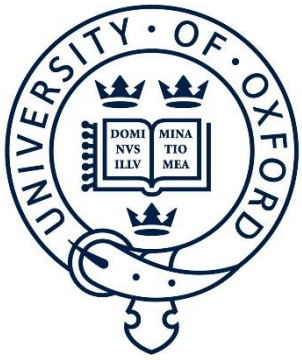


Exploring the spatial epidemiology and population genetics of malaria-protective haemoglobinopathies

Carinna Hockham

St. Catherine's College

Supervised by Prof Sunetra Gupta, Dr Frédéric Piel & Dr Bridget Penman

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2017

Contents

Abstract

Haemoglobinopathies, which include sickle-cell anaemia (SCA) and α- and β-thalassaemia, represent some of our few unequivocal examples of human evolution. The underlying genetic mutations reflect a recurring adaptation against one of the biggest infectious disease killers of humans, *Plasmodium falciparum* malaria. Inheritance of one copy of a sickle-cell or thalassaemic allele leads to protection against death from malaria, while two copies can result in a severe blood disorder. As a result, haemoglobinopathies have risen in frequency through balancing selection and pose a significant public health problem in parts of the world with a history of malaria transmission. Their study therefore lies at the interface between evolutionary biology and public health.

In this thesis, I explore different aspects of the epidemiology and population genetics of haemoglobinopathies around the world. Using pre-existing epidemiological data, statistical and geostatistical methods and Geographic Information System tools, I develop detailed evidence-based maps of the α-thalassaemia allele frequency distribution and genetic diversity in Southeast Asia and sickle-cell allele frequency in India. Pairing these with birth data, I generate sub-national estimates of the number of newborns born with severe forms of α-thalassaemia and SCA in Thailand and India, respectively, together with uncertainty estimates. In addition, I use a flexible population genetic simulation model to explore evolutionary explanations for the contrasting spatial haplotype patterns observed for SCA and the severe form of β-thalassaemia (β^0-thalassaemia) in sub-Saharan Africa and the Middle East, and resurrect a 20-year old question surrounding the genetic origin of sickle-cell.

Understanding the fine-scale geographical heterogeneities in the distributions of malaria-protective haemoglobinopathies is critical for addressing basic science questions and applied public health queries. Working at the interface between evolutionary biology and public health has provided me with the opportunity to build a more complete overview of the neglected increasing public health burden that this group of human disorders represents.

Statement of Contribution

This thesis is comprised of a series of manuscripts prepared for submission to peer-review journals (Chapters 2 and 3) and one published article (Chapter 4). All three results chapters therefore represent collaborative efforts between myself and my supervisors – Dr Frédéric Piel, Dr Bridget Penman and Professor Sunetra Gupta – as well as researchers in the United Kingdom (Chapters 2 and 3), Thailand (Chapter 2) and India (Chapter 3). In this Statement of Contribution, I summarise my personal contribution to each of the results chapters, and confirm that I have been the primary researcher and author for all of them. I also acknowledge and outline each author's contribution to the individual results chapters.

Chapter 2: Epidemiology of α-thalassaemia in Thailand and other Southeast Asian countries

Authors: Carinna Hockham, Supachai Ekwattanakit, Vip Viprakasit, Samir Bhatt, Bridget S. Penman, Sunetra Gupta & Frédéric B. Piel

This chapter has been prepared for submission to a peer-review journal. The work consisted of: (i) a systematic online search of published articles relating to α-thalassaemia prevalence and/or genetic diversity in 10 Southeast Asian countries, (ii) the compilation of a georeferenced database of representative α-thalassaemia surveys in these countries, (iii) the search of additional data sources in Thai peer-reviewed journals, (iv) the creation of detailed regional maps of α-thalassaemia in Southeast Asia, and (v) a detailed description of the epidemiology of α-thalassaemia in Thailand, including continuous

maps of the three major forms of this disorder, along with estimation of the number of newborns with severe α-thalassaemia in the country in 2020.

Dr Piel and I developed the data abstraction protocol together, after which I was solely responsible for compiling a georeferenced database of α-thalassaemia prevalence and/or genetic diversity surveys that adhered to our pre-specified inclusion criteria. Dr Supachai Ekwattanakit from Mahidol University in Bangkok, Thailand, shared additional data that he collated from journal articles found in the local Thai literature. All data cleaning was carried out by myself. Under the guidance of my supervisors, I conducted a detailed descriptive epidemiological analysis of the final database and the geostatistical analysis required for the generation of the continuous maps for Thailand. The R code for implementing the Bayesian geostatistical framework was provided by Dr Samir Bhatt, who provided mathematical expertise to the project. I sourced the demographic data needed to generate estimates of the number of affected newborns in 2020 and carried out the necessary calculations, under the supervision of Dr Piel and Dr Bhatt.

All the figures presented in the chapter were generated by myself and I wrote the full manuscript, with comments from my supervisors throughout. My supervisors and Dr Bhatt provided technical support and guidance throughout the project, whilst my Thai collaborators, Dr Supachai Ekwattanakit and Dr Vip Viprakasit, were available to share their local expertise during the analysis stages. Preliminary results of this work were presented as a poster at the 2nd Pan-Asian Conference on Haemoglobinopathies in Hanoi, Vietnam, in September 2015, organised by the Thalassaemia International Foundation.

Chapter 3: Sickle cell anaemia in India: an updated prevalence map, newborn estimates and the malaria hypothesis revisited

Authors: Carinna Hockham, Samir Bhatt, Malay Mukherjee, Roshan Colah, Bridget S. Penman, Sunetra Gupta & Frédéric B. Piel

This chapter has been prepared for submission to a peer-review journal. For this project, I spent three months at the National Institute of Immunohaematology in Mumbai, India, following a successful application to the Newton Bhabha Fund. Whilst there, I compiled data from the online literature as well as local journals, to which my Indian collaborators, Dr Malay Mukherjee and Dr Roshan Colah, provided me access. I was responsible for all data cleaning tasks and the generation of summary figures and tables for my dataset. Under the guidance of Dr Samir Bhatt, Dr Fred Piel and Dr Bridget Penman, I developed a geostatistical Generalised Additive Model (GAM), which I used to explore associations between β^S frequency in India and various covariates, including *Plasmodium falciparum* malaria endemicity. The fitted model was used to generate an updated map of β^S allele frequency distribution in India, which accounted for the social structure of the Indian population. Dr Bhatt and Mike Thorn, the database manager for the Malaria Atlas Project (MAP), helped me source data on some of the predictor variables. Dr Katherine Battle, also from MAP, provided me with digitised versions of two historical maps of malaria endemicity in India. I sourced all other demographic data myself and used it to calculate state- and district-level estimates of the number of AS and SS neonates born in India in 2020.

All figures presented in the chapter were generated by myself and I wrote the full manuscript, with extensive comments from my supervisors. Dr Malay Mukherjee and Dr

Roshan Colah were on hand to provide in-country expertise and commented on early versions of the manuscript.

Chapter 4: Understanding the contrasting spatial haplotype patterns of malaria-protective β-globin polymorphisms

Authors: Carinna Hockham, Frédéric B. Piel, Sunetra Gupta & Bridget S. Penman

This chapter has been published and is included here in its final form. The purpose of this study was to explore reasons why β^S and β^0-thalassaemia display very different spatial haplotype patterns at the population level, despite having malaria as a common selective pressure. Moreover, I investigated the evolutionary origin of β^S, specifically in relation to the genetic processes that gave rise to multiple observed β^S-associated haplotypes. I used simulation modelling to investigate the conditions under which the two different patterns were likely to form in a subdivided meta-population, and explored a range of demographic and genetic factors (e.g. network structure and mutation rate). I was responsible for developing a stochastic population genetic model that simulated the evolution of the β-globin polymorphisms in a meta-population. Under the guidance of Dr Penman and Dr Piel, I explored a large parameter space involving nine different factors, and developed the figures to summarise our findings. Given the high computational power needed to run the model, the simulations were implemented using the university's Oxford Advanced Research Computing (ARC) service, for which I wrote the required Perl script.

I presented the results of this project as an oral presentation at the Human Genome Meeting in Geneva, Switzerland, in April 2014, and received a grant from Promega

Corporation, a biotechnology company. In addition, I presented a poster of my work at the Challenges in Malaria Research conference in Oxford in September 2014.

I wrote the full manuscript, with extensive feedback from my supervisors throughout. The full citation is as follows:

Hockham, C., Piel, F.B., Gupta, S. and Penman, B.S. (2015). Understanding the contrasting spatial haplotype patterns of malaria-protective β-globin polymorphisms. *Infection, Genetics and Evolution*. 36:174-183 doi: 10.1016/j.meegid.2015.09.018.

* * *

As her supervisors, we certify that the statements of contribution listed here are a fair representation of Carinna Hockham's work.

Prof Sunetra Gupta	Dr Frédéric Piel	Dr Bridget Penman

Acknowledgements

The work of this thesis would not have been possible were it not for the support, guidance and encouragement that I have received from my supervisors over the last four years. Dr Frédéric Piel has been a pillar in the development and execution of my research. His wealth of spatial epidemiological knowledge and understanding has made our discussions productive and stimulating and I have learnt so much from him. I am grateful for his complete generosity with his time and his interest in my all-round development. Dr Bridget Penman's infectious enthusiasm has helped to keep morale high throughout my DPhil. Her extensive modelling expertise and ability to disentangle the most complex of problems have been an ideal combination for guiding my research. I am certain that, were it not for her mentorship over the last seven years, since she supervised my undergraduate project, I would never have gained the conviction to embark on this DPhil. Professor Sunetra Gupta has been a great source of academic support. She has established and nurtured a supportive research group, the Evolutionary Ecology of Infectious Disease (EEID) group, and I am grateful to my fellow postgraduate students for their friendship during my DPhil, especially Andrew Walker and Farania Rangkuti.

A significant proportion of my data for Chapters 2 and 3 was obtained through collaborations with researchers in India and Thailand. I thank Dr Malay Mukherjee and Dr Roshan Colah for hosting me at their institute in Mumbai. My research placement with them has certainly been a highlight of my DPhil. My sincere gratitude also extends to Dr Supachai Ekkanawanit and Dr Vip Viprakasit from Mahidol University Hospital, Bangkok, for their sharing of invaluable data and technical guidance. Both these collaborations have made my DPhil go above and beyond my expectations.

The geostatistical and spatial modelling aspect of my research has been helped along greatly by Dr Samir Bhatt, while Dr Jose Lourenço gave constructive feedback on my modelling work. I gratefully acknowledge their invaluable input. Thank you also to Dr Rosalind Howes and Dr Katherine Battle who have always expressed complete willingness to help, be it through useful discussions on disease mapping or the sharing of data for my analyses. Heather Green and Andrea Kastner work tirelessly to ensure the smooth running of the postgraduate program and EEID research group, respectively. I appreciate their unfailing support and assistance during my DPhil. I also thank the European Research Council for funding my DPhil research and the Newton Bhabha Fund for enabling my research placement in India.

My family and friends have been patient and encouraging. Special thanks to Eleanor Watkins, Georgia Gore-Langton, Liya Assefa and Wendy Carlin for their constant motivation and encouragement. My sisters, Natalie and Selina Hockham, brother-in-law, Stuart Faragher, and nephew, Joseph, have been wonderful sources of fun and respite during the busiest times. Thank you to my mum, who has always helped to keep things into perspective and offset the pressure that I put on myself. My dad has been supportive and encouraging throughout. Finally, but by no means least, I thank my boyfriend, Jonny Glyn, who has willingly spent many an hour listening to me talk through ideas that he could not possibly know anything about and maintained my spirits on a day-to-day basis. I cannot thank him enough.

Abbreviations

α^+	Deletional α^+-thalassaemia
α^0	α^0-thalassaemia
α^{ND}	Non-deletional α^+-thalassaemia
95% CI	Credible interval (Chapter 2) or confidence interval (Chapter 3)
β^C	β-globin variant C
β^D	β-globin variant D
β^E	β-globin variant E
β^S	β-globin variant S
GAM	Generalised additive model
GIS	Geographic information system
GRUMP	Global rural-urban mapping project
Hb	Haemoglobin
HbC	Haemoglobin variant C
Hb CS	Haemoglobin variant Constant Spring
HbD	Haemoglobin variant D
HbE	Haemoglobin variant E
HbS	Haemoglobin variant S
IQR	Interquartile range
Lao PDR	Lao People's Democratic Republic
n	Number
NBS	Newborn screening
PCR	Polymerase chain reaction
PPD	Posterior probability distribution
RBC	Red blood cell
SCA	Sickle-cell anaemia
SCD	Sickle-cell disease
SCT	Sickle-cell trait
SEA	Southeast Asia

SNP	Single nucleotide polymorphism
UN	United Nations
WHO	World Health Organisation

1. Introduction

1.1. Overview

Haemoglobinopathies, which include sickle-cell anaemia (SCA) and thalassaemia, are among the commonest inherited diseases of humans.[1] It is estimated that each year 300,000–400,000 infants are born with a severe haemoglobinopathy, with approximately 80% of these births occurring in developing countries.[2,3] This number is increasing as a result of: (i) population growth, (ii) the ongoing epidemiological transition in low- and middle-income countries such that affected children who would have previously died from other childhood diseases, including malaria, are now surviving long enough to present for diagnosis and management, and (iii) global migration facilitating the spread of the underlying mutations to low-prevalence countries (i.e. where the prevalence is ≤0.01%).[1,2,4,5] The timely development of appropriate programmes for the control, prevention and treatment of these disorders is therefore imperative. However, these disorders have largely been neglected due to competing priorities in high-prevalence countries (i.e. where the prevalence is ≥1%), including poverty, infectious disease and education. Moreover, local heterogeneities in the allele frequencies of the individual causal mutations have been observed in several countries,[6-11] making efficient national health planning difficult.

Aside from their growing public health significance, haemoglobinopathies also represent one of the best examples of recent human evolution.[12] Central to the elevated prevalence of this group of disorders is the malaria protection afforded by the causal alleles in the heterozygous state such that balancing selection has favoured carriers despite homozygotes expressing a severe blood disorder.[13-17] The so-called “malaria hypothesis”

was first put forward by J. B. S. Haldane in 1949,[18] and has since been supported by a multitude of evidence ranging from case-control studies demonstrating a protective effect of certain haemoglobinopathies to severe malaria,[19-22] to a comprehensive global mapping study confirming the geographical concordance between SCA and *Plasmodium falciparum* (*Pf*) malaria.[16] Whilst few dispute the role of malaria in explaining the broad population genetic patterns of haemoglobin mutations, certain features of their combined and respective distributions and population genetics warrant further investigation. This is true both at a global scale and at smaller spatial resolutions, including national and sub-national levels. For instance, Flint *et al.* suggested that no two malaria affected regions of the world contained the same profile of haemoglobinopathies, despite malaria being a common selective pressure.[23] Moreover, many of the haemoglobinopathies represent examples of soft selective sweeps in the human population, whereby ancestral genetic variation around the adaptive site is partially preserved owing to multiple alleles at the site being selected.[24] They exhibit remarkably different patterns in their genetic variation in the population, both at the locus level and with regards to the genetic backgrounds, or haplotypes, on which they occur. It is likely that a range of evolutionary, genetic and population genetic forces have contributed to shaping the patterns we observe today.

The population-level patterns of malaria-protective haemoglobinopathies are therefore interesting from two viewpoints. First, the detailed delineation of their epidemiology is critical for assessing the true burden of these disorders, which is necessary for the design, development and provision of appropriate public health interventions in the populations most affected by them. Second, the study of the population genetics of this group of disorders, and therefore the evolutionary processes that have shaped and maintained their variation within and among populations, is an important component to improving our

understanding of recent human evolution. More detailed knowledge of their population genetics should feed back into enhancing our capacity to predict their absolute health burden, therefore linking these two viewpoints. This thesis draws on concepts and methodologies from the fields of genetic epidemiology, spatial epidemiology and population genetics to address both applied public health queries and basic science questions.

In Chapter 2, I compile a database of pre-existing published prevalence surveys of α-thalassaemia in the 10 southeast Asian countries: Brunei Darussalam, Cambodia, Indonesia, Lao People's Democratic Republic (Lao PDR), Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam. I provide a detailed descriptive epidemiological analysis of the corresponding allele frequency estimates obtained from the surveys. For Thailand, I generate a continuous map of the distribution of the three major forms of α-thalassaemia (α^0-, α^+- and α^{ND}-thalassaemia), though with some important, but insightful, caveats. Where possible, I calculate estimates of the number of newborns born with hydrops fetalis syndrome and HbH disease (deletional and non-deletional) in the country in 2020. In Chapter 3, I explore in more detail the epidemiology and population-level distribution of SCA in India and highlight the importance of a multi-disciplinary approach to assessing the burden of the disorder here. Finally, in Chapter 4, I use a meta-population genetic model to simulate the formation of haplotype patterns for the mutations causing SCA and severe forms of β-thalassaemia (β^S and β^0-thalassaemia, respectively) and explore a range of factors in shaping the patterns observed. I shed light on the likely evolutionary history of these important human genetic polymorphisms and highlight areas of further research that are needed to improve our understanding of their evolution.

1.2. Molecular and genetic basis of normal human haemoglobin

Haemoglobin forms

Several forms of haemoglobin (Hb) are found in humans, with the relative amount of each form varying according to developmental stage.[25] All haemoglobins consist of two dimers of globin polypeptide chains which pair together to create a tetrameric protein molecule. In the fetus, the primary haemoglobin synthesized is fetal haemoglobin, or HbF ($\alpha_2\gamma_2$) In the second trimester and through to the third trimester, HbF levels gradually decline.[26] Adult HbA ($\alpha_2\beta_2$) is also synthesised in the fetus but becomes the principal haemoglobin only after birth, when a rapid increase in its production takes place (Figure 1.1). In a normal human adult, HbA accounts for more than 95% of all haemoglobin, with a second adult form, HbA_2 ($\alpha_2\delta_2$), contributing 1.5-3.5%. HbF is found in small quantities (<1%) in adults.[27]

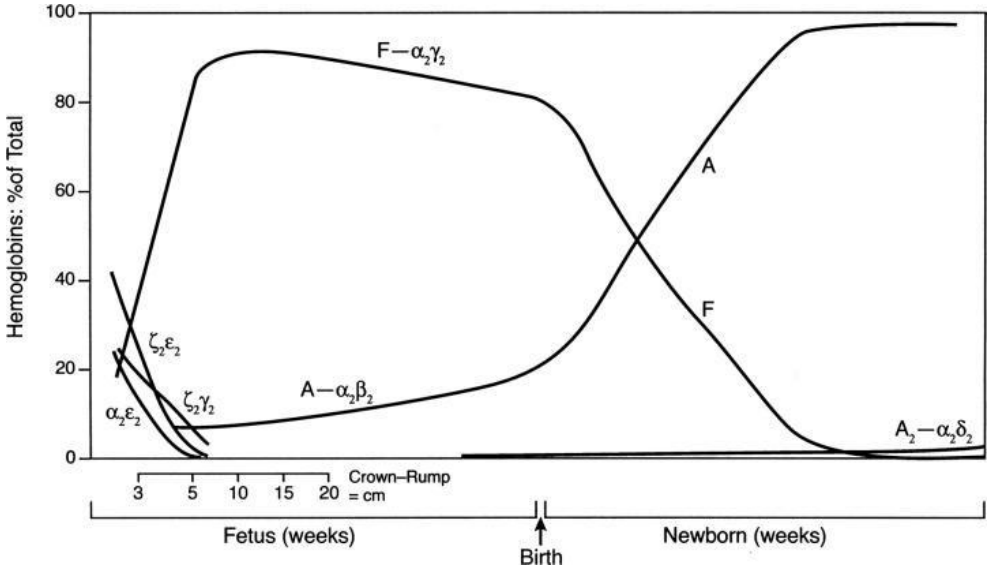


Figure 1.1 Developmental profile of haemoglobin forms in a normal individual. The graph shows the relative proportions (*y*-axis) of the different types of human haemoglobin during gestatation and into the early stages of life (*x*-axis). $\zeta_2\varepsilon_2$, $\alpha_2\varepsilon_2$ and $\zeta_2\gamma_2$ represent the embryonic haemoglobin forms, Hb Gower I, Hb Gower 2 and Hb Portland I, respectively. HbF ($\alpha_2\gamma_2$) is the predominant form during the second and third trimesters of pregnancy, after which adult HbA ($\alpha_2\beta_2$) increases in frequency. The other adult form, HbA_2 ($\alpha_2\delta_2$) also increases around the time of birth, although to a much lesser extent. Reprinted from Bunn and Forget (1986)[28] with permission from Elsevier ©1986.

The human globin gene clusters

Each globin chain (α-, ζ-, ε-, γ- and β-globin) is encoded by a corresponding gene, or genes, in the genome. These genes exist on two separate clusters: α- and ζ-globin on the α-globin gene cluster on chromosome 16, and ε-, γ- and β-globin on the β-globin gene cluster on chromosome 11 (Figure 1.2). Both clusters contain a combination of functional genes, non-expressed pseudogenes and control elements 5' to the encoding regions. The genes encoding α- and γ-globin are duplicated, meaning that there are four copies of each gene in the human genome (α1 and α2 and $^{G}\gamma$ and $^{A}\gamma$, respectively; see Figure 1.2). The remaining chains all have one gene each. During development, the globin genes are expressed according to a strict ontogenetic schedule to ensure the balanced production of the individual globin subunits needed for each haemoglobin form.[26,29]

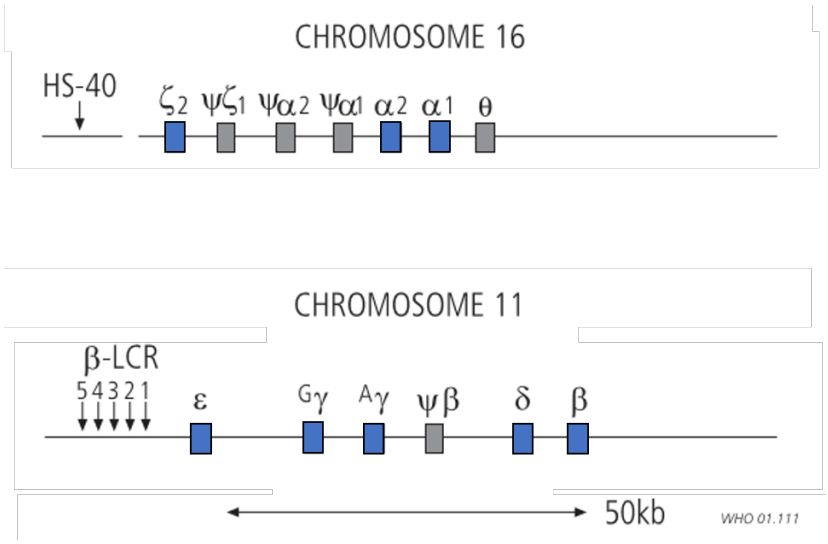


Figure 1.2 The genetic structure of the *α*- (top) and *β*-globin (bottom) gene clusters, which are located on chromosomes 16 and 11, respectively, in the human genome. The β-LCR and HS-40 are the main regulatory regions for these gene clusters. Functional globin genes are displayed in dark blue and pseudogenes in light grey.[25] Figure adapted from Weatherall and Clegg (2001).[1]

1.3. The haemoglobinopathies

Haemoglobinopathies are caused by non-synonymous mutations in any one of the globin genes described above that lead to appreciable changes in the corresponding globin

protein and/or its synthesis.[30,31] These mutations fall into two broad groups: (i) those that structurally alter the globin subunit, thereby affecting the function and/or stability of the compound haemoglobin molecule, and (ii) those that cause a reduction in the rate of synthesis of the corresponding globin protein, leading to an excess of the unaffected globin gene.[32] Notable examples, due to their malaria-protective characteristics, include the mutations responsible for SCA and α- and β-thalassaemia, respectively.

The clinical effects of these mutations are extremely variable, ranging from asymptomatic to fetal death *in utero*.[33] However, given the role of haemoglobin in transporting oxygen around the blood and its location in erythrocytes, they all in some way result in impaired oxygen delivery to tissues and/or ineffective erythropoiesis, which can lead to a myriad of further sequelae. The molecular biology, pathophysiology and clinical implications of these two broad categories of mutations and the major haemoglobinopathies that fall within them, namely those caused by defects in the α- and β-globin genes, are described in turn below.

Structural variants

Structural variants can result from any one of a range of mutations, including missense, nonsense, insertion or deletion events in one globin gene.[34] However, missense mutations, which cause a single amino acid substitution in the corresponding protein, are by far the most common.[34] To date, 1,299 haemoglobin variants have been catalogued on the Globin Gene database (HbVar, http://globin.bx.psu.edu, accessed 23 January 2018), of which a significant proportion are found in the β-globin gene (n = 618).[35] However, the vast majority of the variants reported in HbVar are extremely rare and many are also clinically silent. The most frequent clinically significant haemoglobin variants are haemoglobin S

(HbS), HbC, HbD and HbE, which are all caused by missense mutations in the β-globin gene (β^S, β^C, β^D and β^E, respectively).[34]

β^S is caused by a point mutation at position 6 of the β-globin gene (GAG > GTG), which replaces the hydrophilic glutamic acid amino acid with a hydrophobic valine one.[36] The hydrophobic nature of valine causes the resulting HbS variant, into which it is incorporated, to bind to an equivalent hydrophobic pocket on an adjacent HbS molecule when deoxygenated. The result is the polymerisation of HbS molecules, which distorts the morphology of the red blood cell, triggering a cascade of red blood cell alterations.[36] In the heterozygous state, referred to as sickle-cell trait (SCT), the level of HbS in the blood is sufficiently low that no symptoms are present. However, the homozygous state results in pathologic SCA. Other sickling disorders have also been observed, and these result from the compound inheritance of β^S with other haemoglobin variants, such as β^C, or α- or β-thalassaemia. Collectively, these sickling disorders are referred to as sickle-cell disease (SCD). SCA is the most common and severe form of SCD[37] and the one that I focus on in this thesis.

Individuals with SCA typically present with any combination of severe anaemia; joint, bone and abdominal pain; painful swelling of the hands and feet (dactylitis); acute chest syndrome; acute spleen enlargement; pallor; jaundice; and pneumococcal sepsis or meningitis.[38-41] Chronic pain and disability are a common hallmark of the disease, whilst early sudden death is highly probable in the absence of diagnosis and treatment. The pathophysiology of SCA is reviewed in Steinberg (1998)[42] and, more recently, Odièvre *et al.* (2011).[40] In brief, the polymerisation of deoxygenated HbS into insoluble fibres modifies and weakens the structure of the affected erythrocyte, causing it to sickle and

become less flexible. Sickled red blood cells aggregate and occlude blood vessels, which leads to tissue infarction, dactylitis and painful crises. The accumulation of HbS polymers in the cell also triggers a cascade of other cellular abnormalities, including disrupted ion channel function, cellular dehydration, accumulation of hemichromes on the internal membrane surface and increased cell oxidation. Affected red blood cells are recognised by the immune system and phagocytosed, leading to haemolytic anaemia. In addition, the immune system is profoundly impaired; a combination of overworked macrophages, a diverted blood flow through the spleen and multiple infarcts of spleen tissue reduces the efficiency of the immune system, rendering affected individuals more susceptible to infection. Indeed, infection is one of the main causes of death in patients with SCA, particularly in low- and middle-income countries.[38]

β^C and β^E have also reached polymorphic (defined as ≥1%)[43] allele frequencies in different parts of the world. β^C is caused by a substitution at codon 6 of the β-globin gene in which glutamic acid is replaced with lysine.[44] Similar to β^S, heterozygotes experience few or no symptoms. However, minimal adverse symptoms are also observed in β^C homozygotes, and pathology only occurs when β^C is co-inherited with β^S. In the case of β^E, lysine replaces glutamic acid at codon 26 of the β-globin gene. Heterozygotes are typically asymptomatic while homozygotes experience a mild blood disorder. Similar to β^C, β^E has the greatest clinical impact when it is co-inherited with β-thalassaemia; indeed, β^E/β-thalassaemia represents a substantial proportion of the burden of β-thalassaemia in Southeast Asia.[45] Most symptoms of the resulting disorder stem from the structural variant being synthesised at a lower rate than normal β-globin, similar to a β-thalassaemic mutation.[46]

Thalassaemias

In a healthy individual, the quantitative expression of the genes from each of the globin clusters is balanced and coordinated such that equal amounts of the globin proteins needed at each developmental stage are produced. For example, in the adult, α-globin and β-globin are in equal production. In individuals with thalassaemia, however, the defective synthesis of one or the other of the globin genes disrupts this balance, resulting in α-thalassaemia and β-thalassaemia when there is a deficit or excess of α- and β-globin chains, respectively. Accumulation of the unaffected globin protein forms the pathophysiologic basis of this group of disorders.[47] Worldwide, α- and β-thalassaemia are the most common types of thalassaemia, and the ones I will focus on here.

The thalassaemias are genetically diverse; more than 100 α-thalassaemia-causing mutations and over 250 genetic forms of β-thalassaemia have been described to date (HbVar, http://globin.bx.psu.edu, accessed 12 July 2017). The mutations themselves are remarkably heterogeneous.[48] They may involve the affected globin gene directly or occur in regulatory elements either in *cis* or *trans* to the globin encoding region. They may comprise deletions, which can be small or large and span several genes, or include single nucleotide substitutions. The affected globin gene may be inactivated entirely or its production simply reduced to varying degrees. Non-deletional mutations that alter the structure of the affected globin chain may result in the thalassaemia phenotype either by the resulting globin chain being less efficiently produced (e.g. β^{E} in the β-globin gene) or extremely unstable and therefore functionally deficient (e.g. $\alpha^{\text{Constant Spring}}$ in the α-globin cluster). Finally, different mutations may affect different stages of gene expression, from DNA transcription to RNA processing and translation. They also vary in their inheritance pattern; however, most are inherited in a Mendelian recessive manner.

In α-thalassaemia, mutations in one or more of the α-globin genes leads to a reduced production of α-globin. The extent of the resulting imbalance between α- and β-globin chains depends on the number of α-globin genes affected. The wild-type α-globin haplotype and its corresponding homozygous genotype are indicated as αα and αα/αα, respectively. In α-thalassaemia, deletional mutations can affect just one of the α-globin genes (-α), referred to as α^{+}-thalassaemia, or they can lead to the elimination of both genes in the pair (--), referred to as α^{0}-thalassaemia. The vast majority of the 100+ genetic variants of α-thalassaemia are caused by such deletion events. However, rarer non-deletion variants due to point mutations in either of the α-globin genes, which reduce their rate of synthesis and/or stability, have also been described.[49] These are depicted as $\alpha^{ND}\alpha$ or $\alpha\alpha^{ND}$, depending on the gene affected (α1 and α2, respectively). The most common α-thalassaemia mutations are shown in Table 1.1. Any combination of the above α-globin haplotypes can be inherited, giving rise to 10 possible genotypes at the α-globin locus (Table 1.2).

Table 1.1 The common α-thalassaemia mutations, along with the α-thalassaemia form that they cause and details of the mutation.

Mutation	Type of α-thalassaemia	Details[50,51]
$-\alpha^{3.7}$	α^{+}	3.7kb deletion resulting from reciprocal recombination between highly homologous regions; one α-globin gene is deleted.
$-\alpha^{4.2}$	α^{+}	4.2kb deletion resulting from recombination between mispaired homologous regions; one α-globin gene is deleted.
$--^{SEA}$	α^{0}	~20kb deletion; both α-globin genes on a chromosome are deleted.
$--^{THAI}$	α^{0}	Large 34-38kb deletion; both α-globin genes are deleted.
$--^{FIL}$	α^{0}	Large 30-34kb deletion; both α-globin genes are deleted.
$--^{MED}$	α^{0}	26.5kb deletion; both α-globin genes are deleted.
$\alpha^{\text{Constant Spring}}\alpha$	α^{ND}	Mutation in the termination codon in the α2-globin gene, resulting in an elongated chain.
$\alpha^{\text{Paksé}}\alpha$	α^{ND}	Mutation in the termination codon in the α2-globin gene, resulting in an elongated chain.

The phenotypic profile of α-thalassaemia is similarly variable, ranging from asymptomatic to fetal death *in utero* or stillbirth. The degree of severity correlates largely with the number of functional α-globin genes that are inherited (Table 1.2).[52,53] Individuals who inherit three functional genes (-α/αα or $\alpha\alpha^{ND}$/αα) are silent carriers with no noticeable symptoms, whilst those possessing two α-globin genes, either in *trans* or in *cis* (-α/-α, $\alpha\alpha^{ND}/\alpha\alpha^{ND}$ and --/αα), are deemed to have α-thalassaemia trait. Characteristic to this is the presence of hypochromic microcytic anaemia in the affected individual. Compound heterozygosity for a double deletion and either a single deletion (-α/--) or a non-deletion mutation ($\alpha\alpha^{ND}$/--), whereby only a single functional α-globin gene is present, results in HbH disease of variable clinical and haematological severity. Rarely, HbH disease can arise from homozygosity for severe non-deletional α-thalassaemia.

Notably, patients with non-deletional HbH disease typically present with more severe symptoms than those with deletional HbH disease.[49] Finally, the most severe form of α-thalassaemia, known as Hb Bart's hydrops fetalis, is caused by homozygosity for α^0-thalassaemia (--/--), such that the production of α-globin is completely eliminated. This form of the disorder is typically incompatible with life and death occurs *in utero* or soon after birth, although some cases of survival have been reported.[54]

The clinical outcomes of α-thalassaemia arise from excess β- or γ-globin chains forming pathology-causing β_4 and γ_4 homotetramers, known as HbH and Hb Bart's, respectively.[50] HbH is unstable and precipitates inside mature erythrocytes, resulting in their premature destruction in the spleen and, subsequently, haemolytic anaemia. Neither HbH nor Hb Bart's is able to transport oxygen, leading to tissue hypoxia and other systemic adverse effects. In HbH disease, small amounts of α-globin synthesised from the one functional α-globin gene is sufficient to maintain life, though with considerable morbidity in some cases.[55-57]

Table 1.2 The clinical variability of α-thalassaemia and broad phenotype-genotype relationships.

Genotype	Condition	Clinical characteristics[50]
αα/αα	Normal	None
-α/αα	Silent carrier	Clinically and haematologically normal
α^{ND}/αα	Variable (silent carrier to α-thalassaemia trait)	Variable
--/αα	α-thalassaemia trait	Microcytic hypochromic anaemia (mild)
-α/-α	α-thalassaemia trait	Microcytic hypochromic anaemia (mild)
α^{ND}α/α^{ND}α*	α-thalassaemia trait, or mild HbH disease	Microcytic hypochromic anaemia (mild) Mild HbH disease
-α/α^{ND}α	α-thalassaemia trait	Microcytic hypochromic anaemia (mild)
-α/--	HbH disease (deletional)	Moderate to severe microcytic, hypochromic, haemolytic anaemia Moderate Hepatosplenomegaly Mild jaundice
α^{ND}α/--	HbH disease (non-deletional), or (occasionally) Hb Bart's hydrops fetalis	Moderate to severe microcytic, hypochromic, haemolytic anaemia Moderate Hepatosplenomegaly Mild jaundice
--/--	Hb Bart's hydrops fetalis	Severe anaemia Heart failure, leading to edema, ascites, and pleural and pericardal effusion Extramedullary erythropoiesis, leading to hepatosplenomegaly Skeletal and cardiovascular malformations Death in utero Maternal complications due to enlarged spleen, including pre-eclampsia, haemorrhage and hypertension

*The non-deletion mutation can occur in either of the α-globin genes on a single chromosome and so the haplotype could be $\alpha^{ND}\alpha$ or $\alpha\alpha^{ND}$; however, for simplicity I always refer to a non-deletion mutation-carrying haplotype as $\alpha\alpha^{ND}$.

For β-thalassaemia mutations, the notations β^{++} and β^{+} are used to depict a partial reduction in β-globin synthesis, whilst β^{0} denotes those mutations that completely eliminate β-globin production.[32] The most severe forms of the disease therefore arise when two β^{0}-thalassaemia mutations are inherited. A detailed description of the pathophysiology of β-thalassaemia is provided in Weatherall & Clegg (2001).[47] Similar to α-thalassaemia, pathology stems from a deficiency of HbA and an excess of unaffected

α-globin chains, with the severity of disease largely depending upon the degree to which β-globin production is compromised. However, in this case, α-globin chains remain unpaired within erythroid precursors. They are oxidised and form hemichrome, which precipitates into inclusion bodies in the erythrocyte.[47,58,59] Further oxidative damage results from these inclusion bodies, as well as mechanical damage to the cell membrane. As a result, cell maturation is disturbed, leading to ineffective erythropoiesis. Those cells that do mature have a considerably shorter lifespan, resulting in haemolytic anaemia. To compensate for the reduced blood cell count, the bone marrow becomes overactive and expands, resulting in skeletal deformities characteristic of the disorder.[47] Iron absorption is increased, leading to iron overload, progressive iron deposition and, ultimately, organ failure and death. Affected individuals have a variable clinical course,[60] ranging from no symptoms (thalassaemia minor), to moderate symptoms (thalassaemia intermedia) to lifelong dependence on blood transfusions (thalassaemia major). In the most severe cases of β-thalassaemia major, where no β-globin is synthesised, the majority of cases will die in the first few years of life in the absence of blood transfusion.[1]

1.4. Epidemiology of haemoglobinopathies

In 2006, the World Health Organisation (WHO) recognised SCA and the thalassaemias as an important public health problem,[61,62] and, in 2010, the disorders were included in the Global Burden of Diseases, Injuries and Risk Factors Study.[63] Since then, there has been concerted action by the international haematology research community to improve our knowledge of the epidemiology of this important group of disorders and, crucially, generate estimates of their burden in the countries most affected by them.[4,10,37,64]

This is no easy task, however. A complete understanding of the epidemiology of haemoglobinopathies requires detailed information about their fine-scale distribution, prevalence, genetic diversity, absolute health burden, natural history and interactions with environmental factors. Estimating their current and future burden involves consideration of a myriad of factors, including: (i) the allele frequency of the underlying mutations, (ii) the population density of areas in which the mutations are present, (iii) the birth and population growth rates, (v) the patterns of population admixture (as consanguinity increases the probability of affected newborns), (vi) the impact of other causes of childhood mortality in determining whether children with a haemoglobinopathy survive long enough to present for diagnosis and treatment, (vii) the severity of the disease (morbidity), and (viii) survival rates (mortality). Each of these contributors vary considerably between and within countries, making sub-national estimates of the burden of haemoglobinopathies particularly pertinent. To add further complexity, there is extensive variability in the clinical severity of haemoglobinopathies – in particular, SCD, intermediate forms of β-thalassaemia and HbH disease – which makes assessment of the health provisions that are needed to manage them in each population, and even at the individual level, extremely difficult. Detailed context-specific studies into the effect of genetic modifiers and environmental factors on clinical severity should better enable clinicians and public health practitioners to predict treatment requirements of individual patients and provide appropriate lifestyle advice.

Overview of global distribution of haemoglobinopathies

An overview of the distributions of the major haemoglobinopathies across high-prevalence areas is provided in Figure 1.3. In this thesis, areas of "high" allele frequency are deemed to be those where the frequency of the causal globin mutation is ≥10%, which,

assuming Hardy-Weinberg proportions,[65,66] corresponds to a carrier frequency of ≥18% and a homozygote frequency of ≥1%.[67] Areas of "low" allele frequency are those where the allele frequency of the globin mutation is <1%, the carrier frequency is <2% and/or the homozygote frequency is <0.01%.[67]

Sickle haemoglobin is predominantly found in sub-Saharan Africa, the Mediterranean, the Middle East and parts of the Indian subcontinent,[1,64] where carrier frequencies regions range from 5–40%. In contrast, β^C is mostly prevalent in West Africa, whilst β^E is found almost exclusively in the northeastern states of India and in Southeast Asia. In the latter, β^E carrier rates can reach as high as 60%.[1] Thalassaemia has the most widespread distribution extending from the Mediterranean basin, sub-Saharan Africa (α^+-thalassaemia only), the Middle East and India, to Southeast Asia, Melanesia (α^+-thalassaemia only) and into the Pacific Islands (α^+-thalassaemia only). In these high-prevalence areas, carrier frequencies of α^+-thalassaemia range from 10% to as high as 80%, while those for α^0- and β-thalassaemia can reach 5% and 20%, respectively.[49] Although not shown in Figure 1.3, widespread population migration has dispersed the genetic disorders around the globe and they are now encountered in almost every country in the world, albeit at a lower prevalence.[1]

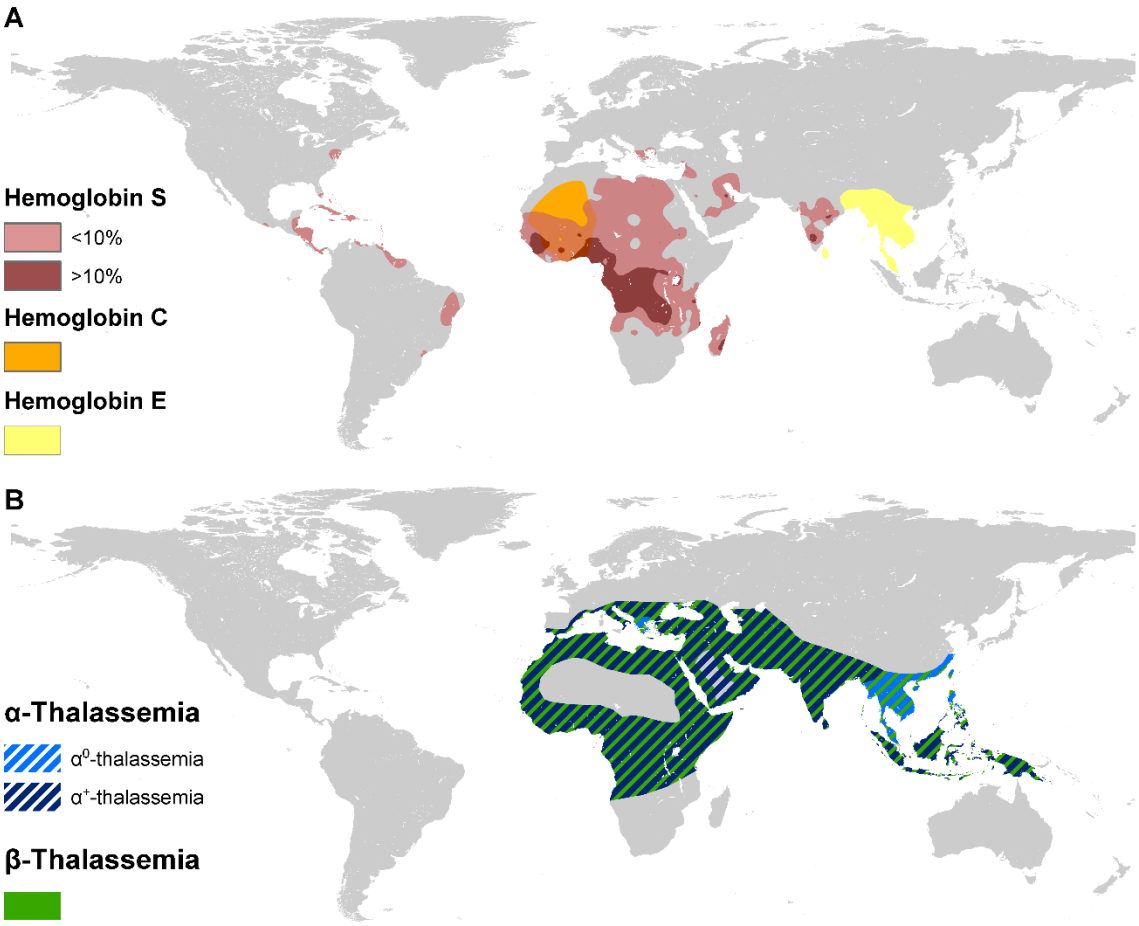


Figure 1.3 Geographic distribution of the major (A) structural haemoglobin variants, β^S, β^C and β^E, and (B) thalassaemias, α^0-, α^+- and β-thalassaemia. Reprinted from Piel (2016)[68] with permission from Elsevier ©2016.

Fine-scale heterogeneities in the allele frequencies of different globin mutations have been documented in several countries, including Sicily,[11] Sri Lanka,[10] Micronesia,[9] India,[6] Senegal,[7] Thailand,[69] and Vietnam.[8] Even within the same geographical location, allele frequencies can vary considerably between different communities, particularly where endogamy is common.[70,71] Optimal approaches for assessing the burden of haemoglobinopathies in individual countries therefore include universal newborn screening and micromapping.[72] However, these are often economically unfeasible, particularly in developing countries where the burden of haemoglobinopathies is greatest. Model-based mapping methods that use georeferenced pre-existing survey data to make

predictions about unsampled areas provide a solution to this.[73] Indeed, the first continuous maps of contemporary β^S and β^C allele frequencies, globally and in Africa, was generated by Piel *et al.* in 2013 using this approach.[44,64] Moreover, by pairing the maps with demographic data, the authors derived contemporary and future estimates of the annual number of affected newborns, including carriers (AS or AC) and homozygotes (SS or CC), nationally and, where possible, sub-nationally.[4] With many fewer resources than would be needed for universal newborn screening or micromapping efforts, these maps captured the spatial heterogeneity of the variants' allele frequencies over short geographical distances from existing surveys and provided a much-needed resource for guiding the design and implementation of appropriate control policies. However, no such maps exist for the thalassaemias and considerable gaps in our knowledge of this group of disorders remain. Furthermore, huge private and public efforts to set up screening programmes in different parts of India have been launched in the last decade, offering the potential to refine existing estimates and to tailor global methodologies mentioned above to a national context.

In order to contribute to the growing effort to assess, update and refine our epidemiological knowledge of the major haemoglobinopathies, in this thesis I have conducted further research into the epidemiology of SCA in India and generated the first detailed maps of the distribution of α-thalssaemia and α-thalassaemia variants in Southeast Asia. The rationale for my focus on the epidemiology of SCA and all major forms of α-thalassaemia (α^+-, α^0- and α^{ND}-thalassaemia) in India and Southeast Asia, respectively, is outlined below.

The epidemiology of SCA in India

The global map by Piel *et al.* (2013) provides the only available continuous map of β^S allele frequency in India.[64] Based on a database of 1,211 spatially unique surveys globally, of which 112 fell within India, the map suggests that the highest β^S allele frequencies in India are in southern Karnataka, varying between 10.0% and 12.5%. Other regions of high allele frequencies (>10%) included small areas of Chhattisgarh and Orissa, while allele frequencies of up to 10% were predicted across a large belt across central India and in large parts of Karnataka and Tamil Nadu (Figure 1.4). Based on available surveys, sickle haemoglobin was considered absent from the southern tip of Kerala and Tamil Nadu, in a large part of Andhra Pradesh, across the whole of the northern part of the country (Rajasthan, Haryana, Punjab, Jammu & Kashmir, Himachal Pradesh, Uttaranchal, Uttar Pradesh), in the eastern states of Bihar, Jharkhand and West Bengal and throughout the northeastern zone (Sikkim, Meghalaya, Tripura, Mizoram, Manipur, Nagaland, Assam and Arunachal Pradesh).

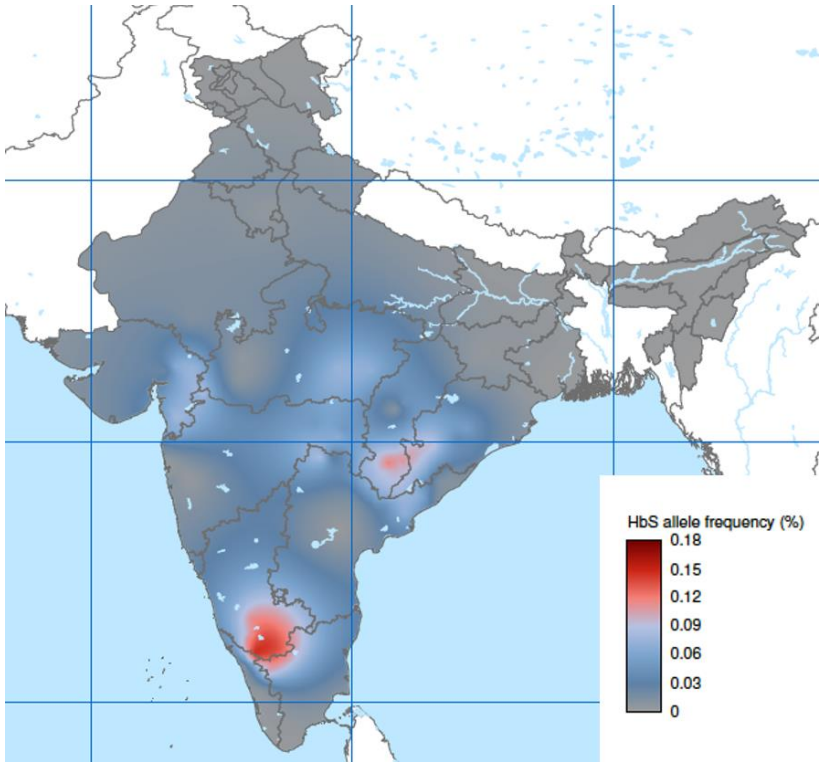


Figure 1.4 Map of the mean predicted β^S allele frequency in each 5km x 5km pixel across India, as calculated in Piel *et al.* (2013).[64] The per-pixel mean β^S allele frequency prediction was calculated from the posterior predictive distribution for each pixel, which was obtained from a Bayesian geostatistical model developed in Piel *et al* (2013). Figure obtained from The Malaria Atlas Project website (https://map.ox.ac.uk/explorer/#/explorer).

After taking into account demographic data, Piel *et al.* estimated that the number of AS and SS neonates born in India in 2010 was 996,563 (IQR 883,210–1,135,290) and 42,016 (IQR 35,346–50,919), respectively (18% and 13% of the global total, respectively). Of the latter, almost 50% were in Karnataka, Tamil Nadu and Maharashtra, with 25% predicted to occur in Karnataka alone (9394 [IQR: 6936–12 890]). By comparison, earlier estimates of β^S homozygous newborns calculated by Modell and Darlison were more conservative;[37] using demographic data from the 2003 United Nations Demographic Yearbook and an updated version of an earlier database of epidemiological studies of haemoglobinopathies, they estimated that 25,768 SCA newborns were born in India in 2003 (12% of the global total). However, these estimates were based on national averages

of β^S allele frequencies, thereby ignoring spatial heterogeneity. Moreover, no uncertainty metrics of their estimates were provided. To my knowledge, no other detailed burden estimates for India exist, although estimates for individual states are available. For instance, it has been estimated that the annual number births of β^S homozygotes in Madhya Pradesh alone is 3,358.[74]

Other attempts to summarise and map the distribution of β^S in India (e.g. Colah *et al.* 2014, 2015; Kaur *et al.* 2013)[71,75,76] have focused on surveys that were carried out in the tribal populations of India and are therefore not representative of the general population. Moreover, whilst they provide an excellent overview of the β^S distribution in tribal groups in sampled areas, there is no indication as to what the situation is in areas where no surveys have been conducted nor what the absolute burden of SCA in these groups is.

Whilst the map generated by Piel *et al.* is undoubtedly a valuable resource, its global nature necessitates that certain intricacies of individual countries are overlooked. For instance, the observation that, in India, SCA is predominantly found in scheduled tribal populations and other groups belonging to socioeconomically disadvantaged castes (known as scheduled castes, or SCs)[71,75] could not be explicitly incorporated into their model. Carrier frequencies as high as 40% and 30% have been reported in the scheduled tribes (STs) and SCs, respectively, compared to much lower carrier frequencies (5-10%) in non-scheduled groups.[75] According to the 2011 Census of India, scheduled groups account for 25% of the Indian population (~9% for STs, ~17% for SCs), corresponding to over 300 million people. To put this into perspective, the total population size of Nigeria, where the highest number of SS babies are predicted to be born,[64] is around 180 million.

The precision of such continuous maps as that generated by Piel *et al.* relies primarily on the size and density of survey data. Since its publication in 2013, a vast number of large-scale screening programmes have been initiated in different states, as illustrated in Colah *et al.*'s survey map. These include both targeted newborn screening (e.g. Upadhye *et al.* 2016),[77] premarital counselling (e.g. Mohanty *et al.*, 2011)[78] and community-based screening (e.g. Patra *et al.*, 2015).[79] This substantial increase in available prevalence data, and therefore β^S allele frequency data, including from very large surveys (>1 million individuals) should: i) contribute to refine model estimates, and ii) serve to reduce the relatively high uncertainties calculated for their estimates (IQR >= 7.5%). Given the large size of the Indian population, even small differences in β^S allele frequency can drastically change the absolute health burden of the associated disease.

In this thesis, I systematically collate this additional data into an up-to-date, georeferenced database to generate the first detailed continuous map of β^S allele frequency that accounts for differences in allele frequencies between scheduled and non-scheduled groups across the Indian subcontinent.

