et al., 2015).



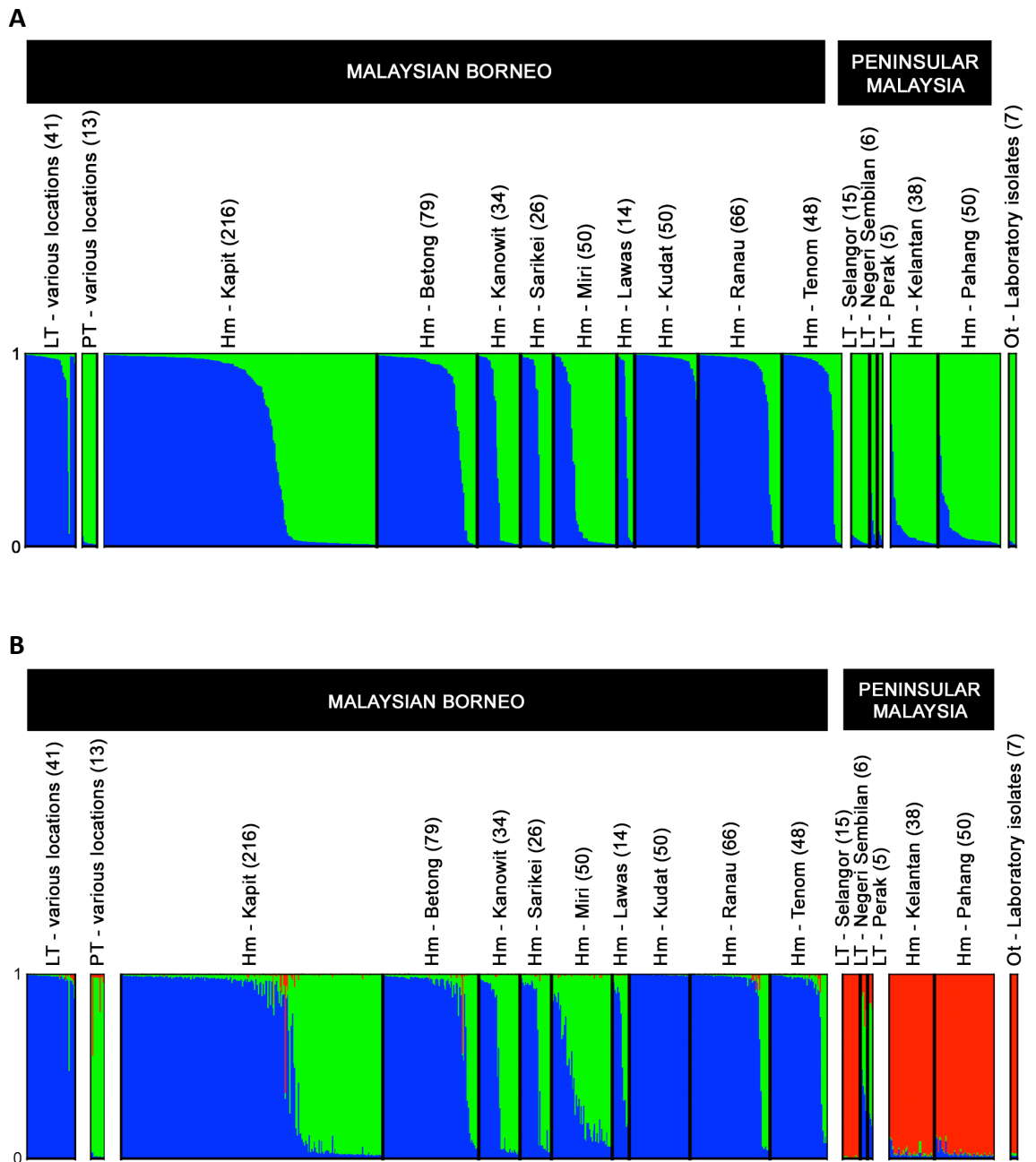
**Figure 4.3.** Subpopulation cluster assignments of 166 individual *P. knowlesi* infections in human and macaque hosts across Malaysia and seven laboratory isolates. The Bayesian-based STRUCTURE analysis (A) with the LOCPRIOR model ( $K = 3$ ;  $\Delta K = 37.72$ ), and (B) without the LOCPRIOR model ( $K = 3$ ;  $\Delta K = 128.51$ ). Ancestral population clusters are referred as Cluster 1 (blue), Cluster 2 (green) and Cluster 3 (red). Abbreviation: LT – long-tailed macaque, PT – pig-tailed macaque, Hm – human, and Ot – various sources

Among the samples from Peninsular Malaysia, those from human cases were all assigned to Cluster 3, along with most of the infections from wild long-tailed macaques sampled in Kelantan, although long-tailed macaque infections from the other two sites had more intermediate cluster assignments suggesting some ancestral affinity with Cluster 2. All laboratory isolates, originating many years ago mainly from Peninsular Malaysia, were clearly assigned to Cluster 3, consistent with results of a recent whole genome sequence analysis (Assefa et al., 2015).

#### **4.3.3 Analysis of population genetic structure incorporating new and previously acquired microsatellite data**

To further evaluate the population structure of *P. knowlesi*, the dataset in this study was collated together with data from samples previously analysed (Chapter 2, (Divis et al., 2015)). This yielded a total of 758 *P. knowlesi* infections with the complete panel of ten microsatellite loci genotyped. This comprised of 166 from the present study, 556 were previously genotyped including 29 that had undergone repeat genotyping for all 10 loci completed here (Appendix 4.1), and seven were derived from Illumina short-read sequence data.

The admixture STRUCTURE analysis without the LOCPRIOR model identified two subpopulation clusters ( $K = 2$ ,  $\Delta K = 255.50$ ; Figure 4.4A; Appendix Figure 4.3). This was consistent with a previous analysis showing that human cases in Malaysian Borneo group into two different genotype clusters, which are also respectively seen in long-tailed and pig-tailed macaque infections, although the current analysis assigned samples from Peninsular Malaysia to Cluster 2 (previously they had been grouped into



**Figure 4.4.** STRUCTURE analysis on 758 *P. knowlesi* genotypes using 10 microsatellite loci.

The Bayesian-based STRUCTURE analysis (A) without the LOCPRIOR model ( $K = 2$ ;  $\Delta K = 255.50$ ), and (B) with the LOCPRIOR model ( $K = 3$ ;  $\Delta K = 98.73$ ). Ancestral population clusters are referred as Cluster 1 (blue), Cluster 2 (green) and Cluster 3 (red).

Abbreviation: LT – long-tailed macaque, PT – pig-tailed macaque, Hm – human, and Ot – various sources.

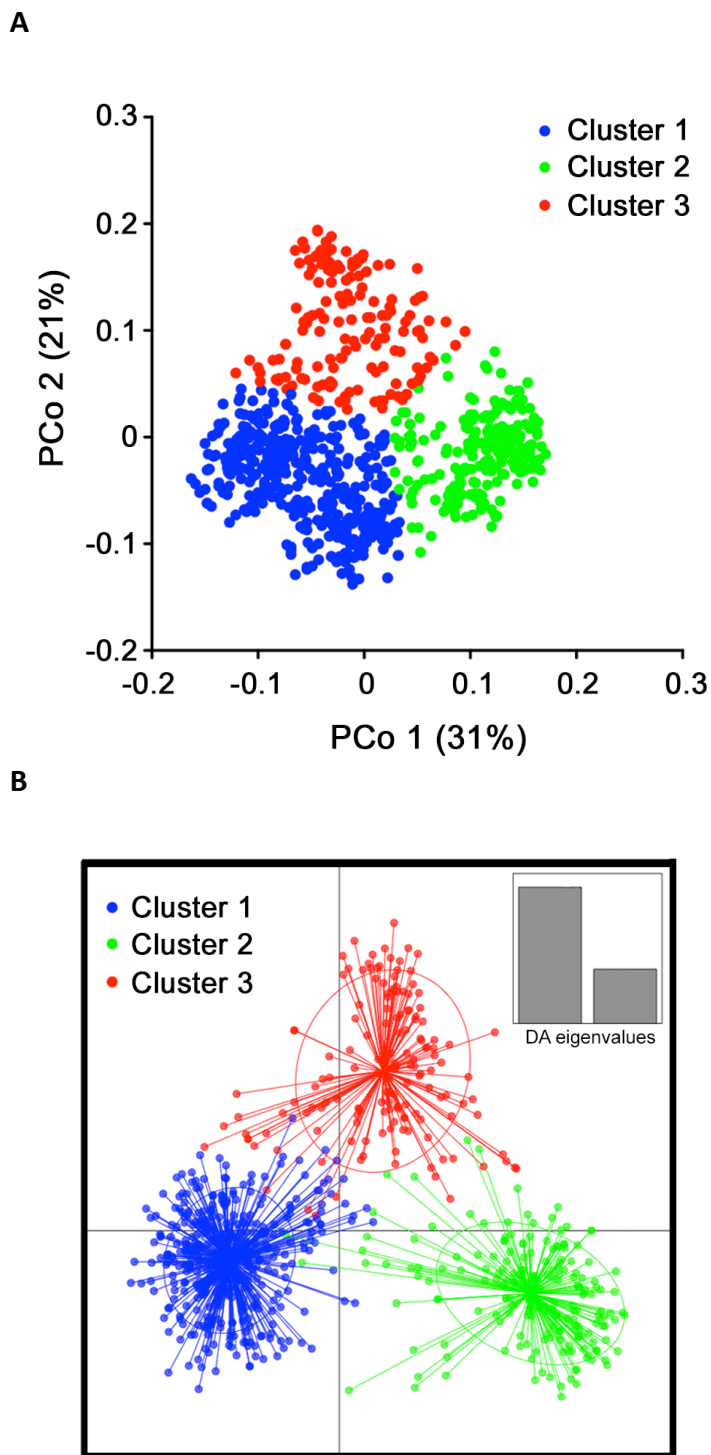
Cluster 1). However, incorporation of the LOCPRIOR model showed three subpopulation clusters ( $K = 3$ ,  $\Delta K = 98.73$ ; Figure 4.4B; Appendix 4.3), with most of the isolates from Peninsular Malaysia belonging to Cluster 3, as also seen with the analysis based solely on the new samples.

Overall, this confirms that human *P. knowlesi* infections in Malaysian Borneo are divided into two different genetic subpopulations that are associated with different macaque reservoir host species, whereas human infections in Peninsular Malaysia belong to a third subpopulation which is also seen in long-tailed macaques at one of the sites in Peninsular Malaysia.

#### **4.3.4 Robustness and divergence of subpopulation clusters**

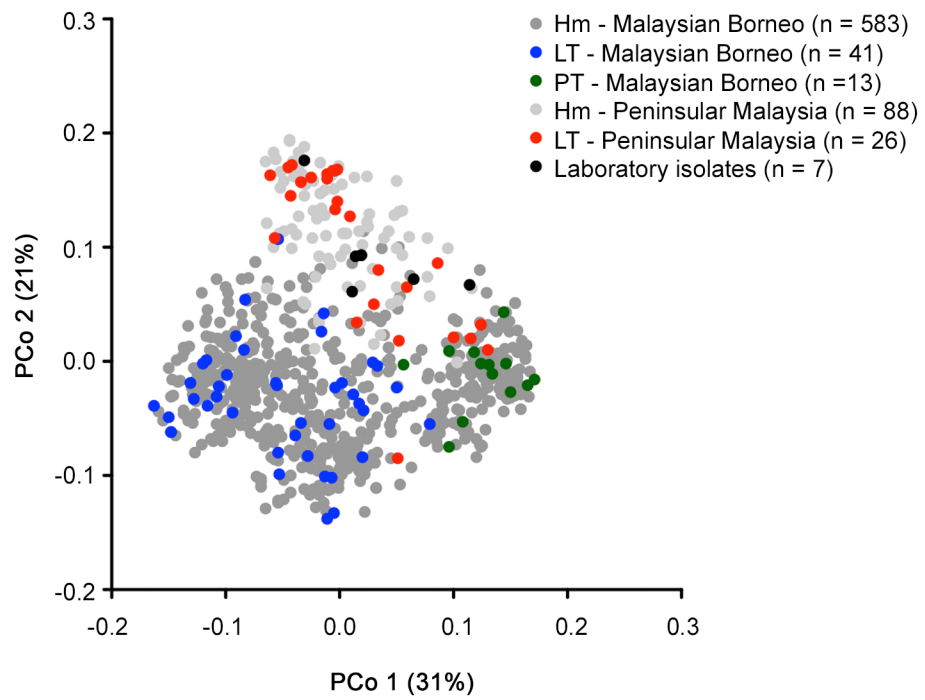
Using an a priori designation of there being three subpopulation clusters ( $K = 3$ ), all 758 infections were independently assigned into clusters using principal coordinate analysis (Figure 4.5A) and discriminant analysis (Figure 4.5B), and the results were compared with those derived from the STRUCTURE analysis (Figure 4.4B). These showed highly concordant results (Appendix 4.1). Principal coordinate analysis indicated that infections in humans were strongly associated with infections in local macaque reservoir hosts for both Malaysian Borneo and Peninsular Malaysia (Figure 4.6). Discriminant analysis also showed clear clustering, with only minimal overlap among the inertia ellipses for the three major clusters.

To test the consistency and robustness of cluster assignment for all 758 infections, across the different methods used (Bayesian analysis using STRUCTURE, principal coordinates analysis, and discriminant analysis), a consensus for each individual was



**Figure 4.5.** Population genetic structure of combined 751 *P. knowlesi* infections across Malaysia and seven laboratory isolates. Using a priori of  $K = 3$  from the STRUCTURE analysis with LOCPRIOR model (Figure 4.4B), individual genotypes were assigned to the most probable subpopulation clusters using independent (A) genetic distance matrix inferred by the principal coordinate analysis and (B) discriminant analysis of principal component (DAPC). In DAPC, clusters depicted as ellipses indicated the variance within the clusters and centred by  $K$ -means.





**Figure 4.6.** Principal coordinate analysis deduced from genetic distance matrix of 10 microsatellite loci in 751 *P. knowlesi* infections across Malaysia and seven laboratory isolates.

Abbreviation: Hm – human; LT – long-tailed macaque; PT – pig-tailed macaque.

**Table 4.2.** Assignment of combined 753 *P. knowlesi* genotypes into three subpopulation clusters determined by minimum of two out of three assignment methods.

<i>P. knowlesi</i> population	Subpopulation cluster	DAPC ∧ PCoA		DAPC ∧ PCoA		Total isolate
		∧ LOC	DAPC ∧ PCoA	∧ LOC	∧ LOC	
LT-various locations in Sarawak	Cluster 1	32	0	4	3	39
	Cluster 2	1	0	0	0	1
	Cluster 3	0	0	0	0	0
PT-various locations in Sarawak	Cluster 1	0	0	0	0	0
	Cluster 2	12	0	0	1	13
	Cluster 3	0	0	0	0	0
Hm-Kapit	Cluster 1	114	0	20	4	138
	Cluster 2	70	1	1	2	74
	Cluster 3	0	0	1	0	1
Hm-Betong	Cluster 1	57	0	10	0	67
	Cluster 2	10	2	0	0	12
	Cluster 3	0	0	0	0	0
Hm-Kanowit	Cluster 1	13	0	3	0	16
	Cluster 2	18	0	0	0	18
	Cluster 3	0	0	0	0	0
Hm-Sarikei	Cluster 1	14	0	0	1	15
	Cluster 2	11	0	0	0	11
	Cluster 3	0	0	0	0	0
Hm-Miri	Cluster 1	16	1	0	0	17
	Cluster 2	28	0	5	0	33
	Cluster 3	0	0	0	0	0
Hm-Lawas	Cluster 1	8	0	0	0	8
	Cluster 2	6	0	0	0	6
	Cluster 3	0	0	0	0	0
Hm-Kudat	Cluster 1	44	0	6	0	50
	Cluster 2	0	0	0	0	0
	Cluster 3	0	0	0	0	0
Hm-Ranau	Cluster 1	49	0	6	2	57
	Cluster 2	8	1	0	0	9
	Cluster 3	0	0	0	0	0
Hm-Tenom	Cluster 1	37	0	5	0	42
	Cluster 2	5	0	1	0	6
	Cluster 3	0	0	0	0	0
LT-Selangor	Cluster 1	0	0	0	0	0
	Cluster 2	0	0	0	0	0
	Cluster 3	14	0	1	0	15

LT-Negeri Sembilan	Cluster 1	0	0	0	0	0
	Cluster 2	2	0	1	1	4
	Cluster 3	0	1	0	0	1
LT-Perak	Cluster 1	0	0	0	0	0
	Cluster 2	4	0	0	0	4
	Cluster 3	1	0	0	0	1
Hm-Kelantan	Cluster 1	0	0	0	0	0
	Cluster 2	0	0	0	0	0
	Cluster 3	33	0	5	0	38
Hm-Pahang	Cluster 1	0	0	0	0	0
	Cluster 2	0	0	0	0	0
	Cluster 3	43	0	5	2	50
Laboratory isolates	Cluster 1	0	0	0	0	0
	Cluster 2	0	0	0	0	0
	Cluster 3	5	0	2	0	7

Five genotypes from long-tailed macaque in Sarawak ( $n = 1$ ), humans in Kapit ( $n = 3$ ) and long-tailed macaque in Negeri Sembilan ( $n = 1$ ) showed inconsistency in cluster assignment methods are not shown in this table. Abbreviation: DAPC – discriminant analysis of principal component; PCoA – principal coordinate analysis based on genetic distance matrix; LOC – STRUCTURE analysis with LOCPRIOR model; LT – long-tailed macaque; PT – pig-tailed macaque; Hm – human.

assessed (Appendix 4.1; Table 4.2). A large majority (86.4%) of infections were assigned into the same cluster by all three methods (Cluster 1,  $n = 384$ ; Cluster 2,  $n = 175$ ; Cluster 3,  $n = 96$ ). Most of the remainder (12.9% of the total) had an agreed assignment for two of the methods (Cluster 1,  $n = 65$ ; Cluster 2,  $n = 16$ ; Cluster 3,  $n = 17$ ), while only 5 (0.7%) showed no agreement across the methods. Omitting the few infections that did not show agreement for two or more methods, this yielded a dataset of 753 *P. knowlesi* infections that grouped into three major subpopulation clusters (Cluster 1,  $n = 449$ ; Cluster 2,  $n = 191$ ; Cluster 3,  $n = 113$ ; Table 4.3).

Analyses of allele frequencies across all ten microsatellite loci confirmed strong genetic differentiation among these clusters ( $F_{ST} = 0.184$  between Clusters 1 and 2;  $F_{ST} = 0.152$  between Clusters 1 and 3;  $F_{ST} = 0.201$  between Clusters 2 and 3;  $P < 3.3 \times 10^{-4}$  for each

**Table 4.3.** Summary of subpopulation cluster assignment on combined 758 *P. knowlesi* genotypes according to host and geographical origins.

Subpopulation cluster	Malaysian Borneo			Peninsular Malaysia		Laboratory isolate	Total isolate
	LT	PT	Hm	LT	Hm		
Cluster 1	39	0	410	0	0	0	449
Cluster 2	1	13	169	8	0	0	191
Cluster 3	0	0	1	17	88	7	113
Unassigned	1	0	3	1	0	0	5

Abbreviation: LT – long-tailed macaque; PT – pig-tailed macaque; Hm – human.

comparison using 3,000 randomised permutations). This indicates deep divergence among the three major parasite subpopulations that infect humans, two of which are sympatric and predominantly associated with different reservoir hosts (long-tailed and pig-tailed macaques in Malaysian Borneo), and one of which is allopatric in a different geographical region (Peninsular Malaysia).

#### 4.4 Discussion

Three major subpopulations of *P. knowlesi* have been demonstrated in natural human infections in Malaysia. These show profound divergence, with pairwise  $F_{ST}$  values of  $\sim 0.2$ , suggesting minimal or no current gene flow between parasites in Malaysian Borneo and Peninsular Malaysia, as well as between parasites in long-tailed and pig-tailed macaque hosts within Malaysian Borneo.

The existence of three divergent clusters was initially indicated from whole genome sequence-based single nucleotide polymorphism analysis of *P. knowlesi* clinical isolates and laboratory lines (Assefa et al., 2015). Whereas two of the clusters of genome

sequences (Clusters 1 and 2) had been seen in clinical infections in Malaysian Borneo, the third (Cluster 3) was only seen in old laboratory lines that were mostly originally isolated from Peninsular Malaysia. Using microsatellite scoring obtained from genome sequences and combined with genotyping of infections from humans and macaques in the current study, it is confirmed that the Cluster 3 subpopulation is widespread in Peninsular Malaysia, and divergent from Clusters 1 and 2 that account for all infections in Malaysian Borneo and apparently a minority of wild macaque infections in Peninsular Malaysia. With smaller numbers of samples, recent studies on sequence diversity in genes encoding the normocyte binding protein (*Pknbp<sub>xa</sub>*) (Ahmed et al., 2016) and the Duffy-binding protein (*PkDBP*) (Putaporntip et al., 2016), as well as the 18S rRNA gene and the mitochondrial *Cox1* gene suggested that parasites in Peninsular Malaysia had probably diverged from those in Malaysian Borneo.

It is likely that allopatric divergence occurred as a result of the ocean barrier between Borneo and mainland Southeast Asia, established at the end of the last ice age ~13,000 years ago, which prevents the movement of wild macaque reservoir hosts (Liedigk et al., 2015). However, one of the old laboratory lines that was recently sequenced is labelled as having originally been isolated from a long-tailed macaque in 'Philippines', and this is clearly assigned to Cluster 3 along with the parasites from Peninsular Malaysia (Assefa et al., 2015), although the islands of the Philippines have never been connected to Peninsular Malaysia or any other part of mainland Southeast Asia (Voris, 2000). Unless there was a historical mislabelling or mix up of parasite material previously, this suggests that wider sampling of *P. knowlesi* in wild macaques will give a more complete understanding of divergence within this zoonotic parasite species (Esselstyn et al., 2004, Meijaard, 2003, Liedigk et al., 2015, Smith et al., 2014).

Similarly, the observation that a minority of *P. knowlesi* parasites in long-tailed macaques from Peninsular Malaysia are assigned to Cluster 2, which has otherwise only been seen in samples from Malaysian Borneo, indicates that additional sampling of macaques from different areas may uncover more features of the parasite population structure.

The sympatric differentiation between Cluster 1 and Cluster 2 parasites in Malaysian Borneo supports the idea that parasite subpopulations are transmitted independently in long-tailed and pig-tailed macaque populations (Muehlenbein et al., 2015, Ziegler et al., 2007). Although pig-tailed macaques occur mostly in forested areas, long-tailed macaques have a broader habitat range in both forested and non-forested areas (Moyes et al., 2016). Due to the absence of parasite samples from pig-tailed macaques in Peninsular Malaysia, it is unknown if there is divergence in *P. knowlesi* between the different macaque host species in this region.

Analysis of genome sequences to derive the frequency distribution of single nucleotide polymorphism alleles indicates that the Cluster 1 subpopulation of *P. knowlesi* has undergone long-term population growth (Assefa et al., 2015). It is unknown whether parasites of Cluster 2 and Cluster 3 subpopulations show similar patterns, as further sampling of genome sequences would be needed to test this.

The observation that most infections in all macaque populations are polyclonal, whereas most human cases contain single parasite genotypes, probably reflects a higher intensity of transmission among macaques than from macaques to humans (Divis et al., 2015). It is not yet known whether there are any significant differences in

the clinical course of infections caused by the three major subpopulations of *P. knowlesi*, and it will be important for this to be investigated in a manner that accounts for any confounding variables between different study sites. In any case, recognition of these divergent subpopulations provides a more accurate basis on which to understand and potentially control the transmission of this zoonosis. Furthermore, obtaining whole genome sequence data from more clinical samples belonging to each of the three major types should enable a more thorough investigation of the genomic divergence, and identify loci at which there are signals of recent adaptation that may relate to differences in virulence or transmission.

