bjh research paper

The evolutionary and clinical implications of the uneven distribution of the frequency of the inherited haemoglobin variants over short geographical distances

Anuja Premawardhena,[1] Angela Allen,[2] Fred Piel,[3] Chris Fisher,[2] Laxman Perera,[1] Rexan Rodrigo,[1] Gayan Goonathilaka,[1] Lebbe Ramees,[1] Tim Peto,[4] Nancy Olivieri[5] and David Weatherall[2]

[1]University of Kelaniya, Colombo, Sri Lanka, [2]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, [3]Imperial College, London, [4]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom, and [5]University of Toronto, Toronto, Canada

Received 9 May 2016; accepted for publication 23 August 2016

Correspondence: Professor Sir David Weatherall FRS, MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DS, UK.
E-mail: liz.rose@imm.ox.ac.uk

Summary

Studies of the frequency of heterozygous carriers for common inherited diseases of haemoglobin in over 7500 adolescent children in 25 districts in Sri Lanka have disclosed a highly significant variation over very short geographical distances. A further analysis of these findings, including their relationship to the past frequency and distribution of malaria, climatic variation, altitude, ethnic origin and consanguinity rates, have provided evidence regarding the evolutionary basis for the variable distribution of these conditions over short distances. It is likely that the complex interplay between malaria and the environment, together with related ethnic and social issues, exists in many countries across the tropical belt. Hence, these observations emphasise the importance of micromapping heterozygote distributions in high-frequency countries in order to define their true burden and the facilities required for the prevention and management of the homozygous and compound heterozygous disorders that result from their interaction.

Keywords: haemoglobinopathies, distribution, malaria, micromapping, consanguinity.

The inherited disorders of haemoglobin (Hb) are the commonest monogenic diseases. They occur at a high frequency throughout the tropical belt and in countries with large numbers of immigrants from this region. However, because of the possibility of their heterogeneous distribution in some high-frequency areas their overall clinical burden, particularly for the developing countries, remains poorly defined.

Normal adult Hb consists of two α and two β globin chains ($\alpha_2\beta_2$), each with an associated haem group, which are regulated by duplicated *HBA1* and *HBA2* genes and a single *HBB* gene, respectively (Hardison, 2013). Two main groups of diseases result from mutations at or related to these loci, structural Hb variants and the thalassaemias, which are caused by defective synthesis of the α or β globin chains. Of the hundreds of inherited structural Hb variants, three, Hbs S, C and E, occur at very high frequencies. The homozygous or compound heterozygous states for HbS and HbC are associated with haemolytic anaemia and a wide variety of vascular complications (Agarwal *et al*, 2009), while HbE, possibly the commonest of the three because the underlying mutation produces a cryptic splice site in *HBB*, is produced at a reduced rate resulting in the phenotype of a mild form of β thalassaemia (Orkin *et al*, 1982).

There are two main forms of α thalassaemia, α^+ and α^0 thalassaemia. The former results from either deletions of one of the duplicated *HBA1* or *HBA2* genes ($-\alpha/\alpha\alpha$) or from a mutation of one of them ($\alpha\alpha^T/\alpha\alpha$ or $\alpha^T\alpha/\alpha\alpha$) (Higgs & Gibbons, 2010). The deletion forms, which result from misalignment crossovers between the *HBA1* and *HBA2* genes, consist of several $-\alpha^{3.7}$ types and a less common $-\alpha^{4.2}$. The superscripts 3·7 and 4·2 describe the size of particular deletions. There are three types of $-\alpha^{3.7}$: $-\alpha^{3.7I}$, $-\alpha^{3.7II}$ and $-\alpha^{3.7III}$ which result from slightly different crossover sites. The α^0 thalassaemias result from deletions of both linked α globin genes ($--/\alpha\alpha$). The deletional forms of α^+ thalassaemia are, by far, the commonest monogenic diseases, occurring at high frequencies right across the tropical belt, almost reaching fixation in parts of South and Southeast Asia (Weatherall &

First published online 29 November 2016
doi: 10.1111/bjh.14437

Clegg, 2001). The α^0 thalassaemias occur sporadically throughout the tropical belt and only reach very high frequencies in Southeast Asia. The homozygous states (−/−) result in stillbirths while the compound heterozygous state with α^+ thalassaemia (−α/− or $\alpha\alpha^T$/−) result in HbH disease, a condition characterised by a haemolytic anaemia of varying severity. Haemoglobin H is an unstable tetramer of β-chains that results from defective α-chain synthesis.

The β thalassaemias, which result from mutations of the *HBB* genes, and which occur at varying frequencies right across the tropical belt, are of two main kinds. First, there are the homozygous or compound heterozygous states for mutations that reduce or abolish the output of the *HBB* genes. Approximately half of the severe forms of β thalassaemia, and those that occur mainly on the east side of South Asia and throughout Southeast Asia, result from compound heterozygosity for HbE and β thalassaemia. The HbE β thalassaemias are characterised by a particularly variable phenotype, ranging from complete transfusion dependence to an almost symptomless disorder (Olivieri *et al*, 2008).

Knowledge of the true frequency of these diseases is still limited however. Much of the published data are based on assessments made in a few centres, which have then been extrapolated to the whole country. Early studies, particularly in the case of the α thalassaemias (O'Shaughnessy *et al*, 1990; Martinson *et al*, 1995; O'Riordan *et al*, 2010), suggested that their frequency might vary widely within relatively short geographical distances. However, apart from one study in India, which reached the same conclusion for the β thalassaemias (Colah *et al*, 2010), there have been no further investigations of this type and no explanation of the reason for the heterogeneous distribution of haemoglobin variants over short distances. While a better understanding of the basis for these observations is highly relevant to determining the mechanisms of the evolutionary biology of the haemoglobin variants, it is also of particular importance in assessing the clinical value of detailed micromapping to determine the true magnitude of the health problems and costs of prevention and management of these diseases, particularly in high-frequency low-income countries.