The epidemiology of α-thalassaemia in Southeast Asia

Southeast Asia consists of 10 countries with a population of over 600 million people: Brunei Darussalam, Cambodia, Indonesia, Lao PDR, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam. It is the region that is worst affected by α-thalassaemia, owing to the co-occurrence of all three major α-thalassaemia forms – α^+-, α^0- and α^{ND}-thalassaemia – at high allele frequencies (Figure 1.5).[60,68] It is estimated that between 10% and 30% of the population in Thailand possesses some form of α-thalassaemia mutation,[80] with the carrier rate of the more severe double deletion variants being as high

as 14% in some areas.[81] α-thalassaemia gene frequencies in Northern Thailand and Lao PDR can reach 40%, whilst lower but notable allele frequencies of up to 5% are observed elsewhere in the region (e.g. Malaysia and Singapore).[45] The $\alpha\alpha^{\text{Constant Spring}}$ variant is the most widespread non-deletion form and is found in Thailand, Cambodia, Vietnam and Lao PDR, with allele frequencies ranging between 1% and 8%.[82] When co-inherited with an α^0-thalassaemia haplotype, it results in the most severe clinical manifestations of HbH disease.[49]

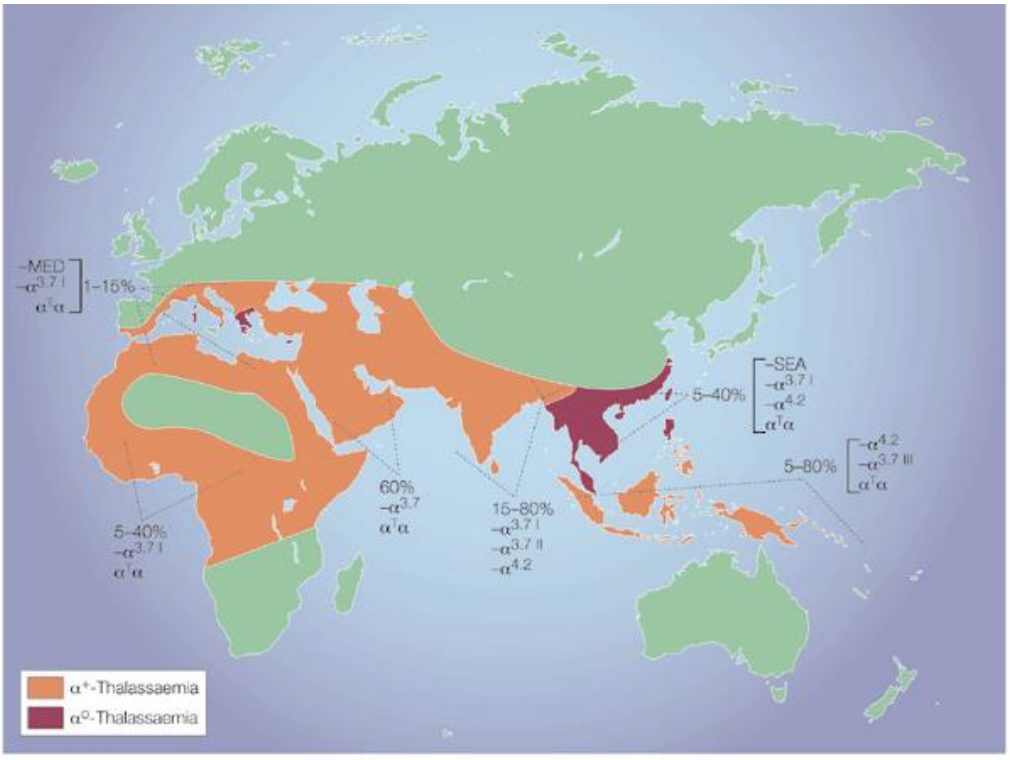


Figure 1.5 Distribution of α^0- and α^+-thalassaemia across Africa, Asia and the Pacific Islands. Different variants, along with the range of allele frequencies observed in each broad location, are indicated in the text. $-\alpha^{3.7}$, its subvarieties, $-\alpha^{3.7\,I}$, $-\alpha^{3.7\,II}$ and $-\alpha^{3.7\,III}$, and $-\alpha^{4.2}$ are all α^+-thalassaemia mutations, while $--^{MED}$ and $--^{SEA}$ are α^0-thalassaemia mutations. α^T indicates non-deletional α-thalassaemia mutations. Figure taken from Weatherall (2001).[53]

Prevalence surveys of α-thalassaemia across Southeast Asia indicate a heterogeneous distribution of the variants even within individual countries.[8,69] For instance, in Thailand, α-thalassaemia gene frequency in the northern part of the country is reported to be double

that of southern Thailand (30% and 16%, respectively).[83] Similarly, a community-based, cross-sectional survey in three different areas of Cambodia found the gene frequency to vary more than twofold (35.4% in Preah Vihar province compared to 14.3% in Phnom Penh). In this study, the highest variation was observed for deletional α^{+}-thalassaemia (10.3% *vs* 26.3%); however, smaller differences were also seen for non-deletional α^{+}-thalassaemia (8% to 3.2%). Since the optimal strategy for the prevention and management of severe forms of α-thalassaemia is based on the economically costly combination of population screening, genetic counselling and prenatal diagnosis with selective abortion of affected pregnancies, repeated blood transfusions and hospitalisations,[49] a detailed overview of areas at highest risk for severe disease is essential for policy makers to target interventions in the most cost-effective way. However, current estimates of the absolute burden of α-thalassaemia are only a very approximate assessment and, in many cases, based on a few, relatively small samples from localised areas in each country.[37]

In addition to better assessing the public health burden of clinically severe forms of α-thalassaemia, understanding the fine-scale distribution of mild forms of α-thalassaemia is also important in the Southeast Asian context. This region has some of the highest rates of both β-thalassaemia and the β^{E} variant globally.[84] There is strong evidence that less severe forms of α-thalassaemia (e.g. α^{+}-thalassaemia), whilst of lower clinical concern, act as genetic modifiers of these inherited conditions affecting the β-globin locus. For instance, the co-inheritance of α- and β-thalassaemia results in a milder disorder than the latter alone as the balance between the production of α- and β-globin chains is partly restored, despite an absolute decrease in the production of both chains.[85] Detailed knowledge of the epidemiology of all forms of α-thalassaemia is therefore important to predict inheritance with other haemoglobinopathies and is thus relevant for the effective

planning and implementation of control strategies for haemoglobinopathies in general. Furthermore, from a biological perspective, Southeast Asia offers an opportunity to assess the population-level interactions between α-thalassaemia and β-globin disorders, a key aspect in the study of their evolution.

Despite the clear need for a detailed delineation of α-thalassaemia and its variants in Southeast Asia, existing global maps of α-thalassaemia (e.g. Figure 1.5) present allele frequency data that is averaged at the national or even regional level, thereby masking any sub-national heterogeneity. In this thesis, I aim to develop a cohesive evidence-base for α-thalassaemia gene frequency in Southeast Asia and provide a detailed description of the current state of our knowledge of its epidemiology here, using the first generated maps of its distribution.

1.5. Population genetics of haemoglobinopathies: evidence for malaria protectiveness of the haemoglobinopathies

It is estimated that malaria began exerting a strong selective pressure on humans between 10,000 and 5,000 years ago, alongside the rise of agriculture,[86,87] although estimates differ markedly between populations, with varying degrees of certainty. Until less than a century ago, the disease was a major cause of mortality throughout most of the world, including in the southern parts of Europe (e.g. Sicily and Sardinia) and North America (e.g. Florida), where it has since been eliminated.[12] Today, it remains one of the most pressing public health issues in sub-Saharan Africa and is found at high prevalences (i.e. when the parasite rate ≥40%)[88] across vast parts of Asia and South America. Despite successive waves of malaria control and a recent push towards malaria elimination, it was estimated to cause more than 400,000 deaths in 2015, of which 70% were in children

under five years (http://who.int/mediacentre/factsheets/fs094/en/). Five species of the protozoan *Plasmodium* parasite are known to cause malaria in humans. These are *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* and *P. knowlesi*.[89,90] However, *P. falciparum* is responsible for the most severe, and potentially fatal, forms of the disease, and the species that most research into the population genetics of haemoglobinopathies has focused upon.

A potential interaction between haemoglobin mutations and malaria was first put forward by J. B. S. Haldane in 1949.[18] After observing that the gene responsible for causing β-thalassaemia major was frequent in Mediterranean countries, where malaria was also prevalent, he purported that the heterozygous state must be fitter than the wild-type genotype and invoked protection from malaria as a possible mechanism.[18] This is referred to as the "malaria hypothesis". Since its formulation, the hypothesis has been extended to other haemoglobin mutations, including β^S and α-thalassaemia, and considerable supporting evidence has been accumulated.

Although the malaria hypothesis was initially expressed with β-thalassaemia-causing mutations in mind, the most comprehensive evidence has been assembled in relation to malaria protection from the β^S variant. Allison (1954) provided the first evidence of a protective effect by showing that *P. falciparum* parasite density in β^S heterozygotes (depicted as AS) was significantly lower than that in normal (AA) individuals.[91] Other clinical evidence comes from case-control and cohort studies in Ghana [19] and Kenya [21], respectively. May *et al.* (2007) found that 1.4% of patients with severe *Pf* malaria carried the β^S gene, compared to 14.8% of matched controls.[19] Through examining malaria incidence in children with and without the β^S variant, Williams *et al.* (2005) found that

the heterozygous state had no effect on the incidence of symptomless parasitaemia.[21] However, AS individuals were half as likely to experience mild clinical malaria than their AA counterparts, whilst the heterozygous state provided 75% protection against hospitalisation for malaria and close to 90% protection against severe malaria. A protective effect has been demonstrated in many other studies, although some with certain shortcomings (e.g. small sample size).[13,14,92-97] Nevertheless, in all studies, AS heterozygotes suffered from malaria less frequently and/or less severely than AA controls. Finally, a global map of the β^S allele frequency in indigenous populations was constructed and compared to a historical map of *Pf* malaria prevalence.[16] A strong geographical concordance between the distributions of the two diseases was clearly demonstrated.

Evidence for an association between the thalassaemias and malaria stems primarily from their geographical coincidence with the infectious disease. α^+-thalassaemia, in particular, is present in all regions of the Old World where malaria has historically been a potent selective force.[85] Its distribution spans three continents, from the Mediterranean in Europe to sub-Saharan Africa, the Middle East and southern and far east Asia. It is also found at high allele frequencies in Melanesia. Microepidemiological studies of α^+-thalassaemia and examination of its relationship with malaria have been carried out in Papua New Guinea and Melanesia [98] and, more recently, Sri Lanka.[10] The first showed a strong positive correlation between α^+-thalassaemia allele frequency and historical malaria endemicity. Crucially, no relationship was found between malaria and other unlinked polymorphisms, thereby eliminating founder effects as a possible explanation for the high allele frequencies of α^+-thalassaemia in certain areas. In Sri Lanka, a significant association between the allele frequency of $-\alpha^{3.7}$ (a single deletion mutation that causes

α^+-thalassaemia, see Table 1.1) and pre-control malaria prevalence was found using two different data sources on malaria prevalence, at both district and school level. Further evidence comes from a comparison of α^+-thalassaemia allele frequency between an indigenous Nepali population and a recently immigrated (and previously unexposed) population, revealing a significantly higher allele frequency in the former (80% vs 10%).[99] Geographical evidence is limited for α^0- and β-thalassaemia, whose distributions are not as widespread as α^+-thalassaemia. Nevertheless, similar to other haemoglobinopathies, both disorders are predominantly found in areas with a history of malaria. In the aforementioned Sri Lankan study, Premawardhena and colleagues found a statistically significant relationship between β-thalassaemia and malaria but only for one of the two historical malaria prevalence datasets used in the analysis.[10]

On the clinical side, it is unclear whether the putative malaria-protective effect of α^+-thalassaemia arises from the homozygous or heterozygous state, particularly given its mild phenotype. Indeed, studies in Kenya and Papua New Guinea have shown homozygotes to be more protected from severe malaria than heterozygotes.[20,22,100] In Ghana, Opoku-Okrah *et al.* found a protective effect against severe malaria parasitaemia in both α^+-thalassaemia homozygotes and heterozygotes, but only amongst microcytic patients.[101] The authors were unable to identify a statistically significant association between α^+-thalassaemia genotype and severe malarial anaemia, however. Other studies that could not find a protective effect in the homozygous state included very small numbers of homozygotes (e.g. May *et al.*, 2007 and Mockenhaupt *et al.*, 2004)[19,102]. To date, there are no clinical studies that have investigated the protective effect, if any, of α^0-thalassaemia. However, given the highly deleterious nature of the mutation in the homozygous state, some form of balancing selection would have had to occur to observe

the allele frequencies of α^0-thalassaemia that we see today. The only available evidence for protection in β-thalassaemia comes from a study in Liberia in which protection against high parasite densities of *P. falciparum* was observed in children with β-thalassaemia trait.[96]

Despite the wealth of evidence in support of the malaria hypothesis, the exact mechanisms behind the malaria protection offered by the various haemoglobin mutations remain unclear.[12,103] Various mechanisms have been proposed, with varying degrees of evidence to support their existence. These include: (i) an inability of the parasite to infect mutant erythrocytes, (ii) impaired parasite growth within mutant cells, (iii) enhanced clearance of infected mutant erythrocytes by the immune system (innate and acquired), (iv) interference with the pathophysiological malaria processes, thereby averting the development of severe disease, and (v) enhanced direct immunity against malaria antigens.[104] Several reviews of the various proposed mechanisms are available.[103,105,106]

1.6. Ongoing questions relating to the population genetics of malaria-protective haemoglobinopathies

As highlighted above, the overarching evolutionary interaction between haemoglobin mutations and *Pf* malaria is largely undisputed.[12,23] However, there are certain features of the distributions of haemoglobinopathies that cannot be explained by malaria selection alone.

First, there are parts of the world where *Pf* malaria is endemic yet haemoglobinopathies occur at very low prevalences or are almost absent (e.g. southern Africa and parts of South America)[9,16,64,107] Other evolutionary forces must have acted upon the causal genes to

preclude their spread in these areas. For instance, the absence of β^S in areas south of the Zambezi river in Africa may be explained by the arrival of the mutation in northern Bantu populations only after the southward expansion of these groups had begun around 2000 years ago.[23] Conversely, the observation that no haemoglobinopathies have ever been reported in indigenous populations of the Americas could be in part due to the recent import of malaria here, estimated to be within the last 500 years.[108] In western Europe, β^S is found exclusively in individuals originating from high β^S-frequency areas, making migration the most plausible explanation for its polymorphic allele frequencies here in the absence of malaria selection. In Polynesia, α^+-thalassaemia has reached allele frequencies as high as 12% on islands where endemic malaria has never been present.[9] A combination of archaeological and linguistic evidence and haplotypic analysis suggests that these unexpectedly high allele frequencies could have arisen as a result of migration by proto-Polynesians from southeast Asia via Melanesia, where they acquired the $-\alpha^{3.7III}$ variant, followed by founder effects and random genetic drift in the small founding population of each Polynesian island.[23] The relatively benign nature of the α^+-thalassaemia variant would have precluded it from being selected out of the populations.

A second intriguing feature of the distribution of the haemoglobinopathies is that no two regions of the world contain the same combination of disorders. The same haemoglobinopathy that predominates in one malarious region may be maintained at a lower prevalence in another, equally malarious region, where another disorder is common. For instance, β^S is only present in local areas in southern European countries (e.g. Italy and Greece) but present in high allele frequencies across large swathes of sub-Saharan Africa, whilst the opposite is true for β-thalassaemia. α^+-thalassaemia, by contrast, is found at high allele frequencies in both regions but never exceeds 50% in sub-

Saharan Africa. Using mathematical modelling methods, Penman *et al.* (2009) showed that epistatic interactions between mutations occurring at the two genetic loci could help to explain these differences in spatial patterns.[109] They cited two types of gene-gene interactions (or epistasis) to be of particular relevance. The first is the amelioration of the blood disorder when α- and β-thalassaemia are inherited together (i.e. positive epistasis), due to the balancing out of impaired globin chain production.[110-113] The second is the cancellation of malaria protection in compound heterozygotes for β^S and α-thalassaemia (i.e. negative epistasis).[114-116] The authors found that inclusion of positive epistasis between α- and β-thalassaemia facilitated the preclusion of β^S from the population. Thus, under appropriate conditions, β^S could have entered early into the Mediterranean and still be at low allele frequencies today, despite it offering greater protection against malaria. This is in contrast to an earlier hypothesis that the low allele frequency of β^S in the Mediterranean is explained by its recent introduction into the region.[117] Moreover, they showed that negative epistasis between β^S and α-thalassaemia could account for the observed 50% cap on the allele frequency of the latter in sub-Saharan Africa.[109] Penman *et al.* have also shown that epistasis may have influenced the allele frequencies obtained by α^+-thalassaemia and β^S in parts of India,[118] and that malaria mortality feedbacks can change the outcomes of the competitive interaction between β^S and β^C, potentially explaining the geographical limitation of β^C to West Africa.[119]

A third intriguing feature of the haemoglobinopathies, and one which I explore in this thesis, is the population-level distributions of their associated globin haplotypic backgrounds. A haplotype is defined as a combination of alleles at multiple loci that, by virtue of their close proximity, are transmitted together. The mutations that cause β^S and β^0-thalassaemia possess contrasting haplotype compositions and spatial haplotype

patterns, despite affecting the same locus and both having been subject to balancing selection.[23] The β^S variant is associated with five classical restriction fragment length polymorphism (RFLP) haplotypes that show high specificity in their geographical distributions across sub-Saharan Africa and in the Middle East and India. These have been named Bantu (or Central African Republic, CAR), Benin, Cameroon, Senegal and Arab-Indian, to reflect the country or region in which they were originally identified.[23] By contrast, β^0-thalassaemia-causing mutations are much more mixed in their distribution, with multiple mutations being found within a single region and each associated with more than one haplotype.[1,120] These differences in their patterns of linked genetic variation at the population level are likely indicative of different evolutionary histories of these two polymorphisms. However, no detailed investigation into this has been carried out.

These haplotype patterns also raise important questions that are still debated surrounding the genetic origins of these two polymorphisms. While individual β^0-thalassaemia mutations clearly represent independent mutational origins, the occurrence of identical mutations on separate haplotypes is generally ascribed to gene conversion.[121] This is in stark contrast to the theory surrounding the origin of β^S, whereby the existence of the five classical β^S-associated haplotypes is deemed indicative of the mutation having arisen independently five times. However, there is no obvious reason why gene conversion could not have generated the patterns that we observe today.[23] Unfortunately, to date, only one study has attempted to examine this possibility further; Livingstone used an old-fashioned stochastic model to simulate the dispersal of β^S- and β^A-associated haplotypes and demonstrated that multiple haplotypes could arise exclusively through gene conversion, with no need for recurrent mutation.

By incorporating a range of genetic and demographic processes into a stochastic population genetic model, I explore the role of migration, population size, network structure, mutation rate, gene conversion rate and haplotype fitness in shaping and maintaining the spatial haplotype patterns that we see today, and investigate the 20-year old question of the genetic origin of the β^S mutation.

1.7. Aims of the thesis and description of chapters

This thesis has two overall themes. These are:

1. The detailed description of the epidemiology of haemoglobinopathies in different parts of the world.
2. The exploration of the role of population genetic factors in determining observed population genetic patterns of haemoglobinopathies.

In Chapter 2, I compile a georeferenced database of representative surveys of α-thalassaemia prevalence in Southeast Asia and use this to provide a detailed epidemiological description of the current state of our knowledge surrounding the epidemiology of this disorder in the region and highlight critical gaps. Despite their limitations, the maps generated in this thesis constitute the first cohesive cartographic representations of the distribution of a clinically relevant but largely neglected haemoglobinopathy. The maps also provide a basis for further research into the evolutionary history of this genetically and phenotypically heterogeneous disorder. Through my collaboration with researchers in Thailand, I demonstrate the importance of cross-country collaboration in closing these gaps.

In Chapter 3, I use a combination of pre-existing allele frequency data and statistical modelling methods to explore in more detail the epidemiology of SCA, and in particular, the β^S variant, in India. I extend the work of a previous global mapping study through the inclusion of additional epidemiological and demographic data to generate an updated continuous map of β^S allele frequency in India and sub-national estimates of the number of newborns born with SCA in 2020. By focusing on a single country, I am able to incorporate some of the intricacies of its population structure into my analysis of the data. However, my research also shows that the inclusion of the additional β^S allele frequency data, which exhibit extensive spatial heterogeneity, makes fine-scale predictions difficult. Given the size and complexity of the Indian population, and the variability in factors that influence the distribution of β^S (e.g. malaria selection, allele frequency of other globin mutations, migration, and so on), it may be necessary to supplement statistical methods with population genetic models to more fully capture the complex distribution of β^S in the Indian context.

In Chapter 4, I shift to the population genetic enquiry of why the β-globin polymorphisms that give rise to SCA and β-thalassaemia display such contrasting patterns of linked genetic variation at the population level, despite occurring at the same locus. In this way, I address a long-standing question surrounding the evolutionary origins and spread of individual β-globin mutations, highlighting the importance of demographic factors in influencing their population genetic patterns. Moreover, I argue that a better understanding of the differences in fitness between genetic backgrounds will further shed light on the evolutionary origins of β^S, the mutation responsible for causing SCA.

2. Epidemiology of α-Thalassaemia in Thailand and Other Southeast Asian Countries

This work was carried out with collaborators from Bangkok, Thailand, who provided a substantial proportion of the data on α-thalassaemia allele frequencies in the compiled database. The work has been written up in the format of a journal article and will be submitted to a peer-review journal.

2.1. Abstract

Background

α-thalassaemia is one of the commonest monogenic disorders worldwide. It is a particularly important public health problem in countries of Southeast Asia, where the most severe forms of this disorder are found. Milder forms, although clinically benign, are important genetic modifiers of other common inherited conditions (e.g. β-thalassaemia). Information on the prevalence, genetic diversity and health burden of α-thalassaemia in Southeast Asia remains limited. There are as yet no available detailed maps of α-thalassaemia gene frequency, while current newborn estimates for the region are largely inconsistent.

Methods

We constructed a geodatabase of published population surveys of α-thalassaemia prevalence and genetic diversity in Southeast Asia, which allowed descriptive maps of the current evidence-base for α-thalassaemia gene frequency and genetic diversity in the region to be generated. As large volumes of additional data for Thailand could be accessed from local journals, a Bayesian geostatistical modelling framework was used to generate continuous allele frequency maps of different α-thalassaemia forms in the country. These national maps were paired with demographic data to yield sub-national estimates of the annual number of newborns born with severe forms of the disorder.

Findings

There is remarkable spatial heterogeneity in α-thalassaemia gene frequency across Southeast Asia. Overall, gene frequencies range from 0% in parts of peninsular Malaysia to 35% in northern Cambodia. Its genetic variant profile is similarly variable. There is an

important publication bias in data that are accessible through international bibliographic databases (e.g. PubMed), and substantial evidence is available from local data sources. The methodology used to calculate aggregated national estimates of disease burden for Thailand provides a refined model, which could be replicated in other countries in the region.

Interpretation

Accurate, fine-scale epidemiological data of α-thalassaemia are necessary to define appropriate national and regional health policies for its control. This is particularly important due to the high heterogeneity of the gene frequency and clinical severity of α-thalassaemia variants in the region. The maps and newborn estimates presented here are an important first step towards this but also highlight critical gaps in our evidence-base. Multi-disciplinary collaborations between local clinicians, field researchers and epidemiologists are essential to ensure maximum use of public health data that are being collected in local settings.

Funding

European Research Council

2.2. Introduction

α-thalassaemia is one of the commonest monogenic disorders in humans, spanning much of the malaria belt, including the Mediterranean, sub-Saharan Africa, Asia and the Pacific. It is estimated that up to 5% of the world's population carries at least one α-thalassaemia variant, with some populations reporting gene frequencies of close to 80%.[1] Central to its elevated frequency is the malaria protection afforded by the underlying genetic mutations such that they have been favoured by natural selection in populations with historically high rates of malaria.[2-5] Due to recent population migrations, α-thalassaemia is now common in other parts of the world, as illustrated by the inclusion of HbH disease (a form of α-thalassaemia) in the newborn screening programme in California.[6,7]

A normal individual inherits four copies of the α-globin gene. In an individual with α-thalassaemia, at least one of these four copies is absent or non-functional. The resulting deficit in α-globin affects the balance between α-globin and β- or γ-globin chains that is necessary to produce normal adult haemoglobin (HbA) and normal foetal haemoglobin (HbF), respectively.[8] A deficit of α-globin leads to the production of γ-globin tetramers, called Hb Bart's, in the foetus or β-globin tetramers, called HbH, in adults. Due to their very high oxygen affinity, neither tetramer is capable of transporting oxygen efficiently.[9] Furthermore, the instability of HbH leads to the production of inclusion bodies in red blood cells and a variable degree of haemolytic anaemia. Overall, the severity of α-thalassaemia is inversely related to the number of functional copies of the α-globin gene. To date, 120 genetic forms of α-thalassemia have been identified (HbVar, http://globin.bx.psu.edu, accessed 23 January 2018). These include: (i) double gene deletions that remove both α-globin copies in a gene pair (α^0-thalassaemia), (ii) single gene deletions that remove one α-globin copy (α^+-thalassaemia), and (iii) non-deletional

(ND) mutations that in some way inactivate the affected gene (α^{ND}-thalassaemia). While deletions constitute the vast majority of these α-thalassaemia variants, non-deletional variants are typically associated with more severe phenotypes.[1] Because the geographical distribution of β-thalassaemia largely overlaps with the distribution of α-thalassaemia, it is important to note that their co-inheritance often leads to a reduced imbalance between α-globin and β-globin chains, resulting in a milder phenotype.[10-13]

From a clinical perspective, α-thalassaemia is mostly a burden in Southeast Asia where α^0-thalassaemia forms (e.g. $--^{SEA}$, $--^{THAI}$) are common and result in HbH disease when inherited with α^+-thalassaemia (e.g. $-\alpha^{3.7}$ or $-\alpha^{4.2}$) or α^{ND}-thalassaemia (e.g. Hb Constant Spring or Hb Paksé), or in Hb Bart's hydrops fetalis when inherited from both parents.[14,15] Previously, HbH disease was considered to be relatively benign; however, recent evidence suggests a spectrum of mild to severe forms of HbH disease, with the worst affected individuals requiring lifelong transfusion.[16-18] The most severe form of α-thalassaemia, Hb Bart's hydrops fetalis, is almost always fatal *in utero* or soon after birth, although intrauterine interventions and perinatal intensive care can lead to survival.[19]

In this context, there is a growing demand for a better understanding of the epidemiology of α-thalassaemia such that burden estimates can be calculated to guide public health decisions and assess the market for new treatments from the pharmaceutical industry. However, whilst several narrative reviews of the epidemiology of α-thalassaemia in Southeast Asian countries are available,[15,20,21] a systematic review for the whole region has not been performed, making the current evidence-base patchy and incohesive. In addition, there appears to be a substantial amount of data that are available only in local journals, which is not being accessed by the international community. Moreover, despite

limited data sources, a lack of precision and various inconsistencies, estimates of the number of newborns with severe forms of α-thalassaemia published by Modell and Darlison currently represent the only comprehensive source of information on the epidemiology of thalassaemias and other inherited haemoglobin disorders.[22] As shown for other genetic conditions, sub-national newborn estimates can be generated by combining continuous allele frequency maps, interpolated from population surveys using geostatistical techniques, with high-resolution demographic and birth rate data.

The aims of this study are therefore two-fold: i) to comprehensively and rigorously evaluate and summarise our knowledge of the distribution of α-thalassaemia and its common genetic variants in Southeast Asia, and ii) to generate the first model-based continuous maps of α-thalassaemia and calculate refined estimates of the annual number of newborns affected by severe forms of α-thalassaemia in Thailand.

2.3. Methods

Compiling a geodatabase of α-thalassaemia allele frequency and genetic diversity

A comprehensive search of three major online bibliographic databases (PubMed, ISI Web of Knowledge and Scopus) was performed to identify published surveys of α-thalassaemia prevalence and/or genetic diversity in Southeast Asia. The 10 member states of the Association of Southeast Asian Nations (ASEAN) were used to define the region under study and include: Brunei Darussalam, Cambodia, Indonesia, Lao PDR, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam (Figure 2.1). In addition, for Thailand, articles published in national journals (in Thai) – not included in international bibliographic databases – were manually searched for local surveys. Consistent and pre-defined sets of inclusion criteria for prevalence/allele frequency data and genetic diversity

data, outlined in the Supplementary methods S2.1, were used to identify relevant surveys. Data extracted from Thai journals were independently validated against the inclusion criteria by two of the authors (SE and CH). The assembled datasets are available on request.

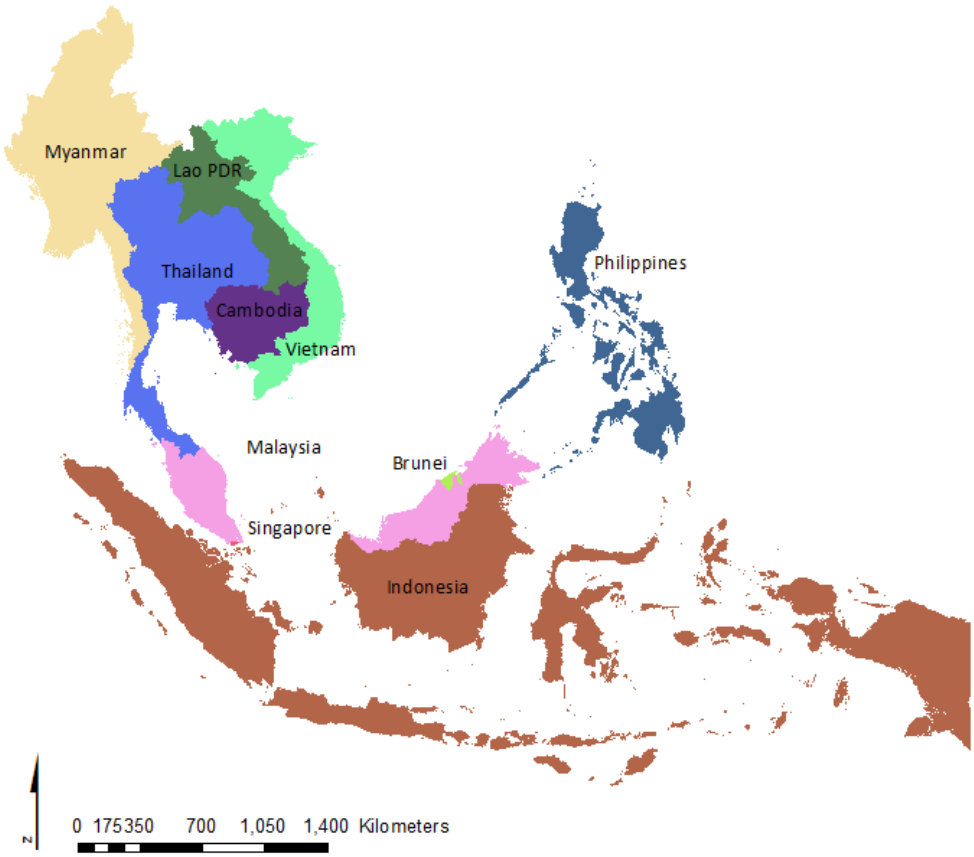


Figure 2.1 A map of the countries included in this study. Here we defined the Southeast Asian region according to the member states of the Association of Southeast Asian Nations (ASEAN) (*http://asean.org/asean/asean-member-states/*).

Summarising the current evidence-base for α-thalassaemia gene frequency in Southeast Asia

Cartographic representations of the identified prevalence surveys were generated using ArcGIS 10.4.1 (ESRI Inc., Redlands, CA, USA). The descriptive maps reflect the spatial distribution of the prevalence surveys, along with their respective sample sizes and observed α^{0}-, α^{+}- and α^{ND}-thalassaemia allele frequencies. Other features of the database,

including the temporal distribution of the surveys, the identity of the populations studied (e.g. community, pregnant women, newborns, etc.) and the contribution of local Thai surveys to the evidence-base, were also examined. In addition, pairwise univariate linear regression was performed to examine the association between the allele frequencies of the different forms of α-thalassaemia in the dataset, transformed on a log scale.

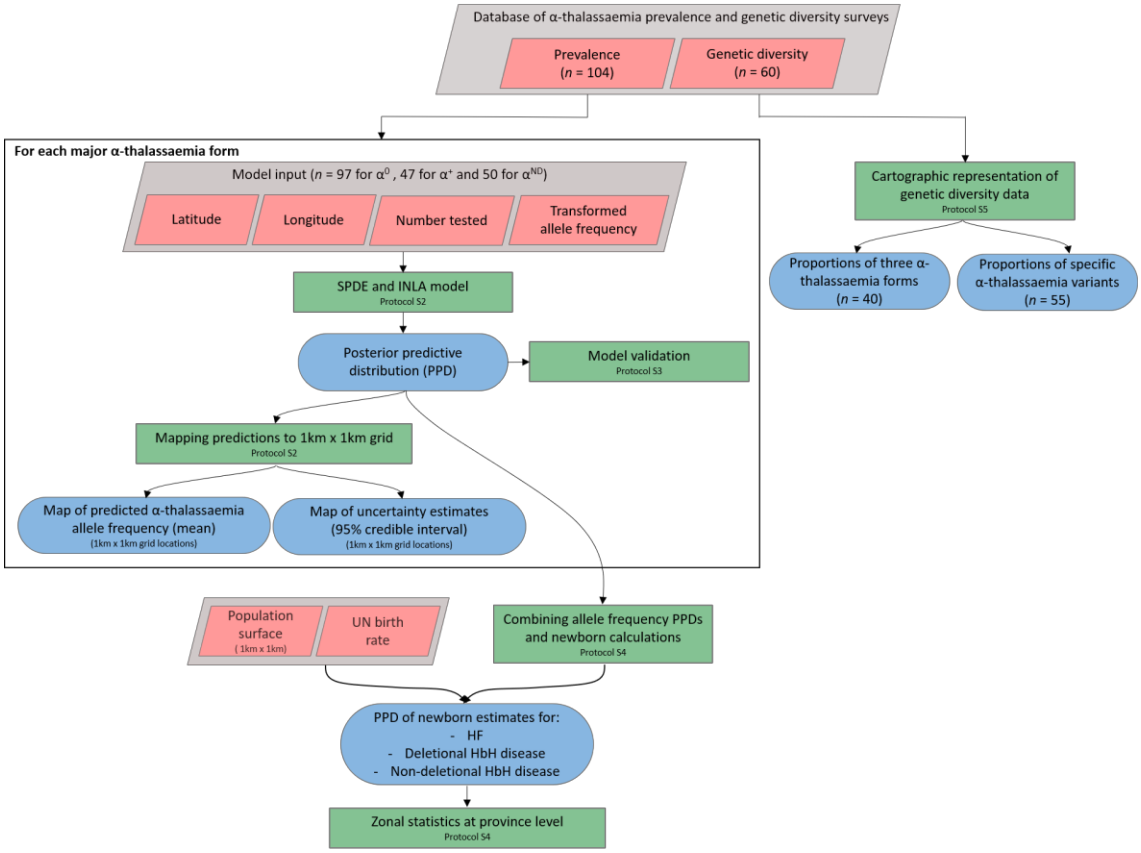


Figure 2.2 A schematic overview of the methodology used in this study and a breakdown of the data types analysed. Pink diamonds indicate the database and input data; green boxes denote model processes and data visualisation steps; blue rods represent study outputs.

Mapping α-thalassaemia genetic diversity

Maps of the genetic diversity of α-thalassaemia across Southeast Asia were also generated (Figure 2.2). Given the heterogeneity in the reporting of different α-thalassaemia genotypes, we divided the genetic diversity data into two subtypes: (i) those surveys that only distinguished between the different α-thalassaemia forms (α^0-, α^+-, and α^{ND}-

thalassaemia), and (ii) those surveys that contained detailed count data for a range of common mutations. We focused on the 11 mutations that are most commonly reported in Southeast Asia or are part of standard multiplex polymerase chain reaction (PCR) methods: $-\alpha^{3.7}$, $-\alpha^{4.2}$, $--^{SEA}$, $--^{THAI}$, $--^{MED}$, $--^{FIL}$, $-(\alpha)^{20.5}$, Hb Adana (HBA2:c.179G > A), Hb Constant Spring (HBA2:c.427T > C). Hb Paksé (HBA2:c.429A > T), Hb Quong Sze (HbA2:c.377T > C).[23] An "Other" category was used for other, rarer α-thalassaemia mutations. For the first data subtype, only surveys that tested for all three α-thalassaemia forms were mapped and the relative proportions of the different forms in the study sample were displayed using pie charts. For the latter, the same approach to that used in Howes *et al.* (2013) was used; the variant proportions were displayed using bar charts in which all variants that were explicitly tested for were included on the *x*-axis.[24] This allowed information regarding the suite of variants that were tested for in the survey to be displayed as well as unambiguous representation of the absence of a variant in the study sample.

Modelling continuous maps of α-thalassaemia allele frequency in Thailand

We employed a Bayesian geostatistical framework to model the allele frequencies of α^{0}-, α^{+}- and α^{ND}-thalassaemia, respectively, in Thailand, where a substantially higher number of surveys were identified. We included data from Thailand and its neighbouring countries (Myanmar, Lao PDR, Cambodia and Malaysia) in order to preclude the possibility of a border effect. Three surveys that were reported only at the national level (one in Thailand and two in Malaysia) were excluded for this part of the analysis. Only geographical location was included as a predictor of α-thalassaemia allele frequency (Figure 2.2).

For each of the three main forms of α-thalassaemia, a model was fitted using a Stochastic Partial Differential Equation (SPDE) approach with Integrated Nested Laplace Approximation (INLA) algorithms developed by Rue *et al*.,[25] available in an R-package (www.r-inla.org). The fitted model was then used to generate predictions at a resolution of 1km x 1km for α-thalassaemia allele frequencies for all unsampled locations in Thailand. Uncertainty estimates, measured as the 95% credible interval, for the predictions were calculated using 1,000 conditionally simulated realisations of the model to generate a posterior predictive distribution (PPD) for each 1km x 1km pixel. Full details of the modelling process and model validation procedures, which involved a 10-fold cross validation, are provided in the Supplementary methods S2.2.

Refining estimates of the annual number of neonates affected by severe disease forms

To generate estimates of the annual number of newborns affected by Hb Bart's hydrops fetalis syndrome (--/--) and deletional and non-deletional HbH disease (-α/-- and $\alpha\alpha^{ND}$/--, respectively) in Thailand in 2020, we paired the predicted allele frequency maps generated using our Bayesian geostatistical framework with high-resolution birth count data. First, we combined the three allele frequency maps to estimate the frequency of each genotype in each pixel, assuming Hardy-Weinberg proportions for a four-allele system (Equation 2.1 and Table 2.1).[26,27]

$$p^2 + 2pq + 2pr + q^2 + 2qr + 2qs + r^2 + 2rs + s^2 = 1 \qquad \text{(Equation 2.1)}$$

where, p is the allele frequency of α a^0-thalassaemia (--), q is the allele frequency of α^+-thalassaemia (-α), r is the allele frequency of α^{ND}-thalassaemia ($\alpha\alpha^{ND}$) and s is the allele frequency of the wild-type α-globin haplotype (αα).

Table 2.1 A breakdown of the genotypes for the three clinically important forms of α-thalassaemia – Hb Bart's hydrops fetalis, deletional HbH disease and non-deletional HbH disease – and the Hardy-Weinberg equilibrium (HWE) proportions used for their calculation. To compare our model output with previous newborn estimates for Hb Bart's hydrops fetalis and deletional HbH disease, we paired our allele frequency maps with 2003 demographic and birth data and included a measure of consanguinity in our calculations.

Genotype	Disorder	HWE proportions	Inclusion of population coefficient of consanguinity (F)
--/--	Hb Bart's hydrops fetalis	p^2	$p^2 + Fp(1 - p)$
-α/--	Deletional HbH disease	$2pq$	$2pq(1 - F)$
$\alpha\alpha^{ND}$/--	Non-deletional HbH disease	$2pr$	$2pr(1 - F)$

To calculate birth counts, the 2015-2020 crude birth rate for Thailand was downloaded from the 2017 United Nations (UN) world population prospects[28] and multiplied with a high-resolution predicted 2020 population surface, adjusted to UN population estimates, obtained from the WorldPop project (www.worldpop.org.uk, last accessed 23 January 2018).[29] The predicted genotype frequencies were then paired with the birth count data over 1,000 conditionally simulated realisations of the geostatistical model and areal estimates at province level calculated, together with 95% credible intervals; their calculation is described in Supplementary methods S2.3. We also applied our maps to 2003 demographic data, and incorporated consanguinity into our calculations (Table 2.1), in order to more directly compare estimates generated using our method with previous estimates.[22]

2.4. Results

The database

Our keyword searches yielded a total of 868 potential sources of data on α-thalassaemia prevalence and/or genetic diversity in the 10 southeast Asian countries included in this study. A further 74 potential data sources were identified from local Thai journals.

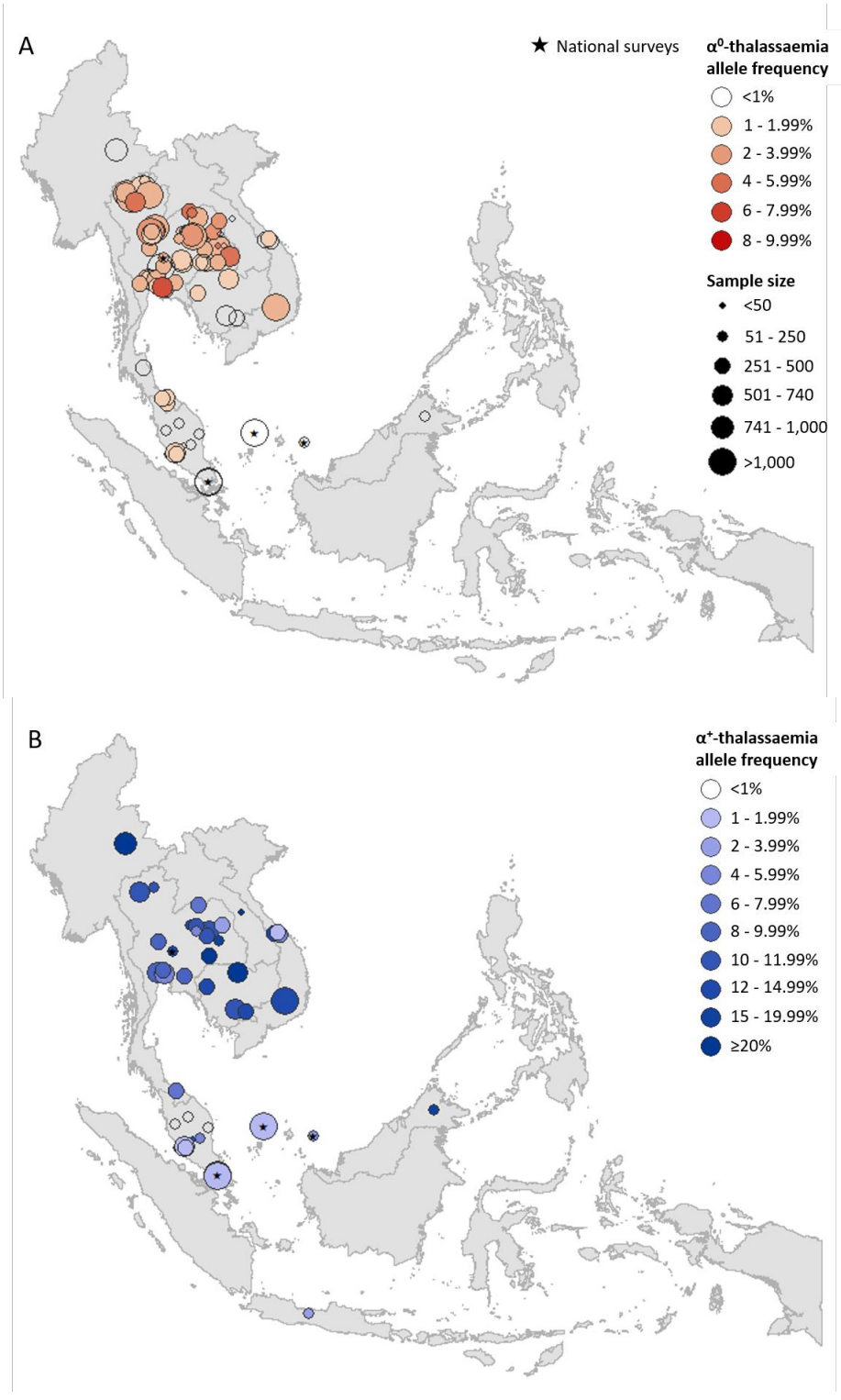
Following careful review, fifty-eight (54.7%) of the included surveys were obtained through our online search of the literature; 48 (45.3%) came from the local Thai literature. A total of 75 sources were therefore included in the final database, which corresponded to 106 population samples due to some sources reporting β^S from multiple locations or population groups. Forty-six surveys provided data on α-thalassaemia gene frequency alone, two provided data relating only to genetic diversity and 58 provided data on both. Four surveys were reported at the national level only (two from Malaysia, one from Singapore and one from Thailand), and were retained for the descriptive analysis.

The spatial and temporal distributions of all the surveys included in the final database are shown in Supplementary figure S2.1. The country for which the highest number of surveys had been published in the international literature (i.e. not including surveys obtained through local searches) was Thailand. No published surveys were identified for Brunei Darussalam or the Philippines. Within the other countries of the region, surveys predominated in certain areas. For instance, within Thailand, the northern and northeastern parts of the country contained >75% of all surveys, presumably due to these areas having the highest burden of α-thalassaemia. Data for southern Thailand came exclusively from Thai journals (n = 4) (Supplementary figure S2.1A). In terms of the population groups studied, 21.7% of surveys were community-based, 59.4% were carried out amongst selected but unbiased population groups (e.g. pregnant women, husbands of pregnant women, neonates and cord blood) and 8.5% in opt-out volunteers. One survey providing data on genetic diversity was carried out in individuals known to have α-thalassaemia. Twelve per cent of surveys provided no detailed information on the sampling methodology but were judged to be unbiased (based on other information provided in the article or personal communication with the corresponding author) and

were included. The majority of studies (95.3%) were carried out after 1995 and in particular from 2005 onwards (78.3%) (Supplementary figure S2.1B). The total number of individuals sampled was 133,649 (the population of the region is estimated to be 647,483,729 in 2017), with a mean sample size of 1,261. Sample size ranged from 4 to 55,796. Supplementary table S2.1 provides a breakdown of the survey features for the overall database and for prevalence surveys and genetic variant surveys, respectively.

Descriptive α-thalassaemia allele frequency maps for Southeast Asia

The identified prevalence surveys varied considerably with regards to the α-thalassaemia alleles and/or genotypes that were tested for or reported upon; whilst some reported allele frequencies for α^0-, α^+- and α^{ND}-thalassaemia, others provided data for only one or two of these. In order to maximise use of the available allele frequency data, whilst avoiding the incorporation of potentially biased estimates for overall α-thalassaemia allele frequency, we generated separate allele frequency maps for each of the three major forms of α-thalassaemia – that is, α^0-, α^+- and α^{ND}-thalassaemia (Figures 2.3A, B and C, respectively). α^0-thalassaemia was the most extensively studied form (n = 97), followed by α^{ND}-thalassaemia (n = 50) and then α^+-thalassaemia (n = 47). The number of surveys that tested for all three forms of α-thalassaemia was 40. Amongst these, the overall α-thalassaemia gene frequency ranged from 0% in populations from peninsular Malaysia to 35.4% in Preah Vihar, Cambodia.[30] A very high allele frequency of 49% was also reported in the So ethnic group from Khammouane Province in Lao PDR, although the sample size for this study was small (N = 50).[31] The mean overall α-thalassaemia gene frequency was 13.63%.


A
National surveys
α0-thalassaemia allele frequency
<1%
1 - 1.99%
2 - 3.99%
4 - 5.99%
6 - 7.99%
8 - 9.99%
Sample size
<50
51 - 250
251 - 500
501 - 740
741 - 1,000
>1,000
B
α+-thalassaemia allele frequency
<1%
1 - 1.99%
2 - 3.99%
4 - 5.99%
6 - 7.99%
8 - 9.99%
10 - 11.99%
12 - 14.99%
15 - 19.99%
≥20%

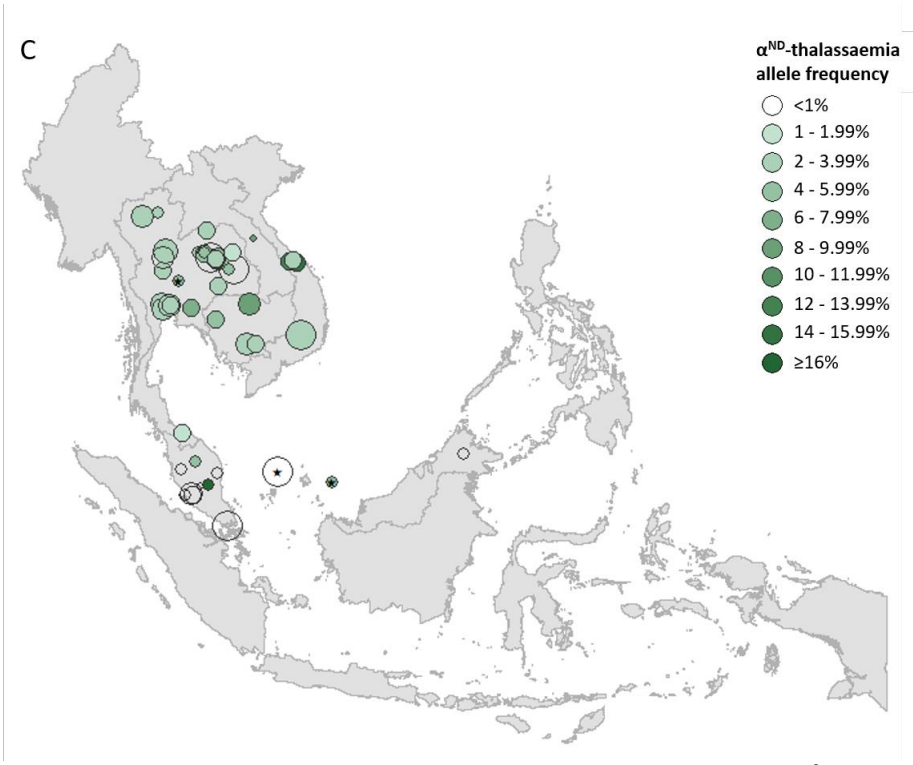


Figure 2.3 Descriptive maps of the observed allele frequencies in the database. (A) α^0-thalassaemia, (B) α^+-thalassaemia and (C) α^{ND}-thalassaemia (on next page). A spatial jitter of up to 0.3^0 latitude and longitude decimal degree coordinates was applied to allow visualisation of spatially duplicated data points. Colour intensity indicates allele frequency; circle size represents the size of the survey size.

For α^0-thalassaemia, the highest allele frequencies were observed in northern and northeastern Thailand and in provinces surrounding Bangkok (Figure 2.3A), although pockets of lower α^0-thalassaemia allele frequency (<2%) in these areas were also present. The highest reported allele frequency was 9.29% in a study amongst all couples attending antenatal screening in Chonburi province.[32] In Lao PDR, surveys along the Lao PDR-Thailand border near Vientiane reported allele frequencies between 4.03% and 7.28%, whilst the survey among the So ethnic group previously mentioned reported an absence of the α^0-thalassaemia allele. Allele frequencies were consistently reasonably low in Cambodia and Vietnam (≤2.66%), although data were sparse (n = 4 in each country). Allele frequencies of up to 1.53% were observed in southern Thailand, whilst the few surveys carried out in Myanmar (n = 1), Malaysia (n = 11) and Singapore (n = 2) reported allele frequencies of around 1%. In Malaysia, the highest allele frequency of α^0-

thalassaemia was 1.92% from a study carried out in newborns in Kuala Lumpur, the capital city.[33]

α^+-thalassaemia, the most prevalent form, reached allele frequencies of 26% in Cambodia (Figure 2.3B).[30] The surveys revealed a clear north-to-south decline in the distribution of α^+-thalassaemia across the region, with a single high allele frequency estimate of 16.8% observed in Sabah in Malaysian Borneo.[34] High allele frequencies (≥10%) were observed in all four surveys in Cambodia, whilst in Vietnam the reported α^+-thalassaemia allele frequency ranged from 1.59% to 14.4%.

α^{ND}-thalassaemia exhibited a heterogeneous distribution, with nearby surveys reporting vastly different α^{ND}-thalassaemia allele frequencies (Figure 2.3C). For instance, in Khon Kaen province, a study amongst couples attending antenatal care reported an allele frequency of 0.15%,[35] whilst a frequency of 7.3% was observed in a study amongst newborns in the same province.[36] In Cambodia and Vietnam, α^{ND}-thalassaemia allele frequencies of up to 8% and 14.3% were reported, respectively, in surveys in which α^0-thalassaemia was found to be absent, whilst in other parts of these countries, the two forms were found to co-exist at similar allele frequencies (e.g. around 2.5% in Binh Phuoc and Khanh Hoa provinces in Vietnam).

Univariate linear regression revealed no significant association between α^0-thalassaemia and α^+-thalassaemia ($p = 0.14$) nor between α^0-thalassaemia and α^{ND}-thalassaemia ($p = 0.82$). A statistically significant positive relationship was found between α^+-thalassaemia and α^{ND}-thalassaemia (Supplementary figure S2.2), with a coefficient of 0.66 (95% CI:

0.38 to 1.1; $p = 0.03$). Given the inter-survey heterogeneity in the variants tested for, the number of surveys included in each of the analyses was 43, 45 and 42, respectively.

Distribution of α-thalassaemia variants across Southeast Asia

Maps of the genetic diversity of α-thalassaemia across Southeast Asia are shown in Figures 2.4-2.6. Unsurprisingly, α^+-thalassaemia most consistently constituted the highest proportion of α-thalassaemia, although there were some surveys in which α^{ND}-thalassaemia was the predominant form (e.g. central Vietnam and in parts of Malaysia). In Figure 2.4, areas where the observed relative proportion of α^0-thalassaemia was greatest include: Chiang Mai and Phayao provinces in northwest Thailand, Kalasin in northeast Thailand, Vientiane in Lao PDR, Kuala Lumpur and Selangor in Malaysia, Singapore and Jakarta in Indonesia. The α^0-thalassaemia allele was absent in the survey from Malaysian Borneo as well as in central Vietnam and central Lao PDR. In certain areas, α^0- and α^{ND}-thalassaemia together accounted for the majority of α-thalassaemia (e.g. ~75% in Kalasin in Thailand, ~60% in Kuala Lumpur and Jakarta and ~53% in Khon Kaen and Chachoensao in Thailand and Vientiane in Lao PDR). Some of these areas also correspond to where the highest allele frequencies of these alleles are found, for example, northeastern Thailand and the Thailand-Lao PDR border.

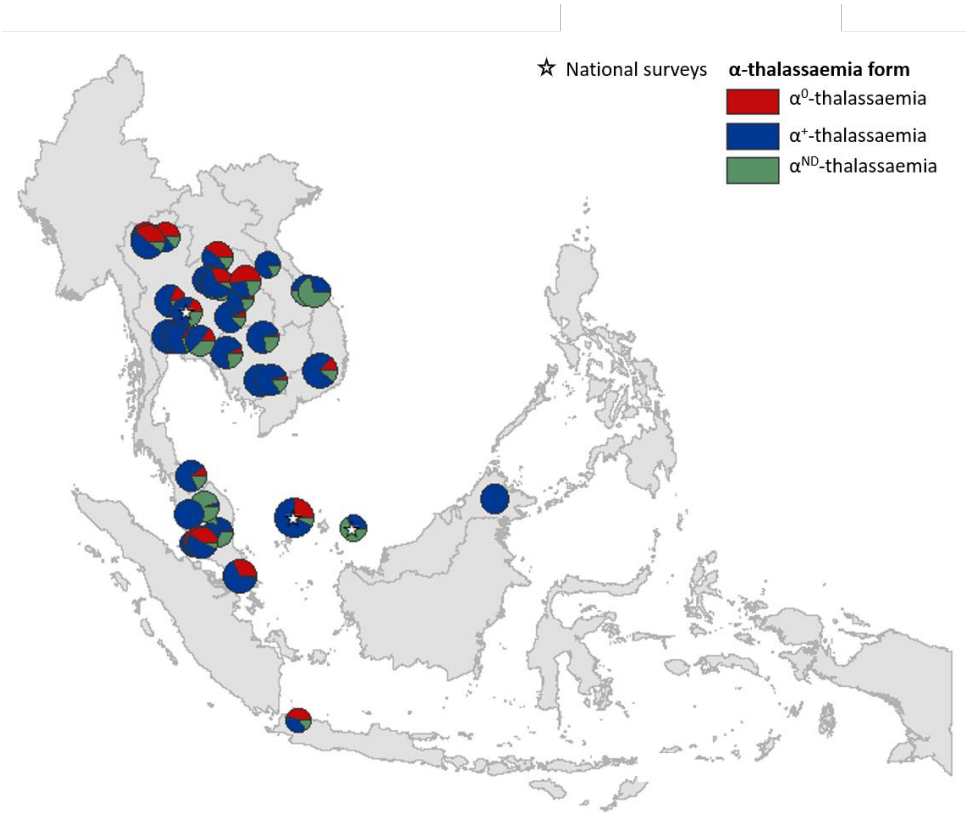


Figure 2.4 Map showing the proportions of α^0-, α^+- and α^{ND}-thalassaemia in Southeast Asia. Three surveys were mapped at the national level (indicated by a black star). The size of the pie charts reflects survey size on a log scale.