# Chapter Five



**Registry**

T: +44(0)20 7299 4646

F: +44(0)20 7299 4656

E: registry@lshtm.ac.uk

**RESEARCH PAPER COVER SHEET**

---

**PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.**

**SECTION A – Student Details**

<b>Student</b>	Paul Cliff Simon Divis
<b>Principal Supervisor</b>	Professor David Conway
<b>Thesis title</b>	Population Genetic Structure and Genomic Divergence in <i>Plasmodium knowlesi</i>

***If the Research Paper has previously been published please complete Section B, if not please move to Section C***

**SECTION C – Prepared for publication, but not yet published**

Where is the work intended to be published?	Molecular Biology and Evolution
Please list the paper's authors in intended authorship order:	Paul CS Divis, Craig W Duffy, Samuel A Assefa, Balbir Singh, David J Conway
Stage of publication	Draft in preparation

**SECTION D – Multi-authored work**

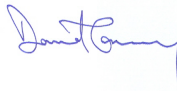
For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper (Attach a further sheet if necessary)	I designed the study, performed laboratory experiments, conducted statistical analyses and draft the manuscript.
---	--

**Student Signature:**



**Date:** 13<sup>th</sup> March 2017

**Supervisor Signature:**



**Date:** 20<sup>th</sup> March 2017

## **Genomic divergence between two sympatric *Plasmodium knowlesi* subpopulations**

### **5.1 Introduction**

The infection of *P. knowlesi* in humans is widespread across the Southeast Asia region (Lubis et al., 2017, Ghinai et al., 2017, Setiadi et al., 2016, Singh and Daneshvar, 2013), even extending to the Nicobar Islands of India (Tyagi et al., 2013). The distribution of human infections corresponds to the sympatric distribution of the natural long-tailed macaque (*Macaca fascicularis*) and pig-tailed macaque (*M. nemestrina*) host, as well as mosquitoes of the *Anopheles leucosphyrus* group as vectors (Vythilingam et al., 2016, Singh and Daneshvar, 2013). Reports of clinical cases have been mainly concentrated in Malaysia (Yusof et al., 2014), particularly Malaysian Borneo (Singh and Daneshvar, 2013), and the two macaque species are incriminated as major reservoirs (Lee et al., 2011).

Comprehensive microsatellite genotyping surveys on *P. knowlesi* infections in humans and macaques across Malaysia recently revealed profound population structure. Two major genetic subpopulations were respectively associated with long-tailed and pig-tailed macaque hosts (Divis et al., 2015), and the divergence between them was further confirmed by whole-genome sequencing analyses (Assefa et al., 2015). The sequence data, and further microsatellite analyses of additional samples, also revealed that three divergent subpopulations of *P. knowlesi* exist (Chapter Four), with two sympatric subpopulations associated with these two macaque species in Borneo

whereas an allopatric subpopulation is identified specifically in mainland Peninsular Malaysia (Divis et al. 2017).

The overall number of clinical infections of the parasite genetic type associated with the long-tailed macaque host (Cluster 1 subpopulation) is higher than those associated with the pig-tailed macaque host (Cluster 2 subpopulation) in a natural situation in Malaysian Borneo (Divis et. al, 2017). Because of this, detailed analyses on population genomics of Cluster 1 subpopulation could be performed first, indicating that this population has undergone long-term population growth with evidence of selection on particular genes (Assefa et al., 2015). There were insufficient data in the initial sequencing study to test whether the Cluster 2 subpopulation had a similar demographic history.

Here, clinical isolates containing parasites of the Cluster 2 subpopulation were identified by screening using the simple allele-specific PCR assays developed in Chapter Three, and this was used to select samples for whole genome sequencing. A new single nucleotide polymorphism (SNP) dataset was obtained in combination with samples sequenced previously. Analyses of this new dataset provide new understanding on the divergence and adaptation of these two sympatric *P. knowlesi* parasites in Malaysian Borneo.

## **5.2 Materials and methods**

### **5.2.1 DNA samples and genotyping assays**

Thirty-five DNA samples received from the Malaria Research Centre, University Malaysia Sarawak (UNIMAS) were obtained from patients infected with *P. knowlesi* in

Kapit Hospital in Sarawak between 2014 and 2015. Initially, human leukocytes were removed by allowing 10 mL of peripheral blood to pass through the CF11 cellulose column, resulting in only parasitised and non-parasitised red blood cells. Genomic DNA was extracted using the QIAamp DNA Mini kit (QIAGEN, Germany) according to the manufacturer's protocol, and was later confirmed to contain only *P. knowlesi* DNA by nested PCR assays to test for all human malaria parasite species (Lee et al., 2011) at University Malaysia Sarawak, UNIMAS. Prior to being sent to the London School of Hygiene and Tropical Medicine, the DNA was preserved dry in the DNASTable® tube (Biomatrica, USA) and stored at room temperature. DNA samples for individual infections were genotyped to determine which subpopulation clusters they belong, using the allele-specific genotyping toolkit as previously described (Chapter 3).

### **5.2.2 DNA library preparation**

DNA libraries were constructed using the TruSeq Nano DNA Library Preparation Kit (Illumina, San Diego, CA, USA) according to manufacturer's instructions. Dried genomic DNA was first resuspended with 30 µl Elution Buffer (QIAGEN, Germany). Physical fragmentation of the genomic DNA was applied using the M220 Focused-ultrasonicator (Covaris, USA) for breaking the DNA into 550 bp. Enrichment of DNA fragments was done with low number of PCR cycles to avoid skewing the representation of the libraries with the following conditions: 95°C for 3 minutes, followed by 8 cycles of 98°C for 20 seconds, 60°C for 15 seconds and 72°C for 30 seconds, and a final 72°C step for 5 minutes. The quality of DNA libraries was assessed by bioanalyzer using the Agilent High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA US) while the quantitation assessment was done using the KAPA Library Quantification Kit for Illumina® platform (KAPA Biosystems, Boston, MA, USA). All

libraries were then normalised to 4 nM, and pooling was performed to include not more than 12 samples on each whole genome sequencing run.

### **5.2.3 Whole genome sequencing and mapping on reference genome**

Paired-end whole genome sequencing was performed on pooled DNA libraries using the MiSeq Chemistry version 3 reagents and run on the MiSeq platform (Illumina, San Diego, CA, USA) with a read length of 300 bp. The quality of raw short read data generated in FASTQ format was filtered using the Trimmomatic software (Bolger et al., 2014) with the following parameters: LEADING:3 TRAILING:3, SLIDINGWINDOW:4:10 MINLEN:36.

Trimmed FASTQ reads for individual isolates were then aligned against the latest version 2.0 of *P. knowlesi* strain H reference genome that was released on GeneDB on March 2014 (accessed at [www.genedb.org/Homepage/Pknowlesi](http://www.genedb.org/Homepage/Pknowlesi) on December 2015) using the Burrows-Wheeler Aligner software version 0.7 with the BWA-MEM algorithm by default parameters (Li, 2013). This generated file in the SAM (sequence alignment/map) format, and followed by the conversion into a BAM (binary alignment/map) format using the SAMtools package version 0.1 (Li et al., 2009a). Due to the possible effect of PCR amplification bias introduced during the DNA library preparations, read duplications were removed using the *MarkDuplicates* command from the java environment Picard toolkit (<https://github.com/broadinstitute/picard>). The summaries of average depth coverage were determined by the BEDTools version 2 package using the *genomeCoverageBed* command (Quinlan and Hall, 2010).

Re-mapping of the short read sequences generated from previous studies (Assefa et al., 2015, Pinheiro et al., 2015) against the version 2.0 of *P. knowlesi* strain H reference genome was also performed in the analysis. These include 48 isolates from Kapit and Betong in Malaysian Borneo (Sequence Read Archive numbers ERR985372 – ERR985419), six isolates from Sarikei in Malaysian Borneo (SRA numbers ERR274221, ERR274222, ERR274224, ERR272225, ERR366425 and ERR366426) and 5 laboratory isolates (“Nuri” SRA numbers ERR019406, “Hackeri” SRR2221468, “Malayan” SRR2225467, “MR4-H” SRR2225571 and “Philippines” SRR2225573).

#### **5.2.4 Single nucleotide polymorphism (SNP) calling and filtration**

The calling of high quality SNPs was performed using several steps, following procedures described previously (Assefa et al., 2015). For each isolate, SNPs were first identified from the BAM file using SAMTools/BCFTools with the following parameters: *mpileup -B -Q 23 -d 2000 -C 50 -ugf; varFilter -d 10 -D 2000*. This would generate a VCF (variant call format) file. A high-quality list of potential variant positions (Phred quality,  $Q > 30$ ) was extracted from this file and a list of unique SNP lists was generated by concatenating all variant positions from all isolates. Using these unique SNP positions, the mapping quality (mq) and base quality (bq) were checked for each isolate to remove positions with an excess of low quality reads with the requirement of the minimum read depth coverage at 10x. The ratio of read depth values at high-quality (mq = 26; bq = 23) and low-quality (mq = 0; bq = 0) thresholds were calculated for each isolate using customized Perl scripts, and any SNP positions with the ratio below 0.5 were discarded.

Further filtration involved the removal of positions that contained ambiguous sequences (represented as a long stretch of unknown nucleotides 'N') in the reference genome. Multigene families of *SICAVars* and *KIR* genes (Pain et al., 2008) and the subtelomeric regions were also filtered out due to the potential of aligning errors which may cause false-positive SNP calls. The subtelomeric regions were determined by visually inspecting the whole genome synteny mapping of *P. knowlesi* with the *P. vivax* homolog using the PlasmoDB GBrowse v2.48 ([plasmodb.org/cg-bin/gbrowse/plasmodb/](http://plasmodb.org/cg-bin/gbrowse/plasmodb/)). The boundaries of subtelomeric regions were defined as sequences before the first conserved protein-coding genes at the left hand end (as oriented in GeneDB) and after the conserved protein-coding genes at the right hand end of each chromosome (Table 5.1). The exclusion of subtelomeric regions and the large multigene families allowed 21.2 Mb (92%) of the 23.0 Mb reference nuclear genome to be analysed.

## **5.2.5 Characterisation of nuclear genome-wide genetic patterns**

### **5.2.5.1 Genomic diversity and population structure**

To measure the amount of polymorphism within a population, the average pairwise nucleotide diversity ( $\pi$ ) was calculated among the individual infection samples. The population demographic history was estimated by the Tajima's *D* statistics. Both statistics were performed using the same genome-wide SNP dataset in non-overlapping window sizes of 10 kb and performed using the DivStat software (Soares et al., 2015). To illustrate the population substructure, the matrix of pairwise DNA distance among individuals was calculated and the Neighbour-Joining tree was constructed using the APE package version 3.4 in the R environment (Paradis et al., 2004). An independent population structure evaluation was also conducted using

**Table 5.1.** Definition of the terminal genes before subtelomeric regions (version 2.0 of *P. knowlesi* H strain).

<b>Chromosome</b>	<b>Left subtelomeric region</b>	<b>Right subtelomeric region</b>
	<b>End position GeneID*</b>	<b>Start position GeneID</b>
PKNH_01_v2	31756 PKNH_0100600	851582 PKNH_0118200
PKNH_02_v2	39903 PKNH_0200500	717562 PKNH_0216200
PKNH_03_v2	40085 PKNH_0300900	984128 PKNH_0321700
PKNH_04_v2	45677 PKNH_0400700	1111994 PKNH_0424800
PKNH_05_v2	40025 PKNH_0500100	735661 PKNH_0516400
PKNH_06_v2	19789 PKNH_0600400	1045508 PKNH_0623600
PKNH_07_v2	9141 PKNH_0700200	1485805 PKNH_0734800
PKNH_08_v2	17009 PKNH_0800400	1874898 PKNH_0841200
PKNH_09_v2	56545 PKNH_0900600	2085215 PKNH_0945900
PKNH_10_v2	69857 PKNH_1001300	1439673 PKNH_1032500
PKNH_11_v2	29067 PKNH_1100500	2317969 PKNH_1149400
PKNH_12_v2	38750 PKNH_1200400	3129934 PKNH_1272100
PKNH_13_v2	24663 PKNH_1300400	2519037 PKNH_1356400
PKNH_14_v2	46265 PKNH_1401100	3204808 PKNH_1472800

\*GeneIDs mark the first and the last conserved protein-coding genes at the ends of each chromosome.



principal coordinate analysis (PCoA) with SNPs having no missing data, using the APE package.

To estimate the divergence between the subpopulations, the genome-wide fixation index ( $F_{ST}$ ) between the two-subpopulation clusters was computed with SNPs at minor allele frequency (MAF) above 0.1, using customised R functions. Elevated  $F_{ST}$  threshold was estimated at the 90th percentiles of the  $F_{ST}$  distributions for all SNPs. Average  $F_{ST}$  values were calculated in windows of 500 SNPs with sliding by 250 SNPs. The  $F_{ST}$  values for each window were tested for high- or low- differentiated regions against the genome-wide mean  $F_{ST}$  value.

#### **5.2.5.2 Defining genomic regions**

Genomic regions were determined empirically by examining the  $F_{ST}$  distribution across the genome at two different MAFs (MAF above 0.1 and 0.3). For each MAF analysis, average  $F_{ST}$  values were calculated in windows of 200 SNPs (sliding by 100 SNPs), 500 SNPs (sliding by 250 SNPs) and 1000 SNPs (sliding by 500 SNPs). Mean global  $F_{ST}$  values and window  $F_{ST}$  values were then converted into standard z-scores in order to standardise the definition of outlier windows for different parameters. Regions of high- or low- $F_{ST}$  windows were observed and compared among the analyses that used different MAF parameters.

Genomic regions were categorised into three main groups: low divergence region (LDR), intermediate divergence region (IDR), and high divergence region (HDR), in which the determination of these regions was tested at different z-score thresholds. To determine the size of these regions in detail, adjacent outlier windows were

merged to form larger adjoining regions. Peak and trough patterns of window z-scores around the thresholds were taken into consideration in determining the range of genomic regions. Each candidate region was demarcated by first and last SNPs that fell within the merged windows, except for HDRs where SNPs with elevated  $F_{ST}$  values were used as starting and end points.

### **5.2.5.3 Diversity and signature of selection in genomic regions**

Patterns of polymorphisms (nucleotide diversity summarised by  $\pi$  and allele frequency spectrum summarised by Tajima's  $D$ ) in all genomic regions were evaluated using DivStat software. Test runs were performed in non-overlapping window sizes of 10-kb for each subpopulations. Nonparametric Kruskal-Wallis tests were used to test for differences among the genomic regions as well as against the genome-wide background.

### **5.2.6 Extra-chromosomal genomes**

Population structure and phylogeny of the sympatric *P. knowlesi* subpopulations were further analysed using the extranuclear DNA, consisting of the genomes of mitochondria and plastid-like apicoplast. The 5.9-kb mitochondrial DNA sequences were obtained from the present whole genome sequence data and previously published sequences (Jongwutiwes et al., 2005, Lee et al., 2011, Pinheiro et al., 2015, Assefa et al., 2015). Complete mitochondrial sequences were obtained from Genbank database, consisting of 26 haplotypes from human isolates (accession numbers EU880446 – EU880470) and 20 haplotypes from macaque isolates (EU880471 – EU880474, EU880477 – EU880486, EU880489 – EU880493 and EU880499) in Kapit of Malaysian Borneo, and one human isolate from Thailand (AY598141). Three species, *P.*

*coatneyi* (AB354575), *P. cynomolgi* (AB434919) and *P. vivax* (AY791551), that have evolutionary relationships with *P. knowlesi* were included in the analysis as outgroups. Each DNA sequence was manually checked for the correct orientation due to the circular form of the genome.

For the apicoplast genome of *P. knowlesi*, limited study has been done with regards to the complete DNA sequences. Therefore, 30.6-kb of the DNA sequences that had clear alignment were solely extracted from the present whole genome dataset as well as from previous data (Assefa et al., 2015, Pinheiro et al., 2015) following mapping and base quality checks as mentioned above.

Extra-chromosomal genome sequences were aligned separately for mitochondrial and apicoplast genomes using the ClustalX programme version 2 (Larkin et al., 2007). DNA sequence alignments were analysed for nucleotide diversity ( $\pi$ ) and haplotype diversity ( $Hd$ ) using the DnaSP version 5 software (Librado and Rozas, 2009). A maximum likelihood tree was inferred with 1,000 bootstrap replicates and gaps treated as missing data using the *phangorn* packages in R (Schliep, 2011). The ModelTest algorithm implemented in the *phangorn* packages was used to determine the best-fit nucleotide substitution model, which was the GTR+I+G model (General Time Reversible model with a proportion of invariable sites and gamma distribution). For the mitochondrial sequences, major haplotypes were determined with gaps treated as missing data, and the statistical parsimony haplotype network was constructed using the TCS version 1.21 software (Clement et al., 2000).

### 5.3 Results

### 5.3.1 Generation of new whole-genome sequence data, assembly and SNP calling

Out of 35 new *P. knowlesi* infections screened here, the initial PCR discrimination assays identified 11 infections as Cluster 1 and 18 infections as Cluster 2, whereas six infections showed positive for both PCR assays. Results were compared to microsatellite genotyping analysis described previously (Chapter 4; Divis et al., 2017). These confirmed that the 18 single Cluster 2 infections were genuine Cluster 2, and that three double positives were predominantly Cluster 2, so these were selected for whole genome sequencing. The short-read sequencing of these 21 samples produced a mean of 6.95 million high-quality reads per infection sample with 53 – 82% read pairs properly mapped against the strain H reference genome (Table 5.2). The mean depth coverage across samples was 52.3-fold  $\pm$  15.4 SD (range from 28.7- to 80.3-fold). Given the high quality of genome assembly and depth coverage, all samples were used for analyses.

For comparative study and consistency in subsequent analyses, short reads of 59 whole genome data of *P. knowlesi* isolates generated previously (Assefa et al., 2015, Pinheiro et al., 2015) were also remapped against the version 2.0 of the reference genome using the same assembly parameters (Appendix 5.1). These whole-genome short reads were mostly generated using the HiSeq sequencing platform, showing higher depth coverage compared to reads generated using the MiSeq platform (mean 106.6-fold  $\pm$  49.9 SD).

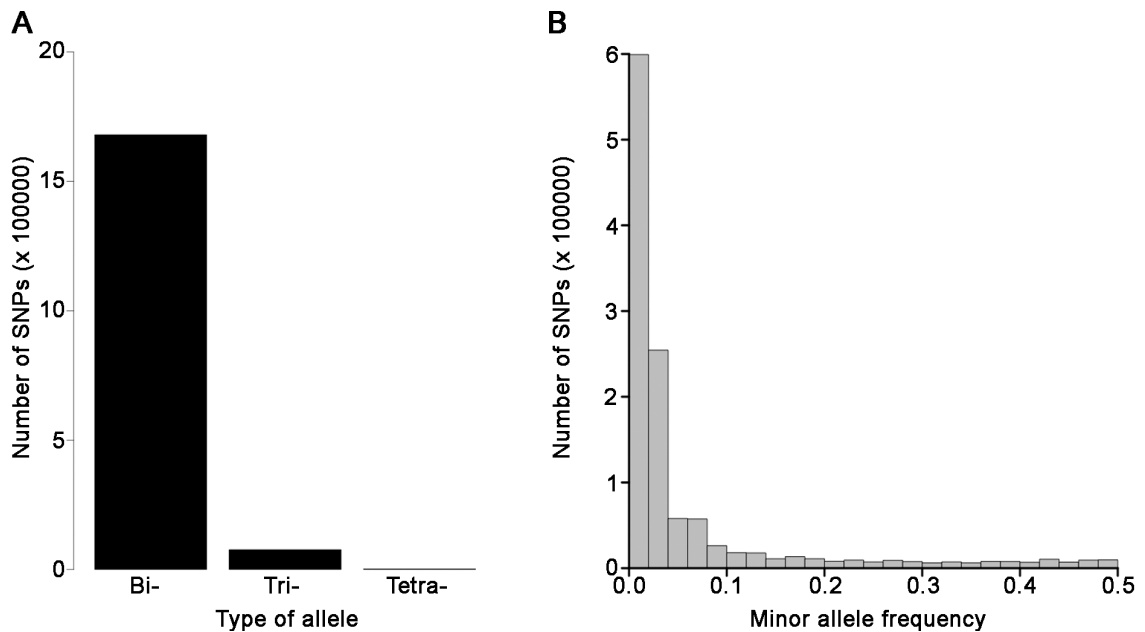
The combination of new and previously generated whole genome data enabled a new SNP list to be produced. A total of 2,123,043 SNPs were identified from the dataset of 80 infections. Of these, 2,109,937 (99%) were located in the nuclear genome, and

**Table 5.2.** Summary of mapping the short read sequences of 21 isolates generated using the MiSeq sequencing platform against the *P. knowlesi* H strain version 2.0 reference genome.

No	Sample ID	QC-passed reads	% mapped	% properly mapped	Mean depth coverage (X)
1	KT133	8,162,583	97.83	79.73	68.43
2	KT143	5,856,181	92.50	76.18	46.80
3	KT147	4,169,555	80.77	66.14	28.66
4	KT151	10,010,749	80.95	65.17	68.03
5	KT161	10,544,255	65.18	53.45	56.19
6	KT165	8,134,708	95.69	78.58	67.36
7	KT172	7,953,445	64.71	52.08	42.17
8	KT176	5,854,135	98.42	81.19	50.11
9	KT186	9,092,261	95.21	78.00	74.47
10	KT198	7,072,188	98.30	82.09	50.96
11	KT217	8,827,711	85.51	69.74	64.04
12	KT221	6,138,392	97.72	80.20	51.73
13	KT223	7,584,970	97.07	79.87	63.69
14	KT224	10,403,400	90.68	74.10	80.27
15	KT226	7,057,397	94.60	77.89	50.92
16	KT231	3,468,323	97.98	80.38	28.99
17	KT233	5,224,563	97.59	79.88	44.10
18	KT243	9,500,408	98.33	81.53	71.51
19	KT263	4,012,790	96.38	79.36	33.64
20	KT266	6,897,861	98.00	81.77	47.88
21	KT305	3,628,341	96.37	78.70	30.26

“QC-passed reads” means number of reads pass the quality control, “% mapped” means proportion of reads are successfully mapped while “% properly mapped” means forward and reverse reads are mapped properly in pairs with correct orientation.

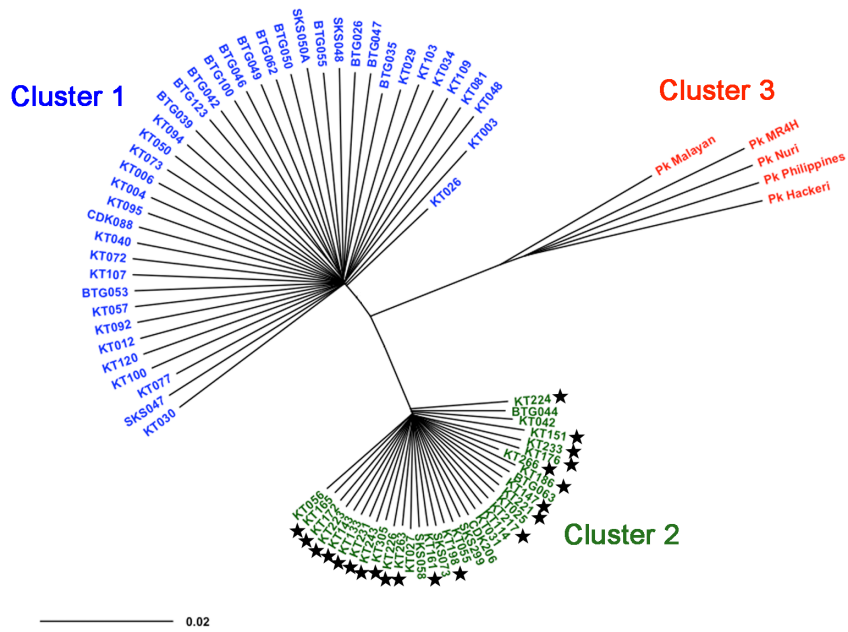
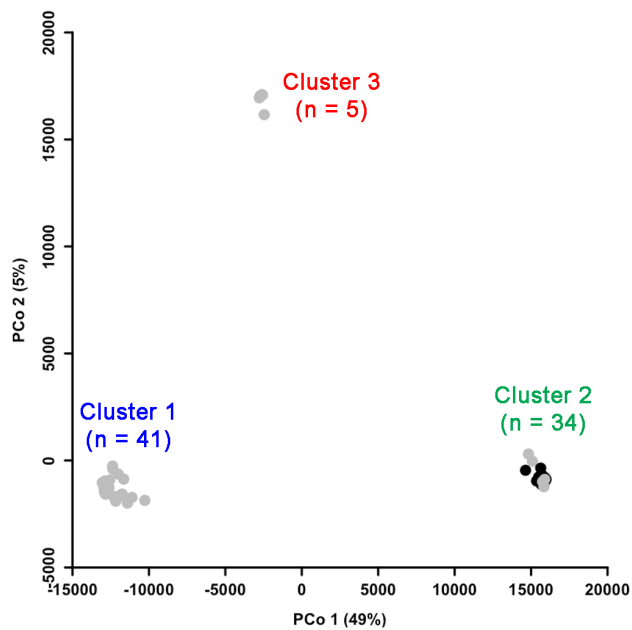
1,669,533 biallelic SNPs remained following exclusion of those positioned at the subtelomeric regions and the multigene families (Figure 5.1A). Of these, 1,186,073 high quality SNPs had less than 10% missingness in the sample of all isolates, and 16.1% (n = 190,901) of these had minor allele frequencies above 10% (Figure 5.1B) in this dataset.