This study analysed the heterozygote frequency of several common haemoglobin disorders in over 7500 adolescent children in 25 different districts in Sri Lanka. The results indicate a remarkable and highly significant difference in the frequency of these conditions within relatively small geographical distances. To attempt to account for these findings a variety of approaches have been pursued, including an analysis of the past history of the distribution of malaria on the island, the effects of altitude and rainfall, the role of different ethnic origins, the role of consanguineous marriage and others. As well as providing preliminary evidence about the evolutionary mechanisms involved in the remarkable geographical variability of these conditions the results of this study have underlined the crucial importance of micromapping the frequency and distribution of the Hb variants before it is possible to determine the true burden of the clinical services required for their management in high-incidence countries.

Methods

Population under study

The study involved 7526 adolescent students from schools across the 25 districts of Sri Lanka. The names of the districts and approximate location of the schools are shown in the Supporting Information (Figure S1). Three schools per district were sampled. Prior permission was obtained from each of the schools and the families involved and the appropriate findings and advice fed back to the families. Ethical approval was obtained for the programme from the Ethical Committee, University of Kelaniya, Sri Lanka and the Oxford University Tropical Research Ethics Committee.

Haematological and haemoglobin analysis

Five millilitre blood samples were collected from each student. Routine haematological measurements were performed using a Coulter Counter (Coulter Electronics, Ltd., Luton, UK). Haemoglobin analysis was performed by high performance liquid chromatography (β thalassaemia short programme, Bio-Rad India, Gurgaon, India). For samples with low red cell indices DNA was extracted from the cell pellet using a QIAamp DNA mini kit 51306 (Qiagen, Manchester, UK) and the α globin genotype determined by multiplex polymerase chain reaction (Liu *et al*, 2000). Because previous studies have shown that there is a slight overlap between reduced and normal red cell indices in the case of α^+ thalassaemia carriers (Weatherall & Clegg, 2001), 817 samples from adolescents with normal indices from different regions were also examined by DNA analysis.

Significance of geographical distribution of haemoglobin variants

The significance of this regional variation was assessed by a chi squared test (24 degrees of freedom).

The relationship with malaria prevalence

A detailed review of historical data on malaria prevalence in Sri Lanka identified two main data sources relevant to this study. A detailed nationwide malaria survey based on splenic examinations (palpation) of random samples of children up to 12 years of age collected in over 500 towns and villages was conducted in 1921–1922. More than 56 000 children were examined. The map presenting the results of this survey in 6 classes (0–5%, 5–10%, 10–20%, 20–40%, 40–60% and >60%), published by Gill (1936), was digitized in ArcGIS

10.3. Point values were extracted for each school surveyed and the mean of all pixels within each district was calculated. Another detailed malaria survey, based on the examination of blood smears of 105 957 Sri Lankans, was conducted before the anti-malaria campaign was launched in 1958. The results were only presented at the district level and therefore not analysed at the school level. The relative accuracy of these data on the distribution of malaria was confirmed by several later publications (Curtis & Rawlings, 1980; Rawlings *et al*, 1981; van der Hoek *et al*, 1997). More recent data sources on malaria prevalence were not considered suitable for this study due to the confounding influence of the malaria control programme and some recent population movements associated with the conflict in the northern part of Sri Lanka. The relationship between the prevalence of each haemoglobin variant and malaria was assessed at District ($n = 25$) and School ($n = 69$) levels using a Pearson correlation test (Eq. 1), in R3.2.2 (R Foundation for Statistical Computing, Vienna, Austria), with a statistical significance of 0·05 (Table I). When spatial autocorrelation was observed we tested for residual autocorrelation and, if statistically significant, used a lag spatial model (Eq. 2) to assess the relationship between the prevalence of inherited disorders of haemoglobin and malaria.

$$Y = \beta X + \varepsilon \quad (1)$$

where Y is the frequency of an inherited blood disorder and X is the frequency of malaria.

$$Y = pWY + \beta X + \varepsilon \quad (2)$$

where W is the weighted spatial proximity matrix; WY expresses the spatial dependence in Y; p is the autocorrelation spatial coefficient.

Table I. *P* values of the Pearson correlations between the prevalence of inherited disorders of haemoglobin and environmental and social factors at the district ($n = 25$) and school ($n = 69$) levels.

	Inherited disorders of haemoglobin					
	$\alpha^{3.7}$-thal $n = 542$	$\alpha^{4.2}$-thal $n = 63$	β-thal $n = 149$	HbE $n = 39$	HbD $n = 13$	HbS $n = 11$
District level ($n = 25$)						
Environmental factors						
Malaria (spleen rate, 1921–1922)	**0·031**	0·094	0·069	0·139	0·299	**0·036**
Malaria (blood smear, 1957)*	**0·012**	0·879	**0·031**	**0·005**	0·294	0·927
Altitude	0·477	0·813	0·696	0·732	0·448	0·258
Mean temperature	0·219	0·628	0·534	0·954	0·373	0·234
Rainfall	0·051	0·123	**0·047**	0·168	0·171	0·100
Ethnicity						
Sinhalese	0·699	0·814	0·277	0·973	0·116	0·379
Sri Lanka tamils	0·876	0·856	0·370	0·705	0·350	0·286
Indian tamils	0·673	0·911	0·400	0·611	0·546	0·444
Sri Lanka moors	0·228	0·741	0·882	0·401	**0·005**	0·861
Others	0·342	0·307	0·548	0·554	0·783	0·225
School level ($n = 69$)						
Environmental factors						
Malaria (spleen rate, 1921–1922)	**0·002**	0·051	**0·001**	0·069	0·405	0·265
Malaria (blood smear, 1957)†	/	/	/	/	/	/
Altitude	0·087	0·334	0·988	0·825	0·886	0·264
Mean temperature	**0·017**	0·212	0·761	0·615	0·754	0·249
Rainfall	**0·006**	**0·036**	**0·013**	**0·050**	0·255	0·263
Ethnicity						
Sinhalese	0·626	0·180	0·225	0·851	0·621	0·490
Sri Lanka tamils	0·689	0·903	0·302	0·535	0·524	0·293
Indian tamils	0·385	0·729	0·510	0·855	0·428	0·500
Sri Lanka moors	**0·019**	0·093	0·523	0·265	0·646	0·666
Others	0·470	0·625	0·430	0·332	0·941	0·696

Statistically significant values ($P < 0·05$) are shown in bold. HbC was excluded from the analyses because only one individual was found with this variant. Burgher, Malay, Sri Lankan Chetty, Bharattha and Other ethnic groups were grouped into a single "Others" group due to the small size of each single group in all districts.