The most commonly tested variant throughout the region was $--^{SEA}$ (n = 44), followed by $-\alpha^{3.7}$ (n = 36) (Figures 2.5 & 2.6). Most of the studies in Thailand tested for a subset of the mutations included in this study; only one survey in Bangkok tested for the whole suite. More than in other countries, surveys in Thailand tested specifically for α^0- or α^{ND}-thalassaemia mutations (n = 7 and 9, respectively). Throughout the region, $--^{SEA}$ was the dominant α^0-thalassaemia mutation, and in the majority of surveys $-\alpha^{3.7}$ was the dominant α^+-thalassaemia and Hb Constant Spring the dominant α^{ND}-thalassaemia mutation. The only exceptions were in Java in Indonesia, where $-\alpha^{3.7}$ and $-\alpha^{4.2}$ were found in equal proportions and in Kelantan in Malaysia, where Hb Adana was the only α^{ND}-thalassaemia mutation identified. The $-(\alpha)^{20.5}$ and $--^{MED}$ mutations were not detected in any of the surveys, whilst the $--^{FIL}$ mutation was found in 2 of the 16 surveys in which it was tested

for and --^{THAI} in 9 of the 31 surveys in which it was included. Consistent with Figure 2.4, α^0-thalassaemia accounted for a small proportion of α-thalassaemia mutations in Myanmar, Cambodia and Vietnam and a high proportion in surveys along the Thailand-Lao PDR border. In Malaysia, Indonesia and Singapore, the proportion of α^0-thalassaemia mutation varied considerably, with it being absent in some areas and the predominant form in others. This is also true for α^{ND}-thalassaemia.

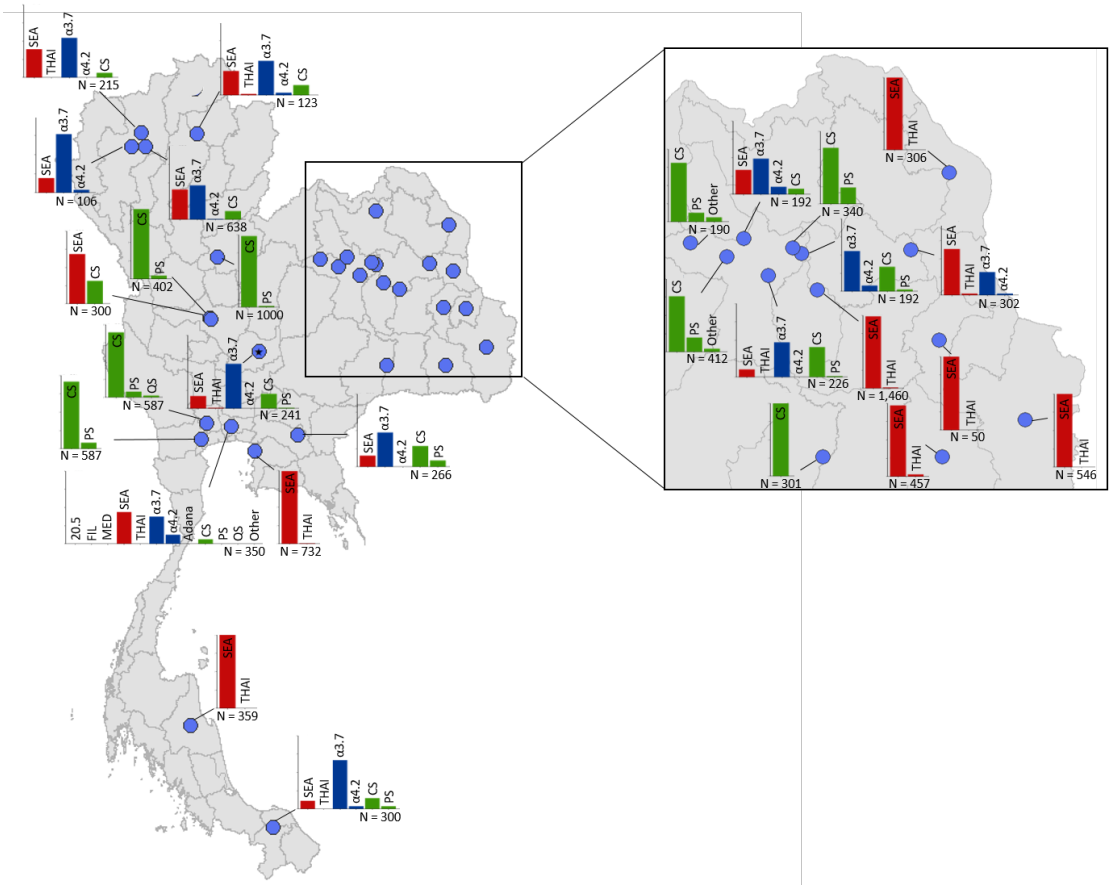


Figure 2.5 Map showing the proportions of specific α-thalassaemia variants in Thailand. Given the high number of surveys in northeast Thailand, this region has been magnified. The *y*-axis scale is the same across all bar charts, ranging from 0 to 1. The variants that were tested for in each survey are indicated above each bar. α^0-thalassaemia mutations are shown in red, α^+-thalassaemia mutations in blue and α^{ND}-thalassaemia mutations in green. Empty spaces along the *x*-axis indicate an absence of the corresponding mutation in the survey sample. The sample size of the survey is given under each plot. Bar charts are connected to their spatial location by a black line. Data points are coloured by country, using the same colour scale as that in Figure 2.1. No jitter in spatial location was included.

Continuous allele frequency maps for Thailand

Data for Thailand and its neighbouring countries (Myanmar, Lao PDR, Cambodia and Malaysia) formed the evidence-base for a Bayesian geostatistical model, which we used to generate spatially continuous maps of allele frequencies of α^0-, α^+- and α^{ND}-thalassaemia in Thailand (Figure 2.8). These are displayed using the mean of predicted values for each 1km x 1km pixel. The 95% credible interval was used to quantify model uncertainty.

The maps for α^0- and α^+-thalassaemia indicate clear spatial heterogeneity in the allele frequencies, although to varying degrees, whilst predicted allele frequencies of α^{ND}-thalassaemia range between 1.57% and 1.65% only. A north-to-south gradient of allele frequencies are predicted for the former two alleles; that is, highest predicted allele frequencies are found in the north and northeastern part of Thailand, whilst the lowest predicted allele frequencies are consistently found in the southern part of the country. The highest predicted allele frequencies of α^{ND}-thalassaemia are also found in the northern half of the country (i.e. north of Ratchaburi province) but are mainly found in central Thailand and the upper part of northeastern Thailand; however, as already mentioned, the model does not predict high heterogeneity in the frequency distribution of α^{ND}-thalassaemia and therefore the highest predicted allele frequencies are only ~1% greater than the lowest predicted allele frequencies.

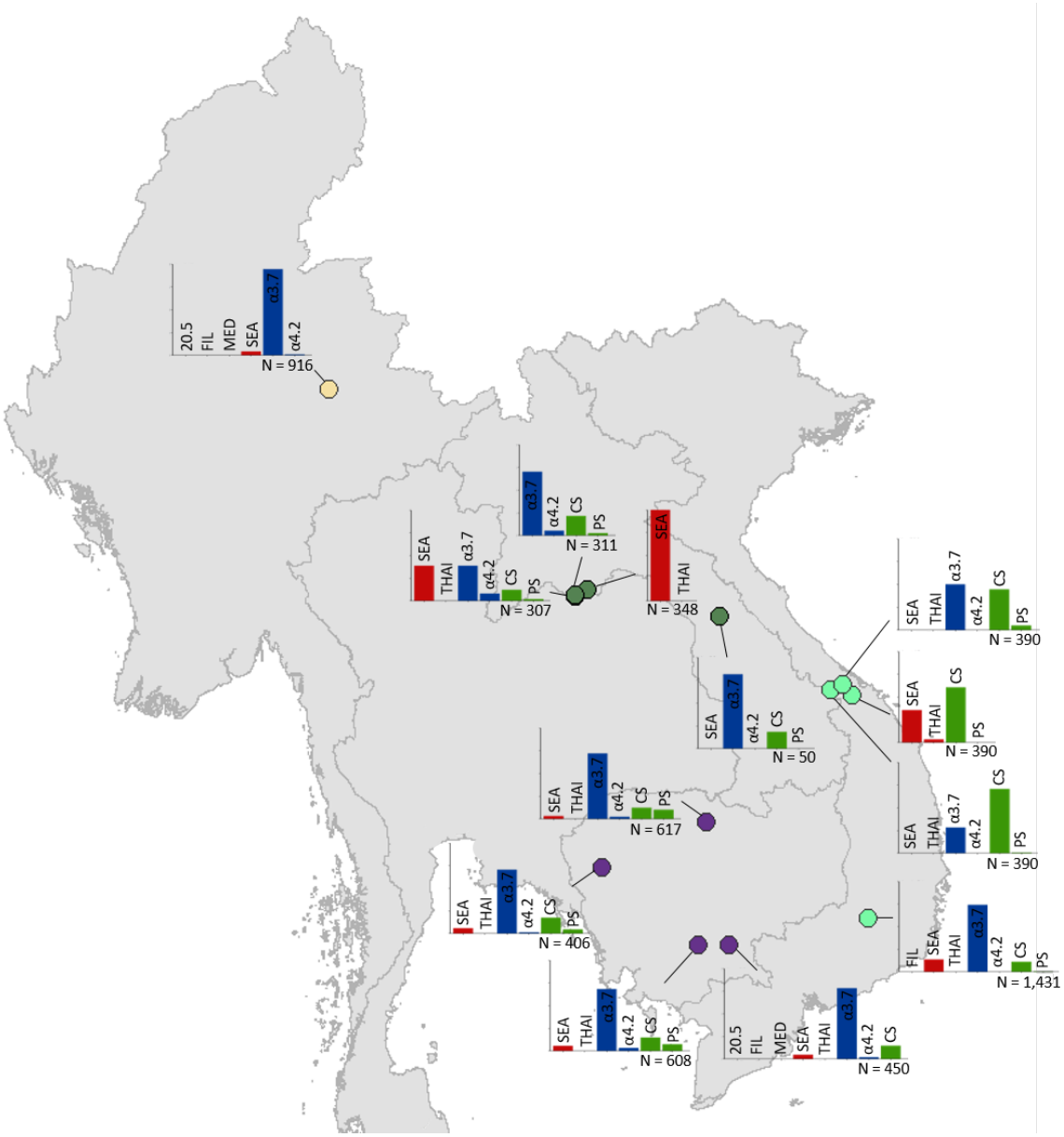


Figure 2.6 Map showing the proportions of specific α-thalassaemia variants in Myanmar, Lao PDR, Cambodia and Vietnam. The *y*-axis scale is the same across all bar charts, ranging from 0 to 1. The variants that were tested for in each survey are indicated above each bar. α^{0}-thalassaemia mutations are shown in red, α^{+}-thalassaemia mutations in blue and α^{ND}-thalassaemia mutations in green. Empty spaces along the *x*-axis indicate an absence of the corresponding mutation in the survey sample. The sample size of the survey is given under each plot. Bar charts are connected to their spatial location by a black line. Data points are coloured by country, using the same colour scale as that in Figure 2.1. No jitter in spatial location was included.

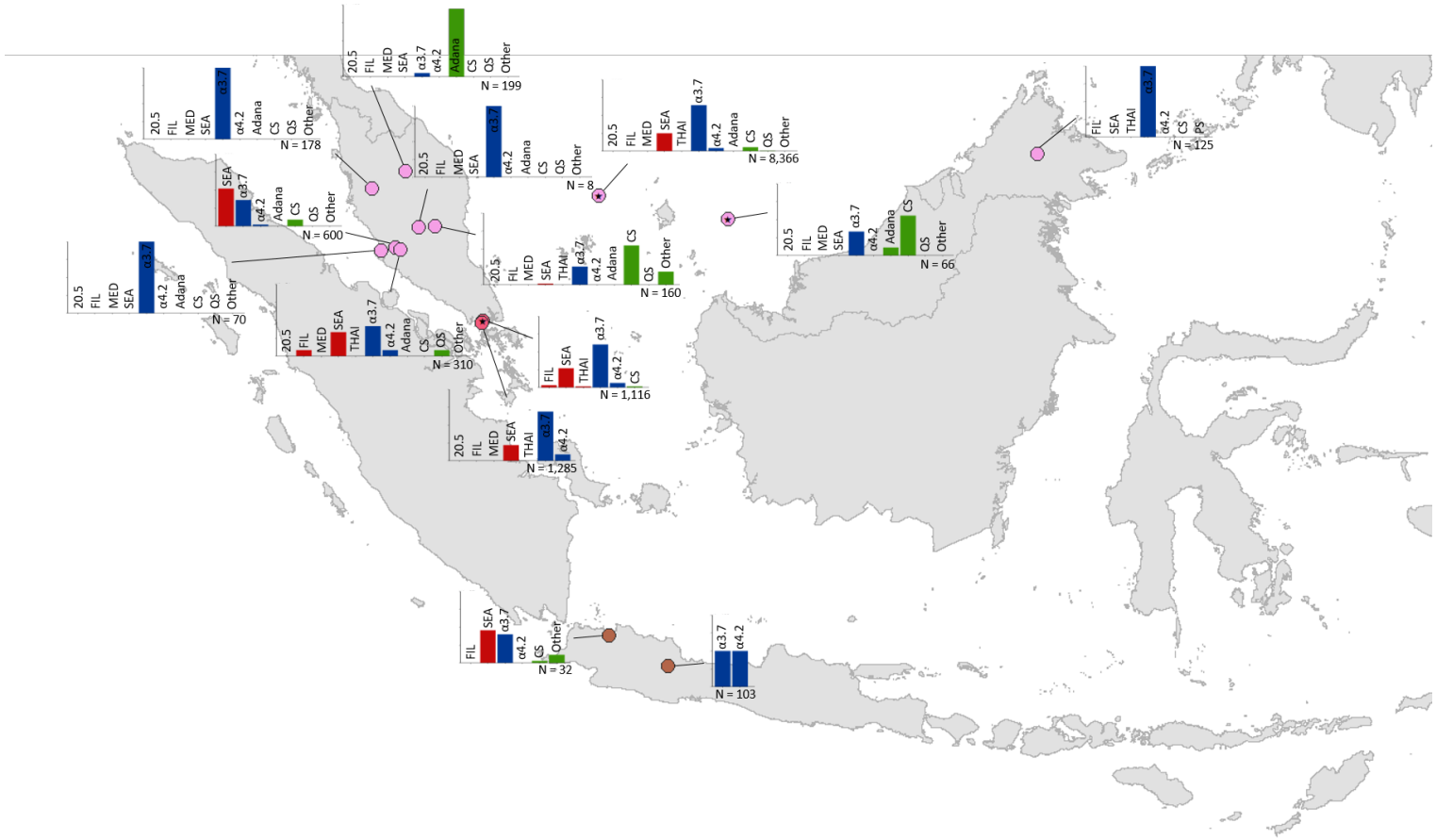


Figure 2.7 Map showing the proportions of specific α-thalassaemia variants in Malaysia, Singapore and Indonesia. The *y*-axis scale is the same across all bar charts, ranging from 0 to 1. The variants that were tested for in each survey are indicated above each bar. α^0-thalassaemia mutations are shown in red, α^+-thalassaemia mutations in blue and α^{ND}-thalassaemia mutations in green. Empty spaces along the *x*-axis indicate an absence of the corresponding mutation in the survey sample. The sample size of the survey is given under each plot. Bar charts are connected to their spatial location by a black line. Data points are coloured by country, using the same colour scale as that in Figure 2.1. No jitter in spatial location was included.

The modelled distribution of α^0-thalassaemia revealed a heterogeneous predicted distribution of the allele, particularly in the north of the country (Figure 2.7A). While large parts of the northernmost provinces of Chiang Rai, Phayao and Nan have predicted allele frequencies of up to 2%, allele frequencies for the neighbouring provinces of Chiang Mai, Lampang and Phrae can be twice as high. The allele is also predicted at allele frequencies of up to 4% in the northeast of the country, along a belt across most of the north of the country and in Chonburi and Rayong provinces in central Thailand. Allele frequencies below 1% are predicted throughout the southern zone. α^+-thalassaemia has its highest predicted allele frequencies across the whole of the north and northeastern zones. Its allele frequency ranges between 2.4% and 15.0%.

Model uncertainty was greatest in areas where data were scarce (e.g. southern Thailand and along the border with Myanmar) or where there was heterogeneity in the available data (e.g. in Chiang Mai and the surrounding area). For α^0-thalassaemia, the highest level of uncertainty was 9% and was found in Chumphon and Ranong provinces in southern Thailand and Kanchanaburi and Tak in the westernmost part of the country. For α^+-thalassaemia, the highest uncertainty (up to 23%) was observed in the northeastern zone and in the north. Very low levels of uncertainty were found in the southernmost tip of the country. Uncertainty for α^{ND}-thalassaemia was patchy and ranged from 1.5% in central and northern Thailand to 2.5% in southern and northeastern Thailand. The results of the 10-fold cross-validation procedure revealed an average mean absolute error of the predictions of 0.93%, 4.10% and 2.30%, for α^0-, α^+- and α^{ND}-thalassaemia, respectively. The average correlation between the observed and predicted values was 0.74 (0.62-0.83), 0.71 (0.49-0.85) and 0.47 (0.17-0.69), respectively.

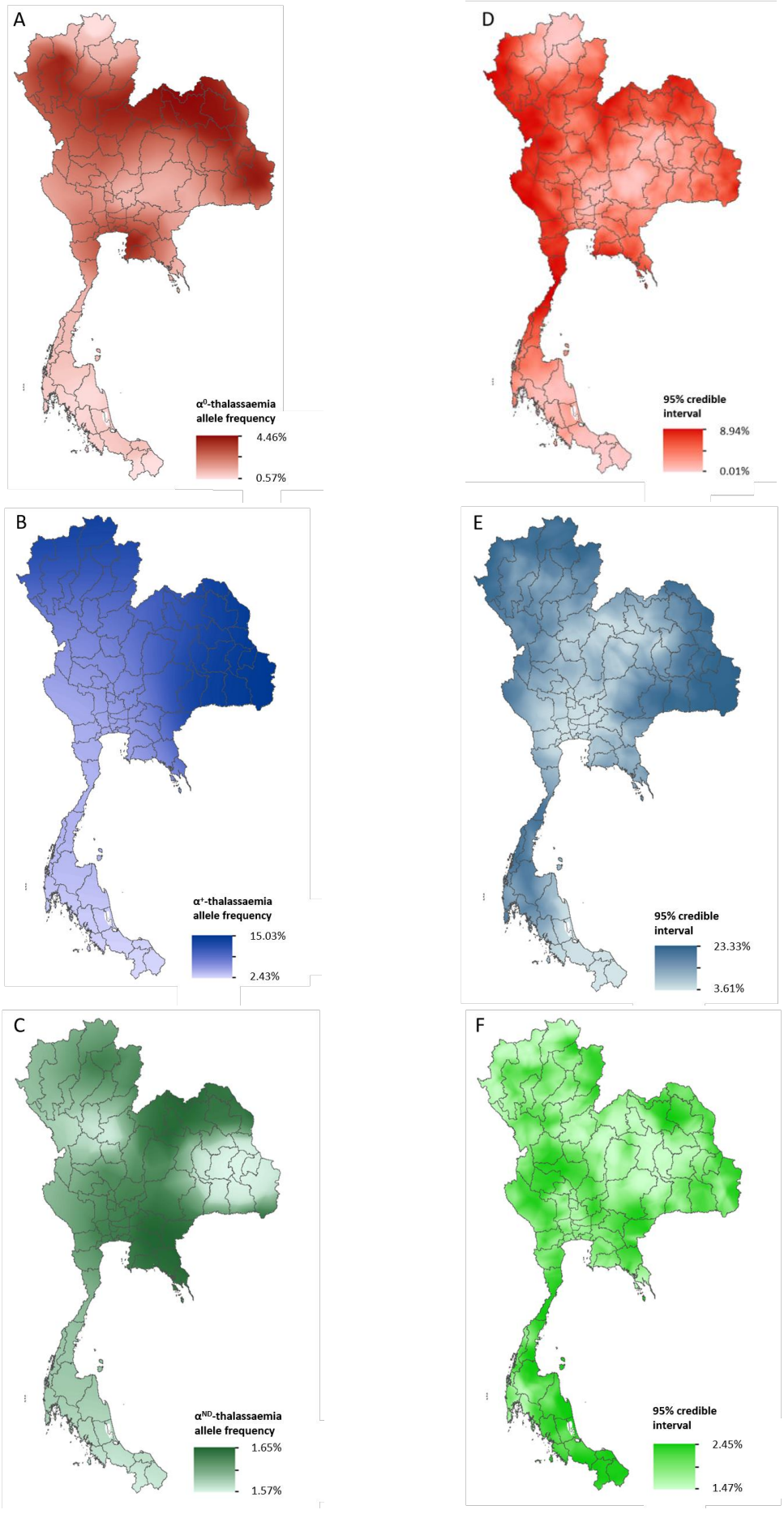

A
α0-thalassaemia allele frequency
4.46%
0.57%
D
95% credible interval
8.94%
0.01%
B
α+-thalassaemia allele frequency
15.03%
2.43%
E
95% credible interval
23.33%
3.61%
C
αND-thalassaemia allele frequency
1.65%
1.57%
F
95% credible interval
2.45%
1.47%

Figure 2.8. Maps of the mean and uncertainty in the predicted α-thalassaemia allele frequencies in Thailand. Panels A to C display the mean of the posterior predictive distribution (PPD). Panels D to F display the 95% credible interval of the PPD. Each row corresponds to a different α-thalassaemia form: α^0-thalassaemia (A and D); α^+-thalassaemia (B and E) and α^{ND}-thalassaemia (C and F).

Estimates of number of affected newborns in Thailand

Estimates of the number of newborns born with a severe form of α-thalassaemia in Thailand in 2020 were generated by pairing our allele frequency predictions to high-resolution demographic data for the country. We estimated that the number of Hb Bart's hydrops fetalis births in the country will be 423 (CI: 184 - 761). Whilst there are currently no estimates of the number of stillbirths that will occur in Thailand in 2020, our estimate of the number of Hb Bart's hyrops fetalis births represents more than 10% of the number of stillbirths estimated for 2015 (3,697).[37] We estimated that the number of new cases of HbH disease will be 3,172, including 2,674 (CI: 1,296 - 4,491) deletional and 498 (CI: 237 - 947) non-deletional cases, respectively. The highest absolute burden of hydrops fetalis was predicted in Bangkok (57 [CI: 13 - 151) and Udon Thani (23 [CI: 6 - 66]) in the northeastern zone, despite the former not having the highest allele frequencies of α^0-thalassaemia. This most likely reflects the high population size in the capital. Other states with a comparatively high burden include: Chiang Mai in the northwest of the country, Khon Kaen, Sakon Nakhon and Ubon Ratchathani in the northeast; and Chon Buri, Samut Prakan and Nonthaburi close to Bangkok. The estimated number of hydrops fetalis births in these provinces ranged between 10 and 19. For HbH disease, the highest burden was predicted in Northeast Thailand for both the deletional and non-deletional forms. Again, Bangkok was predicted to have the highest burden of HbH disease (301 [CI: 94 - 639] for deletional HbH disease and 68 [CI: 25 - 148] for non-deletional HbH disease).

Given the significant time lag between the 2003 estimates by Modell and Darlison and the ones presented here, their direct comparison is only possible by pairing our maps with 2003 population and birth rate data (Supplementary methods S2.3). Modell and Darlison estimated 1,017 and 2,515 births to be affected by Hb Bart's hydrops fetalis and HbH disease, respectively, in 2003. Using population data from the same year paired with our model-based maps, and assuming no consanguinity, we estimated 709 and 5,469 newborns to be born with Hb Bart's hydrops fetalis and HbH disease in the country. These estimates increase to 951 and 5,409 when a population coefficient of consanguinity of 0.01 is incorporated. These findings are consistent for Hb Bart's hydrops fetalis. However, they suggest that the burden of HbH disease in Thailand may have previously been substantially underestimated. Moreover, whilst Modell and Darlison did not estimate the burden of non-deletional forms of HbH disease, our estimates suggest that 848 of the 5,469 neonatal cases were of the non-deletional type.

2.5. Discussion

α-thalassaemia is a neglected public health problem whose burden has, to date, been largely overlooked, but for which morbidity is expected to increase in the coming decades as a result of the epidemiological transition.[1,38] Moreover, country reports (e.g. from Malaysia) indicate a shift in the age distribution of thalassaemia patients towards older ages.[39] As the burden increases, there will be greater demand for resources, including healthcare facilities and staff, genetic counselling and drugs, to treat and manage affected patients. This is particularly true for countries in Southeast Asia, where severe forms of α-thalassaemia are found at high prevalences.

Comparison with existing maps and population estimates

The descriptive maps presented here represent the first detailed cartographic representations of α-thalassaemia allele frequency estimates in Southeast Asia, which take into account the specific geographical location of the surveys in which they were observed. Until now, available maps (e.g. Supplementary figure S2.3) provided only a crude overview of overall α-thalassaemia gene frequency, without any distinction between different α-thalassaemia forms, and extrapolated to the entire region, thereby masking sub-national and even international, variation in allele frequencies.[1] The presented maps are broadly consistent with early narrative reviews of the gene frequency of α-thalassaemia in the region,[20] showing a clear north-to-south trend of decreasing allele frequencies of α^0- and α^+-thalassaemia and a patchier distribution of α^{ND}-thalassaemia. However, the maps also demonstrate a severe lack of data on the allele frequency of α-thalassaemia across large parts of Southeast Asia, including in Myanmar, northern Lao PDR, northern Vietnam, Indonesia, Philippines and Brunei. This impedes our ability to assess the fine-scale burden of α-thalassaemia, making efficient public health planning for its control difficult, and limits our ability to track progress in the prevention and management of the disorder.

The model-based maps for Thailand are, to our knowledge, the first spatially continuous maps of the distribution of α-thalassaemia in the country. Our newborn estimates represent the first evidence-based estimates of specific forms of α-thalassaemia disease amongst newborns since 2003 (although the study in which they were reported was published in 2008)[22] and the first estimates at sub-national level. Comparisons between previous newborn estimates and those generated in this study using our updated database and 2003 demographic data revealed an almost two-fold difference for deletional HbH

disease (2,515 compared to 4,694 in the present study). Reasons for such discrepancies most likely relate to: (i) differences in the inclusion criteria used in the generation of our map and, in turn, our calculation of newborn estimates, (ii) the quantity of survey data used, and iii) the statistical methods employed. For instance, spatial specificity was not a consideration in the study by Modell and Darlison, who used single allele frequency estimates extrapolated to the whole country. As such, the newborn calculations in the present study represent a methodological advance over previous efforts to assess the burden of α-thalassaemia among newborns. We related fine-scale allele frequency data to birth count data of equally high resolution, allowing location-specific estimates to be generated, which could be aggregated to province level. Moreover, the use of model-based maps in our calculations enabled the measurement of uncertainty in our predictions. Finally, by including allele frequency data on α^{ND}-thalassaemia, we were able to estimate the burden of the more severe non-deletional HbH disease.

Our newborn estimates for 2020 are considerably lower than those for 2003; in the absence of consanguinity, we estimate that 423 (CI: 194 – 761) births will be affected by Hb Bart's hydrops fetalis, 2,674 (CI: 1,296 – 4,490) by deletional HbH disease and 498 (CI: 237 – 947) by non-deletional HbH disease. This reduction is most likely due to the lower number of births in Thailand in 2020 compared to 2003 as a result of a decreasing birth rate and population size.[28] It would be interesting to quantify how improvements in the prevention of thalassaemias have affected and will affect these estimates in the future.

Public health implications of patterns of genetic variation

The pattern of genetic diversity observed in this study indicates variable distributions of mild and severe α-thalassaemia forms. Whilst the relative proportion of α^0-thalassaemia

is greatest in far north Thailand (i.e. in Chiang Mai and north of Khon Kaen), the majority of α-thalassaemia in Cambodia, southern Vietnam and central Thailand is of the α^{+}-thalassaemia form. High proportions of α^{0}-thalassaemia are also observed in a few studies from Peninsular Malaysia and a single study from Java in Indonesia. The proportion of α^{ND}-thalassaemia in the reported surveys is also high in parts of Northeast Thailand, central Thailand, northern Cambodia and central Vietnam. One survey in Peninsular Malaysia indicated α^{ND}-thalassaemia to be the predominant form. These spatially variable patterns have important implications for the design of newborn screening programmes with regards to the preferred diagnostic algorithm and allocation of treatment and management service provision. Areas with the highest proportions of co-occurring severe α-thalassaemia forms (i.e. α^{0}-thalassaemia and α^{ND}-thalassaemia) may experience a higher prevalence of the severe non-deletional form of HbH disease. Furthermore, the predominance of Hb Constant Spring in surveys from Malaysia and Vietnam suggests that the health burden of α-thalassaemia in these areas may be greater than previously thought. Hb Constant Spring is a mutation in the 5' α2-globin gene, which, in a normal individual, accounts for three-quarters of overall α-globin production. As a result, α2-globin gene mutations, such as Hb Constant Spring, tend to cause a more severe phenotype.[15]

Model strengths and limitations

The reliability of the model-based predicted maps is intrinsically linked to the quality, quantity and spatial coverage of the data upon which the models are based. Continuous maps for the whole of the Southeast Asian region were deliberately not generated as data were sparse in large areas. We are aware that unpublished surveys are likely to be available for most countries of the region. While we showed that a substantial additional

amount of survey data could be identified in local published sources, such a collaboration with local researchers requires solid foundations and is very time-consuming. As a result, this model could not be replicated in all of the other nine countries of the region.

For Thailand, limitations relating to data sparsity, uneven survey distributions and allele frequency heterogeneity can be quantified in the presented uncertainty intervals. Areas where there is little data or where observed allele frequencies are highly heterogeneous within a small geographical area will have more uncertain predictions, whilst a large amount of data for which there is little heterogeneity will lead to more precise predictions. We identified a serious lack of data in the southern part of Thailand, which is reflected in the larger uncertainty estimates. Other predictions with high associated uncertainty include those along the Thailand-Myanmar border. As there are no surveys on α^0-thalassaemia prevalence from Myanmar, this highlights the arbitrary nature of country borders in mapping studies.

Extensive variation in allele frequencies between different ethnic groups has been observed.[40] However, due to the nature of this study, and the smoothing of allele frequencies over a continuous spatial domain, the predicted allele frequencies might not fully reflect heterogeneity between local ethnic populations. For example, allele frequencies of around 3.65% for the Hb Constant Spring mutation have been reported in the Khmer ethnic group in Surin and Buriram provinces, whilst our model predicts maximum allele frequencies of 1.65% here. It is likely that other factors influence the allele frequencies of the different α-thalassaemia forms, which have not been considered in this mapping study, including ethnicity, consanguinity, rates of malaria (both *Plasmodium falciparum* and *P. vivax*)[41] and population migration patterns. Furthermore,

there is bound to be uncertainty in the geolocation of some of the surveys included in the study due to the lack of details published or available. This uncertainty could not be accounted for. Finally, whilst the inclusion criterion of molecular methods should help to improve the reliability of allele frequency estimates, they are not 100% sensitive[42] and do not cover all possible α-thalassaemia mutations, which may lead to some error in the reported allele frequencies.

Whilst we have calculated the burden of α-thalassaemia in terms of the number newborns born with severe forms in 2020, there are other aspects of the disease burden that would be worth considering pending the availability of more data, for example, milder-forms and their coinheritance with β-thalassaemia, DALY losses from α-thalassaemia, maternal complications (some of which can be life-threatening),[15,43] psychological effects and, in the case of HbH disease, survival data allowing the calculation of all-age population estimates. Furthermore, the estimates presented here do not include compound disorders, such as EA Bart's disease and EF Bart's (HbH disease with heterozygous and homozygous β^E, respectively).[44] Finally, the visualisation of our burden estimates are subject to the modifiable area unit problem, whereby the presentation of estimates at the province level likely masks pockets of high burden.[45]

Future prospects and conclusions

Despite the inclusion of α-thalassaemia in the Global Burden of Disease (GBD) Study, comprehensive and cohesive data on the allele frequency, distribution and burden of α-thalassaemia are still lacking. Here we have presented detailed maps of the reported allele frequencies of the three major forms of α-thalassaemia in Southeast Asia, based on a rigorous and up-to-date review of published surveys, along with a visual overview of the

genetic diversity patterns observed across the region. In addition, we have used Bayesian geostatistical modelling methods to generate evidence-based high-resolution maps of α-thalassaemia allele frequencies in Thailand and estimated the number of newborns that are expected to be born with a severe α-thalassaemia genotype in 2020. The presented maps confirm the heterogeneous spatial distribution of α-thalassaemia as well as the broad north-to-south trend of decreasing allele frequency reported here.[20,21] In addition, this study reveals the substantial amount of epidemiological data available in national data sources that are not accessible through searches of the international literature and highlights the key role of local collaborators for such studies. This illustrates that collaboration between international institutions, national screening programmes and field researchers and clinicians is essential to assembling the most complete evidence-base in the region.

The allele frequency, distribution and genetic variant profile of α-globin forms is only a part of their epidemiological complexity. An improved understanding of the natural history of α-thalassaemia and the factors that modify its clinical outcome will be imperative for establishing better estimates of its burden. This is particularly pertinent in the Southeast Asian region, where the disorder co-exists with β-thalassaemia and β^E. Many studies have shown a positive epistatic interaction between α- and β-thalassaemia, whereby their co-inheritance results in the amelioration of the associated blood disorder. Whether this interaction exists between α-thalassaemia and β^E is yet to be determined.

The presented maps raise certain population genetic questions. The heterogeneity in allele frequencies is likely to be the result of a series of complex environmental factors, including the rate of malaria transmission, and genetic and population genetic processes.

However, it is not immediately apparent as to why allele frequencies for all three forms should be higher in the north of the region compared to the south. It could simply be a reflection of the surveying effort, whereby areas that have historically been considered to have a high burden of α-thalassaemia are the same areas where many of the surveys are conducted. Intriguingly, visual comparisons of the distribution of α-thalassaemia and contemporary *Pf* malaria indicate that the latter has its highest prevalence in southern Thailand, Malaysian Borneo, and the Philippines.[46] A better understanding of these interactions will help in refining predicted allele frequency maps and, subsequently, estimates of populations affected. Population genetic models that incorporate various genetic, demographic and evolutionary processes may also help to shed light on this.

A detailed assessment of current knowledge on the allele frequency of α-thalassaemia and the magnitude of its health burden is needed to develop suitable prevention and control programmes. This study provides a detailed overview of the existing data on the gene frequency and genetic diversity of α-thalassaemia in Southeast Asia. We show that our knowledge of the accurate allele frequency and distribution of this highly complex disease remains somewhat limited. Because of the remarkable geographic heterogeneities in the gene frequency of α-thalassemia, interventions have to be tailored to the specific characteristics of the local population (e.g., prevalence of the disorder in the population, ethnic makeup, and consanguinity) and the local health care system As the epidemiological transition in these countries continues,[38,47,48] it will become increasingly important to regularly update regional and national maps of α-thalassaemia gene frequency and newborn estimates such that health and demographic changes can be properly quantified.[1] Our findings provide a baseline for such endeavours.

2.6. References

1. Piel FB, Weatherall DJ. The alpha-thalassemias. *The New England journal of medicine* 2014; **371**(20): 1908-16.
2. Flint J, Harding RM, Boyce AJ, Clegg JB. The population genetics of the haemoglobinopathies. *Bailliere's clinical haematology* 1998; **11**(1): 1-51.
3. Premawardhena A, Allen A, Piel F, et al. The evolutionary and clinical implications of the uneven distribution of the frequency of the inherited haemoglobin variants over short geographical distances. *British journal of haematology* 2017; **176**(3): 475-84.
4. Weatherall DJ, Clegg JB. Distribution and Population Genetics of the Thalassaemias. The Thalassaemia Syndromes: Blackwell Science Ltd; 2001: 237-84.
5. May J, Evans JA, Timmann C, et al. Hemoglobin variants and disease manifestations in severe falciparum malaria. *Jama* 2007; **297**(20): 2220-6.
6. Vichinsky EP. Clinical Manifestations of α-Thalassemia. *Cold Spring Harbor Perspectives in Medicine* 2013; **3**(5): a011742.
7. Hoppe CC. Newborn screening for non-sickling hemoglobinopathies. *Hematology American Society of Hematology Education Program* 2009: 19-25.
8. Weatherall DJ, Clegg JB. The Pathophysiology of the Thalassaemias. The Thalassaemia Syndromes: Blackwell Science Ltd; 2001: 192-236.
9. Galanello R, Cao A. Alpha-thalassemia.
10. Weatherall DJ, Wood WG, Pressley L, Higgs DR, Clegg JB. MOLECULAR BASIS FOR MILD FORMS OF HOMOZYGOUS BETA-THALASSAEMIA. *The Lancet* 1981; **317**(8219): 527-9.
11. Wainscoat JS, Kanavakis E, Wood WG, et al. Thalassaemia intermedia in Cyprus: the interaction of alpha and beta thalassaemia. *British journal of haematology* 1983; **53**(3): 411-6.
12. Wai Kan Y, G. Nathan D. Mild thalassemia: the result of interactions of alpha and beta thalassemia genes; 1970.
13. Law HY, Chee MK, Tan GP, Ng IS. The simultaneous presence of alpha- and beta-thalassaemia alleles: a pitfall of thalassaemia screening. *Community genetics* 2003; **6**(1): 14-21.
14. Weatherall DJ, Akinyanju O, Fucharoen S, Olivieri N, Musgrove P. Inherited Disorders of Hemoglobin. Disease Control Priorities in Developing Countries. 2nd edition ed. New York: Oxford University Press; 2006.
15. Chui DHK. Alpha-thalassaemia and population health in Southeast Asia. *Annals of Human Biology* 2005; **32**(2): 123-30.
16. Lal A, Goldrich ML, Haines DA, Azimi M, Singer ST, Vichinsky EP. Heterogeneity of Hemoglobin H Disease in Childhood. *New England Journal of Medicine* 2011; **364**(8): 710-8.
17. Fucharoen S, Viprakasit V. Hb H disease: clinical course and disease modifiers. *Hematology American Society of Hematology Education Program* 2009: 26-34.
18. Chui DH, Fucharoen S, Chan V. Hemoglobin H disease: not necessarily a benign disorder. *Blood* 2003; **101**(3): 791-800.
19. Songdej D, Babbs C, Higgs DR. An international registry of survivors with Hb Bart's hydrops fetalis syndrome. *Blood* 2017; **129**(10): 1251-9.
20. Fucharoen S, Winichagoon P. Hemoglobinopathies in Southeast Asia. *Hemoglobin* 1987; **11**(1): 65-88.
21. Fucharoen S, Winichagoon P. Haemoglobinopathies in southeast Asia. *Indian J Med Res* 2011; **134**: 498-506.

22. Modell B, Darlinson M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization* 2008; **86**(6): 480-7.
23. Liu YT, Old JM, Miles K, Fisher CA, Weatherall DJ, Clegg JB. Rapid detection of alpha-thalassaemia deletions and alpha-globin gene triplication by multiplex polymerase chain reactions. *British journal of haematology* 2000; **108**(2): 295-9.
24. Howes RE, Dewi M, Piel FB, et al. Spatial distribution of G6PD deficiency variants across malaria-endemic regions. *Malar J* 2013; **12**: 418.
25. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**(2): 319-92.
26. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science (New York, NY)* 1908; **28**(706): 49-50.
27. Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahresh Wuertt Verh Vaterl Naturkd; 1908: 369-82.
28. World population prospects: The 2017 revision.
29. Tatem AJ. WorldPop, open data for spatial demography. 2017; **4**: 170004.
30. Munkongdee T, Tanakulmas J, Butthep P, et al. Molecular Epidemiology of Hemoglobinopathies in Cambodia. *Hemoglobin* 2016; **40**(3): 163-7.
31. Sengchanh S, Sanguansermsri T, Horst D, Horst J, Flatz G. High frequency of alpha-thalassemia in the So ethnic group of south Laos. *Acta Haematologica* 2005; **114**(3): 164-6.
32. Limsakulsiriratt P, Oncoung W. High Risk Couples for Hb Bart's Hydrops Fetalis in Public Health Region 8 and 9 During 2004-2006. *Chonburi Hospital Journal* 2007; **32**(1): 9-14.
33. Alauddin H, Langa M, Mohd Yusoff M, et al. Detection of alpha-thalassaemia in neonates on cord blood and dried blood spot samples by capillary electrophoresis. *The Malaysian journal of pathology* 2017; **39**(1): 17-23.
34. Tan JAMA, Lee PC, Wee YC, et al. High prevalence of alpha- and beta-thalassemia in the kadazandusuns in east Malaysia: Challenges in providing effective health care for an indigenous group. *Journal of Biomedicine and Biotechnology* 2010; **2010**.
35. Soonklang M, Nonthalee S, Juntharaniyom M. Thalassemia and hemoglobinopathies in couples, Khon Kaen Hospital (Poster abstract). The 20th National Thalassemia Academic Symposium. Bangkok; 2014. p. Page P29.
36. Srivorakun H, Fucharoen G, Changtrakul Y, Komwilaisak P, Fucharoen S. Thalassemia and hemoglobinopathies in Southeast Asian newborns: diagnostic assessment using capillary electrophoresis system. *Clin Biochem* 2011; **44**(5-6): 406-11.
37. Blencowe H, Cousens S, Jassir FB, et al. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. *The Lancet Global Health*; **4**(2): e98-e108.
38. Weatherall D. The inherited disorders of haemoglobin: an increasingly neglected global health burden. *The Indian Journal of Medical Research* 2011; **134**(4): 493-7.
39. Ibrahim H. Current situation in control strategies and health systems in Asia - Malaysia. Malaysia: Ministry of Health 2012.
40. Kulaphisit M, Kampuansai J, Leecharoenkiat K, et al. A comprehensive ethnic-based analysis of alpha thalassaemia allelle frequency in northern Thailand. *Sci Rep* 2017; **7**(1): 4690.
41. Douglas NM, Anstey NM, Buffet PA, et al. The anaemia of Plasmodium vivax malaria. *Malaria Journal* 2012; **11**(1): 135.

42. Old J, Henderson S. Molecular diagnostics for haemoglobinopathies. *Expert Opinion on Medical Diagnostics* 2010; **4**(3): 225-40.
43. Ratanasiri T, Komwilaisak R, Sittivech A, Kleebkeaw P, Seejorn K. Incidence, causes and pregnancy outcomes of hydrops fetalis at Srinagarind Hospital, 1996-2005: a 10-year review. *Journal of the Medical Association of Thailand = Chotmaihet thangphaet* 2009; **92**(5): 594-9.
44. Galanello R. Screening and Diagnosis for Haemoglobin Disorders. In: Old J, ed. Prevention of Thalassaemias and Other Hemoglobin Disorders. Nicosia, Cyprus: Thalassaemia International Federation; 2013.
45. Wong D. The modifiable areal unit problem (MAUP). *The SAGE handbook of spatial analysis* 2009: 105-23.
46. Gething PW, Patil AP, Smith DL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria Journal* 2011; **10**: 378-.
47. Bundhamcharoen K, Odton P, Phulkerd S, Tangcharoensathien V. Burden of disease in Thailand: changes in health gap between 1999 and 2004. *BMC Public Health* 2011; **11**(1): 53.
48. Dhillon PK, Jeemon P, Arora NK, et al. Status of epidemiology in the WHO South-East Asia region: burden of disease, determinants of health and epidemiological research, workforce and training capacity. *International journal of epidemiology* 2012; **41**(3): 847-60.

2.7. Supplementary methods S2.1: Constructing a geodatabase of α-thalassaemia prevalence and/or genetic diversity surveys

Library assembly

Three online biomedical literature databases (PubMed, ISI Web of Knowledge and Scopus) were systematically searched for all records referring to "α-thalassaemia", "alpha-thalassaemia", "α-thal" or alpha-thal". Broad search terms were used to make the search as inclusive as possible. Retrieved articles were imported into the bibliographic database, Endnote X7 (Thomson Reuters, Carlsbad, CA, USA) and grouped by country. The last search was conducted on 21 July 2017. In addition, a non-systematic review of surveys available in local Thai journals was performed.

Survey inclusion criteria

The title and abstract of each reference was reviewed for its suitability to this study and those that were considered not to be so were excluded from further review. This included animal studies, review articles and studies performed elsewhere in the world. The remaining references had their full texts (where available) reviewed using a set of strict inclusion criteria, which varied slightly depending on the type of data being reported (i.e. prevalence data or genetic diversity data).

For prevalence data, details on the number of individuals sampled and the different α-thalassaemia alleles and/or genotypes identified were needed. Thus, surveys that reported allele or genotype frequency without the sample size were excluded as were those that reported an overall α-thalassaemia gene frequency without distinguishing between different α-thalassaemia forms or genotypes. Given the important clinical differences between α^{+}-, α^{0}- and α^{ND}-thalassaemia, many of the surveys focused on specific

genotypes, in particular those containing an α^0-thalassaemia allele (i.e. --/αα, --/--, -α/-- and $\alpha\alpha^{ND}$/--). As a result, allele frequencies of the other α-thalassaemia forms were not always a reliable representation of their true allele frequency in the study population, despite them being tested for. In this instance, the reported frequencies of the alleles of secondary interest were entered into the database as missing values. Similarly, if an α-thalassaemia allele was not tested for, the survey was still included but the frequency value of the untested allele was entered as missing. For surveys where the diagnostic algorithm and genotype reporting were sufficiently detailed (e.g. all seven major variants tested using PCR and a range of genotypes reported), we assumed that the lack of explicit reporting of certain genotypes reflected a zero frequency of them. Data on the absence of α-thalassaemia in the studied populations were also included. Where necessary, allele frequencies were derived from their respective genotype frequencies by assuming Hardy-Weinberg proportions.[1,2]

To be representative of the general population, surveys would take place in the community using random sampling. However, it became apparent early in the review process that very few surveys were conducted in this way. We therefore chose to include surveys that sampled from any of the following population groups, provided sampling was random, consecutive or universal: (i) the general community, (ii) pregnant women attending antenatal care and/or their husbands, (iii) cord blood and/or neonates in the absence of any inclusion/exclusion criteria or initial thalassaemia screening, (iv) individuals attending their yearly health check-up, and (v) opt-out volunteers, e.g. students, blood donors or army personnel. These groups were deemed to contain no inherent bias in α-thalassaemia prevalence and thus suitable for the purpose of this study. By contrast, surveys involving opt-in sampling, relatives of known cases of α-

thalassaemia, putative anaemic cases, patients or individuals possessing a β-globin variant (i.e. β-thalassaemia, β^E or β^S) were excluded. Moreover, surveys that were carried out in specific ethnic groups were only included if the population being investigated was representative of, or indigenous to, the study area.

Whilst a range of diagnostic methods are available for the diagnosis of α-thalassaemia,[3] we only included surveys in which molecular diagnostic methods were used (e.g. polymerase chain reaction (PCR)-based techniques and DNA sequencing).[4] This is because other methods are unreliable for characterising the precise α-thalassaemia genotype of an individual. For instance, haematological indices cannot distinguish between α^+-thalassaemia and the wild-type state nor between α^0-thalassaemia and iron deficiency anaemia (both of which lead to microcytosis).[3] Haemoglobin analysis by capillary electrophoresis (CE), high performance liquid chromatography (HPLC), isoelectric focusing (IEF) and citrate agar electrophoresis (CAE) are all techniques that have been used to identify the quantity of Hb Bart's, Hb H, Hb CS and other types of haemoglobin relevant to α-thalassaemia, and are comparable in performance.[5,6] However, accurate diagnosis of specific α-thalassaemia genotypes is not possible with these techniques and their performance is reduced when there is concomitant inheritance of β-thalassaemia.[5,7] Given that the objective of this study was to describe and map allele frequencies, the inclusion of non-molecular studies could affect the reliability of allele frequency estimates.

The majority of surveys that provided information on the genetic diversity of α-thalassaemia mutations were cross-sectional surveys in which there was no prior knowledge or screening of the α-thalassaemia status of the study sample. All of the same

inclusion criteria as those described above were applied – that is: (i) a clearly reported sample size and allele or genotype frequencies, (ii) a representative sample, and (iii) molecular diagnosis. In other surveys, the underlying mutations among individuals known to have α-thalassaemia were characterised. To avoid potential bias towards more severe variants in these surveys, only those that took place outside of the hospital setting and in the absence of any inclusion or exclusion criteria pertaining to disease severity were included. Surveys amongst other patient groups, for example β-thalassaemics, were also excluded. Again, only those surveys using molecular diagnostic methods were included.

Georeferencing

To capture fine-scale spatial heterogeneities in α-thalassaemia gene frequency and/or genetic diversity, included surveys were mapped to the highest spatial resolution possible based on the information provided in the article. Online geopositioning gazeteers were used to identify the latitude and longitude decimal degree coordinates of the study location. For studies that could be georeferenced to the province or district level only, the centroid of the polygon was extracted from ArcMap 10.4.1 (ESRI Inc., Redlands, CA, USA) and used. For studies that were conducted across multiple locations, the coordinates for each site were obtained from the online gazeteers and the centroid of all the sites calculated using ArcMap. In some cases, this resulted in a data point that fell between two land areas, for example Peninsular Malaysia and Malaysian Borneo.

2.8. Supplementary methods S2.2: Generation of continuous allele frequency maps for Thailand

Continuous maps of the allele frequencies of α^0-, α^+- and α^{ND}-thalassaemia were generated for Thailand only. This is because there was considerably more data for Thailand than for any of the other countries in the region, in part due to the inclusion of data from local journals. Data from Thailand and all of its neighbouring countries (Myanmar, Lao PDR, Cambodia and Malaysia) were used in this part of the analysis to preclude the possibility of a border effect on the predicted allele frequencies. Surveys that were reported only at the national level were excluded. To generate three separate maps of α-thalassaemia allele frequency, data on each α-thalassaemia form were extracted from the database and used as input for three separate models. The observed allele frequency data were transformed through an empirical logit, which, for databases ≥ 20 surveys, can be well approximated by a Gaussian likelihood.[8]

The goal of the geostatistical analysis was to estimate model parameters and generate predictions of α-thalassaemia allele frequencies in Thailand at fine spatial resolution. We employed, using the "r-inla" package, an easily available Bayesian geostatistical framework involving a stochastic partial differential equation (SPDE) approach[9] with an integrated nested Laplace approximation (INLA)[10] algorithm for Bayesian inference. The theoretical principles of this approach have been described in detail elsewhere.[9,11,12] A brief overview is provided below.

Let Y be the observed data, which in this study is the set of allele frequency estimates in Thailand and neighbouring countries. Each value y_i represents the allele frequency y at location i. In general, we can assume that Y is generated by an underlying Gaussian

process, denoted as *S(x)*, such that any evaluations of *S(x)* are multivariate normal distributied with a given mean and covariance function. Therefore, *S(x)* can be interpreted as a continuously indexed Gaussian field (GF), whereby the effect at each location has a multivariate normal as

$$p(y|x,\theta) \sim \prod_{i=1}^{n} MVN(y_i|x_i,\theta)$$

where, *y* denotes allele frequency, *x* is the Gaussian random effect on *y*, *n* represents the number of spatially unique data points, *i* represents location and θ denotes the hyperparameters of *x* (i.e. mean μ and dispersion parameter *k*). Observations *y* are assumed to be conditionally independent, given *x* and θ.

The first law of geography states that: "everything is related to everything else, but near things are more related to distant things."[13] This phenomenon is termed spatial dependency, or autocorrelation, and its inclusion in spatial models can greatly improve predictive performance. Several methods for defining the covariance between spatial points exist; here we use the Matérn class of covariance function in which the correlation function is defined based on the Euclidean distance between locations.[14] Specifically, the stationary Matérn covariance function assumes that if we have two pairs of points separated by the same Euclidean distance *h*, both pairs have same correlation, with correlation monotonically decreasing as a function of *h*.

Matrix algebra operations on a continuously indexed GF are computationally intensive to model. To overcome this, we use a finite element solution of an SPDE approximation to the Matérn covariance function, resulting in a finite dimensional Gaussian Markov random field (GMRF). The finite element representation represents *S(x)* by means of a

basis function representation defined on a triangulation of the domain under study, represented as:[9]

$$S(x) = \sum_{g=1}^{G} \psi_g(x) S_g,$$

where, G is the number of vertices in the triangulation, $\psi_g(\cdot)$ are piecewise polynomial basis functions on each triangle.[15] The SPDE approach replaces the Matérn covariance function's dense covariance matrix by a sparse reduced rank matrix. Following the generation of the GMRF prior, the posterior distribution is approximated by an Integrate Nested Laplace Approximation (INLA), which uses a combination of analytical approximation and numerical integration methods to approximate the posterior distribution at each point in the GMRF. The model's fully Bayesian hierarchical formulation is as follows:

$$\langle Y_i | S, \theta \rangle \sim p\langle Y_i | S, \theta \rangle,$$

$$\langle S | \theta \rangle \sim N(0, Q(\theta)^{-1},$$

$$(\theta) \sim p(\theta),$$

where, again, Y denotes allele frequency (the observation variable), S denotes the underlying Gaussian field, θ denotes a vector of hyperparameters and Q is the sparse precision matrix.

One thousand conditional simulations of the model were performed, whereby all of the pixels in the map were jointly simulated such that spatial autocorrelation between the pixels was accounted for.[16] This generated PPDs of allele frequency for each pixel in a

1km x 1km grid of Thailand, which were then used to calculate the mean and 95% Bayesian credible intervals for the predicted allele frequencies.