**Figure 5.1.** Types of SNP alleles and distribution of minor allele frequency of biallelic SNPs across 14 chromosomes derived from 80 *P. knowlesi* isolate whole genome sequences. (A) Distribution of allele types after removing the subtelomeric regions and multigene gene families (*SICAvar* and *KIR* genes). (B) Distribution of SNPs across minor allele frequencies on 1,186,073 high quality biallelic SNPs of at least 10% missingness of total isolates at any positions. A total of 190,901 (16.1%) of these had minor allele frequencies above 10%.

### 5.3.2 Population structure

The data from the 21 new infections of the Cluster 2 type were combined with data from 59 infections (Appendix 5.1) from previous studies (Assefa et al., 2015, Pinheiro et al., 2015). Analysis of pairwise genetic distance among these isolates inferred by the Neighbour-Joining tree (Figure 5.2A) and the principal coordinate analysis plot (Figure 5.2B) confirmed that the existence of three subpopulation clusters. All 21 new infections assigned to Cluster 2 by microsatellite analysis were confirmed as Cluster 2 by sequence analysis (Table 5.3).

**A****B**

**Figure 5.2.** Population structure of *P. knowlesi* infections of 80 whole genome sequences.

Both (A) unrooted Neighbour-Joining tree and (B) principal coordinate analysis showed three major subpopulation clusters as previously seen elsewhere (Assefa et al., 2015). Data generated in this study (n=21) are marked with stars in the Neighbour-Joining tree and with black dots in the principal coordinate analysis plot.

**Table 5.3.** Subpopulation cluster assignments on 35 *P. knowlesi* infections in humans from Kapit, Malaysian Borneo.

Sample ID	MOI <sup>ref</sup>	Cluster assignments		Whole genome sequencing
		Microsatellite <sup>ref</sup>	Allele-specific PCR	
KT133	Single	C2	C2	C2
KT143	Single	C2	C2	C2
KT146	Mixed	C1	C1	nd
KT147	Single	C2	C2	C2
KT151	Single	C2	C2	C2
KT161	Single	C2	C2	C2
KT163	Single	C1	C1	nd
KT165	Mixed	C2	C2	C2
KT172	Single	C2	C2	C2
KT175	Single	C1	C1	nd
KT176	Single	C2	C1/C2	C2
KT186	Mixed	C2	C2	C2
KT188	Single	nd	C1	nd
KT191	Single	nd	C1	nd
KT198	Single	C2	C2	C2
KT217	Single	C2	C2	C2
KT219	Single	nd	C1	nd
KT221	Single	C2	C2	C2
KT223	Single	C2	C2	C2
KT224	Mixed	C2	C1/C2	C2
KT225	Mixed	C1	C1/C2	nd
KT226	Single	C2	C1/C2	C2
KT227	Single	nd	C1	nd
KT229	Mixed	C1	C1/C2	nd
KT231	Single	C2	C2	C2
KT233	Single	C2	C2	C2
KT238	Single	nd	C1	nd
KT243	Single	C2	C2	C2
KT251	Single	nd	C1	nd
KT255	Single	nd	C1	nd
KT263	Single	C2	C2	C2
KT266	Mixed	C2	C2	C2
KT269	Single	nd	C1	nd
KT278	Mixed	C1	C1/C2	nd
KT305	Single	C2	C2	C2

Cluster assignments of each individual infection deduced by three genotyping methods. The multiplicity of infection and cluster assignment information inferred by microsatellite analyses were derived from previous study (Chapter 4; Divis et al., 2017), marked as 'ref'. Allele-specific PCR assays were performed using primer sets C1A and C2J (refer to Chapter 3; Table 3.1). Results for whole genome sequencing were based on analyses by Neighbour-Joining method and Principal Coordinate Analysis. Abbreviation: nd – not done; C1 – Cluster 1; C2 – Cluster 2.



### 5.3.3 Genome-wide patterns of variation

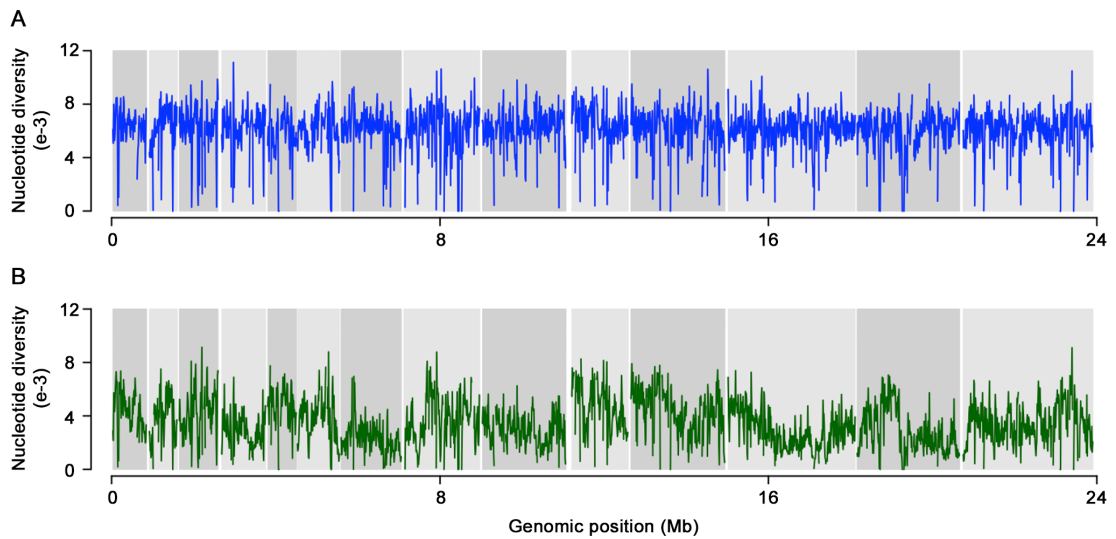
Genome-wide nucleotide diversity indices ( $\pi$ ) were assessed in non-overlapping 10-kb windows for the two-subpopulation clusters from Malaysian Borneo. The Cluster 1 subpopulation demonstrated a homogenous pattern of sequence diversity (Kruskal-Wallis  $P = 0.23$ ) across the 14 chromosomes (Table 5.4; Figure 5.3A). This was in contrast with the Cluster 2 subpopulation in which a heterogeneous pattern (Kruskal-Wallis  $P < 10^{-16}$ ) of diversity was observed across the genome (Figure 5.3B). The Cluster 2 subpopulation showed a wide range of pairwise diversities among the chromosomes ranging between  $2.25 \times 10^{-3}$  in chromosome 7 and  $4.38 \times 10^{-3}$  in chromosome 5 (Table 5.4). Overall, parasites in the Cluster 1 subpopulation showed higher sequence diversity than seen in the Cluster 2 subpopulation (Cluster 1  $\pi = 5.78 \times 10^{-3}$ ; Cluster 2  $\pi = 3.43 \times 10^{-3}$ ; Wilcoxon Signed Rank  $P < 10^{-16}$  for testing of heterogeneity across different chromosomes).

Using the same dataset, the summary of the allele frequency spectrum using the Tajima's  $D$  index was less variable across 14 chromosomes for the Cluster 1 subpopulation (Kruskal-Wallis  $P = 8.4 \times 10^{-5}$ ) (Figure 5.4A) compared to the Cluster 2 subpopulation (Kruskal-Wallis  $P = 1.6 \times 10^{-16}$ ) (Figure 5.4B). Both subpopulations showed strong skew towards low frequency variants, with mean Tajima's  $D$  values for the Cluster 2 subpopulation being even less than for the Cluster 1 subpopulation (Figure 5.5A; Cluster 1 mean  $D = -1.77$ ; Cluster 2 mean  $D = -2.37$ ; Wilcoxon Signed Rank  $P < 10^{-16}$ ), suggesting both parasite subpopulations have undergone previous population expansion. There was a positive correlation in the distribution of Tajima's  $D$  values between the two populations across all 10 kb windows (Figure 5.5B; Spearman's  $\rho = 0.25$ ;  $P < 10^{-16}$ ).

**Table 5.4:** Mean nucleotide diversity ( $\pi$ ) and pairwise differentiation ( $F_{ST}$ ) between Cluster 1 and Cluster 2 subpopulations.

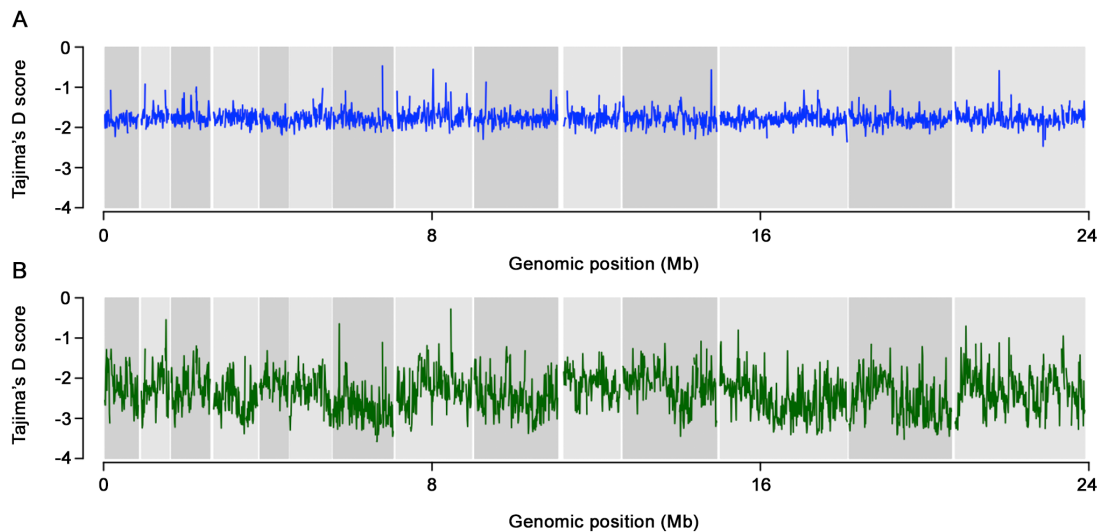
<b>Chromosome</b>	$\pi_{c1}$ ( $\times 10^{-3}$ )	$\pi_{c2}$ ( $\times 10^{-3}$ )	<b>Pairwise</b> $F_{ST}$
1	6.02	4.23	0.19
2	5.99	3.75	0.23
3	5.79	3.91	0.17
4	5.90	2.73	0.27
5	5.38	4.38	0.09
6	5.61	3.53	0.23
7	5.82	2.25	0.40
8	5.50	3.34	0.20
9	5.92	2.81	0.27
10	5.99	4.32	0.15
11	5.63	3.97	0.15
12	5.95	2.80	0.32
13	5.69	2.74	0.33
14	5.77	3.28	0.25

Calculations of  $\pi$  were based on non-overlapping 10-kb windows for each chromosome for the two *P. knowlesi* subpopulation clusters in Malaysian Borneo (Cluster 1 N = 41, Cluster 2 N = 34).



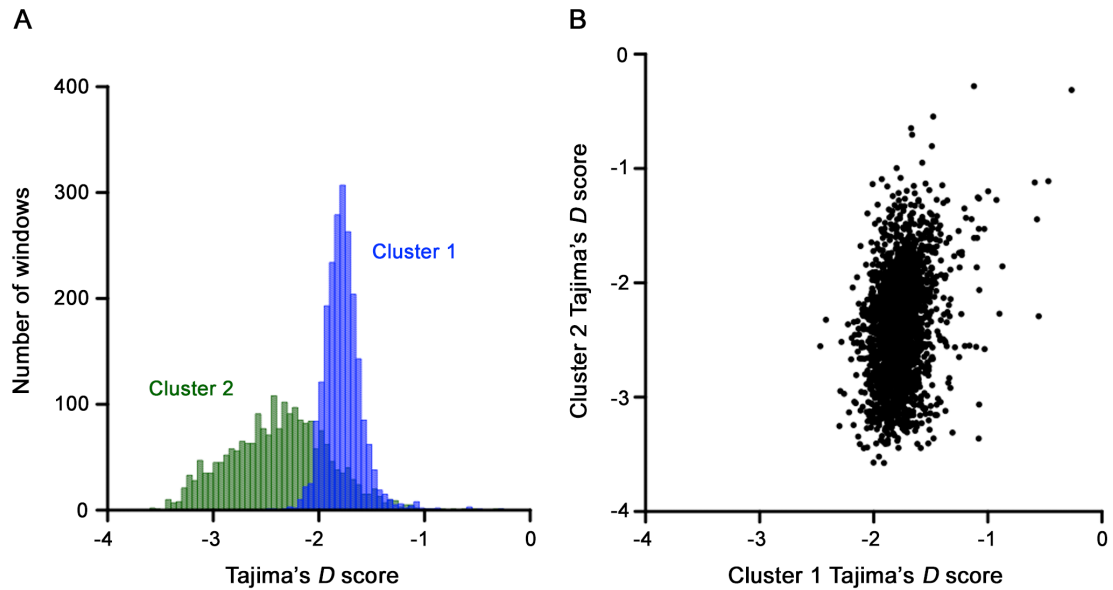
**Figure 5.3:** Genome-wide nucleotide diversity ( $\pi$ ) of two *P. knowlesi* subpopulation clusters in Malaysian Borneo.

The  $\pi$  values for subpopulation (A) Cluster 1 and (B) Cluster 2 were calculated in non-overlapping 10-kb windows for chromosome 1 to 14 as represented in alternate dark and light grey blocks. Gap between chromosomal blocks represents subtelomeric regions and not included in the analysis. The mean genome-wide  $\pi$  for Cluster 1 subpopulation ( $n = 41$ ) is  $5.78 \times 10^{-3}$  and for Cluster 2 subpopulation ( $n = 34$ ) is  $3.43 \times 10^{-3}$ .



**Figure 5.4:** Genome-wide Tajima's  $D$  values of two *P. knowlesi* subpopulation clusters in Malaysia Borneo.

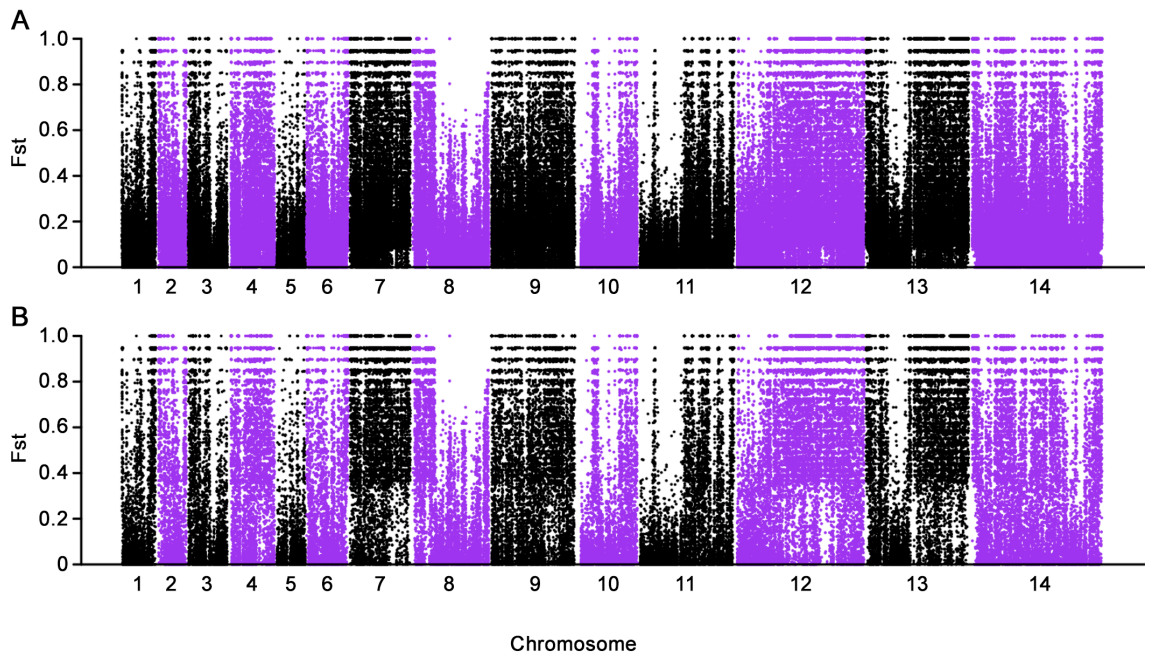
The  $D$  values for subpopulation (A) Cluster 1 and (B) Cluster 2 were calculated in non-overlapping 10-kb windows for chromosome 1 to 14 as presented in alternate dark and light grey blocks. The average genome-wide  $D$  values for Cluster 1 subpopulation is  $-1.77$  and for Cluster 2 subpopulation is  $-2.37$ .



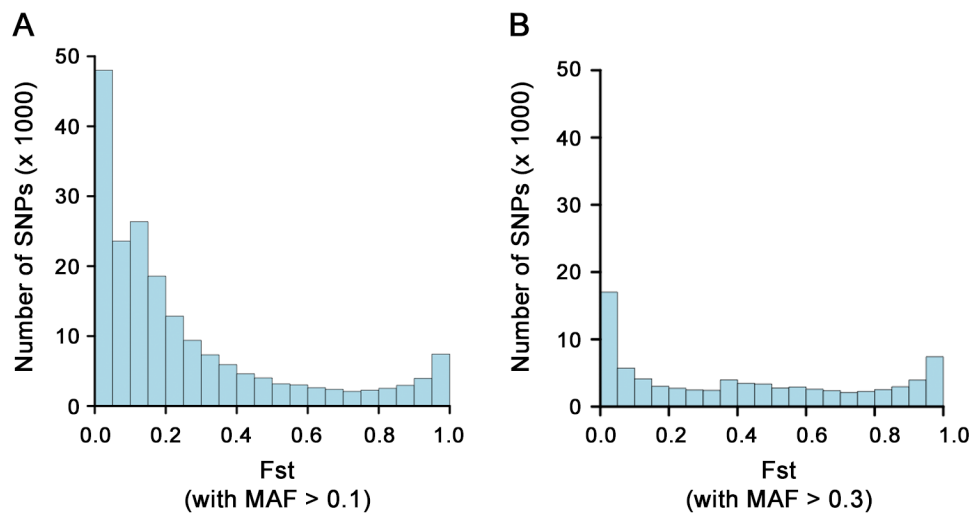
**Figure 5.5:** Comparison of genome-wide Tajima's  $D$  distributions between two *P. knowlesi* subpopulation clusters in Malaysian Borneo. (A) Frequency distribution of Tajima's  $D$  values in non-overlapping 10-kb windows for Cluster 2 subpopulation skews more negatively compared to Cluster 1 subpopulation. (B) Weak association of Tajima's  $D$  values is observed between the two subpopulations even though the correlation is significant (Spearman's  $\sigma = 0.25$ ,  $P < 10^{-16}$ ).

### 5.3.4 Genomic profiles of differentiation between subpopulations

The variation in genomic patterns of diversity between the two subpopulations suggested that there might be variation in levels of divergence between them. Using 193,068 SNPs with minor allele frequency (MAF) above 0.1 between Cluster 1 and Cluster 2 subpopulations in Malaysian Borneo, the mean genome-wide fixation index ( $F_{ST}$ ) indicated substantial overall divergence between the two subpopulations (mean  $F_{ST} = 0.25$ ; Figure 5.6A). Overall frequency distribution of  $F_{ST}$  was bimodal, with a first peak having values just above zero and a second peak having values towards 1.0 indicating complete fixation (Figure 5.7A). Very high  $F_{ST}$  values of  $> 0.8$  was seen in 19,116 SNPs, with 7,415 (3.8%) SNPs showing complete fixation ( $F_{ST} = 1$ ). Each chromosome showed different mean  $F_{ST}$  values, ranging from 0.09 (for chromosome 5)



**Figure 5.6:** Genome-wide plots of  $F_{ST}$  divergence between the sympatric Cluster 1 and Cluster 2 subpopulations of *P. knowlesi* in Malaysian Borneo. Each dot represents value of  $F_{ST}$  for individual SNP. The  $F_{ST}$  plots were assessed with different levels of minor allele frequencies (MAF): (A) MAF > 0.1 with mean  $F_{ST}$  of 0.25, (B) MAF > 0.3 with mean  $F_{ST}$  of 0.42.



**Figure 5.7:** Frequency distribution of  $F_{ST}$  values for SNPs at different levels of minor allele frequencies (MAF). A total of (A) 193,068 SNPs were identified at MAF > 0.1, and (B) 80,168 SNPs at MAF > 0.3.

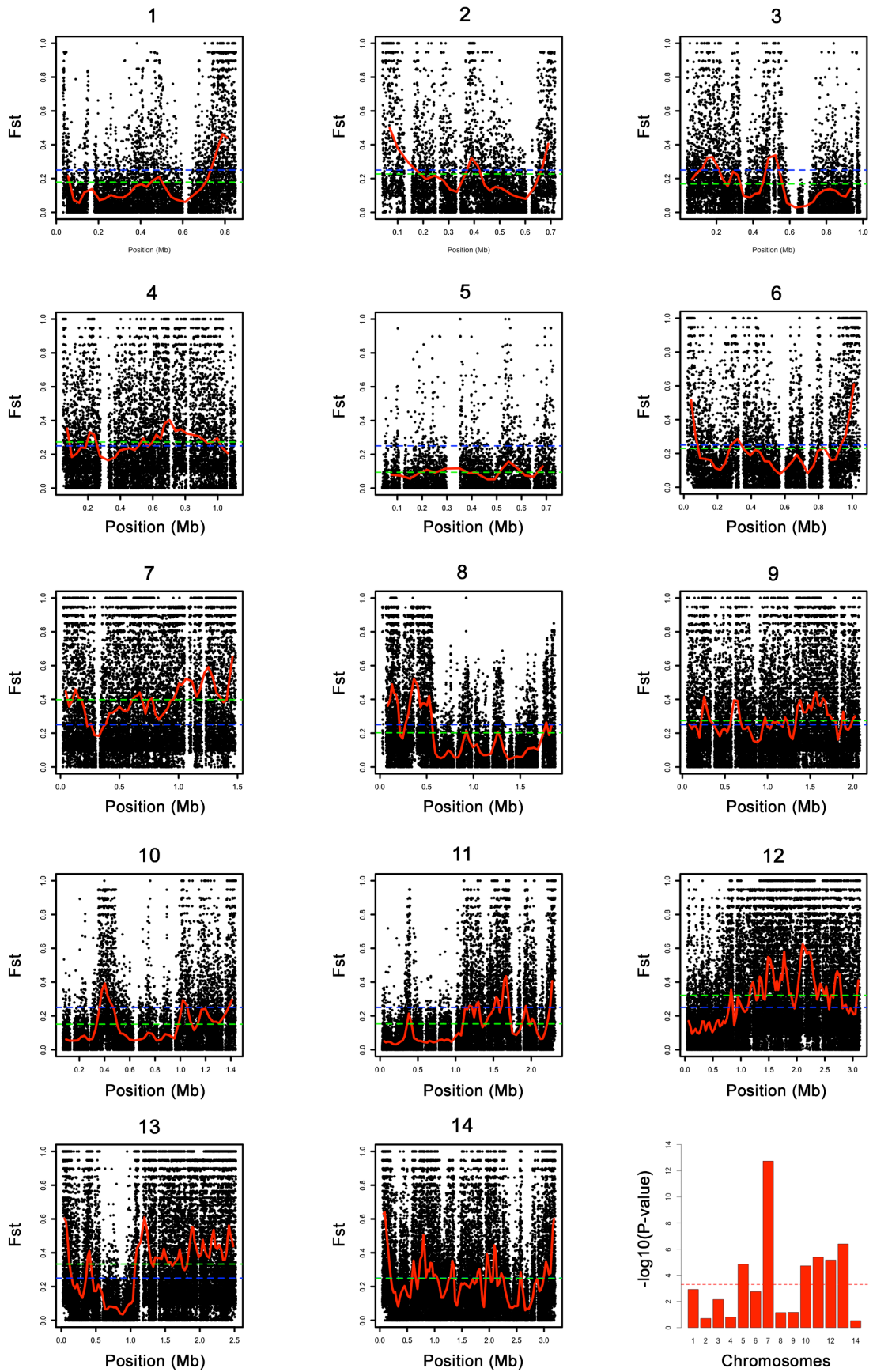
to 0.40 for (chromosome 7) (Figure 5.8; Table 5.4). Mean  $F_{ST}$ , which was calculated using windows of 500 consecutive SNPs and sliding by 250 SNPs revealed significant variations across the genome.

To assess the dependency of  $F_{ST}$  distributions on the underlying SNP allele frequencies, analysis of the fixation indices using different levels of minor allele frequencies was undertaken. It was revealed that most SNPs with low  $F_{ST}$  values were removed when SNPs with low minor allele frequencies were removed (Figure 5.6B, Figure 5.7B).