*Samples from Galle ($n = 4$) and Kalutara ($n = 5$) were removed due to the very small number of blood smears examined in these two districts.

†Data not available at the school level. There was also a significant correlation between the spleen rates and α thalassaemia trait ($P = 0·037$) and β thalassaemia trait ($P = 0·002$) based on early information about spleen rates in the different provinces of the island, not shown in the table.

The relationship with other environmental and social data

The associations were tested between other environmental data often used as a proxy for malaria, including altitude, mean temperature and rainfall, as well as ethnicity data. High-resolution raster layers (~1 km) for altitude, annual mean temperature (BIO1) and annual precipitations (BIO12) were downloaded from the WorldClim website (http://www.worldclim.org). Further details about these datasets have been published in Hijmans *et al* (2005). Specific values were extracted for the location of each of the schools and areal averages were calculated for each of the 25 districts of Sri Lanka. Data on population ethnicity at district and sub-district (DS division) levels from the 2012 Census of Population and Housing were extracted from the Sri Lanka Department of Census and Statistics' website (http://www.statistics.gov.lk/pophousat/cph2011/index.php?fileName=pop32&gp=Activities&tpl=3). Data included the absolute number of individuals from the following ethnic groups: Sinhalese, Sri Lankan Tamil, Indian Tamil, Sri Lankan Moor, Burgher, Malay, Sri Lankan Chetty, Bharatha and Other. Due to the absence or small number of individuals in the Burgher, Malay, Sri Lankan Chetty, Bharatha and Other ethnic groups in all districts, they were merged into a single 'Other' category for statistical analysis. The same approach as described above was used to assess statistically significant relationships (Table I).

Consanguinity assessment

Information about appropriate marriage registrars was obtained for each district from the Registrar General Department. The marriage registrars asked a group of newly married couples to fill in a questionnaire regarding information about their relationship to their families and the patterns of partner selection. The families of the adolescent children with haemoglobin varaints were given genetic counselling but were not questioned about consanguinity.

Population estimates

In order to calculate the annual national number of babies born homozygous for β-thalassemia, compound heterozygous for both HbE and β-thalassemia, and homozygous for α-thalassemia, data were extracted from high-resolution population numbers from the WorldPop website (http://www.worldpop.org.uk/) and the 2010-2015 crude birth rate for Sri Lanka from the United Nations World Population Prospects 2015 Revision (16·4 births/1000 population) (https://esa.un.org/unpd/wpp/DVD/Files/1_Indicators%20(Standard)/EXCEL_FILES/2_Fertility/WPP2015_FERT_F03_CRUDE_BIRTH_RATE.XLS). Thiessen polygons were generated in ArcGIS 3.2 (ESRI, Redlands, CA) in order to extrapolate the frequencies of the relevant inherited disorders of hemoglobin to catchment areas around each of the 69 schools surveyed. Using Hardy-Weinberg assumptions, the observed frequencies within each school area were multiplied by the population of the Thiessen polygon and the national crude birth rate. The sum of the local estimates was then compared to national estimates based on averaged frequencies (Table SI).

Results

Analysis of participants

The average number of adolescents studied from each region was 301·08 (range 231–324). The mean age of the males was 16·24 (±1·27) years and that of the females was 16·0 (±1·72) years. The number of males and females was almost identical.

Frequency and distribution of haemoglobin variants

Out of 7526 blood samples there were 542 heterozygotes for $-\alpha^{3.7}$ thalassaemia, 149 for β thalassaemia, 63 for $-\alpha^{4.2}$, 39 for HbE, 11 for HbS and 13 for HbD Punjab. Eighteen samples showed homozygosity for $-\alpha^{3.7}$ thalassaemia and 6 for compound heterozygosity between $-\alpha^{3.7}$ and $-\alpha^{4.2}$ thalassaemia. One case of HbC heterozygosity was also found. To determine whether there were any cases of deletional forms of α thalassaemia in children with normal red cell indices a further 817 samples from 11 districts were analysed. In six districts no cases were found and in the remainder only 11 cases were observed. In a sample of 31 cases of $-\alpha^{3.7}$ thalassaemia 28 were of the $-\alpha^{3.7I}$ subtype, 2 $-\alpha^{3.7II}$ and 1 $-\alpha^{3.7III}$.

There was a major variation in the frequency of these conditions over short geographical distances (Figs 1 and 2). The overall range in the schools for $-\alpha^{3.7}$ thalassaemia trait was 2·9–20%, for β thalassaemia trait 0–16·4% and for HbE trait 0–3·3% (Figure S2). The mean frequency for each of the schools in every district is shown in Table II. These differences are highly significant in the cases of $-\alpha^{3.7}$ thalassaemia ($P = 0·0001$) and β thalassaemia ($P = 0·001$). As the other traits are less common there is less power to detect heterogeneity. However, there is evidence of heterogeneity in HbE trait ($P = 0·047$) and HbS trait ($P = 0·023$), while the HbD trait and $-\alpha^{4.2}$ trait did not achieve significance.