Given the relatively small number of surveys in each dataset, the predictive ability of the model for each form of α-thalassaemia was assessed using a 10-fold cross validation procedure. For each form of α-thalassaemia, the original dataset was randomly partitioned into 10 equal-sized data subsets. In each of 10 separate cross-validation experiments, one of the data subsets was retained as the test set whilst the remaining nine data subsets were used as the training set. The disparity between the model predictions and the observed allele frequencies in the test set was quantified for each cross-validation iteration using: (i) the mean absolute error (MAE), defined as the average magnitude of errors in the predicted values, and (ii) the correlation. The average across the ten cross-validation experiments was then calculated.

R code for the geostatistical model was generated by SB and is available on request. The model was implemented in R using the "r-inla" package, while the predicted allele frequency and disease burden maps were generated in ArcMap.

2.9. Supplementary methods S2.3: Generating newborn estimates of Hb Bart's hydrops fetalis and HbH disease for Thailand

Newborn estimates of Hb Bart's hydrops fetalis and HbH disease (deletional and non-deletional) were generated by pairing our predicted allele frequency PPDs with population and birth rate data. First, the number of births in each pixel of a 1km x 1km grid of Thailand were calculated by multiplying high-resolution 2020 population count data obtained from the WorldPop project website (www.worldpop.org.uk) with the national birth rate. In the UN World Population Prospects 2017 Revision, birth rates were provided as low-, medium- and high-fertility variant projections for 5-year periods. To obtain 2020 estimates of the number of births in Thailand, we used the average of the two 5-year periods 2015-2020 and 2020-2025.

One-hundred realisations of the geostatistical models for α^0-, α^+- and α^{ND}-thalassaemia allele frequency were run in parallel. For each realisation of the models, the estimated genotype frequencies for Hb Bart's hydrops fetalis disease (--/--), deletional HbH disease (-α/--) and non-deletional HbH disease ($\alpha\alpha^{ND}$/--) were calculated using the predicted allele frequencies, assuming Hardy-Weinberg proportions for a four-allele system (see Equation 1 in the main text). We also examined the effect that the inclusion of consanguinity into the Hardy-Weinberg equations would have on the estimates (see Table 2.1 in the main text). Genotype frequencies were then multiplied by the number of births in the corresponding 1km x 1km pixel of the birth count map described above. A PPD for the number of births in each pixel was therefore generated, from which point estimates and uncertainty around these estimates were computed.

The lower bounds of our credible intervals were calculated using the 25% quantiles for genotype frequency estimates and the low-fertility variant for birth count, whilst the higher bounds were calculated using the 75% quantiles for genotype frequency estimates and birth count data based on the high-fertility variant.

In order to compare the geostatistical method employed in this study with methods used to generate previous α-thalassaemia disease estimates, we obtained 2003 population count data from the Global Rural-Urban Mapping Project (GRUMP) and paired this with the same birth rate as that used in Modell and Darlison (17 births per 1,000 population).[17] The same method described above was used; however, no uncertainty intervals were calculated as no corresponding low- and high-fertility variants for birth rate were reported in their study.

2.10. Supplementary methods S2.4: Generation of maps of genetic variation of α-thalassaemia

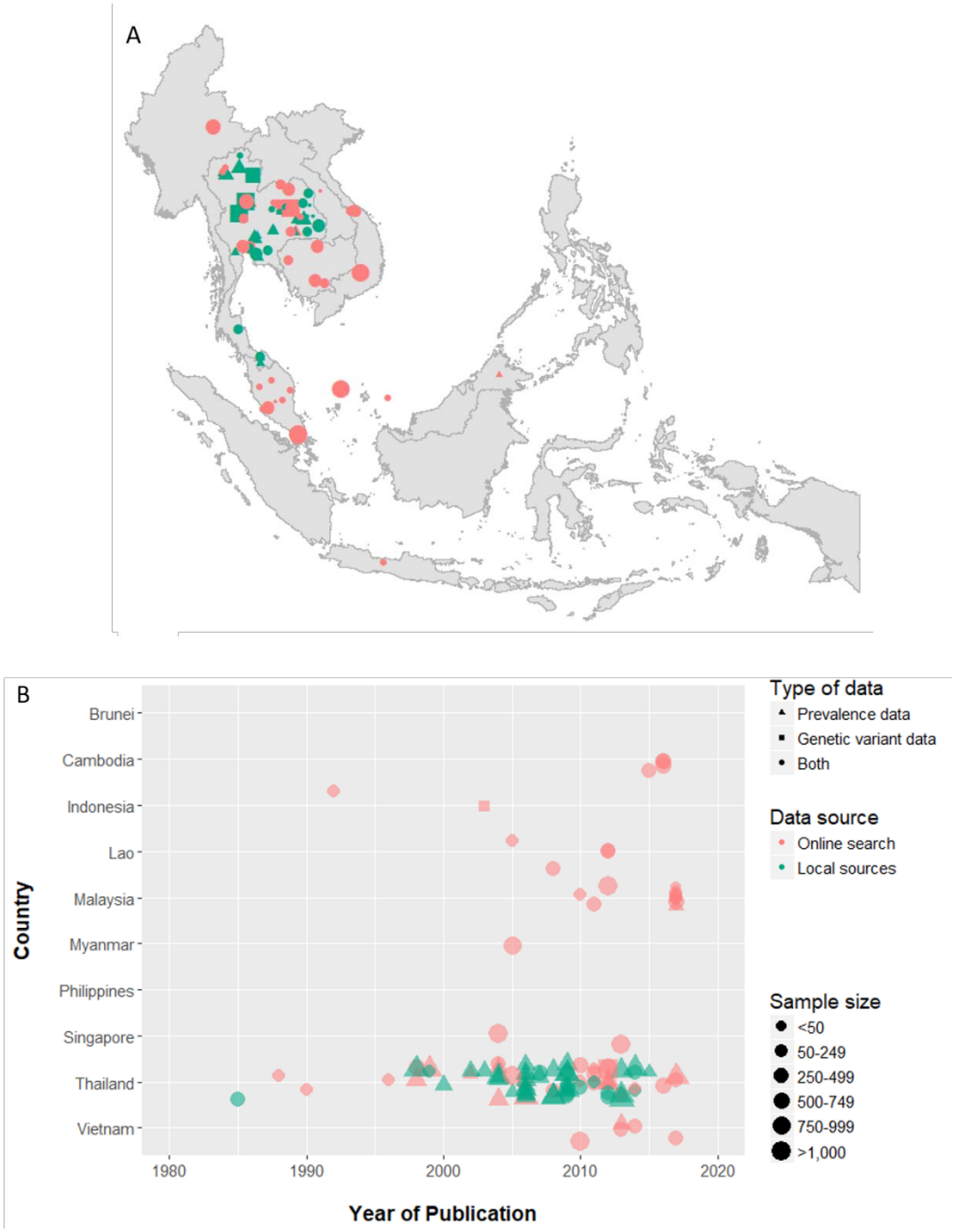
From a clinical perspective, the most important distinction between different α-thalassaemia alleles is the number of α-globin genes affected and whether the mutation is of the deletional or non-deletional type – i.e. the distinction between α^0-, α^+- and α^{ND}-thalassaemia.[18] However, molecular diagnosis of α-thalassaemia is mutation-specific and as such the profile of α-thalassaemia mutations that are found in a population determines which mutations are tested for in screening programmes. A better understanding of the genetic diversity patterns of α-thalassaemia at the mutation level therefore has important implications for programme design.[4,19] In this study, we include surveys that report on genetic diversity at both levels of specificity.

For the descriptive analysis of the relative proportions of each of the three major forms of α-thalassaemia (α^0-, α^+ and α^{ND}-thalassaemia), only surveys which tested for all three forms were included. This was to avoid misinterpretation of untested forms as being absent. The relative proportions of the different forms were represented using pie charts for which the denominator was the number of α-thalassaemia alleles in the study. The size of the pie charts reflects sample size on the log scale to allow clear visualisation of the data points. A jitter of 0.5-1º latitude and longitude decimal degree coordinates was applied to each survey to avoid the overlapping of surveys carried out in the same location.

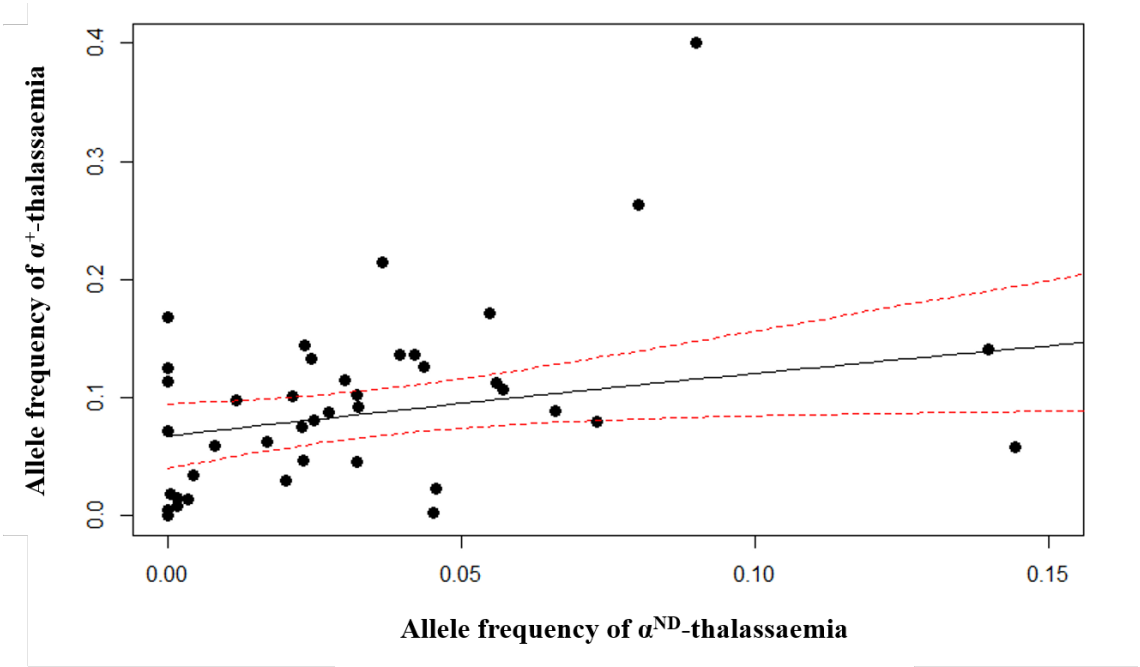
The subset of surveys that reported on specific α-thalassaemia mutations indicated that 11 variants were commonly tested for in the Southeast Asia region: single deletion mutations, $-\alpha^{3.7}$ and $-\alpha^{4.2}$; non-deletion mutations, Hb Adana, Hb Constant Spring, Hb

Paksé and Hb Quong Sze; and double deletion mutations, (α)20.5, FIL, MED, SEA and THAI. Other non-deletion mutations such as Hb Q-Thailand and initiation codon mutations were tested for in very few surveys (n = 13) and were grouped together into a single category as "Other". Surveys were mapped spatially using bar charts to depict variant proportions. All variants that were tested for in a survey were represented along the x-axis of the corresponding chart. Empty spaces are therefore indicative of the variant being absent in the sample. To allow clear visualisation of the bar charts, data for Thailand were displayed separately, whilst the remaining Southeast Asian countries were divided into two maps: (i) Myanmar, Lao PDR, Cambodia and Vietnam in the north, and (ii) Malaysia, Indonesia and Singapore in the south. Bar charts could not be placed at their precise geographical location and so surveys were mapped as points and connected to their corresponding bar chart by a black line.

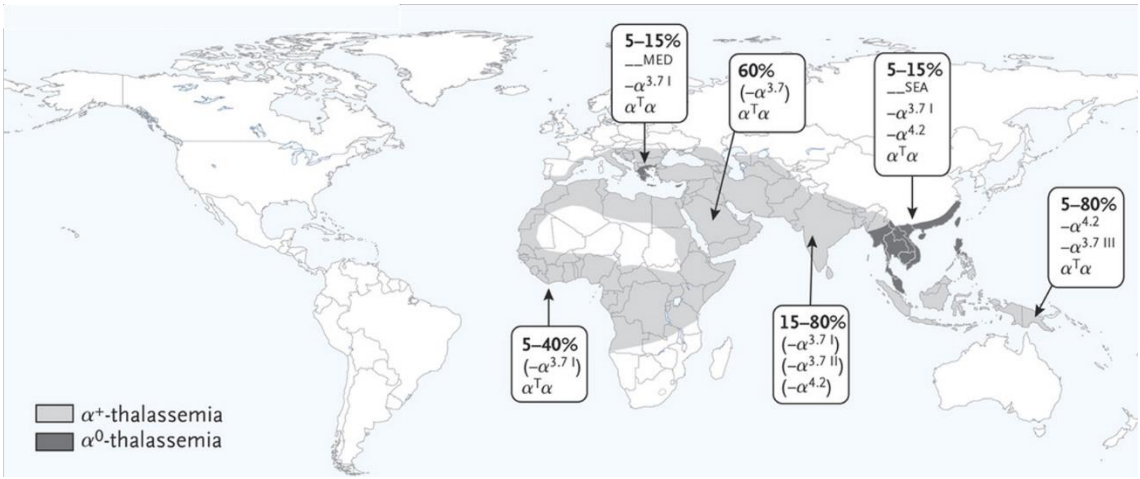
2.11. Supplementary results S2.1: Existing α-thalassaemia maps and characteristics of α-thalassaemia database



Supplementary figure S2.1 Spatial and temporal distributions of the α-thalassaemia surveys included in the final database. In both panels, the shape of the data points indicates the type of data provided by the survey, the colour indicates whether the survey was found in our online literature search or in local journals, and size represents the sample size of the survey. In (A) a spatial jitter of up to 0.3^0 latitude and longitude decimal degree coordinates was applied to allow visualisation of spatially duplicated data points.



Supplementary figure S2.2 Scatterplot showing the relationship between the allele frequency of α^{+}-thalassaemia and that of α^{ND}-thalassaemia in the Southeast Asia dataset, along with the line of best fit and its associated 95% confidence intervals.



Supplementary figure S2.3 A map of our current knowledge of the global distribution, gene frequency and genetic diversity of α-thalassemia. Only the most common variants for α^{+}-thalassemia ($-\alpha^{3.7}$and $-\alpha^{4.2}$) and α^{0}-thalassemia ($--^{MED}$ and $--^{SEA}$) are shown for each region. The variants that appear in parentheses are those for which the data used to make this map are limited. Reprinted with permission from Piel and Weatherall (2014),[18] Copyright Massachusetts Medical Society.

Supplementary table S2.1 Summary of the α-thalassaemia dataset characteristics according to the type of data provided (gene frequency data or genetic variant data), and overall. Numbers correspond to individual surveys that met the study inclusion criteria. As some sources reported more than one survey from multiple locations or in multiple population groups, the number of surveys is greater than the number of references in the Supplementary bibliographies S2.1 and S2.2. Some surveys reported data on both α-thalassaemia prevalence and genetic diversity and are therefore included twice in these columns, but once in the overall column.

	Gene frequency data	**Genetic diversity data**	**Overall**
Total surveys	104	60	106
Number of countries	8	8	8
Publication time			
1959 - 1969	0	0	0
1970 - 1979	0	0	0
1980 - 1989	2	2	2
1990 - 1999	8	5	8
2000 - 2009	46	15	47
2010 - 2017	47	38	48
N/A	1	0	1
Spatial extent			
Admin 0 centroids	4	4	4
Admin 1 centroids	27	12	27
Admin 2 centroids	10	8	9
Admin 3 centroids	3	3	3
Points	46	22	47
Multiple centroids/points	14	11	16
Total individuals sampled	132,157	32,237	133,649
Survey count by sample size			
≤50	3	2	4
51 - 250	22	18	22
251 - 500	38	23	38
501 - 750	21	10	21
751 - 1,000	3	1	3
>1,000	17	6	18

2.12. Supplementary bibliography S2.1

1. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science (New York, NY)* 1908; **28**(706): 49-50.
2. Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahresh Wuertt Verh Vaterl Naturkd; 1908: 369-82.
3. Fucharoen S, Winichagoon P. Haemoglobinopathies in southeast Asia. *Indian J Med Res* 2011; **134**: 498-506.
4. Old J, Henderson S. Molecular diagnostics for haemoglobinopathies. *Expert Opinion on Medical Diagnostics* 2010; **4**(3): 225-40.
5. Alauddin H, Langa M, Mohd Yusoff M, et al. Detection of alpha-thalassaemia in neonates on cord blood and dried blood spot samples by capillary electrophoresis. *The Malaysian journal of pathology* 2017; **39**(1): 17-23.
6. Mantikou E, Arkesteijn Sg Fau - Beckhoven van JM, Beckhoven van Jm Fau - Kerkhoffs J-L, Kerkhoffs Jl Fau - Harteveld CL, Harteveld Cl Fau - Giordano PC, Giordano PC. A brief review on newborn screening methods for hemoglobinopathies and preliminary results selecting beta thalassemia carriers at birth by quantitative estimation of the HbA fraction. (1873-2933 (Electronic)).
7. Chui DHK. Alpha-thalassaemia and population health in Southeast Asia. *Annals of Human Biology* 2005; **32**(2): 123-30.
8. Diggle PJ, Ribeiro PJ. Model-based Geostatistics. 1 ed: Springer-Verlag New York; 2007.
9. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011; **73**(4): 423-98.
10. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**(2): 319-92.
11. Rue H, Riebler A, Sørbye SH, Illian J, Simpson DP, Lindgren F. Bayesian Computing with INLA: A Review. *Annual Review of Statistics and its Application* 2017; **4**(1).
12. Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis* 2013; **97**(2): 109-31.
13. Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 1970; **46**: 234-40.
14. Krainski ET, Lindgren F, Simpson D, Rue R. The R-INLA tutorial on SPDE models. 2017. http://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/R/spde-tutorial-introduction.R.
15. Moraga P, Cano J, Baggaley RF, et al. Modelling the distribution and transmission intensity of lymphatic filariasis in sub-Saharan Africa prior to scaling up interventions: integrated use of geostatistical and mathematical modelling. *Parasites & Vectors* 2015; **8**(1): 560.
16. Gething PW, Patil AP, Hay SI. Quantifying Aggregated Uncertainty in Plasmodium falciparum Malaria Prevalence and Populations at Risk via Efficient Space-Time Geostatistical Joint Simulation. *PLOS Computational Biology* 2010; **6**(4): e1000724.
17. Modell B, Darlinson M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization* 2008; **86**(6): 480-7.
18. Piel FB, Weatherall DJ. The alpha-thalassemias. *The New England journal of medicine* 2014; **371**(20): 1908-16.

19. Galanello R. Screening and Diagnosis for Haemoglobin Disorders. In: Old J, ed. Prevention of Thalassaemias and Other Hemoglobin Disorders. Nicosia, Cyprus: Thalassaemia International Federation; 2013.

2.13. Supplementary bibliography S2.2: Allele frequency data

Sources from which the data points used in the allele maps were identified included:

1. Alauddin H, Langa M, Mohd Yusoff M, et al. Detection of alpha-thalassaemia in neonates on cord blood and dried blood spot samples by capillary electrophoresis. *The Malaysian journal of pathology* 2017; **39**(1): 17-23.
2. Amornkitbamrung S. Screening for Thalassemia Disease, α-thalassemia 1 trait, β-thalassemia trait and Hb E in pregnant women at Nongkhai Hospital. *Chonburi Hospital Journal* 2005; **30**(3): 171-86.
3. Apidechkul T. Prevalence of thalassemia carriers among the Lahu hill tribe population, Chiang Rai, Thailand. *Asian Biomedicine* 2015; **9**(4): 527-33.
4. Ausavarungnirun R, Winichagoon P, Fucharoen S, Epstein N, Simkins R. Detection of zeta-globin chains in the cord blood by ELISA (enzyme-linked immunosorbent assay): Rapid screening for alpha-thalassemia 1 (Southeast Asian type). *Am J Hematol* 1998; **57**(4): 283-6.
5. Chareonkul P, Kraisin J. Prevention and control of thalassemia at Saraburi Regional Hospital. *J Med Assoc Thai* 2004; **1**(8): 8-15.
6. Chinorose S, Toongkam K. Prevalence of Thalassemai Trait and Identification of Couples Risk in Pregnant Women for Severe Thalassemia Disease in Phayao Hospital. *Uttaradit Hospital Medical Bulletin* 2009; **24**(1): 61-9.
7. Dangwibul S, al. e. Efficiency of severe thalassemia screeing program in 5 rural hospitals in Roi Et province (Oral presentation, abstract). The 8th National Thalassemia Academic Symposium. Khon Kaen; 2002. p. 174-5.
8. Fucharoen G, al. e. Thalassemia and iron deficiency in subjects with positive screening OF test and KKU-DCIP-Clear. *Thai J Hematol Transf Med* 1999; **9**: 111-8.
9. Fucharoen G, Sanchaisuriya K, Sae-ung N, Dangwibul S, Fucharoen S. A simplified screening strategy for thalassaemia and haemoglobin E in rural communities in south-east Asia. *Bulletin of the World Health Organization* 2004; **82**(5): 364-72.
10. Fucharoen S, Winichagoon P, Wisedpanichkij R, et al. Prenatal and postnatal diagnoses of thalassemias and hemoglobinopathies by HPLC. *Clin Chem* 1998; **44**(4): 740-8.
11. Hundrieser J, Laig M, Yongvanit P, et al. Study of Alpha-Thalassemia in Northeastern Thailand at the DNA Level. *Hum Hered* 1990; **40**(2): 85-8.
12. Hundrieser J, Sanguansermsri T, Papp T, Flatz G. Alpha-Thalassemia in Northern Thailand - Frequency of Deletional Types Characterized at the DNA Level. *Hum Hered* 1988; **38**(4): 211-5.
13. Jameela S, Sabirah SO, Babam J, et al. Thalassaemia screening among students in a secondary school in Ampang, Malaysia. *The Medical journal of Malaysia* 2011; **66**(5): 522-4.
14. Jearakul W, Khamsaen J. Alpha-thalassemia 1 among Married Couples in Six Northeastern Provinces. *Journal of Health Science* 2009; **18**(5): 728-35.
15. Jindatanmanusan P, Riolueang S, Glomglao W, et al. Diagnostic applications of newborn screening for alpha-thalassaemias, haemoglobins E and H disorders using isoelectric focusing on dry blood spots. *Annals of clinical biochemistry* 2013.
16. Jopang Y. Thalassemia and hemoglobinopathies in anemic schoolchildren The 13th National Thalassemia Academic Symposium. Bangkok; 2007. p. Page 154.

17. Jopang Y. Thalassemia and hemoglobinopathies in anemic schoolchildren. Health Promoting Hospital, Regional Health Promotion Center 5. Nakhonratchasima. *Journal of the Medical Technologist Association of Thailand* 2008; **36**(1): 2235-41.
18. Jopang Y, Mernkratok S, Puangpiruk R. The efficacy of thalassemias and Hb E screening in first-trimester pregnant women at Health Promoting Hospital, Regional Health Promotion Center 5 Nakhonratchasima (poster abstract). The 10th National Thalassemia Academic Symposium. Bangkok; 2004. p. 171.
19. Karakochuk CD, Whitfield KC, Barr SI, et al. Genetic hemoglobin disorders rather than iron deficiency are a major predictor of hemoglobin concentration in women of reproductive age in rural prey veng, Cambodia. *Journal of Nutrition* 2015; **145**(1): 134-42.
20. Karnpean R, Pansuwan A, Fucharoen G, Fucharoen S. Evaluation of the URIT-2900 Automated Hematology Analyzer for screening of thalassemia and hemoglobinopathies in Southeast Asian populations. *Clin Biochem* 2011; **44**(10-11): 889-93.
21. Koh DXR, Raja Sabudin RZA, Mohd Yusoff M, et al. Molecular Characterisation of alpha- and beta-Thalassaemia among Indigenous Senoi Orang Asli Communities in Peninsular Malaysia. *Annals of human genetics* 2017.
22. LemmensZygulska M, Eigel A, Helbig B, Sanguansermsri T, Horst J, Flatz G. Prevalence of alpha-thalassemias in northern Thailand. *Hum Genet* 1996; **98**(3): 345-7.
23. Limsakulsiriratt P, Oncoung W. High Risk Couples for Hb Bart's Hydrops Fetalis in Public Health Region 8 and 9 During 2004-2006. *Chonburi Hospital Journal* 2007; **32**(1): 9-14.
24. Munkongdee T, Pichanun D, Butthep P, et al. Quantitative analysis of Hb Bart's in cord blood by capillary electrophoresis system. *Ann Hematol* 2011; **90**(7): 741-6.
25. Munkongdee T, Tanakulmas J, Butthep P, et al. Molecular Epidemiology of Hemoglobinopathies in Cambodia. *Hemoglobin* 2016; **40**(3): 163-7.
26. Nguyen HV, Sanchaisuriya K, Nguyen D, et al. Thalassemia and Hemoglobinopathies in Thua Thien Hue Province, Central Vietnam. *Hemoglobin* 2013; **37**(4): 333-42.
27. Nguyen NT, Sanchaisuriya K, Sanchaisuriya P, et al. Thalassemia and hemoglobinopathies in an ethnic minority group in Central Vietnam: implications to health burden and relationship between two ethnic minority groups. *Journal of Community Genetics* 2017: 1-8.
28. Nguyen VH, Sanchaisuriya K, Wongprachum K, et al. Hemoglobin Constant Spring is markedly high in women of an ethnic minority group in Vietnam: a community-based survey and hematologic features. *Blood cells, molecules & diseases* 2014; **52**(4): 161-5.
29. Nillakupt K, Nathalang O, Arnutti P, Jindadamrongwech SB, T., Panichkul S, Areekul W. Prevalence and hematological parameters of thalassemia in Tha Kradarn subdistrict Chachoengsao Province, Thailand. *J Med Assoc Thai* 2012; **95**(Suppl 5): S124-S32.
30. O'Riordan S, Hien TT, Miles K, et al. Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. *British journal of haematology* 2010; **150**(3): 359-64.

31. Panomai N, Sanchaisuriya K, Yamsri S, et al. Thalassemia and iron deficiency in a group of northeast Thai school children: relationship to the occurrence of anemia. *Eur J Pediatr* 2010; **169**(11): 1317-22.
32. Panyasai S, Cheechang S. The efficiency of screening for carriers of severe thalassemia in three community hospitals in Nakhon Si Thammarat province, Thailand. *Songkla Med J* 2009; **27**(1): 61-72.
33. Panyasai S, Sringam P, Fucharoen G, Sanchaisuriya K, Fucharoen S. A simplified screening for alpha-thalassemia 1 (SEA type) using a combination of a modified osmotic fragility test and a direct PCR on whole blood cell lysates. *Acta Haematol-Basel* 2002; **108**(2): 74-8.
34. Pharephan S, Sirivatanapa P, Makonkawkeyoon S, Tuntiwechapikul W, Makonkawkeyoon L. Prevalence of α-thalassaemia genotypes in pregnant women in northern Thailand. *Indian Journal of Medical Research* 2016; **143**(MARCH): 315-22.
35. Phollarp P, Tritipsombut J, Worasan C, et al. Thalassemia and iron deficiency among pregnant women attending antenatal care service at Khao Wong Hospital, Kalasin province. *Journal of Medical Technology and Physical Therapy* 2010; **22**(3): 262-70.
36. Pichanun D, Munkongdee T, Klamchuen S, et al. MOLECULAR SCREENING OF THE Hbs CONSTANT SPRING (codon 142, TAA > CAA, alpha 2) AND PAKSE (codon 142, TAA > TAT, alpha 2) MUTATIONS IN THAILAND. *Hemoglobin* 2010; **34**(6): 582-6.
37. Rahimah AN, Nisha S, Safiah B, et al. Distribution of alpha thalassaemia in 16 year old Malaysian Students in Penang, Melaka and Sabah. *The Medical journal of Malaysia* 2012; **67**(6): 565-70.
38. Rasri W. Prevention and control of seven thalassemia in pregnancy at Phayao Hospital. *Journal of Health Science* 2008; **17**: 477-84.
39. Rawangkran A, Janwithee N, Wong P, Jermnim N. Prevalence of Thalassemia Trait from Screening Program in Pregnant Women in the Lower Northern Region of Thailand. *Thai J Genet* 2013; **1**: 156-9.
40. Sanchaisuriya K, Fucharoen S, Ratanasiri T, et al. Thalassemia and hemoglobinopathies rather than iron deficiency are major causes of pregnancy-related anemia in northeast Thailand. *Blood Cell Mol Dis* 2006; **37**(1): 8-11.
41. Sangnark P. Prevalence of thalassemia and hemoglobinopathies in pregnant women at Bangkrathum Hospital, Phitsanulok province. *Buddhachinaraj Med J* 2009; **26**(1): 36-43.
42. Sanguansermsri T, Phumyu N, Chomchuen S, Steger HF. Screening for alpha-thalassemia-1 heterozygotes in expecting couples by the combination of a simple erythrocyte osmotic fragility test and a PCR- based method. *Community Genetics* 1999; **2**(1): 26-9.
43. Sanguansermsri T, Steger HF, Sirivatanapa P, Wanapirak C, Tongsong T. Prevention and Control of Severe Thalassemia Syndrome: Chiang Mai strategy. *Thai J Hematol Transf Med* 1998; **8**: 207-14.
44. Sattaratanamai C, Thongsutti S, Sucharitcheep P, Tuengsaeng D, Chomchuen S. Prevalence of Thalassemia and Hemoglobinopathies in Pregnant women at Surin Hospital. *Medical Journal of Srisaket Surin Buriram Hospitals* 2000; **15**(1): 1-12.
45. Savongsy O, Fucharoen S, Fucharoen G, Sanchaisuriya K, Sae-ung N. Thalassemia and hemoglobinopathies in pregnant Lao women: carrier screening, prevalence and molecular basis. *Ann Hematol* 2008; **87**(8): 647-54.

46. Sengchanh S, Sanguansermsri T, Horst D, Horst J, Flatz G. High frequency of alpha-thalassemia in the So ethnic group of south Laos. *Acta Haematol-Basel* 2005; **114**(3): 164-6.
47. Sirichotiyakul S, Tantipalakorn C, Sanguansermsri T, Wanapirak C, Tongsong T. Erythrocyte osmotic fragility test for screening of alpha-thalassemia-1 and beta-thalassemia trait in pregnancy. *Int J Gynecol Obstet* 2004; **86**(3): 347-50.
48. Soonklang M, Nonthalee S, Juntharaniyom M. Thalassemia and hemoglobinopathies in couples, Khon Kaen Hospital (Poster abstract). The 20th National Thalassemia Academic Symposium. Bangkok; 2014. p. Page P29.
49. Sornkayasit K, al. e. Incidence of Hb Constant Spring and Hb Pakse in Khon Kaen: Using capillary electrophoresis and DNA analysis (Poster abstract). The 18th National Thalassemia Academic Symposium. Nonthaburi; 2012. p. 90.
50. Srivorakun H, Fucharoen G, Changtrakul Y, Komwilaisak P, Fucharoen S. Thalassemia and hemoglobinopathies in Southeast Asian newborns: diagnostic assessment using capillary electrophoresis system. *Clin Biochem* 2011; **44**(5-6): 406-11.
51. Sukrat B, Sirichotiyakul S. The prevalence and causes of anemia during pregnancy in Maharaj Nakorn Chiang Mai Hospital. *J Med Assoc Thai 2006 Oct;89 Suppl 4:S142-6* 2006; **89**(Suppl 4): S142-6.
52. Sutjasung P, Fucharoen G, Fucharoen S, Chattumaruk P, Changtrakun D, Sanchaisuriya K. Effectiveness of thalassemia screening with the use of internal quality control blood samples at Kasetsomboon Hospital, Chaiyaphoom province. *JOURNAL OF MEDICAL TECHNOLOGY AND PHYSICAL THERAPY* 2011; **23**(1): 34-45.
53. Suwannakhon N, Seeratanachot T, Mahingsa K, Namwong P, T. S. Prevalence of Alpha-thalassemia Trait in the Volunteered Personals of University of Phayao. *J Hematol Transfus Med* 2014; **24**: 129-36.
54. Tan JA, Tay JS, Soemantri A, et al. Deletional types of alpha-thalassaemia in central Java. *Hum Hered* 1992; **42**(5): 289-92.
55. Tan JAMA, Lee PC, Wee YC, et al. High prevalence of alpha- and beta-thalassemia in the kadazandusuns in east Malaysia: Challenges in providing effective health care for an indigenous group. *Journal of Biomedicine and Biotechnology* 2010; **2010**.
56. Tangvarasittichai O, Jeenapongsa R, Sitthiworanan C, Sanguansermsri T. Laboratory investigations of Hb Constant Spring. *Clin Lab Haematol* 2005; **27**(1): 47-9.
57. Tangvarasittichai O, Poonanan N, Tangvarasittichai S. Using Red Cell Indices and Reticulocyte Parameters for Carrier Screening of Various Thalassemia Syndromes. *Indian journal of clinical biochemistry : IJCB* 2017; **32**(1): 61-7.
58. Tanphaichitr VS, Pung-amritt P, Puchaiwatananon O, et al. Studies on hemoglobin Bart's and deletion of alpha-globin genes from cord blood in Thailand (poster abstract). The International Conference on Thalassemia. Bangkok; 1985. p. P05.
59. Than AM, Harano T, Harano K, Myint AA, Ogino T, Okada S. High incidence of alpha-thalassemia, hemoglobin E, and glucose-6-phosphate dehydrogenase deficiency in populations of malaria-endemic southern Shan State, Myanmar. *Int J Hematol* 2005; **82**(2): 119-23.
60. Thanomrat P, Wannapira W, Sritippawan S, Boon-eam O. The prevalence of alpha-thalassemia trait in Buddhachinaraj Phitsanulok Hospital preliminary report. *Buddhachinaraj Med J* 2003; **20**(1): 19-25.

61. Thongon R, al. e. Strategy for thalassemia screening at Yala hospital (Poster abstract). The 15 th National Thalassemia Academic Symposium. Udon Thani; 2009. p. 144.
62. Tienthavorn V, Pattanapongsthorn J, Charoensak S, Sae-Tung R, Charoenkwan P, Sanguansermsri T. Prevalence of Thalassemia Carriers in Thailand. *Thai J Hematol Transf Med* 2006; **16**: 307-12.
63. Tongon R, Yunu R, Sanchaisuriya K, et al. Thalassemia and hemoglobinopathies in pregnant women attended antenatal care service at Yala Hospital. *J Med Tech Phy Ther* 2014; **26**(1): 32-9.
64. Traisrisilp K, Jatavan P, Tongsong T. A retrospective comparison of pregnancy outcomes between women with alpha-thalassaemia 1 trait and normal controls. *Journal of obstetrics and gynaecology : the journal of the Institute of Obstetrics and Gynaecology* 2017: 1-4.
65. Trisakul N, Siripulsak P, Chuesupalobol W. Prevalence of Thalassemia, Hemoglobinopathy, At-Risk Couples and Incidence of Thalassemia Major from the Screening Program, Prenatal and Postnatal Diagnosis at Yasothorn Hospital. *J Hematol Transfus Med* 2008; **19**: 285-92.
66. Tritipsombut J, Sanchaisuriya K, Fucharoen S, et al. Hemoglobin Profiles and Hematologic Features of Thalassemic Newborns Application to Screening of alpha-Thalassemia 1 and Hemoglobin E. *Arch Pathol Lab Med* 2008; **132**(11): 1739-45.
67. Tritipsombut J, Sanchaisuriya K, Phollarp P, et al. Micromapping of Thalassemia and Hemoglobinopathies in Diferent Regions of Northeast Thailand and Vientaine, Laos People's Democratic Republic. *Hemoglobin* 2012; **36**(1): 47-56.
68. Uaprasert N, Settapiboon R, Amornsiriwat S, et al. Diagnostic utility of isoelectric focusing and high performance liquid chromatography in neonatal cord blood screening for thalassemia and non-sickling hemoglobinopathies. *Clinica chimica acta; international journal of clinical chemistry* 2014; **427**: 23-6.
69. Wanapirak C, Muninthorn W, Sanguansermsri T, Dhananjayanonda P, Tongsong T. Prevalence of Thalassemia in pregnant women at Maharaj Nakorn Chiang Mai Hospital. *Journal of the Medical Association of Thailand* 2004; **87**(12): 1415-8.
70. Wong P, Thanormrat P, Srithipayawan S, et al. Risk of a couple having a child with severe thalassemia syndrome, prevalence in lower northern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health* 2006; **37**(2): 366-9.
71. Wong P, Thanormrat P, Srithipayawan S, et al. Prevalence of thalassemia trait from screening program in pregnant women of Phitsanulok. *Thai J Hematol Transf Med* 2004; **14**: 181-6.
72. Wongkham J, Ratanasiri T, Komwilaisak R, Saksiriwuttho P, Paibool M, Chatvised P. Thalassemia screening in pregnant women at antenatal care clinic, Srinagarind Hospital. *Srinagarind Med J* 2013; **28**(2): 170-7.
73. Yap ZM, Sun KM, Teo CRL, Tan ASC, Chong SS. Evidence of differential selection for the -alpha(3.7) and -alpha(4.2) single-alpha-globin gene deletions within the same population. *Eur J Haematol* 2013; **90**(3): 210-3.
74. Yin SKK, Chong QT, Mei LA, et al. A molecular epidemiologic study of thalassemia using newborns' cord blood in a multiracial Asian population in Singapore - Results and recommendations for a population screening program. *J Pediat Hematol Onc* 2004; **26**(12): 817-9.

2.14. Supplementary bibliography S2.3: Genetic variant data

Sources from which the data points used in the genetic variation maps were identified included:

1. Alauddin H, Langa M, Mohd Yusoff M, et al. Detection of alpha-thalassaemia in neonates on cord blood and dried blood spot samples by capillary electrophoresis. *The Malaysian journal of pathology* 2017; **39**(1): 17-23.
2. Fucharoen G, al. e. Thalassemia and iron deficiency in subjects with positive screening OF test and KKU-DCIP-Clear. *Thai J Hematol Transf Med* 1999; **9**: 111-8.
3. Fucharoen G, Sanchaisuriya K, Sae-ung N, Dangwibul S, Fucharoen S. A simplified screening strategy for thalassaemia and haemoglobin E in rural communities in south-east Asia. *Bulletin of the World Health Organization* 2004; **82**(5): 364-72.
4. Fucharoen S, Winichagoon P, Wisedpanichkij R, et al. Prenatal and postnatal diagnoses of thalassemias and hemoglobinopathies by HPLC. *Clin Chem* 1998; **44**(4): 740-8.
5. Hundrieser J, Laig M, Yongvanit P, et al. Study of Alpha-Thalassemia in Northeastern Thailand at the DNA Level. *Hum Hered* 1990; **40**(2): 85-8.
6. Hundrieser J, Sanguansermsri T, Papp T, Flatz G. Alpha-Thalassemia in Northern Thailand - Frequency of Deletional Types Characterized at the DNA Level. *Hum Hered* 1988; **38**(4): 211-5.
7. Jameela S, Sabirah SO, Babam J, et al. Thalassaemia screening among students in a secondary school in Ampang, Malaysia. *The Medical journal of Malaysia* 2011; **66**(5): 522-4.
8. Jearakul W, Khamsaen J. Alpha-thalassemia 1 among Married Couples in Six Northeastern Provinces. *Journal of Health Science* 2009; **18**(5): 728-35.
9. Jindatanmanusan P, Riolueang S, Glomglao W, et al. Diagnostic applications of newborn screening for alpha-thalassaemias, haemoglobins E and H disorders using isoelectric focusing on dry blood spots. *Annals of clinical biochemistry* 2013.
10. Karakochuk CD, Whitfield KC, Barr SI, et al. Genetic hemoglobin disorders rather than iron deficiency are a major predictor of hemoglobin concentration in women of reproductive age in rural prey veng, Cambodia. *Journal of Nutrition* 2015; **145**(1): 134-42.
11. Karnpean R, Pansuwan A, Fucharoen G, Fucharoen S. Evaluation of the URIT-2900 Automated Hematology Analyzer for screening of thalassemia and hemoglobinopathies in Southeast Asian populations. *Clin Biochem* 2011; **44**(10-11): 889-93.
12. Koh DXR, Raja Sabudin RZA, Mohd Yusoff M, et al. Molecular Characterisation of alpha- and beta-Thalassaemia among Indigenous Senoi Orang Asli Communities in Peninsular Malaysia. *Annals of human genetics* 2017.
13. LemmensZygulska M, Eigel A, Helbig B, Sanguansermsri T, Horst J, Flatz G. Prevalence of alpha-thalassemias in northern Thailand. *Hum Genet* 1996; **98**(3): 345-7.
14. Limsakulsiriratt P, Oncoung W. High Risk Couples for Hb Bart's Hydrops Fetalis in Public Health Region 8 and 9 During 2004-2006. *Chonburi Hospital Journal* 2007; **32**(1): 9-14.

15. Munkongdee T, Pichanun D, Butthep P, et al. Quantitative analysis of Hb Bart's in cord blood by capillary electrophoresis system. *Ann Hematol* 2011; **90**(7): 741-6.
16. Munkongdee T, Tanakulmas J, Butthep P, et al. Molecular Epidemiology of Hemoglobinopathies in Cambodia. *Hemoglobin* 2016; **40**(3): 163-7.
17. Nguyen HV, Sanchaisuriya K, Nguyen D, et al. Thalassemia and Hemoglobinopathies in Thua Thien Hue Province, Central Vietnam. *Hemoglobin* 2013; **37**(4): 333-42.
18. Nguyen NT, Sanchaisuriya K, Sanchaisuriya P, et al. Thalassemia and hemoglobinopathies in an ethnic minority group in Central Vietnam: implications to health burden and relationship between two ethnic minority groups. *Journal of Community Genetics* 2017: 1-8.
19. Nguyen VH, Sanchaisuriya K, Wongprachum K, et al. Hemoglobin Constant Spring is markedly high in women of an ethnic minority group in Vietnam: a community-based survey and hematologic features. *Blood cells, molecules & diseases* 2014; **52**(4): 161-5.
20. Nillakupt K, Nathalang O, Arnutti P, Jindadamrongwech SB, T., Panichkul S, Areekul W. Prevalence and hematological parameters of thalassemia in Tha Kradarn subdistrict Chachoengsao Province, Thailand. *J Med Assoc Thai* 2012; **95**(Suppl 5): S124-S32.
21. O'Riordan S, Hien TT, Miles K, et al. Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. *British journal of haematology* 2010; **150**(3): 359-64.
22. Panomai N, Sanchaisuriya K, Yamsri S, et al. Thalassemia and iron deficiency in a group of northeast Thai school children: relationship to the occurrence of anemia. *Eur J Pediatr* 2010; **169**(11): 1317-22.
23. Panyasai S, Cheechang S. The efficiency of screening for carriers of severe thalassemia in three community hospitals in Nakhon Si Thammarat province, Thailand. *Songkla Med J* 2009; **27**(1): 61-72.
24. Pharephan S, Sirivatanapa P, Makonkawkeyoon S, Tuntiwechapikul W, Makonkawkeyoon L. Prevalence of α-thalassaemia genotypes in pregnant women in northern Thailand. *Indian Journal of Medical Research* 2016; **143**(MARCH): 315-22.
25. Phollarp P, Tritipsombut J, Worasan C, et al. Thalassemia and iron deficiency among pregnant women attending antenatal care service at Khao Wong Hospital, Kalasin province. *Journal of Medical Technology and Physical Therapy* 2010; **22**(3): 262-70.
26. Pichanun D, Munkongdee T, Klamchuen S, et al. MOLECULAR SCREENING OF THE Hbs CONSTANT SPRING (codon 142, TAA > CAA, alpha 2) AND PAKSE (codon 142, TAA > TAT, alpha 2) MUTATIONS IN THAILAND. *Hemoglobin* 2010; **34**(6): 582-6.
27. Rahimah AN, Nisha S, Safiah B, et al. Distribution of alpha thalassaemia in 16 year old Malaysian Students in Penang, Melaka and Sabah. *The Medical journal of Malaysia* 2012; **67**(6): 565-70.
28. Sanchaisuriya K, Fucharoen S, Ratanasiri T, et al. Thalassemia and hemoglobinopathies rather than iron deficiency are major causes of pregnancy-related anemia in northeast Thailand. *Blood Cell Mol Dis* 2006; **37**(1): 8-11.
29. Savongsy O, Fucharoen S, Fucharoen G, Sanchaisuriya K, Sae-ung N. Thalassemia and hemoglobinopathies in pregnant Lao women: carrier screening, prevalence and molecular basis. *Ann Hematol* 2008; **87**(8): 647-54.

30. Sengchanh S, Sanguansermsri T, Horst D, Horst J, Flatz G. High frequency of alpha-thalassemia in the So ethnic group of south Laos. *Acta Haematol-Basel* 2005; **114**(3): 164-6.
31. Setianingsih I, Harahap A, Nainggolan IM. Alpha thalassaemia in Indonesia: phenotypes and molecular defects. *Advances in experimental medicine and biology* 2003; **531**: 47-56.
32. Sornkayasit K, al. e. Incidence of Hb Constant Spring and Hb Pakse in Khon Kaen: Using capillary electrophoresis and DNA analysis (Poster abstract). The 18th National Thalassemia Academic Symposium. Nonthaburi; 2012. p. 90.
33. Srivorakun H, Fucharoen G, Changtrakul Y, Komwilaisak P, Fucharoen S. Thalassemia and hemoglobinopathies in Southeast Asian newborns: diagnostic assessment using capillary electrophoresis system. *Clin Biochem* 2011; **44**(5-6): 406-11.
34. Sutjasung P, Fucharoen G, Fucharoen S, Chattumaruk P, Changtrakun D, Sanchaisuriya K. Effectiveness of thalassemia screening with the use of internal quality control blood samples at Kasetsomboon Hospital, Chaiyaphoom province. *JOURNAL OF MEDICAL TECHNOLOGY AND PHYSICAL THERAPY* 2011; **23**(1): 34-45.
35. Suwannakhon N, Seeratanachot T, Mahingsa K, Namwong P, T. S. Prevalence of Alpha-thalassemia Trait in the Volunteered Personals of University of Phayao. *J Hematol Transfus Med* 2014; **24**: 129-36.
36. Tan JA, Tay JS, Soemantri A, et al. Deletional types of alpha-thalassaemia in central Java. *Hum Hered* 1992; **42**(5): 289-92.
37. Tan JAMA, Lee PC, Wee YC, et al. High prevalence of alpha- and beta-thalassemia in the kadazandusuns in east Malaysia: Challenges in providing effective health care for an indigenous group. *Journal of Biomedicine and Biotechnology* 2010; **2010**.
38. Tangvarasittichai O, Jeenapongsa R, Sitthiworanan C, Sanguansermsri T. Laboratory investigations of Hb Constant Spring. *Clin Lab Haematol* 2005; **27**(1): 47-9.
39. Tangvarasittichai O, Poonanan N, Tangvarasittichai S. Using Red Cell Indices and Reticulocyte Parameters for Carrier Screening of Various Thalassemia Syndromes. *Indian journal of clinical biochemistry : IJCB* 2017; **32**(1): 61-7.
40. Tanphaichitr VS, Pung-amritt P, Puchaiwatananon O, et al. Studies on hemoglobin Bart's and deletion of alpha-globin genes from cord blood in Thailand (poster abstract). The International Conference on Thalassemia. Bangkok; 1985. p. P05.
41. Than AM, Harano T, Harano K, Myint AA, Ogino T, Okada S. High incidence of alpha-thalassemia, hemoglobin E, and glucose-6-phosphate dehydrogenase deficiency in populations of malaria-endemic southern Shan State, Myanmar. *Int J Hematol* 2005; **82**(2): 119-23.
42. Tongon R, Yunu R, Sanchaisuriya K, et al. Thalassemia and hemoglobinopathies in pregnant women attended antenatal care service at Yala Hospital. *J Med Tech Phy Ther* 2014; **26**(1): 32-9.
43. Tritipsombut J, Sanchaisuriya K, Fucharoen S, et al. Hemoglobin Profiles and Hematologic Features of Thalassemic Newborns Application to Screening of alpha-Thalassemia 1 and Hemoglobin E. *Arch Pathol Lab Med* 2008; **132**(11): 1739-45.

44. Tritipsombut J, Sanchaisuriya K, Phollarp P, et al. Micromapping of Thalassemia and Hemoglobinopathies in Diferent Regions of Northeast Thailand and Vientaine, Laos People's Democratic Republic. *Hemoglobin* 2012; **36**(1): 47-56.
45. Uaprasert N, Settapiboon R, Amornsiriwat S, et al. Diagnostic utility of isoelectric focusing and high performance liquid chromatography in neonatal cord blood screening for thalassemia and non-sickling hemoglobinopathies. *Clinica chimica acta; international journal of clinical chemistry* 2014; **427**: 23-6.
46. Yap ZM, Sun KM, Teo CRL, Tan ASC, Chong SS. Evidence of differential selection for the -alpha(3.7) and -alpha(4.2) single-alpha-globin gene deletions within the same population. *Eur J Haematol* 2013; **90**(3): 210-3.
47. Yin SKK, Chong QT, Mei LA, et al. A molecular epidemiologic study of thalassemia using newborns' cord blood in a multiracial Asian population in Singapore - Results and recommendations for a population screening program. *J Pediat Hematol Onc* 2004; **26**(12): 817-9.

3. Sickle-Cell Anaemia in India: An Updated Allele Frequency Map and Newborn Estimates

This project was conducted through a collaboration with researchers in Mumbai, India, and received funding from the Newton Bhabha Fund. The work has been written up in the format of a journal article and will be submitted to a peer-review journal.

3.1. Abstract

Background

Sickle-cell anaemia (SCA) is a neglected chronic disorder of increasing global health importance. Although the majority of affected births occur in Sub-Saharan Africa, the disease is highly prevalent amongst population subgroups in India, making it one of the worst affected countries worldwide. Over the last decade, extensive efforts to gather epidemiological and clinical data have generated a substantial body of new quantitative evidence on the prevalence of SCA in local communities across India.

Methods

First, we created an up-to-date geodatabase of SCA surveys in India following strict inclusion criteria. Then, we used generalised additive models, accounting for the distribution of population subgroups at the district level, and bootstrapping methods to generate an updated continuous map of sickle-cell allele frequency with confidence intervals. Finally, we calculated sub-national estimates of the number of newborns with SCA expected in 2020, based on projected birth count data.

Findings

The number of data points in our mapping evidence-base increased more than threefold compared to previous studies. However, the data indicated extensive spatial heterogeneity in β^S frequency over short geographical distances and between different population sub-groups, making model predictions from a smoothed surface difficult. Amongst states where the burden of SCA could be estimated with reasonable certainty, the number of newborns varied widely.

Interpretation

Building on a fast growing evidence-base, we created an updated map of sickle-cell allele frequency in India and used this map to generate sub-national estimates of the number of newborns affected, accounting for local population composition. Despite this tailored approach compared to previous mapping studies, we identify important uncertainties in model predictions which point towards a need for a multidisciplinary approach to estimating the distribution and burden of SCA in India. Our findings provide direction for future work into SCA in India, by raising important questions surrounding the best approach for estimating the burden of this heterogeneous disorder in the Indian population.

Funding

Newton-Bhabha Fund, European Research Council

3.2. Introduction

Sickle-cell anaemia (SCA), which results from the inheritance of two copies of the sickle β-globin gene variant (β^S), is the most common form of sickle-cell disease (SCD). SCD refers to a group of inherited disorders affecting the haemoglobin protein responsible for carrying oxygen in the blood.[1] Caused by a single nucleotide substitution at position 6 of the β-globin gene, its pathophysiology stems from the polymerisation of the resulting sickle haemoglobin variant (HbS), triggering a cascade of red blood cell alterations.[2,3] Individuals with SCA experience considerable morbidity from both acute and chronic sequelae. Without effective treatment, the most severe cases can be fatal within the first few years of life.[2,4] The heterozygous state (sickle-cell trait, or SCT) is largely asymptomatic and offers close to 90% protection against severe *Plasmodium falciparum* (*Pf*) malaria.[5,6] As a result, the β^S allele has reached high (≥10%) frequencies in many regions where malaria has historically been common.[7,8]

Due to improved survival and population movements, the global burden of SCA is increasing,[9] with the annual number of newborns with the disease expected to increase from around 300,000 to more than 400,000 between 2010 and 2050.[10] The majority of these births occur in Sub-Saharan Africa. However, some of the highest β^S allele frequencies in the world have been reported in Indian populations,[11-13] and India has been ranked the second worst affected country worldwide, with 42,016 (CI: 35,347 – 50,919) babies estimated born with SCA in 2010.[14] Uncertainty in the estimate of the number of newborn SCA cases in India can be partly explained by a lack of data in some parts of the country as well as by the remarkable heterogeneity in β^S allele frequencies that have been described within geographical areas and between different sub-populations.[14]

In India, β^S is predominantly found amongst scheduled population subgroups (tribes and castes), which are usually the most economically disadvantaged populations[15-17] . This high allele frequency is likely due to a combination of factors, including, but not limited to: (i) the high rate of endogamy that is observed in these groups,[16] (ii) a potentially greater selection pressure on these groups from malaria,[18] and (iii) the competitive evolutionary exclusion of β^S by β-thalassaemia and/or β^E in certain non-scheduled groups.[19-21] Remarkable heterogeneities in the allele frequency of β^S among Indian scheduled populations have been observed, with carrier frequencies ranging from ~1 to 40%.[17,22] Carrier frequencies of up to 12% have also been reported in higher socioeconomic groups,[23,24] though low carrier frequencies of around 1% are more commonly observed.[25-27]

This complex pattern of β^S in India, coupled with India's unique social structure and important pressure on stretched healthcare resources,[28] warrants the generation of an updated map that is specific to the Indian subcontinent. It has been suggested that SCA in India might not be as severe as in the rest of the world, since the majority of SCA cases are caused by the Arab-Indian β^S haplotype, which is associated with higher foetal haemoglobin (HbF) levels, the strongest known genetic modifier of disease severity.[29,30] However, there is growing evidence suggesting that the severity of SCA in India is not as mild as previously reported.[17,31] The need for improved knowledge of the epidemiology of SCD in India is therefore greater than ever.

Over the last decade, various public and private institutions in India have made a remarkable effort to quantify SCA prevalence in different parts of the country, ranging from village-level prevalence surveys to state-wide screening programmes (e.g. Patel *et*

al., 2013, and Patra *et al.*, 2015),[32,33] which offers the opportunity to create more precise maps and newborn estimates.