### **5.3.5 Genomic regions of high and low divergence**

Regions of divergence with elevated  $F_{ST}$  between Cluster 1 and Cluster 2 subpopulations were distributed across the genome. Low frequency allele SNPs potentially obscure the distinguishing of highly differentiated windows (Roesti et al., 2012, Marques et al., 2016), as a minimum MAF of 0.3 is required for a SNP potentially be able to reach an elevated  $F_{ST}$  value of 0.8 or more. The genome-wide average  $F_{ST}$  value increased from 0.25 to 0.42 when SNPs with MAF < 0.3 were removed (Figure 5.6B).

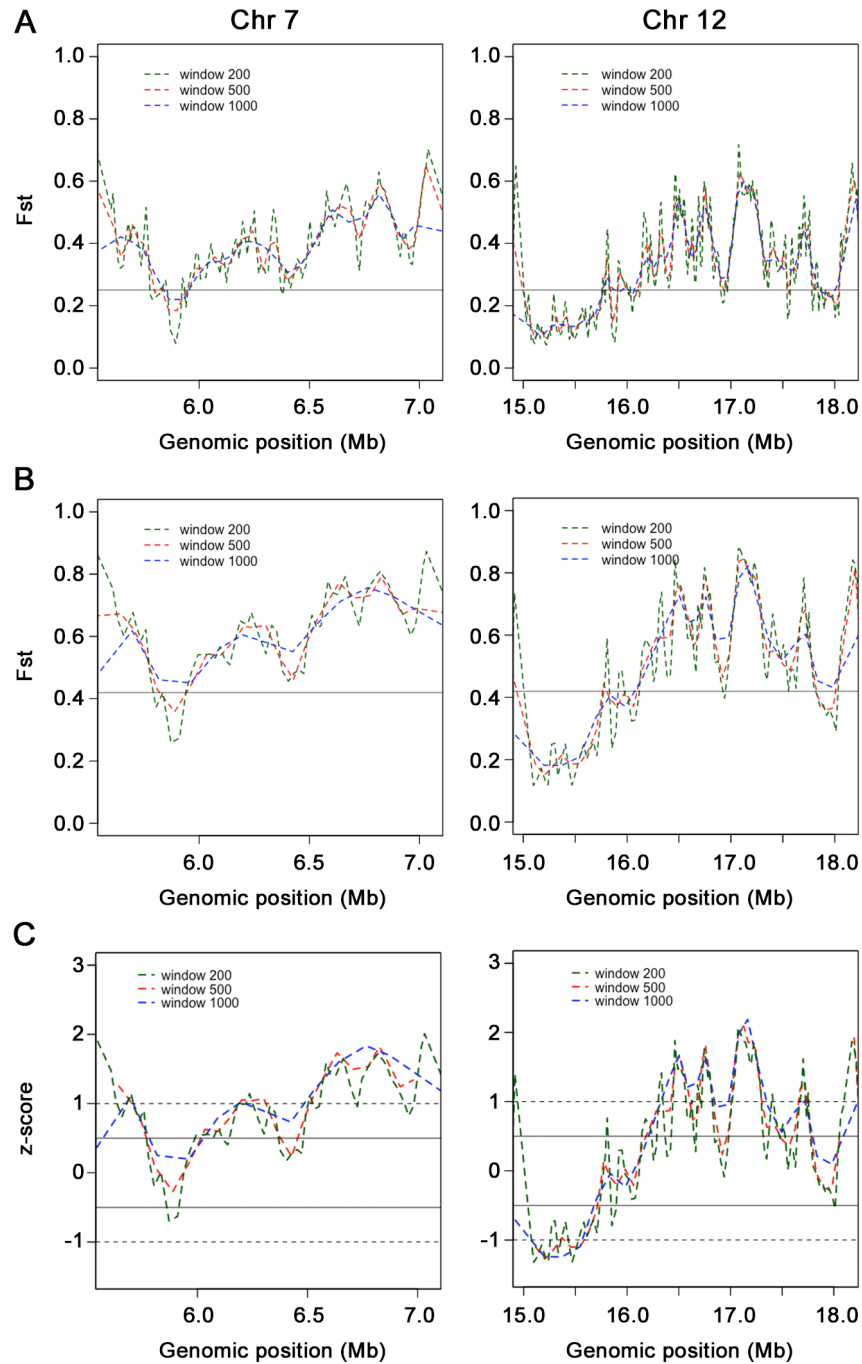
Variation in divergence across the genome was explored using three sliding window sizes (windows of 200, 500 and 1000 consecutive SNPs) at MAF above 0.1 and 0.3 (Figure 5.9A and B). Based on the sliding windows of 500 SNPs, the relative level of population differentiation of all windows were evaluated by calculating standard deviations departing from the mean genome-wide  $F_{ST}$  (z-score). The definition of genomic regions was tested at two z-score thresholds in order to identify contiguous blocks that contain the most fixed SNPs (Figure 5.9C; z-score threshold of 0.5). Regions



**Figure 5.8:** Distribution of genetic differentiation ( $F_{ST}$ ) in 14 chromosomes between Cluster 1 and Cluster 2 *P. knowlesi* subpopulations in Malaysian Borneo.

Black dots represent  $F_{ST}$  values for individual SNPs with minor allele frequency above 10%. The solid red lines represent average  $F_{ST}$  values for sliding windows of 500 consecutive SNPs, with overlapping by 250 SNPs. The blue dashed lines denote average genome-wide  $F_{ST}$  value of 0.25 while green dashed lines denote average  $F_{ST}$  values for each chromosome as shown in Table 5.4. The significance of tests whether more windows are above or below the average genome-wide  $F_{ST}$  (in dashed red line) are shown in the bar plot panel with P values from Fisher's Exact tests.





**Figure 5.9:** Determination of genomic regions of divergence between two subpopulations of *P. knowlesi* in Malaysian Borneo. Results are shown for chromosome 7 and 12 containing mixed peaks and troughs of  $F_{ST}$  windows. Windows of  $F_{ST}$  values departing from the average genome-wide  $F_{ST}$ , shown by the solid horizontal lines, were evaluated by 200, 500 and 1000 consecutive SNPs on (A) minor allele frequency > 0.1, and (B) minor allele frequency > 0.3. (C) The definition of genomic regions was tested on two standard z-scores of 0.5 (shown by the solid horizontal lines) and 1.0 (shown by the dotted horizontal lines). Compared to z-score threshold of 1.0, z-score above 0.5 captures higher numbers of fixed SNPs ( $F_{ST} = 1$ ).

were then demarcated using these standardised z-score thresholds of 0.5 departing from the average genome-wide  $F_{ST}$  value. These regions were then grouped into three categories: low divergence regions (LDR with z-score < -0.5), intermediate divergence regions (IDR with z-score between -0.5 and 0.5) and high divergence regions (HDR with z-score > 0.5), which are summarised in Table 5.5.

Focusing on the high- and low divergence regions (Table 5.5 and Table 5.6), LDRs accounted for more of the genome (10 Mb), but a high proportion of SNPs with elevated  $F_{ST}$  (14.9% of 80,168 SNPs with MAF > 0.3) was densely accumulated in HDRs (Table 5.5). Three chromosomes (7, 12 and 13) had HDRs occupying more than 50% of the entire chromosome while chromosome 3, 5 and 10 showed no HDRs (Figure 5.10A).

### 5.3.6 Diversity of genomic regions between subpopulations

To explore the basis of the divergence, the nucleotide diversity ( $\pi$ ) was analysed at non-overlapping 10-kb windows within the divergence regions. Comparing between the two subpopulations, the differences in nucleotide diversity ( $\pi$ -diff in Figure 5.10B) were higher in the HDRs than in the LDRs or in the rest of the genome (Figure 5.11; Mann-Whitney U P <  $10^{-16}$  for both comparisons). This demonstrates that many of the extremely differentiated regions were those where  $\pi$ -diff was high.

Interestingly, each *P. knowlesi* subpopulation showed variations in the distribution of the HDRs and LDRs. Reduced nucleotide diversity patterns in HDRs were seen particularly in the Cluster 2 subpopulation (Figure 5.12; rho = 0.44, Mann-Whitney U P <  $2.2 \times 10^{-16}$ ; mean  $\pi = 2.08 \times 10^{-3}$ ), compared to Cluster 1 subpopulation where

**Table 5.5:** Summary of mean value indices among low divergence regions (LDR), intermediate divergence regions (IDR) and high divergence regions (HDR) between two divergent of *P. knowlesi* subpopulations in Malaysian Borneo.

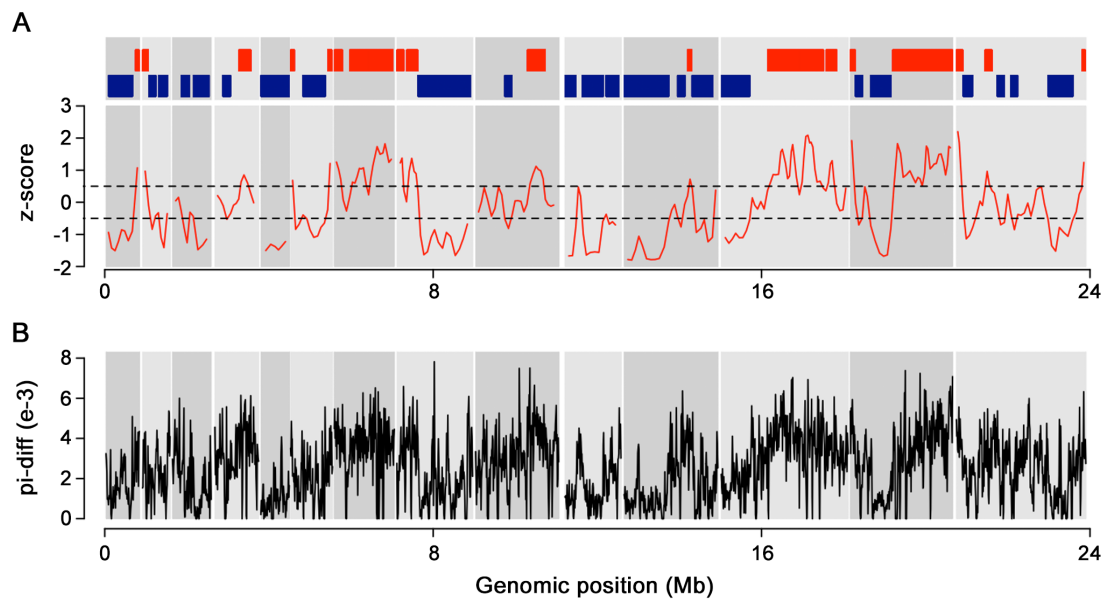
<b>Genomic parameter</b>	<b>LDR</b>	<b>IDR</b>	<b>HDR</b>
$F_{ST}$ - MAF > 0.1	0.14	0.26	0.41
$F_{ST}$ - MAF > 0.3	0.20	0.41	0.64
No. of windows	23	32	20
Smallest window size (kb)	175.2	6.9	82.9
Largest window size (kb)	1,292.7	741.2	1,461.2
Total length (Mb)	10.0	6.0	6.5
No. of fixed SNPs	204	1,157	5,696
SNPs density of elevated $F_{ST}$ ( $\pm$ SD)	0.09 (0.06)	0.60 (0.40)	1.93 (0.66)
$\pi$ - Cluster 1 ( $\times 10^{-3}$ )	5.60	6.20	5.80
$\pi$ - Cluster 2 ( $\times 10^{-3}$ )	4.14	3.24	2.08

Genetic differentiation ( $F_{ST}$ ) was measured in two minor allele frequencies (MAF), while nucleotide diversity ( $\pi$ ) was calculated based on 10-kb non-overlapping windows.

**Table 5.6.** Locations and lengths of high divergence regions (HDRs) and low divergence regions (LDRs) in 14 chromosomes of *P. knowlesi*.

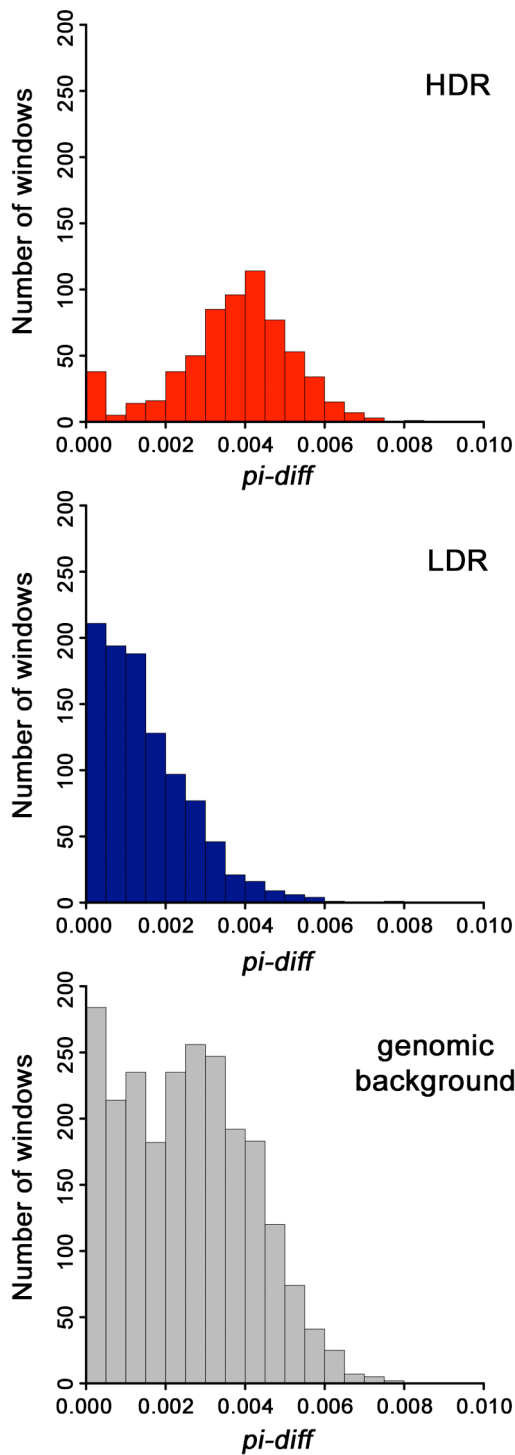
High-diverged regions (HDRs)					Low-diverged regions (LDRs)			
Chr	HDR window	Start position	End position	Length (bp)	HDR window	Start position	End position	Length (bp)
1	HDR01	749494	837835	88342	LDR01	96319	679406	583088
2	HDR02	40006	180420	140415	LDR02	186366	364472	178107
3					LDR03	425017	648577	238648
4	HDR03	625263	919357	294095	LDR04	252738	460344	207607
5					LDR05	546998	937541	390544
6	HDR04	21165	112965	91801	LDR06	234546	421831	187286
	HDR05	921498	1024278	102781	LDR07	40804	733264	692461
7	HDR06	23701	222469	198769	LDR08	318443	864442	546000
	HDR07	397684	853275	455592				
	HDR08	860949	1452166	591218				
8	HDR09	26256	190875	164620	LDR09	534591	1827275	1292685
	HDR10	260538	533984	273447				
9	HDR11	1316694	1754323	437630	LDR10	765850	945549	179700
10					LDR11	70745	338054	267310
					LDR12	481963	1003531	521569
					LDR13	1067530	1385663	318134
11	HDR12	1570780	1676811	106032	LDR14	30861	1117021	1086161
					LDR15	1326585	1507915	181331
					LDR16	1681406	2194542	513137
12	HDR13	1165069	1935432	770364	LDR17	39358	729445	690088
	HDR14	1951856	2524953	573098				
	HDR15	2590082	2849921	259840				
13	HDR16	24680	146791	122112	LDR18	147119	324956	177838
	HDR17	1053698	2514922	1461225	LDR19	515520	1017446	501927
14	HDR18	46282	200282	154001	LDR20	207562	441315	233754
	HDR19	743011	915540	172530	LDR21	1036448	1221194	266306
	HDR20	3103808	3186756	82949	LDR22	1363873	1539047	175175
					LDR23	2280342	2888810	608469

Chr – chromosome

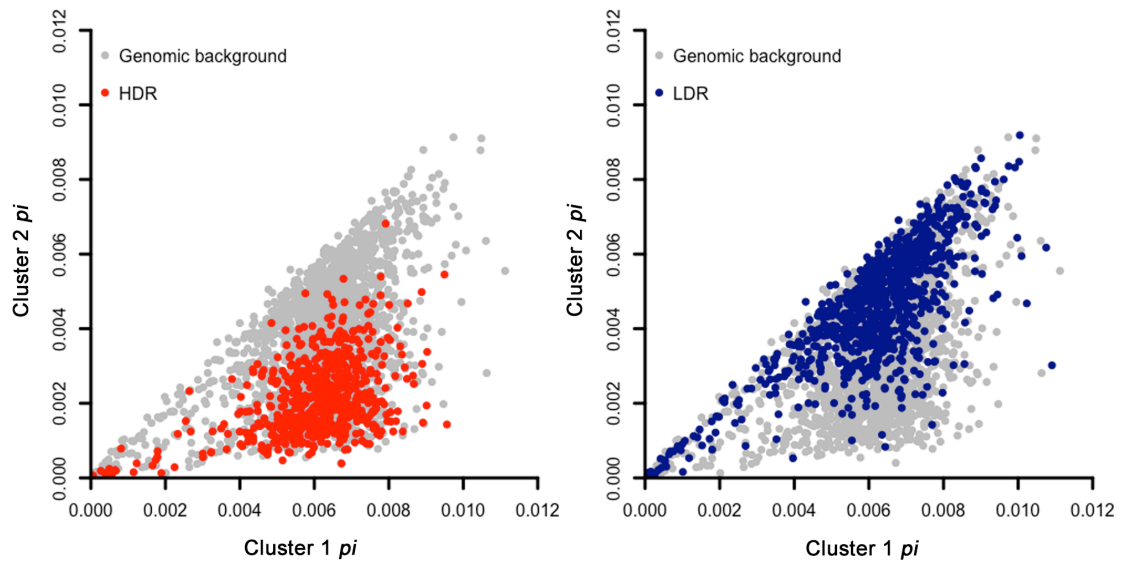


**Figure 5.10:** Genomic landscape between two divergent subpopulations of *P. knowlesi* in Malaysian Borneo.

(A)  $F_{ST}$  values based on windows of 500 consecutive SNPs were converted to standardised z-scores. The genomic thresholds of 0.5 standard deviations above and below the genome-wide average  $F_{ST}$  (z-score = 0) demarcate the high divergence regions (in red blocks) and low divergence regions (in dark blue blocks). (B) Pattern of nucleotide diversity differences between Cluster 1 and Cluster 2 subpopulations, denoted as pi-diff ( $\pi$ -diff), in relation to the genome-wide  $F_{ST}$  windows. Chromosomes 1 to 14 are represented in alternate dark and light grey blocks.



**Figure 5.11.** Spectrum of nucleotide diversity in genomic regions between Cluster 1 and Cluster 2 subpopulations. Distribution of degree of differences in nucleotide diversity ( $\pi$ -diff) based on 10-kb windows between the two subpopulations in high divergence regions (HDR) and low divergence regions (LDR). The genome-wide  $\pi$ -diff, referring as genomic background, is also shown.



**Figure 5.12.** Distribution of 10-kb windows  $\pi$  in high divergence regions (HDR) and low divergence regions (LDR) between Cluster 1 and Cluster 2 subpopulations. For the HDR, the mean  $\pi = 5.80 \times 10^{-3}$  for Cluster 1 and mean  $\pi = 2.08 \times 10^{-3}$  for Cluster 2. For the LDR, and the mean  $\pi = 5.60 \times 10^{-3}$  for Cluster 1 and mean  $\pi = 4.14 \times 10^{-3}$  for Cluster 2.

nucleotide diversity showed no difference between HDRs and the rest of the genome (Figure 5.12, mean  $\pi = 5.80 \times 10^{-3}$ ; Figure 5.3A; Table 5.5; Mann-Whitney U P = 0.25).

For LDRs in the Cluster 2 subpopulation, mean nucleotide diversity was higher compared to the rest of the genome (Table 5.5; Figure 5.12; mean  $\pi = 4.14 \times 10^{-3}$ ;

Mann-Whitney U P =  $2.2 \times 10^{-16}$ ), while the Cluster 1 subpopulation showed

homogeneity in nucleotide diversity across the genome (mean  $\pi = 5.60 \times 10^{-3}$ ; Mann-Whitney U P = 0.77).

### 5.3.7 Phylogeny of extra-chromosomal genomes

The analyses of population structure were further extended using the maternally inherited extra-chromosomal genomes. DNA sequences extracted from the whole

genome data were first identified for their cluster assignments based on the information derived from the nuclear genome analysis as described before.

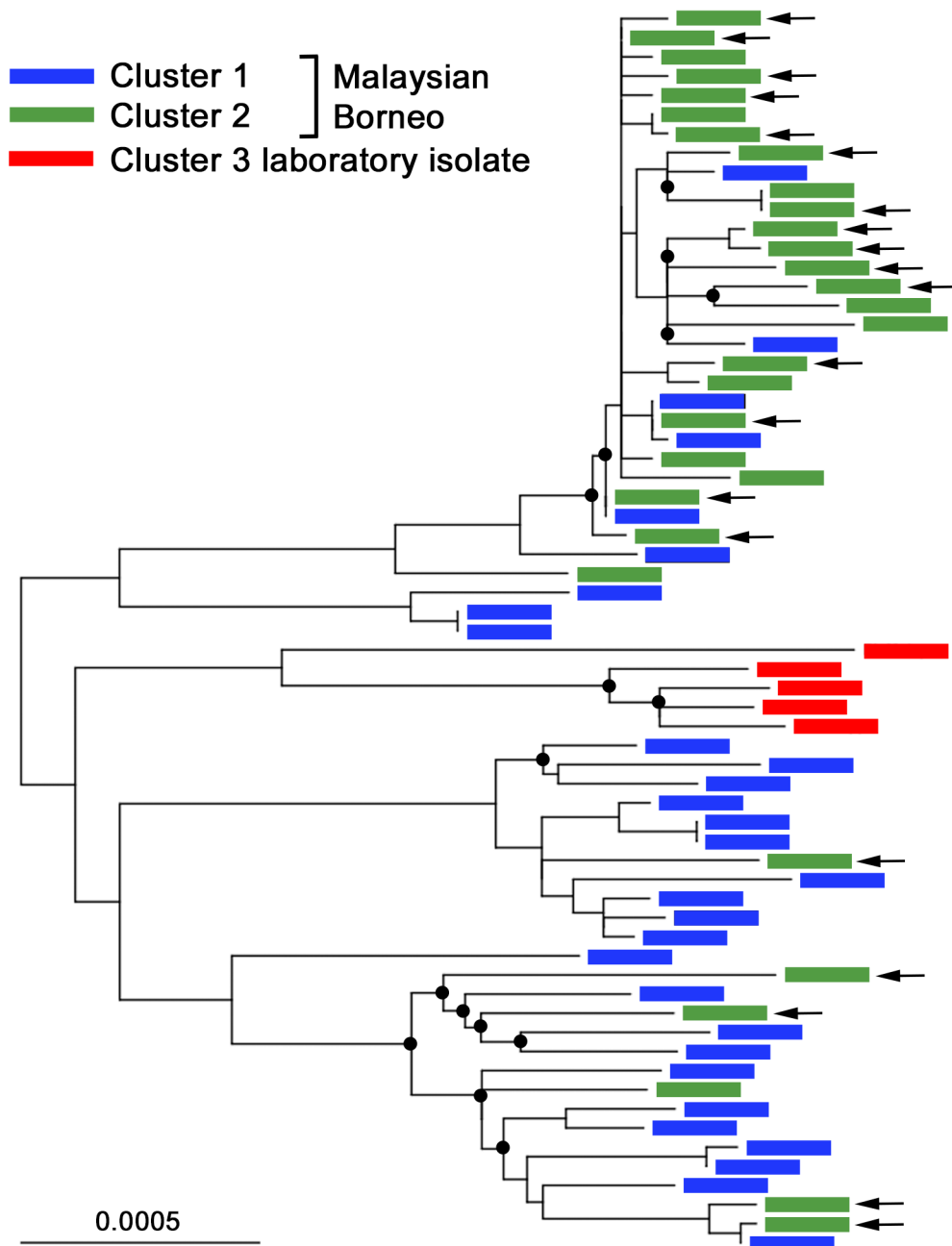
#### **5.3.7.1 Apicoplast genome**

Scanning through the 30.6-kb apicoplast genomes, 15 of the 80 infections were excluded from the analysis due to having 20 – 84% missing SNPs across the genome, allowing 65 infections to be analysed. The alignment of the apicoplast haplotypes identified 520 polymorphic sites with global nucleotide diversity ( $\pi$ ) of  $1.79 \times 10^{-3}$  and haplotype diversity ( $Hd$ ) of 0.998 (SD  $\pm$  0.003). When the haplotypes were grouped into three subpopulation clusters, the average nucleotide diversity among the Cluster 1 haplotypes ( $n = 27$ ;  $\pi = 1.77 \times 10^{-3}$ ) was higher than the Cluster 2 ( $n = 30$ ;  $\pi = 1.12 \times 10^{-3}$ ) and Cluster 3 ( $n = 5$ ;  $\pi = 1.14 \times 10^{-3}$ ) haplotypes. Two major lineages were seen, one of which consisted predominantly of Cluster 1 samples, and the other mainly of Cluster 2 samples (Figure 5.13).