The relationship of the distribution of the haemoglobin variants to malaria

As shown in Fig 2 there was a significant association between elevated spleen rates and the common haemoglobin variants. There were also significant relationships between the frequency of malaria in the 25 districts as judged by parasite levels in the peripheral blood and the common haemoglobin variants based on the published data on blood smear frequencies obtained before the malaria eradication programmes

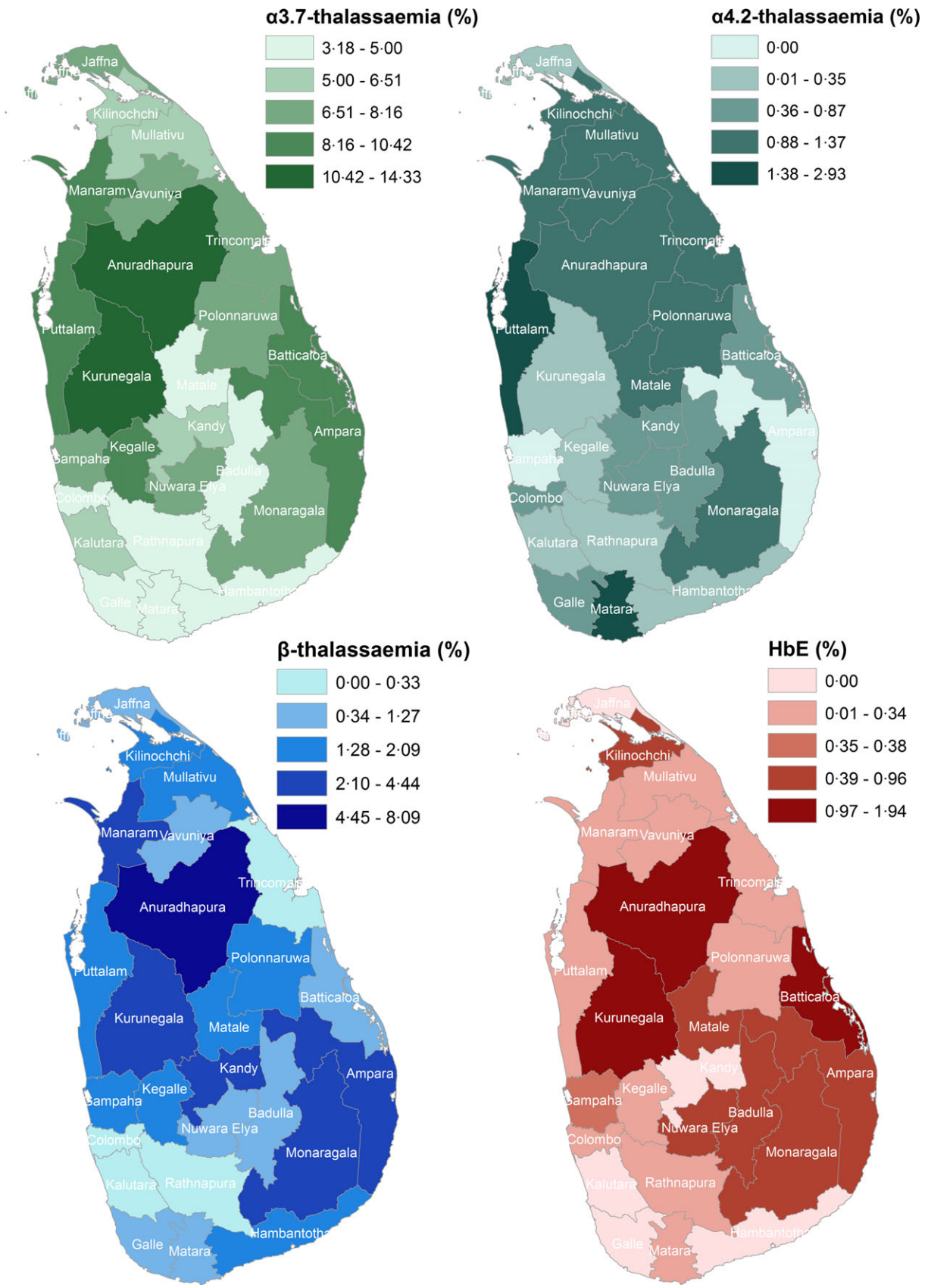

Fig 1. Map of Sri Lanka showing the heterogeneous distribution of the heterozygous states for the inherited haemoglobin disorders.

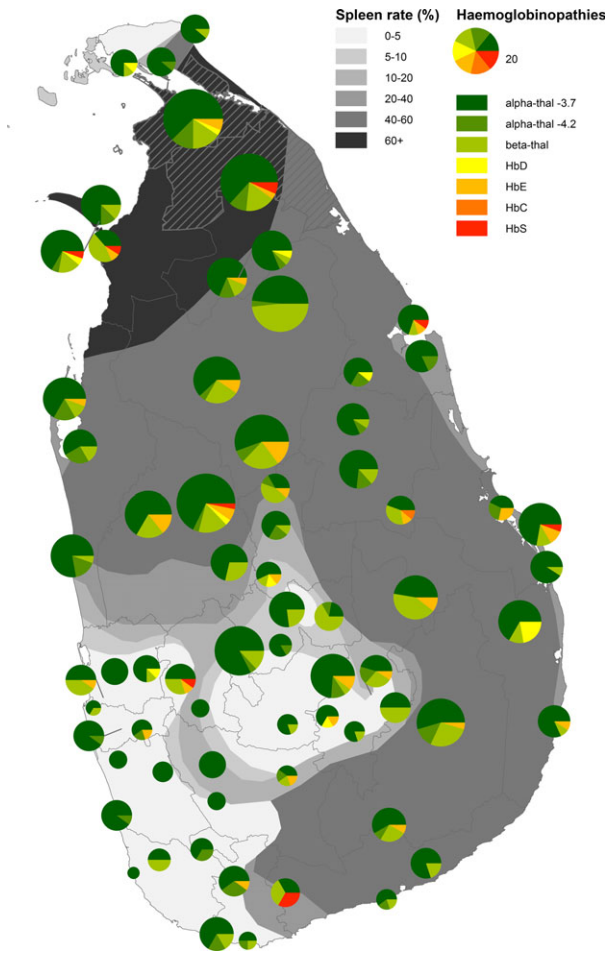

Fig 2. Distribution of inherited disorders of haemoglobin amongst 7524 adolescent children sampled in 69 schools across Sri Lanka. Pie charts are proportional to the number of individuals with an inherited disorder of haemoglobin. The boundaries of the 25 districts and the spleen rates from the 1921–1922 nationwide malaria survey (digitized from Gill, 1936) are shown in the background.