Improved knowledge of the fine-scale geographical distribution and burden of SCA is essential for informing public health policies. In particular, more accurate estimates, which distinguish between affected births in scheduled and non-scheduled groups, will enable a better assessment of the need to develop healthcare infrastructures; screening programmes, and treatments, including hydroxyurea and penicillin prophylaxis, for the prevention and management of SCA. Here we incorporate: (i) large amounts of recent survey data, (ii) an off-the-shelf powerful and reproducible spatial methodology, and (iii) demographic data at the state and district levels, to present the first national evidence-based map of β^S allele frequency in India, along with subnational estimates of the number of affected newborns expected in 2020. Coupled with ongoing extensive efforts to characterise disease survival and clinical severity in different parts of the country, this map represents an important public health tool for developing appropriate models of care nationally and locally.

3.3. Methods

Prevalence survey database assembly and inclusion criteria

We conducted a systematic review of recent sickle-cell surveys in India to generate a compilation of spatially located surveys of β^S allele frequency carried out in the country between 2010 and 2017 (schematic overview in Figure 3.1). Additional surveys older than 2010 that were available only in the local literature were also included, through our UK-India collaboration. This was supplemented by an earlier database published by Piel *et al.* (2013), which included studies from 1950 to 2009. Using procedures outlined in the

Supplementary information S3.1, identified references were reviewed for their suitability to contribute to the evidence-base underpinning our β^S allele frequency mapping analysis. Surveys for which there was stated or suspected selection bias in terms of health status or risk of SCD were excluded. Given the unique structure of the Indian population, and the high variation in β^S allele frequency in different groups, we included surveys conducted in specific ethnic, social or religious groups, and accounted for this in the statistical analysis (Supplementary information S3.2). In addition, we included only surveys that reported both the sample size and the number of SCT individuals identified. When provided, the number of SCA individuals and any compound heterozygotes (e.g. co-inheritance of β^S with β-thalassaemia or with another structural variant such as β^E) were also recorded. Finally, only surveys that could be georeferenced to at least the district level were included. The final dataset is described in detail in the Supplementary information S3.1, and is available on request.

Figure 3.1 Schematic overview of database generation procedures and geostatistical modelling processes. Pink diamonds represent input data; green boxes denote methodological steps; blue rods depict model outputs. * Historical map of malaria endemicity[34] and contemporary map of malaria.[35] **Night-time light data and friction surface (see Supplementary methods S3.2).

Accounting for malaria endemicity

The malaria hypothesis remains unresolved in the Indian context, presumably due to the influence of other population genetic factors, such as migration, consanguinity (as a result of endogamy practices), population structure and importantly, interactions with other globin mutations (specifically, α- and β-thalassaemia and β^E) on the allele frequency of β^S .[36,37] As part of this study, we analysed the geographical association between malaria (historical and contemporary) and β^S allele frequency in India. Three maps of malaria in India – two historical and one contemporary – were sourced (Figure S3.3)[34,35,38] and the malaria prevalence for each survey data point in our database was extracted in ArcMap

10.4.1 (ESRI Inc., Redlands, CA, USA). Univariate linear regression analysis was used to examine the relationship between β^S allele frequency and each of the three malaria estimates separately. Further details on the methods used are presented in the Supplementary information S3.2.

Accounting for population structure

The ethnic composition of the Indian population is highly complex, with the coexistence of more than two thousand isolated ethnic groups[39] and high levels of endogamy observed throughout the country.[40] Detailed data on the precise ethnic composition of local populations are not consistently reported in sickle-cell surveys, making it particularly challenging to account for each individual ethnic group reported in our database. In India, ethnicity is closely linked to social status, and under the Indian reservation system each ethnic group is classified into one of four official social designations.[41] These designations are, for example, used in data from the censuses conducted by the Indian Office of the Registrar General & Census Commissioner. These classes, in order of decreasing socioeconomic deprivation, are: (i) Scheduled Tribes (STs), (ii) Scheduled Castes (SCs), (iii) Other Backward Classes (OBCs), and (iv) General Classes (GCs). It has been observed that β^S in India is particularly prevalent in the STs and SCs. To account for social grouping in our analysis, surveys were included irrespective of the ethnicity of the study sample and categorised as either "scheduled" or "non-scheduled" (Supplementary information S3.2). Predicted continuous maps of β^S allele frequency were generated separately for scheduled and non-scheduled populations using the geostatistical model outlined below. These maps were then combined, by calculating the weighted average of predicted β^S allele frequency using census data on the proportion of scheduled and non-scheduled populations at the district level (n = 667)(

www.censusindia.gov.in), to create a composite map of β^S allele frequencies adjusted for social structure.

Creating an updated contemporary map of β^S allele frequency

We opted for a generalised additive modelling (GAM) approach to examine associations between the observed allelic counts from survey data with a series of predictor variables (or covariates). These modelled associations could then be spatially extrapolated to make predictions of β^S allele frequency at unsampled locations. A range of spatial, demographic and biological covariates were included in our statistical analyses, and are discussed in more detail in Supplementary information S3.2. Covariates of interest included: (i) geographical location, given by latitude and longitude in decimal degrees, (ii) historical rates of malaria, taken from two separate sources,[34,38] (iii) contemporary rates of malaria,[35] and (iv) two urbanisation proxies (Supplementary information S3.2). We also examined the association between social group and β^S allele frequency using the full dataset, before dividing the data into surveys carried out among scheduled groups (STs and SCs) and those among non-scheduled populations (OBCs and GCs).

A detailed description of the model procedures is provided in Supplementary information S3.2. For each of the two data subsets, i.e. scheduled and non-scheduled, we used a backward stepwise selection procedure to decide upon a final GAM. We began with a full model that included all covariates listed above. A two-dimensional smoother was used for the geographical effect to account for spatial autocorrelation. The Generalised Cross Validation (GCV) score, Akaike Information Criterion (AIC) score and mean squared error (MSE) were used as selection criteria, along with *p*-values for individual covariates. The final fitted model for each data subset (scheduled and non-scheduled) was

used to predict the β^S allele frequency at unsampled locations and generate a continuous map at 10km x 10km resolution (Supplementary information S3.2). Bootstrap resampling (with replacement) of the data subsets for 2,500 iterations was performed to generate a predictive probability distribution for each pixel, from which the 95% confidence interval could be calculated as a measure of uncertainty in our predictions. In order to capture information relating to the population distributions of scheduled and non-scheduled groups, we overlaid the maps, together with district-level data for the proportion of scheduled groups in the population, and generated a combined map of the weighted average of β^S allele frequencies for the two groups.

The predictive ability of each model was assessed by comparing the model's predictions with the observed β^S allele frequency for a 10% semi-random hold-out sample of the data subset (Supplementary information S3.3). The mean error (ME) and mean absolute error (MAE) were calculated as an indication of the model's overall bias and accuracy, respectively.

All statistical analyses were performed using the package 'mgcv' (with version R 3.3.2) within the R statistical and programming environment.

Estimating number of newborns affected

The number of newborns with SCA in India in 2020 was estimated by pairing our final continuous map of β^S allele frequency, adjusted for social grouping, with high-resolution birth count data. We used the high-resolution (1km x 1km) population distribution map from WorldPop (http://www.worldpop.org.uk) and multiplied the raster by the average crude birth rate for the two 5-year periods, 2015-2020 and 2020-2025, taken from the UN

world population prospects 2017 database,[42] to calculate the number of newborns in each pixel. After rescaling the birth count dataset to ensure an equivalent spatial resolution (10km x 10km), we calculated the number of newborns born with the SCA genotype (N_{SCA}) based on Hardy-Weinberg assumptions, so that N_{SCA} is given by Bq^2, where B is the number of births in each pixel in the grid and q is the allele frequency of β^S.[43] To calculate areal estimates, with uncertainty intervals, at the state and district level, newborn estimates in each pixel were generated for each of the 2,500 bootstrap repetitions of the model and summed across all pixels falling within an administrative unit. This generated a predictive probability distribution for the number of affected newborns in each administrative unit, which could be used to calculate the median and 95% confidence interval for the newborn estimates. Uncertainty measures for these estimates incorporated the uncertainty in both our β^S allele frequency predictions and the birth count data. Additional detail is provided in the Supplementary information S3.4.

3.4. Results

Prevalence survey database

Our geodatabase represents an amalgamation of: (i) an earlier database of 112 internationally published β^S prevalence surveys conducted between 1950 and 2009 (Piel *et al.*, 2013a) and, (ii) an updated compilation of surveys carried out in India since 2010 as well as those available only in the local literature (Figures 3.1 and S3.1). The full database consists of 340 surveys from 102 sources, spanning 189 spatially unique sites (Figure 3.2A). Surveys were conducted in 22 of the 36 Indian states and union territories. They had an uneven distribution, with more than half (52.35%) falling within just three states: Maharashtra (n = 54), Gujarat (n = 58) and Chhattisgarh (n = 66) (Figure S3.2A). Scheduled populations were the most extensively studied, with 169 surveys carried out

amongst STs, 18 amongst SCs and four amongst the two groups combined. Thirty-one surveys targeted populations belonging to the OBCs, 24 were carried out in the GCs and one in OBCs and GCs together. Thirty-six surveys included a mixture of ethnic and/or social designations, whilst the remaining 57 surveys did not specify the social background of the sample. The overall number of individuals sampled was 10,491,966 (compared to only 34,382 in Piel *et al.*, 2013) and sample size ranged from 2 to 2,297,476. Mean sample size was 30,859 and the median 359. Considerable variation in β^S prevalence was observed across short geographical distances (Figure 3.2A) as well as between social groups. The database is described in more detail in Supplementary information S3.1, including the temporal and spatial distribution and sample size of the surveys (Figure S3.1).

Univariate analyses of the malaria hypothesis

Assuming a linear trend in the malaria endemicity classes used in the two historical maps, univariate linear regression analyses revealed a statistically significant but weak relationship between β^S allele frequency and each of the historical malaria maps (p = 0.0002 and R^2 = 0.04 for Lysenko *et al.* and $p < 0.003$ and R^2 = 0.05 for the map by Hehir). Given the fine-scale heterogeneity of the contemporary malaria map, a univariate GAM analysis was applied, which yielded a similar result (p = 0.05 and R^2 = 0.04).

Univariate analyses of population structure

Of the 340 surveys included in the database, 247 (72.6%) clearly reported the social or ethnic group of their study sample. A univariate analysis of the relationship between β^S allele frequency and social group, classified as scheduled or non-scheduled, in these surveys revealed a highly statistically significant association ($p < 0.0001$). However, there

were 93 surveys for which social grouping was unclear; 36 surveys were conducted in multiple social groups, whilst the remaining 57 included no information relating to the social status of their sample. To determine whether these latter surveys should be categorised as scheduled or non-scheduled, pairwise comparisons of β^S allele frequencies in these surveys and those whose categorisation had already been made, were performed. Significant differences were only observed between the unclassified surveys and surveys carried out in scheduled groups ($p = 0.001$ and $p < 0.0001$, respectively). The surveys were therefore placed in the "non-scheduled" category.

β^S allele frequency map and prediction uncertainty

As described in the Methods, the database was divided into two subsets: (i) surveys carried out in scheduled populations, and (ii) surveys carried out in non-scheduled populations. Separate models were used independently to fit the observed β^S allele frequency in these two data subsets. For both, backward stepwise selection identified the most parsimonious geostatistical GAM to be the one that included geographic location only (Table 3.1); all other covariates were therefore excluded. The *p*-value for the two-dimensional smooth term was highly significant for both models ($p < 0.0001$). The large effective degrees of freedom (edf) indicate a highly non-linear relationship between spatial coordinates and β^S frequency for both scheduled and non-scheduled population groups. The R^2 value was greater for the non-scheduled dataset ($R^2 = 0.633$) than for the scheduled dataset ($R^2 = 0.52$), suggesting that a simple spatial model explains a larger proportion of the variance in the former. However, both the GCV and MSE were greater for the non-scheduled dataset, which suggest a greater prediction error for the model relating to the non-scheduled dataset.

Predicted maps for scheduled and non-scheduled populations were generated and then paired, together with district-level census data on the proportion of scheduled and non-scheduled groups in the general population, to create a final map of β^S allele frequency, adjusted for social structure. The adjusted β^S allele frequency map for India, depicted using equal intervals of the median values of the bootstrapped predictions,[44] is shown in Figure 3.2B. As the map represents only a summary (the median) of the predictive probability distribution, it must be interpreted alongside measures of uncertainty of the predicted values. The measure of uncertainty used here is the 95% confidence interval of the weighted average predictions in each pixel, and is displayed in Figure 3.2C.

Table 3.1 Summary results of the selected GAM for the scheduled and non-scheduled training datasets. The intercept, smoothing term *f* and its corresponding *p*-value, adjusted R^2, GCV score, MSE and AIC are all given.

Scheduled dataset (*n* = 175)		
	Estimate	**p-value**
Intercept	-3.0898	<0.0001
Smooth term	**edf**	**p-value**
***f*(lat,long)**	22.1800	<0.0001
$R^2 = 0.5170$		
GCV = 1.0038		
MSE = 0.7555		

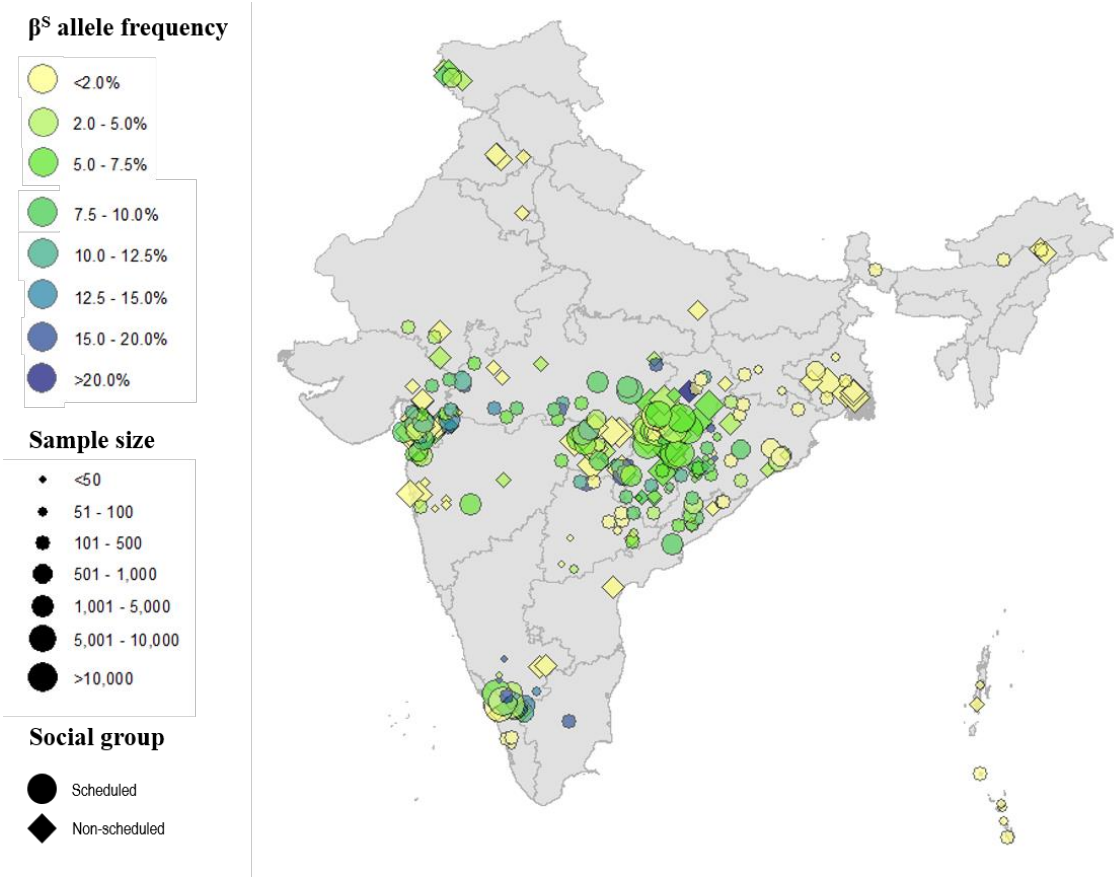
Non-scheduled dataset (*n* = 131)		
	Estimate	**p-value**
Intercept	-4.3033	<0.0001
Smooth term	**edf**	**p-value**
***f*(lat,long)**	25.8700	<0.0001
$R^2 = 0.6330$		
GCV = 1.8342		
MSE = 0.9464		

The highest allele frequencies (up to 10%), where uncertainty was also low (95% CI ≤5%) were found across a belt stretching across central India, extending from southern Gujarat in the west to southwestern West Bengal in the east. Heterogeneity is present within this belt, with lower allele frequencies predicted in the Nagpur Division of Maharashtra and parts of Chhattisgarh. Still higher allele frequencies of up to 21.6% were predicted in the northernmost part of the country, but this is also the area of highest uncertainty (95% CI ~25.0%) and therefore this result must be interpreted with caution. There is also an area of high β^S allele frequency in Southern India along the Karnataka-Tamil Nadu border, extending into Tamil Nadu. However, the level of uncertainty in this area reached 68.0%. Zero or close to zero allele frequencies were found across large swathes of the northwestern and northeastern parts of the country, as well as in the southern states of Karnataka, Telengana, Andhra Pradesh and Kerala.

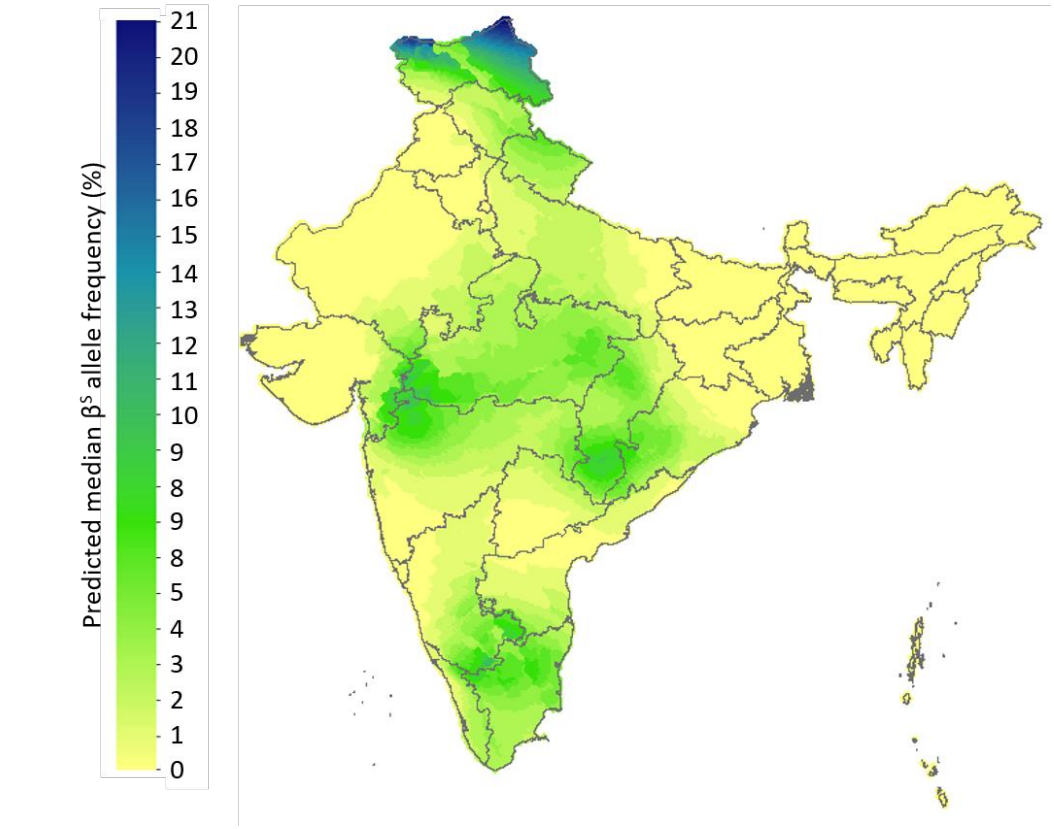
Uncertainty in our predictions is greatest where data are scarce (e.g. in northern and the southern tip of India) and where there is considerable heterogeneity in the observed

prevalence (e.g. in northwestern Maharashtra, eastern Madhya Pradesh and northern Chhattisgarh). Areas of greatest uncertainty include Assam in the northeast and the Andaman & Nicobar islands. Other area of high uncertainty are Jammu & Kashmir, Himachal Pradesh, Uttaranchal and the southern tip of Kerala and Tamil Nadu, where model predictions vary by as much as 68.0%. There is very low uncertainty (<3%) in the northwestern zone and parts of the northeastern zone, despite there being very few data points in these areas, since surveys that were carried out here observed consistent β^S allele frequencies.

A


βS allele frequency
<2.0%
2.0 - 5.0%
5.0 - 7.5%
7.5 - 10.0%
10.0 - 12.5%
12.5 - 15.0%
15.0 - 20.0%
>20.0%
Sample size
<50
51 - 100
101 - 500
501 - 1,000
1,001 - 5,000
5,001 - 10,000
>10,000
Social group
Scheduled
Non-scheduled

B


Predicted median βS allele frequency (%)
21
20
19
18
17
16
15
14
13
12
11
10
9
8
9
8
5
4
3
2
1
0

C

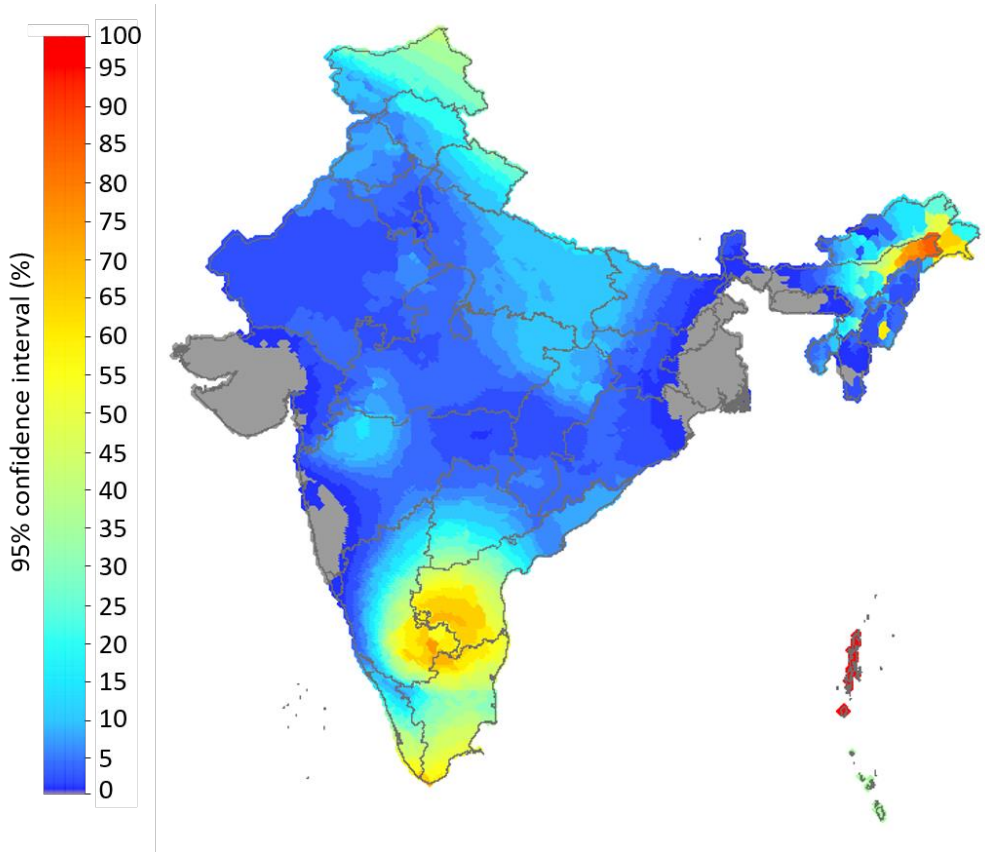


Figure 3.2. (A) A map of the sickle- cell surveys included in our database (n = 340). Data points are coloured according to the β^S allele frequency reported in the study sample. The size of the data points corresponds to their sample size. A spatial jitter of up to 0.3° latitude and longitude decimal degrees coordinates was applied to improve the visualisation of the data. (B) Map of median and (C) map of uncertainty in the predicted β^S allele frequency estimates in India at a resolution of 10km x 10km. State boundaries are displayed in grey.

Validation statistics

For the scheduled and non-scheduled map, we compared predictions from the map generated using the training data with known values in the hold-out subset. Our comparison revealed a mean error in allele frequency prediction of 1.4% and -0.6% in the scheduled and non-scheduled models, respectively. The predicted allele frequencies were therefore only slightly under/overestimated. The mean absolute error for each map was 6.4% and 1.2%, respectively. The predicted allele frequencies in the scheduled map are therefore overestimated (i.e. mean errors were mostly positive), whilst those in the non-scheduled map are slightly underestimated (i.e. mean errors were mostly small and

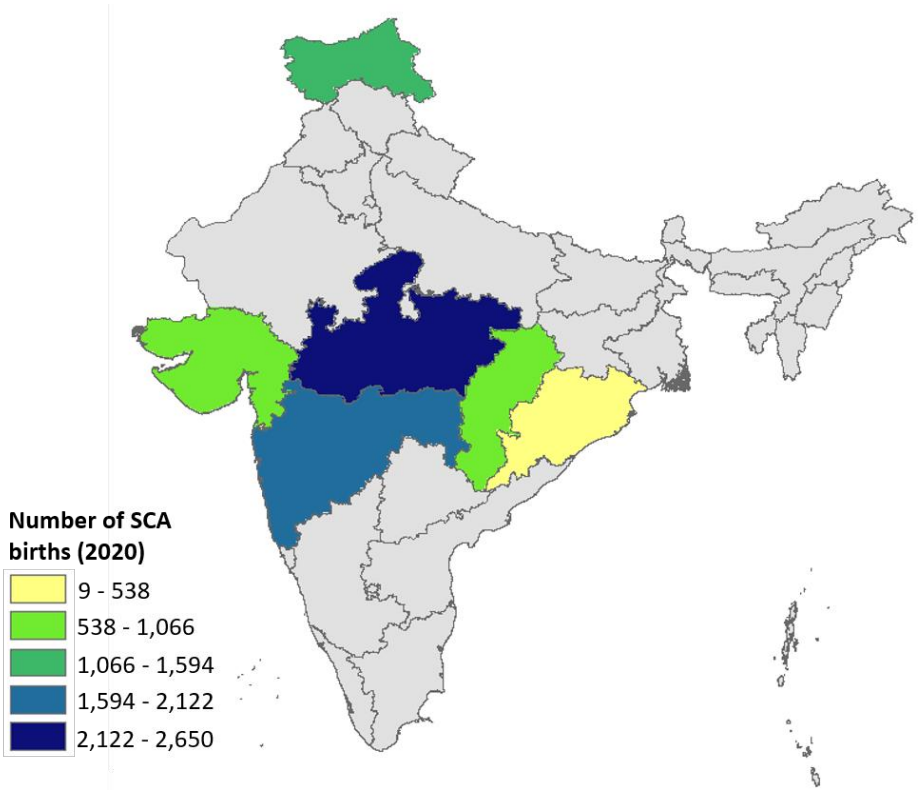
negative). The average magnitude of the error in allele frequency predictions was greater in the scheduled map than in the non-scheduled map. Differences between the observed and predicted values were large in certain areas where there was high short-range heterogeneity in the data. Limitations to the database that lead to relative weaknesses in the predictions are discussed below.

Estimates of newborns affected

The absolute burden of β^S in different parts of the country depends on both the frequency of the β^S allele as well as the size and composition of the population. After pairing our predicted β^S allele frequencies with demographic data, and taking into account the uncertainty in both, the model was able to make reliable predictions in a large part of central India, from Gujarat to Odisha (Figure 3.3), as well as in Jammu & Kashmir. The predicted number of newborns with SCA in these five states in 2020 was 7,938 (CI: 3,209 – 24,948). Together, they account for a quarter of the Indian population. We estimated that the highest number of affected newborns in 2020 will be in Madhya Pradesh (2,650 [CI: 1,190 – 7,401] (Figure 3.3A). In the remaining 75% of the Indian population, the reliability of the data is poorer but suggests a possible sickle-cell hotspot in Tamil Nadu, although the uncertainty associated with this estimate is substantial (5,519 [CI: 245 – 173,351]). Estimates of the total number of newborns with SCA in the less well-surveyed and/or heterogeneous regions range from zero in Andaman & Nicobar (CI: 0 – 2,226) to 5,519 in Tamil Nadu. Reliable estimates at the district level were similarly confined to the central states and Jammu & Kashmir (Figure 3.3B). Within Jammu & Kashmir, the model predicted reliable estimates for the westernmost districts, whilst in central India, estimates for which uncertainty was reasonable was possible for a large proportion of the districts in each state. The same principle of the spatial pattern of model uncertainty

applies to our newborn estimates; uncertainty is greatest where fewer data are available, sample sizes are smaller or observed allele frequency estimates are heterogeneous.

A



B

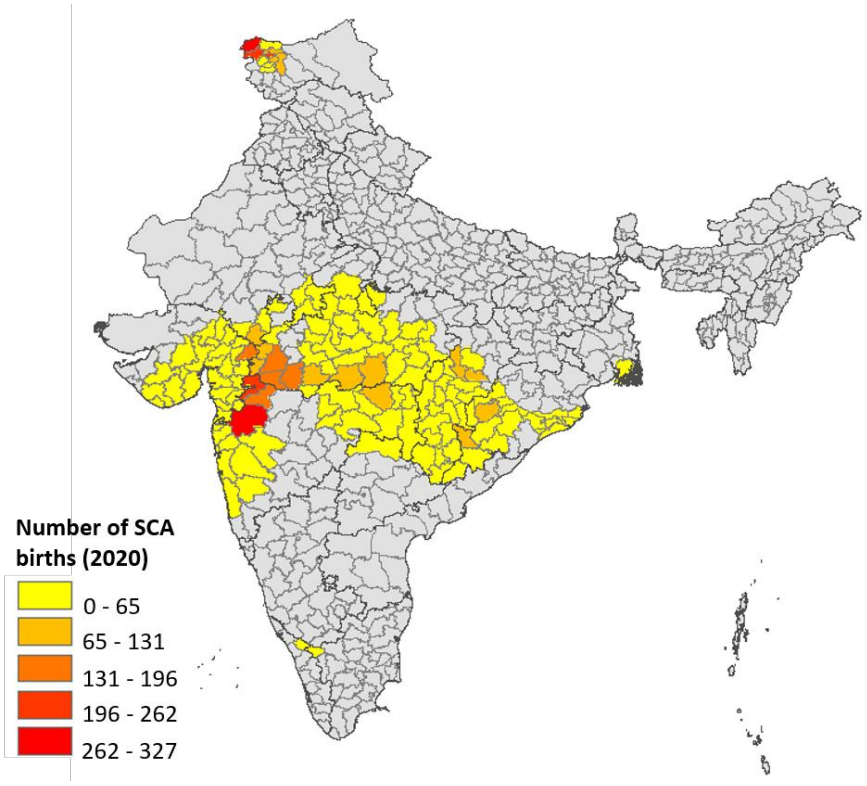


Figure 3.3. Map of the estimated number of newborns born with SCA in India, (A) by province and (B) by district (bottom panel), in 2020. The median of the predictive probability distribution of the areal estimates are displayed. Grey areas depict those states and districts for which uncertainty in our newborn estimates were deemed to be too high to be of reliable value.

3.5. Discussion

It is estimated that India alone accounts for more than 85% of SCA cases in Asia and 13% of cases worldwide.[14] Improvements in nutrition and infectious disease control in the country has caused a decline in childhood mortality over the last three decades, which in turn has led to the increased survival of children, including those with SCA.[45,46] There is also growing evidence that the severity of SCA in India is more heterogeneous than previously thought. As a result, there is an increasing demand on public health resources to identify, treat and manage SCA children in India and tools are needed to ensure effective allocation of healthcare in this vast country. Due to the sheer size of the Indian population, small differences in predicted allele frequencies can lead to substantial differences in national estimates of affected newborns. Here, we have developed an up-to-date, spatially continuous and model-based prevalence map of β^S in India that incorporates all additional usable survey data available from the published literature and in-country reports from the last 10 years as well as information on the distribution of social groups across the country. We have also calculated associated uncertainty metrics, together with sub-national estimates of the number of SCA newborns in 2020.

The reliability of such model-based maps is strongly dependent on both the quality and quantity of data available. Since the publication of the global β^S allele frequency map in 2013, there has been a surge of prevalence surveys, some of considerable size (>1 million individuals), carried out in India. The inclusion of data from these recent screening programmes and population surveys increased the number of data points in our SCA

evidence-base more than threefold. Furthermore, recent surveys provided additional data for areas of the country from which data had previously been sparse or entirely lacking (e.g. northern India and Chhattisgarh state).

Our methodology includes several methodological advances over previous studies. First, we used an off-the-shelf GAM approach, which is easy to implement and reproduce. Second, unlike the Bayesian geostatistical model previously used by Piel *et al.*, the GAM is entirely data-driven and does not require pre-defining of prior values. Third, we have incorporated information on social background into our analysis, which has enabled us to delve further into the intricacies of SCA in the Indian context and generate estimates of the burden of SCA in states where data are abundant and/or uncertainty in β^S allele frequency is low. Our map is broadly consistent with previous survey maps[17,47] or continuous map,[14] with lowest allele frequencies in the northwestern and northeastern parts of the country, highest rates predicted across the central belt, an area of high allele frequency in southern India, and generally heterogeneous distributions across the whole country.

There are some limitations to our map and population estimates. First, as for all model-based maps, their reliability on an evidence-base during model fitting means that predictions are subject to gaps and discrepancies in the underlying database. Despite the substantial expansion of our β^S geodatabase for India, there are still large swathes of the country for which there are no data. The categorisation of ethnic groups into two categories is somewhat reductionist; even within the scheduled groups, there is extensive heterogeneity in β^S allele frequencies and the same is true for non-scheduled groups.[15,47] The inclusion of more specific ethnicities in the analysis would have resulted in a more

tailored map and estimates; however, detailed data on the distribution of all the ethnic groups in India are lacking. Not all sources of uncertainty could be accounted for in the calculated uncertainty intervals for both the predicted allele frequency maps and newborn estimates. The Indian population is highly structured and composed a large number of distinct ethnic groups, many of which have practiced endogamy for millennia. It is therefore possible that communities living in close proximity to one another are genetically distant. Any differences in the allele frequency of β^S between such populations would not be predicted by our model unless data for them were available. Also, whilst recent prevalence data have greatly expanded the size of the evidence-base upon which the model is based, and provided data in previously unsampled areas, there is a definite concentration of surveying effort in areas of the country most affected by β^S. Finally, despite rigorous attempts to geoposition all surveys as precisely as possible, geographical information provided in data sources was sometimes vague and could lead to some uncertainties in the model.

This study has further highlighted the difficulty in generating accurate maps of β^S allele frequency and newborn estimates of its burden in a country as complex and diverse as India. Somewhat paradoxically, the addition of more data into the evidence-base has not increased the precision of allele frequency predictions. This is likely to be in part due to the high amount of spatial heterogeneity in β^S allele frequency that is observed. As a result, we are left with the difficult trade-off between having fewer data points, which makes smoothing by the model easier but masks the spatial variation, and many data points, which illustrates the extensive heterogeneity but makes local predictions by the model difficult. The added complexity of the ethnic composition and structure of the Indian population, the high levels of endogamy, and possibly consanguinity, in different

population sub-groups, variable rates of malaria selection and the co-existence of other globin mutations in certain parts of the country (e.g. β^E in northeast India, β^D in north India and β-thalassaemia in southern India) makes estimating β^S allele frequencies in the country particularly challenging.

Whilst the annual number of newborns born with SCA is expected to decline as a result of a decreasing birth rate,[10] the highly variable distribution of β^S in India, both geographically and across population groups, makes efficient health planning difficult. Reliable subnational estimates of the burden of this disorder, using geostatistical methods, provide a useful tool for the efficient allocation of public health resources for genetic counselling, education programs, and clinical service provision. However, as this study demonstrates, this is a difficult task given the complexity of the Indian context.

3.6. References

1. Serjeant G, Serjeant B. Sickle Cell Disease. Third ed: Oxford University Press; 2001.
2. Rees DC, Williams TN, Gladwin MT. Sickle-cell disease. *Lancet* 2010; **376**(9757): 2018-31.
3. Odièvre M-H, Verger E, Silva-Pinto AC, Elion J. Pathophysiological insights in sickle cell disease. *The Indian Journal of Medical Research* 2011; **134**(4): 532-7.
4. Grosse SD, Odame I, Atrash HK, Amendah DD, Piel FB, Williams TN. Sickle Cell Disease in Africa: A Neglected Cause of Early Childhood Mortality. *American Journal of Preventive Medicine* 2011; **41**(6): S398-S405.
5. Williams TN, Mwangi TW, Roberts DJ, et al. An Immune Basis for Malaria Protection by the Sickle Cell Trait. *PLOS Medicine* 2005; **2**(5): e128.
6. Luzzatto L. Sickle Cell Anaemia and Malaria. *Mediterranean Journal of Hematology and Infectious Diseases* 2012; **4**(1): e2012065.
7. Piel FB, Patil AP, Howes RE, et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications* 2010; **1**: 104-.
8. Allison AC. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *British Medical Journal* 1954; **1**(4857): 290-4.
9. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* 2010; **115**(22): 4331.
10. Piel FB, Hay SI, Gupta S, Weatherall DJ, Williams TN. Global Burden of Sickle Cell Anaemia in Children under Five, 2010–2050: Modelling Based on Demographics, Excess Mortality, and Interventions. *PLOS Medicine* 2013; **10**(7): e1001484.
11. Research ICoM. Intervention Programme for Nutritional Anaemia and Haemoglobinopathies against some Primitive Tribal Populations of India: A National Multicentric Study of ICMR. India: Indian Council of Medical Research, 2010.
12. Godbole S, Gupta RB, Yadav R, et al. High prevalence of sickle haemoglobin in Mehra caste of district Betul, Madhya Pradesh. In: Singh N, Qamra SR, Bhat J, et al., editors. Proceedings of National Symposum on Tribal Health. Jabalpur; 2006.
13. Patra PK, al. e. Sickle Cell Screening Project - At A Glance. 2016. www.scic.cg.nic.in.
14. Piel FB, Patil AP, Howes RE, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *Lancet* 2013; **381**(9861): 142-51.
15. Shrikhande AV, Arjunan A, Agarwal A, et al. Prevalence of the beta(S) gene among scheduled castes, scheduled tribes and other backward class groups in Central India. *Hemoglobin* 2014; **38**(4): 230-5.
16. Dixit S, Sahu P, Kar SK, Negi S. Identification of the hot-spot areas for sickle cell disease using cord blood screening at a district hospital: an Indian perspective. *Journal of Community Genetics* 2015; **6**(4): 383-7.
17. Colah R, Mukherjee M, Ghosh K. Sickle cell disease in India. *Current opinion in hematology* 2014; **21**(3): 215-23.
18. Sharma RK, Thakor HG, Saha KB, Sonal GS, Dhariwal AC, Singh N. Malaria situation in India with special reference to tribal areas. *The Indian Journal of Medical Research* 2015; **141**(5): 537-45.

19. Ghosh K, Colah RB, Mukherjee MB. Haemoglobinopathies in tribal populations of India. *The Indian Journal of Medical Research* 2015; **141**(5): 505-8.
20. Agarwal MB. The burden of haemoglobinopathies in India--time to wake up? *The Journal of the Association of Physicians of India* 2005; **53**: 1017-8.
21. Sinha S, Black ML, Agarwal S, et al. Profiling β-thalassaemia mutations in India at state and regional levels: implications for genetic education, screening and counselling programmes. *The HUGO Journal* 2009; **3**(1-4): 51-62.
22. Italia Y, Krishnamurti L, Mehta V, et al. Feasibility of a Newborn Screening and Follow-up Programme for Sickle Cell Disease among South Gujarat (India) Tribal Populations. *Journal of medical screening* 2015; **22**(1): 1-7.
23. Fareed M, Anwar MA, Ahmad MK, Afzal M. Gene frequency reports of sickle cell trait among six human populations of Jammu and Kashmir, India. *Gene Rep* 2016; **4**: 1-5.
24. Feroze M. Sickle cell disease project in Kerala National Conference on Hemoglobinopathies. Bangalore; 2013.
25. Urade BP. Haemoglobin S and βThal: Their distribution in Maharashtra, India. *Int J Biomed Sci* 2013; **9**(2): 75-81.
26. Dolai TK, Dutta S, Bhattacharyya M, Ghosh MK. Prevalence of hemoglobinopathies in rural Bengal, India. *Hemoglobin* 2012; **36**(1): 57-63.
27. Bhukhanvala DS, Sorathiya SM, Shah AP, Patel AG, Gupte SC. Prevalence and hematological profile of beta-thalassemia and sickle cell anemia in four communities of Surat city. *Indian journal of human genetics* 2012; **18**(2): 167-71.
28. Dandona L, Dandona R, Kumar GA, et al. Nations within a nation: variations in epidemiological transition across the states of India, 1990–2016 in the Global Burden of Disease Study. *The Lancet*; **390**(10111): 2437-60.
29. Menzel S, Thein SL. Genetic architecture of hemoglobin F control. *Current opinion in hematology* 2009; **16**(3): 179-86.
30. Tewari S, Brousse V, Piel FB, Menzel S, Rees DC. Environmental determinants of severity in sickle cell disease. *Haematologica* 2015; **100**(9): 1108-16.
31. Jain D, Warthe V, Dayama P, et al. Sickle Cell Disease in Central India: A Potentially Severe Syndrome. *The Indian Journal of Pediatrics* 2016; **83**(10): 1071-6.
32. Patel J, Patel B, Gamit N, Serjeant GR. Screening for the sickle cell gene in Gujarat, India: A village-based model. *Journal of Community Genetics* 2013; **4**(1): 43-7.
33. Patra PK, Khodiar PK, Hambleton IR, Serjeant GR. The Chhattisgarh state screening programme for the sickle cell gene: a cost-effective approach to a public health problem. *Journal of Community Genetics* 2015; **6**(4): 361-8.
34. Lysenko AJ, Semashko IN. Itogi Nauki: Medicinskaja Geografia: Academy of Sciences, USSR; 1968: 25-146.
35. Gething PW, Patil AP, Smith DL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria Journal* 2011; **10**: 378-.
36. Hamamy H. Consanguineous marriages: Preconception consultation in primary health care settings. *Journal of Community Genetics* 2012; **3**(3): 185-92.
37. Penman BS, Habib S, Kanchan K, Gupta S, Read A. NEGATIVE EPISTASIS BETWEEN α(+) THALASSAEMIA AND SICKLE CELL TRAIT CAN EXPLAIN INTERPOPULATION VARIATION IN SOUTH ASIA. *Evolution; International Journal of Organic Evolution* 2011; **65**(12): 3625-32.
38. Hehir P. Malaria in India: Oxford University Press; 1927.
39. Singh KS. People of India: Introduction: Oxford University Press; 2002.

40. Bittles AH. Endogamy, consanguinity and community genetics. *Journal of genetics* 2002; **81**(3): 91-8.
41. Jangir SK. Reservation Policy and Indian Constitution in India. *American International Journal of Research in Humanities, Arts and Social Sciences* 2013; **3**(1): 126-8.
42. Affairs UNDoEaS. World population prospects, the 2017 Revision. New York: United National Population Division, 2017.
43. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* 2005; **76**(5): 887-93.
44. Patil AP, Gething PW, Piel FB, Hay SI. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends in Parasitology*; **27**(6): 246-53.
45. Koli R, Goli S, Doshi R. Epidemiological Transition in Urban Population of Maharashtra. *Advances in Epidemiology* 2014; **2014**: 11.
46. Yadav S, Arokiasamy P. Understanding epidemiological transition in India. *Global Health Action* 2014; **7**: 10.3402/gha.v7.23248.
47. Colah RB, Mukherjee MB, Martin S, Ghosh K. Sickle cell disease in tribal populations in India. *The Indian Journal of Medical Research* 2015; **141**(5): 509-15.

3.7. Supplementary information 1: Assembling a geodatabase of sickle-cell prevalence surveys

Library assembly

Online searches of two major bibliographic databases, PubMed (http://www.pubmed.gov) and Scopus (http://www.scopus.com) were performed to identify published surveys of β^S allele frequency carried out in India since 2010. To ensure consistency with the search strategy used in Piel *et al.* (2013),[1] the same keyword string was used: "sickle cell" or "haemoglobin S" or "hemoglobin S" or "Hb S". We also included the search term "India" to make it country-specific and limited our search to between 2010 and 2016. Our final search on 14 October 2016 yielded 501 references across all three databases, which were imported to a bibliographic management software (Endnote X7.7.1, Thomson Reuters, Carlsbad, CA, USA). Twenty-four duplicates were identified and manually removed, leaving a total of 477 references to be reviewed. Unpublished sources of data and references available only in the local literature (n = 84) were identified through our UK-India collaboration and reviewed in the same manner. Finally, the dataset used by Piel *et al.* (2013)[1] in their global sickle-cell mapping study was clipped to India and merged. All additional sources from which data were abstracted are listed in the Supplementary Bibliography. Sources included in the earlier database are listed in Piel *et al.* (2013).

Data inclusion criteria and abstraction process

The identified sources were reviewed in detail by a single individual (CH) using a specific set of inclusion criteria, which were outlined in a pre-defined protocol. This ensured maximum objectivity and consistency in the data abstraction process. A random sample

of 10% of the references were reviewed and checked by a second individual (FP), and the survey details extracted by the two authors compared.

Survey representativeness

Only surveys that were conducted among representative population samples were included in the database. Two key considerations when determining the representativeness of a sample were: (i) the source population from which the sample was taken, e.g. pregnant women, newborns, community, etc., and (ii) the sampling methodology used, e.g. random, consecutive, unselected or universal. All family-based studies as well as those conducted amongst patients (e.g. malaria patients or general outpatients) or suspected cases of haemoglobinopathies were deemed to be inherently biased and thus immediately excluded. However, surveys that took place amongst newborns, pregnant women, children and adolescents or in the community, although unbiased in terms of their source population, were scrutinised for their sampling methodology. Those with stated or suspected bias in their selection method were excluded. For instance, studies that took place in the community but which required voluntary participation without any prior education or awareness campaign were excluded due to a potential bias, for example towards individuals with an affected family member.

Ethnicity and social grouping

Previous efforts to map β^S have excluded surveys that target specific ethnic groups due to their lack of representativeness of the general population. However, given the unique population structure of India in the form of highly endogamous population groups[2] and the well-known heterogeneities in β^S frequency across ethnic and social groups,[3, 4] we

included surveys irrespective of the ethnicity of the study sample, and accounted for this in the mapping analysis (Supplementary Information S2). In India, ethnicity is inherently linked to social caste.[3] As such, we categorised ethnic groups according to their social category as defined by the Government of India. The four designations used were: (i) Scheduled Tribe (ST), (ii) Scheduled Caste (SC), (iii) Other Backward Class (OBC), and (iv) General Class (GC). State-wise lists of communities belonging to the STs, SCs, and OBCs were sourced from the websites of the Ministry of Tribal Affairs (http://tribal.nic.in), the Ministry of Social Justice and Empowerment (http://socialjustice.nic.in) and National Commission for Backward Classes (http://www.ncbc.nic.in), respectively. Any ethnic groups that were not included in these lists were considered to belong to the GC category. Subsequently, the data were split into two subsets: (i) surveys carried out in scheduled populations (i.e. STs and/or SCs), and (ii) surveys carried out in non-scheduled population (i.e. OBCs and/or GCs). Those surveys that did not specify the ethnic origin or social grouping of the study sample were ultimately included in the non-scheduled subset, following univariate pairwise comparisons of β^S allele frequencies in these groups with unambiguous scheduled and non-scheduled groups (see Supplementary Information S2).

Survey diagnostic methods

There are several methods available for the diagnosis of SCT and/or SCD.[5-7] These include the sickling test, solubility test, haemoglobin electrophoresis and high performance liquid chromatography (HPLC). In India, the most commonly used methods are sickling test and HPLC, depending on the resources available.[6] Given the good reliability of most diagnostic methods for β^S,[5] no strict inclusion criteria with regards to diagnostics were applied. However, key information regarding the precise diagnostic

algorithm used in a survey was recorded. This included the presence or absence of an initial screening stage and the type of confirmatory diagnostic test used. Where multiple confirmatory diagnostic tests were used, the most reliable method – usually HPLC or, in some cases, DNA sequencing – was recorded.

Sample size and β^S allele frequency

An important prerequisite to the inclusion of a survey in the database was the clear reporting of both the sample size and the number of AS individuals identified. Where available, information on other β^S genotypes, including SS homozygotes and compound heterozygotes (e.g. SD and S/β-thalassaemia), was also recorded. Occasionally, surveys directly reported the frequency of the β^S allele without a breakdown of the different genotypes. These were also included, but only if sample size was provided.

Genotype counts were used to calculate the allele frequency of β^S, a key input for our model, using the formula:

$$q = \frac{(2 * n_{SS}) + n_{AS}}{2N}$$

where q is the β^S allele frequency, n_{SS} and n_{AS} are the number of β^S homozygotes and heterozygotes or compound heterozygotes, respectively, and N is the size of the study sample. A simplifying assumption of this calculation is that the populations are at Hardy-Weinberg equilibrium (HWE).[8, 9]

Sample size varied enormously between surveys. We did not place any constraints on sample size, but rather accounted for it in the statistical model by weighting the contribution of each survey to the fitted model according to its size. As a result, the

contributions of very small samples were down-weighted whilst those of larger samples were strengthened.

Georeferencing

We applied an inclusion criterion of spatial specificity, whereby only those surveys that could be georeferenced to at least the second administrative level were included. In India, this corresponds to the district level (https://censusindia.gov.in). Various geopositioning gazetteers were used to identify the latitude and longitude of all included surveys as precisely as possible. Where a survey described a specific sampling site, the geographic coordinates of that site were used. For surveys only reporting the district in which they were conducted, the centroid of the district was identified and recorded. In instances where the presented survey data came from multiple sites, the centroid between all the survey sites was determined using the Geographic Information Systems (GIS) software, ArcGIS Desktop (ArcMap 10.4.1, ESRI Inc., Redlands, CA, USA). No information on the spatial extent of the surveys was recorded, as this information was rarely available.

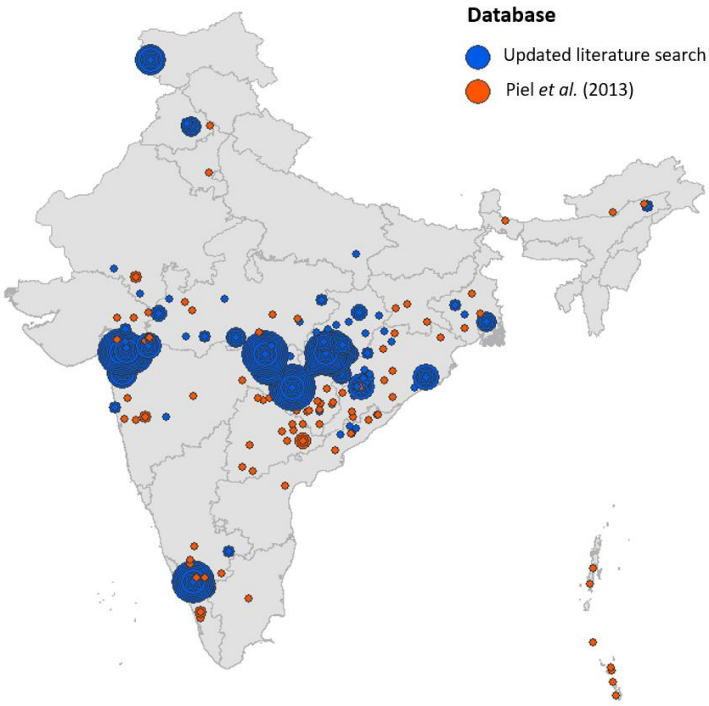
An advantage of our modelling approach, which is described in detail in Supplementary Information S2, is that all spatial duplicates could be retained, thereby maximising the size of the evidence-base for the mapping model.

The final sickle-cell survey dataset

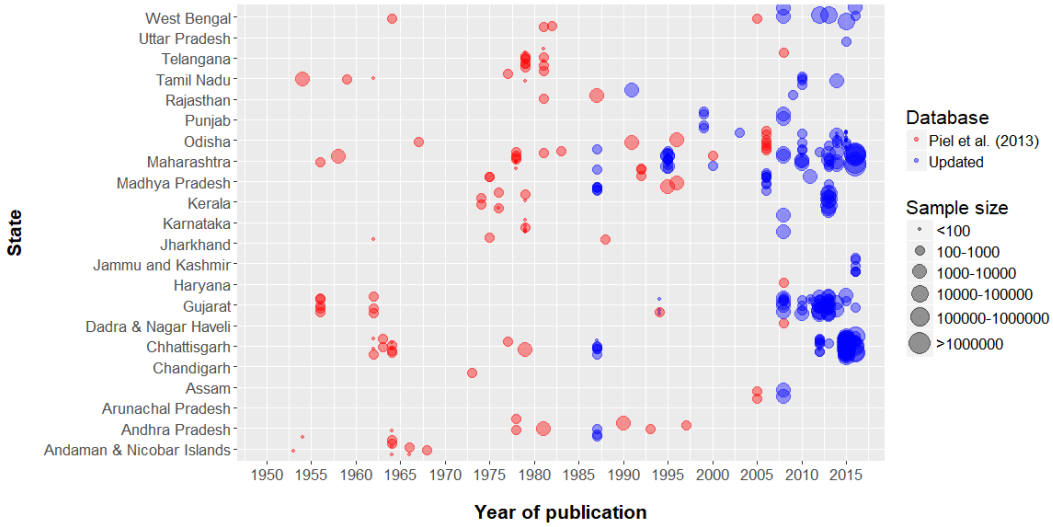
Of the 561 sources reviewed as part of the updated literature search, 228 met the inclusion criteria outlined above. As a result, the inclusion of surveys carried out in India since 2010, as well as surveys available only in the local literature, more than tripled the size of the previous evidence-base upon which a continuous map of β^S frequency was

constructed (n = 112 in Piel *et al.* compared to n = 340 in this study). The additional data increased the number of data points in previously surveyed areas (Supplementary figure S3.1A), rather than expanding the spatial distribution of the data points included in the database. The temporal distribution of the surveys for each state is shown in Supplementary figure S3.1B, along with information on the database in which they were initially recorded and their sample size. Reassuringly, very few surveys were missed by Piel *et al.*'s previous database, illustrating the comprehensiveness of the systematic search and source review process used by them and in this study. The additional surveys conducted prior to 2010 that were found came from 16 sources that could only be accessed from local journals and centre reports.