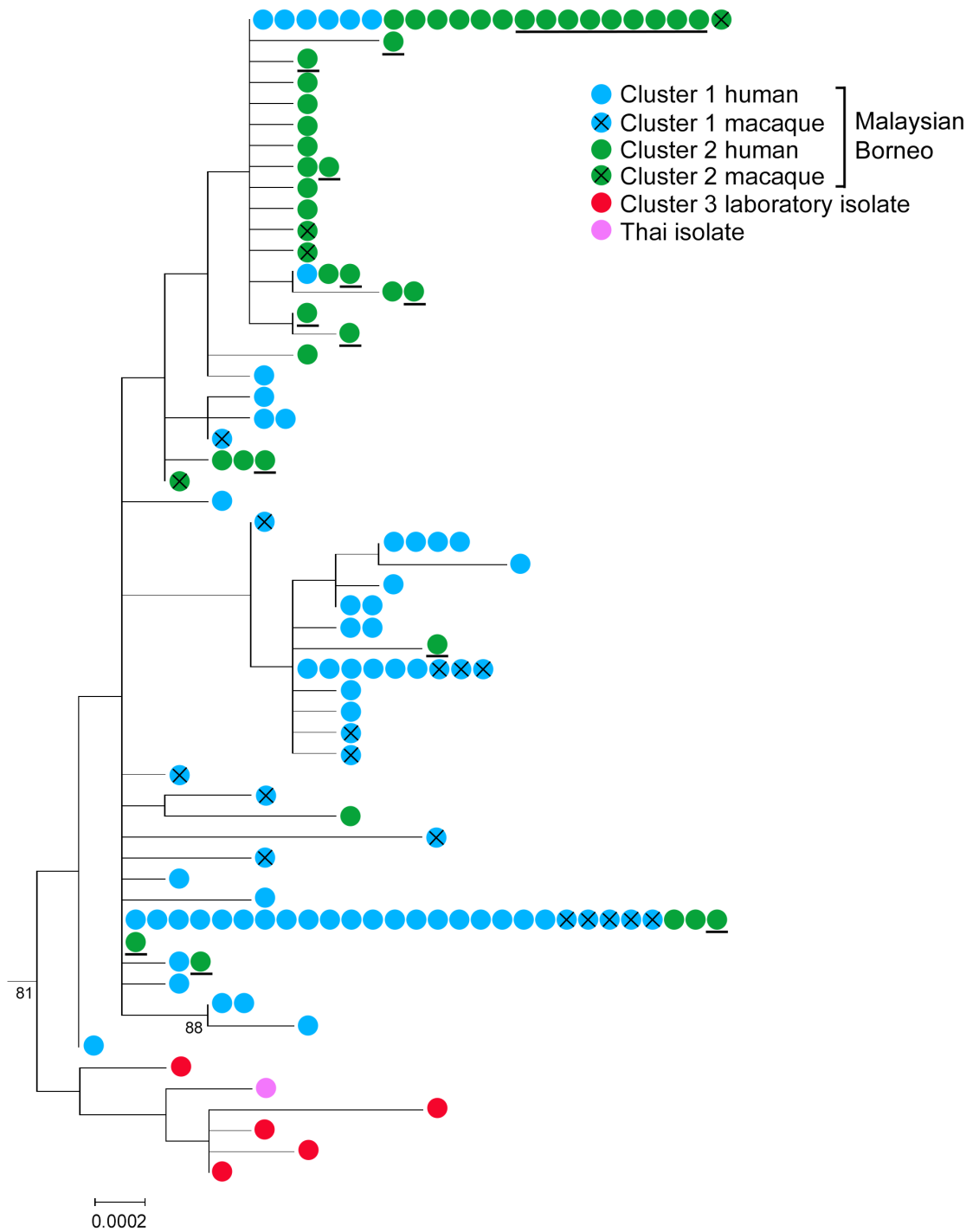
#### **5.3.7.2 Mitochondrial genome**

The combination of 5.9-kb mitochondrial haplotypes generated in this study with previously published haplotypes (total  $n = 129$ ) identified 77 polymorphic sites with global average nucleotide diversity ( $\pi$ ) of  $7.9 \times 10^{-4}$ . Analysis of pairwise nucleotide diversity showed higher values in Cluster 1 ( $n = 74$ ;  $\pi = 6.8 \times 10^{-4}$ ) and Cluster 3 ( $n = 6$ ;  $\pi = 8.8 \times 10^{-4}$ ) haplotypes compared to the Cluster 2 ( $n = 46$ ;  $\pi = 4.9 \times 10^{-4}$ ) haplotypes. The maximum likelihood phylogeny yielded a similar pattern as seen using the apicoplast dataset (Figure 5.14), with no clear fixation of different haplotypes between the Cluster 1 and Cluster 2 subpopulations. Although the fixation of Cluster 1 and Cluster 2 haplotypes were not completely achieved as shown by analyses of both





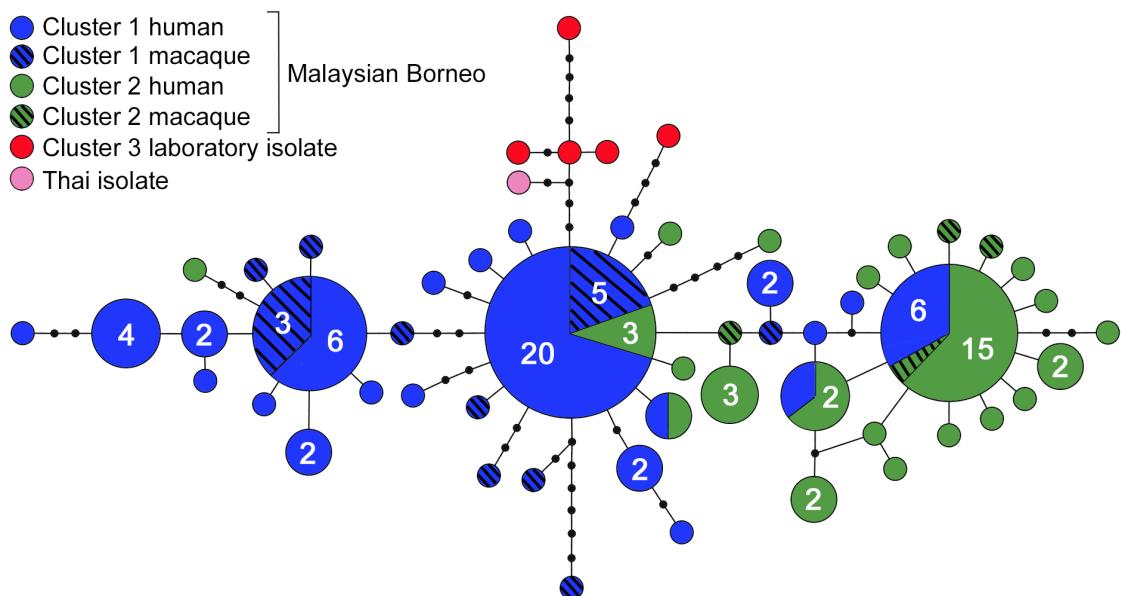
**Figure 5.13.** Maximum likelihood phylogenies inferred by 30.6-kb apicoplast genomes derived from 65 *P. knowlesi* infections. All nodes had bootstrap values above 80% based on 1000 replicates, except for those marked with black dots. The scale horizontal bar indicates nucleotide substitutions per site while vertical branch has no meaning. Data generated in this study (n=21) are marked with arrows.



**Figure 5.14.** Maximum likelihood phylogenies inferred by 5.9-kb mitochondrial genomes of 129 *P. knowlesi* haplotypes. Bootstrap values on nodes are only shown where they are above 80% based on 1000 replicates. The scale horizontal bar indicates nucleotide substitutions per site while vertical branch has no meaning. Data generated in this study (n=21) are marked with underlined circles.

extra-chromosomal genomes, the laboratory isolates (Cluster 3) formed a separate clade with strong bootstrap supports, separating all infections of mainland Southeast Asia from those previously isolated in Malaysian Borneo.

The analysis of genealogical network based on the mitochondrial genomes identified 56 main haplotypes with a complex and diversified topology (Figure 5.15; mean  $Hd = 0.915$ ;  $SD \pm 0.016$ ). The main core haplotype consisted of mainly parasites of Cluster 1 (89%), which then linked to other Cluster 1 and Cluster 2 haplotypes. The second core haplotypes were mainly formed by 70% of Cluster 2 isolates. The branch-out of the laboratory isolates (mostly mainland Southeast Asian) from Malaysian Borneo was expected, supporting the geographical divergence as shown by the maximum likelihood trees.



**Figure 5.15:** Genealogical network based on 129 *P. knowlesi* mitochondrial DNA genomes from Malaysian Borneo and mainland Southeast Asia showing 56 major haplotypes.

The sizes of the circles represent haplotype frequency shown by numbers of samples while those unlabelled indicate one sample. Connectors between the circles represent one mutational step and black dots represent hypothetical missing intermediates.

## 5.4 Discussion

Molecular genomic approaches deliver reliable and powerful tools to evaluate the divergence of related organisms sampled from different ecological environments (Malinsky et al., 2015, Marques et al., 2016). In previous studies, the population structure of *P. knowlesi* infections has been defined as two sympatric divergent subpopulations in Malaysian Borneo and a third separate subpopulation observed particularly in the mainland Peninsular Malaysia (Divis et al., 2015, Assefa et al., 2015, Divis et al., 2017). Using whole genome sequencing approaches, understanding the genomic patterns of divergence between parasites in different subpopulations might increase the knowledge of adaptation mechanisms.

The overall population structure of *P. knowlesi* was re-evaluated using both nuclear and extra-chromosomal genomes. Following the re-mapping of new and old short read data against the version 2.0 of *P. knowlesi* reference sequence, a well-defined 3 subpopulation clusters were seen as previously described (Assefa et al., 2015) based on the nuclear genome. The three divergent clusters might not represent the discovery of new species with the presentation by the nuclear genome alone, as this claim requires additional sequence data sets (Liu et al., 2016). The complete genomic data sets from the organelle genomes give additional information on the three-subpopulation clusters within *P. knowlesi*. The apicoplast and the mitochondrion in *Plasmodium* are inherited maternally as a sticky duo (Lim and McFadden, 2010) with negligible recombination in either genome at the population level. Using this additional datasets, analyses of the extra-chromosomal genomes do not show complete fixation of different haplotypes between the *P. knowlesi* subpopulations in

Malaysian Borneo, suggesting that these sympatric divergent subpopulations do not represent two different species.

If the separation of sympatric Cluster 1 and Cluster 2 subpopulations is the result of independent zoonosis, deep differentiation and profound genetic diversity between the two subpopulations are expected. Measures of allele frequency spectra using the Tajima's D statistics showed negative skew for both subpopulations, signifying long-term population growth. However, it was observed that each subpopulation has undergone different population expansion, since Cluster 2 subpopulation showed more negative Tajima's D value ( $D = -2.37$ ) than Cluster 1 subpopulation ( $D = -1.77$ ).

The pattern of divergence between the sympatric subpopulations in Malaysian Borneo was highly heterogeneous and a large number of SNPs showed complete fixation ( $F_{ST} = 1$ ). This extraordinary phenomenon is not seen in other populations of *Plasmodium* species, such as *P. falciparum* in high endemic areas in West Africa (Miotto et al., 2013), or in other species from different ecological environments, such as *Timema* stick insects (Soria-Carrasco et al., 2014) and Baltic Sea three-spined sticklebacks (Guo et al., 2015), where genome-wide differentiation were generally low ( $F_{ST} < 0.1$ ). The high differentiation between these two *P. knowlesi* subpopulations indicates limited gene flow occurring between them. The architecture of genomic regions exhibits 28% (6.5 Mb) of 23.0 Mb located in the highly diverged regions. Reduced genetic diversity of the Cluster 2 subpopulation observed in highly diverged regions suggest there may have been an initial bottleneck in the formation of this subpopulation (Malinsky et al., 2015).

Overall, results presented here suggest that two sympatric *P. knowlesi* subpopulations in Malaysian Borneo have undergone independent evolution. It is not known whether parasites from Peninsular Malaysia show the same divergence patterns. Because of the small number of old laboratory isolates, further sampling from this region is required to investigate the genomic patterns relative to parasites from Malaysian Borneo. This would provide a complete picture of divergence among these three subpopulations, and understand the evolutionary history of this zoonotic parasite system.

# Chapter Six

## General discussion and perspectives for the future

*P. knowlesi* is a simian parasite that infects humans in the natural environment in Southeast Asia (Singh et al., 2004, Singh and Daneshvar, 2013, Antinori et al., 2013). Infections exhibit a wide spectrum of clinical manifestation and are potentially life-threatening (Daneshvar et al., 2009, Cox-Singh et al., 2010, Rajahram et al., 2012, Rajahram et al., 2013). Prior to this thesis analyses of the circumsporozoite protein gene sequence and the mitochondrial genome showed shared haplotypes in parasites sampled from infections in human and wild macaque hosts (Lee et al., 2011). This suggested long-tailed macaques (*M. fascicularis*) and pig-tailed macaques (*M. nemestrina*) are both reservoirs of infections in humans.

Because of the high incidence of clinical cases in Malaysia compared to other countries in Southeast Asia (Singh and Daneshvar, 2013), a large population genetic survey of *P. knowlesi* was conducted across this country by sampling infections from human and wild macaque hosts. First, a novel genotyping toolkit consisting of 10 microsatellite markers was successfully developed (Chapter Two), suitable for studying diversity within infections, as well as within and between populations (Divis et al., 2015). Application of this genotyping toolkit revealed an extraordinary population genetic structure of *P. knowlesi* that has never been seen in any other *Plasmodium* species.

This study initially indicated that human *P. knowlesi* infections are an admixture of two divergent subpopulations, respectively associated with long-tailed macaques (referred



to as Cluster 1 parasite subpopulation) and pig-tailed macaques (Cluster 2 subpopulation) (Divis et al., 2015). The divergence between these two subpopulations was profound (Chapters Two and Four), signifying that independent zoonoses are occurring sympatrically (Divis et al., 2015, Divis et al., 2017). This finding was further confirmed by SNP analysis after whole genome sequencing of clinical isolates (Assefa et al., 2015), including some of the samples used in the microsatellite genotyping surveys.

Despite the appearance of these two divergent subpopulations by microsatellite analysis, *P. knowlesi* laboratory isolates that had previously been collected primarily from mainland peninsular Malaysia appeared to form a separate third genetically distinct subpopulation cluster by whole genome sequencing analysis (Assefa et al., 2015). The possibility of a third genetically distinct group was also forecasted from different studies. Firstly, evidence of dimorphisms was revealed between peninsular Malaysia and Borneo using unlinked genes encoding the normocyte binding protein (Ahmed et al., 2014, Pinheiro et al., 2015) and the Duffy binding protein (Fong et al., 2014). Secondly, analysis using the 18S rRNA and mitochondrial cytochrome oxidase subunit 1 between the two regions also suggested geographical genetic structure (Yusof et al., 2016). Moreover, in Chapter Two, significant divergence was seen between the two regions by the analysis of microsatellite data (Divis et al., 2015). When the population genetic structure was re-assessed by more samples from human and macaque hosts in peninsular Malaysia (Chapter Four), several independent and complementary statistical tests affirmed two sympatric divergent subpopulations exclusively exist in Malaysian Borneo as described before, while a separate

subpopulation (referred to as Cluster 3) occurs in peninsular Malaysia (Divis et al., 2017).

In order to understand evolution of malaria parasites, it may be relevant to explore the population history of the natural hosts. The evolution of the *Macaca* genus shows four divergent monophyletic groups, with macaques carrying *P. knowlesi* belonging to two of these: the *fascicularis* group (*M. fascicularis*) and the *silenus* group (*M. nemestrina*) (Fa and Lindburg, 1996, Li et al., 2009b). The *silenus* group first diverged from the *fascicularis-sinica* groups circa 5 million years ago (mya) during the Pliocene Epoch when the mainland Southeast Asia was still connected to Borneo and the other Greater Sunda islands (Outlaw and Voelker, 2007). This followed by the separation of the *fascicularis* and *sinica* groups circa 2.5 mya as the sea level rose, around the end of the Pliocene.

*M. nemestrina* separated from the rest of the *silenus* group around 1.5 to 2.6 mya, while the splitting of *M. fascicularis* into subgroups started around the same time (Liedigk et al., 2015, Ziegler et al., 2007, Smith et al., 2014) when the sea level fluctuated due to climatic oscillations (Outlaw and Voelker, 2007). Long-tailed macaques (*M. fascicularis fascicularis*) further diverged into two main clades in concordance with the geography of Southeast Asia, one being associated with the mainland Southeast Asia, Peninsular Malaysia and Sumatera while the other clade is associated with many other islands including Borneo, Java and the Philippines (Liedigk et al., 2015). It is possible that parasites maintained in these macaques underwent similar divergence process.

With the sufficient SNP information derived from the population genomic analysis of *P. knowlesi* (Assefa et al., 2015), development of allele-specific PCR assays for genotyping of different *P. knowlesi* subpopulations were focused on the sympatric subpopulations in Malaysian Borneo (Chapter Three). The high specificity in discriminating the two *P. knowlesi* subpopulations in Malaysia Borneo and not infections from peninsular Malaysia implies the need for more sequence data from the latter area in order to develop a similar allele-specific PCR assay to conveniently distinguish Cluster 3. The application of allele-specific PCR assays enables the identification of Cluster 1 and Cluster 2 subpopulations from field isolates throughout the sympatric distribution, without needing to perform population genetic analysis of multi-locus data on all of these samples.

Following this, study on genomic divergence between the two sympatric *P. knowlesi* subpopulations in Malaysian Borneo was further conducted (Chapter Five). Previous analysis indicated that Cluster 1 subpopulation has undergone long-term population expansion (Assefa et al., 2015). In this present study, long-term population growth was also observed for the Cluster 2 subpopulation, with even more negatively skewed allele frequency spectra. Each subpopulation exhibits unique genomic patterns across all 14 chromosomes with limited gene flow between the subpopulation, further supporting the occurrence of independent zoonoses associated with different macaque host species (Muehlenbein et al., 2015).

This study also illustrates the landscape of genomic divergence between the sympatric subpopulations (Chapter Five). A large number of SNPs showed complete fixation, a

condition that is not seen in *P. falciparum* (Duffy et al., 2015, Mobegi et al., 2014) or *P. vivax* (Hupalo et al., 2016) elsewhere. Despite the differences at the genomic level between the two sympatric subpopulations in Malaysian Borneo, it is not known whether each subpopulation exhibit significant phenotypic differences. Conducting detailed clinical studies on patients infected with each parasite subpopulation type would potentially provide insights, since knowlesi malaria shows a wide spectrum from mild to fatality (Cox-Singh and Singh, 2008, Cox-Singh et al., 2010, Daneshvar et al., 2009, Rajahram et al., 2012, William et al., 2011). Furthermore, high parasitaemia has been consistently associated with disease severity in knowlesi malaria, and determining which subpopulation tends to exhibit high parasitaemia might provide an indication of disease risk.

As clinical manifestations are associated with the erythrocytic cycle of malaria parasites, it is relevant to note that different *Plasmodium* species shows preference in types of red blood cells during erythrocytic invasion. Clinical isolates of *P. vivax* show variations in their level of reticulocyte preference (Lim et al., 2016), and laboratory H strain of *P. knowlesi*, originated from peninsular Malaysia, also preferentially invade reticulocytes (Lim et al., 2013). As the overall population structure of *P. knowlesi* demonstrates three divergent parasite clusters, it would be useful to explore whether these parasites have different erythrocyte preference for invasion, which potentially relates to parasitaemia variations and severity of disease. In a previous study, the dimorphism in the *P. knowlesi* normocyte binding protein (*Pknbp<sub>xa</sub>*) gene sequences in limited number of samples from Malaysian Borneo was considered to be potentially associated with disease severity, as indicated by parasitaemia (Ahmed et al., 2014). Similar samples were also used in this population genetic study (Appendix 4.1) (Divis et

al., 2017) and the reanalysis of the genomic data (Chapter Five), but more detailed analysis of clinical phenotypes will be required to investigate whether the Cluster 1 subpopulation is associated with different virulence compared with the Cluster 2 subpopulation.

There is no clear evidence of direct human-mosquito-human transmission yet. The microsatellite and whole genome sequence data in this study are not in themselves suitable to test whether such a transmission route occurs occasionally. Although this route is possible under experimental conditions (Coatney et al., 1971), it is still difficult to demonstrate this in nature. Patients with knowlesi malaria presented at hospitals are mostly adult farmers and logging camp workers whose work drives them into forests or forest fringes regularly, or travellers entering the risk areas (Singh and Daneshvar, 2013), suggesting that macaque-vector-human transmission is the preferred route (Imai et al., 2014). Since infection in humans is considered accidental and a dead-end of the parasite's life cycle as a result of quick diagnosis and treatment, it is unlikely that humans would commonly act as the host source of potential inter-strain recombinants, although it could occasionally happen.

Since *P. knowlesi* infection in humans is primarily zoonotic, the malaria elimination programme in Malaysia focused mainly on human parasites (*P. falciparum*, *P. vivax* and *P. malariae*) with zero local transmission having been aimed for the Peninsular Malaysia by 2015 and by 2020 for Malaysian Borneo (World Health Organization, 2016, World Health Organization, 2015). Although *P. knowlesi* is not considered as part of this elimination programme, the situation is closely monitored due to the increased incidence particularly in the Malaysian Borneo regions (William et al., 2013, William et

al., 2014). The interruption of *P. knowlesi* transmission would be challenging since it primarily involves wild macaques, together with various factors such as mosquito vectors, ecology and geography. The existence of three groups of *P. knowlesi* (Divis et al., 2017) may complicate the control of this parasite as it requires new strategies by the Malaysia's Ministry of Health. Nonetheless, as part of monitoring the infections, all infected patients will be treated as for *P. falciparum* infections (World Health Organization, 2015).

Like many other vector-borne parasitic diseases, malaria is known for its sensitivity to environmental changes such as deforestation (Confalonieri et al., 2014). A recent report suggests that there is a link between environmental changes and incidence of *P. knowlesi* infections in Sabah state, Malaysian Borneo (Fornace et al., 2016). Moreover, long-tailed macaques and pig-tailed macaques show different habitat ranges in forested and non-forested areas (Moyes et al., 2016), suggesting the possibility to identify hot spots of infection for each subpopulation cluster, and the need to examine changes over time. Monitoring overall patterns of infections, and potential changes ranging from the parasite genomic level to clinical manifestations, in addition to vector studies, would be relevant components of surveillance and public health research on *knowlesi* malaria.

## References

- ABDULLAH, N. R., BARBER, B. E., WILLIAM, T., NORAHMAD, N. A., SATSU, U. R., MUNIANDY, P. K., ISMAIL, Z., GRIGG, M. J., JELIP, J., PIERA, K., VON SEIDLEIN, L., YEO, T. W., ANSTEY, N. M., PRICE, R. N. & AUBURN, S. 2013. *Plasmodium vivax* Population Structure and Transmission Dynamics in Sabah Malaysia. *PLoS One*, 8, e82553.
- AHMED, A. M., PINHEIRO, M. M., DIVIS, P. C., Siner, A., ZAINUDIN, R., WONG, I. T., LU, C. W., SINGH-KHAIRA, S. K., MILLAR, S. B., LYNCH, S., WILLMANN, M., SINGH, B., KRISHNA, S. & COX-SINGH, J. 2014. Disease Progression in *Plasmodium knowlesi* Malaria Is Linked to Variation in Invasion Gene Family Members. *PLoS Negl Trop Dis*, 8, e3086.
- AHMED, M. A., FONG, M. Y., LAU, Y. L. & YUSOF, R. 2016. Clustering and genetic differentiation of the normocyte binding protein (nbp<sub>xa</sub>) of *Plasmodium knowlesi* clinical isolates from Peninsular Malaysia and Malaysia Borneo. *Malar J*, 15, 241.
- ANDERSON, T. J., HAUBOLD, B., WILLIAMS, J. T., ESTRADA-FRANCO, J. G., RICHARDSON, L., MOLLINEDO, R., BOCKARIE, M., MOKILI, J., MHARAKURWA, S., FRENCH, N., WHITWORTH, J., VELEZ, I. D., BROCKMAN, A. H., NOSTEN, F., FERREIRA, M. U. & DAY, K. P. 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol*, 17, 1467-82.
- ANTHONY, T. G., CONWAY, D. J., COX-SINGH, J., MATUSOP, A., RATNAM, S., SHAMSUL, S. & SINGH, B. 2005. Fragmented population structure of *Plasmodium falciparum* in a region of declining endemicity. *Journal of Infectious Diseases*, 191, 1558-1564.
- ANTINORI, S., GALIMBERTI, L., MILAZZO, L. & CORBELLINO, M. 2013. *Plasmodium knowlesi*: the emerging zoonotic malaria parasite. *Acta Trop*, 125, 191-201.
- ASSEFA, S., LIM, C., PRESTON, M. D., DUFFY, C. W., NAIR, M. B., ADROUB, S. A., KADIR, K. A., GOLDBERG, J. M., NEAFSEY, D. E., DIVIS, P., CLARK, T. G., DURASINGH, M. T., CONWAY, D. J., PAIN, A. & SINGH, B. 2015. Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. *Proc Natl Acad Sci U S A*.