in 1958 (Table I). During later years, particularly those of serious internal conflict, more detailed records of malaria transmission were more difficult to obtain. However, during the several short epidemics that occurred the regions with the highest frequency always included Kurunegala and Anuradhapura, those with the highest frequency of the common haemoglobin variants.

Environmental conditions

The relationship between the distribution of haemoglobin variants and different environmental conditions is summarised in Table I. There was no correlation with temperature or altitude but there was a highly significant correlation with rainfall and the distribution of $-\alpha^{3.7}$ thalassaemia ($P = 0{\cdot}006$) and a lesser though significant association with $-\alpha^{4.2}$ ($P = 0{\cdot}035$) and β thalassaemia ($P = 0{\cdot}012$). These findings reflect the relatively high rainfall in the far south of the island in which the rapidly running rivers and changing lake conditions are not ideal environments for the *Anopheles* mosquito whereas the central and northern regions, with low rainfalls and stagnant pools of water, provide an ideal breeding environment (van der Hoek *et al*, 1997). The variation in the frequency and distribution of the haemoglobin variants, illustrated in Fig 2, reflects these differences.

Gender and ethnic background

There was no difference in frequency or distribution of any of the haemoglobin variants between males and females. The main ethnic groups in Sri Lanka are the Sinhalese and the Tamils. Approximately 7% of the population are Moors who are descendants of Arabs and Indian settlers. Together with 0·3% of the population who originated from Malaysia they constitute the Muslim population of the country. There are several other minor populations but they occur at very low frequencies and were not included in this analysis. According to the 2011 Census there has been very little segregation of different ethnic groups except for the Tamils in the North and in a few centres on the East Coast (http://www.statistics.gov.lk/PopHouSat/PopulationAtla_2012/03_DistrictMaps/MapP2.4.1 Population by Ethnicity and District, 2012.pdf). Apart from the Tamil population in the north, these different ethnic groups are distributed fairly evenly round the country. While there were no differences in the gene frequencies for the common haemoglobin variants between the Sinhalese and Tamil populations there was a higher frequency of $-\alpha^{3.7}$ thalassaemia in the Moors ($P = 0{\cdot}019$). The small group of subjects with the sickle cell trait were equally divided between the Tamils and Sinhalese. HbD trait was equally divided among the three main ethnic groups.

Consanguinity

Although consanguineous marriages are known to be an important factor for the frequency of recessive disorders like those due to abnormal haemoglobins, and are thought to be common in many Asian countries, there is very little published data. In order to approach this problem visits were made to marriage registrars in all the districts in Sri Lanka that were covered by this survey. Information was obtained on the frequency of first and second cousin marriages. Using this approach there was an overall average frequency of consanguineous marriage of 7%. However, there was a considerable difference in the frequency, ranging from 2% to 21% among the regions examined. The highest frequency was in the Tamil population.

Birth numbers of homozygotes and compound heterozygotes

As an example of the clinical relevance of these findings the annual birth rates of the severe forms of thalassaemia have

Table II. The frequency of the different haemoglobin variants across the 25 regions of Sri Lanka. The per cent frequency for each region is derived from the mean frequency of the three different schools per region.