A

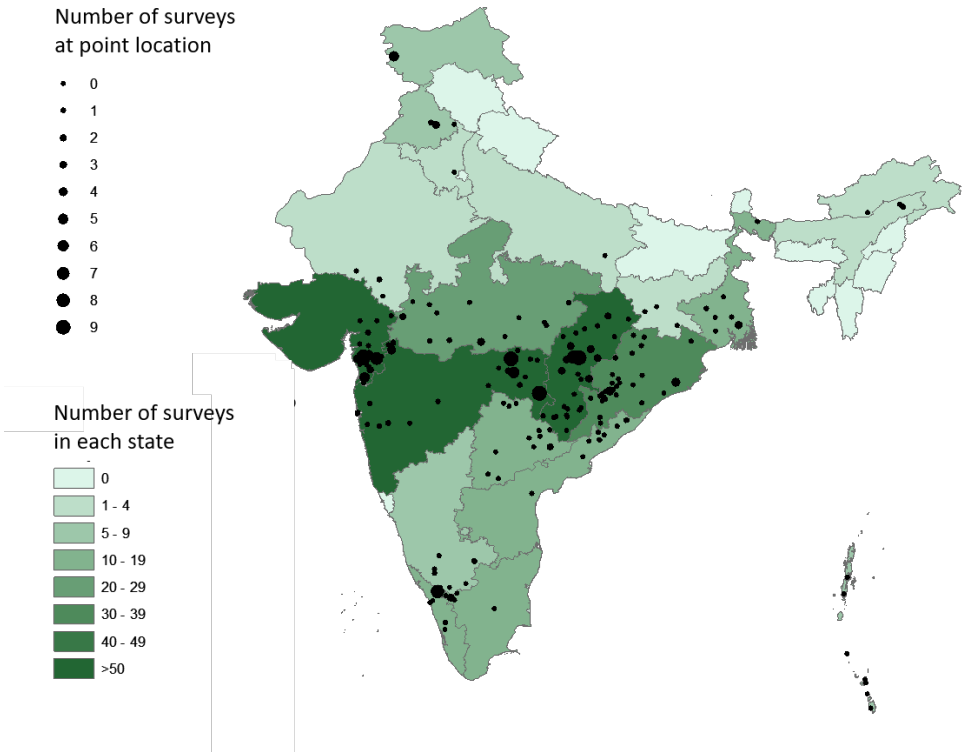


B



Supplementary figure S3.1 Spatial and temporal distribution of β^S surveys in the database. (A) Map showing the distribution of the data points, coloured according to the original database in which they were included. Red circles represent surveys that were originally identified in the geodatabase by Piel *et al.* (2013); blue circles depict surveys obtained through our updated literature search. Concentric circles are used to indicate multiple surveys from the same geographical location, with each circle representing a single survey. Here, the positioning of the data points on the maps reflects their true geographical coordinates. (B) Temporal distribution and sample size of the β^S surveys by state. Again, the colour of the data points indicates the database in which they were originally included, while size depicts the sample size of the survey.

The spatial distribution of surveys and the survey count by state are shown in Supplementary figure S3.2. Three hundred and forty surveys were reported from 102 sources Surveying effort was unevenly distributed; over half of the surveys fell within just three states (Gujarat, Maharashtra and Chhattisgarh) and some of the largest surveys were carried out in two of these states (Maharashtra and Chhattisgarh). Nevertheless, this reflects the areas of highest concern with regards to SCA.[4, 6] The total number of individuals sampled was 10,491,966, with 97.3% of these sampled in these three states.



Supplementary figure S3.2 Survey count by state. The colour intensity of the state indicates the number of surveys that were carried out in it, whilst circle size indicates the number of surveys that were carried out at each point location within the state.

Surveys were carried out in a range of specific population groups, including newborns (*n* = 29), schoolchildren and adolescents (*n* = 82) and pregnant women and/or their husbands (*n* = 17). However, the majority of the surveys were carried out in community samples (*n* = 170). Forty-two surveys lacked did not state their target population group (pregnant

women, newborns, etc.), but did not indicate any potential bias and so were retained. Many of the surveys were conducted within a specific ethnic group (n = 159), whilst 305 studies selected individuals based on their broader socioeconomic status (e.g. scheduled tribe, scheduled caste, other backward class or general class). Thirty-five studies lacked information on the ethnic, social or religious status of their study sample. Scheduled tribes were the most studied group, following by scheduled castes and other backward classes.

Some heterogeneity in the diagnostic algorithm used was also observed. However, the vast majority of the surveys used HPLC as the confirmatory diagnostic test, with less than half of these incorporating an additional screening stage.

3.8. Supplementary information 2: Creating a β^S allele frequency map

Overview of generalised additive models (GAMs)

To generate a continuous map of β^S allele frequency in India, we employed a generalised additive modelling approach. Generalised additive models (GAMs) comprise a collection of nonparametric regression techniques performing regression inference on complex, non-linear, relationships between a given response variable and multiple predictor variables.[10] The advantage of using GAMs over more common linear models is that they can create extremely flexible functional relationships while effectively regularising the objective function - that is, they automatically trade off over/under fitting to provide the best estimation of the underlying predictive patterns.[11] In addition, unlike more black box methods, such as gradient boosting machines,[12] the models are easily interpretable due to the contribution of each independent variable to the prediction being explicitly encoded.

For a given response $Y \in R$ and D associated predictors $X \in R^D$, the GAM estimates an unknown function $Y = f(x)$ through an additive series of smooth functions, i.e.

$$g\big(E(Y)\big) = a + s_1(x_1) + s_2(x_2) + \cdots + s_n(x_n) + \cdots + x_p$$

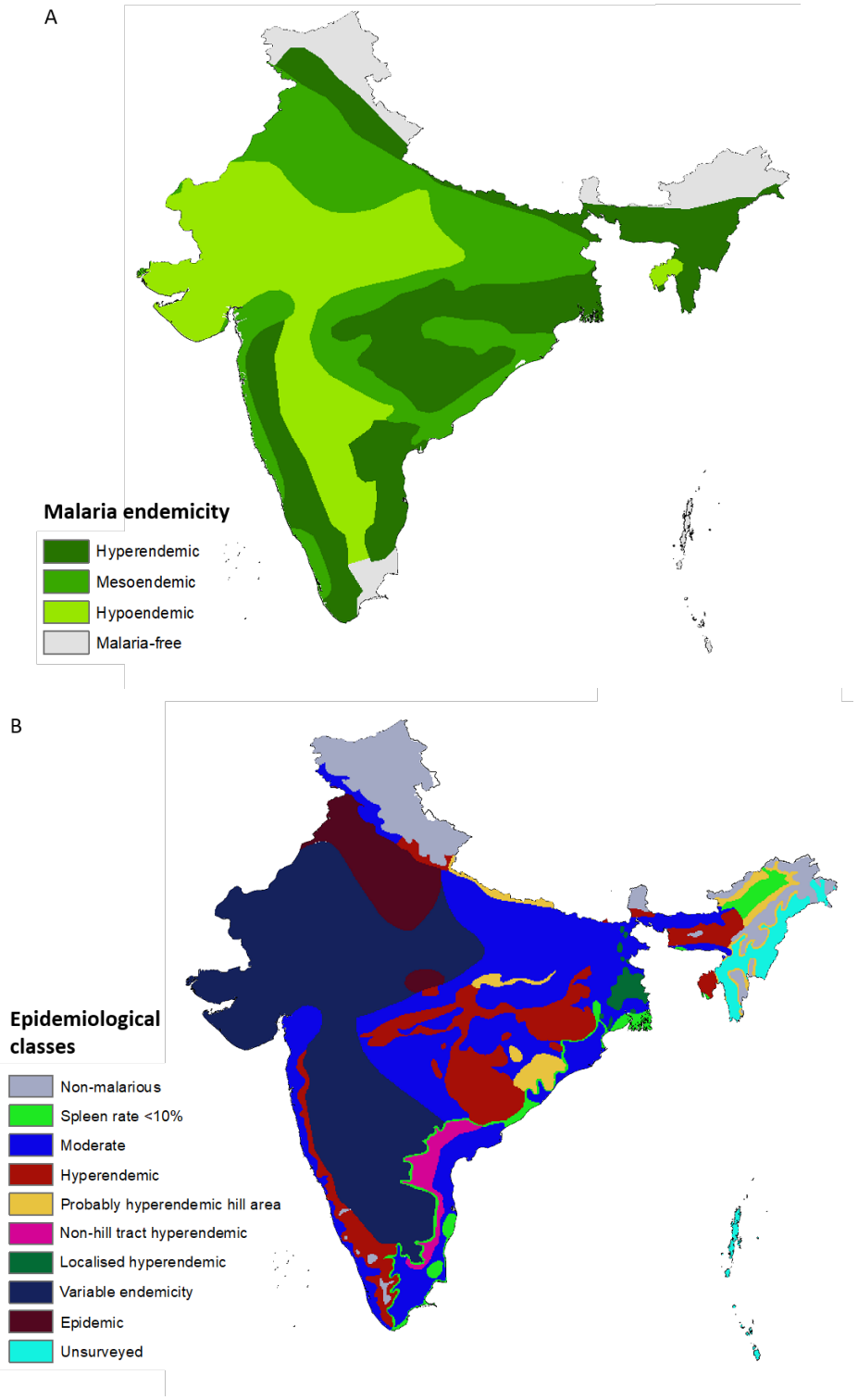
where E(Y) denotes the expected value of the response and g(Y) denotes the link function that allows for the modelling of non-Gaussian likelihoods.[10] The terms $s_1(x_1),\ldots,s_n(x_n)$ denote smooth, nonparametric thin-plate spline functions. GAMs can also be generalised to multiple dimensions to model the interactions between two predictor variables such as spatial coordinates. The general principles of generalised additive modelling have been described previously by Hastie and Tibshirani (1987)[13] and West (2012).[14]

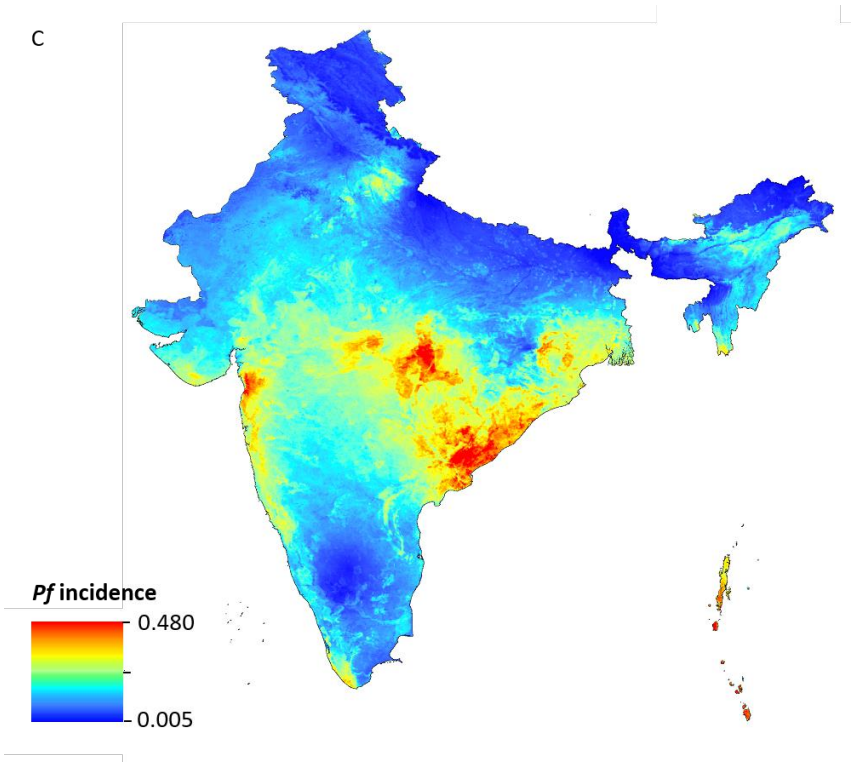
Model covariates and data sources

Whilst the malaria hypothesis is widely accepted and supported by high-quality and varied evidence,[15] the geographical relationship between sickle-cell and historical malaria remains unresolved in the Indian context. This is presumably due to a myriad of other factors influencing the distribution of β^S here, including population structure, consanguinity and migration. To explore this relationship in more detail using our GAM approach, we sourced two maps of the pre-control distribution of malaria in India (Supplementary figure S3.3. The first was a global map produced by a team of researchers in the 1960s and represents the distribution of malaria when it was at its highest, circa 1900 (Supplementary figure S3.3A).[16] The authors synthesised information and data on multiple malariometric indices from a range of sources and combined this with expert opinion as well as temperature and rainfall data to generate the map. To suit the purposes of our study, the map, which was previously digitised by Piel *et al.* (2010), was clipped to India in ArcGIS Desktop. The second map was produced by Sir Patrick Hehir in 1927 and is specific to India (Supplementary figure S3.3B).[17] Whilst there is little information on the evidence-base upon which the map was generated, to our knowledge it is the only available pre-control map of malaria specifically for India. Coupled with the additional detail provided by the map compared to that from Lysenko et al. (1968), this justified its inclusion in our analysis.

To gain insight into the relationship between sickle-cell in India and present-day malaria, we also obtained a contemporary map of malaria from the Malaria Atlas Project map repository (MAP, www.map.ox.ac.uk) (Supplementary figure S3.3C). Generated using a Bayesian geostatistical model that incorporates data from parasite rate surveys as well as from environmental covariates, the map shows the mean value of the range of age-

standardised *Plasmodium falciparum* parasite rates predicted for each location, at a resolution of 1km x 1km.[18]





Supplementary figure S3.3 (A and B) Historical and (C) contemporary maps of malaria that were used in the present analysis. (A) A map of malaria endemicity that was generated by Lysenko *et al.* (1968).[16] (B) A map showing various epidemiological classes of malaria, as defined by Hehir (1927).[17] (C) A model-based geostatistical map of *Pf* incidence, as estimated by Gething *et al.* (2011).[18] The digitised version of the map by Lysenko *et al.* is reprinted from Piel *et al.* (2010)[15] with permission from Elsevier ©2010. The map by Hehir (1927) is licensed under the Creative Commons Attribution 4.0 International license and was obtained from the Wellcome Collection.

It has been previously mentioned that β^s is particularly prevalent in STs and SCs, many of whom reside in rural areas (www.censusindia.gov.in). To see if this pattern could be captured in our model, we sourced data on two metrics for urban accessibility: (i) nighttime lights (NTL), and (ii) travel time to the nearest city of at least 50,000 people by land- or water-based travel. Data were obtained from the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) (https://explorer.earthengine.google.com/) and the European Commission's Joint Research Centre Science Hub (http://forobs.jrc.ec.europa.eu/products/gam/), respectively. Both were accessed on 28 June 2017. The NTL series has a resolution of 1km x 1km and is based on the satellite detection of visible and near-infrared (VNIR)

emission sources at night. The accessibility map has a resolution of 5km x 5km and incorporates information on transport networks (rail, road and sea) as well as the environmental and political factors that affect travel times between locations.

Other covariates explored in our analysis include geographical location, given by latitude and longitude in decimal degrees, and social group, categorised into two main groups: scheduled and non-scheduled. Our treatment of social group is discussed in detail below.

Collinearity

Collinearity refers to the non-independence of predictor variables such that one predictor could be reasonably well modelled as a function of another. This can cause problems for parameter estimation, leading to the incorrect identification of relevant predictors in the model. To test for an association between the two historical malaria maps, both of which presented ordered categorical malaria epidemiological data, we used univariate linear regression, which revealed the maps to be significantly associated ($p < 0.0001$). Univariate linear regression models were also used to examine the collinearity between contemporary parasite rates and historical endemicity classes. We found a significant association between contemporary malaria and the map by Hehir ($p < 0.0001$) and a non-significant relationship between the maps by Lysenko *et al.* and MAP ($p = 0.24$). The latter two maps were therefore retained for the model selection process.

Model selection

The GAM selection process requires consideration of two key model features. First, the covariates to include in the final model must be selected, and a decision about which covariates to include as smoothing terms made. Second, given a model structure, the

degree of smoothness for each smoothing term must be optimally balanced such that under- or over-smoothing is avoided.

Prior to the model selection process, the data were split into two subsets: (i) those surveys that were carried out in scheduled populations, and (ii) those that were carried out in non-scheduled populations or populations of unknown ethnicity. Model selection was then carried for each of the two datasets separately. A backward stepwise approach was used to select the final model for each dataset. The starting model included smooth terms for latitude and longitude (both separately and in combination), historical malaria (Lysenko *et al.*, 1968), contemporary malaria, NTL and accessibility. Cubic regression splines were used for all smooth terms except the two-dimensional geographical term, for which we used thin-plate regression splines. Each survey's contribution to the model was weighted according to its survey size.

The variables associated with the highest *p*-values were successively removed until the generalized cross-validation criterion (GCV, a computationally efficient generalisation of standard out-of-sample predictive performance), Akaike Information Criterion (AIC) and mean square error (MSE) of the model were minimised.[10] The deviance explained and R^2 were also examined. Finally, smoothing parameters, which determine the degree of smoothness of the predictive functions, were estimated from the data via restricted maximum likelihood (REML).

Interpolation to generate two continuous β^S allele frequency surfaces

The final model for each data subset was run repeatedly over 2,000 bootstrap samples of the dataset, generating new model parameter estimates for each bootstrapped sample.

Each iteration of the model was then used to make predictions regarding β^S allele frequency at unsampled locations at a resolution of 10km x 10km. This resulted in a probability distribution of predicted β^S allele frequency for each 10km x 10km pixel. The final β^S allele frequency maps are displayed using the central tendency (median) of these predictions.

The final β^S allele frequency maps for scheduled and non-scheduled populations represent the β^S frequency distribution when the proportion of scheduled groups in the population is 100% and 0%, respectively. This is of course an unrealistic portrayal of the situation in India; rather, a better estimate would fall somewhere between the two estimates, with the weight of each estimate being determined by the proportion of scheduled groups in the population. We obtained data on the district-wise distribution of scheduled groups from the India Census website (www.censusindia.gov.in) and used this to calculate weighted averages of the two estimates in each 10km x 10km pixel, using the formula:

$$E(Y)^{adj} = (E(Y)^S * p^S) + (E(Y)^{NS} * (1 - p^S))$$

where Y denotes β^S allele frequency, $E(Y)^{adj}$ denotes the adjusted predicted value of Y, $E(Y)^S$ denotes the predicted value of Y among scheduled populations, p^s denotes the proportion of scheduled groups in the population, and $E(Y)^{NS}$ denotes the predicted value of Y among non-scheduled populations. The final interpolated map of β^S in India thus represents the median of 2,000 predictions, adjusted for the proportion of scheduled groups in the population.

Uncertainty

All predictions have an associated level of uncertainty. For spatial predictions, this uncertainty varies geographically and is a function of the quality, quantity and sample

size of the available data. In addition, heterogeneity in the observed frequencies of β^S can affect the level of uncertainty. To quantify the uncertainty in our predictions, we measured the interquartile range of the probability distribution of predicted β^S allele frequency for each 10km x 10km pixel.

All analyses were performed using R 3.3.2. Full scripts of the code are available on request.

3.9. Supplementary information 3: Model validation

To assess the model's predictive ability, a validation run was performed to quantify the disparity between the model's predictions and a hold-out subset of the data. The hold-out dataset was obtained by taking a semi-random sample of the data in which there was a minimum distance of 100m between points, thereby improving the coverage of the held-out data points. The models were run with the remaining 90% of the data and their predictions compared with the observed allele frequencies from the hold-out dataset. The prediction's mean error (ME) and mean absolute error (MAE) were used to measure the model's overall bias and accuracy, respectively. The ME is the average distance between the actual data points and the predicted values. The MAE is a measure of the average magnitude of the errors in the predicted values.[1]

3.10. Supplementary information 4: Generating newborn estimates

Whilst an allele frequency map indicates areas of high and low prevalence, it does not relay information regarding the health burden of the disorder in terms of the absolute number of newborns affected. Rather, this is dependent on a combination of allele frequency, population density and birth rate. We combined our adjusted allele frequency map with high-resolution birth count data to generate estimates for the number of newborns born with the AS and SS genotypes in India in 2020. Data on birth counts were obtained by pairing a 100m x 100m population distribution map, downloaded from the WorldPop Project website (http://www.worldpop.org), with United Nations estimates of crude birth rate, averaged between the two 5-year periods, 2015-2020 and 2020-2025. The output is the estimated number of live births per 100m x 100m pixel. To match the resolution of our β^S allele frequency map, we adjusted the birth count data to a resolution of 10km x 10km.

Genotype proportions were calculated from the predicted β^S allele frequency using Hardy-Weinberg proportions. The proportion of SCA newborns were calculated as p^2, where p is the frequency of wild-type β^S alleles. The genotype frequencies were then multiplied by birth count data for each pixel to generate estimates for the number of SCA babies born in 2020. Uncertainty in our newborn estimates stems both from uncertainty in our β^S frequency predictions and uncertainty in birth rate. To indicate the level of uncertainty, we calculated three estimates, corresponding to a lower bound, a central tendency and a higher bound.

3.11. Supplementary bibliography S1

1. Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Dewi M, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. Lancet. 2013; 381:142-51.
2. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian Population History. Nature. 2009; 461:489-94.
3. Shrikhande AV, Arjunan A, Agarwal A, Dani A, Tijare J, Gettig E, et al. Prevalence of the beta(S) gene among scheduled castes, scheduled tribes and other backward class groups in Central India. Hemoglobin. 2014; 38:230-5.
4. Colah R, Mukherjee M, Ghosh K. Sickle cell disease in India. Current opinion in hematology. 2014; 21:215-23.
5. Ashley-Koch A, Yang Q, Olney RS. Sickle Hemoglobin (Hb S) Allele and Sickle Cell Disease: A HuGE Review. American Journal of Epidemiology. 2000; 151:839-45.
6. Colah RB, Mukherjee MB, Martin S, Ghosh K. Sickle cell disease in tribal populations in India. The Indian Journal of Medical Research. 2015; 141:509-15.
7. Naik RP, Haywood C. Sickle cell trait diagnosis: clinical and social implications. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program. 2015; 2015:160-7.
8. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. Science (New York, NY). 1908; 28:49-50.
9. Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahresh Wuertt Verh Vaterl Naturkd1908. p. 369-82.
10. Wood S. Generalized Additive Models: An Introduction with R: Chapman and Hall/CRC 2006.
11. Geman S, Bienenstock E, Doursat R. Neural Networks and the Bias/Variance Dilemma. Neural Computation. 1992; 4:1-58.
12. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist. 2001; 29:1189-232.
13. Hastie T, Tibshirani R. Generalized Additive Models: Some Applications. Journal of the American Statistical Association. 1987; 82:371-86.
14. West RM. Generalised Additive Models. In: Tu Y-K, Greenwood DC, editors. Modern Methods for Epidemiology. Dordrecht: Springer Netherlands; 2012. p. 261-78.
15. Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nature communications. 2010; 1:104-.
16. Lysenko AJ, Semashko IN. Itogi Nauki: Medicinskaja Geografia: Academy of Sciences, USSR; 1968. p. 25-146.
17. Hehir P, cartographer Malaria in India: Wellcome Collection; 1927.
18. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, Johnston GL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. Malaria Journal. 2011; 10:378-.

3.12. Supplementary Bibliography S2: Sickle-cell prevalence surveys

The sources include in the final database are:

1. Balgir RS. Clinical and hematological profiles of hemoglobinopathies in two tribal communities in Sundargarh district in Orissa, India Int J Hum Gen. 2003; 3.
2. Balgir RS. A cross-sectional study of hemoglobin disorders in pregnant women attending two Urban hospitals in eastern coast of Odisha, India. Online J Health Allied Sci. 2013; 12.
3. Bhukhanvala DS, Sorathiya S, Surve R, Nair S, Italia K, Colah R, et al. Hemoglobin variants in Muslim community in South Gujarat, Western India. International journal of laboratory hematology. 2014; 36:e15-e7.
4. Bhukhanvala DS, Sorathiya SM, Sawant P, Colah R, Ghosh K, Gupte SC. Antenatal screening for identification of couples for prenatal diagnosis of severe hemoglobinopathies in Surat, South Gujarat. J Obstet Gynecol India. 2013; 63:123-7.
5. Bhukhanvala DS, Sorathiya SM, Shah AP, Patel AG, Gupte SC. Prevalence and hematological profile of beta-thalassemia and sickle cell anemia in four communities of Surat city. Indian journal of human genetics. 2012; 18:167-71.
6. Centre SRKR. Prevalence of β-thalassaemia trait (BTT) and sickle cell trait (SCT) in the Surat population: Surat Raktadan Kendra & Research Centre2008.
7. Centre SRKR. β-thalassaemia and sickle cell trait in Muslim, Patel and Ghanchi communities of Surat: Surat Raktadan Kendra & Research Centre2008.
8. Chakrabarti S, Mandal K, Pathak S, Patra A, Pal S. Haemoglobinopathies among the tribal and non-tribal antenatal mothers in a tertiary care hospital of rural West Bengal, India. Bangladesh J Med Sci. 2016; 15:90-4.
9. Chirmulay D, Kate SL, Mokashi GD, Bankar MP. Prevalence of sickle cell haemoglobin in Vansda district Valsad Gujarat: A pilot study Indian Journal of Hematology & Blood Transfusion 1994; 12.
10. Choubisa SL. Abnormal Haemoglobins, Thalassaemia and G-6-Pd Enzyme Deficiency in Rajasthan (Western-India) Haematologica. 1991; 24:153-65.
11. Choubisa SL. Sickle cell haemoglobin, thalassaemia and G-6-PD enzyme deficiency genes in Garasiya tribe inhabited malaria endemic areas of Sirohi District, Rajasthan (India). J Commun Dis. 2009; 41:13-8.
12. Choudhuri S, Sen A, Ghosh MK, Misra S, Bhattacharyya M. Effectiveness of Prenatal Screening for Hemoglobinopathies in a Developing Country. Hemoglobin. 2015; 39:380-3.
13. Deore AU, Urade BP. Incidence of sickle cell trait among the Mahar of Raipur, Chhattisgarh. Anthropologist. 2013; 15:377-8.
14. Dey A. ABO, Rh blood groups and sickle cell trait among the Kisans of Sambalpur, Orissa. National Seminar on Mother Child Health Status & XII Annual Conference of Indian Society of Human Genetics; Calcutta1987.
15. Dixit S, Sahu P, Kar SK, Negi S. Identification of the hot-spot areas for sickle cell disease using cord blood screening at a district hospital: an Indian perspective. Journal of community genetics. 2015; 6:383-7.
16. Dolai TK, Dutta S, Bhattacharyya M, Ghosh MK. Prevalence of hemoglobinopathies in rural Bengal, India. Hemoglobin. 2012; 36:57-63.
17. Fareed M, Anwar MA, Ahmad MK, Afzal M. Gene frequency reports of sickle cell trait among six human populations of Jammu and Kashmir, India. Gene Rep. 2016; 4:1-5.

18. Feroze M. Sickle cell disease project in Kerala National Conference on Hemoglobinopathies; Bangalore2013.
19. Fulmali PM, Kate SL, Bankar MP, Mokashi GD, Phadke MA. Genetico-environmental studies amongst tribals of Dhadgoan Tahasil, Dist. Dhule: A survey National Seminar on Mother Child Health Status & XII Annual Conference of Indian Society of Human Genetics; Calcutta1987.
20. Godbole S, Gupta RB, Yadav R, Singh MPSS, Das U, Gupta A, et al. High prevalence of sickle haemoglobin in Mehra caste of district Betul, Madhya Pradesh. In: Singh N, Qamra SR, Bhat J, Sharma RK, Rao KVK, Roy J, et al., editors. Proceedings of National Symposum on Tribal Health; Jabalpur2006.
21. Gunjal Sandeep S, Narlawar Uday W, Humne Arun Y, Chaudhari Vijaya VL. Prevalence of sickle cell disorder and anaemia in tribal school students from central India. Int J Collab Res Intern Med Public Health. 2012; 4:1321-9.
22. Italia Y, Krishnamurti L, Mehta V, Raicha B, Italia K, Mehta P, et al. Feasibility of a Newborn Screening and Follow-up Programme for Sickle Cell Disease among South Gujarat (India) Tribal Populations. Journal of medical screening. 2015; 22:1-7.
23. Jain MK, Padmanabham PBSV, Dhar P, Das K, Rao VR. Distribution of sickle cell disorder among the Nav-Buddhist of Nagpur City, Maharashtra International seminar on Haemoglobinopathies South Asia Problems and Prevention Calcutta2000.
24. Kaur M, Dangi CBS, Singh H. To study the haemoglobinopathies and ratio of copper and zinc in Sindhi Community of Bhopal. Int J Pharma Bio Sci. 2013; 4:672-91.
25. Kaur R, Kaur H. Haemoglobin Variants In Some Ethnic Groups Of Punjab (North India). Indian journal of human genetics. 1999; 5:57-60.
26. Khurana P, Aggarwal A, Huidrom SS, Kshatriya GK. Haptoglobin polymorphism among the tribal groups of southern Gujarat. Indian journal of human genetics. 2011; 17:169-74.
27. Mission NRH. Sickle Cell Disease Control Program. Maharashtra2016.
28. Mohanty D. Current status of sickle cell disease in India: how can you attenuate? Molecular cytogenetics. 2014; 7:I45.
29. Mohanty D, Mohanty N, Kate SL. Sickle Cell Anaemia, Thalassemia and Other Genetic Problems Amongst Tribals of Maharashtra Indian Council of Medical Research1995.
30. Mukhopadhyay D, Saha K, Sengupta M, Mitra S, Datta C, Mitra PK. Spectrum of Hemoglobinopathies in West Bengal, India: A CE-HPLC Study on 10407 Subjects. Indian journal of hematology & blood transfusion : an official journal of Indian Society of Hematology and Blood Transfusion. 2015; 31:98-103.
31. Nagar R, Sinha S, Raman R. Haemoglobinopathies in eastern Indian states: a demographic evaluation. Journal of community genetics. 2015; 6.
32. Nimgaonkar V, Krishnamurti L, Prabhakar H, Menon N. Comprehensive integrated care for patients with sickle cell disease in a remote aboriginal tribal population in Southern India. Pediatr Blood Cancer. 2014; 61:702-5.
33. Panigrahi S, Patra PK, Khodiar PK. Neonatal screening of sickle cell anemia: A preliminary report. Indian journal of pediatrics. 2012; 79:747-50.
34. Panigrahi S, Patra PK, Khodiar PK. The screening and morbidity pattern of sickle cell anemia in chhattisgarh. Indian journal of hematology & blood transfusion : an official journal of Indian Society of Hematology and Blood Transfusion. 2015; 31:104-9.

35. Patel AG, Shah AP, Sorathiya SM, Gupte SC. Hemoglobinopathies in South Gujarat population and incidence of anemia in them. Indian journal of human genetics. 2012; 18:294-8.
36. Patel J, Patel B, Gamit N, Serjeant GR. Screening for the sickle cell gene in Gujarat, India: A village-based model. Journal of community genetics. 2013; 4:43-7.
37. Patra PK, al. e. Sickle Cell Screening Project - At A Glance. Chhattisgarh2016; Available from: www.scic.cg.nic.in.
38. Patra PK, Khodiar PK, Hambleton IR, Serjeant GR. The Chhattisgarh state screening programme for the sickle cell gene: a cost-effective approach to a public health problem. Journal of community genetics. 2015; 6:361-8.
39. Purohit P, Dehury S, Patel S, Patel DK. Prevalence of deletional alpha thalassemia and sickle gene in a tribal dominated malaria endemic area of eastern India. ISRN hematology. 2014; 2014:745245.
40. Qamra S, Roy J, Srivastava P. Impact of sickle cell trait on physical growth in tribal children of Mandla district in Madhya Pradesh, India. Annals of human biology. 2011; 38:685-90.
41. Research ICoM. Jai Vigyan S & T Mission Project on Community Control of Thalassaemia Syndromes – Awareness, Screening, Genetic Counselling and Prevention: A National Multicentric Task Force Study of ICMR India: Indian Council of Medical Research2008.
42. Research ICoM. Intervention Programme for Nutritional Anaemia and Haemoglobinopathies against some Primitive Tribal Populations of India: A National Multicentric Study of ICMR. India: Indian Council of Medical Research2010.
43. Rupani MP, Vasava BC, Mallick KH, Gharat VV, Bansal R. Reaching community through school going children for sickle cell disease in Zankhvav Village of Surat District, Western India. Online J Health Allied Sci. 2012; 11.
44. Sachi Devi S, Naidu JM. Sickle Cell Trait among Bagatha, Konda Dora and Parangi Poroja Tribes of Vishakhapatnam, Andhra Pradesh In: Roy SB, Chowdhury A, Sarkar B, editors. Proceedings of XII Annual Conference of Indian Society of Human Genetics1987.
45. Sahoo SS, Biswal S, Dixit M. Distinctive mutation spectrum of the HBB gene in an urban eastern Indian population. Hemoglobin. 2014; 38:33-8.
46. Sathe M, Gorakshakar A, Rao VR, Mukherjee M, Vasantha K, Bhatia HM. Red Cell Genetic Abnormalities in the Tribes of five districts of Madhya Pradesh In: Roy SB, Chowdhury A, Sarkar B, editors. Proceedings of XII Annual Conference of Indian Society of Human Genetics1987.
47. Shah SP, Shah P, Desai S, Modi D, Desai G, Arora H. Effectiveness and Feasibility of Weekly Iron and Folic Acid Supplementation to Adolescent Girls and Boys through Peer Educators at Community Level in the Tribal Area of Gujarat. Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine. 2016; 41:158-61.
48. Shrikhande AV, Arjunan A, Agarwal A, Dani A, Tijare J, Gettig E, et al. Prevalence of the βs gene among scheduled castes, scheduled tribes and other backward class groups in central India. Hemoglobin. 2014; 38:230-5.
49. Singh MPSS, Gupta RB, Yadav R, Gadge VS, Das U, Godbole S, et al. Prevalence of common haemoglobinopathies among scheduled caste and scheduled tribes of Shahdol and Khandwa (East Nimar) districts of Madhya Pradesh In: Singh N,

Qamra SR, Bhat J, Sharma RK, Rao KVK, Roy J, et al., editors. Proceedings of National Symposum on Tribal Health; Jabalpur2006.
50. Urade BP. Haemoglobin S and βThal: Their distribution in Maharashtra, India. Int J Biomed Sci. 2013; 9:75-81.

4. Understanding the Contrasting Spatial Haplotype Patterns of Malaria-Protective β-Globin Polymorphisms

Chapter 4 relates to the population genetics of β^S and β-thalassaemia mutations and the haplotypes associated with them. This work has been published in the journal, *Infection, Genetics and Evolution*, and is included here in its final form.

Infection, Genetics and Evolution 36 (2015) 174–183

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Understanding the contrasting spatial haplotype patterns of malaria-protective β-globin polymorphisms

Carinna Hockham, Frédéric B. Piel, Sunetra Gupta, Bridget S. Penman *

Department of Zoology, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form 4 September 2015
Accepted 6 September 2015
Available online 21 September 2015

Keywords:
Sickle-cell
β-Thalassaemia
Malaria
Haplotype distribution
Gene conversion

ABSTRACT

The malaria-protective β-globin polymorphisms, sickle-cell (β^S) and β^0-thalassaemia, are canonical examples of human adaptation to infectious disease. Occurring on distinct genetic backgrounds, they vary markedly in their patterns of linked genetic variation at the population level, suggesting different evolutionary histories. β^S is associated with five classical restriction fragment length polymorphism haplotypes that exhibit remarkable specificity in their geographical distributions; by contrast, β^0-thalassaemia mutations are found on haplotypes whose distributions overlap considerably. Here, we explore why these two polymorphisms display contrasting spatial haplotypic distributions, despite having malaria as a common selective pressure. We present a meta-population genetic model, incorporating individual-based processes, which tracks the evolution of β-globin polymorphisms on different haplotypic backgrounds. Our simulations reveal that, depending on the rate of mutation, a large population size and/or high population growth rate are required for both the β^S- and the β^0-thalassaemia-like patterns. However, whilst the β^S-like pattern is more likely when population subdivision is high, migration low and long-distance migration absent, the opposite is true for β^0-thalassaemia. Including gene conversion has little effect on the overall probability of each pattern; however, when inter-haplotype fitness variation exists, gene conversion is more likely to have contributed to the diversity of haplotypes actually present in the population. Our findings highlight how the contrasting spatial haplotype patterns exhibited by β^S and β^0-thalassaemia may provide important indications as to the evolution of these adaptive alleles and the demographic history of the populations in which they have evolved.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The mutations responsible for sickle-cell disease and β^0-thalassaemia represent two unequivocal examples of balanced polymorphisms in the human genome (Hedrick, 2011; Roberts and Williams, 2003; Taylor et al., 2012). Occurring at high frequencies in many populations indigenous to malaria-endemic regions, these variants are subject to balancing selection due to their protective effect against *Plasmodium falciparum* malaria in the heterozygous state (Allison, 1964; Hill et al., 1991; Piel et al., 2010; Rockett et al., 2014; Williams et al., 2005a). Homozygotes suffer severe blood disorders (sickle-cell anaemia and β^0-thalassaemia major, respectively), which, without access to diagnosis and treatment, are often lethal in the first few years of life (Weatherall et al., 2006).

In population genetics theory, it is generally accepted that natural selection results in one of two population genetic outcomes: (i) a *hard selective sweep*, in which a single adaptive allele sweeps rapidly through a population, resulting in the predominance of a single haplotype associated with the adaptive allele in the population, and (ii) a *soft selective sweep*, whereby ancestral genetic variation around the adaptive site is partially preserved owing to multiple alleles at the site being selected (Messer and Petrov, 2013; Ralph and Coop, 2010). In the context of this study, we define a haplotype as a set of DNA variations, including the variant under selection, that are located on a single chromosome and, by virtue of their close proximity, are inherited together. Both the sickle-cell mutation (β^S) and β^0-thalassaemia appear at first glance to be examples of soft selective sweeps. The former, which always results from the replacement of glutamic acid by valine at position 6 of the β-globin gene (*HBB* c. 20 A → T; p. Glu6-Val), is associated with five "classical" restriction fragment length polymorphism (RFLP) haplotypes (Table 1) (Flint et al., 1998). The latter results from any mutation that completely eliminates the production of protein from the β-globin gene (Weatherall and Clegg, 2001a). One-hundred and fifty-eight such mutations are currently reported (http://www.globin.bx.psu.edu/cgi-bin/hbvar/query_vars3, accessed 29 June 2015; Patrinos et al., 2004), and many of these can be found on more than one genetic background (Table 2) (Trabuchet et al., 1991; Weatherall and Clegg, 2001a).

* Corresponding author.
E-mail addresses: carinna.hockham@zoo.ox.ac.uk (C. Hockham), fred.piel@zoo.ox.ac.uk (F.B. Piel), sunetra.gupta@zoo.ox.ac.uk (S. Gupta), bridget.penman@zoo.ox.ac.uk (B.S. Penman).

http://dx.doi.org/10.1016/j.meegid.2015.09.018
1567-1348/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Distribution of β^S haplotypes. Relative percentages (%) of classical β^S-associated haplotypes in different geographical regions. Classical β^S haplotypes are defined as those which cannot be accounted for by reciprocal recombination. Countries where β^S has largely been imported, for example parts of Western Europe and North America, are not included as these do not reflect the early evolutionary history of the variant.

		β^S-associated haplotypes						
Region	Country	Arab-Indian	Benin	Cameroon	Central African Republic	Senegal	Other	Reference
Sub-Saharan Africa	Angola	–	12.0	–	88.0	–	–	(Flint et al., 1998)
	Benin	–	100.0	–	–	–	–	(Gabriel and Przybylski, 2010)
	Burkina Faso	–	100.0	–	–	–	–	(Gabriel and Przybylski, 2010)
	Cameroon	–	83.7	16.3	–	–	–	(Flint et al., 1998)
	CAR	–	6.9	3.4	82.8	3.4	3.5	(Flint et al., 1998)
	Kenya	–	1.3	–	98.2	–	0.5	(Flint et al., 1998)
	Nigeria	–	92.9	3.4	0.7	0.9	2.1	(Flint et al., 1998)
	Senegal	–	14.0	–	1.8	80.7	3.5	(Flint et al., 1998)
	Tanzania	–	–	–	100.0	–	–	(Flint et al., 1998)
	Togo	–	100.0	–	–	–	–	(Gabriel and Przybylski, 2010)
North Africa & Middle East	Algeria	–	100.0	–	–	–	–	(Flint et al., 1998)
	Egypt	–	100.0	–	–	–	–	(Gabriel and Przybylski, 2010)
	Morocco	–	100.0	–	–	–	–	(Flint et al., 1998)
	Saudi Arabia	1.5	98.5	–	–	–	–	(el-Hazmi et al., 1999)
	Saudi Arabia	94.0	–	–	4.0	–	2.0	(Kulozik et al., 1986)
	Kuwait	77.8	16.7	–	–	–	5.5	(Adekile et al., 1994)
	Bahrain	90.0	2.5	–	5.0	–	2.5	(Al Arrayed and Haites, 1995)
	Syria	–	100.0	–	–	–	–	(Flint et al., 1998)
	Tunisia	–	94.8	–	–	–	5.2	(Flint et al., 1998)
	Turkey	0.4	96.3	–	–	–	3.3	(Flint et al., 1998)
	Turkey	–	100.0	–	–	–	–	(Gabriel and Przybylski, 2010)
South Asia	India	100.0	–	–	–	–	–	(Gabriel and Przybylski, 2010)
	India	90.7	–	–	–	–	9.3	(Labie et al., 1989)
	India	98.45	–	1.55	–	–	–	(Oner et al., 1992)
	India	91.67	–	2.78	1.39	–	4.17	(Niranjan et al., 1999)

The precise spatial patterns exhibited by β^S- and β^0-thalassaemia haplotypes are markedly different. For β^S, current data suggests that the five classical haplotypes predominantly occupy geographically separate regions within Sub-Saharan Africa, the Middle East and India (Table 1) (Bitoungui et al., 2015; Flint et al., 1998; Gabriel and Przybylski, 2010; Hanchard et al., 2007). By contrast, whilst β^0-thalassaemia mutations are mostly geographically specific on a cross-continental scale (Cao and Galanello, 2010; Weatherall and Clegg, 2001a) multiple variants can be found in the Mediterranean, the Middle East, and Asia, respectively, with their distributions considerably overlapping (Table 2). Furthermore, for each β^0-thalassaemia variant, various associated genetic backgrounds typically coexist in the population (Table 2). To illustrate, the β^0-thalassaemia mutation IVS-I-1 G → A is found on haplotypes V and II in Sicily (Schilirò et al., 1995), and haplotypes I, III, V, IX and A in Algeria (Bennani et al., 1994; Rouabhi et al., 1988). Similarly, the cd39 C → T mutation is associated with haplotypes I and II in Sicily, Sardinia and Corsica (Falchi et al., 2005), haplotypes I, II and IX in mainland Italy, and haplotypes I, II and B in Algeria (Lemsaddek et al., 2004).

In the case of β^0-thalassaemia, different causal mutations have clearly arisen independently, whilst the occurrence of identical mutations on separate haplotypes is generally ascribed to gene conversion (Chen et al., 2007). For β^S, however, it is commonly believed that each of the five classical β^S-associated haplotypes represents five independent occurrences of the same A → T mutation in codon 6 of β-globin (Chebloune et al., 1988; Kulozik et al., 1986; Pagnier et al., 1984; Trabuchet et al., 1991; Wainscoat et al., 1983). Yet, as suggested by Livingstone in the 1980s, a single β^S mutation, and its subsequent transfer onto different haplotypic backgrounds by gene conversion, could also have generated the same present-day β^S pattern (Flint et al., 1998; Livingstone, 1989a, 1989b).

The different patterns exhibited by β^0-thalassaemia and β^S mutations thus offer a unique opportunity to make a direct comparison between different sub-types of soft selective sweep in humans. Here, we identify the demographic and genetic processes that are more likely to give rise to either a sickle-cell-like or a β^0-thalassaemia-like spatial distribution of haplotypes. Within the context of our spatial framework, we also specifically address the role that recurrent mutation and gene conversion may have played in the evolution of these polymorphisms.

2. Methods

2.1. The model

We simulated a meta-population of *Ne* diploid individuals, divided into *d* demes of equal size and arranged in a network with a varying degree of randomness in its migration connection structure, controlled by parameter *c* (see Penman et al., 2012; Watts and Strogatz, 1998; and Supplementary material for further details). Every generation, *Ne* increases by a percentage drawn from a uniform distribution between zero and a maximum possible population growth rate of *g*% (see Supplementary material). Any increase in total population size is spread equally across all demes.

The meta-population is initially monomorphic for the wild-type (β^A) allele. In every generation *t*, within each deme, a finite set of potential offspring reaching reproductive age is generated according to (i) allele frequencies in that deme in generation *t-1*, (ii) mutation and/or gene conversion events (mutation rate = μ events per chromosome per generation; gene conversion rate = *r* events per chromosome per generation), and (iii) the relative fitness of each genotype, consistent with a standard model of the behaviour of a variant under selection. Following the generation of this offspring pool, individuals are randomly sampled without replacement to create the next generation of $\frac{Ne}{d}$ reproductively active adults in that deme. This step introduces the possibility of genetic drift. The final event to take place per generation is inter-demic gene flow. Each gene flow event involves the migration of *m*% of a deme's population to one of its linked partner demes in the migration network, and *vice versa* (further details in the Supplementary material).

2.2. Key model features/assumptions

2.2.1. Mutation and gene conversion

We are only interested in the haplotypic diversity of β^S- or β^0-thalassaemia-bearing chromosomes. The model thus records only (i) mutation in the $\beta^A \rightarrow \beta^X$ direction, and (ii) the transfer of a β^X allele onto a new haplotypic background by gene conversion in $\beta^A\beta^X$

Table 2
Distribution of β^0-thalassaemia variants and their associated haplotypes. Relative percentages (%) of β^0-thalassaemia variants and their associated haplotypes in a range of geographical settings. Haplotype definitions are given in Antonarakis et al. (1985).

Region	Country	β^0-Thalassaemia variants and their associated haplotypes									Reference
North Africa & Middle East		Cd37 (G→A)	Cd39 (C→T)	IVS-I-1 (G→A)	IVS-I-2 (T→C)	IVS2-1 (G→A)	FSC-8 (-AA)	FSC-6 (-A)	IVS-I-2 (T→G)	Other	
	Algeria	–	44.59 (I, II)	18.91 (I, III, V, IX, A)	5.40	–	–	27.70 (I, IX, A)	–	3.38	(Bennani et al., 1994)
	Algeria	–	50.00	11.90	19.05	–	–	16.67	–	2.38	(Bouhass et al., 1994)
	Morocco	4.67 (I)	39.26 (I, II Nd)	7.48 (V, IX)	7.48 II, IX)	3.74 (III)	20.56 (IV, VI, VII)	8.41 (III, IX)	4.67 (IX)	3.74 (IX, III)	(Agouti et al., 2008)
	Morocco	3.51 (VII, I)	24.56 (1, II)	21.05 (IV, V, IX)	5.26 (II, IX)	1.75 (III)	24.56 (IV, VI)	15.79 (IX, III)		3.51 (IX)	(Lemsaddek et al., 2003)
	Morocco	1.55	37.98 (I, II)	12.40 (IV, V, IX)	3.10	–	13.95 (IV, VI, VII)	19.38 (III, IX, A)	0.78	10.85	(Lemsaddek et al., 2004)
	Tunisia	–	62.07	–	–	–	3.45	3.45	3.45	27.59	(Fattoum et al., 1991)
Mediterranean & Europe		Cd8 (-AA)	Cd39 (C→T)	IVSI-1 (G→A)	IVS-I-2 (T→A)	IVS2-1 (G→A)	Cd6 (-A)	–	–	Other	
	Albania	–	65.00	15.00	–	5.00	–	–	–	15.00	(Boletini et al., 1994)
	Greece	–	50.86	39.66	–	6.03	–	–	–	3.45	(Boletini et al., 1994)
	Macedonia	–	21.43	59.52	–	7.14	–	–	–	11.90	(Boletini et al., 1994)
	Sardinia	–	97.80 (1, 2, 3, 4, 5)	0.035	–	0.035	2.13 (IX)	–	–	–	(Cao et al., 1991)
	Sicily	–	79.35 (I, II)	20.65 (II, V)	–	–	–	–	–	–	(Schilirò et al., 1997)
	Sicily	–	71.92 (I, II)	18.72 (II, V)	3.45	3.94	1.23	–	–	0.74	(Schilirò et al., 1995)
Asia		Cd8/9 (+G)	Cd15 (G→A)	Cd30 (G→C)	Cd41/42 (TCTT)	IVSI-1 (G→T)	IVSII-654 (C→T)	−619 bp del	–	Other	
	India	6.25 (3 haps.)	2.08 (1 hap.)	2.08 (1 hap)	10.42 (3 haps.)	12.50 (2 haps.)	–	2.08 (1 hap.)	–	64.58	(Gupta et al., 2008)
	Maldives	–	–	60.00 (B, C)	5.00 (F)	–	–	–	–	0.35	(Furuumi et al., 1998)

heterozygotes. All mutation and gene conversion rates throughout the manuscript refer to the rates at which these particular processes happen.

Every time a new β^X mutation arises, or an existing β^X mutation undergoes gene conversion, the resulting allele is assigned a unique numerical identifier representing a novel haplotypic background. This approach assumes a high diversity of pre-existing β-globin haplotypic backgrounds, such that each time a rare mutational or gene conversion event occurs it involves a different genetic background to any of those of previous mutations/gene conversions.

It is important to note that, given that the different haplotypes in our model can only arise through mutation or gene conversion (not reciprocal recombination), they are intended as proxies for β^S- and β^0-thalassaemia haplotypes whose occurrence cannot be accounted for by simple reciprocal recombination (*e.g.* the classical β^S haplotypes; Table 1) (Flint et al., 1998).

2.2.2. Assignment of fitness values

Full details of the assignment of fitness values are provided in the Supplementary material. Crucially, the fittest individuals in our simulated populations were always $\beta^A\beta^X$ heterozygotes, who were assumed to experience malaria protection. Throughout all of our simulations, $\beta^X\beta^X$ homozygotes were assigned a fitness of zero and thus were not represented in the potential offspring pool.

It is possible for the haplotypic background of mutations to affect the course of β^0-thalassaemia major or sickle-cell anaemia (Loggetto, 2013; Weatherall and Clegg, 2001b). However, given the absence, historically, of any effective treatment for these disorders, we have assumed that all individuals homozygous for β^S- or β^0-thalassaemia are likely to have been at a considerable disadvantage relative to the wild-type, regardless of whether they possessed a mutation with an ameliorating haplotypic background. We do, however, address the possibility of inter-haplotype fitness variation in our simulations, by incorporating a range, *f*, of possible heterozygote fitnesses into our model.

2.3. Parameterisation and implementation

The model was parameterised using value ranges taken from the literature where available (see Supplementary material). Of particular note, we tested four different allelic mutation rates: (i) 10^{-8} events per chromosome per generation, *i.e.* the average nucleotide substitution rate for the human genome (Ellegren et al., 2003; Xue et al., 2009), although this yielded very few instances of a soft selective sweep in our simulations; (ii) 10^{-7} events per chromosome per generation, accounting for the possibility of a higher-than-average rate of mutation in the β-globin cluster (Nachman and Crowell, 2000), and (iii) 5×10^{-7} events per chromosome per generation and (iv) 10^{-6} events per chromosome per generation to reflect the fact that hundreds of different types of mutations can give rise to a β^0-thalassaemia allele. Results from the latter three mutation rates are presented here. The ranges of values used for all other parameters are described in the Supplementary material.

All simulations shown were run for 500 generations. Assuming a generation time of 15–25 years for humans, this represents 7.5 to 12.5 thousand years of malaria selection, which is consistent with estimates for how long *P. falciparum* is likely to have been a significant cause of human mortality (Carter and Mendis, 2002). Adaptive alleles arose stochastically throughout each simulation. We chose to analyse a "snapshot" of the genetic variation in the meta-population at the 500 generation time point.

Simulations were implemented in Matlab R2012b and performed on a 1728 2.0 GHz cores super-computer part of the Advanced Research Computing (ARC) resources at the University of Oxford.

2.4. Classification of evolutionary outcomes

Based on the reported geographical distributions of β^S and β^0-thalassaemia variants and their associated haplotypes (Tables 1 and 2), we defined a series of possible outcomes within our model (see Sections 2.4.1–2.4.3). In particular, we sought to distinguish a spatially tessellating, 'patchwork' β^S-like pattern and an overlapping β^0-thalassaemia-like pattern.

2.4.1. β^0-Thalassaemia-like, or overlapping, outcome (Fig. 1A)

For this outcome, at least two different β^X-associated haplotypes must be present in the meta-population. In addition, there must be sufficient overlap in the distributions of the different haplotypes that no more than 20% of demes contain a β^X-associated haplotype that accounts for ≥95% of the haplotypic variation in the deme. We refer to such haplotypes henceforth as "dominating" haplotypes.

When assessing the geographical patterns exhibited by β^0-thalassaemia (Table 2), we counted two different β^0-thalassaemia mutations occurring on the same genetic background as two different haplotypes: considering the β^0-thalassaemia mutation itself to be part of the haplotypic diversity in the population.

2.4.2. β^S-type, or patchwork, outcome (Fig. 1B)

For this outcome, there must be at least two different dominating β^X-associated haplotypes in the whole meta-population, and at least 50% of the demes must contain a dominating β^X-associated haplotype.