- ATROOSH, W. M., AL-MEKHLAFI, H. M., MAHDY, M. A., SAIF-ALI, R., AL-MEKHLAFI, A. M. & SURIN, J. 2011. Genetic diversity of *Plasmodium falciparum* isolates from Pahang, Malaysia based on MSP-1 and MSP-2 genes. *Parasit Vectors*, 4, 233.
- BALLOUX, F. & LUGON-MOULIN, N. 2002. The estimation of population differentiation with microsatellite markers. *Mol Ecol*, 11, 155-65.
- BARBER, B. E., WILLIAM, T., DHARARAJ, P., ANDERIOS, F., GRIGG, M. J., YEO, T. W. & ANSTEY, N. M. 2012. Epidemiology of *Plasmodium knowlesi* malaria in north-east Sabah, Malaysia: family clusters and wide age distribution. *Malar J*, 11, 401.
- BARBER, B. E., WILLIAM, T., GRIGG, M. J., MENON, J., AUBURN, S., MARFURT, J., ANSTEY, N. M. & YEO, T. W. 2013. A prospective comparative study of *knowlesi*, *falciparum*, and *vivax* malaria in Sabah, Malaysia: high proportion with severe disease from *Plasmodium knowlesi* and *Plasmodium vivax* but no mortality with early referral and artesunate therapy. *Clin Infect Dis*, 56, 383-97.
- BARTOLONI, A. & ZAMMARCHI, L. 2012. Clinical aspects of uncomplicated and severe malaria. *Mediterr J Hematol Infect Dis*, 4, e2012026.
- BERG, P. R., JENTOFT, S., STAR, B., RING, K. H., KNUTSEN, H., LIEN, S., JAKOBSEN, K. S. & ANDRE, C. 2015. Adaptation to Low Salinity Promotes Genomic Divergence in Atlantic Cod (*Gadus morhua* L.). *Genome Biol Evol*, 7, 1644-63.
- BERRY, A., IRIART, X., WILHELM, N., VALENTIN, A., CASSAING, S., WITKOWSKI, B., BENOIT-VICAL, F., MENARD, S., OLAGNIER, D., FILLAUX, J., SIRE, S., LE COUSTUMIER, A. & MAGNAVAL, J. F. 2011. Imported *Plasmodium knowlesi* malaria in a French tourist returning from Thailand. *Am J Trop Med Hyg*, 84, 535-8.
- BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114-20.
- BRONNER, U., DIVIS, P. C., FARNERT, A. & SINGH, B. 2009. Swedish traveller with *Plasmodium knowlesi* malaria after visiting Malaysian Borneo. *Malar J*, 8, 15.
- CARLTON, J. M., ADAMS, J. H., SILVA, J. C., BIDWELL, S. L., LORENZI, H., CALER, E., CRABTREE, J., ANGIUOLI, S. V., MERINO, E. F., AMEDEO, P., CHENG, Q., COULSON, R. M. R., CRABB, B. S., DEL PORTILLO, H. A., ESSIEN, K., FELDBLYUM, T. V., FERNANDEZ-BECERRA, C., GILSON, P. R., GUEYE, A. H., GUO, X., KANG'A, S., KOUIJ, T. W. A., KORSINCZKY, M., MEYER, E. V. S., NENE, V., PAULSEN, I.,



- WHITE, O., RALPH, S. A., REN, Q. H., SARGEANT, T. J., SALZBERG, S. L., STOECKERT, C. J., SULLIVAN, S. A., YAMAMOTO, M. M., HOFFMAN, S. L., WORTMAN, J. R., GARDNER, M. J., GALINSKI, M. R., BARNWELL, J. W. & FRASER-LIGGETT, C. M. 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455, 757-763.
- CHANG, H. H., MOSS, E. L., PARK, D. J., NDIAYE, D., MBOUP, S., VOLKMAN, S. K., SABETI, P. C., WIRTH, D. F., NEAFSEY, D. E. & HARTL, D. L. 2013. Malaria life cycle intensifies both natural selection and random genetic drift. *Proc Natl Acad Sci U S A*, 110, 20129-34.
- CHENUIL, A. 2006. Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica*, 127, 101-20.
- CHIN, W., CONTACOS, P. G., COATNEY, G. R. & KIMBALL, H. R. 1965. A Naturally Acquired Quotidian-Type Malaria in Man Transferable to Monkeys. *Science*, 149, 865.
- CLEMENT, M., POSADA, D. & CRANDALL, K. A. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol*, 9, 1657-9.
- COATNEY, G. R., COLLIN, W. E., WARREN, M. & CONTACOS, P. G. 1971. *The Primate Malarias*, Washington, U.S. Government Printing Office.
- CONFALONIERI, U. E., MARGONARI, C. & QUINTAO, A. F. 2014. Environmental change and the dynamics of parasitic diseases in the Amazon. *Acta Trop*, 129, 33-41.
- CORDINA, C. J., CULLETON, R., JONES, B. L., SMITH, C. C., MACCONNACHIE, A. A., COYNE, M. J. & ALEXANDER, C. L. 2014. *Plasmodium knowlesi*: Clinical Presentation and Laboratory Diagnosis of the First Human Case in a Scottish Traveler. *J Travel Med*.
- COX-SINGH, J. 2012. Zoonotic malaria: *Plasmodium knowlesi*, an emerging pathogen. *Curr Opin Infect Dis*, 25, 530-6.
- COX-SINGH, J., DAVIS, T. M., LEE, K. S., SHAMSUL, S. S., MATUSOP, A., RATNAM, S., RAHMAN, H. A., CONWAY, D. J. & SINGH, B. 2008. *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis*, 46, 165-71.
- COX-SINGH, J., HIU, J., LUCAS, S. B., DIVIS, P. C., ZULKARNAEN, M., CHANDRAN, P., WONG, K. T., ADEM, P., ZAKI, S. R., SINGH, B. & KRISHNA, S. 2010. Severe

- malaria - a case of fatal *Plasmodium knowlesi* infection with post-mortem findings: a case report. *Malar J*, 9, 10.
- COX-SINGH, J. & SINGH, B. 2008. Knowlesi malaria: newly emergent and of public health importance? *Trends Parasitol*, 24, 406-10.
- CULLETON, R. L. & ABKALLO, H. M. 2015. Malaria parasite genetics: doing something useful. *Parasitol Int*, 64, 244-53.
- DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GENOMES PROJECT ANALYSIS, G. 2011. The variant call format and VCFtools. *Bioinformatics*, 27, 2156-8.
- DANESHVAR, C., DAVIS, T. M., COX-SINGH, J., RAFA'EE, M. Z., ZAKARIA, S. K., DIVIS, P. C. & SINGH, B. 2009. Clinical and laboratory features of human *Plasmodium knowlesi* infection. *Clin Infect Dis*, 49, 852-60.
- DIVIS, P. C., LIN, L. C., ROVIE-RYAN, J. J., KADIR, K. A., ANDERIOS, F., HISAM, S., SHARMA, R. S. K., SINGH, B. & CONWAY, D. J. 2017. Three divergent subpopulations of the malaria parasite *Plasmodium knowlesi*. *Emerg Infect Dis*, 23.
- DIVIS, P. C., SINGH, B., ANDERIOS, F., HISAM, S., MATUSOP, A., KOCKEN, C. H., ASSEFA, S. A., DUFFY, C. W. & CONWAY, D. J. 2015. Admixture in Humans of Two Divergent *Plasmodium knowlesi* Populations Associated with Different Macaque Host Species. *PLoS Pathog*, 11, e1004888.
- DUFFY, C. W., ASSEFA, S. A., ABUGRI, J., AMOAKO, N., OWUSU-AGYEI, S., ANYORIGIYA, T., MACINNIS, B., KWIATKOWSKI, D. P., CONWAY, D. J. & AWANDARE, G. A. 2015. Comparison of genomic signatures of selection on *Plasmodium falciparum* between different regions of a country with high malaria endemicity. *BMC Genomics*, 16, 527.
- EARL, D. A. & VONHOLDT, B. M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4, 359-361.
- ELLEGREN, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*, 5, 435-45.
- ESSELSTYN, J. A., WIDMANN, P. & HEANEY, L. R. 2004. The mammals of Palawan Island, Philippines. *Proc Biol Soc Wash*, 117, 271-302.

- EVANNO, G., REGNAUT, S. & GOUDET, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 14, 2611-20.
- FA, J. E. & LINDBURG, D. G. 1996. *Evolution and Ecology of Macaque Societies*, Cambridge University Press.
- FAIRCLOTH, B. C. 2008. MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol Ecol Resour*, 8, 92-4.
- FARIA, N. R., RAMBAUT, A., SUCHARD, M. A., BAELE, G., BEDFORD, T., WARD, M. J., TATEM, A. J., SOUSA, J. D., ARINAMINPATHY, N., PEPIN, J., POSADA, D., PEETERS, M., PYBUS, O. G. & LEMEY, P. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346, 56-61.
- FIGTREE, M., LEE, R., BAIN, L., KENNEDY, T., MACKERTICH, S., URBAN, M., CHENG, Q. & HUDSON, B. J. 2010. *Plasmodium knowlesi* in human, Indonesian Borneo. *Emerg Infect Dis*, 16, 672-4.
- FONG, M. Y., LAU, Y. L., CHANG, P. Y. & ANTHONY, C. N. 2014. Genetic diversity, haplotypes and allele groups of Duffy binding protein (PkDBPalphall) of *Plasmodium knowlesi* clinical isolates from Peninsular Malaysia. *Parasit Vectors*, 7, 161.
- FONG, M. Y., RASHDI, S. A., YUSOF, R. & LAU, Y. L. 2015. Distinct genetic difference between the Duffy binding protein (PkDBPalphall) of *Plasmodium knowlesi* clinical isolates from North Borneo and Peninsular Malaysia. *Malar J*, 14, 91.
- FONG, Y. L., CADIGAN, F. C. & COATNEY, G. R. 1971. A presumptive case of naturally occurring *Plasmodium knowlesi* malaria in man in Malaysia. *Trans R Soc Trop Med Hyg*, 65, 839-840.
- FORNACE, K. M., ABIDIN, T. R., ALEXANDER, N., BROCK, P., GRIGG, M. J., MURPHY, A., WILLIAM, T., MENON, J., DRAKELEY, C. J. & COX, J. 2016. Association between Landscape Factors and Spatial Patterns of *Plasmodium knowlesi* Infections in Sabah, Malaysia. *Emerg Infect Dis*, 22, 201-8.
- FORNACE, K. M., NUIN, N. A., BETSON, M., GRIGG, M. J., WILLIAM, T., ANSTEY, N. M., YEO, T. W., COX, J., YING, L. T. & DRAKELEY, C. J. 2015. Asymptomatic and submicroscopic carriage of *Plasmodium knowlesi* malaria in household and

- community members of clinical cases in Sabah, Malaysia. *J Infect Dis*, 213, 784-7.
- FOSTER, D., COX-SINGH, J., MOHAMAD, D. S., KRISHNA, S., CHIN, P. P. & SINGH, B. 2014. Evaluation of three rapid diagnostic tests for the detection of human infections with *Plasmodium knowlesi*. *Malar J*, 13, 60.
- GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R. W., CARLTON, J. M., PAIN, A., NELSON, K. E., BOWMAN, S., PAULSEN, I. T., JAMES, K., EISEN, J. A., RUTHERFORD, K., SALZBERG, S. L., CRAIG, A., KYES, S., CHAN, M. S., NENE, V., SHALLOM, S. J., SUH, B., PETERSON, J., ANGIUOLI, S., PERTEA, M., ALLEN, J., SELENGUT, J., HAFT, D., MATHER, M. W., VAIDYA, A. B., MARTIN, D. M., FAIRLAMB, A. H., FRAUNHOLZ, M. J., ROOS, D. S., RALPH, S. A., MCFADDEN, G. I., CUMMINGS, L. M., SUBRAMANIAN, G. M., MUNGALL, C., VENTER, J. C., CARUCCI, D. J., HOFFMAN, S. L., NEWBOLD, C., DAVIS, R. W., FRASER, C. M. & BARRELL, B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419, 498-511.
- GARNHAM, P. C. C. 1966. *Malaria parasites and other haemosporidia*, Oxford, Blackwell Scientific Publication.
- GHINAI, I., COOK, J., HLA, T. T., HTET, H. M., HALL, T., LUBIS, I. N., GHINAI, R., HESKETH, T., NAUNG, Y., LWIN, M. M., LATT, T. S., HEYMANN, D. L., SUTHERLAND, C. J., DRAKELEY, C. & FIELD, N. 2017. Malaria epidemiology in central Myanmar: identification of a multi-species asymptomatic reservoir of infection. *Malar J*, 16, 16.
- GIRE, S. K., GOBA, A., ANDERSEN, K. G., SEALFON, R. S., PARK, D. J., KANNEH, L., JALLOH, S., MOMOH, M., FULLAH, M., DUDAS, G., WOHL, S., MOSES, L. M., YOZWIAK, N. L., WINNICKI, S., MATRANGA, C. B., MALBOEUF, C. M., QU, J., GLADDEN, A. D., SCHAFFNER, S. F., YANG, X., JIANG, P. P., NEKOUI, M., COLUBRI, A., COOMBER, M. R., FONNIE, M., MOIGBOI, A., GBAKIE, M., KAMARA, F. K., TUCKER, V., KONUWA, E., SAFFA, S., SELLU, J., JALLOH, A. A., KOVOMA, A., KONINGA, J., MUSTAPHA, I., KARGBO, K., FODAY, M., YILLAH, M., KANNEH, F., ROBERT, W., MASSALLY, J. L., CHAPMAN, S. B., BOCHICCHIO, J., MURPHY, C., NUSBAUM, C., YOUNG, S., BIRREN, B. W., GRANT, D. S., SCHEIFFELIN, J. S., LANDER, E. S., HAPPI, C., GEVAO, S. M., GNIRKE, A., RAMBAUT, A., GARRY, R. F., KHAN, S. H. & SABETI, P. C. 2014. Genomic

- surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345, 1369-72.
- GOODHEAD, I., CAPEWELL, P., BAILEY, J. W., BEAMENT, T., CHANCE, M., KAY, S., FORRESTER, S., MACLEOD, A., TAYLOR, M., NOYES, H. & HALL, N. 2013. Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. *MBio*, 4.
- GOUDET, J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity*, 86, 485-486.
- GUICHOUX, E., LAGACHE, L., WAGNER, S., CHAUMEIL, P., LEGER, P., LEPAIS, O., LEPOITTEVIN, C., MALAUSA, T., REVARDEL, E., SALIN, F. & PETIT, R. J. 2011. Current trends in microsatellite genotyping. *Mol Ecol Resour*, 11, 591-611.
- GUO, B., DEFAVERI, J., SOTELO, G., NAIR, A. & MERILA, J. 2015. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biol*, 13, 19.
- HAMID, M. M., REMARQUE, E. J., EL HASSAN, I. M., HUSSAIN, A. A., NARUM, D. L., THOMAS, A. W., KOCKEN, C. H., WEISS, W. R. & FABER, B. W. 2011. Malaria infection by sporozoite challenge induces high functional antibody titres against blood stage antigens after a DNA prime, poxvirus boost vaccination strategy in Rhesus macaques. *Malar J*, 10, 29.
- HAUBOLD, B. & HUDSON, R. R. 2000. LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics*, 16, 847-848.
- HUBISZ, M. J., FALUSH, D., STEPHENS, M. & PRITCHARD, J. K. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour*, 9, 1322-32.
- HUPALO, D. N., LUO, Z., MELNIKOV, A., SUTTON, P. L., ROGOV, P., ESCALANTE, A., VALLEJO, A. F., HERRERA, S., AREVALO-HERRERA, M., FAN, Q., WANG, Y., CUI, L., LUCAS, C. M., DURAND, S., SANCHEZ, J. F., BALDEVIANO, G. C., LESCANO, A. G., LAMAN, M., BARNADAS, C., BARRY, A., MUELLER, I., KAZURA, J. W., EAPEN, A., KANAGARAJ, D., VALECHA, N., FERREIRA, M. U., ROOBOSONG, W., NGUITRAGOOL, W., SATTABONKOT, J., GAMBOA, D., KOSEK, M., VINETZ, J. M., GONZALEZ-CERON, L., BIRREN, B. W., NEAFSEY, D. E. & CARLTON, J. M. 2016. Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat Genet*.

- IMAI, N., WHITE, M. T., GHANI, A. C. & DRAKELEY, C. J. 2014. Transmission and Control of *Plasmodium knowlesi*: A Mathematical Modelling Study. *PLoS Negl Trop Dis*, 8, e2978.
- IMWONG, M., NAIR, S., PUKRITTAYAKAMEE, S., SUDIMACK, D., WILLIAMS, J. T., MAYXAY, M., NEWTON, P. N., KIM, J. R., NANDY, A., OSORIO, L., CARLTON, J. M., WHITE, N. J., DAY, N. P. J. & ANDERSON, T. J. C. 2007. Contrasting genetic structure in *Plasmodium vivax* populations from Asia and south America. *International Journal for Parasitology*, 37, 1013-1022.
- JAKOBSSON, M. & ROSENBERG, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23, 1801-6.
- JARNE, P. & LAGODA, P. J. L. 1996. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, 11, 424-429.
- JESLYN, W. P., HUAT, T. C., VERNON, L., IRENE, L. M., SUNG, L. K., JARROD, L. P., SINGH, B. & CHING, N. L. 2011. Molecular epidemiological investigation of *Plasmodium knowlesi* in humans and macaques in Singapore. *Vector Borne Zoonotic Dis*, 11, 131-5.
- JIANG, N., CHANG, Q., SUN, X., LU, H., YIN, J., ZHANG, Z., WAHLGREN, M. & CHEN, Q. 2010. Co-infections with *Plasmodium knowlesi* and other malaria parasites, Myanmar. *Emerg Infect Dis*, 16, 1476-8.
- JIMOH, A., SOFOLA, O., PETU, A. & OKOROSOBO, T. 2007. Quantifying the economic burden of malaria in Nigeria using the willingness to pay approach. *Cost Eff Resour Alloc*, 5, 6.
- JIRAM, A. I., VYTHILINGAM, I., NOORAZIAN, Y. M., YUSOF, Y. M., AZAHARI, A. H. & FONG, M. Y. 2012. Entomologic investigation of *Plasmodium knowlesi* vectors in Kuala Lipis, Pahang, Malaysia. *Malar J*, 11, 213.
- JOMBART, T., DEVILLARD, S. & BALLOUX, F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*, 11, 94.
- JONGWUTIWES, S., BUPPAN, P., KOSUVIN, R., SEETHAMCHAI, S., PATTANAWONG, U., SIRICHAISINTHOP, J. & PUTAPORNTIP, C. 2011. *Plasmodium knowlesi* Malaria in humans and macaques, Thailand. *Emerg Infect Dis*, 17, 1799-806.

- JONGWUTIWES, S., PUTAPORNTIP, C., IWASAKI, T., FERREIRA, M. U., KANBARA, H. & HUGHES, A. L. 2005. Mitochondrial genome sequences support ancient population expansion in *Plasmodium vivax*. *Mol Biol Evol*, 22, 1733-9.
- JONGWUTIWES, S., PUTAPORNTIP, C., IWASAKI, T., SATA, T. & KANBARA, H. 2004. Naturally acquired *Plasmodium knowlesi* malaria in human, Thailand. *Emerg Infect Dis*, 10, 2211-3.
- JOY, D. A., GONZALEZ-CERON, L., CARLTON, J. M., GUEYE, A., FAY, M., MCCUTCHAN, T. F. & SU, X. Z. 2008. Local adaptation and vector-mediated population structure in *Plasmodium vivax* malaria. *Molecular Biology and Evolution*, 25, 1245-1252.
- KANTELE, A., MARTI, H., FELGER, I., MULLER, D. & JOKIRANTA, T. S. 2008. Monkey malaria in a European traveler returning from Malaysia. *Emerg Infect Dis*, 14, 1434-6.
- KHIM, N., SIV, S., KIM, S., MUELLER, T., FLEISCHMANN, E., SINGH, B., DIVIS, P. C., STEENKESTE, N., DUVAL, L., BOUCHIER, C., DUONG, S., ARIEY, F. & MENARD, D. 2011. *Plasmodium knowlesi* infection in humans, Cambodia, 2007-2010. *Emerg Infect Dis*, 17, 1900-2.
- KOCKEN, C. H., OZWARA, H., VAN DER WEL, A., BEETSMA, A. L., MWENDA, J. M. & THOMAS, A. W. 2002. *Plasmodium knowlesi* provides a rapid in vitro and in vivo transfection system that enables double-crossover gene knockout studies. *Infect Immun*, 70, 655-60.
- KOLPAKOV, R., BANA, G. & KUCHEROV, G. 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res*, 31, 3672-8.
- KORBIE, D. J. & MATTICK, J. S. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat Protoc*, 3, 1452-6.
- KOSUWIN, R., PUTAPORNTIP, C., TACHIBANA, H. & JONGWUTIWES, S. 2014. Spatial variation in genetic diversity and natural selection on the thrombospondin-related adhesive protein locus of *Plasmodium vivax* (PvTRAP). *PLoS One*, 9, e110463.
- LAPP, S. A., KORIR-MORRISON, C., JIANG, J., BAI, Y., CORREDOR, V. & GALINSKI, M. R. 2013. Spleen-dependent regulation of antigenic variation in malaria parasites: *Plasmodium knowlesi* SICAvax expression profiles in splenic and asplenic hosts. *PLoS One*, 8, e78014.

- LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-8.
- LEE, K. S., COX-SINGH, J., BROOKE, G., MATUSOP, A. & SINGH, B. 2009a. *Plasmodium knowlesi* from archival blood films: further evidence that human infections are widely distributed and not newly emergent in Malaysian Borneo. *Int J Parasitol*, 39, 1125-8.
- LEE, K. S., COX-SINGH, J. & SINGH, B. 2009b. Morphological features and differential counts of *Plasmodium knowlesi* parasites in naturally acquired human infections. *Malar J*, 8, 73.
- LEE, K. S., DIVIS, P. C., ZAKARIA, S. K., MATUSOP, A., JULIN, R. A., CONWAY, D. J., COX-SINGH, J. & SINGH, B. 2011. *Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog*, 7, e1002015.
- LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXivorg*, arXiv:1303.3997v2 [q-bio.GN].
- LI, H., HANDSAKER, B., WYSOKER, A., FENNEL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9.
- LI, J., HAN, K., XING, J., KIM, H. S., ROGERS, J., RYDER, O. A., DISOTELL, T., YUE, B. & BATZER, M. A. 2009b. Phylogeny of the macaques (Cercopithecidae: Macaca) based on Alu elements. *Gene*, 448, 242-9.
- LIBRADO, P. & ROZAS, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25, 1451-2.
- LIEDIGK, R., KOLLECK, J., BOKER, K. O., MEIJAARD, E., MD-ZAIN, B. M., ABDUL-LATIFF, M. A., AMPENG, A., LAKIM, M., ABDUL-PATAH, P., TOSI, A. J., BRAMEIER, M., ZINNER, D. & ROOS, C. 2015. Mitogenomic phylogeny of the common long-tailed macaque (*Macaca fascicularis fascicularis*). *BMC Genomics*, 16, 222.
- LIM, C., HANSEN, E., DESIMONE, T. M., MORENO, Y., JUNKER, K., BEI, A., BRUGNARA, C., BUCKEE, C. O. & DURAISINGH, M. T. 2013. Expansion of host cellular niche can drive adaptation of a zoonotic malaria parasite to humans. *Nat Commun*, 4, 1638.