District	Total	$-\alpha^{3.7}/\alpha\alpha$ *n* (%)	$-\alpha^{4.2}/\alpha\alpha$ *n* (%)	β trait *n* (%)	HbE trait *n* (%)	HbS trait *n* (%)	HbD trait *n* (%)
Gampaha	265	19 (7·17%)	0 (0·0%)	5 (1·89%)	1 (0·38%)	0 (0·0%)	1 (0·37%)
Colombo	309	14 (4·53&)	2 (0·65%)	1 (0·32%)	1 (0·32%)	0 (0·0%)	0 (0·0%)
Kalutara	295	18 (6·10%)	1 (0·34%)	0 (0·0%)	0 (0·0%)	0 (0·0%)	0 (0·0%)
Kegalle	306	26 (8·50%)	1 (0·33%)	6 (1·96%)	1 (0·33%)	1 (0·33%)	0 (0·0%)
Kurunegala	300	41 (13·67%)	1 (0·33%)	13 (4·33%)	5 (1·67%)	1 (0·33%)	1 (0·33%)
Kandy	315	17 (5·40%)	2 (0·63%)	9 (2·86%)	0 (0·0%)	0 (0·0%)	0 (0·0%)
Galle	314	9 (2·87%)	1 (0·32%)	3 (0·96%)	0 (0·0%)	0 (0·0%)	0 (0·0%)
Matara	300	14 (4·67%)	6 (2·0%)	3 (1·0%)	1 (0·33%)	0 (0·0%)	0 (0·0%)
Hambantota	287	12 (4·18%)	1 (0·35%)	6 (2·09%)	0 (0·0%)	3 (1·05%)	0 (0·0%)
Ratnapura	299	14 (4·68%)	1 (0·33%)	1 (0·33%)	1 (0·33%)	0 (0·0%)	0 (0·0%)
Kilinochchi	324	17 (5·25%)	4 (1·23%)	5 (1·54%)	2 (0·62%)	0 (0·0%)	1 (0·31%)
Moneragala	315	20 (6·25%)	2 (0·63%)	14 (4·44%)	2 (0·63%)	0 (0·0%)	0 (0·0%)
Mullaitivu	308	21 (6·82%)	3 (0·97%)	5 (1·62%)	1 (0·32%)	2 (0·65%)	0 (0·0%)
Batticaloa	307	26 (8·47%)	2 (0·65%)	3 (0·98%)	4 (1·30%)	1 (0·33%)	0 (0·0%)
Trincomaleee	305	22 (7·21%)	4 (1·31%)	1 (0·33%)	1 (0·33%)	1 (0·33%)	1 (0·33%)
Ampara	314	26 (8·28%)	0 (0·0%)	11 (3·50%)	3 (0;96%)	0 (0·0%)	4 (1·3%)
Jaffna	305	20 (6·56%)	1 (0·33%)	2 (0·66%)	0 (0·0%)	0 (0·0%)	1 (0·33%)
Vavuniya	294	24 (8·16%)	3 (1·02%)	3 (1·02%)	1 (0·34%)	0 (0·0%)	1 (0·34%)
Mannar	301	26 (8·64%)	3 (1·0%)	10 (3·32%)	1 (0·33%)	2 (0·66%)	1 (0·33%)
Badulla	314	13 (4·14%)	2 (0·64%)	4 (1·27%)	2 (0·64%)	0 (0·0%)	1 (0·32%)
Nuwara Eliya	231	18 (7·79%)	2 (0·87%)	2 (0·87%)	2 (0·87%)	0 (0·0%)	0 (0·0%)
Anuradhapura	309	40 (12·94%)	3 (0·97%)	25 (8·09%)	6 (1·94%)	0 (0·0%)	0 (0·0%)
Matale	293	13 (4·44%)	3 (1·02%)	6 (2·05%)	2 (0·68%)	0 (0·0%)	1 (0·34%)
Puttalam	307	29 (9·45%)	7 (2·28%)	5 (1·63%)	1 (0·33%)	0 (0·0%)	0 (0·0%)
Polonnaruwa	310	24 (7·74%)	3 (0·97%)	6 (1·9%)4	1 (0·32%)	0 (0·0%)	0 (0·0%)

been assessed (Table SI). The first approach was to compare birth estimates calculated from the average frequency of different haemoglobin variants amongst the overall sample and then to extrapolate to the entire birth rates nationwide. The other approach were estimations using the local frequency at the 69 schools extrapolated to the number of births within the appropriate catchment area of each school and then summed up for the whole country. It is clear that a single average figure for the overall survey for each haemoglobin variant tends to significantly underestimate the number of homozygotes and compound heterozygotes with an abnormal variant.

Discussion

The inherited disorders of haemoglobin are the commonest monogenic Mendelian diseases. It is currently estimated that between 300 000 and 400 000 babies are born each year, 80% in the low or medium income countries of the tropical belt or in countries that house relatively large numbers of immigrants from this region (Christianson *et al*, 2006). An early study in Sri Lanka, directed mainly at the different forms of severe thalassaemia and their molecular basis, included a small survey of the frequency of β thalassaemia trait and HbE trait and suggested that they might be rather heterogeneously distributed though the numbers were too small to be certain. The α thalassaemia trait also showed signs of heterogeneity although this was only performed in four different centres (de Silva *et al*, 2000). Early studies in Southeast Asia and the Pacific Islands suggested that, particularly in the case of the α thalassaemias, there might be significant differences in the frequency between relatively small geographical distances (Ganczakowski *et al*, 1995; O'Riordan *et al*, 2010). While a later study in west India confirmed that this was the case for the β thalassaemias (Colah *et al*, 2010) no further studies of this type have been reported and there has been no attempt to try to understand the underlying basis or clinical significance of these findings.

The present study has shown that there is a highly significant variation in the gene frequency over short geographical distances for the common forms of haemoglobin disorders in Sri Lanka. Because of the relatively large number of individuals studied in each part of the island it was possible to obtain a high degree of statistical significance for these different frequencies. These findings have important implications for determining the clinical load created by the number of births of homozygotes or compound heterozygotes for the different forms of thalassaemia, particularly those in the high frequency low or medium income countries of the tropical belt. The usual approach is to determine the heterozygote

frequency in a few centres and to extrapolate the findings to the rest of the country. In the present study it was found that this approach, when compared with estimations using local frequency at the 69 schools extrapolated to the number of births within the appropriate catchment area of each school and then summed up for the whole country, significantly under estimated the births of the severe forms of thalassaemia.

Because there is strong evidence that the common forms of haemoglobin disorder have reached a high frequency due to heterozygote protection against *P. falciparum* malaria (Williams & Weatherall, 2012), the relationship between the frequency of the variants and the distribution of malaria has been examined. Because malaria had come under almost complete control at the time of this study this was only possible because of the excellent records that have been kept on the frequency of malaria in Sri Lanka over the last 100 years (Briercliffe & Dalrymple-Champneys, 1936; Gill, 1936). The early studies in Sri Lanka utilised spleen rates, that is the frequency of a palpable spleen, as an indication of malarial infection. Although more recent studies have shown that the spleen may be enlarged in heterozygotes for different forms of thalassaemia, this is only detectable by ultrasonography or related techniques and the spleen is not palpable in these conditions unless there are associated complications, malaria infection for example (Williams *et al*, 1998; Premawardhena *et al*, 2008; Karimi *et al*, 2009). It was found that there was a significant correlation between the early data on spleen rates and the distribution of the haemoglobin variants. These findings were supported after comparing the frequency of malaria as judged by parasite counts in the peripheral blood and relating them to the heterozygote frequency of the haemoglobin disorders. The parasite data were obtained from records of estimates made shortly before the first major malaria eradication programme in 1958. The frequency of the haemoglobin disorders in the absence of a selective factor occurs extremely slowly over many years (Weatherall & Clegg, 2001).