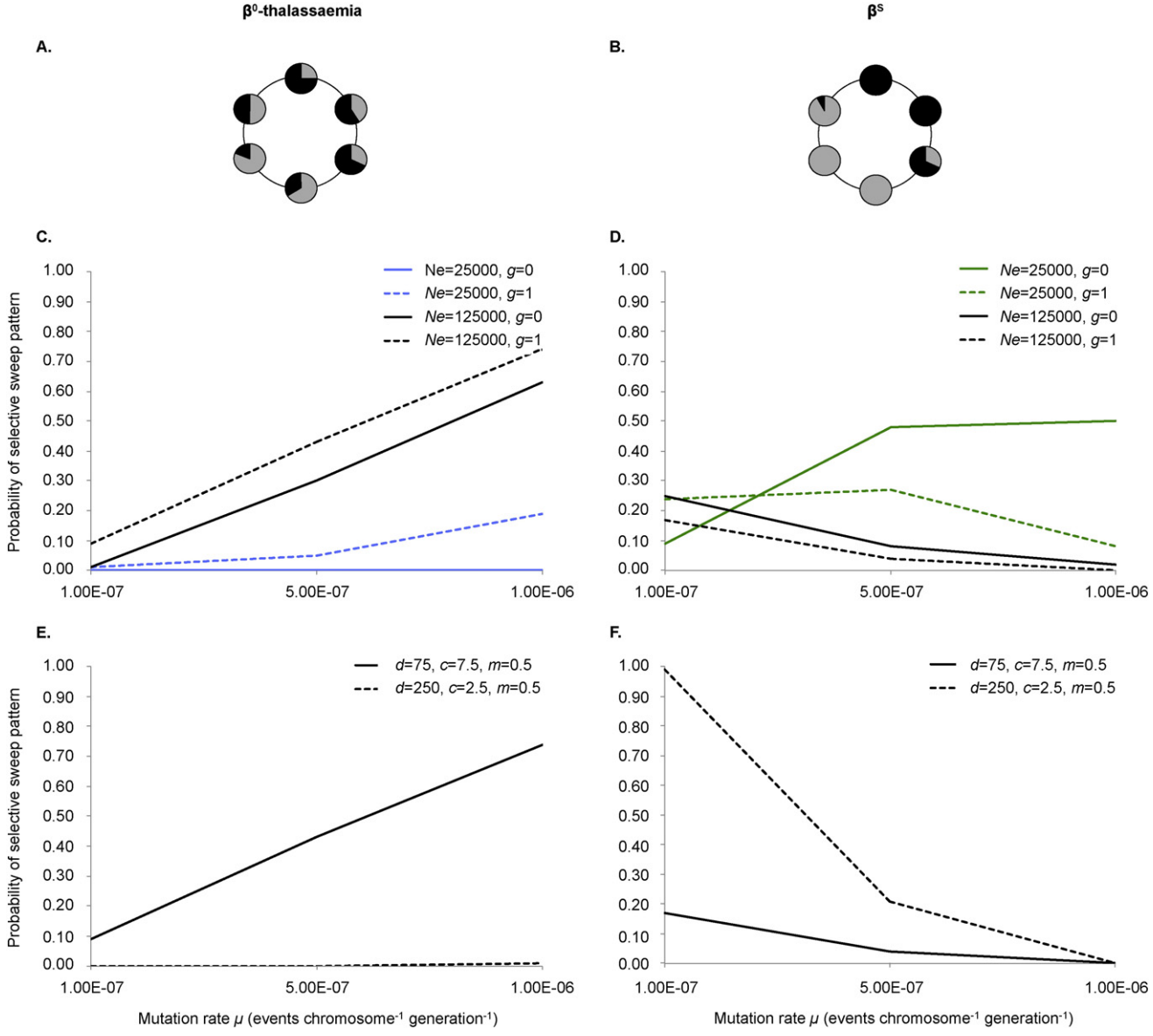


Fig. 1. The β^0-thalassaemia- and β^S-like patterns and the effect of mutation rate on their likelihood of emergence. Panels (A) and (B) are abstract representations of the β^0-thalassaemia- and β^S-like outcomes of our model; each circle represents a deme, whilst different colours indicate whether one (one colour) or multiple (more than one colour) β^X haplotypes are present in a deme. In panels (C–F), each graph indicates how the probability of observing a β^0-thalassaemia- (panels C and E) or β^S-like (panels D and F) pattern changes with changing mutation rate (x-axis). Each data point is based on 100 simulations. In panels (C) and (D), the different lines represent different demographic scenarios relating to population size and growth rate. Other parameters were fixed as follows: $d = 75$, c = 7.5, $m = 0.5$, $f = 0$ and $r = 10^{-6}$. Increasing the mutation rate increases the probability of the β^0-thalassaemia-like pattern, provided that initial total population size, *Ne*, and/or population growth rate, *g*, is sufficiently high. In panels (E) and (F), the different lines represent (i) a relatively weak population structure (fewer demes, more random connections: solid line), or (ii) a relatively strong population structure (many demes, fewer random connections: dashed line). Other parameter values were fixed as follows: $Ne = 125{,}000$, $G = 1\%$, $f = 0$ and $r = 10^{-6}$. The effect of mutation rate on the probability of observing the β^0-thalassaemia-like pattern is strongest for a weak population structure, whilst the opposite is true for the β^S-like pattern.

2.4.3. Other possible model outcomes

Other possible model outcomes include: (i) no malaria-protective variation at the β-globin locus; (ii) a hard selective sweep, whereby malaria-protective variation is associated with only a single haplotype in the meta-population; (iii) the co-occurrence of malaria-protective variation on multiple haplotypes in the meta-population where haplotypes are completely deme-specific; and (iv) the co-occurrence of multiple haplotypes in the meta-population whose distributions are not deme-specific but do not reflect closely enough the overlapping (β^0-thalassaemia) or patchwork (β^S) spatial patterns defined above.

3. Results

3.1. Different mutation rates favour β^S- or β^0-thalassaemia-like patterns, subject to population size and structure

β^0-Thalassaemic mutations can result from deletions, insertions and point mutations anywhere in the coding or regulatory region of β-globin. The β^S mutation, by contrast, is the result of the replacement of a specific, single nucleotide with another. There are therefore strong biological reasons to suppose that β^0-thalassaemic mutations arise much more frequently than β^S mutations.

In our simulations, increasing the mutation rate did increase the probability of observing the β^0-thalassaemia-like pattern, but only if the overall population size was high (Fig. 1C). If the population was too small and/or too highly structured (*i.e.* with high population subdivision, a strong connection structure and low migration between demes), the probability of the β^0-thalassaemia-like pattern remained low (<0.20) even at the highest tested mutation rate. This was due to there being insufficient genetic variation or population movement to facilitate overlap in the haplotypes' distributions (Fig. 1C, E, respectively). The probability of a β^S-like (patchwork) haplotype pattern, by contrast, was negatively correlated with mutation rate (Fig. 1D, F), except in the case of a low initial total population size ($Ne = 25{,}000$) and minimal population growth ($g \leq 0.5\%$) (Fig. 1D).

We also observed an interaction between the effects of initial total population size and population growth on the probability of the β^S-like pattern at low mutation rates (Supplementary Fig. S1A). Population growth had a positive effect on the probability of the β^S-like pattern when the initial total population size was small ($Ne = 25000$); such growth increased the overall size of the population, and thus the chances of more than one β^S haplotype arising anywhere. However, for a much larger starting population size ($Ne = 125000$) population growth had a slight negative effect. This is because increasing the size of an already large population led to too many β^S haplotypes arising through mutation and intermingling, thereby preventing a patchwork β^S-like pattern. At a higher mutation rate ($\mu = 10^{-6}$), population growth rate had a negative effect on the likelihood of the β^S-like pattern across all initial population sizes.

3.2. A β^0-thalassaemia-like pattern is most likely in a meta-population with a highly random connection structure, comprising few demes between which there is considerable gene flow; the opposite is true for a β^S-like pattern

As illustrated in Fig. 2A and C, for a given initial total population size and rate of population growth, the β^0-thalassaemia-like pattern was more likely under conditions of low subdivision, high migration and when the migration network contained more random connections. By contrast, the probability of obtaining the β^S-like pattern was highest when population subdivision was high, the migration network was non-random and migration was low (Fig. 2B, D).

For both patterns of selective sweep, population subdivision and the degree of randomness in the migration network had a weaker effect at smaller initial population sizes when $\mu = 10^{-7}$ events per chromosome per generation (Supplementary Figs. S2 and S3). Presumably this is because, even if the spread of alleles in the meta-population is slowed by high population subdivision or a highly non-random migration network, this has little impact if opportunities for new copies of the allele to arise are few. The converse is true for β^S when $\mu = 10^{-6}$ events per chromosome per generation; in this case, the effects of population subdivision and connection structure are strongest when the initial total population size is small. Moreover, the effect of population subdivision was greatest when the connectivity network was non-random (Fig. 2A, B), indicating that even at high levels of subdivision a more random migration network is sufficient to allow the "small world" phenomenon (Watts and Strogatz, 1998; Watts, 1999) to occur, minimising the number of migratory steps that it takes for an allele to access to the entire network.

3.3. When haplotype fitnesses vary, gene conversion is more likely to contribute to the haplotypic diversity of either β^0-thalassaemia or β^S

The β-globin cluster incorporates the γ- and δ-genes as well as an extensive locus control region (Forget and Hardison, 2009), mutations in any of which are capable of affecting the phenotypic outcome of a mutation in the coding region of β-globin itself. This phenomenon could be deemed epistasis, although the tight linkage between all elements of the β-globin cluster makes it equally acceptable to conceive of different haplotypes as allelic to one another. In either case, it is entirely plausible that β^S and β^0-thalassaemia mutations could have different fitnesses according to their haplotypic background. We addressed the possibility of haplotypic fitness variation by randomly assigning a heterozygote fitness value, drawn from a predefined range of width f, to each haplotype as it arose (see Supplementary material). Including inter-haplotype fitness variation decreased the probability of observing the β^0-thalassaemia-like pattern (Fig. 3A and C). This was also true for the β^S-like pattern when $\mu = 10^{-7}$ (Fig. 3B), although when $\mu = 10^{-6}$, inter-haplotype fitness variation increased the probability of observing the β^S-like pattern (Fig. 3D).

The rate of gene conversion had almost no effect on the overall probability of either the β^S- or β^0-thalassaemia-like pattern (Fig. 3). However, whenever gene conversion was allowed to occur, inter-haplotype fitness variation increased the proportion of scenarios in which haplotypes resulting from gene conversion formed part of the *final* haplotypic diversity of the population (Fig. 3). For example, for the β^0-thalassaemia-like pattern, when the gene conversion rate was 10^{-5} events per chromosome per generation and the mutation rate was 10^{-7} events per chromosome per generation, gene conversion contributed to the final haplotypic diversity 97% of the time if $f > 0$, but <1% of the time if $f = 0$ (Fig. 3A).

There has so far been no attempt to quantify the relative heterozygote fitnesses of different β^S- and β^0-thalassaemia-associated haplotypes, although clinical evidence does suggest that the severity of sickle-cell anaemia and β^0-thalassaemia in homozygotes can vary according to haplotypic background (Ashley-Koch et al., 2000; Tsaras et al., 2009). It is therefore difficult to know whether the maximum fitness range included in Fig. 3 ($f = 0.2$) is plausible. Our model additionally shows that the haplotypes that coexisted in the long-term tended to be relatively similar in their heterozygote fitness, (Fig. 4), so a study carried out today may not provide a fair picture of what fitness variation could have existed in the past. Amongst *all* β^0-thalassaemia-like results where $f > 0$, the average within-simulation heterozygote fitness range of haplotypes that coexisted in the meta-population after 500 generations was 0.03, compared to 0.18 for *all* haplotypes that arose during the 500 generations. For β^S-like results, these values were 0.05 and 0.16, respectively. The average fitness range of dominating haplotypes was 0.01 for β^0-thalassaemia-like repeats and 0.04 for β^S-like repeats. The fittest haplotypes to arise contributed to the final haplotypic diversity in only 15% and 26% of all of the repeats exhibiting the β^S- and β^0-thalassaemia-like patterns, respectively, when $f > 0$.

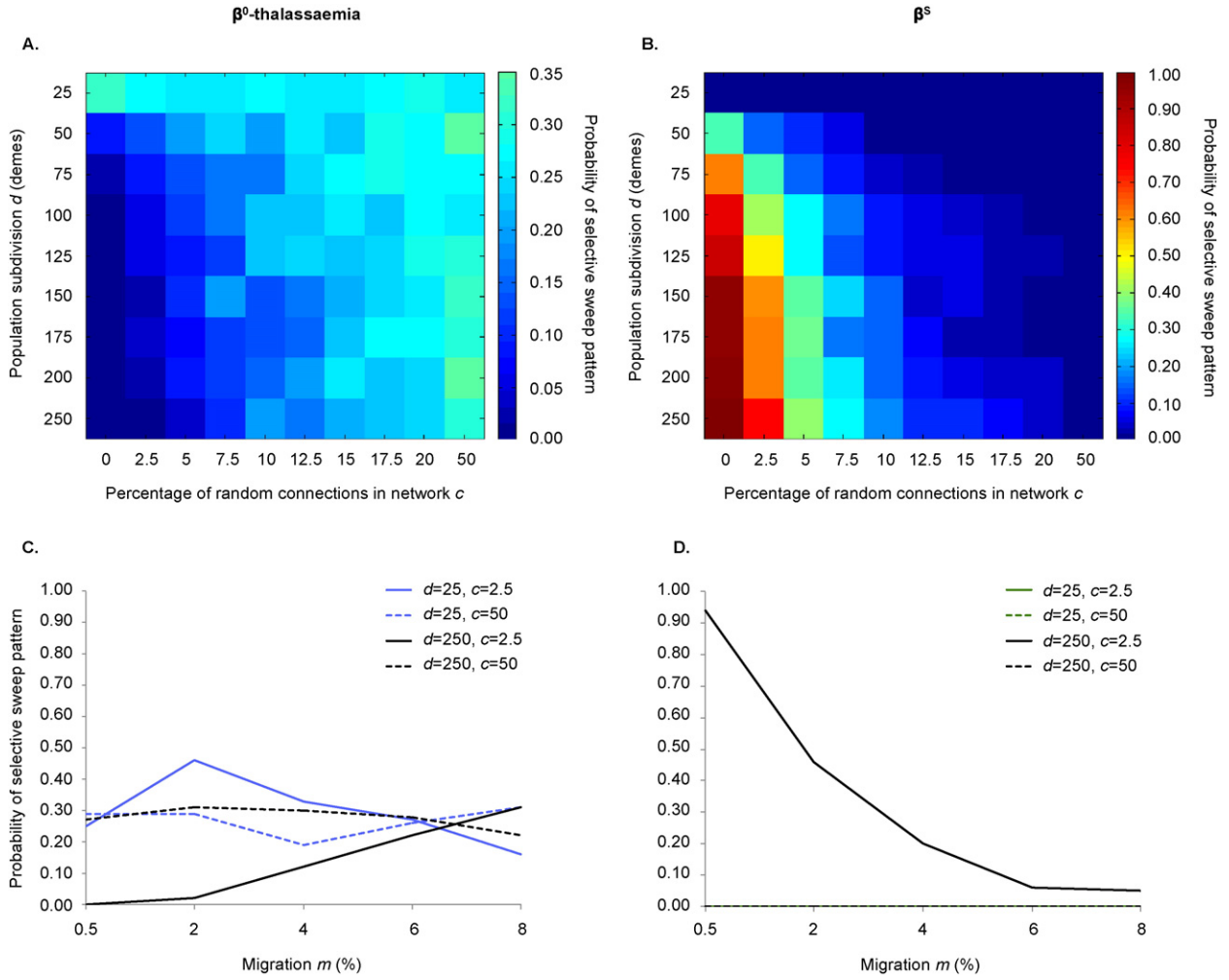


Fig. 2. The combined effects of population subdivision, connection structure and migration on the probabilities of different types of selective sweep. The heatmaps indicate the effects of population subdivision and connection structure on the probability of observing β^0-thalassaemia- (panel A) or β^S-like (panel B) patterns. The colour of each square depicts the proportion of repeats (100 in total) that exhibit the pattern of interest. The x-axis of each panel indicates the percentage of connections in the network that are random, which corresponds directly to parameter c (see Supplementary material). Panels (C) and (D) indicate the additional impact of varying the level of gene flow (x-axis). Each data point is based on 100 simulations. Unless otherwise indicated in each panel, parameters were fixed as follows: $Ne = 125{,}000$, $g = 1$, $f = 0$, $r = 0$ and $\mu = 10^{-7}$. In panels (A) and (B), results were averaged across two rates of migration: $m = 0.5\%$ and 2%. The β^0-thalassaemia-like pattern is most likely when population subdivision is low and the connection structure is random; by contrast the β^S-like pattern requires high population subdivision, a non-random connection structure and minimal gene flow to maximise its occurrence.

4. Discussion

Across all of our simulations, once a deme had come to contain a β^S- or β^0-thalassaemia-associated haplotype at a high frequency, that deme was rarely taken over by another haplotype. This is because alleles that arose later contributed only a very small fraction of the adaptive variation in question and thus were vulnerable to loss by genetic drift. We refer to this phenomenon as "allelic exclusion" to coincide with the terminology used by Ralph and Coop (Ralph and Coop, 2010). Importantly, we found that the generation of either β^S- or β^0-thalassaemia-like patterns requires allelic exclusion to be undermined, but to different spatial extents; the β^S-like pattern requires that allelic exclusion be maintained in parts of the meta-population but not the entire network, whilst the near complete avoidance of allelic exclusion is necessary for the β^0-thalassaemia-like pattern.

Allelic exclusion can be avoided if a pre-existing haplotype has not yet reached a threshold frequency in a deme when subsequent haplotypes arrive. It follows that the timing of mutation, gene conversion and/or migration events within each deme is important in determining the evolutionary trajectory of the meta-population. Alternatively, allelic exclusion is undermined when genetic drift is weakened for incoming alleles, either through population growth or fitness variation between haplotypes. As illustrated in Figs. 1–3, the balance between all of these factors determines the probability of either a β^0-thalassaemia- or β^S-like pattern occurring. However, generally speaking, we expect the β^S-like pattern to emerge when mutation rate is low and the population is highly subdivided, with low connectivity and little gene flow. The β^0-thalassaemia-like pattern is more likely when mutation rate is high and the population is less subdivided, with high connectivity and high gene flow.

Several previous theoretical treatments of soft selective sweeps have delivered important insights into the genetic and demographic factors influencing the probability of adaptation by soft selective sweep *versus* hard selective sweep (see Messer and Petrov, 2013 for a full review). In a series of papers, Pennings and Hermisson used coalescent theory to show that soft sweeps are most likely when population size is large and/or allelic mutation rate is high (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006). More recently, Ralph and Coop (2010) demonstrated that soft sweeps, specifically of the patchwork type, are likely to be common in species whose distributions are widespread and whose populations are geographically structured (Ralph and Coop, 2010). The behaviour of our model is consistent with these previous studies. By modelling a meta-population where we do not assume that different alleles exclude one another when they meet, we

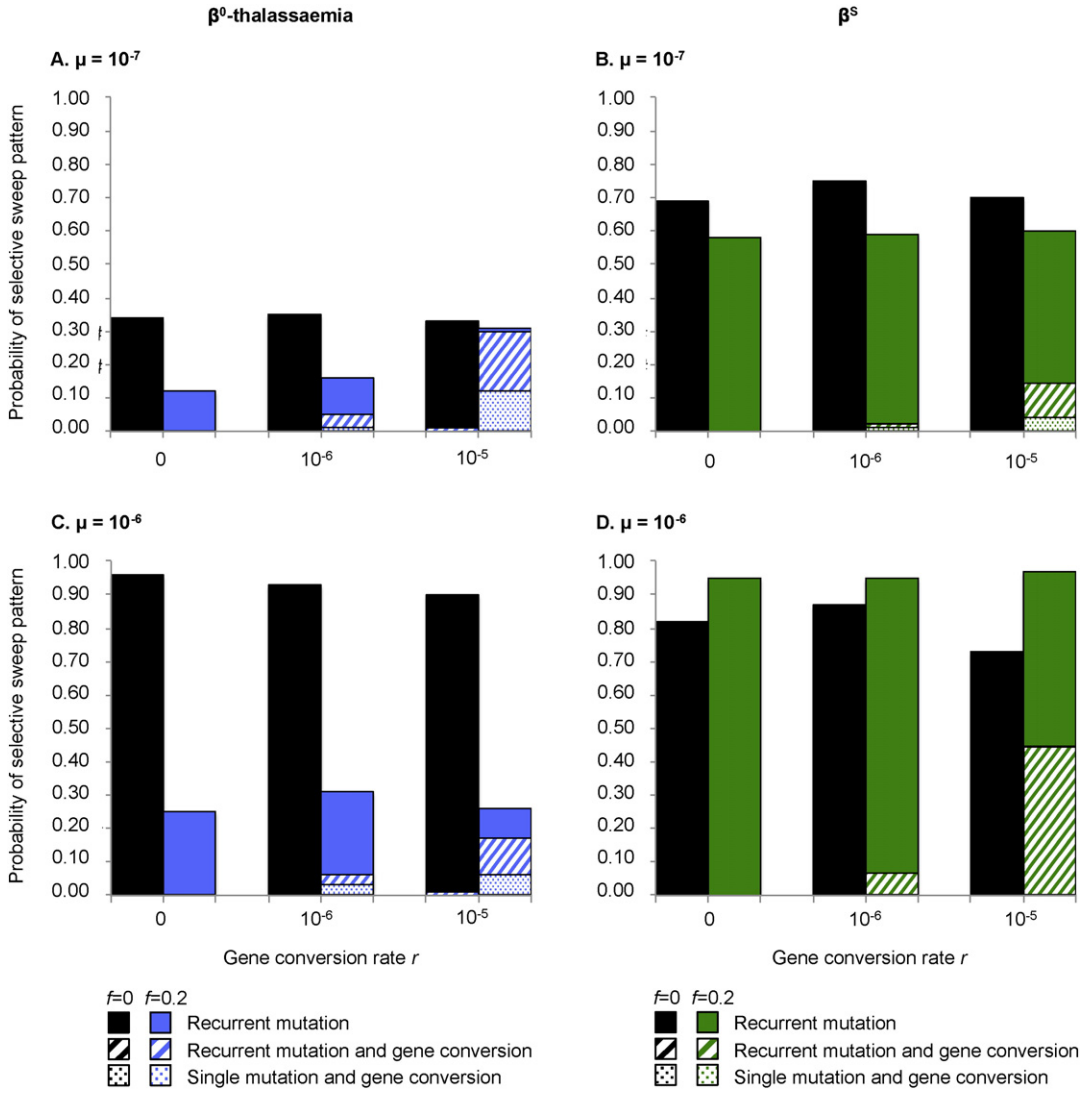


Fig. 3. The effects of gene conversion and fitness variation on the probabilities of different types of selective sweep. The total height of each bar indicates the overall probability of observing a β^0-thalassaemia-like (A,C) or β^S-like (B,D) pattern at different levels of gene conversion (x-axis), and in the absence (black bars) or presence (blue or green bars) of fitness variation amongst heterozygotes. Each bar is based on 100 simulations, and is subdivided to indicate whether gene conversion and/or recurrent mutation was responsible for the haplotypic diversity present in the population at the end of each simulation (see legend to each graph). Results from two different mutation rates are shown: 10^{-7} (A,B), and 10^{-6} (C,D). Demographic parameter values in panels (A) and (C) were fixed as follows, to maximise the probability of a β^0-thalassaemia-like scenario: $Ne = 125{,}000$, $g = 1$, $d = 125$, $c = 15$ and $m = 2$. Demographic parameter values in panels (B) and (D) were fixed as follows, to maximise the probability of a β^S-like scenario: $Ne = 25000$, $g = 1$, $d = 200$, $c = 2.5$, $m = 0.5$. The inclusion of inter-haplotype fitness variation consistently decreases the probability of the β^0-thalassaemia-like pattern. Whilst gene conversion has no effect on the overall probability of either patterns with or without inter-haplotype fitness variation, its relative contribution to the final haplotypic diversity increases with the rate of gene conversion if fitness variation is also present.

are also able to show that a mutation rate that is too high precludes the possibility of a β^S-like soft selective sweep pattern, whilst a weaker geographical structure is important for the formation of a β^0-thalassaemia-like pattern.

As noted in the Introduction, the occurrence of the same β^0-thalassaemia variant on multiple haplotypes is generally attributed to gene conversion. Our results imply that gene conversion can contribute to haplotypic diversity only if inter-haplotype fitness is sufficiently variable. β^0-Thalassaemia variants certainly vary in their clinical severity (Weatherall and Clegg, 2001a), often due to factors such as different levels of expression of foetal haemoglobin. No study has yet compared the relative level of malaria protection that is afforded by heterozygosity for different β^0-thalassaemia haplotypes, but it is entirely possible that variable maintenance of foetal haemoglobin might affect malaria susceptibility (Amaratunga et al., 2011; Billig et al., 2012). Curiously, however, we found that, whilst including fitness variation made it more likely that gene conversion contributes to long-term haplotypic diversity, it simultaneously made the β^0-thalassaemia-like pattern less likely (although not impossible). A specific combination of demographic conditions, gene conversion rate and inter-haplotype fitness variation, which increases the probability of observing a β^0-thalassaemia-like pattern where the haplotypic diversity is partly derived from gene conversion, may yet be discovered.

Present consensus seems to be that gene conversion has had no role in the generation of the classical β^S haplotypes. This is despite modelling work by Livingstone in 1989, in which he used a stochastic model of the diffusion of different β^S- and β^A-associated chromosomes to demonstrate that reciprocal recombination and gene conversion readily give rise to multiple β^S haplotypes, with no need for recurrent mutation (Livingstone, 1989a). Our model demonstrates that a patchwork haplotype pattern that is at least partly derived from gene conversion is difficult to obtain unless inter-haplotype fitness variation exists. It is therefore possible that, until we understand what fitness variation is possible amongst β^S haplotypes, we will not be able to judge properly the role of gene conversion in its evolution. However, as indicated by our simulations, it is important to bear in mind that the observed

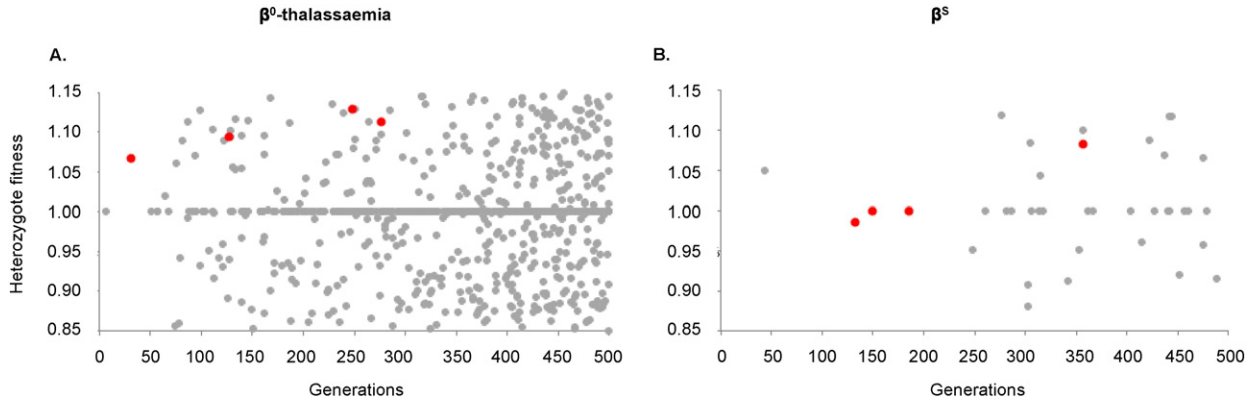


Fig. 4. Comparing the fitnesses of all haplotypes that arose *versus* those that succeeded. Each panel indicates the heterozygote fitness values assigned to all haplotypes that arose over the course of a single simulation (grey), compared to those haplotypes that remained in the meta-population at a frequency ≥1% after 500 generations (red). The range of fitness values for the latter is considerably smaller than that for the former. Panel (A) illustrates a β^0-thalassaemia-like scenario, for which the parameter values were: $Ne = 125{,}000$, $g = 1$, $d = 125$, $c = 15$, $m = 2$, $f = 0.3$, $r = 10^{-5}$ and $\mu = 10^{-7}$. Panel (B) illustrates a β^S-like scenario, for which the parameter values were $Ne = 75{,}000$, $g = 0$, $d = 75$, $c = 5$, $m = 0.5$, $f = 0.3$, $r = 5 \times 10^{-6}$ and $\mu = 10^{-7}$.

present-day inter-haplotype fitness variation for both β^S- and β^0-thalassaemia may not necessarily reflect the full fitness range of all haplotypes that have arisen over the course of human evolutionary history and may not include the fittest haplotypes to have ever existed.

There is good evidence for past gene conversion events in the β-globin cluster (reviewed in Papadakis and Patrinos, 1999). Further and improved sequence data from this region of the genome will continue to provide insight into these processes, and may be able to indicate whether gene conversion has played a role in the generation of the classical β^S haplotypes. However, given that gene conversion events can involve a few hundred bases, which for a conversion event involving the β^S mutation is likely to include the highly conserved coding region of the β-globin gene, it may not always be possible to distinguish between *de novo* mutation and gene conversion at the β^S locus using sequence data alone.

We defined a patchwork β^S-like pattern based on the geographical distribution of the classical β^S haplotypes. As noted in the Introduction, classical β^S haplotypes derive from RFLP analyses, which continue to be used in present-day studies of sickle-cell diversity (*e.g.* Bitoungui et al., 2015). Using SNP markers, Hanchard and colleagues showed that the classical Benin and Senegal haplotypes both exhibit a high degree of long-range haplotypic similarity extending across more than 400 kb in three separate populations (Hanchard et al., 2007). Similar results were found in a Ghanaian population (Ghansah et al., 2012). Fine-scale sequence analysis has revealed heterogeneities within the classical Benin and Bantu haplotypes (Bouhassira and Nagel, 1990; Patrinos et al., 2005). However, the distribution of the observed polymorphisms suggests that these differences evolved after the emergence of β^S on the distinct classical haplotypes, so as such, the broad pattern of the classical RFLP haplotypes remains.

β^0-Thalassaemia mutations completely eliminate β-globin production from the affected gene. Other mutations exist which reduce but do not eliminate the production of β-globin, designated β^+-thalassaemia alleles. Like β^0-thalassaemia, β^+-thalassaemia mutations are associated with multiple haplotypic backgrounds whose distributions have been found to overlap. To some extent, therefore, the results we present here for β^0-thalassaemia also apply to β^+-thalassaemia. However, our present assumption of zero fitness for homozygotes for the relevant mutation is less reasonable for certain milder β^+-thalassaemic variants. We predict that allowing for milder homozygous, or compound heterozygous, phenotypes will allow an overlapping haplotypic pattern to be obtained over a still wider range of parameter space. We propose to explore this in the future as part of a model that allows a wider range of malaria-protective globin mutations to compete with one another.

Our work so far has focused on the β-globin locus. The study of the haplotypic evolution of α-globin will require a different modelling approach, incorporating duplicated α-globin genes in the α-globin cluster; the possible occurrence of the same variants in paralogous genes (Moradkhani et al., 2009); and the wide array of α-thalassaemia variants that are observed in human populations, including both single and double gene deletions (Weatherall and Clegg, 2001b). In this way, the relative contributions of recurrent mutation, gene conversion and unequal crossover in generating complex genetic variation in the α-globin gene cluster can be explored, along with the roles of malaria selection and demographic factors in shaping the spatial pattern of this diversity.

Our theoretical framework can be extended in a number of ways. One informative next step will be to allow β^S and β^0-thalassaemia mutations to compete within the same interconnected network, alongside β^+-thalassaemia mutations and other malaria-protective alleles with less severe clinical outcomes, for example HbC in Africa and HbE in Asia (Fucharoen and Weatherall, 2012; Piel et al., 2013). Further investigation into the origin, maintenance and fate of different β^0-thalassaemia and β^S haplotypes will also need to consider the possible influence of epistasis between mutations at the α- and β-globin loci (Williams et al., 2005b; Penman et al., 2009, 2011); as well as interactions with other malaria-protective genetic variants elsewhere in the genome (reviewed in Hedrick, 2011; Kwiatkowski, 2005; Williams, 2006). Finally, Wilson and colleagues recently showed that, depending on their severity and frequency of recurrence, population bottlenecks can cause a soft selective sweep to become hard (Wilson et al., 2014). It would be interesting to see how their results relate to the specific case of β-globin polymorphisms under malaria selection.

5. Conclusion

Sickle-cell trait and β^0-thalassaemia are two of our best examples of recent human evolution. Here we have shown that their differing selective sweep patterns may be just as much a product of different demographic conditions as they are of different mutation rates. Our results also suggest that inter-haplotypic fitness variation – a very real possibility for β-globin variants – both affects the probability of observing specific haplotype patterns and increases the probability of gene conversion having contributed to the variation present today. A better understanding of the fitness variation that is possible amongst β^0-thalassaemia- or β^S-associated haplotypes, will therefore be critical in determining the role of gene conversion in their evolution.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.meegid.2015.09.018.

Acknowledgements

The authors would like to acknowledge the use of the Oxford Advanced Research Computing (ARC) in carrying out this work. We also thank José Lourenço for comments on earlier versions of the manuscript. B.S.P. is a Sir Henry Wellcome Postdoctoral Fellow (Grant 096063/Z/11/Z) and a Junior Research Fellow at Merton College. S.G. is a Royal Society Wolfson Research Fellow and an ERC advanced investigator (DIVERSITY).

References

Adekile, A.D., Gu, L.H., Baysal, E., Haider, M.Z., Al-Fuzae, L., Aboobacker, K.C., Al-Rashied, A., Huisman, T.H., 1994. Molecular characterization of alpha-thalassemia determinants, beta-thalassemia alleles, and beta S haplotypes among Kuwaiti Arabs. Acta Haematol. 92, 176–181.

Agouti, I., Badens, C., Abouyoub, A., Levy, N., Bennani, M., 2008. Molecular basis of beta-thalassemia in Morocco: possible origins of the molecular heterogeneity. Genet. Test. 12, 563–568.

Al Arrayed, S., Haites, N., 1995. Features of sickle-cell disease in Bahrain. East Mediterr. Health J. 1, 112–119.

Allison, A.C., 1964. Polymorphism and natural selection in human populations. Cold Spring Harb. Symp. Quant. Biol. 29, 137–149. http://dx.doi.org/10.1101/SQB.1964.029.01.018.

Amaratunga, C., Lopera-Mesa, T.M., Brittain, N.J., Cholera, R., Arie, T., Fujioka, H., Keefer, J.R., Fairhurst, R.M., 2011. A role for fetal hemoglobin and maternal immune IgG in infant resistance to *Plasmodium falciparum* malaria. PLoS One 6, e14798.

Antonarakis, S., Kazazian, H., Orkin, S., 1985. DNA polymorphism and molecular pathology of the human globin gene clusters. Hum. Genet. 69, 1–14.

Ashley-Koch, A., Yang, Q., Olney, R.S., 2000. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am. J. Epidemiol. 151, 839–845.

Bennani, C., Bouhass, R., Perrin-Pecontal, P., Tamouza, R., Malou, M., Elion, J., Trabuchet, G., Beldjord, C., Benabadji, M., Labie, D., 1994. Anthropological approach to the heterogeneity of beta-thalassemia mutations in northern Africa. Hum. Biol. 66, 369–382.

Billig, E.M.W., McQueen, P.G., McKenzie, F.E., 2012. Foetal haemoglobin and the dynamics of paediatric malaria. Malar. J. 11, 396.

Bitoungui, V.J.N., Pule, G.D., Hanchard, N., Ngogang, J., Wonkam, A., 2015. Beta-globin gene haplotypes among Cameroonians and review of the global distribution: is there a case for a single sickle mutation origin in Africa? OMICS 19, 171–179.

Boletini, E., Svobodova, M., Divoky, V., Baysal, E., Cürük, M.A., Dimovski, A.J., Liang, R., Adekile, A.D., Huisman, T.H., 1994. Sickle cell anemia, sickle cell beta-thalassemia, and thalassemia major in Albania: characterization of mutations. Hum. Genet. 93, 182–187.

Bouhass, R., Perrin, P., Trabuchet, G., 1994. The spectrum of beta-thalassemia mutations in the Oran region of Algeria. Hemoglobin 18, 211–219.

Bouhassira, E.E., Nagel, R.L., 1990. A 6-bp deletion 5′ to the G gamma globin gene in beta S chromosomes bearing the Bantu haplotype. Am. J. Hum. Genet. 47, 161–163.

Cao, A., Galanello, R., 2010. Beta-thalassemia. Genet. Med. 12, 62–73.

Cao, A., Rosatelli, C., Pirastu, M., Galanello, R., 1991. Thalassemias in Sardinia: molecular pathology, phenotype–genotype correlation, and prevention. Am. J. Pediatr. Hematol. Oncol. 13, 179–188.

Carter, R., Mendis, K.N., 2002. Evolutionary and historical aspects of the burden of malaria. Clin. Microbiol. Rev. 15, 564–594.

Chebloune, Y., Pagnier, J., Trabuchet, G., Faure, C., Verdier, G., Labie, D., Nigon, V., 1988. Structural analysis of the 5′ flanking region of the beta-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. Proc. Natl. Acad. Sci. U. S. A. 85, 4431–4435.

Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C., Patrinos, G.P., 2007. Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8, 762–775.

el-Hazmi, M.A., Warsy, A.S., Bashir, N., Beshlawi, A., Hussain, I.R., Temtamy, S., Qubaili, F., 1999. Haplotypes of the beta-globin gene as prognostic factors in sickle-cell disease. East. Mediterr. Health J. 5, 1154–1158.

Ellegren, H., Smith, N.G., Webster, M.T., 2003. Mutation rate variation in the mammalian genome. Curr. Opin. Genet. Dev. 13, 562–568.

Falchi, A., Giovannoni, L., Vacca, L., Latini, V., Vona, G., Varesi, L., 2005. Beta-globin gene cluster haplotypes associated with beta-thalassemia on Corsica island. Am. J. Hematol. 78, 27–32.

Fattoum, S., Guemira, F., Oner, C., Oner, R., Li, H.W., Kutlar, F., Huisman, T.H., 1991. Beta-thalassemia, HB S-beta-thalassemia and sickle cell anemia among Tunisians. Hemoglobin 15, 11–21.

Flint, J., Harding, R.M., Boyce, A.J., Clegg, J.B., 1998. The population genetics of the haemoglobinopathies. Baillieres Clin. Haematol. 11, 1–51.

Forget, B., Hardison, R., 2009. The normal structure and regulation of human globin gene clusters. Disorders of Hemoglobin. Cambridge University Press, pp. 46–61. http://dx.doi.org/10.1017/CBO9780511596582.007.

Fucharoen, S., Weatherall, D.J., 2012. The hemoglobin E thalassemias. Cold Spring Harb. Perspect. Med. 2, a011734–.

Furuumi, H., Firdous, N., Inoue, T., Ohta, H., Winichagoon, P., Fucharoen, S., Fukumaki, Y., 1998. Molecular basis of beta-thalassemia in the Maldives. Hemoglobin 22, 141–151.

Gabriel, A., Przybylski, J., 2010. Sickle-cell anemia: a look at global haplotype distribution. Nat. Educ. 3, 2.

Ghansah, A., Rockett, K.A., Clark, T.G., Wilson, M.D., Koram, K.A., Oduro, A.R., Amenga-Etego, L., Anyorigiya, T., Hodgson, A., Milligan, P., Rogers, W.O., Kwiatkowski, D.P., 2012. Haplotype analyses of haemoglobin C and haemoglobin S and the dynamics of the evolutionary response to malaria in Kassena-Nankana District of Ghana. PLoS One 7, e34565.

Gupta, A., Sarwai, S., Pathak, N., Agarwal, S., 2008. Beta-globin gene mutations in India and their linkage to β-haplotypes. Int. J. Hum. Genet. 8, 237–241.

Hanchard, N., Elzein, A., Trafford, C., Rockett, K., Pinder, M., Jallow, M., Harding, R., Kwiatkowski, D., McKenzie, C., 2007. Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. BMC Genet. 8, 52.

Hedrick, P.W., 2011. Population genetics of malaria resistance in humans. Heredity (Edinb) 107, 283–304.

Hermisson, J., Pennings, P.S., 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169, 2335–2352.

Hill, A.V., Allsopp, C.E., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J., Greenwood, B.M., 1991. Common west African HLA antigens are associated with protection from severe malaria. Nature 352, 595–600.

Kulozik, A.E., Wainscoat, J.S., Serjeant, G.R., Kar, B.C., Al-Awamy, B., Essan, G.J., Falusi, A.G., Haque, S.K., Hilali, A.M., Kate, S., 1986. Geographical survey of beta S-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. Am. J. Hum. Genet. 39, 239–244.

Kwiatkowski, D.P., 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77, 171–192.

Labie, D., Srinivas, R., Dunda, O., Dode, C., Lapoumeroulie, C., Devi, V., Devi, S., Ramasami, K., Elion, J., Ducrocq, R., Krishnamoorthy, R., Nagel, R.L., 1989. Haplotypes in tribal Indians bearing the sickle gene: evidence for the unicentric origin of the beta S mutation and the unicentric origin of the tribal populations of India. Hum. Biol. 61, 479–491.

Lemsaddek, W., Picanço, I., Seuanes, F., Mahmal, L., Benchekroun, S., Khattab, M., Nogueira, P., Osório-Almeida, L., 2003. Spectrum of beta thalassemia mutations and HbF levels in the heterozygous Moroccan population. Am. J. Hematol. 73, 161–168.

Lemsaddek, W., Picanço, I., Seuanes, F., Nogueira, P., Mahmal, L., Benchekroun, S., Khattab, M., Osório-Almeida, L., 2004. The β-thalassemia mutation/haplotype distribution in the Moroccan population. Hemoglobin 28, 25–37.

Livingstone, F., 1989a. Who gave whom hemoglobin S: the use of restriction site haplotype variation for the interpretation of the evolution of the βS-globin gene. Am. J. Hum. Biol. 1, 289–302.

Livingstone, F., 1989b. Simulation of the diffusion of the beta-globin variants in the Old World. Hum. Biol. 61, 297–309.

Loggetto, S.R., 2013. Sickle cell anemia: clinical diversity and beta S-globin haplotypes. Rev. Bras. Hematol. Hemoter. 35, 155–157.

Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28, 659–669.

Moradkhani, K., Préhu, C., Old, J., Henderson, S., Balamitsa, V., Luo, H.-Y., Poon, M.-C., Chui, D.H.K., Wajcman, H., Patrinos, G.P., 2009. Mutations in the paralogous human alpha-globin genes yielding identical hemoglobin variants. Ann. Hematol. 88, 535–543.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304.

Niranjan, Y., Chandak, G.R., Veerraju, P., Singh, L., 1999. Some atypical and rare sickle cell gene haplotypes in populations of Andhra Pradesh, India. Hum. Biol. 71, 333–340.

Oner, C., Dimovski, A., Olivieri, N., Schiliro, G., Codrington, J., Fattoum, S., Adekile, A., Oner, R., Yuregir, G., Altay, C., Gurgey, A., Gupta, R., Jogessar, V., Kitundu, M., Loukopoulos, D., Tamagnini, G., Ribeiro, M.L., Kutlar, F., Gu, L.-H., Lanclos, K., Huisman, T.J., 1992. Beta S haplotypes in various world populations. Hum. Genet. 89, 99–104.

Pagnier, J., Mears, J.G., Dunda-Belkhodja, O., Schaefer-Rego, K.E., Beldjord, C., Nagel, R.L., Labie, D., 1984. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc. Natl. Acad. Sci. 81, 1771–1773.

Papadakis, M.N., Patrinos, G.P., 1999. Contribution of gene conversion in the evolution of the human beta-like globin gene family. Hum. Genet. 104, 117–125.

Patrinos, G.P., Giardine, B., Riemer, C., Miller, W., Chui, D.H.K., Anagnou, N.P., Wajcman, H., Hardison, R.C., 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res. 32, D537–D541.

Patrinos, G.P., Samperi, P., Lo Nigro, L., Kollia, P., Schiliro, G., Papadakis, M.N., 2005. Evidence for the molecular heterogeneity of sickle cell anemia chromosomes bearing the betaS/Benin haplotype. Am. J. Hematol. 80, 79–80.

Penman, B.S., Pybus, O.G., Weatherall, D.J., Gupta, S., 2009. Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean. Proc. Natl. Acad. Sci. U. S. A. 106, 21242–21246.

Penman, B.S., Habib, S., Kanchan, K., Gupta, S., 2011. Negative epistasis between α+ thalassaemia and sickle cell trait can explain interpopulation variation in South Asia. Evolution 65, 3625–3632.

Penman, B.S., Gupta, S., Buckee, C.O., 2012. The emergence and maintenance of sickle cell hotspots in the Mediterranean. Infect. Genet. Evol. 12, 1543–1550.

Pennings, P.S., Hermisson, J., 2006. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23, 1076–1084.

Piel, F.B., Patil, A.P., Howes, R.E., Nyangiri, O.A., Gething, P.W., Williams, T.N., Weatherall, D.J., Hay, S.I., 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat. Commun. 1, 104.

Piel, F.B., Howes, R.E., Patil, A.P., Nyangiri, O.A., Gething, P.W., Bhatt, S., Williams, T.N., Weatherall, D.J., Hay, S.I., 2013. The distribution of haemoglobin C and its prevalence in newborns in Africa. Sci. Rep. 3, 1671.

Ralph, P., Coop, G., 2010. Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 186, 647–668.
Roberts, D.J., Williams, T.N., 2003. Haemoglobinopathies and resistance to malaria. Redox Rep. 8, 304–310.
Rockett, K.A., Clarke, G.M., Fitzpatrick, K., Hubbart, C., Jeffreys, A.E., Rowlands, K., Craik, R., Jallow, M., Conway, D.J., Bojang, K.A., Pinder, M., Usen, S., Sisay-Joof, F., Sirugo, G., Tour, O., Taylor, T., Kwiatkowski, D.P., 2014. Reappraisal of known malaria resistance loci in a large multicenter study. Nat. Genet. 46, 1197–1204.
Rouabhi, F., Lapoumèroulie, C., Amselem, S., Krishnamoorthy, R., Adjrad, L., Girot, R., Chardin, P., Benabdji, M., Labie, D., Beldjord, C., 1988. DNA haplotype distribution in Algerian beta thalassaemia patients. An extended evaluation by family studies and representative molecular characterization. Hum. Genet. 79, 373–376.
Schilirò, G., Di Gregorio, F., Samperi, P., Mirabile, E., Liang, R., Cürük, M.A., Ye, Z., Huisman, T.H., 1995. Genetic heterogeneity of beta-thalassemia in southeast Sicily. Am. J. Hematol. 48, 5–11.
Schilirò, G., Mirabile, E., Testa, R., Russo-Mancuso, G., Dibenedetto, S.P., 1997. Presence of hemoglobinopathies in Sicily: a historic perspective. Am. J. Med. Genet. 69, 200–206.
Taylor, S.M., Parobek, C.M., Fairhurst, R.M., 2012. Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. Lancet Infect. Dis. 12, 457–468.
Trabuchet, G., Elion, J., Baudot, G., Pagnier, J., Bouhass, R., Nigon, V.M., Labie, D., Krishnamoorthy, R., 1991. Origin and spread of beta-globin gene mutations in India, Africa, and Mediterranea: analysis of the 5′ flanking and intragenic sequences of beta S and beta C genes. Hum. Biol. 63, 241–252.
Tsaras, G., Owusu-Ansah, A., Boateng, F.O., Amoateng-Adjepong, Y., 2009. Complications associated with sickle cell trait: a brief narrative review. Am. J. Med. 122, 507–512.
Wainscoat, J.S., Bell, J.I., Thein, S.L., Higgs, D.R., Sarjeant, G.R., Peto, T.E., Weatherall, D.J., 1983. Multiple origins of the sickle mutation: evidence from beta S globin gene cluster polymorphisms. Mol. Biol. Med. 1, 191–197.
Watts, D.J., 1999. Networks, dynamics, and the small-world phenomenon. Am. J. Sociol. 105, 493–527.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–442.
Weatherall, D.J., Clegg, J.B., 2001a. Inherited haemoglobin disorders: an increasing global health problem. Bull. World Health Organ. 79, 704–712.
Weatherall, D.J., Clegg, J.B., 2001b. The Thalassaemia Syndromes. 4th ed. Blackwell Science Ltd, Oxford.
Weatherall, D., Akinyanju, O., Fucharoen, S., Olivieri, N., Musgrove, P., 2006. Inherited disorders of hemoglobin. Disease Control Priorities in Developing Countries, pp. 663–680.
Williams, T.N., 2006. Human red blood cell polymorphisms and malaria. Curr. Opin. Microbiol. 9, 388–394.
Williams, T.N., Mwangi, T.W., Roberts, D.J., Alexander, N.D., Weatherall, D.J., Wambua, S., Kortok, M., Snow, R.W., Marsh, K., 2005a. An immune basis for malaria protection by the sickle cell trait. PLoS Med. 2, e128.
Williams, T.N., Mwangi, T.W., Wambua, S., Peto, T.E.A., Weatherall, D.J., Gupta, S., Recker, M., Penman, B.S., Uyoga, S., Macharia, A., Mwacharo, J.K., Snow, R.W., Marsh, K., 2005b. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. Nat. Genet. 37, 1253–1257. http://dx.doi.org/10.1038/ng1660.
Wilson, B.A., Petrov, D., Messer, P.W., 2014. Soft selective sweeps in complex demographic scenarios. Genetics 198, 669–684.
Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H., Tyler-Smith, C., 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. 19, 1453–1457.

Supplementary material: A detailed description of the model and its processes

As mentioned in *Methods*, our model follows a standard model of the behaviour of a variant under selection, and can be represented by the following equation:

$$G_{het_i(t)} = \frac{2.p_{i(t)}(1-\sum_{j=1}^{n} p_{j(t)}).w_{het_i}}{(1-\sum_{j=1}^{n} p_{j(t)})^2.w_{AA} + \sum_{k=1}^{n}(2.p_{k(t)}(1-\sum_{j=1}^{n} p_{j(t)}).w_{het_k})} \quad \textbf{(1)}$$

where $G_{het_i(t)}$ represents the frequency of $\beta^A\beta^X{}_i$ heterozygotes in the deme's offspring pool in generation *t*; $p_{i(t)}$ is the frequency of mutant haplotype $\beta^X{}_i$ in the deme in generation *t* following mutation and/or gene conversion; w_{het_i} is the relative fitness of genotype $\beta^A\beta^X{}_i$; w_{AA} is the relative fitness of genotype $\beta^A\beta^A$; and *n* is the total number of β^X haplotypes present in the deme. In scenarios where w_{het_i} is varied, the maximum range of possible heterozygote fitnesses ($w_{het_i(\max)} - w_{het_i(\min)}$) was *f*. Further details of the assignment of fitness values are provided below.

Sickle-cell anaemia and β^0-thalassaemia are both lethal if untreated (Weatherall et al., 2006). Given the absence of effective treatment in early human populations, we have assumed the fitness of any β^X homozygote to be equal to 0, therefore:

$$G_{AA(t)} = 1 - \sum_{j=1}^{n} G_{het_j(t)} \quad \textbf{(2)}$$

where $G_{AA(t)}$ represents the frequency of $\beta^A\beta^A$ homozygotes in the deme in generation *t*.

Population size and population growth

We tested different values of initial total population size between 25,000 and 125,000 in increments of 25,000, to reflect likely historic population sizes (Atkinson et al., 2009). Parameter *g*, the maximum possible population growth rate in any given simulation, was varied between 0% and 1%. This is based on reported estimates of historic human population growth (McEvedy and Jones, 1978).

Population subdivision and connection network

The number of demes *d* in the meta-population was fixed for each simulation. We tested values of *d* between 25 and 250 in increments of 25-50. As in Penman et al. (2012), we used the method outlined by Watts and Strogatz (1998) to generate population networks demarcating the migratory connections (edges) between the constituent demes (nodes) (Penman et al., 2012; Watts and Strogatz, 1998). Briefly, the level of connectivity within a meta-population is determined by the percentage *c* of demes that are connected to a randomly chosen partner deme. When $c = 0$, each deme is connected to its two nearest neighbors in a ring-shaped network, and the average shortest path length between any two demes is at a maximum. As *c* increases, randomly chosen connections are re-wired and the average shortest path length between any two demes decreases. We varied the value of *c* between 0% and 50%, since preliminary simulations showed that higher *c* values yielded similar results to $c = 50$.

Mutation and gene conversion

Different haplotypes in our model can only arise through mutation or gene conversion, not reciprocal recombination. They are therefore intended as proxies for

haplotypes whose occurrence cannot be accounted for by simple recombination, for example the major β^S haplotypes.

The meta-population is initially monomorphic for the ancestral β^A allele. In every generation, within each deme, the number of mutant β^X alleles entering the deme is drawn from a Binomial distribution, where the number of "trials" equals the number of wild-type chromosomes in that deme and the probability of success is determined by the per chromosome mutation rate μ. Four different allelic mutation rates are explored: (i) 10^{-8} events per chromosome per generation, (ii) 10^{-7} events per chromosome per generation, (iii) $5x10^{-7}$ events per chromosome per generation, and (iv) 10^{-6} events per chromosome per generation. Only results from the latter three are presented, as the lowest mutation rate yielded very few instances of the β^S- or β^0-thalassaemia-like patterns.

If a β^X-associated chromosome is present in a deme, the number of gene conversion events involving the transfer the β^X_i mutation onto haplotypic background $i+1$ is again drawn from a Binomial distribution. In this instance, the number of "trials" is the number of $\beta^A\beta^X$ heterozygotes in the deme and the probability of success is the per chromosome gene conversion rate r. Crucially, the number of "trials" equalling the number of $\beta^A\beta^X$ heterozygotes captures the fact that gene conversion can only occur within heterozygotes, making the deme-level gene conversion rate markedly lower than it would be if gene conversion occurred independently to an individual's genotype. Given the lack of evidence for the per chromosome gene conversion rate at the β-globin locus, we varied this value between 0 events per chromosome per generation and the extreme value of $5x10^{-5}$ events per chromosome per generation.

Each β^{X}_{i} allele is labeled according to its associated haplotypic background, with any mutation arising *de novo* or undergoing gene conversion being assigned a new haplotypic background. Although a simplification, this allowed individual mutant alleles to be tracked separately within the meta-population.

Gene flow

During every iteration of each simulation, following reproduction and the ascertainment of successful offspring in each deme, *d* demes were chosen to undergo bi-directional gene flow with a randomly chosen partner deme to which they were directly linked in the migration network. These *d* demes were sampled with replacement from all of the demes present in the meta-population, thus a gene flow event may not occur in every deme in every generation or may, conversely, occur more than once per generation for some demes. Every gene flow event between two demes involves the exchange of a randomly chosen *m*% of the individuals in each of the partner demes. The value of *m* was varied between 0.5% and 8% in increments of 1.5-2%.

Malaria selection and genotype fitness

Genotypes $\beta^{X}\beta^{X}$ were always assigned a fitness of 0, reflecting their historical lethality. As in Penman et al. (2009, 2012), we calculate genotype fitness values for $\beta^{A}\beta^{A}$ and $\beta^{A}\beta^{X}$ based on presumed mortality rates that incorporate mortality due to *falciparum* malaria. Each genotype was assigned a mortality rate *of* $q + \gamma\varphi$, where q is the baseline mortality rate of the $\beta^{A}\beta^{A}$ and $\beta^{A}\beta^{X}$ genotypes in the absence of malaria, γ is the excess mortality due to malaria and φ is the relative risk of death from malaria for that genotype. The fitnesses of the genotypes are then calculated as the ratio of the

average survival time of each genotype (the reciprocal of the total mortality rate of the genotype) to the maximum total survival time of any genotype in the population. A value of q of 0.04 years^{-1} is assigned to both $\beta^A\beta^A$ and $\beta^A\beta^X$ to reflect the absence of any pathology in $\beta^A\beta^X$ heterozygotes. The values of γ and φ are: $\gamma = 0.008$ years^{-1}, $\varphi = 1$ for $\beta^A\beta^A$ and $\varphi = 0.06$ for $\beta^A\beta^X$.