- LIM, C., PEREIRA, L., SALIBA, K. S., MASCARENHAS, A., MAKI, J. N., CHERY, L., GOMES, E., RATHOD, P. K. & DURAISINGH, M. T. 2016. Reticulocyte Preference and Stage Development of *Plasmodium vivax* Isolates. *J Infect Dis*, 214, 1081-4.
- LIM, L. & MCFADDEN, G. I. 2010. The evolution, metabolism and functions of the apicoplast. *Philos Trans R Soc Lond B Biol Sci*, 365, 749-63.
- LINK, L., BART, A., VERHAAR, N., VAN GOOL, T., PRONK, M. & SCHARNHORST, V. 2012. Molecular detection of *Plasmodium knowlesi* in a Dutch traveler by real-time PCR. *J Clin Microbiol*, 50, 2523-4.
- LIU, W., LI, Y., LEARN, G. H., RUDICELL, R. S., ROBERTSON, J. D., KEELE, B. F., NDJANGO, J. B., SANZ, C. M., MORGAN, D. B., LOCATELLI, S., GONDER, M. K., KRANZUSCH, P. J., WALSH, P. D., DELAPORTE, E., MPOUDI-NGOLE, E., GEORGIEV, A. V., MULLER, M. N., SHAW, G. M., PEETERS, M., SHARP, P. M., RAYNER, J. C. & HAHN, B. H. 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature*, 467, 420-5.
- LIU, W., LI, Y., SHAW, K. S., LEARN, G. H., PLENDERLEITH, L. J., MALENKE, J. A., SUNDARARAMAN, S. A., RAMIREZ, M. A., CRYSTAL, P. A., SMITH, A. G., BIBOLLET-RUCHE, F., AYOUBA, A., LOCATELLI, S., ESTEBAN, A., MOUACHA, F., GUICHET, E., BUTEL, C., AHUKA-MUNDEKE, S., INOGWABINI, B. I., NDJANGO, J. B., SPEEDE, S., SANZ, C. M., MORGAN, D. B., GONDER, M. K., KRANZUSCH, P. J., WALSH, P. D., GEORGIEV, A. V., MULLER, M. N., PIEL, A. K., STEWART, F. A., WILSON, M. L., PUSEY, A. E., CUI, L., WANG, Z., FARNERT, A., SUTHERLAND, C. J., NOLDER, D., HART, J. A., HART, T. B., BERTOLANI, P., GILLIS, A., LEBRETON, M., TAFON, B., KIYANG, J., DJOKO, C. F., SCHNEIDER, B. S., WOLFE, N. D., MPOUDI-NGOLE, E., DELAPORTE, E., CARTER, R., CULLETON, R. L., SHAW, G. M., RAYNER, J. C., PEETERS, M., HAHN, B. H. & SHARP, P. M. 2014. African origin of the malaria parasite *Plasmodium vivax*. *Nat Commun*, 5, 3346.
- LIU, W., SUNDARARAMAN, S. A., LOY, D. E., LEARN, G. H., LI, Y., PLENDERLEITH, L. J., NDJANGO, J. N., SPEEDE, S., ATENCIA, R., COX, D., SHAW, G. M., AYOUBA, A., PEETERS, M., RAYNER, J. C., HAHN, B. H. & SHARP, P. M. 2016. Multigenomic Delineation of *Plasmodium* Species of the *Laverania* Subgenus Infecting Wild-living Chimpanzees and Gorillas. *Genome Biol Evol*.

- LUBIS, I. N., WIJAYA, H., LUBIS, M., LUBIS, C. P., DIVIS, P. C., BESHIR, K. B. & SUTHERLAND, C. J. 2017. Contribution of *Plasmodium knowlesi* to multi-species human malaria infections in North Sumatera, Indonesia. *J Infect Dis*.
- LUCHAVEZ, J., ESPINO, F., CURAMENG, P., ESPINA, R., BELL, D., CHIODINI, P., NOLDER, D., SUTHERLAND, C., LEE, K. S. & SINGH, B. 2008. Human Infections with *Plasmodium knowlesi*, the Philippines. *Emerg Infect Dis*, 14, 811-3.
- MALINSKY, M., CHALLIS, R. J., TYERS, A. M., SCHIFFELS, S., TERAJ, Y., NGATUNGA, B. P., MISKA, E. A., DURBIN, R., GENNER, M. J. & TURNER, G. F. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 350, 1493-8.
- MARCHAND, R. P., CULLETON, R., MAENO, Y., QUANG, N. T. & NAKAZAWA, S. 2011. Co-infections of *Plasmodium knowlesi*, *P. falciparum*, and *P. vivax* among Humans and *Anopheles dirus* Mosquitoes, Southern Vietnam. *Emerg Infect Dis*, 17, 1232-9.
- MARQUES, D. A., LUCEK, K., MEIER, J. I., MWAIKO, S., WAGNER, C. E., EXCOFFIER, L. & SEEHAUSEN, O. 2016. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLoS Genet*, 12, e1005887.
- MEIJAARD, E. 2003. Mammals of South-East Asian Islands and their Late Pleistocene environments. *J Biogeog*, 30, 1245-1257.
- MEIRMANS, P. G. & HEDRICK, P. W. 2011. Assessing population structure: F(ST) and related measures. *Mol Ecol Resour*, 11, 5-18.
- MENARD, R., TAVARES, J., COCKBURN, I., MARKUS, M., ZAVALA, F. & AMINO, R. 2013. Looking under the skin: the first steps in malarial infection and immunity. *Nat Rev Microbiol*, 11, 701-12.
- MIAO, M., YANG, Z. Q., PATCH, H., HUANG, Y. M., ESCALANTE, A. A. & CUI, L. W. 2012. *Plasmodium vivax* populations revisited: mitochondrial genomes of temperate strains in Asia suggest ancient population expansion. *Bmc Evolutionary Biology*, 12.
- MIOTTO, O., ALMAGRO-GARCIA, J., MANSKE, M., MACINNIS, B., CAMPINO, S., ROCKETT, K. A., AMARATUNGA, C., LIM, P., SUON, S., SRENG, S., ANDERSON, J. M., DUONG, S., NGUON, C., CHUOR, C. M., SAUNDERS, D., SE, Y., LON, C., FUKUDA, M. M., AMENGA-ETEGO, L., HODGSON, A. V., ASOALA, V., IMWONG, M., TAKALA-HARRISON, S., NOSTEN, F., SU, X. Z., RINGWALD, P., ARIEY, F.,

- DOLECEK, C., HIEN, T. T., BONI, M. F., THAI, C. Q., AMAMBUA-NGWA, A., CONWAY, D. J., DJIMDE, A. A., DOUMBO, O. K., ZONGO, I., OUEDRAOGO, J. B., ALCOCK, D., DRURY, E., AUBURN, S., KOCH, O., SANDERS, M., HUBBART, C., MASLEN, G., RUANO-RUBIO, V., JYOTHI, D., MILES, A., O'BRIEN, J., GAMBLE, C., OYOLA, S. O., RAYNER, J. C., NEWBOLD, C. I., BERRIMAN, M., SPENCER, C. C., MCVEAN, G., DAY, N. P., WHITE, N. J., BETHELL, D., DONDORP, A. M., PLOWE, C. V., FAIRHURST, R. M. & KWIATKOWSKI, D. P. 2013. Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat Genet*, 45, 648-55.
- MOBEGI, V. A., DUFFY, C. W., AMAMBUA-NGWA, A., LOUA, K. M., LAMAN, E., NWAKANMA, D. C., MACINNIS, B., ASPELING-JONES, H., MURRAY, L., CLARK, T. G., KWIATKOWSKI, D. P. & CONWAY, D. J. 2014. Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Mol Biol Evol*, 31, 1490-9.
- MOBEGI, V. A., LOUA, K. M., AHOUIDI, A. D., SATOGUINA, J., NWAKANMA, D. C., AMAMBUA-NGWA, A. & CONWAY, D. J. 2012. Population genetic structure of *Plasmodium falciparum* across a region of diverse endemicity in West Africa. *Malar J*, 11, 223.
- MOHD ABD RAZAK, M. R., SASTU, U. R., NORAHMAD, N. A., ABDUL-KARIM, A., MUHAMMAD, A., MUNIANDY, P. K., JELIP, J., RUNDI, C., IMWONG, M., MUDIN, R. N. & ABDULLAH, N. R. 2016. Genetic Diversity of *Plasmodium falciparum* Populations in Malaria Declining Areas of Sabah, East Malaysia. *PLoS One*, 11, e0152415.
- MOON, R. W., HALL, J., RANGKUTI, F., HO, Y. S., ALMOND, N., MITCHELL, G. H., PAIN, A., HOLDER, A. A. & BLACKMAN, M. J. 2013. Adaptation of the genetically tractable malaria pathogen *Plasmodium knowlesi* to continuous culture in human erythrocytes. *Proc Natl Acad Sci U S A*, 110, 531-6.
- MOYES, C. L., SHEARER, F. M., HUANG, Z., WIEBE, A., GIBSON, H. S., NIJMAN, V., MOHD-AZLAN, J., BRODIE, J. F., MALAIVIJITNOND, S., LINKIE, M., SAMEJIMA, H., O'BRIEN, T. G., TRAINOR, C. R., HAMADA, Y., GIORDANO, A. J., KINNAIRD, M. F., ELYAZAR, I. R., SINKA, M. E., VYTHILINGAM, I., BANGS, M. J., PIGOTT, D. M., WEISS, D. J., GOLDING, N. & HAY, S. I. 2016. Predicting the geographical

- distributions of the macaque hosts and mosquito vectors of *Plasmodium knowlesi* malaria in forested and non-forested areas. *Parasit Vectors*, 9, 242.
- MUDUNURI, S. B. & NAGARAJARAM, H. A. 2007. IMEx: Imperfect Microsatellite Extractor. *Bioinformatics*, 23, 1181-7.
- MUEHLENBEIN, M. P., PACHECO, M. A., TAYLOR, J. E., PRALL, S. P., AMBU, L., NATHAN, S., ALSISTO, S., RAMIREZ, D. & ESCALANTE, A. A. 2015. Accelerated diversification of nonhuman primate malarias in southeast Asia: adaptive radiation or geographic speciation? *Mol Biol Evol*, 32, 422-39.
- MURPHY, J. R., WEISS, W. R., FRYAUFF, D., DOWLER, M., SAVRANSKY, T., STOYANOV, C., MURATOVA, O., LAMBERT, L., ORR-GONZALEZ, S., ZELESKI, K. L., HINDERER, J., FAY, M. P., JOSHI, G., GWADZ, R. W., RICHIE, T. L., VILLASANTE, E. F., RICHARDSON, J. H., DUFFY, P. E. & CHEN, J. 2014. Using infective mosquitoes to challenge monkeys with *Plasmodium knowlesi* in malaria vaccine studies. *Malar J*, 13, 215.
- NAIR, S., NKHOMA, S. C., SERRE, D., ZIMMERMAN, P. A., GORENA, K., DANIEL, B. J., NOSTEN, F., ANDERSON, T. J. & CHEESEMAN, I. H. 2014. Single-cell genomics for dissection of complex malaria infections. *Genome Res*.
- NG, O. T., OOI, E. E., LEE, C. C., LEE, P. J., NG, L. C., PEI, S. W., TU, T. M., LOH, J. P. & LEO, Y. S. 2008. Naturally acquired human *Plasmodium knowlesi* infection, Singapore. *Emerg Infect Dis*, 14, 814-6.
- OHASHI, J., SUZUKI, Y., NAKA, I., HANANANTACHAI, H. & PATARAPOTIKUL, J. 2014. Diversifying selection on the thrombospondin-related adhesive protein (TRAP) gene of *Plasmodium falciparum* in Thailand. *PLoS One*, 9, e90522.
- OLIVEIRA, E. J., PADUA, J. G., ZUCCHI, M. I., VENCOSKY, R. & VIEIRA, M. L. C. 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, 29, 294-307.
- OUTLAW, D. C. & VOELKER, G. 2007. Pliocene climatic change in insular Southeast Asia as an engine of diversification in *Ficedula* flycatchers. *J Biogeog*, doi:10.1111/j.1365-2699.2007.01821.x.
- PAIN, A., BOHME, U., BERRY, A. E., MUNGALL, K., FINN, R. D., JACKSON, A. P., MOURIER, T., MISTRY, J., PASINI, E. M., ASLETT, M. A., BALASUBRAMMANIAM, S., BORGWARDT, K., BROOKS, K., CARRET, C., CARVER, T. J., CHEREVACH, I., CHILLINGWORTH, T., CLARK, T. G., GALINSKI, M. R., HALL, N., HARPER, D.,

- HARRIS, D., HAUSER, H., IVENS, A., JANSSEN, C. S., KEANE, T., LARKE, N., LAPP, S., MARTI, M., MOULE, S., MEYER, I. M., ORMOND, D., PETERS, N., SANDERS, M., SANDERS, S., SARGEANT, T. J., SIMMONDS, M., SMITH, F., SQUARES, R., THURSTON, S., TIVEY, A. R., WALKER, D., WHITE, B., ZUIDERWIJK, E., CHURCHER, C., QUAIL, M. A., COWMAN, A. F., TURNER, C. M., RAJANDREAM, M. A., KOCKEN, C. H., THOMAS, A. W., NEWBOLD, C. I., BARRELL, B. G. & BERRIMAN, M. 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 455, 799-803.
- PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20, 289-90.
- PEAKALL, R. & SMOUSE, P. E. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Resources*, 6, 288-295.
- PINHEIRO, M. M., AHMED, M. A., MILLAR, S. B., SANDERSON, T., OTTO, T. D., LU, W. C., KRISHNA, S., RAYNER, J. C. & COX-SINGH, J. 2015. *Plasmodium knowlesi* Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism. *PLoS One*, 10, e0121303.
- PRADEL, G., GARAPATY, S. & FREVERT, U. 2004. Kupffer and stellate cell proteoglycans mediate malaria sporozoite targeting to the liver. *Comp Hepatol*, 3 Suppl 1, S47.
- PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-59.
- PUTAPORNTIP, C., HONGSRIMUANG, T., SEETHAMCHAI, S., KOBASA, T., LIMKITTIKUL, K., CUI, L. & JONGWUTIWES, S. 2009. Differential prevalence of *Plasmodium* infections and cryptic *Plasmodium knowlesi* malaria in humans in Thailand. *J Infect Dis*, 199, 1143-50.
- PUTAPORNTIP, C., KUAMSAB, N. & JONGWUTIWES, S. 2016. Sequence diversity and positive selection at the Duffy-binding protein genes of *Plasmodium knowlesi* and *P. cynomolgi*: Analysis of the complete coding sequences of Thai isolates. *Infect Genet Evol*, 44, 367-375.
- PUTAPORNTIP, C., THONGAREE, S. & JONGWUTIWES, S. 2013. Differential sequence diversity at merozoite surface protein-1 locus of *Plasmodium knowlesi* from humans and macaques in Thailand. *Infect Genet Evol*, 18, 213-9.

- QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841-2.
- RAJAHRAM, G. S., BARBER, B. E., WILLIAM, T., MENON, J., ANSTEY, N. M. & YEO, T. W. 2012. Deaths due to *Plasmodium knowlesi* malaria in Sabah, Malaysia: association with reporting as *Plasmodium malariae* and delayed parenteral artesunate. *Malar J*, 11, 284.
- RAJAHRAM, G. S., BARBER, B. E., YEO, T. W., TAN, W. W. & WILLIAM, T. 2013. Case Report: Fatal *Plasmodium Knowlesi* Malaria Following an Atypical Clinical Presentation and Delayed Diagnosis. *Med J Malaysia*, 68, 71-72.
- RAYMOND, M. & ROUSSET, F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Heredity*, 86, 248-249.
- RICKLEFS, R. E., OUTLAW, D. C., SVENSSON-COELHO, M., MEDEIROS, M. C., ELLIS, V. A. & LATTA, S. 2014. Species formation by host shifting in avian malaria parasites. *Proc Natl Acad Sci U S A*, 111, 14816-21.
- ROESTI, M., SALZBURGER, W. & BERNER, D. 2012. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol Biol*, 12, 94.
- ROUSSET, F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour*, 8, 103-6.
- RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics*, 16, 944-5.
- SACHS, J. & MALANEY, P. 2002. The economic and social burden of malaria. *Nature*, 415, 680-5.
- SBONER, A., MU, X. J., GREENBAUM, D., AUERBACH, R. K. & GERSTEIN, M. B. 2011. The real cost of sequencing: higher than you think! *Genome Biol*, 12, 125.
- SCHAER, J., PERKINS, S. L., DECHER, J., LEENDERTZ, F. H., FAHR, J., WEBER, N. & MATUSCHEWSKI, K. 2013. High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc Natl Acad Sci U S A*, 110, 17415-9.
- SCHLIEP, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27, 592-3.
- SEILMAIER, M., HARTMANN, W., BEISSNER, M., FENZL, T., HALLER, C., GUGGEMOS, W., HESSE, J., HARLE, A., BRETZEL, G., SACK, S., WENDTNER, C., LOSCHER, T. &

- BERENS-RIHA, N. 2014. Severe *Plasmodium knowlesi* infection with multi-organ failure imported to Germany from Thailand/Myanmar. *Malar J*, 13, 422.
- SERMWITTAYAWONG, N., SINGH, B., NISHIBUCHI, M., SAWANGJAROEN, N. & VUDDHAKUL, V. 2012. Human *Plasmodium knowlesi* infection in Ranong province, southwestern border of Thailand. *Malar J*, 11, 36.
- SETIADI, W., SUDOYO, H., TRIMARSANTO, H., SIHITE, B. A., SARAGIH, R. J., JULIAWATY, R., WANGSAMUDA, S., ASIH, P. B. & SYAFRUDDIN, D. 2016. A zoonotic human infection with simian malaria, *Plasmodium knowlesi*, in Central Kalimantan, Indonesia. *Malar J*, 15, 218.
- SIFFT, K. C., GEUS, D., MUKAMPUNGA, C., MUGISHA, J. C., HABARUGIRA, F., FRAUNDORFER, K., BAYINGANA, C., NDOLI, J., UMULISA, I., KAREMA, C., VON SAMSON-HIMMELSTJERNA, G., AEBISCHER, T., MARTUS, P., SENDEGEYA, A., GAHUTU, J. B. & MOCKENHAUPT, F. P. 2016. Asymptomatic only at first sight: malaria infection among schoolchildren in highland Rwanda. *Malar J*, 15, 553.
- SINGH, B. & DANESHVAR, C. 2013. Human infections and detection of *Plasmodium knowlesi*. *Clin Microbiol Rev*, 26, 165-84.
- SINGH, B., KIM SUNG, L., MATUSOP, A., RADHAKRISHNAN, A., SHAMSUL, S. S., COX-SINGH, J., THOMAS, A. & CONWAY, D. J. 2004. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet*, 363, 1017-24.
- SMITH, D. G., NG, J., GEORGE, D., TRASK, J. S., HOUGHTON, P., SINGH, B., VILLANO, J. & KANTHASWAMY, S. 2014. A genetic comparison of two alleged subspecies of Philippine cynomolgus macaques. *Am J Phys Anthropol*, 155, 136-48.
- SOARES, I., MOLEIRINHO, A., OLIVEIRA, G. N. & AMORIM, A. 2015. DivStat: a user-friendly tool for single nucleotide polymorphism analysis of genomic diversity. *PLoS One*, 10, e0119851.
- SORIA-CARRASCO, V., GOMPERT, Z., COMEAULT, A. A., FARKAS, T. E., PARCHMAN, T. L., JOHNSTON, J. S., BUERKLE, C. A., FEDER, J. L., BAST, J., SCHWANDER, T., EGAN, S. P., CRESPI, B. J. & NOSIL, P. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, 344, 738-42.
- STUKENBROCK, E. H., CHRISTIANSEN, F. B., HANSEN, T. T., DUTHEIL, J. Y. & SCHIERUP, M. H. 2012. Fusion of two divergent fungal individuals led to the recent

- emergence of a unique widespread pathogen species. *Proc Natl Acad Sci U S A*, 109, 10954-9.
- SULISTYANINGSIH, E., FITRI, L. E., LOSCHER, T. & BERENS-RIHA, N. 2010. Diagnostic difficulties with *Plasmodium knowlesi* infection in humans. *Emerg Infect Dis*, 16, 1033-4.
- SUTHERLAND, C. J., TANOMSING, N., NOLDER, D., OGUIKE, M., JENNISON, C., PUKRITTAYAKAMEE, S., DOLECEK, C., HIEN, T. T., DO ROSARIO, V. E., AREZ, A. P., PINTO, J., MICHON, P., ESCALANTE, A. A., NOSTEN, F., BURKE, M., LEE, R., BLAZE, M., OTTO, T. D., BARNWELL, J. W., PAIN, A., WILLIAMS, J., WHITE, N. J., DAY, N. P., SNOUNOU, G., LOCKHART, P. J., CHIODINI, P. L., IMWONG, M. & POLLEY, S. D. 2010. Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium ovale* occur globally. *J Infect Dis*, 201, 1544-50.
- TA, T. H., HISAM, S., LANZA, M., JIRAM, A. I., ISMAIL, N. & RUBIO, J. M. 2014. First case of a naturally acquired human infection with *Plasmodium cynomolgi*. *Malar J*, 13, 68.
- TA, T. T., SALAS, A., ALI-TAMMAM, M., MARTINEZ MDEL, C., LANZA, M., ARROYO, E. & RUBIO, J. M. 2010. First case of detection of *Plasmodium knowlesi* in Spain by Real Time PCR in a traveller from Southeast Asia. *Malar J*, 9, 219.
- TACHIBANA, S., SULLIVAN, S. A., KAWAI, S., NAKAMURA, S., KIM, H. R., GOTO, N., ARISUE, N., PALACPAC, N. M., HONMA, H., YAGI, M., TOUGAN, T., KATAKAI, Y., KANEKO, O., MITA, T., KITA, K., YASUTOMI, Y., SUTTON, P. L., SHAKHBATYAN, R., HORII, T., YASUNAGA, T., BARNWELL, J. W., ESCALANTE, A. A., CARLTON, J. M. & TANABE, K. 2012. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet*, 44, 1051-5.
- TAN, C. H., VYTHILINGAM, I., MATUSOP, A., CHAN, S. T. & SINGH, B. 2008. Bionomics of *Anopheles latens* in Kapit, Sarawak, Malaysian Borneo in relation to the transmission of zoonotic simian malaria parasite *Plasmodium knowlesi*. *Malar J*, 7, 52.
- TANIZAKI, R., UJIIE, M., KATO, Y., IWAGAMI, M., HASHIMOTO, A., KUTSUNA, S., TAKESHITA, N., HAYAKAWA, K., KANAGAWA, S., KANO, S. & OHMAGARI, N. 2013. First case of *Plasmodium knowlesi* infection in a Japanese traveller returning from Malaysia. *Malar J*, 12, 128.