The strong relationship between the variation in frequency of the haemoglobin variants and the pattern of rainfall suggests a further possible malaria-related mechanism for their heterogeneous distribution. The variants are significantly less frequent in the southern parts of the island where rainfall is very heavy and the rivers and lakes are being constantly refilled. This environment compares with the high haemoglobin variant frequency areas in the centre and north of the island where rainfall is limited and hence where sluggish pools of water and lakes provide a much more effective breeding ground for mosquitoes. Although there are limited data on the dispersal and flight range of malaria vectors in Sri Lanka (Curtis & Rawlings, 1980; Rawlings *et al*, 1981) such data as there are suggests that the commonest vector for *P. falciparum*, *Anopheles culicifacies*, probably can travel only 1–2 km. This limited dispersal distance could also be a factor in the different frequency of haemoglobin variants over short geographical distances.

In short, although there are many problems and uncertainties about the use of a historical approach to determining the frequency and distribution of malaria it is hoped that the findings in this study might provide an incentive for further analyses of this type carried out across the tropical zone and particularly in countries in which malaria is still active.

Apart from the higher level of α thalassaemia in the Moors, which constitute only about 7% of the population of Sri Lanka, the frequency of the different haemoglobin variants did not show any significant differences across the other ethnic groups of the island. A history of the island (Meyer, 2006) suggests that there has been very limited movement between different ethnic groups over the years. The reason for the curious distribution of the small number of cases of HbS trait round the periphery of the island remains unexplained. There is no doubt that consanguineous marriages are a high risk for the transmission of Mendelian recessive disorders like the haemoglobin diseases. A previously published survey of the frequency of cousin marriages provided a value which was remarkably close to the 7% frequency found in the present study (Reid, 1976). There was considerable difference in the frequency of consanguineous marriage across the island, the highest frequency occurring in the Tamil population.

In conclusion, by providing some possible insights into the evolutionary biology that accounts for the remarkable variation in the frequency of the common haemoglobin disorders over short geographical distances, this study has potential implications for the future control and management of these disorders. They occur at their highest frequency right across the tropical belt in countries of low or medium income. It seems possible therefore that the complex interplay between malaria and subtle changes in the environment, ethnicity and consanguinity that have been described in Sri Lanka may also occur in many other tropical countries. Most importantly however, the study has provided strong evidence that micromapping and the use of regional birth rates provides a significantly more accurate approach to determining the true burden of the severe haemoglobin disorders than other currently applied approaches. Micromapping, as described here, is relatively inexpensive, particularly if applied to identify the most serious forms of haemoglobin disorder. In particular, it will provide invaluable information for the governments of high-frequency countries and international health agencies regarding the true burden of the prevention and management of these still neglected diseases.

Acknowledgements

We thank Dr. Karmini Mendis of the World Health Organization for her help with the malaria aspects of the study. We also thank Liz Rose of the University of Oxford for her invaluable help in preparing this manuscript. The study was supported by grants from The Wellcome Trust, The Medical

Research Council and The Anthony Cerami and Ann Dunne Foundation for World Health.

Author contributions

AP, LP, RR, GG & LR organised the collection of blood samples and their initial analysis; AA and CF carried out further extensive analysis in the UK; FP performed an extensive statistic analysis of the data; TP also helped with the statistical analysis; NO helped to develop the long-term research and clinical programme in Sri Lanka; and DW helped to devise and analyse the study and wrote the paper.

Disclosure of conflicts of interest

No conflicts of interest.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig S1. A map of Sri Lanka showing the main regions and the approximate positions of the schools from which the adolescent children were analysed for different haemoglobin variants.

Fig S2. Boxplot of the proportion of school children sampled with an inherited disorder of haemoglobin within the 69 schools surveyed across Sri Lanka in this study. The median and interquartile range (IQR) are shown by the bold lines and the blue boxes, respectively. The blue boxes contain values between 25th and 75th percentile. The top/bottom whiskers (upper/lower marks) denote the maximum/minimum or the 3rd/1st quartile plus/minus 1.5 times the IQR, whichever is smaller/larger. Black circles represent values falling outside the lower-upper mark range, i.e. outliers. Red dots represent precise values for the 69 schools, overlaid over the boxplots with a jitter for better visualisation.

Table SI. Frequencies of the most common inherited disorders of hemoglobin in the 69 schools surveyed across Sri Lanka and national estimates of the number of babies born homozygote for β-thalassemia, compound heterozygote for HbE and β-thalassemia, and homozygote for α-thalassemia (−3.7 and −4.2) based on Hardy-Weinberg Equilibrium. Estimates calculated as the sum of local estimates (69 Thiessen polygons) are compared to estimates based on national averaged frequencies of inherited disorders of hemoglobin.

References

Agarwal, N., Nagel, R.L. & Prchal, J.T. (2009) Dyshemoglobinemias. In: Disorders of Hemoglobin, 2nd edn. (ed. by M.H. Steinberg, B.G. Forget, D.R. Higgs & D.J. Weatherall), pp. 607–622. Cambridge University Press, New York.

Briercliffe, R. & Dalrymple-Champneys, W. (1936) The malaria epidemic in Ceylon 1934–1935: (Section of Epidemiology and State Medicine with Section of Tropical Diseases and Parasitology), Joint Discussion No. 1. *Proceedings of the Royal Society of Medicine*, **29**, 537–562.

Christianson, A., Howson, C.P. & Modell, B. (2006) March of Dimes Global Report on Birth Defects. March of Dimes Birth Defects Foundation, New York.