When inter-haplotype fitness variation was absent, any $\beta^A\beta^X{}_i$ heterozygote was assigned the maximum possible fitness in the population (w_i always = 1). However, as mentioned previously, we also sought to address the possibility of variation in fitness between the heterozygotes carrying different β^S or β^0-thalassamia haplotypes (Ashley-Koch et al., 2000; Tsaras et al., 2009). When this was allowed to occur, there was an equal probability that a specific heterozyote should be assigned a fitness of 1 ($w_i = 1$) or that the fitness of a specific heterozygote should be drawn from a uniform distribution between $1 - f/2$ and $1 + f/2$ ($1 - f/2 < w_i < 1 + f/2$). Thus, the majority of haplotypes had a heterozygote fitness of 1 and the total possible range of heterozygote fitnesses was f. Due to uncertainty in the degree of fitness variation for β^S and β^0-thalassaemia, we varied f between 0 and 0.3 to allow for a wide spectrum of evolutionary scenarios.

References for Supplementary Text

Ashley-Koch, A., Yang, Q., Olney, R.S., 2000. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am. J. Epidemiol. 151, 839–45.

Atkinson, Q.D., Gray, R.D., Drummond, A.J., 2009. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. Proc. Biol. Sci. 276, 367–73. doi:10.1098/rspb.2008.0785

Ellegren, H., Smith, N.G., Webster, M.T., 2003. Mutation rate variation in the mammalian genome. Curr. Opin. Genet. Dev. 13, 562–568.

McEvedy, C., Jones, R., 1978. Atlas of world population history. Penguin.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the Mutation Rate per Nucleotide in Humans. Genetics 156, 297–304.

Penman, B.S., Gupta, S., Buckee, C.O., 2012. The emergence and maintenance of sickle cell hotspots in the Mediterranean. Infect. Genet. Evol. 12, 1543–50.

Penman, B.S., Pybus, O.G., Weatherall, D.J., Gupta, S., 2009. Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean. Proc. Natl. Acad. Sci. U. S. A. 106, 21242–6.

Tsaras, G., Owusu-Ansah, A., Boateng, F.O., Amoateng-Adjepong, Y., 2009. Complications associated with sickle cell trait: a brief narrative review. Am. J. Med. 122, 507–12. doi:10.1016/j.amjmed.2008.12.020

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–2.

Weatherall, D., Akinyanju, O., Fucharoen, S., Olivieri, N., Musgrove, P., 2006. Inherited Disorders of Hemoglobin.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H., Tyler-Smith, C., 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. 19, 1453–7.

5. General Discussion

Haemoglobinopathies are an emerging global health problem and some of the best studied examples of human evolution.[12,68,122,123] However, many aspects of their epidemiology and population genetics remain ambiguous. In this thesis, I have used statistical, geostatistical and mathematical modelling tools to explore, describe, predict and better understand the epidemiology and population genetic patterns of some of the major haemoglobinopathies around the world. By working at the interface between public health and evolutionary biology, the research presented contributes to a more holistic and multidisciplinary approach to the study of this important group of disorders.

In Chapters 2 and 3, I apply systematic review methods compile two geodatabases of prevalence and/or genetic diversity surveys for α-thalassaemia and β^S in Southeast Asia and India, respectively. In Chapter 2, I present the first spatially referenced maps of α-thalassaemia in Southeast Asia and provide a thorough overview of current knowledge of the distributions and genetic diversity patterns of its major forms (α^0-, α^+- and α^{ND}-thalassaemia) there. In addition, I generate the first model-based continuous maps of α-thalassaemia in Thailand, where the majority of the epidemiological data came from, as well as refined estimates of the number of newborns born with a severe form of the disorder in 2020. In Chapter 3, I take advantage of recent surveying efforts to more than triple the amount of data in the evidence-base used for mapping SCA in India. Using this database, I provide an updated map of the distribution and burden of this disorder in the highly populated country, whilst taking into account information on the social structure of the population. Both chapters, which involved local collaborators in Thailand and India, respectively, include valuable novel quantitative data and tools to inform clinicians

and public health policy makers on public health policies for the prevention and management of haemoglobinopathies in Southeast Asia and India.

In both Chapters 2 and 3, I employed, Geographic Information System (GIS) tools to generate detailed maps of the surveys as well as summary maps of predicted allele frequencies and sub-national estimates of the number of affected newborns. Two different approaches to the spatial component of this work were used: a Bayesian framework that uses a Stochastic Partial Differential Equation (SPDE) approach[124] with Integrated Nested Laplace Approximation (INLA)[125,126] (Chapter 2); and a non-Bayesian Generalised Additive Model approach (Chapter 3). There are important differences between these methods, which are discussed below.

In Chapter 4, I cross over to the population genetic enquiry of why severe forms of β-thalassaemia and SCD display contrasting spatial haplotypic patterns, despite sharing a common malaria selection pressure and both causing a severe blood disorder in the homozygous state. In addition, I explore one of the oldest, but still controversial, questions of the origin of β^S – that is whether the occurrence of the mutation on five classical haplotypes is indicative of it having arisen on five separate occasions. For this work, I developed a meta-population model, implemented in Matlab 2012b, which simulated the spread of malaria-protective globin mutations on distinct haplotypes. By varying the input values of a range of demographic and genetic factors, I delineated the conditions under which: (i) the contrasting population-level spatial haplotype patterns were likely to arise, and (ii) gene conversion could have contributed to the final haplotypic variation in a subpopulation.

In this General Discussion, I first provide a comparison and discussion of the two epidemiological studies conducted in Chapters 2 and 3, which shared methodological similarities (e.g. the systematic construction of a geodatabase and predictive modelling) but focused on different disorders (α-thalassaemia and SCA) in different geographical areas (Southeast Asia and India). Following this, I share my views on the benefits of bridging epidemiological studies, which focus on disease burden for public health purposes, with theoretical population genetic modelling methods for improving our understanding of the evolutionary interactions between different genetic disorders and their public health consequences.

5.1. Construction of the geodatabases for α-thalassaemia and SCA prevalence: similarities and differences

Whilst there is an obvious overlap in the methodological steps required to generate the two geodatabases for α-thalassaemia in Southeast Asia and sickle haemoglobin in India, differences in the biology, genetics, population genetics and diagnosis of these two disorders required the tailoring of survey inclusion criteria and data abstraction strategies.

For both studies, avoidance of any bias in the database was key and as such survey inclusion criteria were overlapping. For instance, only those surveys that were carried out in representative population groups, such as the community, newborns and pregnant women, were included, provided that there was no sampling bias. In addition, sufficient information regarding sample size and either the number of individuals identified with a particular genotype or the allele frequency observed in the sample was required. Where possible, data from local sources were inputted into the database, helped along by collaborations with clinicians and public health researchers in Bangkok and Mumbai.

This inclusion of locally obtained data was invaluable to the work, particularly in Thailand. However, their inclusion did not come without caveats and care had to be taken to ensure reliability of the data as well as traceability of the data sources. Finally, data abstraction was done according to pre-defined protocols; for β^S, the same protocol as that used by Piel *et al.* (2013) was used, whilst for α-thalassaemia a new protocol was developed.[64]

There are several key differences between the two databases. For instance, for the α-thalassaemia database, the minimum administrative level at which included surveys could be reported was the national level. This was because one of the main aims of the study was to describe current knowledge regarding α-thalassaemia allele frequency in Southeast Asia, along with an overview of surveying effort there. Nevertheless, all surveys that were only reported at the national level were excluded from the geostatistical component of this work. By contrast, for India, only surveys that could be georeferenced to at least the district level were included; this was important given the large size of India (six times that of Thailand) and that the primary aim of the study was to generate an updated map of β^S allele frequency and refined newborn estimates in the country.

One of the main challenges of constructing the geodatabase for α-thalassaemia and generating reliable and informative maps of the data contained within it, was handling the complexity of α-thalassaemia genotypes in Southeast Asia. Indeed, given the range of possible α-thalassaemia genotypes, their reporting was sometimes more ambiguous than for β^S, and concerted effort had to be made to accurately record whether an α-thalassaemia genotype: (i) had been tested for and was absent, (ii) was not tested for, or (iii) was tested for but not included in the results due to the focus of the article being on a specific

genotypic outcome. This was not an issue for β^S. Incomplete reporting of genotypes was also the reason why an overall map of α-thalassaemia was not generated and why separate maps of the major α-thalassaemia forms were generated using subsets of the database.

A third key difference between the two studies is the explicit inclusion of ethnicity in the sickle-cell database and social grouping in the subsequent statistical analysis. India has a unique population structure that is composed of a large number (several thousands) of highly endogamous groups, partly as a result of its social caste system.[127] It has long been recognised that the highest allele frequencies of β^S occur in tribal groups and populations belonging to the scheduled castes, a designation used to describe the most socioeconomically disadvantaged classes. However, due to their global focus, previous maps of β^S allele frequency do not capture this. As such, a main aim of this study was to incorporate social grouping into the final predicted β^S allele frequency map. By contrast, for α-thalassaemia in Southeast Asia, there has been less of a discourse surrounding the allele frequency of α-thalassaemia in different ethnic groups – that is, until very recently[69] – and ethnicity was not a major consideration in our study. Whilst ethnicity was reported in the database, it was only used as an indicator of whether the study sample was representative of the general population in which the study took place.

For the generation of continuous α-thalassaemia allele frequency maps in Thailand, I modelled data from Thailand and its neighbouring countries (Myanmar, Lao PDR, Cambodia, Vietnam and Malaysia), whilst for India, I restricted my analysis to surveys carried out within the country. National borders are often irrelevant demarcations that bear no, or very minimal, influence on the distribution of the disease. Nevertheless, this border effect is largely driven by the geography of individual countries. India's borders

are mostly made of coastlines and the Himalaya represent an important natural border towards the North of the country. Furthermore, only one survey on β^S allele frequency was identified in Pakistan and two in Nepal in Piel *et al.* (2013).[64] The geography of Thailand is very different, with substantial (and passable) borders with Myanmar, Lao PDR, Cambodia and, to a lesser extent, Malaysia. Thus, surveys from countries neighbouring Thailand were deemed important contributions to the model, in particular with regards to predictions made along the Thailand border.

Finally, different inclusion criteria were applied in relation to the diagnosis of the two disorders. Accurate diagnosis of specific α-thalassaemia genotypes or alleles can only be done at the DNA level, for example through PCR analysis or DNA sequencing[49,128] Other diagnostic methods, including complete blood count, histological analysis and haemoglobin typing by high performance liquid chromatography (HPLC) or cellulose acetate electrophoresis, do not reliably detect all α-thalassaemia genotypes and are unable to distinguish between heterozygotes and homozygotes. In addition, their results can be altered by compound heterozygosity for another haemoglobinopathy. As a result, we only included data on α-thalassaemia variants for which molecular methods had been used to confirm their presence. By contrast, the multitude of tests that are available for β^S all have similar reliability and so no restriction on diagnostic method was made.

5.2. A comparison of the different statistical modelling methods used

Cartographic representations of the spatial patterns of disease prevalence, allele frequencies of the causal mutations and other aspects of a disease's epidemiology and/or biology are an invaluable public health tool. They can be purely descriptive (Chapter 2), include an analytical component (Chapter 3) or be inferential, whereby predictions for

unsampled locations are made (Chapters 2 & 3). In all cases, they provide a powerful visual tool for identifying both differences and similarities in population health.

Geostatistics is a sub-branch of spatial statistics that is intended for analysing and predicting the values associated with spatially continuous phenomena by explicitly modelling spatial autocorrelation (i.e. dependence) in the data. An overarching question in geostatistical analysis is: "How can measured responses Y_i at i locations in a continuous study region be used to predict the underlying response surface S(x) for that region?"[129] Predictions in unsampled areas can also be made using non-geostatistical models in which a spatial component is incorporated through the inclusion of latitude and longitude as model covariates. Spatial autocorrelation is not an inherent feature of non-geostatistical models.

In this thesis, I use two different approaches to the geospatial analysis of globin mutations in Thailand and India: (i) a Bayesian geostatistical framework that uses a Stochastic Partial Differential Equation (SPDE) approach,[124] with Integrated Nested Laplace Approximation (INLA) to model the observed allele frequency of α-thalassaemia mutations,[125] and (ii) a non-Bayesian non-geostatistical Generalised Additive Model (GAM) to model β^S allele frequency in India. Both methods are available as off-the-shelf packages in R that are easy to implement and reproducible. In both methods, a model is formulated, fitted to the data and subsequently used to make predictions of allele frequencies at unsampled locations. Uncertainty in the predictions are also quantified. However, the treatment of model parameters, how the data were fitted, the model predictive process and uncertainty measurement calculations differ between these two approaches.

Bayesian geostatistical models are one of the most widely used methods in modern disease mapping, and formed the basis of previous efforts to map the distribution of haemoglobinopathy-causing mutations around the world, namely β^S and β^C.[16,44,64] These models have been fitted through Markov Chain Monte Carlo (MCMC) sampling, which is computationally intensive, requiring hours, or even days, to run. Recently, Integrated Nested Laplace Approximation has been developed to overcome some of the challenges of MCMC, and is the approach I used in my analysis of α-thalassaemia in Southeast Asia. It has the benefit of being considerably more computationally efficient to MCMC, taking just a few seconds to perform.[130] Furthermore, studies have shown that INLA and MCMC performed similarly in terms of parameter estimation, unless the disease under study is rare.

However, whilst INLA provides a fast and efficient alternative to MCMC, its use with multivariate data is limited. This is because the method focuses on the posterior marginals and not on the joint posterior distribution, making multivariate posterior inferences on the parameters of the model difficult.[131] In Chapter 3, my aim was to not only model the spatial structure of the data but also to test the relationships between β^S and a range of possible predictor variables. As a result, for this project, I selected a Generalised Additive Modelling approach, primarily for the reason that the results of the GAM are easily interpretable in terms of understanding the relationship between the response variable and the covariates. Nevertheless, this was not without a trade-off, as this method took longer to run, which required me to reduce the spatial resolution of the predicted map. Nevertheless, given the complexity of the β^S distribution in India and the other caveats with the dataset, this was deemed to be of less significant importance.

The Bayesian and frequentist approaches to statistics are the topic of intense debate, and their use in geostatistics is no exception. A key advantage of the Bayesian approach is that priors allow information obtained through previous studies or from expert opinion to be formally included in the model process. However, a counter-argument is that the specification of these priors can sometimes be subjective. The relative strengths and weaknesses of these two schools of thought are not the primary concern of this thesis and methods from both have been employed in two different contexts. Nevertheless, the approach used has important implications for the correct interpretation of the models' outputs and were kept in mind throughout.

General limitations to geostatistical modelling and geospatial analysis

Limitations of the geostatistical and spatial models used in this thesis are discussed in the respective chapters. In general, limitations stem from weaknesses in the underlying database of surveys, be it in relation to their coverage, short-range variability or inclusion of confounding factors (e.g. non-random mating or interactions with other haemoglobinopathies). This is primarily the result of the retrospective nature of the analyses, which use data from a range of different sources, each with divergent goals. In addition, there may be uncertainties that are introduced by errors in the georeferencing of surveys. Only some of these limitations can be accounted for in the uncertainty estimates. Finally, in the case of the geostatistical INLA technique, the model is based on the assumption that the spatial correlation structure does not vary across the study area. However, it is possible that certain populations may be more connected than others.

5.3. The use of population genetic models in the study of haemoglobinopathy evolution and epidemiology

The haemoglobinopathies provide a useful case study in human evolution and genetics.[132] They were the first adaptive traits to be uncovered in humans[133] and still today represent some of the best examples of selection in humans, with their high prevalences in historically malarious regions areas of the world being testament to the profound impact that infectious disease can have on the human genome. Moreover, the occurrence of β^S on five major haplotypes has been cited as a potential example of a soft selective sweep in humans, whilst haemoglobinopathies in general provide an excellent example of the role that epistatic interactions can have in determining the distributions of genetic mutations.[109] Thus, the population genetics of haemoglobinopathies are fascinatingly complex and have long been discussed in relation to their broad distributions and overall prevalences around the world; however, the explicit consideration of the population genetic patterns of haemoglobinopathies and, crucially, how they arose and are maintained, has typically been overlooked in efforts to quantify the burden of haemoglobinopathies in different parts of the world.[37,64]

First, I further discuss my findings in Chapter 4 and how they provide insight into the evolution of β^S and β^0-thalassaemia and human evolution in general. Following this, I comment on the role that I believe population genetic models, such as that employed in Chapter 4, could play in generating refined estimates of the current and future burden of different haemoglobinopathies, singly and in combination.

Evolutionary insights gained from population genetic simulation models of β^S and β^0-thalassaemia

Population genetic simulation models are a useful tool for understanding the varied demographic (e.g. migration), genetic (e.g. mutation rate) and evolutionary (e.g. selection) processes that drive diverse haemoglobinopathy patterns around the world, both at a cross-continental and national scale.[109,118,134] In particular, such models allow us to study the multidimensional interactions of these processes as well as explore a large parameter space covering a range of possible combinations of values for all the processes under study, many for which data would be difficult to ascertain. In this thesis, I have used a stochastic meta-population genetic model to identify the demographic and genetic scenarios under which the differing spatial haplotype patterns of β^S and β^0-thalassaemia are likely to occur. I explored the evolutionary consequences of complex processes whose interactions would be difficult to predict analytically.

The work in Chapter 4 expands on previous research into soft selective sweeps (described in Chapter 4) by developing a unified population genetic model that incorporates a multitude of genetic and demographic processes and conditions within a single framework. These include mutation, population size and growth, population subdivision, network structure and inter-deme gene flow. In addition, the model includes two factors not previously explored – that is, gene conversion and inter-haplotype fitness variation. Inter-haplotype differences in clinical severity have been shown for β^S.[135] This has important implications for our understanding of the origin of β^S in humans, as our model shows that the contribution of gene conversion to the haplotypic diversity of β^S is only possible if inter-haplotype fitness variation exists. However, a second feature revealed by the model that would be difficult to detect through empirical research is the finding that

observed inter-haplotype fitness variation of β-globin variants may not reflect the full fitness range of all haplotypes that have arisen over the course of human evolutionary history. As such, I argue that we cannot assume that the presence of five classical β^S haplotypes necessarily means the mutation has arisen five times. Further exploration of the possible fitness differences between sickle-cell mutations on different haplotypic backgrounds will shed additional light on this.

If we are ever to invoke a single origin for β^S, then we must also consider where it first arose. Most have assumed an African origin, but Livingstone noted several features of the distributions of the major β^S haplotypes in Africa that are incompatible with this view. Specifically, the Benin, Cameroon and Senegal haplotypes all have surprisingly disjointed distributions, whereby adjacent populations, whose languages are mutually intelligible, differ starkly in terms of the amount of β^S that they possess (see Livingstone 1989a for further details).[136] By contrast, the distribution of β^S in Africa closely parallels the penetration of Arab influence into the continent, with only those populations that have had extensive contact with Arabic groups exhibiting high allele frequencies of β^S (Livingstone 1989a,b).[136,137] From these observations, Livingstone postulated a single origin of β^S in the Middle East, possibly on the Arab-Indian haplotype, and the diffusion of the gene throughout Africa as a result of Arab expansion. Future modelling work that accounts for disparities in the migration routes of a large, geographically distant population into isolated subpopulations within a network may serve to substantiate this theory. In addition, if the hypothesis of a single origin of β^S in the Middle East is true, this raises the question of how the mutation penetrated the Indian population, where its distribution is highly variable (as shown in Chapter 3 and previous mapping studies).[64,71,75] Building on previous modelling work into the role that negative epistasis

between α-thalassaemia and β^S and malaria selection have had in influencing the distribution of β^S in India,[118] it would be interesting to develop a model that also incorporates long- and short-distance migration patterns. In this way, it may be possible to simulate the early dispersal patterns of the mutation across India, thereby shedding light on its origin here. The maps generated in Chapter 4 will serve as a useful resource with which to compare the model outputs.

This modelling work also highlighted potential aspects of the evolution of the β-globin polymorphisms that are not immediately clear from genetic or haplotypic analysis alone. For instance, allelic exclusion, whereby an incoming allele is unable to become established in a population if a preceding allele has already reached equilibrium allele frequencies, was a consistent observation in our simulations and one which provides a mechanistic explanation for the effect of various processes on the observed haplotype patterns. This notion of allele exclusion is not a new one and the phenomenon has been a key assumption in theoretical treatments of soft selective sweeps (e.g. Ralph and Coop, 2010), described in Chapter 4.[138] However, the fact that it was not imposed as an *a priori* assumption on the population genetic model used in this thesis meant that I was able to explore in more detail the kind of soft selective sweep that arises from β^0-thalassaemia, whereby the pattern of a soft selective sweep is present both at a meta-population scale and within sub-populations. Given the growing evidence that adaptation via a soft selective sweep might be common across a range of species, from viruses to higher order mammals (reviewed in Messer and Petrov, 2013),[24] a better understanding of the processes that give rise to sub-types of soft selective sweeps is needed. It is possible that lessons learnt from haemoglobinopathies might be extended to other systems. Models

will be an efficient tool for generating hypotheses, which can feed into the design of genetic studies of selective sweeps in the population.

The role of population genetic models in epidemiological studies

In Chapters 2 and 3, a large amount of new epidemiological data were available compared to previous efforts to describe and/or map the distribution of α-thalassaemia and β^S. These additional data greatly expand our understanding of the epidemiology of these disorders in the two regions studied and its utility should not be understated. However, due to the complexity of the population structure in India and the diversity of α-thalassaemia variants in Southeast Asia, the amount of data available remains relatively limited to fully map the heterogeneities of the disorders. Population genetic models offer a way to understand such heterogeneities and to contextualise the available data.

For instance, it would be interesting to use population genetic models to more fully explore the influence of consanguinity and/or endogamy in India in determining the SCA health burden there. In Chapter 3, newborn estimates were calculated assuming Hardy-Weinberg proportions. At present, no fine-scale data on the pattern of consanguinity in Indian sub-populations exist and as such it is impossible to empirically assess. Population genetic models that directly simulate population structure, endogamy and consanguinity would enable us to assess how much of an impact these factors are likely to have on allele frequency estimates, and help us refine the confidence limits of our estimates to account for maximal and minimal likely levels of consanguinity, even in the absence of complete information.

As mentioned in the General Introduction, several modelling studies have already shown that epistatic interactions are likely to have been critical factors in determining the spatial patterns of haemoglobinopathies, along with malaria selection pressure, migration and other demographic factors.[109,118,134] Indeed, it has been postulated that the variable distribution of β^S in India could, in part, be due to the negative epistatic interactions between the β-globin mutation and α-thalassaemia. In addition, while no modelling studies have examined epistasis between α- and β-thalassaemia in Southeast Asia specifically, the findings of Penman *et al.* (2009) for the Mediterranean,[109] where α- and β-thalassaemia are also present at high prevalences, could reasonably be applied to this region. Penman *et al.* (2009) reported that the maintenance of the severe α^0-thalassaemia form in the Mediterranean could be facilitated by its co-occurrence with β^+-thalassemia, provided that the co-inheritance of deletional HbH disease (i.e. α^+-thalassaemia with α^0-thalassaemia) with β^+-thalassaemia resulted in a milder HbH disease phenotype with no effect on malaria protection. It is possible that the same epistatic interactions have influenced the distribution of the thalassaemias in Southeast Asia. However, the influence of the added β^E mutation in Southeast Asia within this system remains unclear. The empirical study of the population-level outcomes of these interactions is difficult. Mathematical models based on population genetics offer an efficient alternative to examine and assess the possible influence of these interactions and their impact on the public health burden of these disorders.

5.4. Future directions

There is vast scope for further research into the epidemiology and population genetics of haemoglobinopathies. An obvious next step would be to replicate our UK-Thailand collaboration with researchers in other Southeast Asian countries in order to expand the

evidence-base, gain a more complete picture of α-thalassaemia and generate model-based continuous maps that span the whole region. A similar study in the Mediterranean, where α^0-thalassaemia is also present, would also be of value. The ultimate aim will be to generate a global map of the distribution of α-thalassaemia, which can more reliably inform public health policies surrounding the diagnosis, prevention and management of the disorder. Even in countries where α-thalassaemia itself is not a public health burden due to severe forms being absent, knowing its distribution will be useful for assessing the severity of the burden of β-globin disorders in overlapping areas. Furthermore, it will provide a baseline for tracking spatio-temporal changes in the distribution of α-thalassaemia as a result of global migration and population admixture. Population genetic models could be an important component of this.

However, the generation of a global map of α-thalassaemia will be no easy feat. Unlike structural haemoglobin variants that have been previously mapped,[44,64] α-thalassaemia is highly genetically diverse, and its major forms display striking patterns of geographic differentiation. Both inter-study and inter-region variation will therefore need to be accounted for. Careful database design will be needed to ensure that reported genotype and/or allele frequencies are reliably captured everywhere. In addition, the development of more sophisticated geostatistical and spatial models that can simultaneously apply varied data inputs to output predictions will extend our current approach of mapping α-thalassaemia forms in isolation of each other using subsets of the available data. The widespread use of molecular diagnostic methods such as gap-PCR will be necessary to reliably identify individual mutations in many more populations.

My research in India has demonstrated the rapid expansion of available data on β^S allele frequency in the country since 2013 and the utility of their inclusion in updated maps. However, I also found that the extensive short-range variability of the allele frequency of β^S observed in some areas in India increased prediction uncertainty in those areas. In future, it may be necessary to develop local models that account for the observed fine-scale heterogeneity in β^S allele frequency. In addition, taking into account the timing of β^S surveys in India will help researchers to keep abreast of its changing burden in the highly population country.[139] Ideally, a standardised platform with which to collect spatial and temporal data would be developed but this is difficult to implement as it would require not only the cooperation of local researchers but also regular quality control to ensure that biased estimates are not being entered.

An important aspect of the disease burden of SCA that is not captured in the presented maps is the wide variation in the clinical severity of the disorder.[71,75,140] Recent evidence suggests that the disorder is more clinically severe than previously considered,[141] but that there is profound variability in its clinical profile, including age of onset of symptoms, number of vaso-inclusive crises per year, varying involvement of different organs and transfusion needs.[140] This clearly has important implications for the care and management of SCA patients, particularly given the chronic nature of the disorder. In turn, the demand on clinical and public health resources will also vary. Putative factors contributing to the high variability include the level of foetal haemoglobin, presence or absence of the Xmn1 polymorphism and co-inheritance of α-thalassaemia, although this is not an exhaustive list.[142] It is also likely that the severity of the disease is affected by environmental factors.[143,144] Given the vast size of India and its population, the development of locally appropriate models of care, and assessment of the necessary resources, is needed.

Consideration of both clinical severity patterns in the population and the general frequency distribution of β^S in the country will be important.[145] Personal communication with local collaborators has revealed that, as part of ongoing sickle cell screening programmes in India, carefully planned follow-up of SCA cases has been conducted. It would be interesting to use individual-level clinical and geographical data from such programmes to compare and contrast the clinical outcomes of SCA in individuals hailing from different parts of the country. Such a spatial point analysis of Indian SCA patients from different areas of the country will help to: (i) identify whether there is a spatial trend in the pattern of clinical severity, and (ii) hypothesise on the role of local environmental factors, and possibly genetic ones, in influencing clinical outcomes.

Amongst this complexity, one thing that is certain is that co-inheritance of more than one haemoglobinopathy can significantly alter the clinical outcome of the blood disorder.[111,114,146] As clinical understanding of these epistatic interactions improves, combined maps of the respective distributions of different haemoglobinopathies will become increasingly important. For instance, combined maps of α-thalassaemia and β-globin disorders may serve as a tool to predict areas with not only the highest but also most severe disease burden. Similar methods used to map α-thalassaemia in Southeast Asia in this thesis should also apply to β-thalassaemia. Such tools will also facilitate estimation of the burden of compound disorders, such as β^E/β-thalassaemia in Southeast Asia and HbSE disease in India.

Maps of the distributions of different haemoglobinopathies will also be useful from a population genetic perspective. Previous modelling studies have demonstrated that different haemoglobinopathy profiles in different parts of the world likely reflect specific

combinations that work well together, rather than chance assemblies. This notion may explain the patterns that are seen elsewhere in the world, for instance the distribution of β^E in Southeast Asia. The generation of maps of different haemoglobinopathies will make it easier to test these ideas.

6. Conclusions

The epidemiology and population genetics of haemoglobinopathies are fascinatingly complex, with observed patterns likely being the result of a multitude of factors, ranging from genetic (e.g. epistasis) to demographic (e.g. migration) to environmental (malaria selection). Understanding their population genetics and evolutionary history should feed back into the design of public health policies for their control. Indeed, the more we understand the haemoglobinopathies, their distributions, evolution and interactions with each other and the environment, the better placed we may be to predict where particular combinations of them are likely to occur and their public health consequences, both currently and in the future. Epidemiological data will be integral to building a complete picture of the distributions of these disorders, not only to guide public health policies in the short-term, but also provide data upon which population genetic models can be constructed and used to make predictions long into the future. The study of haemoglobinopathies may also offer insights into the evolutionary and population biology of other genetic diseases and the mechanisms that underlie the phenotypic diversity of monogenic disease. In addition, their inherent link to malaria protection in humans may be informative in understanding malaria pathogenesis.

References

1. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. *Bulletin of the World Health Organization* 2001; **79**(8): 704-12.
2. Weatherall DJ. The challenge of haemoglobinopathies in resource-poor countries. *British journal of haematology* 2011; **154**(6): 736-44.
3. Christianson A, Howson CP, Modell B. March of Dimes Global Report on Birth Defects. New York: Mach of Dimes Birth Defects Foundation, 2006.
4. Piel FB, Hay SI, Gupta S, Weatherall DJ, Williams TN. Global Burden of Sickle Cell Anaemia in Children under Five, 2010–2050: Modelling Based on Demographics, Excess Mortality, and Interventions. *PLOS Medicine* 2013; **10**(7): e1001484.
5. Weatherall D, Hofman K, Rodgers G, Ruffin J, Hrynkow S. A case for developing North-South partnerships for research in sickle cell disease. *Blood* 2005; **105**(3): 921-3.
6. Colah R, Gorakshakar A, Phanasgaonkar S, et al. Epidemiology of beta-thalassaemia in Western India: mapping the frequencies and mutations in sub-regions of Maharashtra and Gujarat. *British journal of haematology* 2010; **149**(5): 739-47.
7. Martinson JJ, Excoffier L, Swinburn C, et al. High diversity of alpha-globin haplotypes in a Senegalese population, including many previously unreported variants. *American Journal of Human Genetics* 1995; **57**(5): 1186-98.
8. O'Riordan S, Hien TT, Miles K, et al. Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. 2010; (1365-2141 (Electronic)).
9. O'Shaughnessy DF, Hill AVS, Bowden DK, Weatherall DJ, Clegg JB. Globin genes in Micronesia: origins and affinities of Pacific Island peoples. 1990; (0002-9297 (Print)).
10. Premawardhena A, Allen A, Piel F, et al. The evolutionary and clinical implications of the uneven distribution of the frequency of the inherited haemoglobin variants over short geographical distances. *British journal of haematology* 2017; **176**(3): 475-84.
11. Schiliro G, Gregorio FD, Romeo MA, Testa R, Russo A. Incidence of Hemoglobin S Carriers in Sicily. *Hemoglobin* 1986; **10**(1): 95-9.

12. Hedrick PW. Population genetics of malaria resistance in humans. *Heredity* 2011; **107**(4): 283-304.
13. Allison AC. POLYMORPHISM AND NATURAL SELECTION IN HUMAN POPULATIONS. *Cold Spring Harbor symposia on quantitative biology* 1964; **29**: 137-49.
14. Hill AVS, Allsopp CEM, Kwiatkowski D, et al. Common West African HLA antigens are associated with protection from severe malaria. *Nature* 1991; **352**(6336): 595-600.
15. Malaria Genomic Epidemiology N, Rockett KA, Clarke GM, et al. Reappraisal of known malaria resistance loci in a large multi-centre study. *Nature genetics* 2014; **46**(11): 1197-204.
16. Piel FB, Patil AP, Howes RE, et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications* 2010; **1**: 104-.
17. Williams TN, Mwangi TW, Roberts DJ, et al. An Immune Basis for Malaria Protection by the Sickle Cell Trait. *PLOS Medicine* 2005; **2**(5): e128.
18. Haldane JBS. THE RATE OF MUTATION OF HUMAN GENES. *Hereditas* 1949; **35**(S1): 267-73.
19. May J, Evans JA, Timmann C, et al. Hemoglobin variants and disease manifestations in severe falciparum malaria. *Jama* 2007; **297**(20): 2220-6.
20. Wambua S, Mwangi TW, Kortok M, et al. The effect of alpha+-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya. *PLoS Med* 2006; **3**(5): e158.
21. Williams TN, Mwangi TW, Wambua S, et al. Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *The Journal of infectious diseases* 2005; **192**(1): 178-86.
22. Williams TN, Wambua S, Uyoga S, et al. Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood* 2005; **106**(1): 368-71.
23. Flint J, Harding RM, Boyce AJ, Clegg JB. The population genetics of the haemoglobinopathies. *Bailliere's clinical haematology* 1998; **11**(1): 1-51.
24. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution* 2013; **28**(11): 659-69.

25. Schechter AN. Hemoglobin research and the origins of molecular medicine. *Blood* 2008; **112**(10): 3927-38.
26. Gleason CA, Juul SE. Avery's Diseases of the Newborn E-Book: Elsevier Health Sciences; 2011.
27. Rochette J, Craig JE, Thein SL. Fetal hemoglobin levels in adults. *Blood Rev* 1994; **8**(4): 213-24.
28. Bunn HF, Forget BG. Hemoglobin: Molecular, genetic and clinical aspects. Philadelphia, PA: W.B. Saunders; 1986.
29. Bard H. Haemoglobin synthesis and metabolism during the neonatal period. In: Christensen RD, ed. Hematologic Problems of the Neonate. Philadelphia: WB Saunders; 2000: 365-88.
30. Forget BG, Bunn HF. Classification of the Disorders of Hemoglobin. *Cold Spring Harbor Perspectives in Medicine* 2013; **3**(2): a011684.
31. Kohne E. Hemoglobinopathies: clinical manifestations, diagnosis, and treatment. *Deutsches Arzteblatt international* 2011; **108**(31-32): 532-40.
32. Weatherall DJ, Clegg JB. Thalassaemia: Classification, Genetics and Relationship to Other Inherited Disorders of Haemoglobin. The Thalassaemia Syndromes: Blackwell Science Ltd; 2001: 121-32.
33. Weatherall DJ. Single gene disorders or complex traits: lessons from the thalassaemias and other monogenic diseases. *BMJ* 2000; **321**(7269): 1117.
34. Thom CS, Dickson CF, Gell DA, Weiss MJ. Hemoglobin Variants: Biochemical Properties and Clinical Correlates. *Cold Spring Harbor Perspectives in Medicine* 2013; **3**(3): a011858.
35. Hardison RC, Chui DH, Giardine B, et al. HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Human mutation* 2002; **19**(3): 225-33.
36. Marengo-Rowe AJ. Structure-function relations of human hemoglobins. *Proceedings (Baylor University Medical Center)* 2006; **19**(3): 239-45.
37. Modell B, Darlinson M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization* 2008; **86**(6): 480-7.
38. Booth C, Inusa B, Obaro SK. Infection in sickle cell disease: A review. *International Journal of Infectious Diseases* 2010; **14**(1): e2-e12.

39. Italia Y, Krishnamurti L, Mehta V, et al. Feasibility of a newborn screening and follow-up programme for sickle cell disease among South Gujarat (India) tribal populations. *Journal of medical screening* 2015; **22**(1): 1-7.
40. Odièvre M-H, Verger E, Silva-Pinto AC, Elion J. Pathophysiological insights in sickle cell disease. *The Indian Journal of Medical Research* 2011; **134**(4): 532-7.
41. Chakravorty S, Williams TN. Sickle cell disease: a neglected chronic disease of increasing global health importance. *Archives of Disease in Childhood* 2014; **100**(1): 48.
42. Steinberg MH. Pathophysiology of sickle cell disease. *Bailliere's clinical haematology* 1998; **11**(1): 163-84.
43. Hedrick PW. Genetics of Populations. 4th ed. Sudbury, Massachusetts: Jones and Bartlett; 2011.
44. Piel FB, Howes RE, Patil AP, et al. The distribution of haemoglobin C and its prevalence in newborns in Africa. *Scientific Reports* 2013; **3**: 1671.
45. Fucharoen S, Winichagoon P. Haemoglobinopathies in Southeast Asia. *The Indian Journal of Medical Research* 2011; **134**(4): 498-506.
46. Olivieri NF, Muraca GM, O'Donnell A, Premawardhena A, Fisher C, Weatherall DJ. Studies in haemoglobin E beta-thalassaemia. *British journal of haematology* 2008; **141**(3): 388-97.
47. Weatherall DJ, Clegg JB. The Pathophysiology of the Thalassaemias. The Thalassaemia Syndromes: Blackwell Science Ltd; 2001: 192-236.
48. Thein SL. The Molecular Basis of β-Thalassemia. *Cold Spring Harbor Perspectives in Medicine* 2013; **3**(5): a011700.
49. Piel FB, Weatherall DJ. The alpha-thalassemias. *The New England journal of medicine* 2014; **371**(20): 1908-16.
50. Galanello R, Cao A. Alpha-thalassemia.
51. Wiwanitkit V. Tertiary structural analysis of the elongated part of an abnormal hemoglobin, hemoglobin Pakse. *International Journal of Nanomedicine* 2006; **1**(1): 105-7.
52. Purohit P, Dehury S, Patel S, Patel DK. Prevalence of Deletional Alpha Thalassemia and Sickle Gene in a Tribal Dominated Malaria Endemic Area of Eastern India. *ISRN Hematology* 2014; **2014**: 6.
53. Weatherall DJ. Phenotype[mdash]genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat Rev Genet* 2001; **2**(4): 245-55.

54. Songdej D, Babbs C, Higgs DR. An international registry of survivors with Hb Bart's hydrops fetalis syndrome. *Blood* 2017; **129**(10): 1251-9.
55. Chui DH, Fucharoen S, Chan V. Hemoglobin H disease: not necessarily a benign disorder. *Blood* 2003; **101**(3): 791-800.
56. Lal A, Goldrich ML, Haines DA, Azimi M, Singer ST, Vichinsky EP. Heterogeneity of Hemoglobin H Disease in Childhood. *New England Journal of Medicine* 2011; **364**(8): 710-8.
57. Fucharoen S, Viprakasit V. Hb H disease: clinical course and disease modifiers. *Hematology American Society of Hematology Education Program* 2009: 26-34.
58. Origa R, Galanello R. Pathophysiology of beta thalassaemia. *Pediatric endocrinology reviews : PER* 2011; **8 Suppl 2**: 263-70.
59. Thein SL. Pathophysiology of β Thalassemia—A Guide to Molecular Therapies. *ASH Education Program Book* 2005; **2005**(1): 31-7.
60. Weatherall DJ, Akinyanju O, Fucharoen S, Olivieri N, Musgrove P. Inherited Disorders of Hemoglobin. Disease Control Priorities in Developing Countries. 2nd edition ed. New York: Oxford University Press; 2006.
61. World Health Organization. Fifty-ninth World Health Assembly: resolutions and decisions, annexes. *WHA59/2006/REC/1* 2006; **Geneva**(World Health Organisation).
62. Organization WH. 118th Session of the Executive Board: resolutions and decisions, annexes. *EBSS-EB118/2006/REC/1* 2006; **Geneva**(World Health Organisation).
63. Murray CJL, Ezzati M, Flaxman AD, et al. GBD 2010: a multi-investigator collaboration for global comparative descriptive epidemiology. *The Lancet* 2012; **380**(9859): 2055-8.
64. Piel FB, Patil AP, Howes RE, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. *Lancet* 2013; **381**(9861): 142-51.
65. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science (New York, NY)* 1908; **28**(706): 49-50.
66. Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahresh Wuertt Verh Vaterl Naturkd; 1908: 369-82.
67. Lobo I. Multifactorial Inheritance and Genetic Disease. *Nature Education* 2008; **1**(1).

68. Piel FB. The Present and Future Global Burden of the Inherited Disorders of Hemoglobin. *Hematology/Oncology Clinics* 2016; **30**(2): 327-41.
69. Kulaphisit M, Kampuansai J, Leecharoenkiat K, et al. A comprehensive ethnic-based analysis of alpha thalassaemia allelle frequency in northern Thailand. *Sci Rep* 2017; **7**(1): 4690.
70. Sabahelzain MM, Hamamy H. The ethnic distribution of sickle cell disease in Sudan. *The Pan African Medical Journal* 2014; **18**: 13.
71. Colah RB, Mukherjee MB, Martin S, Ghosh K. Sickle cell disease in tribal populations in India. *The Indian Journal of Medical Research* 2015; **141**(5): 509-15.
72. Weatherall DJ. The importance of micromapping the gene frequencies for the common inherited disorders of haemoglobin. *British journal of haematology* 2010; **149**(5): 635-7.
73. Patil AP, Gething PW, Piel FB, Hay SI. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends in Parasitology*; **27**(6): 246-53.
74. Gupta RB. Sickle cell disease load in Madhya Pradesh. *RMRCT Update: A Biannual Newsletter of RMRCT* 2006; **3**(1): 1-6.
75. Colah R, Mukherjee M, Ghosh K. Sickle cell disease in India. *Current opinion in hematology* 2014; **21**(3): 215-23.
76. Kaur M, Dangi CBS, Singh M, Singh H, Kapoor S. Burden of sickle cell diseases among tribes of India - A burning problem. *International Research Journal of Pharmaceutical and Applied Sciences* 2013; **3**(1): 60-80.
77. Upadhye DS, Jain DL, Trivedi YL, Nadkarni AH, Ghosh K, Colah RB. Neonatal Screening and the Clinical Outcome in Children with Sickle Cell Disease in Central India. *PLOS ONE* 2016; **11**(1): e0147081.
78. Mohanty D, Das K. Genetic counselling in tribals in India. *The Indian Journal of Medical Research* 2011; **134**(4): 561-71.
79. Patra PK, Khodiar PK, Hambleton IR, Serjeant GR. The Chhattisgarh state screening programme for the sickle cell gene: a cost-effective approach to a public health problem. *Journal of Community Genetics* 2015; **6**(4): 361-8.
80. Tritipsombut J, Sanchaisuriya K, Phollarp P, et al. Micromapping of Thalassemia and Hemoglobinopathies in Diferent Regions of Northeast Thailand and Vientaine, Laos People's Democratic Republic. *Hemoglobin* 2012; **36**(1): 47-56.

81. Chui DHK. Alpha-thalassaemia and population health in Southeast Asia. *Annals of Human Biology* 2005; **32**(2): 123-30.
82. Panomai N, Sanchaisuriya K, Yamsri S, et al. Thalassemia and iron deficiency in a group of northeast Thai school children: relationship to the occurrence of anemia. *Eur J Pediatr* 2010; **169**(11): 1317-22.
83. Munkongdee T, Tanakulmas J, Butthep P, et al. Molecular Epidemiology of Hemoglobinopathies in Cambodia. *Hemoglobin* 2016; **40**(3): 163-7.
84. Fucharoen S, Winichagoon P. Haemoglobinopathies in southeast Asia. *Indian J Med Res* 2011; **134**: 498-506.
85. Weatherall DJ, Clegg JB. Distribution and Population Genetics of the Thalassaemias. The Thalassaemia Syndromes: Blackwell Science Ltd; 2001: 237-84.
86. Carter R, Mendis KN. Evolutionary and Historical Aspects of the Burden of Malaria. *Clinical Microbiology Reviews* 2002; **15**(4): 564-94.
87. Joy DA, Feng X, Mu J, et al. Early origin and recent expansion of Plasmodium falciparum. *Science (New York, NY)* 2003; **300**(5617): 318-21.
88. Gething PW, Patil AP, Smith DL, et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malaria Journal* 2011; **10**: 378-.
89. Snounou G, Viriyakosol S, Jarra W, Thaithong S, Brown KN. Identification of the four human malaria parasite species in field samples by the polymerase chain reaction and detection of a high prevalence of mixed infections. *Molecular and Biochemical Parasitology* 1993; **58**(2): 283-92.
90. White NJ. Plasmodium knowlesi: The Fifth Human Malaria Parasite. *Clinical Infectious Diseases* 2008; **46**(2): 172-3.
91. Allison AC. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *British Medical Journal* 1954; **1**(4857): 290-4.
92. Aidoo M, Terlouw DJ, Kolczak MS, et al. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 2002; **359**(9314): 1311-2.
93. Fleming AF, Storey J, Molineaux L, Iroko EA, Attai ED. Abnormal haemoglobins in the Sudan savanna of Nigeria. I. Prevalence of haemoglobins and relationships between sickle cell trait, malaria and survival. *Annals of tropical medicine and parasitology* 1979; **73**(2): 161-72.

94. Le Hesran JY, Personne I, Personne P, et al. Longitudinal study of Plasmodium falciparum infection and immune responses in infants with or without the sickle cell trait. *International journal of epidemiology* 1999; **28**(4): 793-8.

95. Marsh K, Otoo L, Hayes RJ, Carson DC, Greenwood BM. Antibodies to blood stage antigens of Plasmodium falciparum in rural Gambians and their relation to protection against infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1989; **83**(3): 293-303.

96. Willcox M, Bjorkman A, Brohult J, Pehrson PO, Rombo L, Bengtsson E. A case-control study in northern Liberia of Plasmodium falciparum malaria in haemoglobin S and beta-thalassaemia traits. *Annals of tropical medicine and parasitology* 1983; **77**(3): 239-46.

97. Billo MA, Johnson ES, Doumbia SO, et al. Sickle Cell Trait Protects Against Plasmodium falciparum Infection. *American Journal of Epidemiology* 2012; **176**(suppl_7): S175-S85.

98. Flint J, Hill AV, Bowden DK, et al. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* 1986; **321**(6072): 744-50.

99. Modiano G, Morpurgo G, Terrenato L, et al. Protection against malaria morbidity: near-fixation of the alpha-thalassemia gene in a Nepalese population. *Am J Hum Genet* 1991; **48**(2): 390-7.

100. Allen SJ, O'Donnell A, Alexander NDE, et al. α(+)-Thalassemia protects children against disease caused by other infections as well as malaria. *Proceedings of the National Academy of Sciences of the United States of America* 1997; **94**(26): 14736-41.

101. Opoku-Okrah C, Gordge M, Kweku Nakua E, et al. An investigation of the protective effect of alpha+-thalassaemia against severe Plasmodium falciparum amongst children in Kumasi, Ghana. *International journal of laboratory hematology* 2014; **36**(1): 62-70.

102. Mockenhaupt FP, Ehrhardt S, Gellert S, et al. Alpha(+)-thalassemia protects African children from severe malaria. *Blood* 2004; **104**(7): 2003-6.

103. Gong L, Parikh S, Rosenthal PJ, Greenhouse B. Biochemical and immunological mechanisms by which sickle cell trait protects against malaria. *Malaria Journal* 2013; **12**(1): 317.

104. Daou M, Kituma E, Kavishe R, et al. α-Thalassaemia trait is associated with Antibody prevalence against Malaria Antigens AMA-1 and MSP-1. *Journal of Tropical Pediatrics* 2015; **61**(2): 139-42.
105. Lopez C, Saravia C, Gomez A, Hoebeke J, Patarroyo MA. Mechanisms of genetically-based resistance to malaria. *Gene* 2010; **467**(1-2): 1-12.
106. Verra F, Mangano VD, Modiano D. Genetics of susceptibility to Plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. *Parasite immunology* 2009; **31**(5): 234-53.
107. Trent RJ, Mickleson KNP, Wilkinson T, et al. α Globin gene rearrangements in polynesians are not associated with malaria. *American Journal of Hematology* 1985; **18**(4): 431-3.
108. Molina-Cruz A, Barillas-Mury C. The remarkable journey of adaptation of the Plasmodium falciparum malaria parasite to New World anopheline mosquitoes. *Memórias do Instituto Oswaldo Cruz* 2014; **109**(5): 662-7.
109. Penman BS, Pybus OG, Weatherall DJ, Gupta S. Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean. *Proceedings of the National Academy of Sciences of the United States of America* 2009; **106**(50): 21242-6.
110. Weatherall DJ, Wood WG, Pressley L, Higgs DR, Clegg JB. MOLECULAR BASIS FOR MILD FORMS OF HOMOZYGOUS BETA-THALASSAEMIA. *The Lancet* 1981; **317**(8219): 527-9.
111. Wainscoat JS, Kanavakis E, Wood WG, et al. Thalassaemia intermedia in Cyprus: the interaction of alpha and beta thalassaemia. *British journal of haematology* 1983; **53**(3): 411-6.
112. Wai Kan Y, G. Nathan D. Mild thalassemia: the result of interactions of alpha and beta thalassemia genes; 1970.
113. Law HY, Chee MK, Tan GP, Ng IS. The simultaneous presence of alpha- and beta-thalassaemia alleles: a pitfall of thalassaemia screening. *Community genetics* 2003; **6**(1): 14-21.
114. Williams TN, Mwangi TW, Wambua S, et al. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. *Nat Genet* 2005; **37**(11): 1253-7.

115. Herbert Opi D, Ochola LB, Tendwa M, et al. Mechanistic Studies of the Negative Epistatic Malaria-protective Interaction Between Sickle Cell Trait and α(+)thalassemia. *EBioMedicine* 2014; **1**(1): 29-36.
116. Crompton PD, Traore B, Kayentao K, et al. Sickle cell trait is associated with a delayed onset of malaria: implications for time-to-event analysis in clinical studies of malaria. *The Journal of infectious diseases* 2008; **198**(9): 1265-75.
117. Barrail I, Schiliro G, Beretta M, Mazzetti P, Russo A, Mancusoz GR. Population structure of Sicily: beta-thalassaemia and HbS. *Human Genetics* 1987; **75**: 1-3.
118. Penman BS, Habib S, Kanchan K, Gupta S, Read A. NEGATIVE EPISTASIS BETWEEN α(+) THALASSAEMIA AND SICKLE CELL TRAIT CAN EXPLAIN INTERPOPULATION VARIATION IN SOUTH ASIA. *Evolution; International Journal of Organic Evolution* 2011; **65**(12): 3625-32.
119. Gonçalves BP, Gupta S, Penman BS. Sickle haemoglobin, haemoglobin C and malaria mortality feedbacks. *Malaria Journal* 2016; **15**(1): 26.
120. Cao A, Galanello R. Beta-thalassemia. *Genet Med* 2010; **12**(2): 61-76.
121. Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 2007; **8**(10): 762-75.
122. Hedrick PW. Resistance to malaria in humans: the impact of strong, recent selection. *Malaria Journal* 2012; **11**(1): 349.
123. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* 2010; **115**(22): 4331.
124. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011; **73**(4): 423-98.
125. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**(2): 319-92.
126. Rue H, Riebler A, Sørbye SH, Illian J, Simpson DP, Lindgren F. Bayesian Computing with INLA: A Review. *Annual Review of Statistics and its Application* 2017; **4**(1).
127. Majumder PP, Basu A. A Genomic View of the Peopling and Population Structure of India. *Cold Spring Harbor Perspectives in Biology* 2015; **7**(4): a008540.

128. Leung WC, Leung KY, Lau ET, Tang MH, Chan V. Alpha-thalassaemia. *Seminars in fetal & neonatal medicine* 2008; **13**(4): 215-22.
129. Diggle PJ, Ribeiro PJ. Model-based Geostatistics. 1 ed: Springer-Verlag New York; 2007.
130. De Smedt T, Simons K, Van Nieuwenhuyse A, Molenberghs G. Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Archives of Public Health* 2015; **73**(Suppl 1): O2-O.
131. Gómez-Rubio V, Bivand RS, Rue H. Spatial Models Using Laplace Approximation Methods. In: Fischer MM, Nijkamp P, eds. Handbook of Regional Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014: 1401-17.
132. Weatherall DJ. The Role of the Inherited Disorders of Hemoglobin, the First "Molecular Diseases," in the Future of Human Genetics. *Annual Review of Genomics and Human Genetics* 2013; **14**(1): 1-24.
133. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. *Annual review of genetics* 2013; **47**: 97-120.
134. Penman BS, Gupta S, Buckee CO. The emergence and maintenance of sickle cell hotspots in the Mediterranean. *Infection, Genetics and Evolution* 2012; **12**(7): 1543-50.
135. Loggetto SR. Sickle cell anemia: clinical diversity and beta S-globin haplotypes. *Revista Brasileira de Hematologia e Hemoterapia* 2013; **35**(3): 155-7.
136. Livingstone FB. Who gave whom hemoglobin S: The use of restriction site haplotype variation for the interpretation of the evolution of the βS-globin gene. *American Journal of Human Biology* 1989; **1**(3): 289-302.
137. Livingstone FB. Simulation of the diffusion of the beta-globin variants in the Old World. *Human biology* 1989; **61**(3): 297-309.
138. Ralph P, Coop G. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics* 2010; **186**(2): 647-68.
139. Musa GJ, Chiang P-H, Sylk T, et al. Use of GIS Mapping as a Public Health Tool—From Cholera to Cancer. *Health Services Insights* 2013; **6**: 111-6.
140. Italia K, Kangne H, Shanmukaiah C, Nadkarni AH, Ghosh K, Colah RB. Variable phenotypes of sickle cell disease in India with the Arab-Indian haplotype. *British journal of haematology* 2015; **168**(1): 156-9.
141. Jain D, Warthe V, Dayama P, et al. Sickle Cell Disease in Central India: A Potentially Severe Syndrome. *Indian J Pediatr* 2016: 1-6.

142. Steinberg MH, Sebastiani P. Genetic Modifiers of Sickle Cell Disease. *American journal of hematology* 2012; **87**(8): 795-803.
143. Tewari S, Brousse V, Piel FB, Menzel S, Rees DC. Environmental determinants of severity in sickle cell disease. *Haematologica* 2015; **100**(9): 1108-16.
144. Piel FB, Tewari S, Brousse V, et al. Associations between environmental factors and hospital admissions for sickle cell disease. *Haematologica* 2016.
145. Serjeant GR. Evolving locally appropriate models of care for Indian sickle cell disease. *The Indian Journal of Medical Research* 2016; **143**(4): 405-13.
146. Galanello R, Dessi E, Melis MA, et al. Molecular analysis of beta zero-thalassemia intermedia in Sardinia. *Blood* 1989; **74**(2): 823-7.