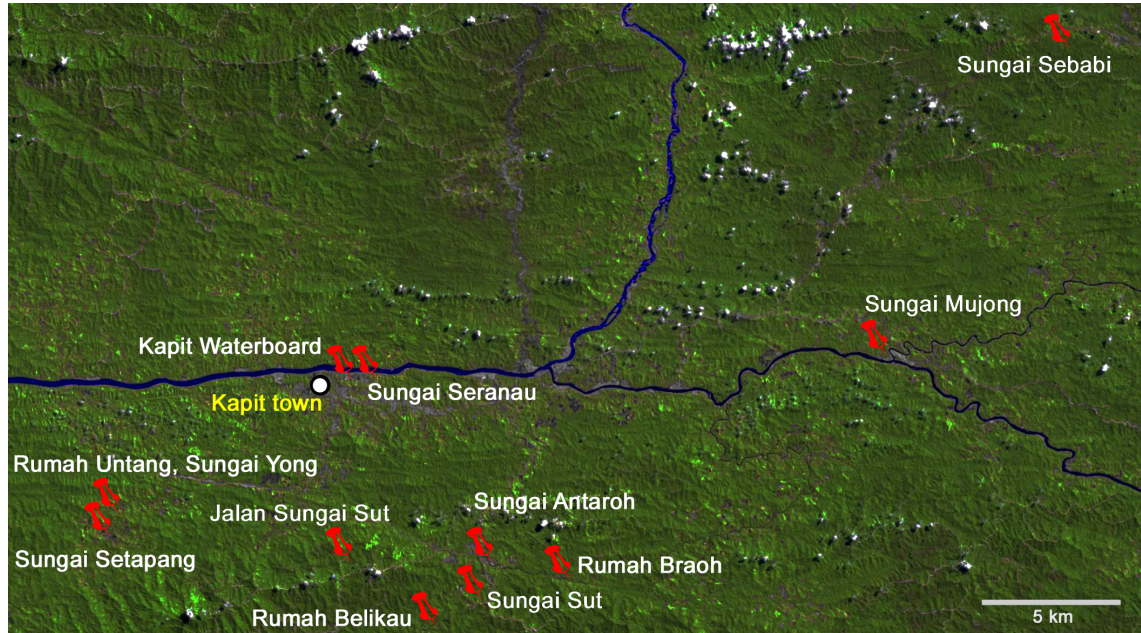
- TYAGI, R. K., DAS, M. K., SINGH, S. S. & SHARMA, Y. D. 2013. Discordance in drug resistance-associated mutation patterns in marker genes of *Plasmodium falciparum* and *Plasmodium knowlesi* during coinfections. *J Antimicrob Chemother*, 68, 1081-8.
- TYAGI, S., SHARMA, M. & DAS, A. 2011. Comparative genomic analysis of simple sequence repeats in three *Plasmodium* species. *Parasitol Res*, 108, 451-8.
- VAN DEN EEDE, P., VAN DER AUWERA, G., DELGADO, C., HUYSE, T., SOTO-CALLE, V. E., GAMBOA, D., GRANDE, T., RODRIGUEZ, H., LLANOS, A., ANNE, J., ERHART, A. & D'ALESSANDRO, U. 2010. Multilocus genotyping reveals high heterogeneity and strong local population structure of the *Plasmodium vivax* population in the Peruvian Amazon. *Malar J*, 9, 151.
- VAN DEN EEDE, P., VAN, H. N., VAN OVERMEIR, C., VYTHILINGAM, I., DUC, T. N., HUNG LE, X., MANH, H. N., ANNE, J., D'ALESSANDRO, U. & ERHART, A. 2009. Human *Plasmodium knowlesi* infections in young children in central Vietnam. *Malar J*, 8, 249.
- VARGAS-SERRATO, E., CORREDOR, V. & GALINSKI, M. R. 2003. Phylogenetic analysis of CSP and MSP-9 gene sequences demonstrates the close relationship of *Plasmodium coatneyi* to *Plasmodium knowlesi*. *Infect Genet Evol*, 3, 67-73.
- VORIS, H. K. 2000. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J Biogeog*, 27.
- VYTHILINGAM, I., LIM, Y. A., VENUGOPALAN, B., NGUI, R., LEONG, C. S., WONG, M. L., KHAW, L., GOH, X., YAP, N., SULAIMAN, W. Y., JEFFERY, J., ZAWIAH, A. G., NOR ASZLINA, I., SHARMA, R. S., YEE LING, L. & MAHMUD, R. 2014. *Plasmodium knowlesi* malaria an emerging public health problem in Hulu Selangor, Selangor, Malaysia (2009-2013): epidemiologic and entomologic analysis. *Parasit Vectors*, 7, 436.
- VYTHILINGAM, I., NOORAZIAN, Y. M., HUAT, T. C., JIRAM, A. I., YUSRI, Y. M., AZAHARI, A. H., NORPARINA, I., NOORRAIN, A. & LOKMANHAKIM, S. 2008. *Plasmodium knowlesi* in humans, macaques and mosquitoes in peninsular Malaysia. *Parasit Vectors*, 1, 26.
- VYTHILINGAM, I., WONG, M. L. & WAN-YUSSOF, W. S. 2016. Current status of *Plasmodium knowlesi* vectors: a public health concern? *Parasitology*, 1-9.

- WEBSTER, M. T., SMITH, N. G. & ELLEGREN, H. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci U S A*, 99, 8748-53.
- WILLIAM, T., JELIP, J., MENON, J., ANDERIOS, F., MOHAMMAD, R., AWANG MOHAMMAD, T. A., GRIGG, M. J., YEO, T. W., ANSTEY, N. M. & BARBER, B. E. 2014. Changing epidemiology of malaria in Sabah, Malaysia: increasing incidence of *Plasmodium knowlesi*. *Malar J*, 13, 390.
- WILLIAM, T., MENON, J., RAJAHRAM, G., CHAN, L., MA, G., DONALDSON, S., KHOO, S., FREDERICK, C., JELIP, J., ANSTEY, N. M. & YEO, T. W. 2011. Severe *Plasmodium knowlesi* malaria in a tertiary care hospital, Sabah, Malaysia. *Emerg Infect Dis*, 17, 1248-55.
- WILLIAM, T., RAHMAN, H. A., JELIP, J., IBRAHIM, M. Y., MENON, J., GRIGG, M. J., YEO, T. W., ANSTEY, N. M. & BARBER, B. E. 2013. Increasing incidence of *Plasmodium knowlesi* malaria following control of *P. falciparum* and *P. vivax* Malaria in Sabah, Malaysia. *PLoS Negl Trop Dis*, 7, e2026.
- WONG, M. L., CHUA, T. H., LEONG, C. S., KHAW, L. T., FORNACE, K., WAN-SULAIMAN, W. Y., WILLIAM, T., DRAKELEY, C., FERGUSON, H. M. & VYTHILINGAM, I. 2015. Seasonal and Spatial Dynamics of the Primary Vector of *Plasmodium knowlesi* within a Major Transmission Focus in Sabah, Malaysia. *PLoS Negl Trop Dis*, 9, e0004135.
- WORLD HEALTH ORGANIZATION 2013. World malaria report. Geneva, Switzerland: World Health Organization.
- WORLD HEALTH ORGANIZATION 2015. Eliminating Malaria: Case Study 8. Progress Towards Elimination in Malaria. San Francisco: University of California.
- WORLD HEALTH ORGANIZATION 2016. World malaria report. Geneva, Switzerland: World Health Organization.
- YUSOF, R., AHMED, M. A., JELIP, J., NGIAN, H. U., MUSTAKIM, S., HUSSIN, H. M., FONG, M. Y., MAHMUD, R., SITAM, F. A., JAPNING, J. R., SNOUNOU, G., ESCALANTE, A. A. & LAU, Y. L. 2016. Phylogeographic Evidence for 2 Genetically Distinct Zoonotic *Plasmodium knowlesi* Parasites, Malaysia. *Emerg Infect Dis*, 22, 1371-80.

- YUSOF, R., LAU, Y. L., MAHMUD, R., FONG, M. Y., JELIP, J., NGIAN, H. U., MUSTAKIM, S., HUSSIN, H. M., MARZUKI, N. & MOHD ALI, M. 2014. High proportion of knowlesi malaria in recent malaria cases in Malaysia. *Malar J*, 13, 168.
- ZHONG, D. B., BONIZZONI, M., ZHOU, G. F., WANG, G. Z., CHEN, B., VARDOS-ZALIK, A., CUI, L. W., YAN, G. Y. & ZHENG, B. 2011. Genetic diversity of *Plasmodium vivax* malaria in China and Myanmar. *Infection Genetics and Evolution*, 11, 1419-1425.
- ZIEGLER, T., ABEGG, C., MEIJAARD, E., PERWITASARI-FARAJALLAH, D., WALTER, L., HODGES, J. K. & ROOS, C. 2007. Molecular phylogeny and evolutionary history of Southeast Asian macaques forming the *M. silenus* group. *Mol Phylogenet Evol*, 42, 807-16.

## Appendices

### Appendix 2.1



Map illustrates the macaque sampling sites from locations within a 30 km radius of Kapit town, Sarawak.

## **Appendix 2.2**

Dataset – Genotypes at *P. knowlesi* microsatellite loci in infections from macaques (n = 47) and humans (n = 552).

Available online at

<http://journals.plos.org/plospathogens/article/file?type=supplementary&id=info:doi/10.1371/journal.ppat.1004888.s014>

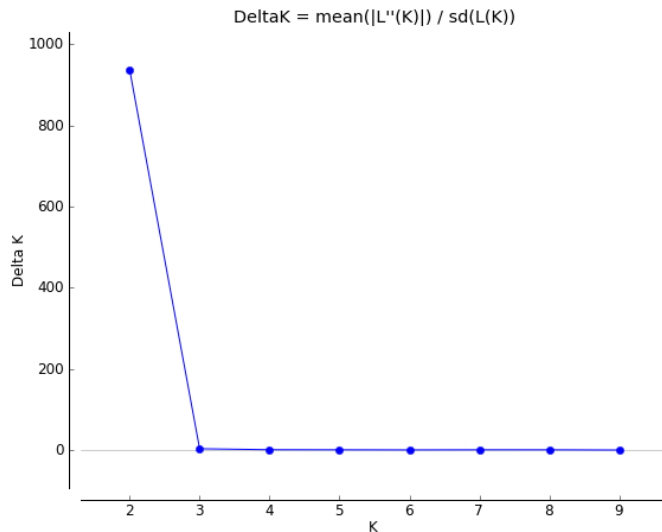
### Appendix 2.3

Genotypic data of 10 pairs and one triplet of identical haplotypes detected in six geographical locations. Out of 11 haplotypes, all were unique with no identical genotype at all complete loci.

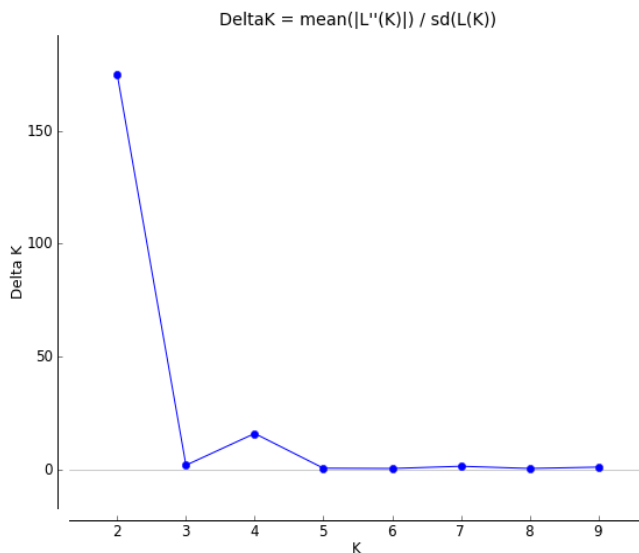
Haplotypes	Sample ID	Origin	Date of collection	NC12_2	NC03_2	NC09_1	NC12_4	NC10_1	CD08_61	CD11_157	CD13_61	CD05_06	CD13_107	Inferred cluster
1	KT033		2013											
	KT034	Kapit	2013	316	129	293	226	258	213	248	170	266	186	1
	KT052		2013											
2	CDK158	Kapit	16/7/2007	319	129	305	229	261	228	251	170	266	186	1
	CDK185		11/11/2007											
3	BTG095	Betong	2013	313	129	278	241	261	228	242	167	251	189	1
	BTG098		2013											
4	BTG090	Betong	2013	316	129	284	238	258	213	248	170	251	186	1
	BTG091		2013											
5	BTG059	Betong	2012	316	132	278	229	264	213	254	158	248	183	2
	BTG063		6/11/2012											
6	DFS612	Sarikei	2011	319	132	290	226	258	213	263	158	248	174	2
	DFS613		2011											
7	MRI/028/03	Miri	23/9/2003	316	129	278	232	264	213	248	158	248	189	2
	MRI/097/04		7/6/2004											
8	MRI/021/02	Miri	27/2/2002	322	132	287	226	258	213	248	167	248	189	2
	MRI/083/04		24/4/2004											
9	MRI/093/04	Miri	17/6/2004	334	129	287	232	258	213	248	158	248	180	2
	MRI/096/04		30/6/2004											
10	ML146/13	Tenom	2013	322	132	284	235	258	219	245	170	248	189	1
	ML148/13		2013											
11	KEL25	Kelantan	2013	310	132	281	229	261	213	248	182	248	174	2
	KEL26		2013											

## Appendix 2.4

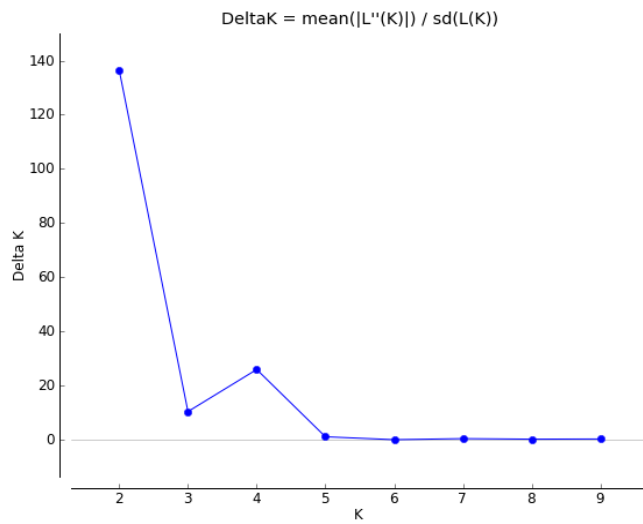
Plots of delta  $K$  ( $\Delta K$ ) based on Evanno's method for the determination of hypothetical ancestral population cluster ( $K$ ) from the STRUCTURE analysis extracted using the STRUCTURE Harvester.



**A.** STRUCTURE analysis of *P. knowlesi* from 44 macaque and 167 human isolates from Kapit, Sarawak strongly indicates the existence of two separate ancestral clusters throughout the populations ( $K = 2$ ,  $\Delta K = 936.75$ ).

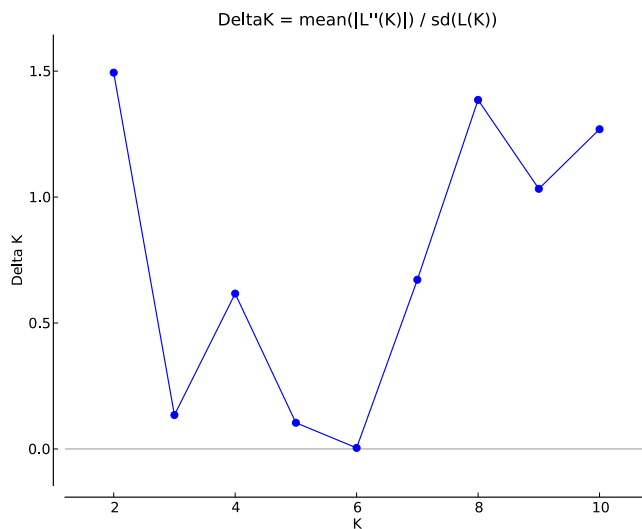


**B.** STRUCTURE analysis of *P. knowlesi* from 512 human isolates from all 10 sampling sites strongly indicates the existence of two separate ancestral clusters throughout the populations ( $K = 2$ ,  $\Delta K = 174.94$ ).



**C.** STRUCTURE analysis of *P. knowlesi* from all 44 macaque and 512 human isolates from Malaysia strongly indicates the existence of two separate ancestral clusters throughout the populations ( $K = 2, \Delta K = 136.39$ ).





# K	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	Ln''(K)	Delta K
1	-517.95	0.55	NA	NA	NA
2	-519.01	1.78	-1.06	2.66	1.49
3	-522.73	5.79	-3.72	0.78	0.13
4	-525.67	6.47	-2.94	3.99	0.62
5	-524.62	8.84	1.05	0.92	0.10
6	-524.49	9.52	0.13	0.04	0.01
7	-524.40	5.61	0.09	3.77	0.67
8	-520.54	3.17	3.86	4.39	1.39
9	-521.07	4.75	-0.53	4.91	1.03
10	-526.51	7.71	-5.44	9.79	1.27

**D.** In a similar analysis, reanalysis of 404 macaque and human isolates from the Cluster 1 population did not resolve any further population clusters, as indicated by the very low values of delta K.

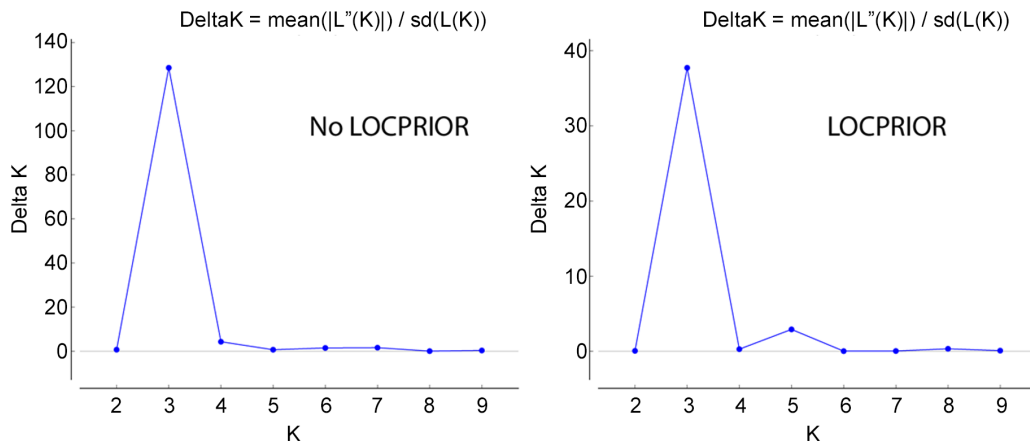
## **Appendix 4.1**

Dataset - Additional datasets regarding genotypes in humans and macaques studied in Malaysian Borneo and peninsular Malaysia.

Available online at <https://wwwnc.cdc.gov/eid/article/23/4/16-1738-techapp1.xlsx>.

## Appendix 4.2

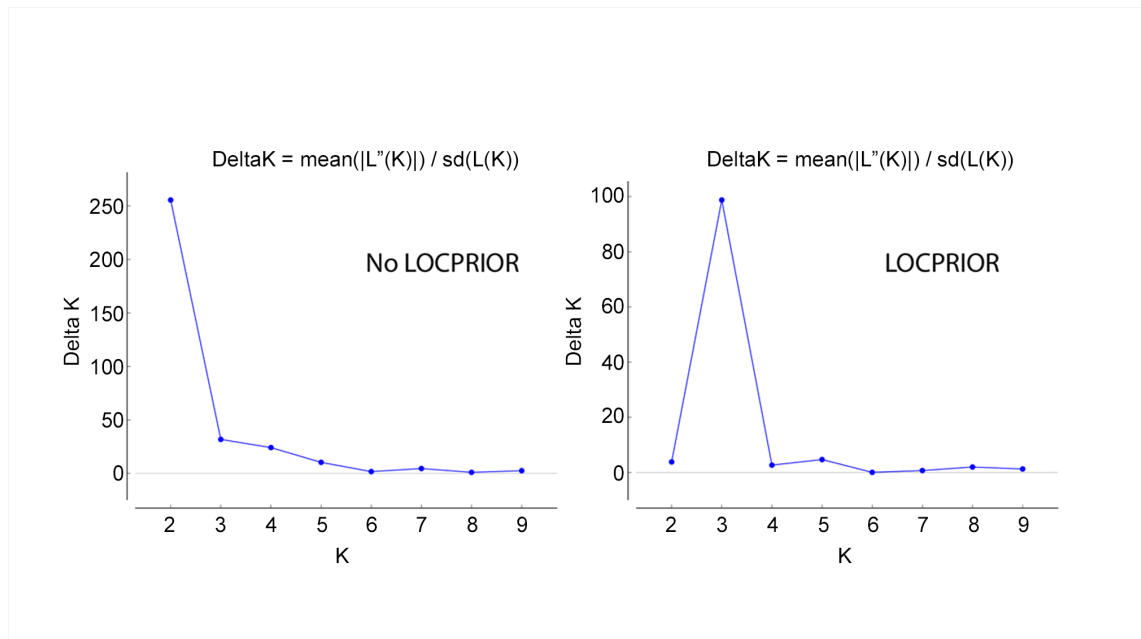
STRUCTURE analysis on 166 *P. knowlesi* infections across Malaysia and seven laboratory isolates obtained by 10 microsatellite loci.



Both admixtures with and without the LOCPRIOR models indicate the existence of three subpopulation clusters as estimated by Evanno's method ( $K = 3$ ;  $\Delta K = 128.51$  for the non-LOCPRIOR model and  $\Delta K = 37.72$  for the LOCPRIOR model).

### Appendix 4.3

STRUCTURE analysis on 758 *P. knowlesi* genotypes obtained by 10 microsatellite loci.



Estimated by Evanno's method, analysis of admixture without the LOCPRIOR model shows two subpopulation clusters ( $K = 2$ ,  $\Delta K = 255.50$ ), whereas admixture with LOCPRIOR model shows three subpopulation clusters ( $K = 3$ ,  $\Delta K = 98.73$ ).

## Appendix 5.1

Summary of remapping the previously generated short read sequences of *P. knowlesi* against the version 2.0 reference genome. The Sequence Read Archive accession numbers for each sample ID are mentioned in the Materials and Methods section 5.2.3.

No	Sample ID	Sequencing platform	QC-passed reads	% mapped	% properly mapped	Average depth coverage (X)
1	BTG026	HiSeq	41,313,779	96.04	84.36	146.64
2	BTG035	HiSeq	52,463,104	10.02	8.27	17.77
3	BTG039	HiSeq	45,761,870	79.59	70.55	134.72
4	BTG042	HiSeq	42,081,816	93.66	81.97	138.64
5	BTG044	HiSeq	29,325,669	97.40	86.28	106.17
6	BTG046	HiSeq	38,670,821	95.42	83.46	136.54
7	BTG047	HiSeq	32,277,653	95.26	83.92	113.71
8	BTG049	HiSeq	53,221,390	97.49	85.81	190.93
9	BTG050	HiSeq	33,519,780	94.01	80.75	115.50
10	BTG053	HiSeq	35,283,028	34.90	30.03	43.21
11	BTG055	HiSeq	21,829,275	77.64	66.57	59.30
12	BTG062	HiSeq	16,994,448	70.22	57.78	40.84
13	BTG063	HiSeq	24,682,482	62.87	54.77	54.42
14	BTG100	HiSeq	27,267,930	67.12	58.67	69.15
15	BTG123	HiSeq	25,430,647	89.20	80.91	86.82
16	CDK088	HiSeq	66,751,502	60.00	53.24	138.42
17	CDK206	HiSeq	38,122,706	35.21	30.06	46.09
18	KT003	HiSeq	42,777,884	17.76	15.00	27.47
19	KT004	HiSeq	42,025,428	92.31	81.48	146.56
20	KT006	HiSeq	60,281,143	92.74	79.15	211.51
21	KT012	HiSeq	22,442,197	95.63	85.76	81.86
22	KT025	HiSeq	56,208,103	94.16	83.79	201.71
23	KT026	HiSeq	33,338,952	83.74	73.31	105.78
24	KT027	HiSeq	52,853,787	92.96	81.42	187.58
25	KT029	HiSeq	28,872,280	62.58	54.37	68.46
26	KT030	HiSeq	28,525,157	90.61	78.79	98.13
27	KT031	HiSeq	50,367,159	90.81	81.98	173.76
28	KT034	HiSeq	23,838,990	53.04	46.11	47.20
29	KT040	HiSeq	30,655,543	97.60	86.47	113.01
30	KT042	HiSeq	28,468,946	95.15	86.27	103.16
31	KT048	HiSeq	20,775,707	77.46	68.02	60.40
32	KT050	HiSeq	19,424,602	93.49	82.97	68.46
33	KT055	HiSeq	46,135,860	90.10	84.64	156.67
34	KT056	HiSeq	48,992,338	81.75	76.99	151.34
35	KT057	HiSeq	45,715,366	92.63	85.97	159.30
36	KT072	HiSeq	45,016,497	95.70	88.65	161.03
37	KT073	HiSeq	46,334,104	95.13	88.03	164.87
38	KT077	HiSeq	26,284,061	95.29	86.43	95.54
39	KT081	HiSeq	28,061,976	95.01	86.35	101.66
40	KT092	HiSeq	26,779,552	96.35	87.61	98.40
41	KT094	HiSeq	29,726,016	97.46	88.18	110.38

42	KT095	HiSeq	26,349,673	96.90	87.82	97.27
43	KT100	HiSeq	24,359,451	97.65	88.17	90.80
44	KT103	HiSeq	22,730,860	97.59	88.17	84.31
45	KT107	HiSeq	30,273,348	94.21	85.17	108.66
46	KT109	HiSeq	26,247,181	96.40	87.09	96.43
47	KT114	HiSeq	24,462,899	88.82	81.43	83.28
48	KT120	HiSeq	20,937,609	79.02	71.65	63.15
49	SKS047	MiSeq	6,249,430	94.99	76.20	32.62
50	SKS048	HiSeq	44,612,191	97.59	86.07	165.11
51	SKS050A	HiSeq	57,817,253	98.31	86.79	215.82
52	SKS058	HiSeq	40,794,756	97.63	87.74	152.48
53	SKS073	MiSeq	6,378,060	97.36	81.60	34.48
54	SKS299	HiSeq	51,056,847	97.64	87.24	190.48
55	Pk_Hackeri	HiSeq	20,636,601	96.47	90.28	74.96
56	Pk_Malayan	HiSeq	16,464,035	87.42	86.12	55.99
57	Pk_MR4H	HiSeq	20,571,229	77.39	72.99	59.00
58	Pk_Nuri	Illumina Genome Analyzer II	23,348,408	97.30	91.52	90.08
59	Pk_Philippines	HiSeq	32,660,672	97.40	92.16	85.40

*"QC-passed reads"* means number of reads pass the quality control, *"% mapped"* means proportion of reads are successfully mapped while *"% properly mapped"* means forward and reverse reads are mapped properly in pairs with correct orientation.