Colah, R., Gorakshakar, A., Phanasgaonkar, S., D'Souza, E., Nadkarni, A., Surve, R., Sawant, P., Master, D., Patel, R., Ghosh, K. & Mohanty, D. (2010) Epidemiology of beta-thalassaemia in Western India: mapping the frequncies and mutations in sub-regions of Maharashtra and Gujarat. *British Journal of Haematology*, **149**, 739–747.

Curtis, C.F. & Rawlings, P. (1980) A preliminary study of dispersal and survival of Anopheles culicifacies in relation to the possibility of inhibiting the spread of insecticide resistance. *Ecological Entomology*, **5**, 11–17.

Ganczakowski, M., Bowden, D.K., Maitland, K., Williams, T.N., O'Shaughnessy, D., Viji, J., Lucassen, A., Clegg, J.B. & Weatherall, D.J. (1995) Thalassaemia in Vanuatu, SW Pacific: frequency and haematological phenotypes of young children. *British Journal of Haematology*, **89**, 485–495.

Gill, C.A. (1936) Some points in the epidemiology of malaria arising out of the study of the malaria epidemic in Ceylon in 1934-5. *Transaction of the Royal Society of Tropical Medicine and Hygiene*, **29**, 427–480.

Hardison, R.C. (2013) Function of hemoglobin and its genes. In: Hemoglobin and Its Diseases (ed. by D.J. Weatherall, A.N. Schechter & D.G. Nathan), pp. 49–66. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Higgs, D.R. & Gibbons, R.J. (2010) The molecular basis of alpha-thalassemia: a model for understanding human molecular genetics. *Hematology/Oncology Clinics of North America*, **24**, 1033–1054.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

van der Hoek, W., Konradsen, F., Perera, D., Amerasinghe, P.H. & Amerasinghe, F.P. (1997) Correlation between rainfall and malaria in the dry zone of Sri Lanka. *Annals of Tropical Medicine and Parasitology*, **91**, 945–949.

Karimi, M., Bagheri, M.H., Tahmtan, M., Shakibafard, A. & Rashid, M. (2009) Prevalence of hepatosplenomegaly in beta thalassemia minor subjects in Iran. *European Journal of Radiology*, **69**, 120–122.

Liu, Y.T., Old, J.M., Miles, K., Fisher, C.A., Weatherall, D.J. & Clegg, J.B. (2000) Rapid detection of alpha-thalassaemia deletions and alpha-globin gene triplication by multiplex polymerase chain reactions. *British Journal of Haematology*, **108**, 295–299.

Martinson, J.J., Swinburn, C., Excoffier, L., Harding, R.M., Boyce, A.J., Langaney, A. & Clegg, J.B. (1995) High diversity of a-globin haplotypes in a Senegalese population, including many novel variants. *American Journal of Human Genetics*, **57**, 1186–1198.

Meyer, E. (2006) Sri Lanka. Biography of an Island. Printed by Laballery, Clamecy, France, Viator Publications, Clamecy, France.

Olivieri, N.F., Muraca, G.M., O'Donnell, A., Premawardhena, A., Fisher, C. & Weatherall, D.J. (2008) Studies in haemoglobin E beta-thalassaemia. *British Journal of Haematology*, **141**, 388–397.

O'Riordan, S., Hien, T.T., Miles, K., Allen, A., Quyen, N.N., Hung, N.Q., Anh do, Q., Tuyen, L.N., Khoa, D.B., Thai, C.Q., Triet, D.M., Phu, N.H., Dunstan, S., Peto, T., Clegg, J., Farrar, J. & Weatherall, D. (2010) Large scale screening for haemoglobin disorders in southern Vietnam: implications for avoidance and management. *British Journal of Haematology*, **150**, 359–364.

Orkin, S.H., Kazazian, H.H., Antonarakis, S.E., Ostrer, H., Goff, S.C. & Sexton, J.P. (1982) Abnormal RNA processing due to the exon mutation of b^E-globin gene. *Nature*, **300**, 768–769.

O'Shaughnessy, D.F., Hill, A.V.S., Bowden, D.K., Weatherall, D.J. & Clegg, J.B.; with collaborators. (1990) Globin genes in Micronesia: origins and affinities of Pacific Island peoples. *American Journal of Human Genetics*, **46**, 144–155.

Premawardhena, A., Arambepola, M., Katugaha, N. & Weatherall, D.J. (2008) Is the beta thalassaemia trait of clinical importance? *British Journal of Haematology*, **141**, 407–410.

Rawlings, P., Curtis, C.F., Wickremasinghe, M.B. & Lines, J. (1981) The influence of age and season on dispersal and recapture of Anopheles culicifacies in Sri Lanka. *Ecological Entomology*, **6**, 307–319.

Reid, R.M. (1976) Effects of consanguineous marriage and inbreeding on couple fertility and offspring mortality in rural Sri Lanka. *Human Biology*, **48**, 139–146.

de Silva, S., Fisher, C.A., Premawardhena, A., Lamabadusuriya, S.P., Peto, T.E., Perera, G., Old, J.M., Clegg, J.B., Olivieri, N.F. & Weatherall, D.J. (2000) Thalassaemia in Sri Lanka: implications for the future health burden of Asian populations. Sri Lanka Thalassaemia Study Group. *Lancet*, **355**, 786–791.

Weatherall, D.J. & Clegg, J.B. (2001) The Thalassaemia Syndromes, 4th edn. Blackwell Science, Oxford.

Williams, T.N. & Weatherall, D.J. (2012) World distribution, population genetics, and health burden of the hemoglobinopathies. In: Cold Spring Harbor Perspectives in Medicine, 2nd edn. (ed. by D.J. Weatherall, A.N. Schechter & D.G. Nathan), pp. a011692. Cold Spring Harbor, Laboratory Press, New York, Cold Spring Harbor.

Williams, T.N., Maitland, K., Martin, P.M., Weatherall, D.J. & Clegg, J.B. (1998) Splenic size in homozygous alpha+ thalassaemia. *British Journal of Haematology*, **100**, 611–612.